ID N. 13

UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA

DEPARTMENT OF ENGINEERING

UNIVERSITA' DEGLI STUDI DI ROMA "TOR VERGATA"

DEPARTMENT OF BIOMEDICINE AND PREVENTION

Italian National Ph.D. in Artificial Intelligence
Health and Life Sciences
XXXVII Cycle

Multimodal brain decoding through deep learning

Supervisors

Nicola Toschi

Andrea Duggento

Candidate

Matteo Ferrante

01, 2024

University Campus Bio-Medico of Rome, Department of Engineering
University of Rome Tor Vergata, Department of Biomedicine and prevention

Multimodal brain decoding through deep learning

Matteo Ferrante

ID 13

Supervisor

Nicola Toschi

University of Rome, Tor Vergata

Health and Life Sciences

XXXVII Cycle

Academic Year 2024/2025

To my parents
and friends

Contents

Abstract

Understanding the neural mechanisms underlying cognition and perception is a fundamental pursuit in neuroscience. This thesis addresses the challenge of decoding brain activity to interpret and reconstruct human cognitive processes using functional magnetic resonance imaging (fMRI). By leveraging advancements in deep learning, we develop robust pipelines for decoding various sensory and cognitive modalities, including vision, language, music, and video.

Central to our approach is the alignment of brain representations with computational models, assuming brain representations can be mapped on vectorial spaces of pretraining multimodal models, enabling seamless mappings between neural data and external stimuli. Through the integration of encoding and decoding models, we explore tasks ranging from image reconstruction and cross-modal decoding to language generation and video retrieval.

Furthermore, this thesis delves into the concept of "brain algebra," examining how neural representations adhere to compositional and transformational principles akin to vector spaces. By perturbing brain activity in this high-dimensional semantic space, we uncover insights into the brain's capacity for concept manipulation and compositionality.

This work highlights the synergy between neuroscience and artificial intelligence, showcasing how multimodal, data-driven approaches can deepen our understanding of brain function and pave the way for innovative applications in brain-computer interfaces and cognitive modeling.

Glossary

- **fMRI:** Functional Magnetic Resonance Imaging, a non-invasive technique to measure brain activity based on changes in blood oxygenation levels.
- **EEG:** Electroencephalography, a technique for recording electrical activity of the brain with high temporal resolution.
- **MEG:** Magnetoencephalography, a technique that measures magnetic fields generated by neuronal activity.
- **BCI:** Brain-Computer Interface, a system that translates brain activity into external commands for communication or device control.
- **CLIP:** Contrastive Language-Image Pre-training, a multimodal model aligning text and image representations.
- **HRF:** Hemodynamic Response Function, a model describing the blood flow changes measured in fMRI.

Acronyms

AI: Artificial Intelligence
AUC: Area Under the Curve
BCI: Brain-Computer Interface
BERT: Bidirectional Encoder Representations from Transformers
BLEU: Bilingual Evaluation Understudy (metric for language tasks)
BOLD: Blood Oxygenation Level Dependent
BS: Batch Size
CEBRA: Contrastive Embedding for Brain Activity
CL: Contrastive Learning
CLIP: Contrastive Language-Image Pretraining
CNN: Convolutional Neural Network
COCO: Common Objects in Context (image dataset)
CV: Computer Vision
DALL·E: Generative AI Model for Image Synthesis
DINO: Self-Supervised Vision Transformer Model
DKL: Kullback-Leibler Divergence
DNN: Deep Neural Network
EEG: Electroencephalography
ELBO: Evidence Lower Bound
ERP: Event-Related Potential
FID: Fréchet Inception Distance (image evaluation metric)
FLIRT: FMRIB's Linear Image Registration Tool
FSL: FMRIB Software Library
GAN: Generative Adversarial Network
GELU: Gaussian Error Linear Unit (activation function)
GLM: General Linear Model
GPT: Generative Pre-trained Trans-

former
GPU: Graphics Processing Unit
HCP: Human Connectome Project
HRF: Hemodynamic Response Function
ICA: Independent Component Analysis
JEPA: Joint Embedding Predictive Architecture
LLM: Large Language Model
LFP: Local Field Potential
LSTM: Long Short-Term Memory
MNI: Montreal Neurological Institute (standard brain template)
MRI: Magnetic Resonance Imaging
MSE: Mean Squared Error
NLP: Natural Language Processing
NSD: Natural Scenes Dataset
OLS: Ordinary Least Squares
PCA: Principal Component Analysis
PPA: Parahippocampal Place Area
RAM: Random Access Memory
RMS: Root Mean Square
RMSE: Root Mean Squared Error
RNN: Recurrent Neural Network
ROI: Region of Interest
ROUGE: Recall-Oriented Understudy for Gisting Evaluation (language metric)
RSA: Representational Similarity Analysis
SNR: Signal-to-Noise Ratio
SVD: Singular Value Decomposition
SSIM: Structural Similarity Index Measure (image quality metric)
SSL: Self-Supervised Learning
STFT: Short-Time Fourier Transform
VAE: Variational Autoencoder
VAEGAN: Variational Autoencoder with GAN extension
VDVAE: Very Deep Variational Autoencoder
VI: Variational Inference
VIT: Vision Transformer
WER: Word Error Rate (language evaluation metric)
XCLIP: Cross-Modal CLIP Model

Part I

Introduction

Connecting minds and machines

One of the fundamental quests in neuroscience is to correlate neural activity with mental states and thus develop models able to map between external stimuli or internal processes and their neural representations [142, 135, 62, 182, 195, 134, 84, 185, 261, 100, 129].

It could be posited that neuroscience seeks to understand the "language" of the brain or thoughts [82], thus advancing our comprehension of the organ that produces the mind and connects us to the world, other people, and all living beings.

To achieve this, the brain can be studied by measuring electrical activity directly inside the brain with probes or by using noninvasive techniques that rely on neural correlates such as functional MRI, EEG or MEG. This, in turn, requires tools to interpret and analyze these data. Traditionally, such investigations have been performed at the population level [260, 200, 174, 32, 234, 252]. However, recent advances in technology, the availability of larger publicly accessible datasets, and the rise of AI-based methods have enabled a true paradigm shift. Deep analyses of brain activity are now feasible at the subject level, facilitating much more nuanced mappings between the external world and neural representations.

In this thesis, I present an extensive body of research on non-invasive decoding of brain representations of images, language, and music. The common thread across all the contributions presented is the idea that brain activity can be linked to latent representation of external stimuli obtained through pre-trained computational models, in turn enabling tasks such as neural encoding and decoding (see Fig. 1 for a visual representation of these concepts).

Understanding the mapping between neural activity and mental states is not just a matter of correlation; it is a critical step toward unraveling the fundamental principles of cognition. These insights have the potential to revolutionize brain-

computer interfaces (BCIs), enabling true mind-machine communication, and to inspire advances in AI by mimicking the computational strategies of the brain. This thesis explores these connections to contribute to both theoretical neuroscience and AI-based technological applications and propose of AI based pipelines for neuroscience research is one of the main contributions.

This approach falls under the umbrella of computational cognitive neuroscience [142]. Cognitive computational neuroscience is an interdisciplinary field that aims to understand the neural and computational mechanisms that underlie cognitive processes, such as perception, learning, memory, decision-making, and language. It combines insights from neuroscience, cognitive science, psychology, and AI to create computational models that simulate and explain how the brain processes information to generate cognition and behavior. By integrating methods and theories from all these disciplines, cognitive computational neuroscience addresses questions such as: *How do neural networks in the brain process and store information? What are the mechanisms of perception, learning, and memory in biological systems? How can insights from neural data improve artificial intelligence systems, and vice versa? How can we decode internal representations of the world from neural data?*

In our daily lives, we constantly receive a multitude of stimuli—visual, auditory, tactile, olfactory, and more—that are continuously processed by our brain. The brain, as the central control system, dictates how we react to external stimuli, making it essential to investigate the mechanisms underlying this complex processing. To understand how the brain performs this processing, which brain regions are involved in specific computations, and how these regions process on the input, we need to adopt systematic approaches.

A classical approach in neuroscience has been to examine behavioral data in controlled settings, often through psychological and psychophysical experiments. Such studies have been pivotal in understanding psychological processes and high-level input-output relationships at a population level. For example, behavioral experiments can reveal patterns in how individuals respond to stimuli, shedding light on the principles that govern perception and action [129, 127, 73, 246, 256, 253, 243]. However, this level of analysis does not localize the regions of the brain responsible for specific computations.

To identify the regions of the brain involved in specific computations, researchers have historically relied on studying pathological cases in which damage to specific areas of the brain results in functional impairments. These studies allow us to infer the importance of certain regions for particular types of computation [22, 51, 272]. Although insightful, this method provides only a coarse understanding of the functional architecture of the brain. Furthermore, evi-

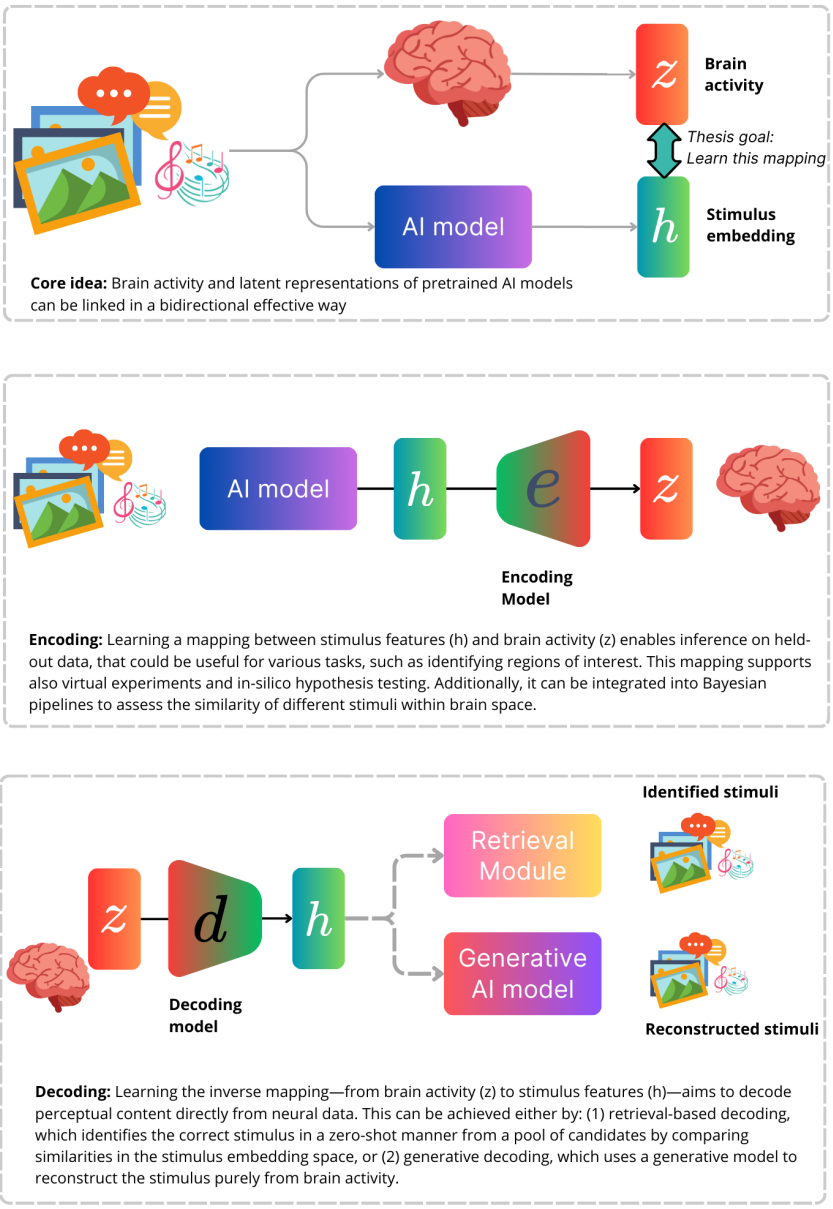


Figure 1: Illustration of the core ideas of this thesis—linking brain activity (z) and latent representations (h) from pre-trained AI models to study neural representations of external stimuli. The top panel highlights the goal of learning a bidirectional mapping between brain activity and computational embeddings. The middle panel shows Encoding Models, which map stimulus features (h) to brain activity (z). The bottom panel illustrates Decoding Models, which predict stimulus features (h) or reconstruct stimuli (x) directly from brain activity (z).

dence suggests that brain computation is not confined to isolated regions, but often involves networks of interconnected areas working together, either in serial or parallel, to process the same input and extract relevant information. [172]. When we dive deeper into the functional workings of the brain to address questions such as *'How do neural networks in the brain process and store information?'*, we require more sophisticated approaches. One powerful method is to measure functional brain activity, which is believed to reflect the computations required to process external information, and try to relate it with this information. By designing controlled experiments in which participants are exposed to specific stimuli (e.g., images, text, audio) or tasked with particular activities, we can collect paired datasets of stimuli and corresponding brain activity.

Initially, analyses may focus on aggregating data across many participants to perform statistical contrasts, revealing population-level patterns. Such studies have been instrumental in identifying functional areas that encode information about specific types of stimuli, such as faces, places, or the distinction between animate and inanimate objects [98, 100, 129]. This approach, grounded in robust and controllable methodologies, has led to numerous breakthroughs in neuroscience.

However, the complexity and variability of stimuli encountered in daily life far exceed what can be captured by these analyses. Humans effortlessly process this complexity, suggesting that the brain encodes highly detailed information about each type of stimulus it encounters. To uncover these intricate representations, computational neuroscience increasingly relies on **encoding models** [183, 14]. These models aim to reproduce the computational mechanisms by which the brain constructs, combines, and processes representations of the world.

Encoding models predict brain activity from specific stimulus features, enabling researchers to approximate the computations performed by the brain. This approach not only provides information on the functional architecture of the brain, but also establishes a connection between neuroscience and artificial intelligence, as both fields explore how information can be represented and processed efficiently (see Fig. 2 for a visual description of some main neuroscientific findings for different investigation modalities) .

0.1 Encoding

Consider an experiment where a subject is exposed to a naturalistic, complex stimulus x, and their brain activity is recorded using a neural recording technique such as fMRI, fNIRS, EEG, MEG, ECoG, or iEEG. To address fundamental questions like *"How do neural networks in the brain process and store information?"*

or *"What are the mechanisms of perception in biological systems?"*, we need to establish a relationship between x (the stimulus) and z (the measured brain activity). Specifically, by presenting diverse stimuli x, we aim to uncover which features of the stimuli are captured and processed by the brain.

For example, if x is an image, what aspects of it are processed by the brain? Does the brain primarily process low-level features, such as shapes, edges, colors, and contrasts? Does it also capture higher-level, abstract information, such as object semantics or relationships between elements in the scene? Intuitively, the brain likely processes all these aspects and possibly more, extracting salient information hierarchically or specialized across different regions [159].

To explore this hypothesis, we can extractfeatures h whcih represent the stimulus using a pretrained computational model. We hypothesize that the brain relates z and x through an unknown internal function $z = f(x)$, where f represents the neural mechanisms underlying the encoding of stimulus information. By approximating this relationship, we can fit an encoding model e, such that $z \approx e(x)$. This approach allows us to model how specific stimulus features are represented in the brain, providing insights into its functional organization and computational principles. Typically, we can explore various models m to extract features h from stimulus $h = m(x)$ and test different hypotheses by examining which encoding model best predicts brain activity $z \approx e(m(x))$. In other words, we would be building an encoding model on top of stimuli features such that $z \approx e(h)$ on held-out data. When an encoding model demonstrates a strong predictive power for brain activity, it provides evidence that the input characteristics h are relevant to the computations performed by the brain. This approach not only validates the significance of specific stimulus features but also offers insights into the brain's processing mechanisms. For example, if a model demonstrates strong predictive and generalization capabilities, it could serve as a foundation for virtual experiments, replacing costly initial exploratory experiments to test scientific hypotheses *in silico*. In addition, such models can be valuable tools for gaining deeper insights into brain processing. This can be achieved through explainability pipelines, which help identify the features most relevant to neural computations, or by integrating the model into Bayesian decoding pipelines. In the latter case, the model could rank potential candidate stimuli based on their similarity to true brain activity, aiding in the reconstruction of perceived or imagined stimuli [181, 183, 251].

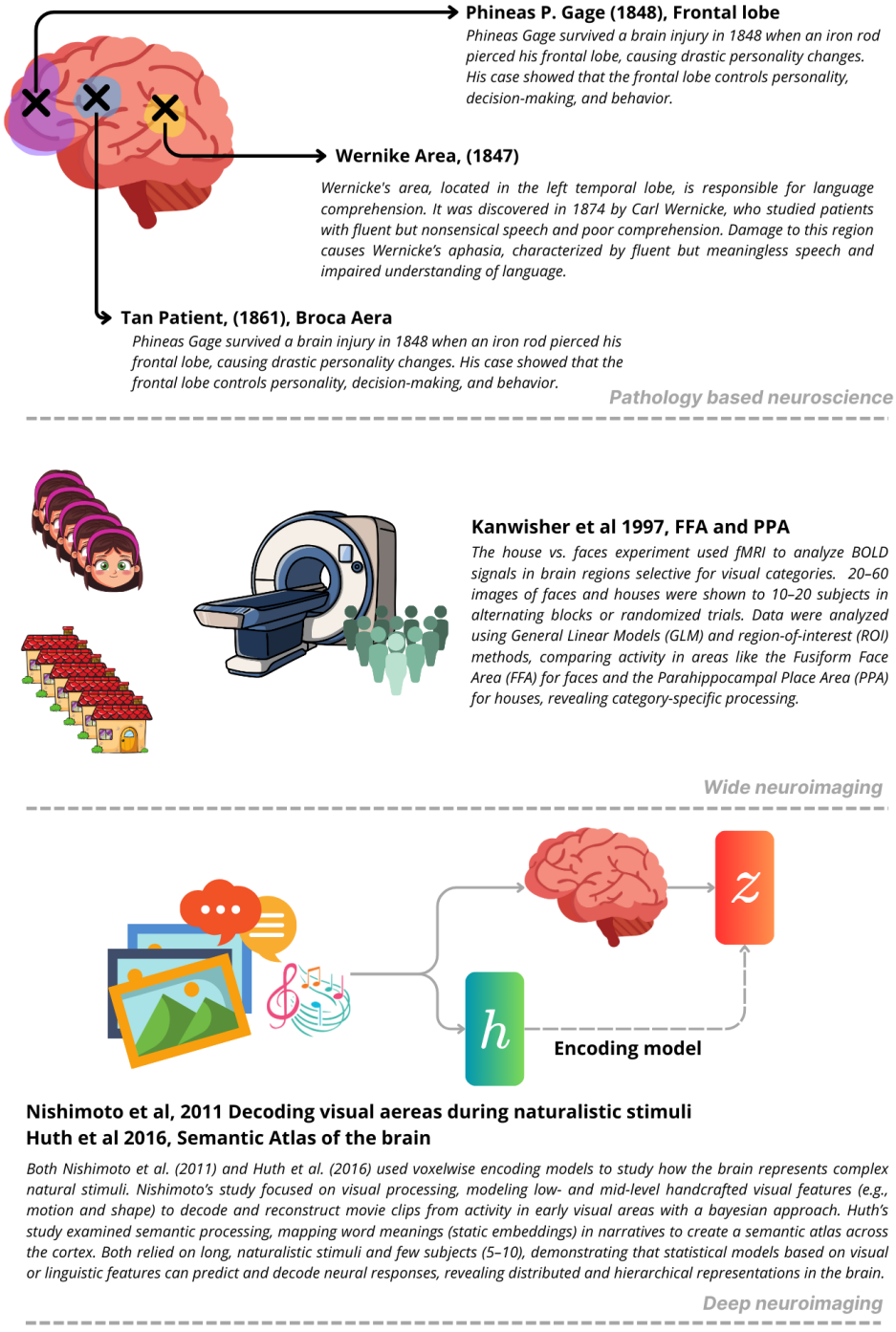


Figure 2: Non- exhaustive illustration of the evolution of computational neuroscience research, from pathology-based approaches like Phineas Gage (1848) and Broca's discoveries (1861) that linked brain lesions to specific functions, to wide neuroimaging studies such as Kanwisher et al. (1997) using fMRI and statistical modeling to reveal category-specific processing, and deep neuroimaging methods like Nishimoto et al. (2011) and Huth et al. (2016), which employed voxelwise encoding models to link stimuli features and analyze distributed neural representations of complex stimuli.

0.2 Decoding

Beyond fundamental science, deciphering how the brain processes external stimuli has substantial technological implications. If we can uncover these internal mappings and use them for **decoding**, i.e. translating brain activity into external representations of stimuli, we could develop brain-computer interfaces (BCIs) that connect minds and machines. Such technology holds the promise of assisting individuals with disabilities or impairments by enhancing accessibility and restoring visual, language, movement, or ability to control external devices. In essence, BCIs have the potential to restore the capacity to interact with the world when it has been diminished or lost due to external factors or even enhance human abilities in healthy subjects. Decoding refers to the process of translating patterns of brain activity z into external representations of stimuli $\hat{x}$. This approach complements encoding by aiming to reverse-engineer how information is processed in the brain. decoding is driven by a fundamental question: *How can we decrypt and interpret the representation of the external world from neural data?*

Decoding models operate by learning a mapping from the neural activity space z to a feature space h, or directly to an approximation or reconstruction of the original stimulus x, which we can call $\hat{x}$. Mathematically, decoding involves approximating an inverse function $\hat{x} = d(z)$, where d represents the decoding model. This mapping can be learned using various techniques, such as regression-based methods, classification frameworks, or generative approaches. In classification-based decoding, brain activity z is mapped to discrete categories y that represent the type or class of the stimulus x. For example, given neural data, a decoder might predict whether a subject was viewing an image of a face or a place. Such methods have been successfully used in studies in which stimuli belong to well-defined categories, such as emotional states, object categories, or types of motion [116, 79].

Identification decoding involves matching neural activity z to one of a set of known stimuli. Here, the decoder identifies the stimulus x from a predefined library of options based on its similarity to brain activity. This approach has been widely applied in experiments using controlled stimulus sets, such as visual or auditory stimuli, where the task is to determine which specific stimulus corresponds to the observed brain activity.

Reconstruction decoding aims to recreate a detailed representation of the original stimulus $\hat{x}$, such as an image, a sound, or a continuous sequence, based on brain activity z. This technique is particularly challenging, as it requires the capture of both low- and high-level features encoded in the neural data. Recent advances in generative models and pre-trained feature extractors (e.g.,

deep learning models like Stable Diffusion or CLIP for vision) have significantly improved the fidelity of neural activity reconstructions [49].

Decoding models are typically trained using paired stimuli data x and recorded brain activity z. The training process involves optimizing a model to minimize the difference between the decoded output $\hat{x}$ (or its features $\hat{h}$) and the actual stimulus x (or actual features h), often using techniques like regression, Bayesian inference, or deep learning. For example: Regression-based models aim to predict continuous features h of x directly from the neural activity z, providing a straightforward approach to mapping stimuli to brain activity. Bayesian models, on the other hand, incorporate prior knowledge about the stimulus space to guide the decoding process, enhancing robustness and performance, particularly in scenarios where data are limited [181], often reconstructing stimuli $\hat{x}$ such as $p(\hat{x}|z)$ is maximized. Given the complexity of this task, usually a bayesian decomposition of the problem is chosen by computing this probability as $p(z|x)p(x)$, where $p(z|x)$ is an encoding model which is usually easier to train and $p(x)$ a generative model for stimuli. In this context, deep learning models use neural networks to learn complex, non-linear relationships between z and $\hat{h}$, making them well suited for the reconstruction of high-dimensional stimuli with complex semantic features.

So, decoding models provide a powerful framework for investigating brain function. Decoding has been used to study vision, language, and even imagination, revealing how different regions of the brain contribute to these processes [185, 14]. Decoding also plays a central role in BCIs, enabling direct communication between the brain and external devices. For example, decoding models can be used to translate brain activity into commands for prosthetic limbs, communication aids, or even text generation for individuals with speech impairments [241]. Decoding models with high accuracy and real-time capabilities are essential for making BCIs practical and accessible.

Although decoding has achieved significant progress, several challenges remain. One major hurdle is the high dimensionality of both brain activity z and stimuli x, which requires the use of sophisticated models and large datasets to effectively capture their mappings. Additionally, neural representations can vary substantially between individuals, leading to intersubject variability that complicates the generalization of decoding models. The inherent noise and artifacts in neural data further exacerbate these challenges, requiring robust preprocessing techniques to ensure the reliability and accuracy of the decoding outcomes. In addition, as decoding technologies advance, ethical considerations surrounding privacy, consent, and potential misuse of neural data become increasingly critical and must be addressed proactively.

Despite these obstacles, decoding models hold tremendous potential for deepening our understanding of the brain and enabling transformative applications, from brain-computer interfaces to virtual experiments and beyond.

0.3 Thesis organization and contribution

The primary focus of this work is on functional magnetic resonance imaging (fMRI) data due to its non-invasive nature, high spatial resolution, and ability to capture whole brain activity. In recent years, a paradigm shift has been observed in fMRI studies, transitioning from *wide fMRI* approaches (characterized by many subjects and few samples per subject) to *deep fMRI* studies (fewer subjects, often fewer than ten, each exposed to a large number of stimuli, allowing the creation of subject-specific encoding and decoding models. Furthermore, this shift has extended to *intensive fMRI*, a broader approach in which stimuli are designed to span a wide hypothesis space, generating extensive data that can help to address multiple scientific questions[144] .

The central aim of this thesis is to develop AI pipelines for decoding brain activity across various human cognitive and perceptual tasks, including vision, language comprehension, video processing, and music perception. The chapters in this thesis reflect the chronological progression of this rapidly evolving field. In most cases, the goal is multifaceted: On the one hand, the objective is to demonstrate the feasibility of decoding meaningful information from brain activity recorded through fMRI. On the other hand, a significant focus is placed on developing novel pipelines, exploring cross- and multimodal representations, and investigating how brain representations are encoded. In addition, this research seeks to uncover potential neuroscientific insights by combining advanced AI analyses with large-scale neural data.

The common thread is the idea that the brain processes semantic representations of the external world in a manner strikingly similar to computational models. This similarity allows us to approximate the two processes as homomorphic, enabling a direct mapping between brain representations and model representations. This mapping can often be achieved using a simple function, such as a linear or nonlinear model, trained in a data-driven manner. Specifically, we propose that the model representations can be obtained as $\tilde{h} = d(z)$, where d is the decoding model. This approach assumes that the brain performs extensive non-linear transformations on the stimuli it receives, much like computational models do. Once the brain has processed these stimuli into sufficiently abstract and high-level representations, mapping between the brain's representations and the model's becomes simpler and more effective (see fig. 1). Furthermore,

these representations tend to be more compact and organized than the stimuli in their original space, reflecting the brain's ability to distill complex inputs into meaningful, structured forms under this assumption of similarity between the brain and computational representation processing. [201].

Typically, the decoding pipeline involves two stages. The first stage learns a mapping from brain activity to the model feature space. The second stage depends on the specific decoding objective. If the goal is stimulus identification, the pipeline compares the estimated features to the ground truth features, allowing for the identification of the most likely stimulus. In this case, the decoded stimulus x_i can be expressed as:

$$x_i = \arg\max_j \text{sim}(\tilde{h}_i, H)$$

where $\tilde{h}_i$ represents the brain-estimated features derived from a sample z_i, and $H = [h_j]$ is a dataset of ground-truth characteristics of candidate stimuli x_j, with $h_j = m(x_j)$ being the features extracted from a computational model m.

Alternatively, if the focus is on stimulus generation or reconstruction, the second stage involves a generative model g conditioned on the estimated features $\tilde{h}$. In this case, the reconstructed stimulus can be expressed as:

$$\tilde{x}_i = g(\tilde{h}_i)$$

This generative approach enables the synthesis of stimuli that resemble the original inputs, providing a detailed visualization representation of the decoded brain activity and ability to generalize better since we are not restricted to a set of pre-defined pool of candidates.

The thesis is structured as follows (See Fig 3 for a visual scheme of chapters and leading scientific questions and Table 1 for a textual descriptions of main contributions of each chapter): Part II focuses on decoding visual stimuli, specifically images. Vision was chosen as the starting point for this exploration due to its historical importance in neuroscience research and the availability of data during the early stages of this work.

Chapter 1 focuses on semantic decoding using the Generic Object Decoding (GOD) dataset [116]. In this dataset, five subjects viewed 1,200 ImageNet images from 150 distinct semantic categories while undergoing fMRI recordings. To evaluate the model's ability to generalize beyond the classes used during training, 50 images—each representing a different novel category—were set aside as the test set, posing the challenge of out-of-distribution classification.

Although much of the previous literature has concentrated on pixel-level image reconstruction, this chapter takes a semantic-based approach [188], empha-

Figure 3: Structure of this thesis and key research questions. The work is organized into multiple thematic areas, each addressing specific scientific challenges in decoding brain activity. Starting with semantic brain decoding (1), the work explores decoding visual information, cross-modal decoding (2), and cross-subject decoding (3) to overcome individual variability. Multimodal representation alignment (4) examines aligning neural data from different neuroimaging techniques. Language decoding is investigated through contrastive learning approaches (5) and direct fMRI-based language generation using large language models (6). Extensions to other modalities include decoding musical tracks (7) and video identification (8) from brain activity. The final section (9) focuses on brain algebra, exploring how semantic representations are organized in the brain and whether they follow an algebraic framework. Each area is represented by its corresponding modality: images (part I), language (part II), music and videos (grouped together in part III), and compositionality (part IV).

sizing the classification of unseen semantic classes and demonstrating strong generalization capabilities. We implemented both semantic reconstruction and retrieval tasks, using features from convolutional neural networks (CNNs) and vision transformers, plus a k-nearest neighbors algorithm to perform this classification and condition a generative model.

Chapter 2 builds on vision work by exploring cross-modal decoding using the NSD dataset [5]. This chapter focuses on decoding both language and low-level image features from brain activity and uses these decoded features to condition diffusion models for generating final images. The NSD data set offers significant advantages over previous datasets, with approximately 10,000 stimuli per subject, fMRI data acquired at 7T (as opposed to 3T), and stimuli drawn from the COCO image dataset, which includes images with more complex and diverse contextual elements. Due to these strengths, the NSD dataset serves as the basis for subsequent work involving image-related decoding tasks.

Chapter 3 addresses a critical challenge in fine-grained decoding: Functional differences between subjects. Up to this point, all decoding pipelines were subject-specific, as inter-individual differences at the fine-grained voxel level outweighed similarities. This limitation highlighted the need for methods to bridge these individual variations. In this chapter, we introduce a Ridge Regression-based functional alignment approach to harmonize brain activity between subjects. This method demonstrates that it is possible to decode visual perception from one individual's brain activity using data from another, while requiring only 10% of the original data typically needed to train the decoding model. Remarkably, this reduction in training data comes without a significant loss in decoding performance or image reconstruction quality, showcasing the potential of cross-subject decoding to significantly enhance the efficiency and scalability of decoding pipelines.

Chapter 4 extends the scope of decoding to other neuroimaging modalities including fMRI, EEG, and MEG during vision tasks, presenting a unified model capable of handling multiple tasks, such as encoding, decoding, and modality conversion, through a single framework based on contrastive learning.

Part III shifts the focus to decoding language representations. This section highlights how brain activity can be used to decode language, leveraging both retrieval-based and generative decoding approaches. Chapter 5 showcases a retrieval-based method, which employs contrastive learning architectures to decode fragments of sentences. Chapter 6 takes a more ambitious step by introducing a generative approach, aiming to move toward non-invasive direct brain-computer interfaces (BCIs). This involves adapting and fine-tuning large language models (LLMs) to work with fMRI data, paving the way for context-

aware language decoding and sentence reconstruction. Both approaches emphasize the critical role of language encoding models in identifying brain regions involved in language processing.

Part IV explores the extension of these methods to other types of stimuli, such as music (Chapter 7) and videos (Chapter 8). In these works, we build on techniques developed in earlier chapters, including functional alignment, region-of-interest (ROI) selection, direct ridge mapping between brain and model representations, and retrieval-based decoding. These studies demonstrate the versatility of the proposed methods and their applicability in diverse sensory modalities.

Part V moves beyond decoding per se and explores how decoding can be used as a lens to better understand brain function and representation organization. This section introduces the concept of brain algebra, where the focus shifts to how perturbations or manipulations of brain activity relate to cognitive processes and representations. By using decoding as a tool, we examine how the brain processes, combines, and transforms information, offering novel insights into the brain's functional architecture. The final Appendix A explores how the methods and considerations presented in this work can be extended to other neuroimaging modalities, such as EEG, for decoding semantic visual stimuli. It highlights the need for specialized architectures tailored to address the domain-specific challenges of EEG data as opposed to fMRI. In particular, this section builds upon the objectives outlined in Chapter 1, focusing on semantic image reconstruction. A knowledge distillation pipeline is employed to guide the EEG decoder in learning CLIP-like representations, enabling effective semantic decoding.

The next sections of the introduction provide foundational, high level context for the work presented in detail in subsequenc chapters. These include a primer on neuroimaging techniques (Section 0.4), an overview of AI models used in decoding (Section 0.5), and a brief review of relevant literature (Section 0.6). This background serves to situate the contributions of this thesis within the broader fields of artificial intelligence and neuroscience.

0.4 NeuroImaging

The human brain encodes information through complex neural processes involving individual neurons and populations of neurons. Neurons communicate via electrical impulses, or spikes, that reflect their activation. These activations collectively encode sensory input, thoughts, actions, and memories. One of the fundamental questions in neuroscience is how this activity can be mea-

sured and interpreted to understand the mechanisms underlying brain function. Specifically, the debate revolves around whether spikes, firing rates, population patterns, or a combination of these features hold the key to decoding neural information.

To address this question, a variety of neuroimaging techniques have been developed. These techniques can broadly be categorized into invasive and non-invasive methods. Noninvasive approaches, such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), magnetoencephalography (MEG), and functional near-infrared spectroscopy (fNIRS), enable the study of brain activity without physical access to the brain itself. In contrast, invasive methods, including electrocorticography (ECoG), intracranial EEG (iEEG) and stereoelectroencephalography (sEEG), require direct interaction with brain tissue, offering high-resolution insights, but necessitating surgical intervention [121, 48] (see Fig. 4 for a quick overview of the most common neuroimaging modalities).

Among non-invasive techniques, fMRI has become one of the most widely used due to its ability to measure whole-brain activity with high spatial resolution. fMRI indirectly captures neuronal activity by detecting changes in blood oxygenation levels, known as BOLD signals. When neurons become active, local changes occur in the ratio of oxygenated to deoxygenated hemoglobin, which can be measured using the magnetic properties of blood [187, 121]. fMRI's primary strengths lie in its non-invasive nature, millimeter-scale spatial resolution, and comprehensive brain coverage. However, it suffers from low temporal resolution, as BOLD signals lag behind neuronal activity by several seconds, and it provides only an indirect measure of neural activity.

EEG, on the other hand, measures electrical activity from the scalp using electrodes and provides a direct window into the rapid dynamics of brain activity. It captures the combined activity of large populations of neurons, particularly in cortical regions, with millisecond temporal resolution [48]. While EEG excels at capturing fast neural processes, its spatial resolution is limited due to the blurring effects of the skull and scalp, and it primarily detects activity from cortical regions of the brain. Moreover, EEG is extremely subject to artifacts and usually lead to very noisy recordings. Despite these limitations, EEG remains a cornerstone of neuroimaging, particularly for studying time-sensitive cognitive processes such as attention and perception.

MEG complements EEG by measuring the magnetic fields generated by neuronal currents. Because magnetic fields are less distorted by the skull than electrical signals, MEG offers another kind of signal while maintaining the millisecond-level temporal precision of EEG. However, MEG systems are costly

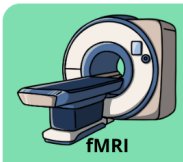
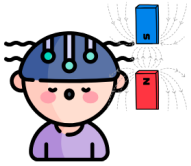
	What is measured?	Pro	Cons
 fMRI	*Blood-oxygen-level-dependent (BOLD) signals, reflecting neural activity indirectly via blood flow*	*High spatial resolution (1-3 mm); non-invasive.*	*Poor temporal resolution (0-1 Hz); expensive and immobile, neural signal response is delayed and spread in time due to hemodynamic response function.*
 EEG	*Electrical activity from neural populations via scalp electrodes.*	*High temporal resolution (1000 Hz); portable, non-invasive and inexpensive.*	*Poor spatial resolution (few cm); highly susceptible to noise and artifacts*
 MEG	*Magnetic fields generated by neural electrical currents*	*High temporal resolution and better spatial resolution than EEG; non-invasive.*	*Expensive, requires magnetically shielded rooms, and limited to superficial brain activity.*
 invasive EEG	*Electrical activity directly from the brain using implanted electrodes.*	*Excellent spatial and temporal resolution; highly precise for localization.*	*Invasive procedure, limited to clinical use cases (e.g., epilepsy monitoring).*

Figure 4: Comparison of neuroimaging modalities—fMRI, EEG, MEG, and invasive EEG—highlighting what each technique measures, their advantages (Pros), and limitations (Cons). These methods provide complementary insights into brain activity, balancing trade-offs between spatial and temporal resolution, invasiveness, and practical considerations.fMRI is highlighted since this is the main modality investigated for this thesis work.

and require magnetically shielded environments, which limit their accessibility. Like EEG, MEG is primarily sensitive to cortical activity, making it a valuable tool for investigating fast neural dynamics but less effective for studying deeper brain structures.

fNIRS is a more portable and accessible noninvasive technique that measures changes in blood oxygenation by emitting near-infrared light into the scalp and detecting the reflected light. Like fMRI, it relies on hemodynamic responses, making it useful for studying brain function in naturalistic and ecologically valid settings [55]. However, fNIRS has limited spatial resolution and cannot measure activity beyond the cortical surface, which restricts its applicability. Despite

these constraints, its low cost and portability make it increasingly popular for studies with special populations, such as infants and older adults. However, large scale datasets of naturalistic stimuli are not yet available to the public, limiting for now, it's use within the scope of this thesis.

Invasive methods, including ECoG, iEEG, and sEEG, provide unparalleled resolution by directly recording neuronal activity. ECoG involves placing electrodes on the cortical surface, while iEEG and sEEG use depth electrodes to measure the activity of specific brain regions, including subcortical structures [241]. These techniques offer high spatial and temporal resolution, making them ideal for studying fine-grained neural processes. However, their invasive nature restricts their use to clinical settings, typically in patients undergoing treatment for epilepsy or other neurological conditions. These methods are often used to validate findings from noninvasive studies and provide ground truth data for decoding models.

Each neuroimaging modality comes with trade-offs between spatial resolution, temporal resolution, and invasiveness. By integrating data across multiple techniques, researchers can gain a more complete understanding of brain function.

0.5 AI models

Artificial intelligence (AI) models form the foundation for many tasks in the encoding or decoding of neural activity, providing tools to analyze, model, and approximate complex data and relationships. This section introduces the key concepts, training paradigms, and architectures relevant to the work in this thesis. Fig 5 shows the main training strategies and models that are explored in the various chapters of this thesis to connect the brain and the stimuli representations.

0.5.a Training AI Models

AI models are a family of parametric functions that approximate unknown relationships in a data-driven way. They are typically trained to minimize a loss function $\mathcal{L}$ that measures the difference between the predicted output of the model and the ground truth. The optimization problem can be expressed as:

$$\hat{\theta} = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f_\theta(x), y)],$$

where $\mathcal{D}$ is the dataset, x is the input, y is the target output and $f_\theta(x)$ is the model parameterized by θ. Depending on the nature of y, training paradigms

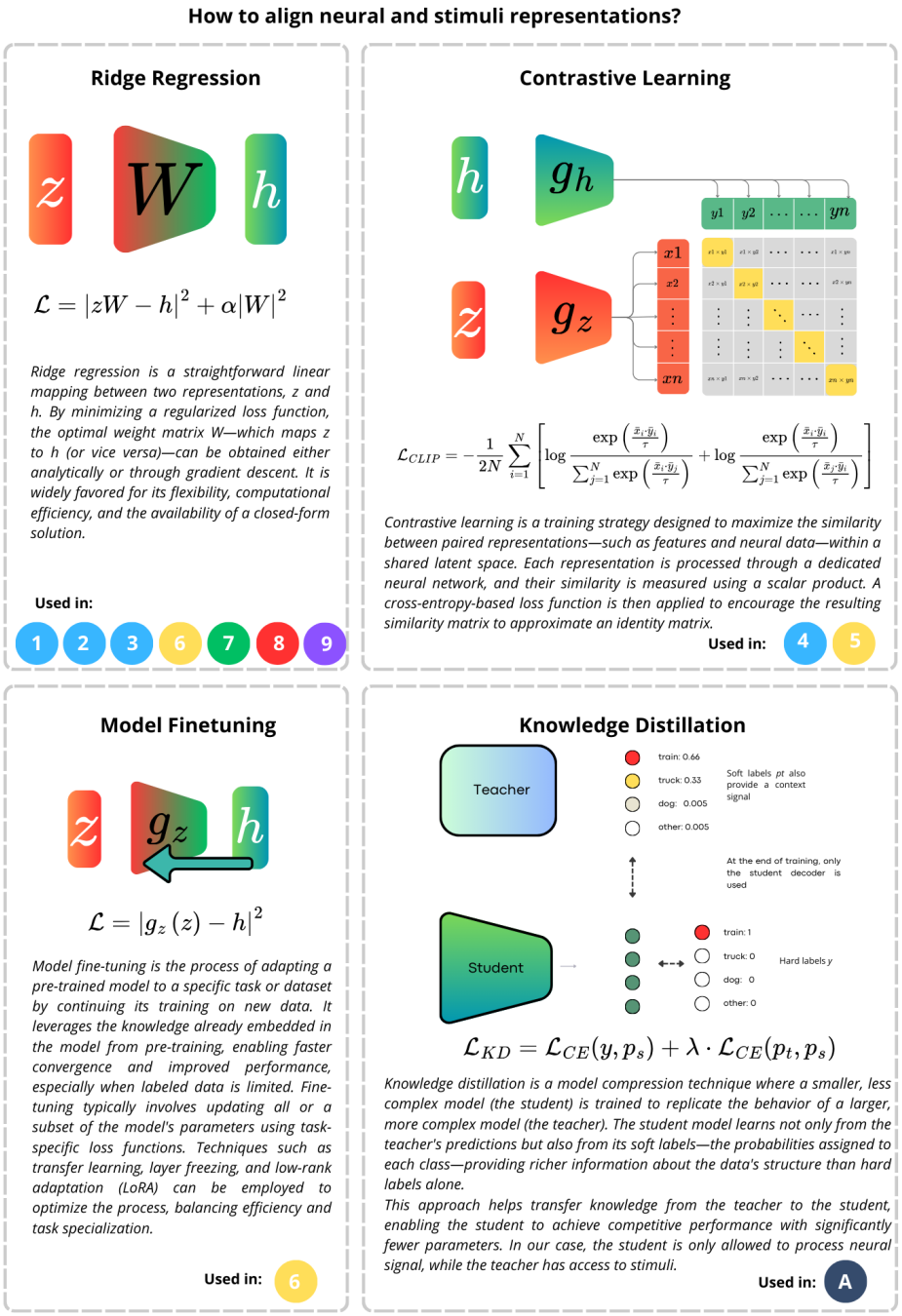


Figure 5: This figure outlines key methods used in this thesis. Ridge Regression provides efficient linear mapping for high-dimensional data. Contrastive Learning aligns neural and stimulus features in a shared latent space by maximizing similarity between pairs. Model Fine-Tuning adapts pre-trained models to new tasks, and Knowledge Distillation transfers knowledge from a larger model (teacher) to a smaller one (student) for efficient decoding. These methods support encoding, decoding, and multimodal alignment pipelines.

can be classified as follows:

Supervised Learning

In supervised learning, y represents labels or ground truth values. Thus, there is the need to have access to pairs of data (x, y) (i.e. we need some examples of input-output pairs that we can leverage to infer the relation between the two).

The model learns a mapping $f_\theta : \mathcal{X} \to \mathcal{Y}$ by minimizing a task-specific loss, such as the **Mean squared error** for regression tasks:

$$\mathcal{L}(f_\theta(x), y) = \|f_\theta(x) - y\|_2^2,$$

Or, e.g., a Cross-entropy loss for classification tasks:

$$\mathcal{L}(f_\theta(x), y) = -\sum_{i=1}^{C} y_i \log(f_\theta(x)_i),$$

where C is the number of classes. Minimizing task-specific loss functions involves adapting the model to effectively solve the given task by reducing errors, either in number (for classification tasks) or in magnitude (for regression tasks). The model f_θ, can take various forms, such as a linear layer, a decision tree, or a neural network. In general, f_θ represents any function capable of approximating the mapping between inputs x and outputs y, parameterized by θ. Through an iterative optimization process, the parameters θ are updated to minimize the loss function chosen, allowing the model to better align its predictions with the ground truth and improve performance in the task at hand.

Unsupervised Learning

Unsupervised learning aims to uncover the underlying structure of data without requiring explicit labels or predefined targets. This paradigm focuses solely on a data set of input characteristics x and seeks to identify meaningful patterns, relationships, or representations within the data. Unlike supervised learning, where the objective is to map inputs to corresponding outputs, unsupervised learning relies entirely on intrinsic properties of the data, making it especially valuable in domains where labeled data is scarce or unavailable.

One of the primary goals of unsupervised learning is to learn meaningful representations of stimuli without the need for external labels. These representations tend to be useful also for clustering, where the objective is to group data points into clusters such that points within the same cluster exhibit higher similarity compared to those in different clusters. Mathematically, clustering can

be expressed as an optimization problem. For example, in k-means clustering [163], the goal is to minimize the within-cluster variance:

$$\min_{C_1,\ldots,C_k} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2,$$

where C_i represents the i-th cluster and μ_i is its centroid. Advanced methods, such as hierarchical clustering and spectral clustering, go beyond simple distance-based metrics by leveraging hierarchical tree structures or spectral embeddings derived from similarity graphs.

Another key objective of unsupervised learning is dimensionality reduction, where high-dimensional data are projected onto a lower-dimensional space while preserving its most salient features. Principal Component Analysis (PCA) [197, 117, 126] is classic approach which identifies orthogonal directions (principal components) that maximize the variance in the data in that direction. The PCA optimization minimizes the reconstruction error:

$$\mathcal{L}_{\text{PCA}} = \|x - WW^\top x\|_2^2,$$

where W is a projection matrix composed of the top k eigenvectors of the covariance matrix of x where k can be at maximum equal to the rank of the matrix W. PCA is effective for linear transformations, but more complex relationships in data often require non-linear techniques such as t-distributed stochastic neighbor embedding (t-SNE) [259], which preserves local relationships in the data, making it particularly useful for visualization tasks.

Additionally, unsupervised learning includes methods for density estimation, which model the probability distribution $p(x)$ of the data. For instance, Gaussian Mixture Models (GMMs) approximate $p(x)$ as a weighted sum of Gaussian components [219]:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k),$$

where π_k, μ_k, and Σ_k represent the mixture weights, means, and covariances of the components, respectively. Such models enable tasks like anomaly detection, where low-probability regions in $p(x)$ indicate potential outliers.

Modern advances in unsupervised learning have introduced neural network-based techniques, such as auto-encoders [227, 150, 21, 110], which are designed to learn compressed representations of data. These models consist of an encoder f_θ, which maps the input x to a latent representation z, and a decoder g_ϕ, which reconstructs x from z. The objective of an autoencoder is to minimize the

reconstruction loss:

$$\mathcal{L}_{\text{AE}} = \|x - g_\phi(f_\theta(x))\|_2^2.$$

Variations such as denoising autoencoders[18], which reconstruct x from a corrupted version $\tilde{x}$, enhance robustness by encouraging the model to learn meaningful representations.

Another significant advancement in the evolution of autoencoders is the probabilistic variant known as the Variational Autoencoder (VAE) [136]. Unlike traditional autoencoders, VAEs model the data distribution through latent variables, enabling a probabilistic interpretation of the latent space. The training objective of a VAE is to maximize the evidence lower bound (ELBO), expressed as:

$$\mathcal{L} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{\text{KL}}(q_\phi(z|x)\|p(z)),$$

where $q_\phi(z|x)$ is the encoder, which approximates the posterior distribution of the latent variables, $p_\theta(x|z)$ is the decoder, which reconstructs the input data, and D_{KL} represents the Kullback-Leibler divergence between the approximate posterior $q_\phi(z|x)$ and the prior $p(z)$, often chosen as a standard Gaussian distribution.

Through this formulation, each input sample x is mapped to a distribution in the latent space rather than a single point, ensuring organization and continuity in the latent space during training. This structured latent space allows for smooth transitions between points, enabling the generation of new samples and interpolation between data points. By traversing specific directions in the latent space, one can often uncover meaningful properties or variations in the data x. This ability to explore and manipulate the latent space makes VAEs a powerful tool for understanding and leveraging the learned representations.

Unsupervised learning provides a foundational framework for many downstream tasks, providing insight into the structure of data without reliance on external labels. Its flexibility and adaptability make it a crucial tool for exploratory data analysis and representation learning. These tools are used throughout the thesis, especially in Part II where diffusion models rely on VAEs to reduce the computational complexity of the diffusion process in the latent space, clustering algorithms are used to identify semantic clusters and perform pseudo-labeling of datasets.

Self-Supervised Learning, Masked autoencoders and multimodal representation alignment

Self-supervised learning (SSL) bridges the gap between unsupervised and supervised learning by leveraging the inherent structure of data to create pseudo-

labels, enabling models to learn meaningful representations without the need for manual annotations. This paradigm opens up new possibilities for pretraining models, particularly in domains where labeled datasets are scarce or expensive to generate.

A prominent framework within SSL is **contrastive learning**, which aims to structure the representation space so that similar data points (positive pairs) are brought closer together, while dissimilar data points (negative pairs) are pushed apart. This is achieved through a contrastive loss function defined as:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(z_i, z_k)/\tau)},$$

where z_i and z_j represent the latent embeddings of a positive pair (e.g., augmented versions of the same image, text or any kind of input), sim is a similarity metric (commonly cosine similarity), τ is a temperature parameter controlling the sharpness of the distribution and N is the batch size. The numerator strengthens the similarity between positive pairs, while the denominator ensures that negative samples are contrasted effectively. In the context of this work, contrastive learning is explored as a tool to align neural and stimuli representations of images and text in Chapters 4 and 5.

Self-supervised learning (SSL) methods often involve auxiliary tasks, known as pretext tasks, which are specifically designed to help models extract general-purpose features from data. These tasks encourage the model to focus on learning meaningful and transferable representations without requiring manual annotations. One prominent example is context prediction, where the model learns to predict the spatial relationships between patches of an image, as introduced by Doersch et al. [63]. Another widely used pretext task is masked prediction, where parts of the input are intentionally masked and the model is trained to predict missing content. This approach has been effectively applied in text modeling with BERT [60], where masked tokens in a sentence are reconstructed, and in vision tasks with Masked Autoencoders (MAE) [102], where random patches of an image are hidden and reconstructed.

In particular, masked autoencoders (MAEs) have been proven to be highly effective in self-supervised learning (SSL). By masking a significant portion of the input data (e.g., random patches in an image or words in a sentence) and tasking the model with reconstructing the missing content, MAEs force the network to learn robust and generalized features. This approach is closely aligned with traditional autoencoders, but introduces masking as a mechanism to improve data efficiency, robustness, and the model's ability to capture meaningful patterns in the input space [103, 59].

Recently, the masked autoencoder approach has been scaled to train foundation models in a wide range of modalities and tasks. In the domain of vision, models such as DINO [29] and MAE [103] have demonstrated the effectiveness of masked prediction in learning general-purpose visual characteristics. These models have been pivotal in the advancement of SSL for image recognition, segmentation, and other vision-related tasks, achieving performance competitive with supervised learning on large-scale benchmarks.

In natural language processing (NLP), masked autoencoder-based models like BERT [59], BART [154], and recent large language models such as GPT [23] and LLaMA [255] have revolutionized the field. These models are trained using masked prediction tasks, where certain tokens in the input sequence are masked and the model is tasked with reconstructing them. This approach not only enables efficient learning of linguistic patterns but also allows these models to generalize effectively to downstream NLP tasks, such as question answering, summarization, and translation. These kind of models where particular useful to extract useful embedding and stimulus representations for language or images such as done in Chapters 2,5,6, and 1.

Extending beyond text and vision, similar ideas have been applied to other modalities, such as audio and multimodal tasks. For example, models like BEATs [41] and SpeechMAE [120] leverage masking in the audio domain to learn representations for speech recognition, speaker identification, and emotion detection. In multimodal learning, approaches such as Joint Embedding Predictive Architectures (JEPA) and Vision-JEPA (V-JEPA) [149] have emerged, which extend the concept of masked prediction to align representations across different data modalities, such as vision, text, and audio.

This scaling of masked autoencoder-based approaches highlights their flexibility and adaptability in learning foundational representations across diverse data types. These methods have established themselves as cornerstone techniques in the development of modern SSL pipelines, pushing the boundaries of what can be achieved with unlabeled data.

Beyond single-modality tasks, SSL has been extended to multimodal learning, where the objective is to learn shared representations across different data modalities. Given two modalities, x_1 and x_2, the goal is to learn a common latent space representation h such that:

$$h = f_\theta(x_1) = g_\phi(x_2),$$

where f_θ and g_ϕ are modality-specific encoders. By aligning the representations from different modalities, the model facilitates cross-modal tasks such as text-to-image translation, speech-to-text conversion, or brain-to-image decoding [209].

A notable example of multimodal alignment is CLIP [209], which trains on pairs of text and images to learn a shared embedding space. This alignment enables zero-shot learning, where a model can perform tasks it has not explicitly been trained on by leveraging relationships between modalities.

SSL and multimodal learning are integral to the development of modern machine learning pipelines. They enable the use of vast amounts of unlabeled data and foster the creation of representations that generalize effectively across tasks and modalities. Their success in pretraining models for downstream applications has revolutionized fields such as natural language processing, computer vision, and neuroscience, bridging gaps between modalities and advancing the frontier of AI research. In particular, CLIP has shown very strong capabilities when it is combined with linear models to predict brain activity and it is used in Part II, IV and V of this thesis.

0.5.b Linear Models and Ridge Regression

Linear models are foundational tools in statistical learning, especially for decoding tasks where simplicity, interpretability, and computational efficiency are prioritized. Ridge regression, a widely used linear model, extends ordinary least squares (OLS) regression by incorporating regularization, addressing multicollinearity and overfitting in high-dimensional datasets. As we will see later on in various chapters and in the discussion section, linear models demonstrate superior performance in mapping stimulus representations to and from brain activity measured via fMRI, making them a simple yet effective tool for brain encoding and decoding.

Ordinary Least Squares (OLS)

In the standard OLS framework, our objective is to find a linear mapping β between predictors X and responses Y by minimizing the residual sum of squares:

$$\mathcal{L}_{\text{OLS}} = \|Y - X\beta\|_2^2,$$

where $X \in \mathbb{R}^{n \times d}$ is the design matrix, $Y \in \mathbb{R}^n$ is the response vector and $\beta \in \mathbb{R}^d$ represents the coefficients. The OLS solution is given analytically as

$$\beta_{\text{OLS}} = (X^\top X)^{-1} X^\top Y,$$

provided that $X^\top X$ is invertible.

Ridge Regression

Ridge regression introduces a regularization term to penalize large coefficients, stabilizing the solution in ill-conditioned or high-dimensional settings. The ridge optimization problem is:

$$\mathcal{L}_{\text{Ridge}} = \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

where $\lambda \geq 0$ is a regularization parameter controlling the trade-off between fit and complexity. The closed-form solution is:

$$\beta_{\text{Ridge}} = (X^\top X + \lambda I)^{-1} X^\top Y,$$

where I is the identity matrix. Ridge regression shrinks coefficients towards zero, with the amount of shrinkage depending on λ.

A principal component analysis (PCA) perspective highlights ridge regression's bias towards high-variance directions. Using the singular value decomposition (SVD) of X, $X = U\Sigma V^\top$, the ridge solution can be expressed as:

$$\beta_{\text{Ridge}} = V(\Sigma^2 + \lambda I)^{-1}\Sigma U^\top Y.$$

In this form, ridge regression scales contributions from principal components u_i by $\sigma_i^2/(\sigma_i^2 + \lambda)$, where σ_i are the singular values. Components with small singular values, corresponding to low variance, are penalized more heavily.

This selective shrinkage improves generalization by reducing the influence of noisy, low-variance components while preserving the signal in high-variance components.

Ridge regression is particularly effective in brain decoding tasks due to its linearity and regularization properties, which align well with the characteristics of high-dimensional fMRI data. Despite the inherent non-linear complexity of the brain, voxel-level fMRI signals exhibit surprising linearity owing to several factors. Spatial averaging in fMRI aggregates signals from neuronal populations, reducing nonlinearities, while temporal smoothing by the hemodynamic response function (HRF) dampens high-frequency noise and nonlinear fluctuations. In addition, instrumental and physiological noise contribute additively to the observed signal, supporting linear approximations for robust feature extraction. However, ridge regression has notable limitations. High regularization parameters (λ) can lead to over-regularization, shrinking coefficients excessively and potentially discarding meaningful information. Its fixed linear mapping assumes a linear relationship, which may fail to capture complex dependencies in neural data. Furthermore, its performance heavily depends on the quality of

input features, often necessitating dimensionality reduction or functional alignment to optimize decoding accuracy.

0.5.c Neural Network Architectures

Modern AI models rely on diverse neural network architectures [228, 151, 115, 225, 170], each designed to address specific tasks and data structures. These architectures form the backbone of deep learning and enable models to process a variety of input types, including structured, sequential, and high-dimensional data. In the following, we explore the most common architectures and their core functionalities (please refer to Fig 5 for an overview of the adoption of these techniques in this thesis).

Multi-Layer Perceptrons (MLPs)

Multi-Layer Perceptrons (MLPs) represent the simplest and most fundamental neural network architecture. They consist of fully connected layers where every neuron in one layer is connected to every neuron in the next. The computation within each layer is defined as:

$$h^{(l+1)} = \sigma(W^{(l)}h^{(l)} + b^{(l)}),$$

where $W^{(l)}$ and $b^{(l)}$ denote the weights and biases of layer l, $h^{(l)}$ is the input to the layer, and σ is a nonlinear activation function such as ReLU, sigmoid, or tanh.

MLPs excel at modeling simple relationships between input features and output labels, making them effective for small-scale datasets and straightforward regression or classification problems. However, their fully connected nature leads to a high number of parameters, making them computationally expensive and less suitable for high-dimensional data like images or sequences. This limitation has driven the development of more specialized architectures, such as CNNs and transformers.

Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are specifically designed to process spatially structured data, such as images, by using local connectivity patterns [151]. Unlike MLPs, CNNs use convolutional layers to extract features hierarchically, starting with low-level patterns like edges and progressing to more complex structures such as shapes or objects. A single convolution operation for the filter k is defined as:

$$h_{i,j}^{(k)} = \sum_{m,n} x_{i+m,j+n} \cdot w_{m,n}^{(k)},$$

where $w_{m,n}^{(k)}$ is the filter kernel, $x_{i+m,j+n}$ represents a local patch of the input, and $h_{i,j}^{(k)}$ is the output feature map at position (i, j).

CNNs are highly efficient due to their shared weights in the convolutional filters, significantly reducing the number of parameters compared to fully connected layers. Pooling layers, such as max-pooling or average-pooling, further reduce the spatial dimensions of feature maps, enhancing computational efficiency and promoting translation invariance. CNNs are widely used in computer vision tasks, including image recognition, segmentation, and object detection. Interestingly, CNNs also showed some degree of similarity to the human and primate visual cortex in the computations done[159].

Transformers

Transformers have revolutionized AI by providing a highly versatile architecture for sequential and structured data. Originally designed for natural language processing (NLP) tasks, transformers rely on self-attention mechanisms to capture dependencies between elements in a sequence, regardless of their distance. The self-attention mechanism computes weights as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

where Q, K, V represent the query, key, and value matrices derived from the input, and d_k is the dimensionality of the keys.

Unlike recurrent networks, transformers process input sequences in parallel, which leads to significant computational advantages. They have been widely adopted beyond NLP, including in vision tasks (Vision Transformers, or ViTs [65]) and multimodal learning. Their scalability and effectiveness make them a cornerstone of modern AI research [263].

0.5.d Generative Models

Generative models are a class of machine learning algorithms designed to model the underlying data distribution $p(x)$ and generate new samples that resemble the original data. These models are essential for tasks such as image synthesis, text generation, and data augmentation.

Large Language Models (LLMs)

Large Language Models (LLMs) are a type of generative model designed for text processing and generation. They are typically based on transformer architectures

and predict the next token in a sequence by modeling the conditional probability:

$$P(x_t|x_{t-1}, x_{t-2}, \ldots, x_1) = \text{softmax}(Wh_t),$$

where h_t is the hidden state at time t, and W represents the weights of the output layer.

Models like GPT and LLama [23, 220, 68] are trained on massive text corpora and have demonstrated remarkable abilities in tasks such as writing coherent paragraphs, summarizing documents, and answering questions. By leveraging pre-training on large-scale unlabeled data followed by fine-tuning on task-specific datasets, LLMs have set new benchmarks in natural language understanding and generation.

Diffusion Models

Diffusion models are a class of generative models that create data by iteratively refining a noisy sample until it resembles the target data distribution [111]. Starting from a sample drawn from a Gaussian distribution, these models gradually "denoise" the sample using learned noise patterns. The objective during training is to minimize the difference between predicted noise and true noise in each step:

$$\mathcal{L} = \mathbb{E}_{x,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right],$$

where x_t is the noisy sample at timestep t, ϵ is the true noise added to the sample, and ϵ_θ is the predicted noise from the model.

Recently, latent diffusion models, such as VersatileDiffusion or Stable Diffusion [278, 224], that are combinations of VAEs and diffusion in latent space, have shown state-of-the-art performance in image synthesis, surpassing even GANs in some benchmarks. They have also been extended to other domains, including text-to-image generation (e.g., DALL-E [215]) and audio synthesis, demonstrating their versatility and potential for creative applications.

0.6 Related work

Understanding and decoding neural representations has been an active area of research in various sensory modalities. This section reviews key contributions on decoding images, videos, language, and music from neural activity, highlighting approaches using both invasive and noninvasive neural recordings. These studies provide a foundation for advances in encoding and decoding, as well as cross-modal representation alignment. Here we provide a brief and high-level summary to help the reader place the contributions of this thesis into context.

For more detailed reviews we refer the interested reader to exhaustive review papers such as [188, 241].

In recent years, several trends have emerged in brain decoding and encoding models. Encoding models typically combine pre-trained (often multimodal) models with a linear regression, which can be trained independently or jointly with the base model during fine-tuning. Fine-tuning often uses objectives like regression or contrastive learning to map stimuli to predicted brain activity.

For decoding, the methodologies are also diverse [283, 188]. Early approaches focused on classifiers trained on neural data to predict semantic categories. As the goal shifted to reconstruction, Bayesian methods became prominent. These methods pair an encoding model, which is generally easier to train (with respect a decoding one)—with a generative model [283]. The generative model produces candidate reconstructions, which are then evaluated according to their similarity to the target brain activity using the encoding model. Iterative optimization refines these candidates until their estimated brain activity closely matches the target, yielding the decoded output.

Pushing the boundaries further, recent research has focused on estimating model embeddings directly from neural activity. This is often achieved using contrastive learning-based neural networks or Ridge Regression. The estimated embeddings are then paired with retrieval modules or generative models, such as diffusion models or large-language models (LLMs), to reconstruct the stimuli with higher fidelity.

0.6.a Images and Videos

So far, the visual modality has been the main focus of decoding studies due to the structured nature of visual data and well-established models of the visual system. Classical encoding models, such as those by Kay et al. [134], demonstrated the prediction of voxel-wise brain activity from visual stimuli by using Gabor wavelet filters to model responses in the early visual cortex. On the decoding side, Nishimoto et al. [185] reconstructed natural movies from fMRI activity by employing a combination of voxel-wise encoding models and Bayesian inference on a large dataset of natural video clips. Challeges like Algonauts challenge [86] where large scale vision fMRI dataset of people watching short videos or images further improved the quality of encoding models [47, 86, 184] that rapidly converged in using latent representations from multimodal models such as CLIP [209, 46] by projecting them to brain space with Ridge regression or by finetuning the whole model with a linear probe on top.

Recent studies have taken advantage of the capabilities of deep neural networks (DNN) and large data sets to improve the accuracy of encoding and

decoding models. Horikawa and Kamitani [116] used sparse linear regression to predict features of convolutional neural networks (CNN) from fMRI data, allowing category classification. Their study employed the Generic Object Decoding (GOD) [116] dataset, with 1,200 training image stimuli spanning 150 classes and 50 unique test classes, demonstrating high accuracy in decoding visual categories.

Ren et al. [218] and Mozafari et al. [175] contributed by reconstructing images using pre-trained models as priors, highlighting the use of brain-optimized latent spaces for decoding.

VanRullen et al. [262] developed a novel framework to reconstruct faces from fMRI activity by estimating the latent space of a VAE-GAN model. Participants viewed images from the CelebA dataset, a collection of facial images with varying attributes, while fMRI signals were recorded. The reconstructed images were visually realistic, showing significant advances in the decoding of complex visual stimuli. Allen et al. [4] introduced the Natural Scenes Dataset (NSD), a benchmark dataset for semantic-level decoding and reconstruction of natural images. It features unprecedented scale and quality, including 7T fMRI acquisitions from eight participants, who had been exposed to 10,000 unique images sourced from the COCO dataset. The data set has facilitated the training of subject-specific models and cross-subject decoding approaches, enabling researchers to decode complex visual features such as object categories, spatial arrangements, and scene semantics.

Diffusion models, renowned for their generative capabilities, have become instrumental in achieving semantic-level decoding of brain activity. A seminal contribution in this domain is the Brain-Diffuser by Ozcelik et al. [190], which introduced a dual-pipeline approach. This method linearly estimates vision and text embeddings from fMRI data to condition the Versatile Diffusion model [278] for semantic decoding. Simultaneously, it estimates the latent space of the VDVAE model [45] with a linear model of brain activity to generate low-resolution image representations, providing additional conditioning to enhance the generative process.

Building on this foundation, numerous subsequent studies have adopted similar approaches, integrating advancements and refinements. These models typically map fMRI data into multimodal latent spaces, often leveraging the CLIP model, which aligns image and text embeddings, to facilitate robust cross-modal decoding. Ferrante et al. [79, 78] (See chapters 1 and 2) used a multi-step decoding pipeline combining diffusion models, CLIP embeddings and caption (both estimated from brain activity) to retrieve and reconstruct semantic content of images and generate realistic reconstructions. Similar works, such as those by

Takagi et al. [250] and Scotti et al. [236, 235], showcased the potential of contrastive learning methods in brain decoding. The DREAM pipeline by Xia et al. [277] further pushed boundaries, accurately reconstructing visual scenes while mimicking hierarchical visual processing.

Most decoding pipelines are subject-specific, meaning that a separate decoding model must be trained for each individual due to anatomical and functional variations between subjects. To address this limitation, several approaches have been proposed, starting with functional alignment techniques such as hyperalignment, shared response modeling (SRM), and optimal transport methods [98, 39, 15, 220]. Recent advances by Ho et al. and Ferrante et al. [113, 80] demonstrated that functional alignment can be outperformed by training a Ridge regression to map brain activities between subjects. This approach enables the development of models that generalize across individuals.

In particular, Ferrante et al. [80] (see chapter 3) showed that it is possible to align data from different experiments (cross-dataset experiments), effectively overcoming discrepancies in MRI machines, magnetic field strengths, and experimental paradigms.

An intriguing recent approach to video decoding is presented in the Mind-Video study by Chen et al. [42], which addresses the task of reconstructing video content from fMRI data using a generative variant of Stable Diffusion augmented with temporal and spatial attention mechanisms to ensure frame-to-frame consistency. This methodology was further refined in a subsequent study by Sun et al. [248], which introduced advancements in the generative model, enhancing the fidelity of decoded video sequences.

On a larger dataset[145], Ferrante et al [75] (see chapter 8) introduced a novel multistream decoding pipeline that integrates audio, visual, and semantic information for video retrieval tasks.

So far, this review of the literature has focused primarily on fMRI due to the rapid advancements driven by the availability of high-quality, large-scale datasets and increased computational resources. While fMRI offers high spatial resolution and a rich dataset for understanding neural representations, its limitations include non-portability, long experimental durations, and high costs, making it less suitable for real-world brain-computer interface (BCI) applications. This has spurred significant interest in more portable (EEG) and temporally precise recording technologies like MEG.

Electroencephalography (EEG) has been widely explored in the context of BCIs for a variety of tasks [283, 239, 264, 26, 176]. Early work, such as Kavasidis et al. [132], used EEG data recorded during ImageNet image presentations to generate class-level images using LSTM models combined with variational au-

toencoders or GANs. Spampinato et al. [244] further analyzed EEG responses to ImageNet stimuli by training LSTM encoders for image category classification and employing CNN regressors to predict EEG features from images. Extending these approaches, Palazzo et al. [191] utilized contrastive learning to align EEG and visual features, focusing on improving image classification. Initially, some results seemed extremely promising, however, it was later shown that performances in some of these works were inflated due to unindentend information leakage between train and test set due to preprocessing and data splitting [156, 155], especially on works based on the [131] dataset. This originated a debate, and in more recent works it appears that semantic image decoding from EEG is feasible with performances above chance level but the noise in the data introduces a lot of misclassifications.

Recently, Singh et al. [242] introduced an EEG-to-image GAN framework, demonstrating its applicability to smaller datasets with simpler stimuli, such as characters and shapes. Ferrante et al [77] propose a semantic image reconstruction pipeline based on EEG encoding and CLIP-based knowledge distillation showing above-chance-level performance in image decoding on [131] and [105] datasets.

In parallel, advances in magnetoencephalography (MEG) have emerged, with notable contributions from Meta's work on MEG-based image reconstruction [17] with an approach similar to Brain-Diffuser applied on MEG data. These studies underscore the growing potential of EEG and MEG for portable and high-temporal-resolution decoding applications.

Again on images, Ferrante et al. [81] (See chapter 4) proposed a contribution based on contrastive learning, where different neural datasets (fMRI, EEG and MEG) during vision tasks are aligned in the same CLIP space with a single model. Based on a retrieval approach, it's possible to solve decoding, encoding, and modality conversion all within a single model.

Invasive methods such as electrocorticography (ECoG) have further advanced visual decoding. Liu et al. [52] reconstructed high-fidelity static images from ECoG recordings, while similar approaches have also achieved video reconstruction. The use of high-temporal-resolution techniques, combined with DNN embeddings, has led to significant improvements in fidelity and realism. For example, the CEBRA framework [231] integrated latent space alignment techniques to decode detailed video sequences directly from intracranial recordings of a mouse.

How does mental imagery factor into this? In everyday life, we often and easily visualize objects, scenes, or concepts mentally, highlighting the importance of extending vision-based models to decode imagery. Although the amount of

available data is currently limited, there have been notable attempts to explore this area [116, 141, 226]. These studies suggest that in fMRI, the signal-to-noise ratio for mental imagery is significantly lower than that of visual perception. Additionally, while principal component analysis (PCA) reveals some shared components between imagery and visual perception, not all components overlap. This indicates potential opportunities for investigating scaling laws and exploring the similarities and differences between visual perception and mental imagery to advance this field further.

0.6.b Language

Language decoding has immense potential to transform our understanding of human communication and facilitate the development of neuroprosthetic technologies. By analyzing neural recordings, researchers aim to unravel the intricate relationship between brain activity and language representations, paving the way for innovative decoding methods.

Noninvasive techniques, particularly functional magnetic resonance imaging (fMRI), have been instrumental in mapping distributed semantic representations in the brain. Pioneering studies by Huth et al. [123, 122] revealed that semantic concepts are encoded in spatially distributed cortical patterns. Using naturalistic stimuli, they successfully mapped these representations across the brain. Recently, Tang et al. [251] advanced the field by proposing a Bayesian approach to semantic text decoding. This method integrates an encoding model of brain activity with contextual sentence embeddings, a language model to generate candidate sequences based on natural language statistics, and a noise model to disentangle voxel contributions. Their approach decoded the content of narratives listened to during fMRI experiments with remarkable precision.

The rise of large language models (LLMs) has significantly improved the accuracy and generalizability of brain decoding methodologies. Schrimpf et al. [233] demonstrated the alignment between transformer-based models, such as GPT, and neural data, showcasing how these models serve as effective scaffolds for decoding brain activity into semantic representations. Encoding models have become central to research on language understanding and processing, measuring alignment between task-related language model embeddings and brain activity. Larger models consistently predict brain activity more effectively [7, 199, 254], suggesting that superior embeddings better approximate the brain's computational mechanisms. This observation aligns with scaling laws observed in brain encoding research. Several papers on language understanding in the brain followed a similar approach [35, 36, 33] that led to neuroscientific insights on language processing. Ferrante et al. ([76], Chapter 6) (see Chapters 5 and 6)

proposed a contrastive-based approach to retrieve sentences from fMRI data and LLM adaptations to direct perform language decoding by fine-tuning this kind of models with specific data augmentations. Language decoding faces challenges due to the multifaceted ways in which humans engage with language, such as reading, listening, thinking, and speaking. Each mode activates distinct yet overlapping brain regions, complicating decoding efforts. While invasive neural recording techniques, such as electrocorticography (ECoG), have achieved high accuracies approaching natural spoken rates of up to 60 words per minute [106, 206, 273], non-invasive methods are safer and more broadly applicable. Recent advances in deep learning, coupled with richer datasets encompassing diverse stimuli per subject, are reshaping noninvasive language decoding research by using recurrent neural networks or LLMs to have strong priors on the statistics of language. High-temporal-resolution methods, including electroencephalography (EEG) and magnetoencephalography (MEG), have also been extensively explored. While MEG has shown promising correlations with language processing [53], the results of EEG work remains debated [67, 125, 280]. Scaling laws appear critical in this context; evidence suggests that surpassing chance-level decoding with EEG may require datasets with more than 100 hours of subject data, indicating a logarithmic scaling relationship [230]. Interestingly, it seems that decoding imagined speech is possible both from fMRI [251] and iEEG [206].

0.6.c Music

Music has also been explored to investigate his perception in the brain, providing a foundation for further exploration of neural mechanisms underlying auditory processing and creative applications in music generation and retrieval.

A pivotal study by Denk et al. [57] on music retrieval and generation from fMRI data proposes subject-specific pipelines that rely on anatomical atlases and proprietary backbone models such as MuLAN and MusicLM [3, 119] on the GTZan fMRI dataset (5 subjects each listening to 540 songs belonging to 10 different musical genres) [180]. In contrast, Ferrante et al. [74] (see chapter 7) aimed to establish a generalized framework based on music ROI identification, CLAP based music embeddings, and cross-subject alignment to train a decoding module able to retrieve musical tracks from fMRI activity.

Using invasive recordings of neural activity, [16] demonstrate that time-frequency decompositions can be effective representations for this type of task, and that they can be performed using both linear and nonlinear approaches to decode the auditory experience using invasive iEEG data, nicely decoding a Pink Floyd song.

Table 1: Summary of Thesis Chapters, Datasets, and Contributions

Chapter	Dataset Used	Main Contribution	Modality
1 Semantic Decoding	Generic Object Decoding (GOD) Dataset [116]	Developed a semantic decoding pipeline using CLIP embeddings and linear ridge regression. Demonstrated classification of unseen semantic categories and semantic reconstruction.	fMRI
2 Brain Captioning	Natural Scenes Dataset (NSD) [4]	Introduced a cross-modal decoding approach combining fMRI data and diffusion models to generate captions and reconstructed images from neural data.	fMRI
3 Cross-Subject Decoding	BOLD5000 [38] + NSD	Proposed a Ridge Regression-based functional alignment method for cross-subject decoding, reducing training data requirements by 90%.	fMRI
4 Multimodal Decoding	NSD, ImageNetEEG [131], THINGS-MEG [104]	Developed a unified contrastive learning-based framework for encoding, decoding, and modality conversion tasks, aligning neural data from multiple modalities in a shared space.	fMRI, EEG, MEG
5 Contrastive Language Decoding	Voxelwise langauge dataset [148]	Created a retrieval-based language decoding method leveraging contrastive learning to decode fragments of sentences from fMRI activity.	fMRI
6 Direct Language Decoding	Same as above	Adapted large language models (LLMs) to directly generate sentences from fMRI data, demonstrating generative decoding capabilities.	fMRI
7 Music Decoding	GTZan [180]	Proposed a decoding pipeline for music retrieval and reconstruction, integrating CLAP embeddings and region-of-interest (ROI) selection.	fMRI
8 Video Decoding	Large-scale video-fMRI BOLDmoments dataset [145]	Introduced a multi-stream decoding framework integrating audio, visual, and semantic information for video reconstruction and retrieval tasks.	fMRI
9 Neuroscience Insights	NSD Dataset	Explored the theoretical underpinnings of brain function, proposing a "brain algebra" framework to analyze semantic representations and their transformations in neural spaces.	fMRI

Part II

Decoding vision

1 Semantic Brain Decoding

In this chapter[1], we propose an approach to brain decoding that relies on semantic and contextual similarity. We use several fMRI datasets of natural images as stimuli and create a deep learning decoding pipeline inspired by the bottom-up and top-down processes in human vision.

Our pipeline includes a linear brain-to-feature model that maps fMRI activity to semantic visual stimuli features. We assume that the brain projects visual information onto a space that is homeomorphic to the latent space represented the last layer of a pretrained neural network, which summarizes and highlights similarities and differences between concepts. These features are categorized in the latent space using a nearest-neighbor strategy, and the results are used to retrieve images or condition a generative latent diffusion model to create novel images. We demonstrate semantic classification and image retrieval on three different fMRI datasets, GOD (vision perception and imagination), BOLD5000 and NSD. In all cases a simple mapping between fMRI and a deep semantic representation of the visual stimulus resulted in meaningful classification and retrieved or generated images. We assessed quality using quantitative metrics and a human evaluation experiment that reproduces the multiplicity of conscious and unconscious criteria that humans use to evaluate image similarity. Our method achieved correct evaluation in over 80% of the test set. Our results demonstrate that measurable neural correlates can be linearly mapped onto the latent space of a neural network to synthesize images that match the original content. The findings have implications for both cognitive neuroscience and artificial intelligence.

[1]The work presented in this chapter has been presented at ISMRM 2023 and published at "Journal of Neural Engineering" [79].

1.1 Introduction

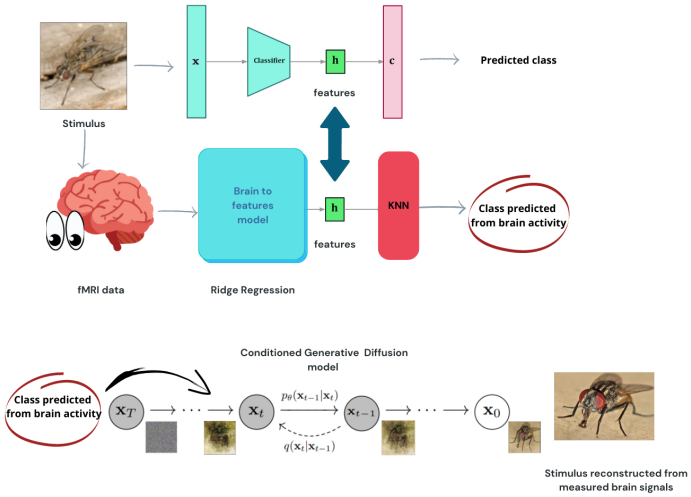


Figure 1.1: Our proposed architecture. According to our hypothesis, the brain processes information by extracting visual features from images and projecting them onto a latent semantic space similar to the one formed by a neural network, termed "classifier" in this figure (in this paper we employed CLIP as backbone architecture) when trained for object categorization. We developed a regression model that maps fMRI brain data to the CNN's latent space and used a k-nearest-neighbor (kNN) method to predict the related classes. Finally, we conditioned a latent diffusion model to generate novel images that are semantically similar to the visual stimuli from the predicted classes.

1.1.a Background and Motivation

Brain decoding attempts to infer internal representations of perceptual stimuli from measurable brain activity. Isolated attempts have been made to use deep learning to 1) identify complex brain data patterns and 2) reconstruct the stimuli that have generated such patterns using noninvasive neuromonitoring data such as functional magnetic resonance imaging (fMRI) or electroencephalography (EEG) [283]. While these activities are in very early stages, they also carry great promise for the development of novel strategies to diagnose and treat neurological or neuropsychiatric conditions. However, such endeavors carry many challenges. Noninvasive data, for example, have lower temporal or spatial resolution than that of neural firing, resulting in a potential upper limit on the granularity of information that may be retrieved. The latter is also degraded by physiological noise and signal/image artifacts.

1.1.b Visual Cortex and Pathways

Vision has been extensively studied along with its brain representations (i.e., the visual cortex). The latter are organized hierarchically into sections that respond to specific stimuli (commonly termed V1, V2, V3, V4, and the lower and upper visual cortices). Simple visual inputs tend to elicit V1 responses, while V2 responds to texture, color, and more complex outlines. There is strong evidence that information flows from the visual cortex (VC) to the rest of the brain through two separate routes, the *what* and *where* pathways [13, 258, 97, 94]. The *what* pathway connects the VC to the inferior temporal lobe (IT) and is involved in object recognition, whereas the *where* pathway connects the VC to the parietal lobe and is primarily involved in movement and position recognition.

1.1.c Semantic Representations in the Brain

In vision, the bottom-up information extraction described above is accompanied by a top-down mechanism [89] where semantic prior knowledge of the world is exploited to create internal representations of external stimuli. This results in a combination of context-given prediction and purely external signals relayed from the retina to the brain. According to the 'hub-and-spoke' theory of semantic representation, conceptual knowledge arises from the progressive learning of the statistical regularities of our multi-sensorial experiences. In other words, we learn how to recognize an ever-changing environment by systematically linking apparently separate aspects of our experiences (e.g., color, motion, sounds, sensory-motor actions associated with an object, etc.) that tend to co-occur. Such learning processes transform a sensory 'cacophony' into a coherent, context-specific, and behaviorally-relevant semantic representation of the stimuli.

1.1.d Semantic Cognition and Modality-Specific Brain Areas

The brain mechanisms underlying semantic cognition have not been fully elucidated, but the prevailing hypothesis suggests that modality-specific brain areas, also known as the 'spokes' (e.g., visual cortices, auditory cortices, motor areas, emotional systems), interact via a central and a-modal 'hub' region (the anterior temporal lobe) to form conceptual knowledge. This process shapes the semantic representation through various experiences, such as visual, auditory, verbal, and tactile, and critically promotes the ability to generalize across different items and variable contexts. Interestingly, there are indications of the existence of a continuous semantic space representation [123] in the human brain. Though the structure and topology of this putative semantic space have been poorly inves-

tigated, there is evidence that fMRI data from occipital brain regions collected during a visual task can be linked to features learned by a convolutional neural network (CNN) [159], with a particular focus on the early and middle CNN layers.

1.1.e Decoding Visual Stimuli from fMRI Data

In this paper, we tackle the problem of decoding (i.e., reconstructing or retrieve) visual stimuli (images) from fMRI data only, by leveraging the hypothesis that deep convolutional layers can operate as a proxy for parts of the brain that extract semantic features from images [188]. Specifically, we used CLIP [210] as frozen deep learning backbone to extract latent image representation. Previous experience [78] and literature [161, 46] show evidence that large multimodal models can better capture the semantic aspect of the images and connect them with brain activity. We propose a cascade of deep learning models that builds convincing semantic reconstructions of the stimulus presented at acquisition time. It is important to note that the aim of this paper is not to create exact reconstructions of the images presented under fMRI. Instead, our objectives are to either a) retrieve in the dataset realistic visual representations that capture the main concepts contained in the original stimulus, or b) create synthetic images that can trigger similar brain activity when employed as stimuli. Achieving either of these results can pave the way for a more general understanding of cognitive-visual information storage and retrieval. Specifically, we focused on the problem of decoding the semantic category of the seen image from brain activity, performing classification and image retrieval. When possible, we extended our results to image generation guided by predicted semantic activity.

1.2 Related Work

1.2.a Reconstructing Information from fMRI Data

In recent years, several attempts have been made to reconstruct information from noninvasively acquired brain data, particularly fMRI data. This has been fueled by the increasing availability of public datasets, advances in computational power, and more sophisticated nonlinear analytic approaches, such as deep neural networks. While challenges related to signal-to-noise ratio (SNR), duration of acquisition session, and HRF variability remain, fMRI appears capable of extracting useful information in a wide range of situations and tasks, including vision and visual stimulus classification.

1.2.b Existing Approaches and Challenges

Various modeling frameworks have been employed in brain decoding literature, where the input is usually preprocessed fMRI time series. These data are referred to as "fMRI data", "fMRI patterns", and "fMRI activations", terms used interchangeably in this paper. Existing approaches to brain decoding include:

- Variational autoencoder with a generative adversarial component (VAE-GAN) for encoding latent representations of human faces [262].
- Sparse linear regression over preprocessed fMRI data for predicting features extracted by multiple early convolutional layers from a pretrained CNN [116].
- An adversarial strategy employing a generator and discriminator to differentiate between real and reconstructed images, further improved by a perceptual loss and a comparator network [237].
- A dual VAEGAN consisting of two linked variational autoencoders for representing both stimuli and fMRI patterns [218].
- An unsupervised technique using two encoders and two decoders learning separately how to reconstruct fMRI data and stimuli, bound by a supervised loss [85].
- Optimizing pretrained architectures' latent spaces, such as BigBiGAN [175] and IC-GAN [189], to reconstruct high-quality images from fMRI patterns.
- [43] performed a direct estimation of the latent space of a latent diffusion model from fMRI data, employing a pre-trained autoencoder to reduce the dimensionality of fMRI representations. By combining the HCP [260] (1,200 subjects) and GOD datasets, they achieved a substantial sample size to learn self-supervised representations and fine-tune them for inferring the latent representations of images with limited labeled pairs. Our work is closely related due to the utilization of the same GOD dataset and latent diffusion models for image reconstruction; however, our main distinction lies in the development of an ad-hoc pipeline to address the small sample size of the GOD dataset independently, whereas they relied on external fMRI acquisitions to learn self-supervised representations.
- Latent diffusion models have been recently employed as image generators in [250] and [190], where the authors used the Natural Scenes Dataset [5], containing 70,000 images acquired with a 7T scanner. This extensive dataset significantly enhances the quality and quantity of input data for brain decoding tasks. In the first study, the authors directly optimized the latent space of the diffusion model, while in the second, an initial guess image reconstruction was obtained by mapping fMRI data into the latent space of a deep variational autoencoder trained for image reconstruction.

These initial guesses encompass information about shape, color, and pose of images, and can be combined with predicted conditioning in latent diffusion models through image-to-image pipelines to improve reconstruction quality.

Complete lists of excellent other works can be found in reviews like [267]. The main innovation that we are proposing i this paper is a shift of focus on semantics. Wew build a semantic decoder of brain activity with the objective of retrieving seen images based on their semantic content. Our approach is simple and flexible, and we demonstrate that reach good perfomances on three different datasets.

1.2.c Focus on Semantic Content

Most of the research in brain decoding has focused on extracting either low-level visual stimulus characteristics or reconstructing whole images in pixel space. While these studies capture forms, colors, or images that look similar to the original stimuli, reconstructions are often blurred and mix elements from unrelated concepts. In this paper, we focus on context, i.e., the semantic content of presented stimuli, with the aim of reconstructing images that resemble the original ones and can elicit the same fMRI activity. We hypothesize that this approach may add ecological relevance to our findings in terms of future applications for understanding visual information representation in the brain. Our approach was implemented as follows: For the Generic Object Decoding (GOD) dataset (Horikawa et al. [116]), which includes images with specific classes akin to those in ImageNet, we used a methodology involving the linear regression of latent image representations from fMRI data, succeeded by the use of a k-nearest neighbors (kNN) algorithm for the purposes of image retrieval and classification. Following the classification phase, we expanded our examination to include the generation of images and the assessment of these images by human evaluators. Regarding the Natural Scenes Dataset and BOLD5000 [5, 38], which exposed participants to more intricate visual scenes, we initially re-categorized the dataset by employing a clustering strategy on image embeddings derived from the CLIP classification token, thereby deriving a refined assortment of semantic classes. Subsequently, we applied the identical pipeline—regressing embeddings from fMRI data followed by kNN-based retrieval—for classification and image retrieval.

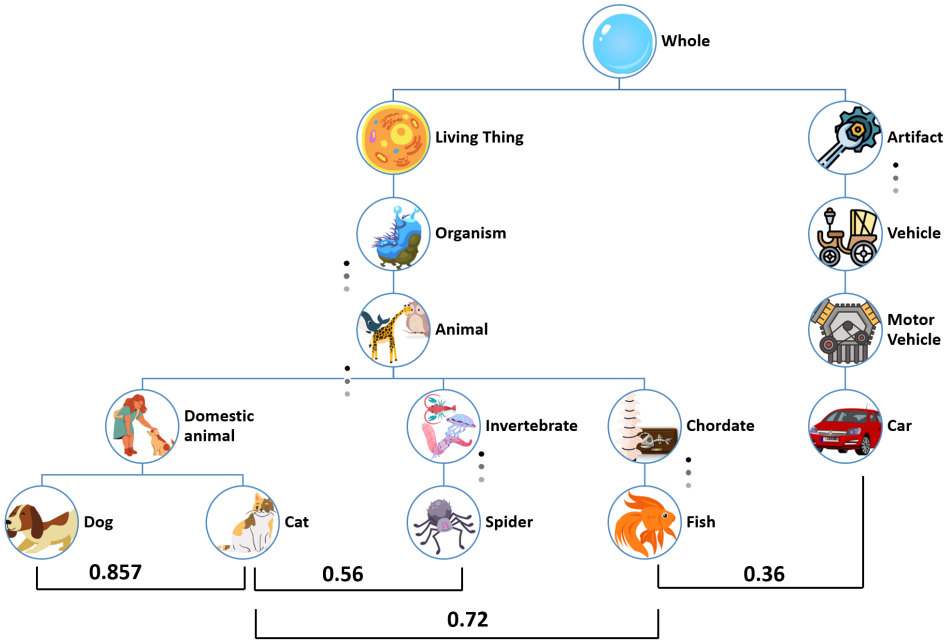


Figure 1.2: Simplified depiction of the hierarchical representation of semantics and concepts in the WordNet lexicon. Dotted lines indicate that there are additional nodes between the ones visualized in the figure (but no ramifications). Wu-Palmer distances between nodes are represented by numbers over solid lines.

1.3 Methods

In this section, we describe the implementation aspects of our study, summarized in Fig. 1.1. We used Python 3.9 along with the PyTorch and scikit-learn libraries to develop our models. The experiments were conducted on a server equipped with two Intel Xeon Gold processors, 512 GB RAM, and an NVIDIA A6000 GPU with 48 GB RAM. Our code is available at https://github.com/matteoferrante/semantic-brain-decoding, and the preprocessed data can be accessed at https://figshare.com/articles/dataset/Generic_Object_Decoding/7387130. Unprocessed fMRI data is available at https://openneuro.org/datasets/ds001246/versions/1.2.1. As reported in the original paper that described data [116] all subjects provided written informed consent for participation in the experiments, and the study protocol was approved by the Ethics Committee of ATR Brain Activity Imaging Center.

1.3.a Data and Preprocessing

In this section we describe the datasets used in this work. We reproduced the same pipeline on several fMRI vision data to dmonstrate the generalizability and applicability of our method. In particular, we started our analysis on the GOD dataset, which comes with actual classes from ImageNet, and further extended to the BOLD5000 [38] dataset as well as to the large Natural Scenes Dataset [5], acquired at 7T. The latter two datasets exposed subjects to complex scenes derived from the COCO set of natural images, thus increasing the complexity of the problem that we want to tackle.

Generic Object Decoding (GOD)

We utilize the publicly accessible Generic Object Decoding (GOD) dataset [116], which comprises fMRI data from 5 subjects who participated in either an image presentation experiment or an imagery experiment. The GOD dataset has been instrumental in developing previous brain decoding models and is emerging as a valuable benchmark for decoding visual stimuli from fMRI data. All visual stimuli in the GOD dataset originate from the ImageNet database (http://www.image-net.org/, Fall 2011 release), which is categorized into various classes, including animals (e.g., "goldfish," "swarm," and "tiger") and objects (e.g., "airplane," "hat," or "knife").

The image presentation experiment involved separate training and test sessions. In the training session, 1,200 images from 150 object categories (8 images per category) were presented once. In the test session, 50 images from 50 object

categories (1 image per category) were shown 35 times each. Each stimulus was displayed for nine seconds. No overlap existed between the categories of training and test images. In this dataset, a single fMRI acquisition is called a "run," with 24 runs for training images and 35 runs for testing images performed for each subject. The fMRI protocol was based on an EPI sequence with $TR = 3000$ ms, $TE = 30$ ms, flip angle=80°, and a voxel size of 3 mm^3.

Data were preprocessed in native subject space by performing 3D motion correction, linear trend removal, and coregistration to a high-resolution common anatomical template. Reference masks for the visual cortex (VC) and several other brain areas, such as the face fusiform area (FFA), the high VC (HVC), and the low VC (LVC), were provided for each subject. In this study, we used data extracted from the VC (approximately 4,500 voxels per subject) as our input space. The data were normalized runwise, ensuring each voxel-specific timeseries had a zero mean and unit variance. Subsequently, data were averaged over time using nonoverlapping 9-second windows and effectively shifted forward by 3 seconds (i.e., three volumes per average, corresponding to the length of a stimulus presentation). This process helped reduce complexity and account for delays induced by the hemodynamic response function (HRF) convolution.

Natural Scenes Dataset (NSD)

We employed the Natural Scenes Dataset (NSD) [5], which consists of extensive fMRI data from eight individuals who were shown images from the COCO21 collection for 2 seconds with one second of pause between subsequent stimuli presentations. Our analysis concentrated on a subset of four participants, resulting in a specialized training dataset comprising 8,859 images and 24,980 fMRI trials, along with a shared dataset of 982 images and 2,770 trials. To diminish spatial dimensionality, a mask was used on the fMRI signals (with a resolution of 1.8mm isotropic), targeting the NSDGeneral ROI which covers various visual regions (around 15000 voxels). This selective focus on specific regions of interest (ROIs) improved the signal-to-noise ratio and reduced the complexity of the data, thereby facilitating the investigation of visual features at both the basic and advanced levels. The temporal dimensionality was minimized through the use of precomputed coefficients (beta values) of a general linear model (GLM) that included a fitted hemodynamic response function (HRF) and a denoising strategy as outlined in the NSD documentation.

BOLD5000 dataset

The BOLD5000 dataset [38] consists of fMRI data from five subjects who were exposed to a collection of 5,000 COCO images. This dataset was obtained using a 3T magnetic field strength. The methodologies employed in these datasets also vary; NSD uses a rapid-event related protocol with images shown for two seconds followed by a one-second break, whereas BOLD5000 presents images for one second, followed by a nine-second period of cross fixation. Despite these differences, both datasets were processed in the same manner to extract task-related voxel coefficients, particularly within the visual cortex masks.

1.3.b Re-labelling of NSD and BOLD5000 datasets

To enhance the utility and interpretability of the NSD (Natural Scenes Dataset) and BOLD5000 datasets, which predominantly feature complex scenes from the COCO dataset, we implemented a re-labeling strategy. This strategy was designed to introduce semantic pseudo-labels that better capture the rich content within these images, thereby facilitating the application of our pipeline to these datasets. The first step in our re-labeling process involved the computation of the CLIP CLS (classification token) embeddings for each image in the training datasets. CLIP embeddings are particularly suited for this task because they are designed to capture a wide array of visual features in a manner that correlates well with natural language descriptions, making them ideal for understanding complex scenes and obtaining semantic clusters. We then applied the K-Means clustering algorithm to group the images into clusters based on the similarity of their CLIP embeddings. The K-Means algorithm partitions the data into K distinct clusters, with each cluster represented by the mean of the embeddings belonging to it. To determine the optimal number of clusters, K, we employed the k-elbow method, which assesses the variance explained as the number of clusters increases and identifies the point at which the marginal gain in explained variance begins to diminish. Through this analysis, we identified 25 as the optimal number of clusters within the range of 4 to 40. The resulting clusters effectively served as semantic pseudo-labels for the images in the datasets. These labels offer a more nuanced and interpretable categorization of the scenes than traditional labeling methods, reflecting the complex and multifaceted nature of the images. Leveraging these newly assigned semantic pseudo-labels, we adopted a combined approach of fMRI regression and k-Nearest Neighbors (kNN) for image retrieval on these datasets. This approach involves using fMRI data to predict the cluster (or semantic pseudo-label) an unseen image belongs to, facilitating the retrieval of images from the dataset that are semantically related

to the brain activity patterns observed. This method represents a significant advancement in our ability to connect neural responses with complex visual stimuli, opening new avenues for research in visual neuroscience and beyond.

1.3.c Subject-specific Brain Activity Models

We developed individual models for each subject to decode their brain activity, as intersubject functional variability could be greater than the impact we aim to extract. Our hypothesis is that the brain processes sensory input in the VC to extract relevant features from images for object recognition, employing a hierarchical approach similar to convolutional neural networks (CNNs) or multimodal models like CLIP.

We propose a linear mapping between processed fMRI data and the last convolutional layer of a frozen CLIP [210] architecture, pretrained on large image dataset. The objective is to find the optimal weights W that minimize the regularized loss described in Eq. (1.1):

$$\min(|Wx(s) - f(s)|^2 + \lambda|W|^2) \tag{1.1}$$

Here, s represents the image/stimulus presented during the experiment, f is the neural network that projects s into the latent space, and $x(s)$ is the preprocessed brain activity associated with viewing the stimulus. W maps fMRI data into image features in the latent space generated by the deep learning backbone. λ is a hyperparameter for $L2$ regularization on the weights. We optimized λ using a $90 - 10$ training/validation split and grid search.

Subsequently, we generated the conditioning for the generative model that synthesizes the final output. We used the frozen pretrained deep learning backbone (CLIP) to compute the latent representations of images shown during the experiment and stored their latent representation and ground truth labels. From the image features $\tilde{h} = Wx(s)$ predicted from brain activity, we identified the five nearest neighbors in the latent space and used their labels as candidates for classification.

This strategy accounts for the poor signal-to-noise ratio in fMRI data and the limited dataset size. Assuming that similar semantic concepts lead to similar features within the deep learning backbone latent space, the features generated by our brain-to-features model (ridge regression) are likely to be close to concepts semantically close to the target.

1.3.d Bottom-up and Top-down Processes

We now discuss the combination of predicted features that simulate the bottom-up process in vision (where the brain computes stimuli) and the use of a nearest-neighbor-based algorithm to mimic top-down connections that modulate the signal we perceive according to our knowledge of the world. We also address the domain adaptation technique employed to predict the test set features from brain activity, the use of latent diffusion models as image generators, and the evaluation of semantic content through metrics such as the Wu-Palmer distance. In this study, we combine predicted features to simulate the bottom-up process in vision, where the brain computes stimuli, while using a nearest-neighbor-based algorithm to mimic top-down connections that modulate the signal we perceive according to our knowledge of the world [112, 61].

1.3.e Output normalization

There is no overlap between training and test categories in the GOD dataset, and test images are displayed numerous times to achieve a higher SNR. Since the brain-to-feature model is trained using training data, we employ a simple domain adaptation technique to predict test set features from brain activity, which involves replacing the mean and standard deviation of predicted features from the test set with those from the training set [112, 61].

1.3.f Image Retrieval

Upon mapping the fMRI data to the CLIP image embeddings, the retrieval process is initiated for all datasets. This process employs a kNN algorithm, with k set to 3, to identify the closest matches between the estimated image features (derived from the fMRI data) and the actual image features (obtained directly from the CLIP embeddings). The distance metric used in the k-NN search is primarily based on Euclidean distance, offering a straightforward yet effective measure of similarity between the high-dimensional feature vectors. For each set of estimated image features, the algorithm searches the entire test set of real image features. It then selects the three images whose CLIP embeddings are most closely aligned with the estimated embeddings. This process ensures that the retrieved images are those that most closely resemble, in a high-dimensional feature space, the visual content that corresponds to the recorded brain activity.

1.3.g Latent Diffusion Models as Image Generators

Since for the GOD dataset, we have a limited number of classes with specific and explicit semantic content, we could use the predicted labels as conditioning for reconstructing images. To generate images (i.e., reconstruct visual stimuli), we rely on a powerful, recent pretrained image generator belonging to the family of denoising probabilistic diffusion models [112]. Diffusion models are generative architectures that learn how to reverse a diffusion process, which in this context refers to the progressive addition of Gaussian noise to an image. This family of models is far more robust in training than other generative models, such as generative adversarial networks (GANs), and has greater mode coverage [61].

1.3.h Evaluating Semantic Content

In our study, the overarching goal is to assess and ensure semantic closeness between the visual stimuli shown to participants during fMRI experiments and the images generated by our decoding model. Given the nature of our objectives, it is crucial to employ evaluation metrics that transcend pixel-based similarity and capture the deeper semantic essence of the images. Therefore, our analysis primarily focuses on metrics that reflect the semantic content of the images. We therefore evaluated the ability of our model to correctly classify images based on predefined categories. For the GOD dataset, the images were classified according to ground-truth object descriptions. For the other two datasets, we used pseudo-labels generated from cluster analysis to classify the images. The classification performance was quantified by measuring the top-1 and top-3 accuracy rates, providing an initial indication of how well our model can distinguish between different semantic categories. Extending our analysis beyond initial classification, we examined the images retrieved through our retrieval pipeline. This involved assessing a variety of metrics for the closest decoded image. Low-level metrics such as Pixel Correlation (PixCorr) and the Structural Similarity Index (SSIM) were considered to evaluate basic image integrity and similarity. Furthermore, high-level metrics were also employed to measure semantic accuracy. This included 2-way accuracy assessments conducted in the feature spaces of AlexNet, Inception, and CLIP, which are designed to capture more abstract and semantically relevant image characteristics. For the GOD dataset, the analysis was deepened by integrating sophisticated semantic similarity metrics. We measured the Wu-Palmer distance metric [198] between the real and predicted classes in the WordNet lexicon to estimate a quantifiable measure of semantic similarity. This is a well-established metric that measures the similarity of two different nodes (i.e., synsets) in the WordNet graph and can be computed as de-

scribed in Eq (1.2), where s is the similarity metric, lcs stands for "least common subsumer" and is a function that returns the deepest common ancestor in the taxonomy between the two synsets $s1, s2$ and $depth$ is a function that computes the depth in the graph. This metric is bounded in the interval $[0, 1]$, where higher values mean that two synsets are more similar. A simplified graphical representation of the WordNet subgraph is shown in Fig. 1.2 along with some examples of Wu-Palmer distances. Additionally, we used the Fréchet Inception Distance (FID) score to further assess the quality of the generated images. The FID compares the multivariate Gaussian distributions of real and generated images in the feature space of a pretrained neural network (InceptionV3). This metric not only reflects the visual quality of the images but also incorporates elements of semantic similarity due to the feature space in which the comparisons are made. The implementation details and results of these metrics are described in the subsequent section of this paper, reinforcing our comprehensive approach to evaluating semantic similarity in image generation models.

$$s_{wup} = \frac{depth(lcs(s1, s2))}{depth(s1) + depth(s2)} \tag{1.2}$$

In addition to the Wu-Palmer distance metric, we conducted a human evaluation to assess the semantic similarity of the reconstructed images.

1.3.i Human Evaluation of Image Reconstructions

We designed a human evaluation paradigm as follows. A local web page was created, which displayed the original image alongside five model-generated reconstructions in one row and five random reconstructions in another row (Fig. 1.3). Volunteers were instructed to examine the similarities between the images and select the row (first or second) that appeared closest to the original image. To minimize priming, the row positions were continuously randomized between "top" and "bottom."

Seven observers (5 males, 2 females, aged 25-33, with normal eyesight) participated in this evaluation, covering all subjects in the GOD dataset. Each observer assessed the 50 images in the test set and a common random subset of 50 images from the training set, resulting in a total of 350 evaluations. When performing this task, the human observers likely focused on various elements, including broad features like shapes and colors, as well as more semantically related aspects, such as "wild animals" or "furnishings." We believe this natural flexibility in judgment is relevant to our study, as the model utilizes features extracted by a classifier trained on the ImageNet dataset. These features can represent different levels of complexity based on the difficulty of the task, and similar comparison

Figure 1.3: Example taken from the local human assessment local web page. The target image is presented on the left. The subject is instructed to assess the overall resemblance of the original stimulus (left) to the 5 images in the top and bottom rows on the right and to pick "TOP" or "BOTTOM " accordingly.

operations might be performed by our brains in everyday life. To further minimize priming, the row positions were continuously randomized between "top" and "bottom."

1.4 Results

Reference	Dataset	Modality	# Classes	Top_1_accuracy	Top_3_accuracy
Our Work	GOD	Visual	50	0.2000 (0.0583)	0.4080 (0.0832)
Koide et al. [140]	GOD	Vision	50	-	-
Our Work	GOD	Imagery	50	0.0800 (0.0316)	0.2000 (0.0693)
Koide et al. [140]	GOD	Imagery	10	-	-
Mind-vis [43]	BOLD5000	Visual	50	0.334	-
Our Work	BOLD5000	Visual	25	0.5773 (0.0390)	0.8257 (0.0445)
Our Work	NSD	Visual	25	0.7225 (0.0271)	0.8854 (0.0190)
Lin et al. [158]	NSD	Visual	-	-	-
Takagi et al. [250]	NSD	Visual	-	-	-
Gu et al. [218]	NSD	Visual	-	-	-
Ozcelik et al. [190]	NSD	Visual	-	-	-
Scotti et al. [236]	NSD	Visual	-	-	-

1.4.a Visual Comparison and Qualitative Results

The primary objective of our study is to generate images that are realistic reconstructions of visual inputs that semantically match the target image, which is the image used as a stimulus in the fMRI experiment. Regarding image generation

Reference	pix_corr	ssim_score	alex2	alex5	inception	clip_acc
Our Work	0.32 (0.06)	0.38 (0.05)	0.67 (0.06)	0.68 (0.04)	0.66 (0.06)	0.79 (0.07)
Koide et al. [140]	-	-	-	-	-	0.900
Our Work	0.22 (0.03)	0.29 (0.02)	0.56 (0.05)	0.56 (0.03)	0.58 (0.03)	0.63 (0.06)
Koide et al. [140]	-	-	-	-	-	0.756
Mind-vis [43]	-	-	-	-	-	-
Our Work	0.18 (0.01)	0.23 (0.01)	0.57 (0.02)	0.71 (0.03)	0.58 (0.02)	0.81 (0.01)
Our Work	0.31 (0.02)	0.34 (0.02)	0.70 (0.01)	0.86 (0.00)	0.83 (0.02)	0.93 (0.01)
Lin et al. [158]	-	-	-	-	0.782	-
Takagi et al. [250]	-	-	0.830	0.830	0.760	0.770
Gu et al. [218]	0.150	0.325	-	-	-	-
Ozcelik et al. [190]	0.254	0.356	0.942	0.962	0.872	0.915
Scotti et al. [236]	0.309	0.323	0.947	0.978	0.938	0.941

Table 1.1: Comparison of various metrics across different studies for semantic image retrieval from fMRI data, with performance metrics normalized to a 0-1 scale and the number of classes considered in each study.

(only on the GOD dataset) Fig. 1.4 presents a comparison with state-of-the-art reconstruction approaches over the same dataset, demonstrating qualitative differences between our approach and others. Our diffusion model generates images that are crisp and sharp and convey clear and specific content, which helps recognize similarities between images and distinguish between failed and successful semantic reconstructions.

Fig 1.5, Fig 1.7 and Fig 1.8 present a visual comparison of images shown to participants (left column) versus images retrieved by the decoding model for the different datasets. Each row displays the originally shown image alongside the top three images retrieved by the model for each subject. The retrieval is based on semantic similarity and the decoded images are broadly similar in content to the shown image. For instance, across all the datasets, the decoder has retrieved images of different animals and objects that share some visual characteristics with the target image, such as form and context. The retrieved images do not always match the target perfectly, highlighting the challenges of decoding visual perception from fMRI data. However, the model does show some success in capturing the semantic content, as many of the retrieved images are from the correct category (e.g., animals, musical instruments). Fig 1.6 shows a similar experiment run for the imagery part. Since imagery involves some common process with vision but brain activation could be different, performances are slightly lower as compared to vision perception. On the left column, we show

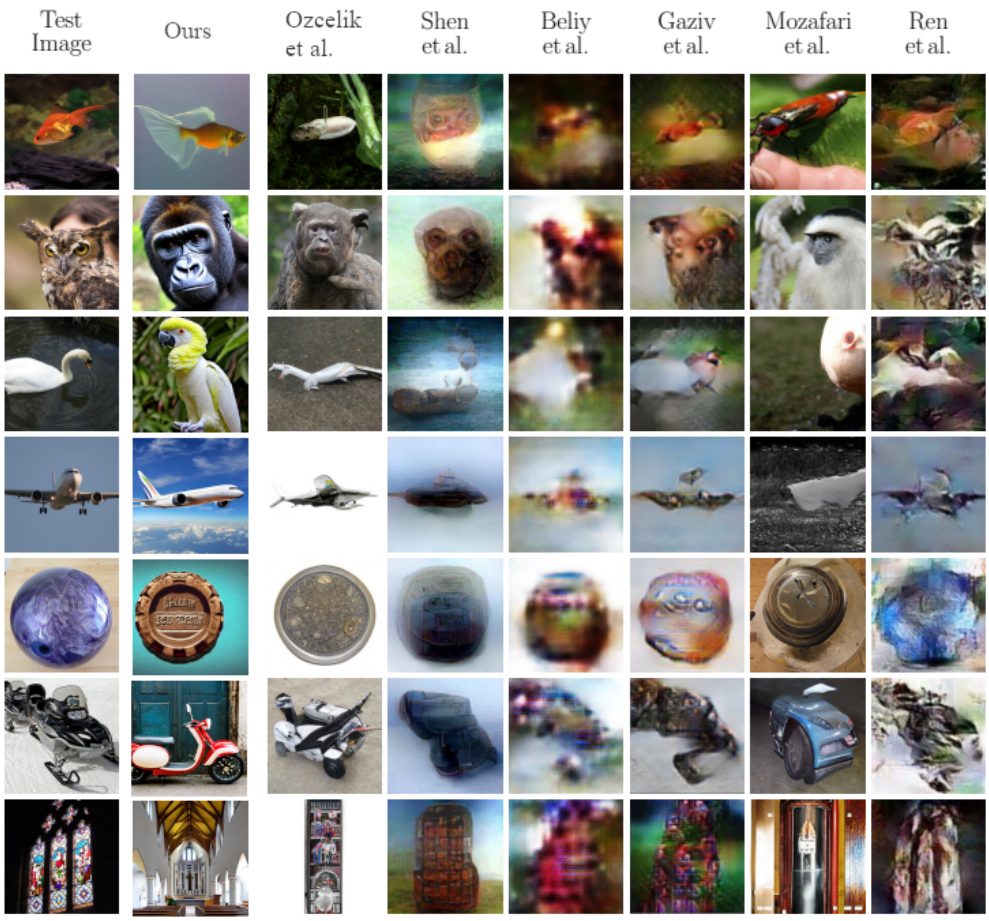


Figure 1.4: Comparison with previous approaches in brain decoding (image generation) of visual stimuli over the GOD dataset. The first column shows original images used as stimuli, while other columns are reconstructions from different works. Our results are depicted in the second column.

candidate images that subject in the GOD dataset are required to imagine and recall and the other groups of columns show results of our retrieval procedure. It is worth noting that a broad semantic category can be decoded even for imagery (animals, objects) and sometimes the right images is found by the algorithm. We propose a paradigm shift in our approach to reconstruction. Rather than focusing on obtaining accurate reconstructions in pixel space, we aim to produce novel images that are semantically and contextually as close to the target visual stimulus as possible. For instance, reconstructions of "fish" and "airplane" (see Fig. 1.4, first and fourth rows, with the first column showing original images and the second column showing our reconstructions) are among our best results, as

they clearly portray the same concepts as the original image. Other images that match the stimulus on a semantic level, such as the swan that is reconstructed as a parrot (both birds), the snowmobile that is reconstructed as a motorbike (both vehicles), or the colorful church window reconstructed as a church, are instances of visuals that match the content and context without being exact pixelwise reconstructions.

In the Supplementary Material we show additional reconstruction examples for all subjects. One can see that our model provides a plausible reconstruction that matches the original at some contextual level in the majority of cases, although with a natural degree of variation that reflects the breadth of possible semantic similarities.

1.4.b Quantitative semantic distance

The performance evaluation of semantic image retrieval from functional Magnetic Resonance Imaging (fMRI) data is provided in Table 1.1 Our methodology demonstrates competitive performance as compared to state-of-the-art methods across varied datasets. Notably, our approach employs a straightforward yet flexible strategy that prioritizes the semantic content of the images. Focusing on the GOD dataset under the Visual modality, our model achieved a Top-1 accuracy of 20% (chance level 2%) for image retrieval and a Top-3 accuracy of 40% (chance level 6%), revealing a robust capability to capture the essence of the visual stimuli presented to participants. Within the realm of Imagery, our approach yielded a Top-1 accuracy of 8% (chance 2%) and a Top-3 accuracy of 20% (chance 6%), which, despite the apparent lower accuracy in comparison to visual modalities, reflects a significant performance given the complexity of the task. Our model's performance is compellingly close to that of [140], who reported an identification accuracy of 0.756, however using 10 classes only in a different version of the GOD dataset. we tackled 50 possible imagery classes, achieving an identification accuracy of 0.63. The BOLD5000 dataset, explored under the Visual modality yielded a Top-1 accuracy of 0.5773 (chance 0.04) and a Top-3 accuracy of 0.8257 (chance 0.12); our method outperformed Mind-Vis's Top-1 accuracy of 0.334. Such a comparison underscores the efficacy of our method in managing datasets with a reduced number of classes, thereby highlighting its adaptability and precision. In the evaluation on the NSD dataset, we achieved a Top-1 accuracy of 0.7225 (chance 0.04) and a Top-3 accuracy of 0.8854 (chance 0.12). All the other metrics were measured on the first retrieved image for each test sample. While focusing only on the semantics of the image, our results remain competitive when compared to other works on the same dataset [158, 250, 218, 190, 236]. Please note that, while we reported all the metrics to enable better

comparisons, PixCorr and SSIM only reflect retrieval performance (1 = perfect retrieval, with a decreasing value as images that differ in shape and structure are picked).

About generated images for the GOD dataset, we achieved a FID score of 10.58 ± 1.95 (mean $\pm$ standard deviation, test set) and an average Wu-Palmer distance of 0.811 ± 0.204 over the training set and 0.571 ± 0.157 over the test set (Fig 1.9). It is important to note that the images in the test set correspond to categories that do not overlap with those in the training set. Therefore, the quality of prediction in the test set is determined by the number of features shared by the two sets. However, there is a notable factor of similarity between original and generated images, even in the test dataset, suggesting that the brain-to-feature model can estimate semantic features related to groups of objects, such as wings, fur, and buildings, correctly. This result holds even though the model is trained on data with different categories and data distribution. In other words, our model performs well in spite of the non-overlap between training and test categories. While a simple classifier would likely not be able to generalize to this particular test set, our model performs well and demonstrates the potential for brain decoding to generalize to new categories and data distributions.

1.4.c Human Evaluation

Humans perform well in complex assessments with wide criteria and can naturally examine images at numerous levels of semantic information as well as shapes, colors, and many more. Fig. A1 and Table 1.2 show the results of human evaluation for both the training and test sets. On average, human observers selected the images generated from the model (as opposed to the randomly generated images) in $95 \pm 3\%$ of the cases for images from the training set and in $81 \pm 4\%$ of the cases for images from the test set. In all cases, human observers chose the model-generated images far more frequently than what would have been the chance level, supporting the hypothesis that our computational approach can correctly capture various semantic features of the images in a manner that corresponds well to the way the human brain evaluates this type of content and context. Furthermore, the inter-rater reliability, as measured by the Fleiss Kappa coefficient, provides additional validation of the consistency of human evaluations. Fleiss Kappa values were calculated for both training and test datasets, with coefficients ranging from 0.131 to 0.809 for the training set and 0.142 to 0.625 for the test set. These Kappa values indicate a range from slight to substantial agreement among raters.

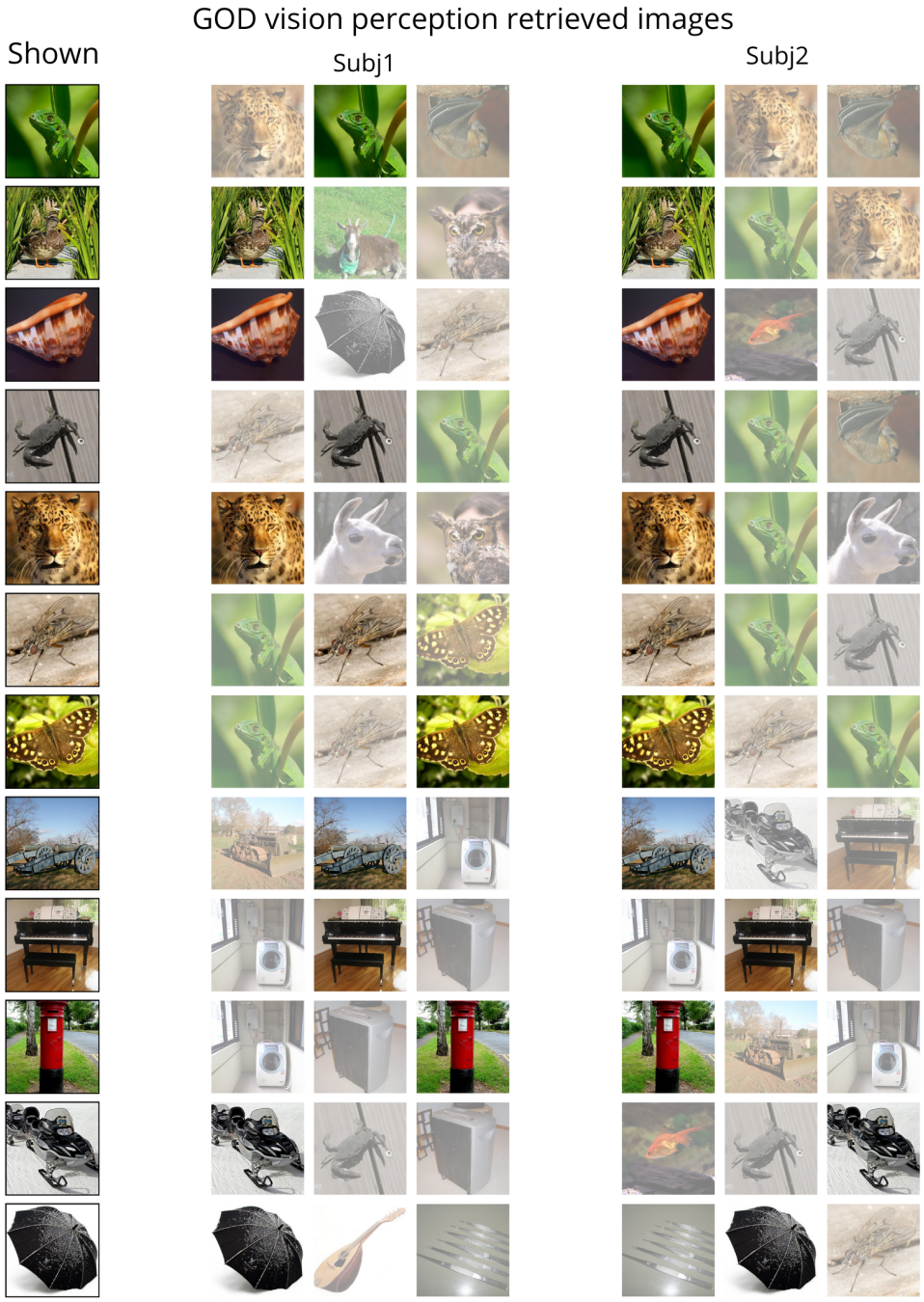


Figure 1.5: Examples of retrieved images for the GOD dataset. The left column show the stimuli presented during the experiment, while the other two groups of columns show the retrieved images from fMRI activity for two subjects.

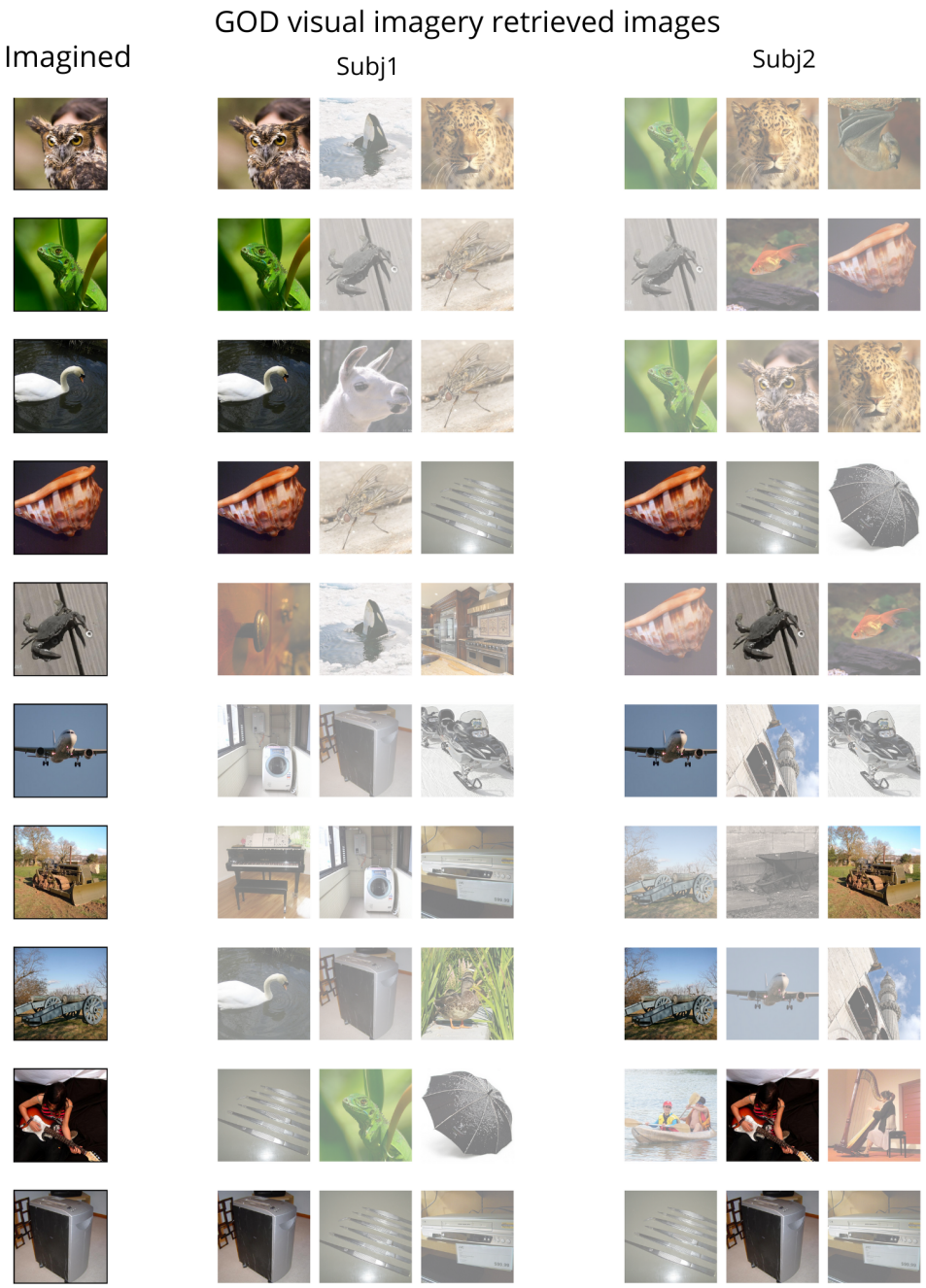


Figure 1.6: Examples of retrieved images for the GOD dataset imagery experiment. The left column show the stimuli that subject are required to imagine during the experiment, while the other two groups of columns show the retrieved images from fMRI activity for two subjects.

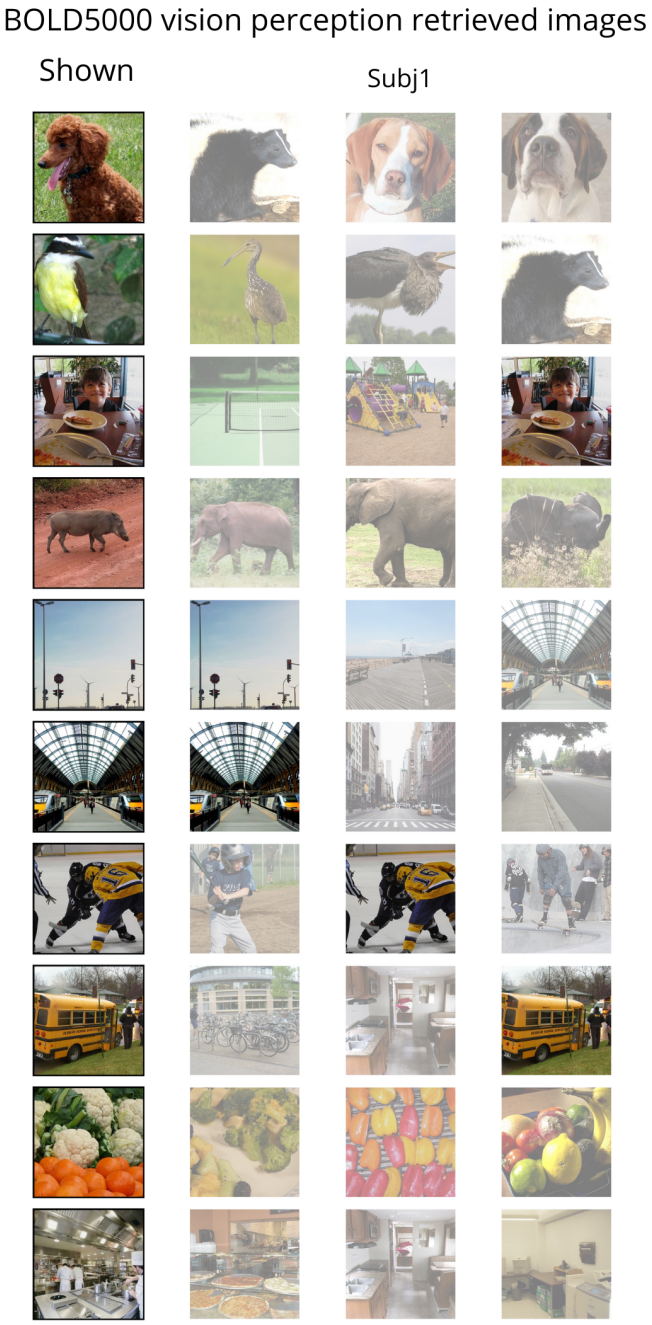


Figure 1.7: Examples of retrieved images for the BOLD5000 dataset. The left column show the stimuli presented during the experiment, while the other columns show the retrieved images from fMRI activity for an example subject.

Figure 1.8: Examples of retrieved images for the NSD dataset. The left column show the stimuli presented during the experiment, while the other columns show the retrieved images from fMRI activity for an example subject.

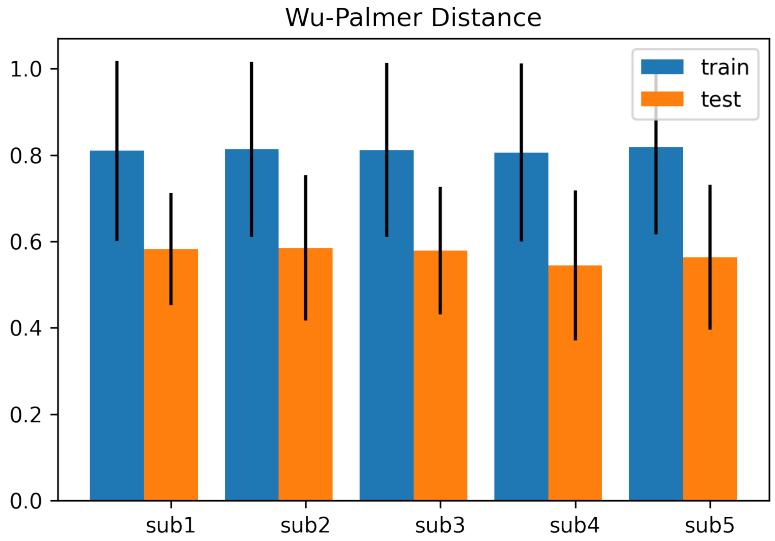


Figure 1.9: Wu-Palmer distances (mean +/- s.d.) between original image stimuli shown to the subjects under fMRI for all subjects for both training (blue) and test (orange) sets.

Subject	Human Evaluation Training Dataset	Human Evaluation Test Dataset	Fleiss Kappa (Train)	Fleiss Kappa (Test)
1	0.960 ± 0.031	0.778 ± 0.031	0.130952	0.524635
2	0.945 ± 0.022	0.880 ± 0.043	0.809244	0.212121
3	0.940 ± 0.028	0.834 ± 0.043	0.690307	0.142340
4	0.943 ± 0.031	0.745 ± 0.042	0.464646	0.624909
5	0.954 ± 0.031	0.797 ± 0.059	0.523787	0.376001

Table 1.2: Results of human evaluation. Rate of selection of images generated by our model versus random images from human evaluators, alongside Fleiss Kappa coefficients reflecting inter-rater agreement.

1.5 Discussion

1.5.a Developing the Brain-to-Feature Model and Reconstruction Pipeline

Grounded in the assumption that fMRI data from the VC during a visual task can be used as a proxy for the last layer of a convolutional neural network (CNN) trained for image classification, we developed a brain-to-feature model. This model is a trained ridge regression between fMRI and image features extracted from the original visual stimuli images through CLIP, establishing univocal relationships between fMRI data and the CLIP features [159, 169, 123, 46].

We subsequently employed a nearest neighbor-like technique to map these features into object semantic "categories." Output of this part can be used for retrieval and decode possible candidates of what was seen during the expeiment from the dataset. For the GOD dataset, these categories were then used to condition a pretrained latent diffusion model to produce novel images from text prompts corresponding to the synset name of the related WordNet class. Our reconstruction pipeline incorporates these hypotheses through the mapping between fMRI and CLIP latent space, the use of the k-nearest neighbors algorithm, and reliance on a powerful image generator.

1.5.b Bottom-Up and Top-Down Processes in Vision

Our brain-to-feature model represents the bottom-up process in vision, a rapid initial estimate of relevant features. This estimate is refined by our top-down approach, represented by the choice of the nearest neighbor in the latent space to condition the generative model. This component of our architecture is supported by prior knowledge of the world, contained in the CLIP latent space representation. This, in turn, allows us to evaluate the "distance" between concepts.

1.5.c Evaluating Performance Through Semantic-Related Measures

We assessed our work both qualitatively (visually) and quantitatively through semantic-related measures. We employed the Wu-Palmer distance to analyze similarities between concepts in the WordNet lexicon, discovering a good average similarity. Additionally, we included an assessment of the contextual distance between original and reconstructed stimuli by naïve human observers to allow for additional flexibility and human-like semantic evaluation. Our results suggested that the model performed well in selecting relevant features and producing images closer to the original than any other image.

1.5.d Reconstruction Performance and Categories

We found that with all assessment techniques, reconstructed images are rarely noticeably distant from the target, similar to the results reported in [116, 43, 236, 140, 190]. Specifically, original images of animals generated reconstructions that accurately depicted other animals, with striking accuracy in high-level features such as "species". Similarly, original images of non-animated objects, such as vehicles, exhibited comparable behavior, giving rise to accurate renderings of planes, motorbikes, tractors, and carriages. While a similar behavior occurred for most visual stimuli, some categories appeared to be “misunderstood” by our model, such as the cowboy hat or the guitar (see Supplementary Material).

In this context, it is possible that the traits associated with certain test images are underrepresented in the training set, increasing the difficulty of capturing relevant semantics.

1.5.e Brain and Deep Learning Models

Our brain can be thought of as a prediction machine that utilizes past knowledge in the form of top-down processing of external inputs. We found that in the VC, this might produce a feature space that is homeomorphic to the latent space of a pretrained neural network. Notably, a linear (ridge regression) model was sufficient to achieve convincing reconstruction results. These findings are in line with evidence that deep learning models and brain activity prompted by language converge [36, 35, 92, 169] in terms of behavioral, physiological, and fMRI data, supporting our key hypothesis that context and semantics play a significant role in how we process sensory information. These ideas bear similarities to the concepts of attention-based deep learning models with convolutional layers.

1.5.f Semantic Cognition and Reconstruction

Semantic cognition refers to a group of neuropsychological processes that sustain not only conceptual representation and formation but also the manipulation of semantic knowledge to influence context-relevant behavior. These brain mechanisms are thought to depend on a constant flow of top-down and bottom-up interactions between posterior and anterior areas, including occipito-temporal cortices and prefrontal networks. In the visual domain, the 'bottom-up' and 'top-down' interplay between multiple occipitotemporal cortices might allow the 'distillation' of a latent space of features that are believed to be at the 'core' of semantic representation. Our reconstruction approach, which used a combination of brain-to-feature and generative models, allowed us to recreate the original visual stimuli and obtain reconstructions of the images that surpass the state-of-the-art in the literature, particularly at the semantic level of reconstruction. This supports our approach's validity and its ability to mimic the way the human brain extracts, categorizes, and internally represents visually acquired information. We employed a deep latent diffusion model to generate novel images that could evoke similar brain activity, featuring images with congruent semantic content. This capacity to synthesize images with precise content directly from brain activity lays the foundation for more advanced analyses and reconstructions. For instance, using an image-to-image diffusion model that starts with an initial guess containing low-level aspects such as colors and shapes can lead to more accurate and plausible reconstructions.

1.5.g Neurobiological considerations

It is important to note that our deep learning architecture is conceptually inspired by the current understanding of the neural mechanisms underlying semantic cognition. However, our model only employed fMRI data from a group of visual cortices (V1, V2, V3, V4) due to practical and computational considerations [213]. This choice does not deny the critical role of other brain regions, such as the anterior temporal lobe (ATL), in semantic cognition. The spoke-hub theory of semantic cognition clearly states that semantic cognition arises from the interplay of modality-specific (sensory, motor cortices) and a-modal regions (ATL, prefrontal cortex, etc.) [196]. Future investigations could explore the role of other brain regions, such as the ATL, and determine whether the features extracted from those regions are superior to those of other regions of the brain when decoding the "mental states" associated with visual processing.

1.5.h Limitations

The fMRI experiments used to collect the data were restricted in length because individuals need to be exposed to images slowly enough for the brain response to stabilize. As a result, the applicability of end-to-end deep learning algorithms is limited. In addition, because the categories in the training and test sets in the dataset we used do not overlap, the model's performance depends on the relationship between the fMRI data and image features in the training set. The assumption is that this relationship is sufficient to detect variations in unseen categories. Our model demonstrated good generalization capabilities, suggesting that semantic feature content, rather than precise train/test class overlap, may be predominant in determining performance. However, if the categories are highly dissimilar between the test and training set, it is conceivable that their essential properties are underrepresented in the training set, limiting the model's performance capabilities in the test . Also, it is important to note that we outperformed all models trained on the same 3T dataset, hence potentially widening the applicability of our methods to an extremely large number of centers wich do not have access to ultra-high-filed (7T or more) scanners.

Furthermore, there are numerous potential sources of error that can arise between the vision process and the generation of the image feature space. These include fMRI acquisition noise, bias in the feature space of the ResNet50 architecture, bias introduced by the limited sample size in the brain-to-feature model, and errors introduced by the conditioning algorithm. These circumstances can be responsible for cases where the performance of our model in reconstructing context is poor.

Additionally, mental attention may warp the semantic space in the human brain [50]. When subjects become tired or bored during fMRI sessions, the encoded stimuli may change, introducing another source of variability that is not under experimental control. These limitations and sources of error should be taken into account when interpreting the results and considering future research directions.

1.5.i Future directions

Future research could delve into the role of mental attention in semantic cognition and examine whether attentional states can modulate the distributed neural representations of semantic concepts in the visual cortex [50]. Such investigations would contribute to our understanding of how attention influences decoding accuracy and the neural mechanisms underlying semantic cognition. A critical element of this study pertains to interpretability. Our methodology employs neural networks as opaque mechanisms to extract latent representations of stimuli, followed by a model to approximate these representations based on brain activity. Enhancing interpretability in future works may provide insights into the underlying mechanisms of the brain. One potential approach to investigate brain processes involves utilizing encoding models, which reverse the pipeline to identify brain regions that may be effectively modeled as responsive to latent image representations. Employing such models to categorize images could facilitate virtual experiments and enable exploration of brain patterns. The growing availability of extensive open fMRI datasets will likely enable us to enhance brain decoding results using diffusion models as image generators, by conditioning these models in various ways. Interestingly, the majority of work in this field, including our own, currently focuses on subject-wise reconstruction. It would be intriguing to develop models capable of decoding intra-subject activity. This could pave the way for large-scale decoding on new subjects by merely fine-tuning a more extensive model, thus bypassing the need for lengthy fMRI acquisitions for each individual. Another crucial next step is improving on decoding of imagery activity, reconstructing examples of images seen exclusively by the mind's eye. Naturally, this raises ethical concerns regarding privacy and confidentiality, as decoding brain activity entails accessing an individual's internal mental state, potentially revealing sensitive information about their thoughts, emotions, and behavior. There is a risk that such information could be misused or disclosed without the person's consent, leading to privacy breaches. Ethical questions also arise concerning the accuracy of decoded images, which may produce a distorted version of a person's perception due to model imperfections. Nonetheless, this type of research can lead to numerous beneficial applications.

For instance, a completely new form of art could emerge from the interaction between the physics of fMRI acquisition, the artist's thoughts and perceptions, and the artificial intelligence used for decoding. This technology could also enable individuals with locked-in syndrome to communicate through images.

Moreover, future investigations could employ other brain imaging modalities, such as EEG or MEG, to investigate the temporal dynamics of the neural representations of semantic concepts and how they evolve over time during visual processing. Additionally, future studies could employ multi-modal data fusion methods to combine fMRI data with other modalities, such as behavioral data or natural language descriptions of visual stimuli, to gain a more comprehensive understanding of the neural basis of semantic cognition.

1.6 Conclusions

Our study proposes a pipeline based on semantic in brain space that can help us to retrieve or synthesize images that are conceptually and similar to the original stimuli, starting from fMRI data only. We reported experiments with three public available independent datasets, observing good performances across vision and imagery among all datasets despite different image acquisition paradigms, MRI field strenghts, subjects and other source of variability. We assume that measurable neural correlates can be linearly mapped onto the latent space of a convolutional neural network that represents a semantic description of the image. The overall objective is to replicate the way humans process information by combining bottom-up visual inputs with top-down cognitive descriptions of the environment, which is known to aid in "classification" processes in the brain. In summary, our study provides evidence that measurable neural correlates can be linearly mapped onto the latent space of a multimodal neural network to retrieve and synthesize images that are conceptually and semantically similar to the original stimuli. The findings have implications for both cognitive neuroscience and artificial intelligence, as they shed light on the neural mechanisms underlying visual perception and suggest promising avenues for future research.

Appendix

In the presented table A1, we report the findings of our comparative study which aimed to evaluate the efficacy of linear versus nonlinear mapping methods for decoding fMRI activity into semantic image features. The Root Mean Square Error (RMSE) values for both Multi-Layer Perceptron (MLP) and Ridge regression models are listed across three distinct datasets: GOD, BOLD5000, and NSD. Both models were optimized with a cross-validation procedure. The RMSE values indicate the average deviation between the predicted and the actual values, with lower values suggesting a better fit. For all datasets, the linear Ridge regression model yielded a significantly lower RMSE compared to the nonlinear MLP model. This suggests that in the context of mapping fMRI activity to semantic image features, a linear approach (Ridge regression) was superior. The p-values reported in the table are derived from a statistical paired t-test and support the conclusion that the differences observed are statistically significant. The p-value threshold for significance was not specified, but given all p-values are below 0.01, we can infer that the results are significant at least at the 1% level. This evidence suggests that a linear mapping between fMRI activity and semantic image features is more effective than a nonlinear mapping for the datasets examined. This is notable because it challenges the often held belief that the complex nature of brain activity requires equally complex models for accurate decoding. Instead, our findings underscore the potential of linear models in capturing the essence of neural representations of visual stimuli.

Dataset	RMSE MLP	RMSE Ridge	P-Value	Significant
GOD	1.0899 (0.0602)	0.9659 (0.0327)	0.001851	Yes
BOLD5000	1.0488 (0.0295)	1.0161 (0.0228)	0.008519	Yes
NSD	0.6676 (0.0227)	0.7291 (0.0222)	0.000172	Yes

Table A1: Comparison of RMSE values between MLP and Ridge regression models across three datasets.

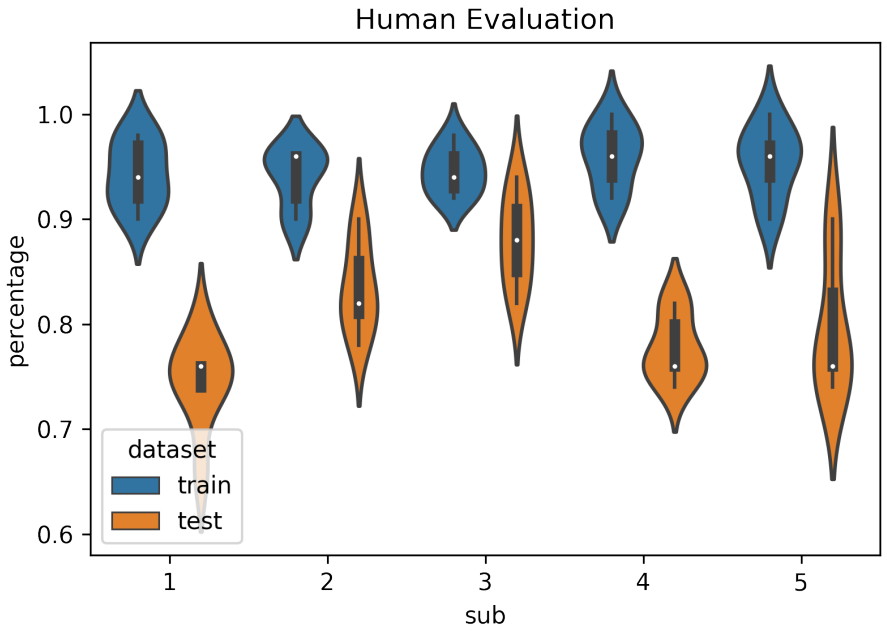


Figure A1: Human evaluation: Rate of selection (mean $\pm$ std) of images generated by our model versus random images from human evaluators for images in the training (blue) and testing (orange) set for the GOD image generation experiment.

2 Cross-modal Brain Decoding

Every day, the human brain processes an immense volume of visual information, relying on intricate neural mechanisms to perceive and interpret these stimuli. Recent breakthroughs in functional magnetic resonance imaging (fMRI) have enabled scientists to extract visual information from human brain activity patterns. In this Chapter[1], we present an innovative method for decoding brain activity into meaningful images and captions, with a specific focus on brain captioning due to its enhanced flexibility as compared to brain decoding into images. Our approach takes advantage of cutting-edge image captioning models and incorporates a unique image reconstruction pipeline that utilizes latent diffusion models and depth estimation.

We analyzed the Natural Scenes Dataset, a comprehensive fMRI dataset from eight subjects who viewed images from the COCO dataset. We employed the Generative Image-to-text Transformer (GIT) as our backbone for captioning and propose a new image reconstruction pipeline based on latent diffusion models. The method involves training regularized linear regression models between brain activity and extracted features. Additionally, we incorporated depth maps from the ControlNet model to further guide the reconstruction process.

We propose a multimodal based approach that leverages similarities between neural and deep learning representations and by learning alignment between these spaces, we produce textual description and image reconstruction from brain activity.

We evaluate our methods using quantitative metrics for both generated captions and images. Our brain captioning approach outperforms existing methods, while our image reconstruction pipeline generates plausible images with improved spatial relationships.

In conclusion, we demonstrate significant progress in brain decoding, showcas-

ing the enormous potential of integrating vision and language to better understand human cognition. Our approach provides a flexible platform for future research, with potential applications based on a combination of high-level semantic information coming from text and low-level image shape information coming from depth maps and initial guess images.

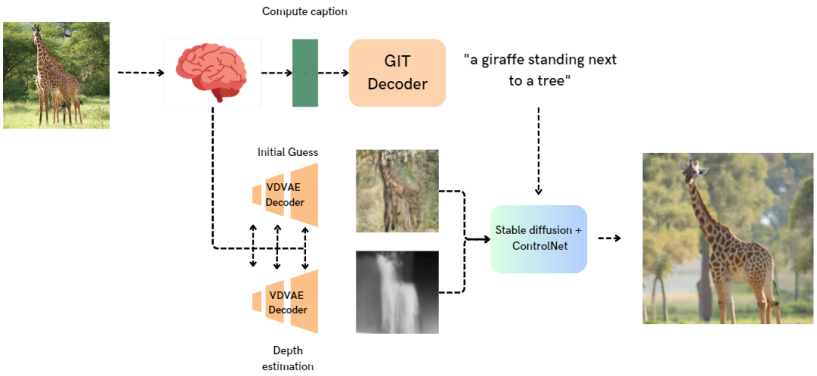


Figure 2.1: Our model utilizes fMRI measurements to extract features for GIT captioning and VDVAE initial and depth image estimation using linear models. Image captions serve as the primary general result, used in the second stage alongside other conditioning to generate plausible reconstructions with a latent diffusion model. GIT and VDVAE models are pre-trained and frozen, while linear regressions are trained from fMRI to their latent spaces.

2.1 Introduction

The human visual system is an extraordinary product of evolution, enabling us to navigate and interact with our surroundings. From basic patterns to intricate scenes, our brains persistently process and interpret visual information. A central challenge in neuroscience is comprehending how these elaborate processes occur at the neural activity level. Functional magnetic resonance imaging (fMRI) has emerged as an essential tool for studying neural activity associated with visual perception, by measuring blood oxygen level-dependent (BOLD) signals. Brain decoding has progressed significantly, employing fMRI data to reconstruct visual stimuli from brain activity patterns. This has the potential to revolutionize our understanding of the neural code underlying visual perception with possible applications in brain-computer interfaces and clinical diagnostics. The increasing interest in reconstructing information from noninvasive brain data is driven by enhanced data availability, improved computational power, and

[1]The work presented in this chapter has been presented at UniReps workshop at NeurIPS 2023 and published in the relative "Proceedings of Machine Learning Research" [78].

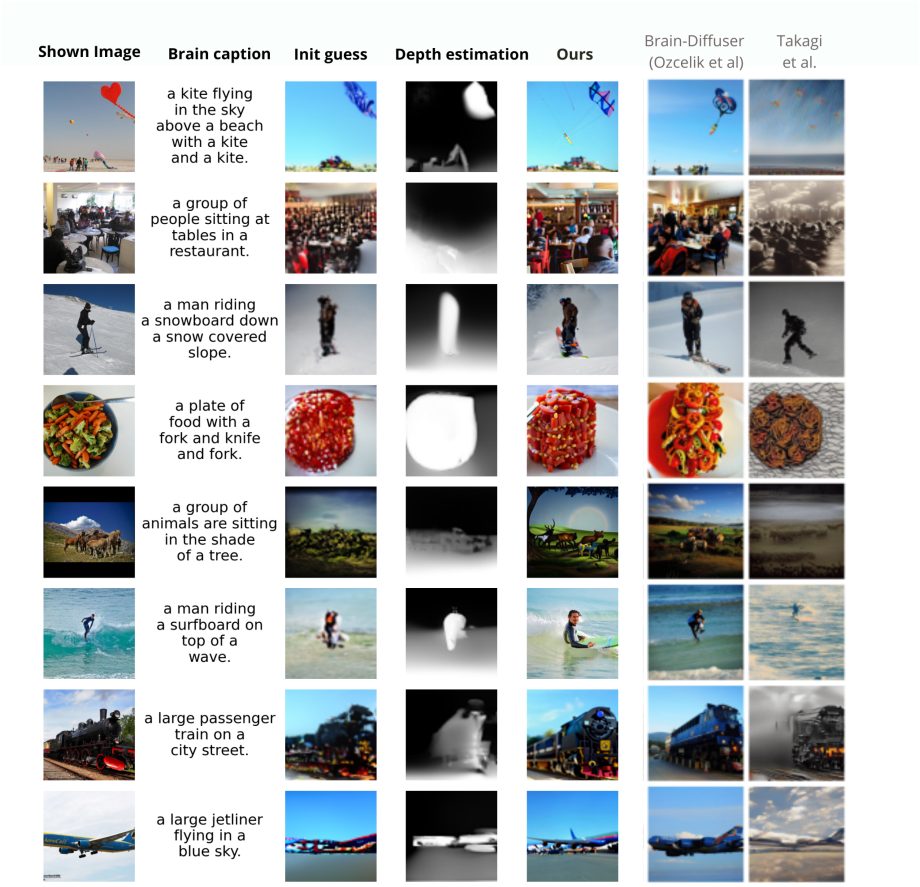


Figure 2.2: Comparison of our results (Columns 2-4) with the shown stimuli and reconstructions from other works. The second column displays the caption computed from the brain activity, the third column presents the initial guess image, the fourth column shows the depth estimated images, and the fifth column reports our final reconstruction. The last two columns showcase reconstructions from two recent works. All results are from subj01.

sophisticated deep learning methods. Despite challenges with signal-to-noise ratio, session duration, and hemodynamic response function variability, fMRI has proven effective in various tasks such as visual stimulus and text classification and reconstruction [231, 283, 159, 9].

In this work, our first contribution is shifting the prediction from images to text, aiming to generate a caption of the observed scene from brain activity. To compare with prior work, we propose a new model for image captioning from brain activity and propose a new image reconstruction pipeline based on a conditioned and controlled version of the latent diffusion model, Stable Diffusion. Predicting a caption instead of the image in brain decoding from fMRI of visual stimuli offers several advantages. Captions naturally represent a higher level of

abstraction, requiring a more advanced interpretation and summarization of visual information than merely predicting the image itself. As a result, predicting captions can help us understand how the brain processes and represents complex visual information. In real-world situations, humans often describe visual scenes with words, so predicting captions instead of images may better capture an important aspect of visual information processing. Recent neuroscience research has shown substantial evidence that large language models can be correlated with brain activity and that it is possible to predict one representation from the other [37, 251]. Finally, predicting text from fMRI could lead to better generalization across modalities. Natural language is our main tool as humans to interact with each other and nowadays even with foundation models. We can exploit large language models to condition other models to generate images, videos, audio, and more. Predicting text from brain helps us rapidly change the reconstruction model, leveraging state-of-the-art text-to-image models to generate realistic images from brain activity. In summary, our contributions in this paper are two-fold: We propose a method to generate image captions from brain activity using a multimodal large language model [266] and introduce a novel image reconstruction pipeline based on predicted text and estimated initial and depth maps from brain activity. The main novelty proposed in our work is a pipeline that leverage aligned representations of brain, text and images for visual stimuli. Fig 2.1 is a scheme of the entire procedure that we propose, while Fig 2.2 shows generated captions and images from brain activity compared to other image reconstruction methods.

2.1.a Related Works

In the field of brain decoding, researchers have utilized various modeling frameworks with preprocessed fMRI time series as input. These data have served as the basis for numerous decoding approaches. Some examples include employing a variational autoencoder with a generative adversarial component (VAE-GAN) to encode latent representations of human faces [262] and applying sparse linear regression on preprocessed fMRI data to predict features extracted from early convolutional layers in a pre-trained CNN [116] for natural images. Unsupervised and adversarial strategies have been used to reconstruct images, incorporating dual VAE-GAN and unsupervised methods for fMRI stimuli decoding with various encoders and decoders trained in different ways [237, 218, 85]. Optimizing the latent spaces of pretrained architectures, such as BigBiGAN and IC-GAN, can facilitate reconstructing high-quality images from fMRI patterns [64, 31, 175, 189]. Recently, diffusion models have become a significant component of the decoding pipeline due to their improved performance in image generation [250,

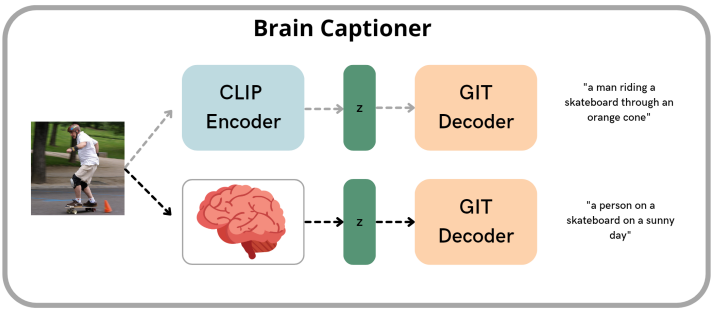


Figure 2.3: Image captioning from brain activity pipeline: Gray dotted lines are only used during training, and only orange boxes are used during inference, replacing their inputs with those estimated from brain activity.

43], also incorporating semantic-based strategies like [79] or multi-step decoding strategies as in [190, 44, 236, 251]. To the best of our knowledge, only a few works [249, 169, 207] have attempted brain captioning, utilizing a combination of a pre-trained convolutional neural network and recurrent neural network for captioning and estimating the convolutional features from brain activity. The primary differences between our work and previous research are the shift in paradigm from direct image estimation to brain captioning and leveraging multimodal transformer-based language models, which have been shown to better describe brain activity [46].

2.2 Methods

In this section, we describe the proposed method and the data we used. The data are publicly available and can be requested at `https://naturalscenesdataset.org/`. All experiments and models were trained on a server equipped with four A100 GPU cards and 2 TB of RAM. The entire analysis took approximately 16 hours per subject. The pipelines are based on pre-trained versions of deep learning models used as proxies for brain activity, generating latent representations that could be similar (and thus linearly mapped) to brain activity and vice versa.

Data:We employed the Natural Scenes Dataset (NSD) [5], a comprehensive fMRI dataset featuring eight subjects who viewed images from the COCO dataset. Our analysis concentrated on four subjects (same used in other decoding works for comparison), yielding a training set of 8,859 images and 24,980 fMRI trials, and a test set of 982 images and 2,770 fMRI trials per subject. Images are repeated up to three times and their trials were averaged to increase signal-to-noise ratio. To reduce spatial dimensionality to approximately 15,000 voxels, the fMRI signal (1.8mm resolution) was masked using the NSDGeneral

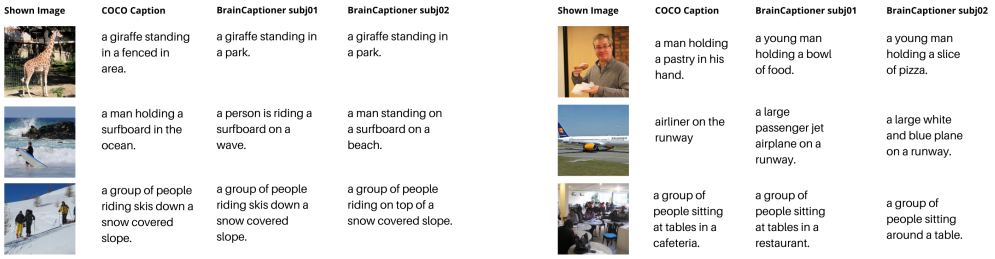


Figure 2.4: Examples of generated caption with our BrainCaptioner pipeline. Shown images are test set stimuli used for subj01 and subj02 during the fmri experiment. COCO Caption column report the first annotations for the original COCO image, while the other two columns are the output of our model for the two subjects.

ROI mask, which covers numerous visual areas. This ROI selection is vital for enhancing the signal-to-noise ratio and minimizing data complexity. The chosen ROI mask facilitated the investigation of both low-level and high-level visual features. To decrease temporal dimensionality, we employed precomputed betas from a GLM with fitted HRF and denoised as described in the NSD paper.

Captioning model and renormalization: For brain captioning, we utilized the state-of-the-art image captioning model, GIT [266], as our backbone. GIT (Generative Image-to-text Transformer) is an innovative model designed to integrate vision and language tasks. In contrast to conventional approaches that depend on intricate architectures and external modules, GIT adopts a streamlined structure consisting of a single image encoder and a text decoder, unified under one language modeling task. Leveraging large-scale pre-training data and model size, GIT outperforms existing models on 12 benchmarks and even surpasses human performance on TextCaps. Essentially, GIT comprises a CLIP Vision encoder [210] followed by a GPT decoder, trained on large-scale datasets. For the stimuli in the train set, we computed features from images and trained a regularized linear regression to map between brain activity and these features. We used cross-validation to select the best regularization parameter α and discovered that a value of 50,000 performed optimally using the negative mean squared error as a scoring function. This is our brain-to-features model, which serves as the core component of our method for brain captioning. Before feeding estimated features to the decoder, we required a normalization pass. Thus, we computed the mean and standard deviation of features from images and those predicted by the model over the training set, replacing their values during inference on the test set to match the real feature distributions. A schematic representation of the overall pipeline can be seen in Fig. 2.3 and generated captions from this pipeline for both subjects are shown in Fig 2.4.

Reconstruction pipeline: Recent research in brain decoding has focused on developing image reconstruction techniques [190, 189, 250, 158, 43]. Studies have demonstrated that high SNR fMRI data of visual stimuli enables effective brain decoding using diffusion models. Various approaches have been proposed to enhance these models' performance, with the optimal method for image reconstruction remaining an open question. One approach to improve low-level detail generation and increase the similarity between original and decoded images is to provide the network with an initial guess image or an estimated latent space.

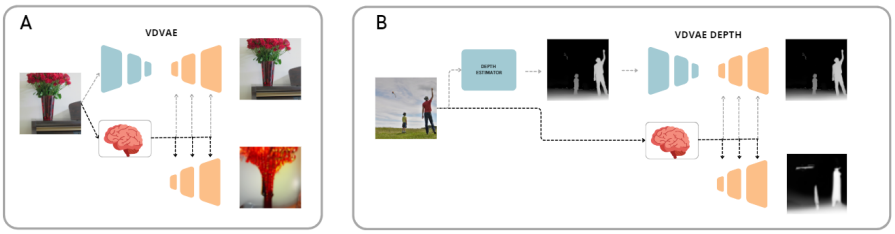


Figure 2.5: A: Pipeline for initial images capturing 2D RGB pixel information. **B**: Pipeline for inferred depth estimates. Both depth image and the initial image are estimated from brain activity. Gray dotted lines are only used during training, while only orange boxes are used during inference, replacing their inputs with the ones estimated from brain activity.

Initial Guess: To compare our approach with existing research on brain decoding, we augmented our method by proposing an image reconstruction pipeline based on latent diffusion models. Following the approach described in [190], we initially estimate a "guess image" to generate an approximate initial image with colors and shapes. To achieve this, we computed the latent representations of the first 31 layers of the very deep variational autoencoder model [45] (VDVAE), pre-trained on natural images, and kept frozen. In a VDVAE, the encoder network maps the input data onto a lower-dimensional latent space, while the decoder network maps the latent space back to the original data space. The architecture of the VAE is hierarchical. In other words, the hidden units in each layer depend not only on the input data but also on the outputs of the previous layer. This conditional dependence allows the VAE to capture complex relationships between the input data and the latent space, resulting in a more powerful and expressive model. Consequently, we trained a regularized linear regression between brain activity and estimated features for each of the first 31 layers, using the renormalization procedure described in the previous section to match the target distribution. During inference over the test set, these features are estimated from brain activity, renormalized, and passed to the VDVAE

decoder to reconstruct an initial image, as depicted in Fig. 2.5.

Depth estimation: We propose using ControlNet [284] to augment Stable Diffusion [223], a state-of-the-art latent diffusion model, for improving foreground-background matching in reconstructed images by incorporating depth information. We first compute grayscale depth images for all training stimuli using Dense Vision Transformer and the Huggingface library [216, 275]. We then pass these depth images into the Variational Diffusion Autoencoder (VDVAE) model and train a regularized linear regression from brain activity to the model's latent, as illustrated in Fig 2.5.The VDVAE is the same used before (pre-trained on natural images and kept frozen), however here it is here to generate latent representation of the estimated depth images, which are our target for regression.

Whole Reconstruction pipeline: The pipeline (Fig 2.1) first decodes brain activity into a latent space to generate captions for test stimuli using learned ridge regression. Then, the initial guess and depth images are computed from brain activity to condition the latent model. Stable Diffusion v2 + ControlNet is used for implementation, with 30 inference steps, guidance scale 9, and control net weight 0.8. The negative prompt sentence *is* also included to improve quality.

Evaluation: We compared our brain captioning work with existing methods by re-implementing the architecture from [249], consisting of a CNN followed by an LSTM. We used Ridge regression to map brain activity to the CNN's final convolutional layer and applied renormalization before feeding the LSTM. We evaluated the generated captions using metrics such as METEOR, CLIP similarity, and SentenceTransformer similarity. Additionally, we assessed our image reconstruction pipeline using low-level and high-level metrics like PixCorr, SSIM, 2-way accuracy in AlexNet, Inception, and CLIP latent spaces, and FID, allowing comparison with other brain decoding studies.

2.3 Results

Table 2.1 presents the results of the evaluation of the proposed approach compared to the baseline models and previous works. This table reports text-based metrics, including Meteor score, CLIP, and SentenceTransformer similarity, computed for the reference captions, captions generated from images by both models (baseline and proposed), and captions generated from brain activity using the proposed approach. Results show that our approach outperforms the baseline models on all metrics and achieves significantly higher scores than previous works, indicating the effectiveness of the approach in generating accurate and meaningful captions from brain activity.
The table 2.2 reports image-based metrics, including PixCorr, SSIM, accuracy in

Metric	Baselines				**Ours**			
	subj01	subj02	subj05	subj07	subj01	subj02	subj05	subj07
Meteor (image vs human)	0.176	0.174	0.177	0.175	0.404	0.404	0.404	0.404
Meteor (brain vs image)	0.163	0.166	0.166	0.166	**0.305**	**0.298**	**0.303**	**0.291**
Sentence (image vs human)	0.319	0.315	0.321	0.315	0.703	0.703	0.703	0.703
Sentence (brainvs image)	0.280	0.281	0.282	0.281	**0.447**	**0.418**	**0.443**	**0.413**
CLIP (image vs human)	0.672	0.673	0.676	0.673	0.831	0.831	0.831	0.831
CLIP (brain vs image)	0.624	0.627	0.626	0.627	**0.705**	**0.688**	**0.702**	**0.693**

Table 2.1: Text Metrics Comparison: This table reports the values of various metrics for each subject, both for the baseline and our model (columns). Each row represents a different metric. Metrics labeled with "(image captions and human captions)" evaluate the model-generated captions from images against the original COCO captions, serving as a comparison of the model's performance. Metrics labeled with "(brain captions and image captions)" pertain to captions computed from brain activity.

various layers of AlexNet and Inception, CLIP similarity, and FID score. Results show that the proposed approach outperforms the previous works in low-level metrics, including PixCorr, SSIM and the lower layer of AlexNet. High level metrics are on par or slightly lower than state-of-the-art methods, probably due to a bottleneck in text predictions. If a word is predicted wrongly, this error is propagated in the image reconstruction pipeline and impacts on high-level metrics. Overall, the results demonstrate the effectiveness of the proposed approach in decoding brain activity into meaningful images and captions, performing on par on even outperforming state-of-the-art in several metrics. Fig 2.2, 2.6, 2.4 and figures in the supplementary material show some visual comparison with other works for a qualitative comparison. Qualitatively, the captions represent plausible descriptions of images matching the high-level semantic content in most of cases. Sometimes, captions are more general with descriptions like "animals in the grass" instead of the specific type of animal. In other cases, only details are missing (or wrong). For example, in Fig 2.4 for the surfer image for one subject, the model adds "on a wave" while for the other the model specifies "on a beach". Similarly, in the first image of the right part, the pastry in the man's hand is changed to "a bowl of foods" or "slice of pizza". This could support the hypothesis that our pipeline is able to capture the main characteristic of the images from brain activity and the GIT decoder help in plausible sentence decoding.

2.4 Discussion

In this study, we proposed a method to generate captions from brain activity measured during a vision task. The primary motivation for shifting from image reconstruction to image captions is the flexibility of manipulating text prompts

Model	Low level metrics			High level		
	PixCorr	SSIM	AlexNet (2)	AlexNet (5)	Inception	CLIP
Lin et al (2022)	-	-	-	-	0.782	-
Takagi et al (2022)	-	-	0.83	0.83	0.76	0.77
Gu et al (2023)	0.15	0.325	-	-	-	-
Ozcelik et al (2023)	0.30	0.28	**0.89**	**0.98**	**0.92**	**0.94**
Our Model	**0.353**	**0.327**	**0.89**	0.97	0.84	0.90

Table 2.2: Image Metrics Analysis: Metrics from Ozcelik et al were recomputed by requesting images from subj01 and subj02 from the authors and averaging them to facilitate comparison with our results. Metrics from other works are cited directly from the original articles.

and the ease of modifying the image reconstruction pipeline as separate modules. We also proposed an image reconstruction pipeline that incorporates depth maps and initial guesses to generate plausible images. Depth maps provide information about the spatial relationships between objects in a scene injecting information that could improve the overall quality of the reconstructed images.

Neural Art and Examples: Our approach has potential applications in neural art and style transfer. By leveraging our image reconstruction pipeline, we can explore the creative space of combining content and style from different text prompts. This could lead to the generation of visually captivating art, expanding the possibilities for artistic expression using AI. For example, modifying inputs by adding specific styles could drive the diffusion process toward an image with the same content but a different style. This approach represents a novel type of art that combines artificial intelligence, neuroscience, and creativity, starting from the decoded activity of the brain that could be modulated by a text description of the scene.

Ethics: As brain decoding research advances, ethical considerations must be addressed. For instance, the potential misuse of image reconstruction and generative models to create misleading or harmful content raises concerns, given that decoded activity is related to the mental and internal states of someone. It is crucial to develop guidelines and policies that ensure responsible use and prevent the exploitation of this technology for malicious purposes. Additionally, we must consider potential biases in the training data, as these can propagate and influence the generated output, perpetuating stereotypes and unfair representations, unrelated to thoughts of the specific subject. There are also possible concerns about privacy, given that brain decoding models are able to decode language, thoughts, and perceptions [231, 251]. From early experiments, it seems

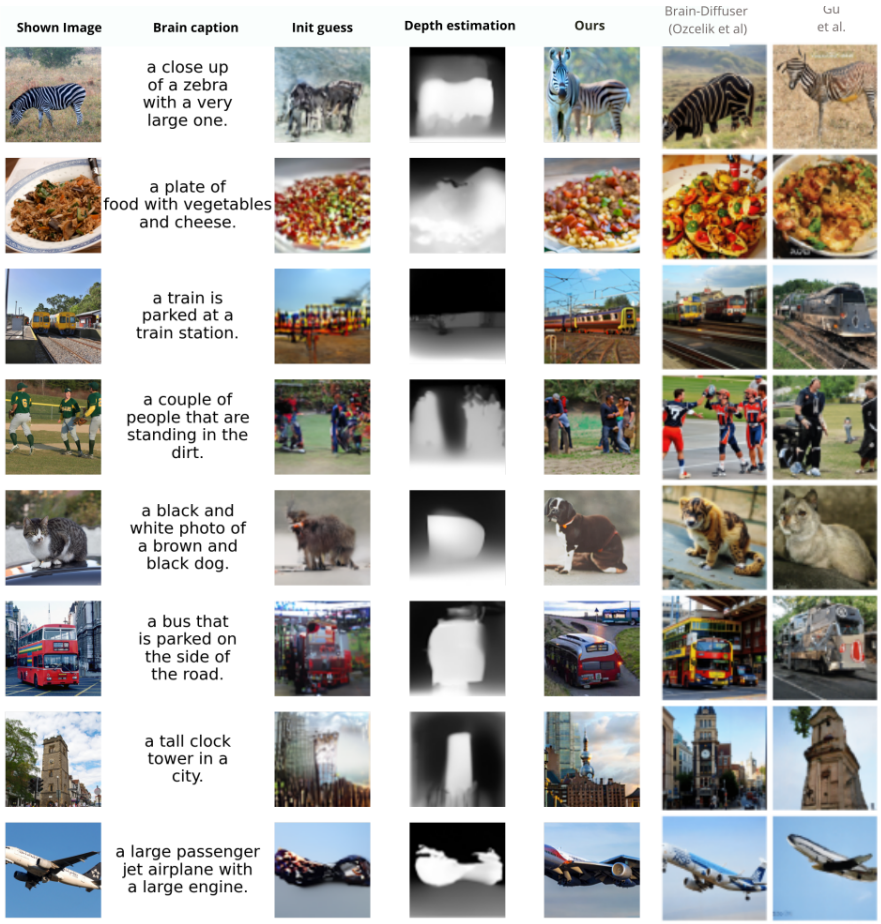


Figure 2.6: Comparison of our results (Columns 2-4) with the presented stimuli and other reconstruction works. The second column displays the caption derived from brain activity, the third column presents the initial guess image, the fourth column exhibits the depth-estimated images, and the fifth column showcases our final reconstruction. The last two columns demonstrate reconstructions from two recent works. All results are from subj01.

that high-level performances are only achievable when subjects are collaborating because the attention process can warp [50] the semantic representation in the brain, which is the primary target of these deep learning multimodal models used as a proxy for brain activity [46]. **Limitations**: In our investigation of brain decoding, we have identified several key limitations that impact the efficacy and generalizability of our findings. The following discussion aims to elaborate on these constraints, establish their interconnections, and provide a deeper understanding of the challenges we face in advancing this field of research. A major limitation in brain decoding work is the necessity for subject-specific models. In-

dividual differences in brain structure, function, and cognitive processing make it challenging to develop a universal decoding model. This specificity hinders the broader applicability of our findings and demands the development of personalized models for each subject.
Even for subject-specific models, to achieve reliable and accurate decoding, a significant amount of high-quality data is needed. Obtaining such data is often time-consuming and resource-intensive, limiting the scalability of brain decoding studies. Additionally, low SNR data can introduce errors and inconsistencies in the decoding process, further compromising the reliability of the results. In this work we used a 7T dataset, that inherently has higher SNR with respect to previous 3T datasets [116], enhancing the quality of our results. In our work, the image captioning model acts as an upper limit: the performance of our brain captioning pipeline is inherently limited by the GIT image captioning model employed. Any inaccuracies or biases present in the model will directly impact the quality of decoded information, setting an upper bound on the performance that can be achieved. Also, the quality of the mapping between neural activity and external stimuli representation in latent spaces is another critical factor influencing the performance of our approach. This determines the accuracy and resolution of the decoded information. Current methods, however, are often limited by the complexity and variability of brain activity, as well as the constraints imposed by the data acquisition techniques, and usually rely on simple regression techniques. Addressing these challenges is essential for refining the mapping process and improving decoding outcomes. Regarding image reconstruction, generating images from text could be another bottleneck. If the text contains errors, these will be propagated and/or enhanced by a separate image reconstruction pipeline. This represents the price for increased flexibility and independence from the specific image reconstruction pipeline used. Finally, the brain decoding process may involve multiple areas, including temporal poles, which further impact of performances. Different brain regions may process and represent information differently, and understanding these variations is crucial for developing accurate and comprehensive decoding models. With the aim of reducing spatial dimensionality, we used only a visual responding region defined by the NSDGeneral ROI, however other brain areas could also encode relevant pieces of information that are relevant to improve performances. Exploring performances as a function of different input regions could be an interesting field of future research.

2.5 Conclusions

Our approach builds upon neuroscientific and AI concepts, leveraging multimodal models to generate captions from brain activity related to the vision of different scenes. We augmented our brain captioning with a pipeline for image reconstruction that uses predicted text and initial information about colors and depth also estimated by brain activity. In conclusion, our approach demonstrates promising results in image captioning and reconstruction from brain activity, with potential applications in a number of cross-disciplinary fields. By drawing on these foundations, we could further our understanding of the human brain's processing of visual and language information, ultimately improving related AI algorithms as well as applications. As we refine our approach, we can continue to explore the intricate relationship between neuroscience and AI, potentially uncovering novel insights and fostering interdisciplinary collaboration.

Appendix

In this section, more comparisons of captions and reconstructed images are provided, compared with state-of-the-art brain decoding pipelines.

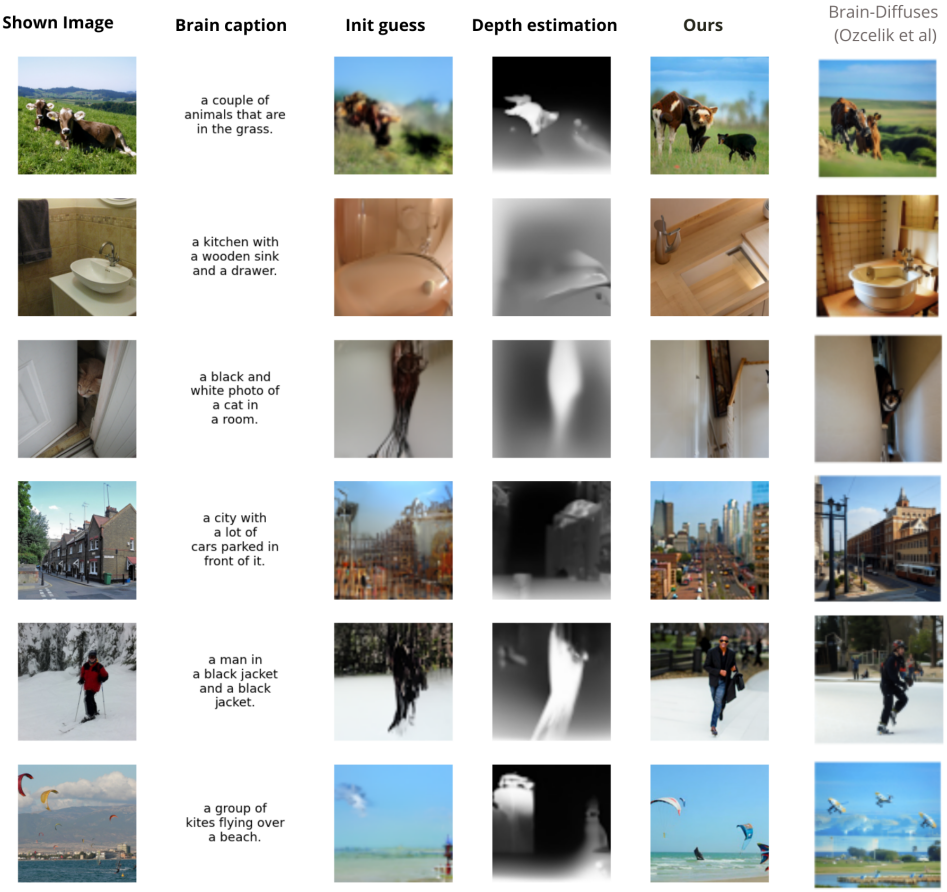


Figure A1: Comparison of our results (Columns 2-4) with the presented stimuli and other reconstruction works. The second column displays the caption derived from brain activity, the third column presents the initial guess image, the fourth column exhibits the depth-estimated images, and the fifth column showcases our final reconstruction. The last column demonstrates reconstructions from the recent BrainDiffuser work. All results are from subj01.

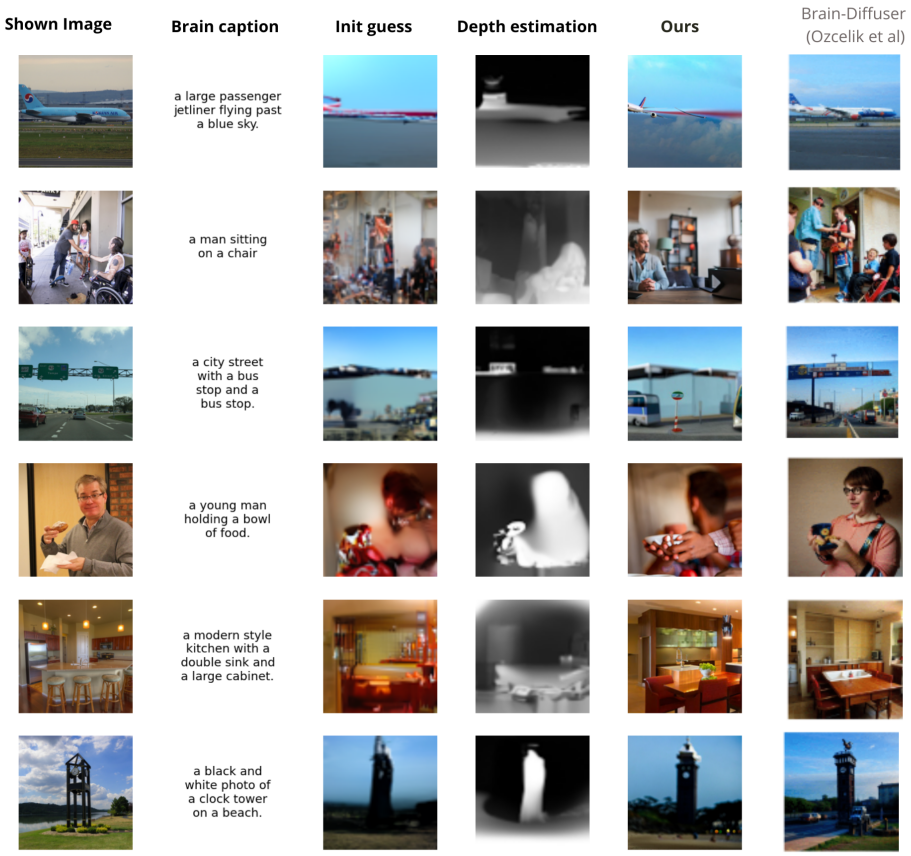


Figure A2: Comparison of our results (Columns 2-4) with the presented stimuli and other reconstruction works. The second column displays the caption derived from brain activity, the third column presents the initial guess image, the fourth column exhibits the depth-estimated images, and the fifth column showcases our final reconstruction. The last column demonstrates reconstructions from the recent BrainDiffuser work. All results are from subj01.

Ablation Study

To validate the contributions of our proposed extensions, we conducted ablation studies analyzing the impact of the depth estimation component. As shown in the attached table, we compared three model variations: 1) a baseline Stable Diffusion Img2Img pipeline using only the initial guess image, 2) a Depth2Image pipeline using only the estimated depth map, and 3) our full approach combin-

Ablation study	Low level metrics			High level		
Variant	PixCorr	SSIM	AlexNet (2)	AlexNet (5)	Inception	CLIP
Text + init	0.1204	0.1941	0.5815	0.7454	0.7974	0.8768
Stable Diffusion depth	0.3333	0.3106	0.8493	0.9654	0.8248	0.8778
ControlNet	0.**3379**	**0.3178**	**0.8707**	**0.9674**	**0.8238**	**0.8788**

Table A1: Ablation Study: Performance Metrics of Different Model Variants. Text + init is the plain Stable Diffusion Img2Img pipeline with initial guess image and captions predicted by the brain. Stable Diffusion depth is a variant pipeline that takes as input the initial guess image and captions and internally tries to estimate a depth map from the initial guess. ControlNet is external conditioning for the StableDiffusion Img2Img pipeline, so the inputs are the initial guess, the captions, and the depth maps estimated from the brain. This latter method is the one used in the paper and values (higher is better) show that this particular combination improves performance. Overall, this ablation study shows that including information about depth improves performances, particularly on low-level features.

ing Stable Diffusion and ControlNet with both initial images and depth maps. Across low-level metrics like PixelCOrr and SSIM, the addition of depth information provided a consistent boost in performance. This aligns with the hypothesis that depth cues aid in capturing spatial relationships between objects and foreground-background segmentation. The full model with both initial images and depth performed the best, indicating that the two components are complementary. Qualitatively, the depth maps appeared to enhance object boundaries and 3D perspective. These results suggest that incorporating depth estimates helps the model reconstruct more accurate and realistic representations of the visual stimuli. The depth component specifically seems to benefit lower-level aspects like shapes and spatial relationships, which are critical for humans to perceive two images as highly similar [107]. By guiding the image reconstruction process with depth information extracted from brain activity, our approach can generate images that better match human perceptual judgments.

3 Cross-subject Brain Decoding

To-date, brain decoding literature has focused on single-subject studies, i.e. reconstructing stimuli presented to a subject under fMRI acquisition from the fMRI activity of the same subject. The objective of this Chapter[1] is to introduce a generalization technique that enables the decoding of a subject's brain based on fMRI activity of another subject, i.e. cross-subject brain decoding. To this end, we also explore cross-subject data alignment techniques. Data alignment is the attempt to register different subjects in a common anatomical or functional space for further and more general analysis.

We utilized the Natural Scenes Dataset, a comprehensive 7T fMRI experiment focused on vision of natural images. The dataset contains fMRI data from multiple subjects exposed to 9841 images, where 982 images have been viewed by all subjects. Our method involved training a decoding model on one subject's data, aligning new data from other subjects to this space, and testing the decoding on the second subject based on information aligned to first subject. We also compared different techniques for fMRI data alignment, specifically ridge regression, hyper alignment, and anatomical alignment.

We found that cross-subject brain decoding is possible, even with a small subset of the dataset, specifically, using the common data, which are around 10% of the total data, namely 982 images, with performances in decoding comparable to the ones achieved by single subject decoding. Cross-subject decoding is still feasible using half or a quarter of this number of images with slightly lower performances. Ridge regression emerged as the best method for functional alignment in fine-grained information decoding, outperforming all other

[1] The work presented in this chapter has been presented at AAAI 2024 Human Centric Learning Representation Learning Workshop, and published at "Imaging Neuroscience" journal [80].

techniques.

By aligning multiple subjects, we achieved high-quality brain decoding and a potential reduction in scan time by 90%. This substantial decrease in scan time could open up unprecedented opportunities for more efficient experiment execution and further advancements in the field, which commonly requires prohibitive (20 hours) scan time per subject.

3.1 Introduction

Deep learning has revolutionized numerous fields, including neuroscience [221, 147, 265, 10]. The application of deep learning techniques in neuroscience has led to significant advancements in understanding brain function and decoding the intricate workings of the human mind [188]. Brain decoding, in particular, has emerged as a crucial area where deep learning plays a pivotal role.

Brain decoding involves the extraction of meaningful information from recorded brain activity, allowing researchers to infer mental states, perceptual experiences, or cognitive processes. For example, deep learning algorithms have been used to decode brain activity and predict whether an individual is looking at a face or an object based on their neural responses [283, 9, 188].

The potential applications of brain decoding are vast, from understanding various aspects of brain function, such as information processing strategies, decision-making, memory formation, and consolidation, to potential uses in neurofeedback, neuroaesthetics, or neuromarketing strategies [66]. Moreover, successful brain decoding could lead to novel strategies for diagnosing and treating neurological or neuropsychiatric conditions, and potentially contribute to the development of radically new algorithmic learning strategies. However, these promising endeavors are not without challenges. Noninvasive data, for instance, have lower temporal or spatial resolution than neural firing, which may limit the granularity of information that can be retrieved. Furthermore, physiological noise and signal/image artifacts can affect both fMRI and EEG data, which can only be imperfectly removed after the data are acquired. Nevertheless, several brain decoding studies have achieved impressive results [283, 9].

One of the key challenges in brain decoding is the subject-specific nature of all models developed thus far. This means that the models are tailored to individual subjects, which can lead to significant variability in the results, given that the amount of data collection per subject could be limited by external factors like time and acquisition costs. Moreover, intrinsic inter-individual variability poses further challenges and every model has to be built from scratch for each new subject. This variability is a consequence of the unique functional and anatomi-

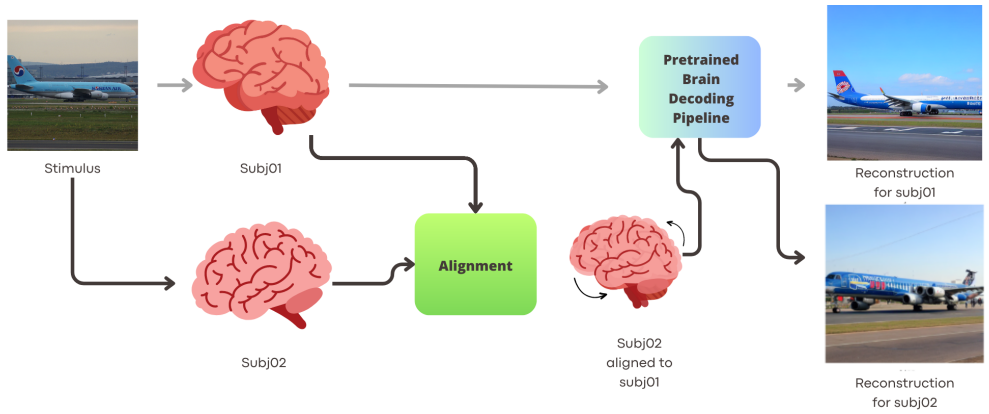


Figure 3.1: Scheme for cross-subject decoding: The procedure involves the following steps: In the first step, Subj01 (top row) is selected as the target subject. A decoding model is trained to reconstruct seen images based on the brain activity of Subj01 on the training set of Subj01 images (8859 images). These images are shown only to this subject. Next, we decode the brain activity of a second subject, Subj02, who was exposed to a share of the same stimuli that Subj01 was exposed to (982 images). Using the shared images we can compare the brain activity related to the same stimuli across different subjects. We used this shared information to align the functional activity of Subj02 with that of the target subject. Once the alignment transformation is learned, we can align the complete dataset (including data not used for alignment nor for decoder training) and we can utilize the pretrained decoder to reconstruct images from Subj02 without training a decoding model specifically for Subj02.

cal structure of each individual's brain, and implies the need to acquire an entire dataset and train an individual model for each subject. This technology's use is limited by a bottleneck requiring extensive data collection—typically thousands of stimulus images—to function properly. This complexity stems from the unique brain anatomy, information processing methods, and functional responses each individual possesses, complicating the training of a universally applicable brain activity decoding model. Despite individual differences in brain anatomy and function, common structures enable reliable neuroscience analysis using template matching techniques, such as anatomical and functional alignment. Anatomical alignment transforms individual brain images to match a standard 'average' brain template or 'atlas', aligning the size, shape, and orientation of brain images. This facilitates meaningful cross-comparisons of brain images, although it is more effective for larger, well-defined structures and may lack precision for smaller, variable regions. Additionally, it does not account for functional differences across brains. Thus, anatomical alignment is often supple-

mented with functional alignment, which synchronizes brain activity patterns across individuals, aiding the comparison and analysis of functional data. This method is vital as activity locations can differ among individuals. Numerous functional alignment methods exist, each with distinct applications and limitations.

Figure 3.2: Example results: The first column, "Stimulus", presents the stimuli from the fMRI experiment. The "Subj01" column (in gray) displays the decoded activity from Subj01, providing an upper performance baseline using a subject-specific decoder model. All other columns (in red) show results from functional alignment using Ridge Regression with 100% common data (952 images), meaning subjects were functionally aligned to Subj01 and decoded using Subj01's trained decoder. To ensure robust visual comparisons, none of the displayed images were used in learning alignment transformation, demonstrating functional alignment on unseen data not used for decoder training or alignment function learning.

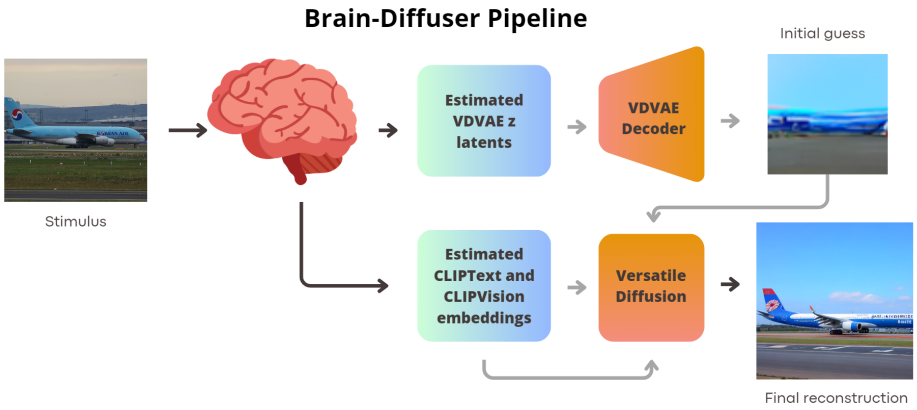


Figure 3.3: The Brain-Diffuser pipeline,the decoder for brain activity used in this study, begins with brain activity from viewing an image stimulus. A model is trained to estimate the latent representation of the VDVAE autoencoder as well as the text and visual embeddings of the CLIP model, using linear models. These estimated vectors and an initial guess image obtained by decoding the autoencoder latents, are fed into Versatile Diffusion—a latent diffusion model—to reconstruct the final image.

In this study, we explore and contrast three methods for cross-subject brain decoding of visual stimuli, using the established, cutting-edge Brain-Diffuser decoding procedure. This approach mitigates variability from new decoding procedures, enabling straightforward quantitative and qualitative comparisons of cross-subject decoding outcomes.

We train a visual stimuli decoding model on one subject (Subj01 usually used as template, except in comparison between target subject section) and employ anatomical alignment, hyper alignment, and functional alignment with ridge regression for three others, decoding their activity using the pretrained model. Our goal is to demonstrate the feasibility of fine-grained cross-subject decoding for visual stimuli reconstruction, potentially reducing scan times significantly by only acquiring data necessary for alignment, thereby reaching state-of-the-art in image reconstruction. Figure 3.1 outlines our proposed pipeline, while Figure 3.2 provides examples of cross-subject decoding results.

3.2 Related Work

In the evolving field of deep learning-based brain decoding, a range of models has been utilized to scrutinize preprocessed fMRI time series as input, particularly focusing on visual stimuli decoding. This involves reconstructing images

that could have triggered specific fMRI patterns—termed brain activity. Here, we review major works in this domain. Some methods have employed variational autoencoders with a generative adversarial component (VAE-GAN) to encode latent human face representations, estimating these encoded representations from fMRI activity using a linear model [262]. Sparse linear regression has also been utilized on preprocessed fMRI data to predict features from the early convolutional layers of a pre-trained convolutional neural network (CNN) for natural images [116]. Unsupervised and adversarial strategies have been used for image reconstruction, including dual VAEGAN and unsupervised methods for decoding fMRI stimuli, utilizing multiple encoder and decoder approaches [237, 218, 85]. Pretrained architectures like BigBiGAN and IC-GAN have optimized latent spaces, significantly enhancing high-fidelity image reconstruction from fMRI patterns [64, 31]. Recently, diffusion models have become prominent in the decoding pipeline, providing superior image generation performance [250, 43]. These models often incorporate semantic-based strategies [79] and multi-step decoding strategies [190, 78].

Interesting approaches also include MindEye [236], a reconstruction tool that effectively maps brain activity to multimodal latent spaces with a contrastive approach for accurate image retrieval and reconstruction, and DREAM [277] that reconstructs images from brain activities using fMRI, closely mirroring the human visual system's structure. In general, the field of bidirectional brain computer interface research has opened avenues for diverse and innovative studies aimed at mapping inputs from and to the brain. Notable examples of such groundbreaking work include those by Mai et al. [167], Luo et al. [164], and Liu et al. [161], among others. Additionally, this research domain has expanded to encompass a variety of stimuli beyond traditional inputs. The Mind-Video study by Chen et al. [42], which tackles the task of decoding video content from fMRI data. Moreover, a particularly captivating branch of research is dedicated to unraveling language processing within the brain through fMRI data, focusing on both the encoding of language into brain activity and the decoding of brain activity back into language. This dual approach has seen significant recent contributions, as evidenced by works such as those by Caucheteux et al [36, 35], Tang et al. [251, 7], and Huth et al. [123], marking crucial advancements in understanding and interpreting the neural substrates of language. Decoding language with non-invasive measurements has also been extended to other modalities, including MEG and EEG [53, 67, 268], yielding promising results. Recent work from Benchetrit et al. [17] is specifically related because the architecture proposed to decode MEG signals includes an alignment module trained specifically for each subject. This is an elegant solution to train a model across subjects tak-

ing into consideration individual differences. In contrast to these approaches, in this paper we explore avenues to leverage pretrained single subject models with a separate alignment procedure, making it more flexible and agnostic to the specific decoding pipeline used. Finally, a recent survey by [188] compiles numerous works on encoding and decoding across different modalities, offering a comprehensive review for interested readers. Notably, most advanced image decoding techniques from fMRI (which is the focus of this study) employ a subject-specific approach requiring training or fine-tuning on individual subject data. Our main aim was therefore to enhance this approach by proposing a universal method that leverages functional alignment of neural data.

Regarding alignment techniques, there are several approaches [14, 15]. Hyperalignment [98, 99, 25] aligns functional brain activity across individuals in a high-dimensional space, enhancing the precision of cross-subject brain activity predictions, but requires extensive high-quality data and complex computational resources. The Shared Response Model (SRM) [39, 40, 220] aligns brain activity by identifying a common response pattern across subjects, ideal for shared experiences, but assumes uniform responses, which individual perception and cognition differences may contradict. Independent Component Analysis (ICA) [27], separates multivariate signals into additive subcomponents, identifying common brain activity patterns, but requires statistical independence of subcomponents, which may not always apply to brain data.

While functional alignment methods provide powerful tools for comparing brain activity, their limitations and assumptions require careful result interpretation. These methods align and compare functional brain data, complementing, not replacing, anatomical alignment. Various other methods, each with its pros and cons, have been proposed. The work most similar to ours in addressing decoding through functional alignment is [113], which demonstrates the potential for cross-subject decoding using linear regression between subjects. Our study contrasts anatomical alignment, hyperalignment-based functional alignment, and ridge regression-based alignment methods for cross-subject brain decoding across various dataset sizes. We specifically investigate the feasibility of cross-dataset decoding using a high-quality 7T dataset, offering robust evidence that simple methods can achieve high performance in cross-subject decoding.

3.3 Material and Methods

In this section, we describe the proposed method and the data we used. The data are publicly available and can be requested at `https://naturalscenesdataset.org/`. All experiments and models were trained on a server equipped with four

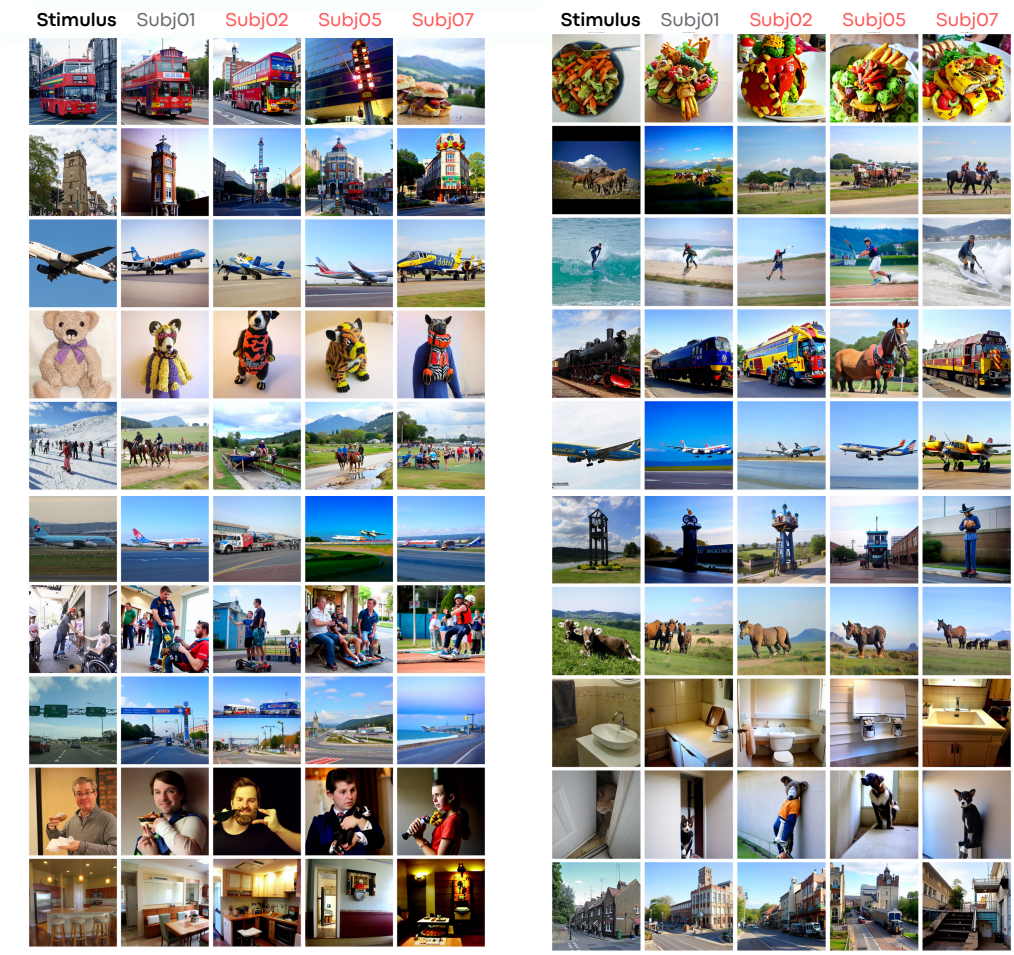


Figure 3.4: More example results. Format and conventions as in Figure 2

NVIDIA A100 GPU cards (80GB RAM each connected through NVLINK) and 2 TB of System RAM.

The study utilizes the Natural Scenes Dataset (NSD) [5], a vast fMRI data set from eight subjects exposed to images from the COCO21 dataset. We focused on four subjects, forming a unique training dataset of 8,859 images and 24,980 fMRI trials, and a common dataset of 982 images and 2,770 trials. To reduce spatial dimensionality, we applied a mask to the fMRI signal (resolution of 1.8mm isotropic) using the NSDGeneral ROI, targeting various visual areas. This strategic ROI selection enhanced the signal-to-noise ratio and simplified data complexity, enabling exploration of both low-level and high-level visual features. Temporal dimensionality was reduced using precomputed betas from

a general linear model (GLM [205, 133]) with a fitted hemodynamic response function (HRF) and a denoising process as detailed in the NSD paper. Data from Subj01, Subj02, Subj05, Subj07, warped into the Montreal Neurological Institute common space (MNI) and downsampled at 2mm, represented the brain activity of each subject and helped decrease computational time and cost. We used the common dataset as alignment, keeping out 30 images for visual comparison, so there are 8859 unique images for each subject. We only used them for training the decoding model for Subj01. Then there are 952 common images across all subjects that were used to functionally align them to the activity of Subj01, and 30 common images kept out for visual comparison on images neither used in the training nor in the alignment procedure. These 30 images were chosen because they're used as visual qualitative evaluation of decoding results in other papers and could help the reader to compare results across different methods. Decoding metrics are evaluated on the 952 images which correspond to our test set for each one of the subjects, since these images are never seen by the decoder model, so the evaluation is still fair and on unseen images. When we refer to 100% of common data we are pointing to these 952 images.

3.3.a Decoding Pipeline: Brain-Diffuser

The "Brain-Diffuser" [190] model is a two-stage framework for reconstructing natural scenes from fMRI signals. Initially, a Very Deep Variational Autoencoder (VDVAE) provides an "initial guess" of the reconstruction, focusing on low-level details. This guess is refined using high-level semantic features from CLIP-Text and CLIP-Vision models, employing a latent diffusion model (Versatile Diffusion) for final image generation. The model, represented in Fig. 3.3, takes fMRI signals as input and generates reconstructed images, capturing low-level properties and overall layout. As a state-of-the-art procedure, Brain-Diffuser was trained using data from Subj01 in the MNI space (cross-subject decoding requires of a common space). Further details about the decoding model are available in the original paper. Aside from the choice of this specific pipeline, our method is universally applicable and can enhance any single subject decoding pipeline. It serves as a versatile tool that can seamlessly integrate with new, cutting-edge pipelines. By focusing solely on preprocessing the input data, our approach allows the underlying pipeline, regardless of its specifics, to function effectively with single subject fMRI data, thereby generating images without needing direct modifications to the pipeline itself. This "plug and play" capability ensures our method remains adaptable and effective in evolving research landscapes.

3.3.b Alignment

This study investigates three alignment strategies to evaluate cross-subject fine-grained brain decoding's feasibility: anatomical alignment, functional alignment via hyper alignment, and functional alignment through ridge regression.

Anatomical alignment, our baseline, relies solely on anatomical details, transforming functional aspects using pre-computed structural image transformations. On the other hand, functional alignment necessitates a more comprehensive approach. Consider the scenario where the brain activity of a source subject $\mathbf{S}$ needs to align with a target subject $\mathbf{T}$. These activities, responses to numerous stimuli, are matrices of shape *(# stimuli, # voxels)*. Given that subjects encounter several common stimuli (i.e., they view identical images in the fMRI scanner), we can divide the datasets into $\mathbf{T}_{\text{common}}$, $\mathbf{T}_{\text{different}}$ and $\mathbf{S}_{\text{common}}$, $\mathbf{S}_{\text{different}}$. Our goal is to leverage the $common$ dataset portion to learn a mapping from S to T, aligning the entire $\mathbf{S}$ dataset with the $\mathbf{T}$ functional space. The NSD experiment's structure, with separate training and test sets (the latter containing identical images for each subject), provides a common stimuli set for alignment purposes.

Anatomical Alignment

Anatomical alignment, a common neuroscience method, aligns to a standard template, here, the MNI space, facilitating anatomical structure comparison. This alignment typically involves a linear coregistration of anatomical images between native and common spaces, followed by a nonlinear warping to match common brain structures. Several software options like FSL and ANTs can perform this task. The NSDData authors [5] elaborate on this process in their released code, providing betas (i.e. coefficients obtained by regressing the stimulus waveform against the fMRI data) for all subjects in the MNI common space (1mm). We downsampled these to 2mm to approximate the resolution used in the original Brain-Diffuser decoding paper (1.8mm) and to reduce spatial dimensionality.

HyperAlignment

HyperAlignment [98, 99], a functional data alignment technique, models functional data as high-dimensional points, with each voxel representing a dimension with betas ranging in $\mathcal{R}$. This method, based on Procrustes Analysis [95], presents a high-dimensional model of the representational space in the human ventral temporal (VT) cortex, wherein dimensions are response-tuning functions common across individuals.

To perform the Procrustes analysis for functional brain alignment, we aim to find a rotation matrix $\mathbf{R}$ and a scale factor c such that the difference $|c\mathbf{SR} - \mathbf{T}|^2$ is minimized.

This is achieved by computing the matrix product $\mathbf{P} = \mathbf{S}_{common}^T \mathbf{T}_{common}$, Performing the singular value decomposition of $\mathbf{P}$ to obtain left and right eigenvector matrices $\mathbf{U}$ and $\mathbf{V}$, Computing $\mathbf{R} = \mathbf{U}\mathbf{V}^T$ and the scaling factor $c = \frac{trace(\mathbf{T_{common}}^T(\mathbf{S_{common}}\mathbf{R}))}{trace(\mathbf{S}_{common}^T\mathbf{S}_{common})}$. Finally, we can apply the matrix $\mathbf{R}$ and the scaling c to both common and non-common source data to align them with the target subject. We computed these values for Subj02, Subj05, and Subj07 as source subjects, using Subj01 as the target, to align all subjects to the functional space of the first one. For detailed mathematical proofs and other insights, please refer to the original articles [99, 98, 95].

Ridge Regression

Our third approach embraces a simple assumption: even in different subjects, all functional data contain the information for the same stimuli, albeit possibly spread across different voxels. This suggests that one subject's activity (source) might be expressed as a linear combination of the activity of another subject (target) for the same stimuli. By deriving a linear combination for each voxel of the target from all possible voxels of the source, we can create a linear map from the source to the target, facilitating functional alignment. The target subject activity can be expressed as $\mathbf{t}i = \sum j\mathbf{w}j\mathbf{s}j$, where $\mathbf{t}i$ is the i-th activity of the target voxel for each common dataset stimulus. Here, $\mathbf{t}i$ represents the i-th column of $\mathbf{T}common$, expressed as a linear combination of all $\mathbf{S}common$ columns. The challenge lies in finding the vector of $\mathbf{w}$ values. When extended to all the target subject voxels, the w vector morphs into a square matrix $\mathbf{W}$, each column of which contains weights to estimate one target subject voxel from a combination of source values. The objective can be redefined as minimizing $|\mathbf{S}common\mathbf{W}^T - \mathbf{T}common|^2$.

We employed Ridge Regression from [24] to determine the $\mathbf{W}$ matrix, conducting a 5-fold cross-validation to select the optimal hyper-parameter α from the values $[0, 1, 10, 1e2, 1e3, 1e4]$. Our findings indicated that $\alpha = 1000$ yields superior performance, hence we adopted it as our final regularization parameter. We computed these values to align all subjects to the initial functional space For the sources Subj02, Subj05, and Subj07, and using Subj01 as the target.

3.3.c Evaluation

Our research seeks to evaluate visual stimuli's detailed brain decoding feasibility, scrutinizing the alignment methods and shared data ratio at play. We examined how the alignment performance fluctuates when the shared data makes up 10%, 25%, 50%, and 100% of the total common data (952 images).

Our shared dataset, or "alignment dataset," comprises 982 images, all viewed by every subject. In order to allow visual comparison, we excluded 30 images from the original Brain-Diffuser paper. Thus, these excluded images neither influenced the training of the decoding pipeline nor the alignment process. The remaining 952 images serve as the shared dataset. We computed transformations for each alignment method (anatomical, hyperalignment, ridge regression) and shared dataset proportion, applying the linear transformation to the complete dataset. This procedure aligns the images with Subject 01's functional space. We then used the pre-trained Brain-Diffuser pipeline for decoding the aligned fMRI activity and reconstructing the images. We assessed our image reconstruction process through both basic and advanced metrics, including PixCorr, SSIM, and 2-way accuracy in AlexNet, Inception, and CLIP latent spaces. This comprehensive evaluation approach allows us to benchmark our results against other brain decoding studies. However, the goal here is not merely comparison, but rather the examination of performance in relation to the shared data fraction and alignment method, given a fixed decoding pipeline, trained solely on Subj01 as a reference target. These metrics enable us to benchmark our results against prior studies in decoding research. However, it is important to note that they rely on an extensive pipeline comprising multiple components, as outlined previously. Given our focus on the input data for the decoding model, which employs a linear decoder within the CLIP space, we also employ a critical evaluation metric: the Direct CLIP 2-way accuracy, assessed directly on the regressed CLIP Classification (CLS) embeddings, that is also part of our decoding linear layer. We refer to this measure as "Direct CLIP" to differentiate it from previously mentioned metrics.

3.3.d Cross-Dataset decoding experiment

In order to evaluate the generalizability of our decoding pipeline, we undertook cross-dataset decoding between different fMRI datasets.

Firstly, we undertook cross-dataset decoding between the BOLD5000[38] and the Natural Scenes Dataset (NSD). The BOLD5000 dataset comprises fMRI data from five subjects who viewed 5,000 images. This data was acquired at a field strength of 3T, providing inherently lower signal to noise ratio than the 7T data

present in the NSD. Additionally, BOLD5000 encompasses a narrower range of semantic categories in comparison to the diverse stimuli of the NSD. The protocols between the two datasets also differ: the NSD employs a rapid-event related protocol where images are displayed for 2 seconds followed by a 1-second pause, whereas in BOLD5000, images are displayed for 1 second and succeeded by a 9-second cross fixation. Both datasets underwent identical processing to extract the task-related voxel coefficients, specifically within the visual cortex masks. For this particular experiment, our attention was centered on the first three BOLD5000 subjects, which shared 1,000 images with the NSD subjects. This overlap facilitated a direct neural response comparison to the same stimuli across both datasets. The primary objective was to train a decoder using NSD Subject 1, align the common data from BOLD5000 subjects, and then proceed with cross-dataset decoding. For comparative purposes, we implemented the Brain-Diffuser pipeline on the training data (comprising 80% of non-common data) from the BOLD5000 subject to gauge within-dataset and within-subject decoding performances. Subsequently, we employed Ridge Regression to serve as the alignment matrix bridging the neural spaces of BOLD5000 and NSD. After applying this transformation to the BOLD5000 test data, decoding was performed using the decoder trained on NSD Subj01 data. Metrics were computed across the decoded test sets, both within and between datasets. The overarching aim of this experiment was to investigate the feasibility of addressing the intricate task of achieving high-quality cross-dataset and cross-field fMRI data decoding. To further test this cross-dataset experiment, we also leveraged another 3T fMRI dataset, part of the THINGS collection [105]. This database, which is richly annotated, consists of 1,854 object concepts representative of the American English language and contains 26,107 manually-curated naturalistic object images. In the fMRI study, a representative subset of images from the THINGS database was presented to participants across 12 separate sessions, involving 8740 unique images of 720 objects. The images were displayed in rapid succession (4.5 seconds), with participants instructed to focus on central fixation. To maintain participant engagement, an oddball detection task was incorporated, requiring responses to occasional artificially-generated images. Additionally, a specific subset of images (100 in total) was repeatedly shown in each session. This task presents a greater challenge in our cross-dataset experiment due to the absence of explicitly common images. To address this issue, we implemented a retrieval system leveraging CLIP embeddings of images and established an alignment dataset consisting of 1,000 images that are semantically similar between the THINGS and NSD datasets. By using these images along with their associated fMRI data, we were able to learn the alignment between the datasets

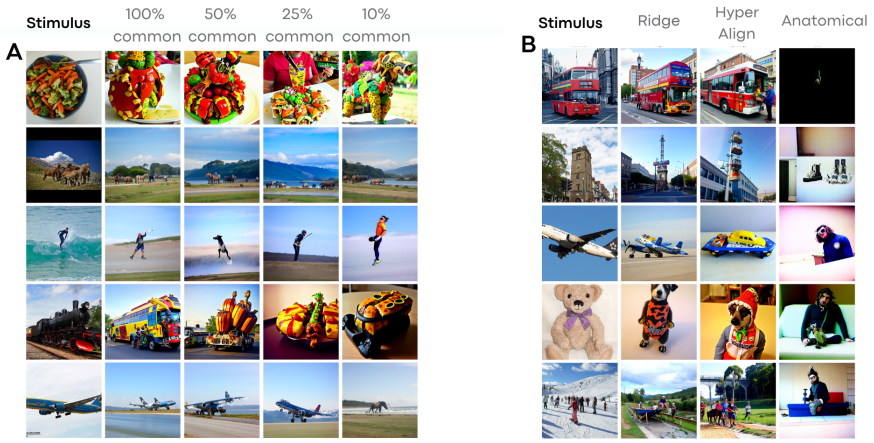


Figure 3.5: **A**: Functional alignment comparison using Ridge Regression across varying fractions of shared data. The "Stimulus" column showcases experimental images, while subsequent columns depict the decoded and aligned activity of Subj02 based on Subj01. **B**: A comparison of distinct alignment techniques. The "Stimulus" column again presents the experimental images, with the remaining columns illustrating the decoded activity of Subj02, aligned to Subj01 through various methodologies.

and perform our cross-dataset decoding task.

3.4 Results

3.4.a Cross-Subject decoding experiment

Figures 3.2 and 3.4 provide a comparison between stimuli and decoded images from Subj01 (on which the decoding model is trained). These figures also display the aligned activity of all other subjects using Ridge Regression. Figure 3.5 compares fractions of common data used for Ridge Regression-based alignment and different alignment methods. Lastly, Figure 3.8 illustrates each quantitative metric, computed and averaged over the entire test set for each aligned subject (2,5,7). Metrics are expressed as a fraction of the entire dataset, which contains approximately 10k images per subject (8859 unique + 982 common across subjects). As common images are necessary, the maximum amount of images that can be included in the alignment process is 952 (termed the "alignment set", except 30 images left out for visualization purposes), representing around 10% of the dataset. This represents the maximum data that can be incorporated into the procedure, and we experimented with half, a quarter, and a tenth of this data.

Anatomical alignment fails to yield satisfactory results, demonstrating just

above chance performance levels for 2-way classification accuracy and poor performance for low-level metrics such as SSIM and PixCorr. However, Ridge Regression exhibits an increasing performance based on the volume of data used for alignment mapping function learning. This method reaches performance levels comparable with the within-subject decoder in both low-level and high-level metrics, using all the common data (approximately 10% of the entire dataset). An extended version of this figure can be found in the Appendix, where alignment techniques are compared with within-subject decoders trained on different amounts of data, confirming that performing functional alignment improves performances in decoding when only sparse data are available.

Our findings are encapsulated in the following key points:

Functional alignment's critical role in fine-grained brain decoding: Our research emphasizes the pivotal role of functional alignment in fine-grained brain decoding. This process, which interprets neural signals to reconstruct perceived images or thoughts, greatly benefits from precise functional alignment of brain activity. Accurate alignment ensures that neural signals are matched correctly to their corresponding brain regions, thus enhancing the decoding accuracy.

Anatomical method's inefficacy: As corroborated by previous studies [99], our research found that anatomical methods for brain decoding are ineffective. Relying on the physical structure of the brain for alignment and decoding does not deliver the requisite precision for fine-grained decoding tasks. This could be attributed to inherent anatomical variability across different individuals, which may not necessarily align with functional differences. The specialized areas in the brain with functional selectivity can sometimes yield performance above chance levels. However, in most cases, decoded images do not correlate with the stimulus, undermining the reliability of this method for cross-subject brain decoding.

Hyperalignment and Ridge regression efficacy: We found that hyperalignment and ridge regression can both be successfully used to perform cross-subject functional alignment and decoding. Our results demonstrate that Ridge Regression-based methods for brain decoding can achieve above-chance performance levels with as little as 1% of the entire dataset. Furthermore, these methods have performances close to baseline levels using around the 10% of the dataset. This crucial finding implies that reliable brain decoding results can be achieved while significantly reducing scan time. This efficiency could be instrumental in making brain decoding research and applications more feasible and cost-effective. These results contribute to our comprehension of the challenges and potential remedies in brain decoding and emphasize the need for additional research to refine these techniques and augment their effectiveness.

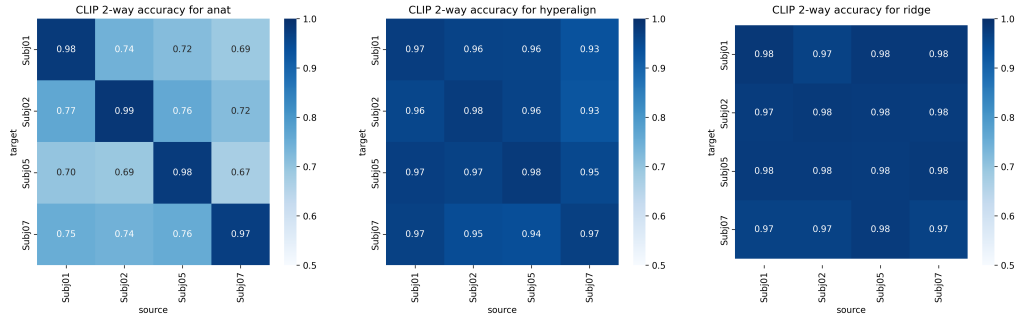


Figure 3.6: This figure illustrates decoding results with direct CLIP 2-way regressed space, in order to be independent from the actual latent diffusion model used for image generation.

Comparable Qualitative Results from Varying Training and Alignment Subjects:

To systematically assess the influence of selecting distinct subjects for model training and alignment, we experimented with multiple combinations. For instance, the decoder was trained on Subject 1, followed by alignment of Subject 2 to this target, and subsequent decoding of Subject 2. This procedure was also executed with the decoder trained on Subject 2, alignment of Subject 1, and decoding of Subject 1. We extended this approach to encompass all potential combinations of our four subjects. As evidenced in figures 3.7, the qualitative nature of the decoded images remained consistent irrespective of the subjects chosen for training and alignment. These figures distinctly captured high-level content and foundational shapes across varying subject combinations, yielding analogous visual decoding results. Figure 3.6 presents the direct CLIP 2-way accuracy across all combinations of source and target subjects, utilizing both anatomical and functional alignment. The left side of the figure illustrates that only the diagonal elements yield high decoding accuracy. In contrast, both hyperalignment and ridge regression exhibit off-diagonal values that are comparable to, and on par with, the diagonal elements. This indicates that cross-subject decoding is as effective as within-subject decoding. Within these figures, the diagonal cells represent within-subject decoding, wherein the model undergoes both training and testing on an identical subject. In contrast, off-diagonal cells signify cross-subject decoding, where distinct subjects are employed for training as opposed to alignment and decoding. Despite the variations in quantitative metrics, the visual reconstructions derived from different combinations are qualitatively analogous. This underscores the decoder's proficiency in generalizing across diverse subjects. The presence of shared neural representations, even amidst individual disparities, facilitates precise cross-subject decoding across a

spectrum of training and alignment configurations.

3.4.b Cross-Dataset Decoding Experiment

Despite the inherent disparities in acquisition protocols and magnetic field strengths between datasets, we observed successful cross-dataset decoding. Qualitative analysis revealed that reconstructions derived from the aligned BOLD5000 subject were often comparable to those from within-dataset decoding. The same principles apply to the reconstruction of within and cross decoding for the THINGS dataset, highlighting the viability of applying a decoding model across distinct fMRI datasets sharing common stimuli. The primary challenges of cross-dataset application arise from variations in experimental protocols, magnetic field strengths, and even minor details in the images from the THINGS dataset, while maintaining the same semantic content. Nevertheless, our methodology successfully addressed these challenges, achieving effective functional alignment. Figure 3.9 offers a qualitative comparison of reconstructions generated by a model either trained on the same subject or aligned to another subject's brain activity from the NSD dataset, using its decoder. The alignment technique produced reconstructions that were semantically aligned with the original stimuli, demonstrating a significant advantage of this method. Qualitatively, the performance of decoded image reconstructions was found to be similar when comparing within and across dataset decoding. This observation is supported by quantitative metrics, as presented in Tables 3.1 and 3.2, which demonstrate that performance on all evaluated parameters using the cross-dataset decoder exceeds chance, thereby validating the feasibility of cross-subject decoding. For the BOLD5000 dataset, a comparison of top-1 accuracy with state-of-the-art decoding methodologies cited in [43] is also included. Although the decoding approach differs significantly, requiring a pretraining phase on the Human Connectome Project dataset, which includes around 1000 subjects, followed by fine-tuning on specific BOLD5000 data, it outperforms other methods. The emphasis here is to demonstrate the feasibility of decoding at an individual subject level within the BOLD5000 dataset and through the use of our alignment method and a pretrained decoder on an NSD subject. Comparative metrics indicate that cross-dataset decoding sometimes surpasses within-dataset decoding, suggesting a successful transfer of information despite differences between datasets and subjects. Extending this analysis to the THINGS dataset (Table 3.2), it was observed that within-subject decoding consistently outperforms cross-dataset decoding, although the latter still achieves performance levels above chance. This further suggests the potential for information transfer. Notably, this task is more challenging than transferring from BOLD to NSD

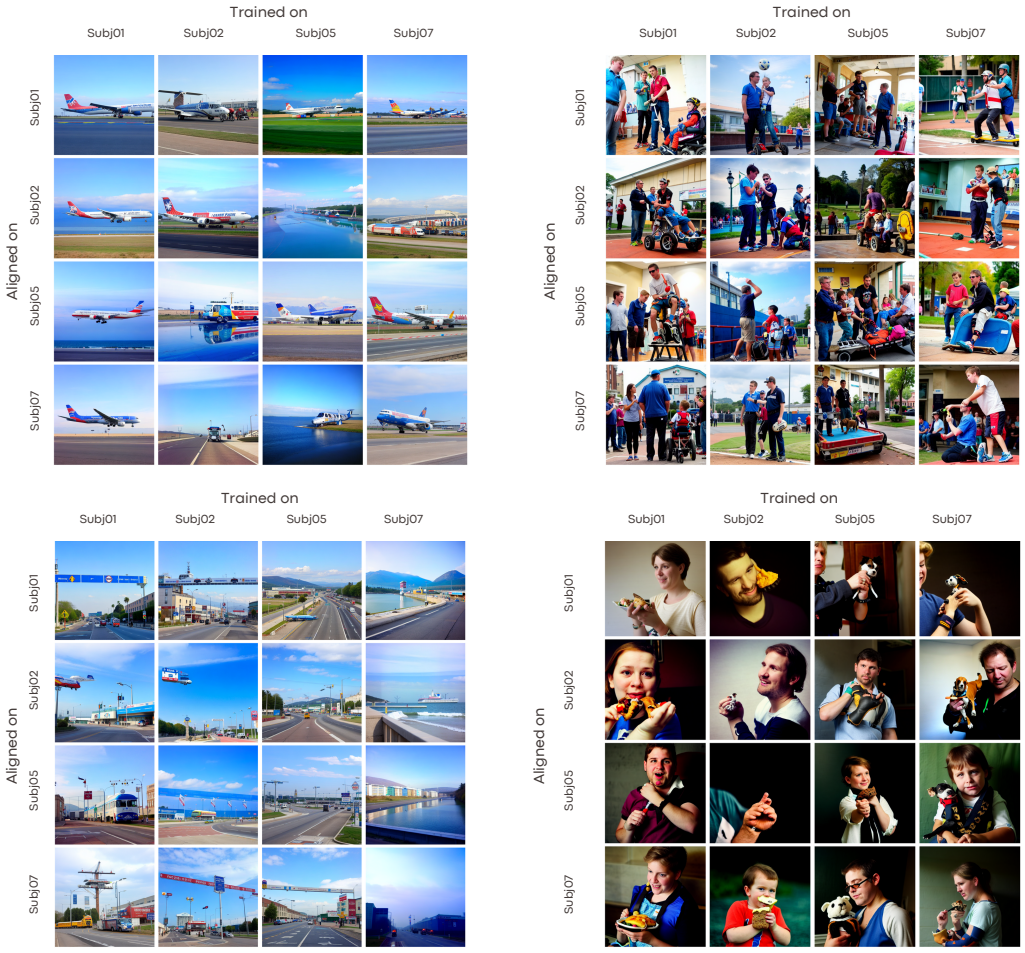


Figure 3.7: Decoding results from different combinations of subjects used for model training versus alignment. The columns represent decoders trained on individual target subjects. The rows show each remaining subject aligned to the target space of the column subject for decoding.

because it relies on semantically similar stimuli rather than identical images.

Subj	Pixcorr	SSIM	AlexNet 2nd	AlexNet 5th	Inception	CLIP	Direct CLIP	Top 1 acc
CSI1 (within)	0.1650	0.2201	**0.6009**	**0.7469**	**0.6642**	**0.8126**	0.7104	0.1535
CSI1 (cross)	**0.1736**	**0.2314**	0.5693	0.6691	0.5669	0.7201	**0.7420**	0.115
CSI1 (mind-vis) [43]	-	-	-	-	-	-	-	**0.3345**
CSI2 (within)	0.1448	0.2114	**0.5669**	**0.7323**	**0.6350**	**0.7761**	0.6909	0.1240
CSI2 (cross)	**0.1560**	**0.2286**	0.5474	0.6593	0.6082	0.6593	**0.7299**	0.0861
CSI2 (mind-vis) [43]	-	-	-	-	-	-	-	**0.185**
CSI3 (within)	0.1479	0.2221	**0.5693**	**0.7493**	**0.6642**	**0.7591**	0.7250	0.129
CSI3 (cross)	**0.1485**	**0.2321**	0.5304	0.6155	0.5109	0.6715	**0.7323**	0.095
CSI3 (mind-vis) [43]	-	-	-	-	-	-	-	**0.210**

Table 3.1: Expanded table showing quantitative metrics across different datasets and models. Bold values indicate superior performance. "Within Data" pertains to results from a decoder trained on BOLD5000 data, while "Cross Data" involves results from reconstructed images using the NSD decoder and aligned BOLD5000 activity.

Subj	Pixcorr	SSIM	AlexNet 2nd	AlexNet 5th	Inception	CLIP	Direct CLIP
sub-01 (within)	**0.1680**	**0.2820**	**0.6633**	**0.775**	**0.64833**	**0.76**	**0.8166**
sub-01 (cross)	0.1496	0.2288	0.53166	0.6000	0.5733	0.6425	0.6524
sub-02 (within)	**0.1642**	**0.2854**	**0.6516**	**0.7916**	**0.65416**	**0.77166**	**0.82999**
sub-02 (cross)	0.14478	0.24148	0.5508	0.62333	0.5608	0.62583	0.6216
sub-03 (within)	**0.15852**	**0.2672**	**0.62583**	**0.73083**	**0.62**	**0.73833**	**0.78083**
sub-03 (cross)	0.15807	0.2282	0.5541666	0.617	0.5425	0.58	0.6233

Table 3.2: Expanded table showing quantitative metrics across different datasets and models. Bold values indicate superior performance. "Within Data" pertains to results from a decoder trained on THINGS data, while "Cross Data" involves results from reconstructed images using the NSD decoder and aligned THINGS activity.

3.5 Discussion

Our study underscores the intricacies and potential of cross-subject fine-grained brain decoding, a field promising to enhance our understanding of the human brain and cognition.

We identified the criticality of functional alignment for successfully executing brain decoding. This alignment, which maps neural signals to their corresponding brain regions, is vital for accurately decoding neural activity from other individuals using a model pre-trained on separate subject data. This insight holds promise for constructing large studies with a high-accuracy decoding pipeline, subsequently requiring only alignment data acquisition for new subjects. This approach negates the need for an entire experimental reproduction each time, streamlining the process.

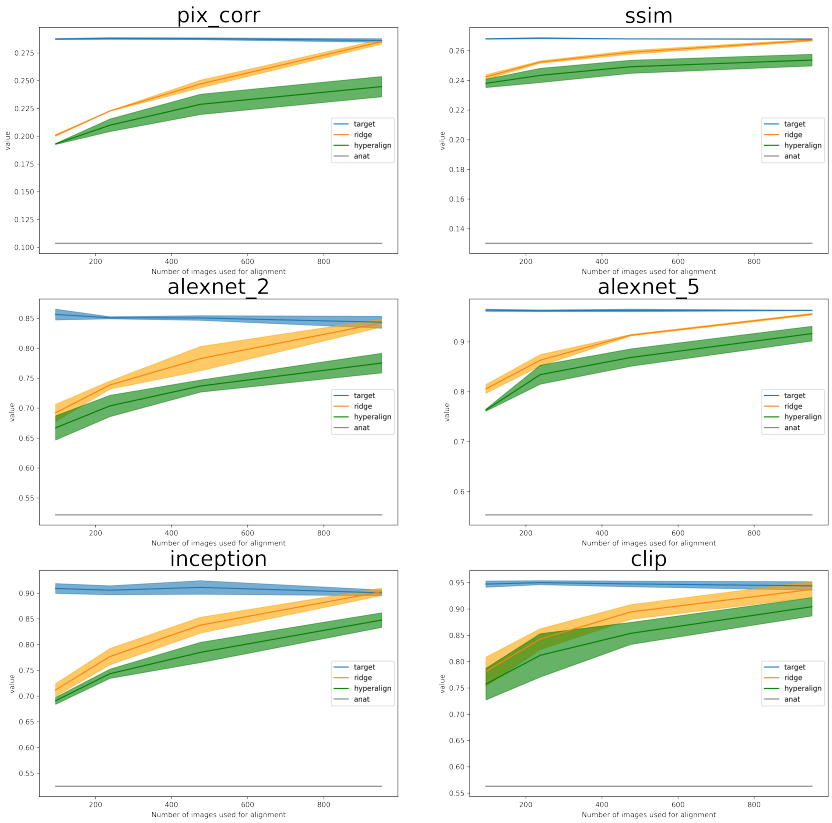


Figure 3.8: This figure illustrates the performance of alignment methods evaluated using different metrics. Blue lines represent metrics from the target subject's decoded images, derived from their test set brain activity. Green lines denote the mean and standard deviation (std) of performance on test sets from other subjects, aligned using hyperalignment. Gray lines present results achieved using anatomical alignment, while orange lines display outcomes using Ridge Regression. Remarkably, the Ridge Regression approach yields positive results even when using a tiny fraction of the entire dataset. Furthermore, as this fraction approaches roughly 10% of the total set, resulting in 952 images the performance becomes comparable with that obtained by the within-subject model.

Our research reveals the limitations of anatomical methods for brain decoding, which rely on the physical brain structure for alignment and decoding. These methods underperformed due to inherent brain anatomical variability across individuals, which may not align with functional differences. Thus, this study

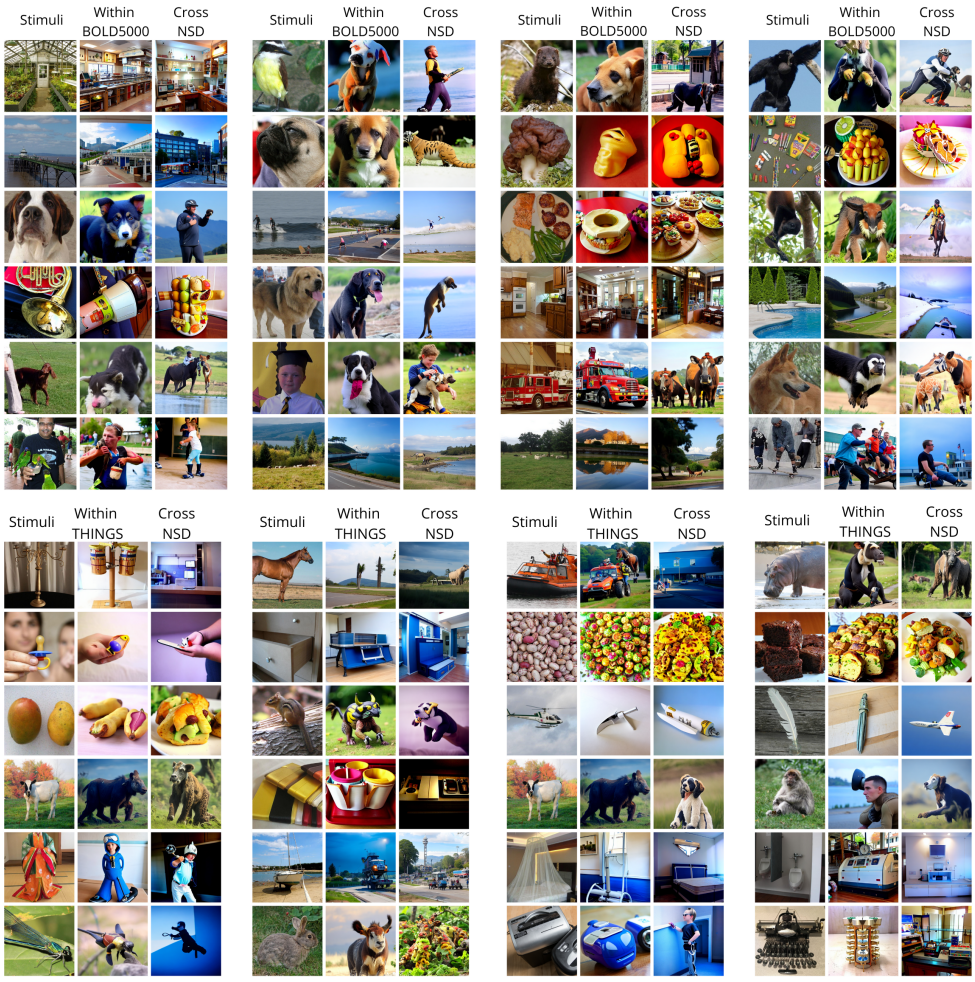


Figure 3.9: Cross-Dataset qualitative results. Each panel illustrate different random examples from test set. The first column "Stimuli" shows the original stimulus shown during the experiment. The column "Within BOLD5000" shows reconstruction from a decoding model trained on the BOLD5000 subject. "Cross NSD" column show results decoded from aligned activity on the NSD dataset and decoded using a decoder trained on the NSD data. Qualitatively, images from this latter column are more semantically similar then the results in the second column. Same happens for the second set of images (panel at the bottom) with the THINGS dataset.

emphasizes the need for functional, not merely anatomical, considerations in decoding studies.

Excitingly, our results suggest significant reductions in scan-time are possible. Ridge Regression-based methods were found to provide reliable brain decoding results with just a fraction of the entire dataset, implying practical implications for brain decoding research feasibility and cost-effectiveness.

Our study also highlights the qualitative similarities in decoded images across subjects. While these images largely match high-level semantic content, intra-subject differences appear minimized. This observation prompts us to consider whether the decoding procedure is fully captured. Given that high-level concepts are generally aligned, we propose a possible brain activity decomposition into *brain activity* = *concept* + *individual perception*. Such a model might only capture the *concept* while treating differences as noise, offering new research directions to explore fine-grained inter-subject differences. This observation might also explain why the metrics for aligned subjects are comparable to or even surpass those of the target subject. The alignment process, which is designed to capture only the visual content common among individuals, could also serve as a form of denoising. This makes it easier for the decoder associated with the target subject to extract information from brain patterns.

Beyond mere conjecture, our results indicate that despite the differences in individual brain structures and functions, it is possible to decode shared neural activity patterns across individuals. This revelation presents compelling prospects for the creation of generalized brain decoding models that can be applied across a diverse range of individuals. Nonetheless, our research also uncovers the constraints inherent in current functional alignment methodologies. As a progression, future investigations might delve into the utilization of more sophisticated models, such as neural networks. These networks are renowned for their capacity to discern intricate, non-linear associations, which could potentially enhance functional alignment.

As we project into the future, several promising avenues for research become evident. These include training models across a spectrum of subjects and devices, potentially culminating in the development of more resilient and universally applicable brain decoding models. For example, using a shared response model, mapping different subjects representations into a common low-dimensional subspace could be a preprocessing step useful to train multisubject large fMRI decoding models. Or, with the advent of larger datasets, future approach could follow the idea described in the work from Benchetrit et al. [17] of a shared network with subject-specific alignment layers that at beginning maps all the subjects into a common representation and then elaborate over this shared

space. The inception of novel techniques and methodologies might pave the way for surmounting the existing challenges in brain decoding, ushering in a new era of precise and efficient brain activity interpretation.

In our study, we embarked on a preliminary exploration of this domain through our cross-dataset experiment. The successful cross-dataset decoding achieved between the BOLD5000 or THINGS fMRI and the Natural Scenes Dataset attests to the viability of this approach, even in the face of disparities in acquisition protocols. However, challenges emanate from differences in factors such as magnetic field strength, resolution, stimuli distributions, and more. Designing models that can either account for these variations or remain resilient to them is a formidable task. In forthcoming research, the adoption of intricate techniques, like neural networks equipped with contrastive losses, might aid in discerning invariant feature representations. When trained on a diverse array of fMRI datasets, these methods could effectively identify shared neural patterns, thereby enhancing generalization across varied acquisition specifics. The realm of cross-dataset decoding is burgeoning with potential, offering the prospect of harnessing multiple resources and minimizing the need for fresh data acquisition. The establishment of standardization protocols and shared evaluation benchmarks would significantly bolster this endeavor. Our research underscores the immense potential inherent in cross-dataset brain decoding, paving the way for transcending the constraints of individual datasets.

A potential limitation of our study could be the inherent methodological challenge associated with using the Natural Scenes Dataset, specifically in creating truly independent training and test sets due to the distribution of trials across runs and the computation of GLM's beta coefficients, which is the standard input for fMRI decoding models in recent literature. This structure could theoretically lead to an overestimation of decoding performance metrics, due to potential information leakage from test trials given that fMRI data are time-series and during the computation of beta coefficients all information in a run is used. However, the implementation of advanced analytical methods, such as GLMsingle [205, 133], substantially mitigates this concern given the use of internal and nested cross-validation as well as regularization. This model uses CV to select voxel-wise optimal hemodynamic function, followed by Ridge Regression within the GLM used to obtain regression coefficients, again using a cross-validation procedure. Furthermore, the practice of averaging coefficients for identical stimuli across multiple presentations in different runs further ensures the minimization of possible information leakage between the training and test data. Thus, while the dataset's structure presents this practical and potential challenge, commonly used methodological precautions and the application

of rigorous cross-validation techniques affirm the reliability and validity of our and recent literature findings within the context of this limitation. In our specific case, showcasing model generalization across different subjects and even datasets can be considered robust evidence for true ability of generalization and mitigate concern about potential overfitting.

Furthermore, as we delve into fine-grained brain decoding, addressing potential privacy concerns and ethical implications is paramount. Current research suggests that while certain brain activity aspects can be decoded across subjects, the process is not yet a comprehensive or intrusive 'mind-reading' tool. A key finding highlights the disruptive role of attention mechanisms, suggesting that brain decoding is only possible with actively participating, aware subjects.

While our methods currently prevent involuntary or covert 'mind reading', as the field advances, maintaining strong ethical frameworks for brain decoding research becomes even more critical. Informed consent, strict data privacy protocols, and potential societal implications consideration remain key. Decoding brain activity raises broader ethical questions, such as its potential use to enhance communication for individuals with speech or motor impairments or its potential misuse for coercive or manipulative purposes. These critical questions must be confronted by the scientific community and society as we continue to explore brain decoding potential.

3.6 Conclusions

In this research, we detailed a method for brain decoding of visual stimuli across multiple subjects. Our exploration highlighted the importance of functional alignment in decoding neural signals and brought attention to the challenges associated with anatomical methods and complex decoding techniques that might be prone to overfitting.

A key aspect of our study was the application of Ridge Regression-based methods. This approach was effective in decoding neural activity using a subset of the dataset, suggesting potential for reduced scanning durations, possibly nearing 90%. Such reductions could have implications for the efficiency of brain decoding research and its applications.

Our work achieved cross-subject and cross-dataset brain decoding by training the decoding framework on one subject and decoding neural activity of other individuals. This result indicates the presence of shared neural activity patterns, which could be foundational for future generalized brain decoding models. Additionally, we could speculate that results can be explained by a hierarchical structure in the brain's processing and representation of information, distin-

guishing brain decoding into concept and perception components. While these findings are promising, we also acknowledge the current limitations in functional alignment methods and see value in exploring other research directions, such as training models across different subjects and devices.

In summary, our study provides insights into the challenges and potential avenues in the realm of cross-subject brain decoding, especially concerning visual stimuli.

3.7 Data and Code Availability

Data are publicly available at https://naturalscenesdataset.org/. As described in the original data paper [5] informed written consent was obtained from all subjects, and the experimental protocol was approved by the University of Minnesota Institutional Review Board. The code to perform the cross-subject decoding analysis and reproduce results is available at https://github.com/matteoferrante/cross-subject-decoding.git

Appendix

In this section, we offer further insights and numerical data for all figures presented, to facilitate a comprehensive comparison. In addition to detailed tables, we present findings from an additional experiment. This experiment assesses the efficacy of cross-subject alignment versus within-subject decoding when trained with identical data volumes. It demonstrates that applying cross-subject decoding through a straightforward linear transformation into the target space—where a proficient decoding model is already established—yields superior results compared to direct training on a limited dataset. For this purpose, we replicated the analysis shown in Figure 3.8 from the main manuscript in Fig A1, incorporating an updated dotted line to depict the performance metrics. The latter refer to a decoder trained on the same subject but with varying data quantities. We re-trained the Brain-Diffuser decoding pipeline for all subjects on the same data used for alignment and then measured performances on generated images. The black dotted line in the picture shows an increasing trend with the increase of the dataset size, but performances are always lower than functional alignment.

Target \ Source	Method	Subj01	Subj02	Subj05	Subj07
Subj01	anat	0.978992	0.742647	0.718487	0.690126
Subj02	anat	0.766807	0.985294	0.763655	0.718487
Subj05	anat	0.700630	0.692227	0.977941	0.673319
Subj07	anat	0.753151	0.743697	0.756303	0.971639
Subj01	hyperalign	0.973740	0.964286	0.961134	0.934874
Subj02	hyperalign	0.957983	0.975840	0.963235	0.929622
Subj05	hyperalign	0.971639	0.965336	0.980042	0.954832
Subj07	hyperalign	0.968487	0.945378	0.943277	0.969538
Subj01	ridge	0.984244	0.966387	0.981092	0.981092
Subj02	ridge	0.974790	0.980042	0.976891	0.975840
Subj05	ridge	0.977941	0.977941	0.975840	0.978992
Subj07	ridge	0.966387	0.966387	0.977941	0.966387

Table A1: Detail of the CLIP 2-way accuracy with direct regression for all alignment methods and subjects.

	Method/# of images	95	238	476	952
0	Target	0.287453	0.287965	0.287736	0.286151
1	Anat	0.103591	0.103591	0.103591	0.103591
2	Hyperalignment	0.193022	0.209869	0.228685	0.244612
3	Ridge	0.200703	0.222746	0.247006	0.285059

Table A2: Tabular results depicted in Fig 3.8 for Pixel Correlation

	Method/# of images	95	238	476	952
0	Target	0.268018	0.268481	0.267975	0.267864
1	Anat	0.130182	0.130182	0.130182	0.130182
2	Hyperalignment	0.238031	0.243387	0.249211	0.253708
3	Ridge	0.242306	0.252384	0.258937	0.267115

Table A3: Tabular results depicted in Fig 3.8 for SSIM

	Method/# of images	95	238	476	952
0	Target	0.856415	0.850984	0.850645	0.843517
1	Anat	0.521724	0.521724	0.521724	0.521724
2	Hyperalignment	0.667006	0.703666	0.736931	0.775289
3	Ridge	0.692125	0.738629	0.782756	0.842159

Table A4: Tabular results depicted in Fig 3.8 for AlexNet 2nd layer 2-way accuracy

	Method/# of images	95	238	476	952
0	Target	0.963340	0.962322	0.963001	0.963001
1	Anat	0.553632	0.553632	0.553632	0.553632
2	Hyperalignment	0.763408	0.834352	0.868635	0.916497
3	Ridge	0.805838	0.863204	0.913442	0.955533

Table A5: Tabular results depicted in Fig 3.8 for AlexNet 5th layer 2-way accuracy

	Method/# of images	95	238	476	952
0	Target	0.909029	0.905635	0.911066	0.900543
1	Anat	0.524440	0.524440	0.524440	0.524440
2	Hyperalignment	0.690767	0.743381	0.784793	0.847590
3	Ridge	0.711813	0.776646	0.837746	0.902580

Table A6: Tabular results depicted in Fig 3.8 for Inception 2-way accuracy

	Method/# of images	95	238	476	952
0	Target	0.947386	0.950102	0.947726	0.943992
1	Anat	0.562797	0.562797	0.562797	0.562797
2	Hyperalignment	0.757298	0.811948	0.853700	0.904277
3	Ridge	0.781059	0.842838	0.894433	0.937203

Table A7: Tabular results depicted in Fig 3.8 for CLIP 2-way accuracy

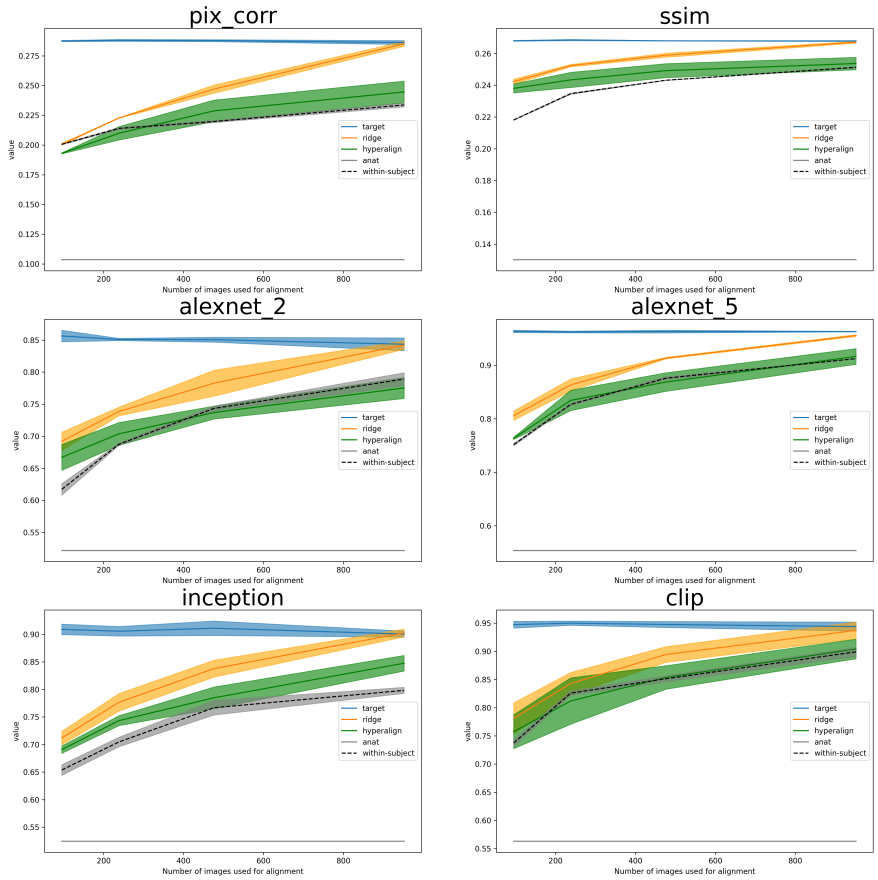


Figure A1: This figure illustrates the performance of alignment methods evaluated using different metrics. Blue lines represent metrics from the target subject's decoded images, derived from their test set brain activity. Green lines denote the mean and standard deviation (std) of performance on test sets from other subjects, aligned using hyperalignment. Gray lines present results achieved using anatomical alignment, while orange lines display outcomes using Ridge Regression. The black dotted line signifies the performance outcomes of a within-subject decoder that was trained on a small subset of the dataset. Remarkably, the Ridge Regression approach yields positive results even when using a tiny fraction of the entire dataset. Furthermore, as this fraction approaches roughly 10% of the total set, resulting in 952 images the performance becomes comparable with that obtained by the within-subject model.

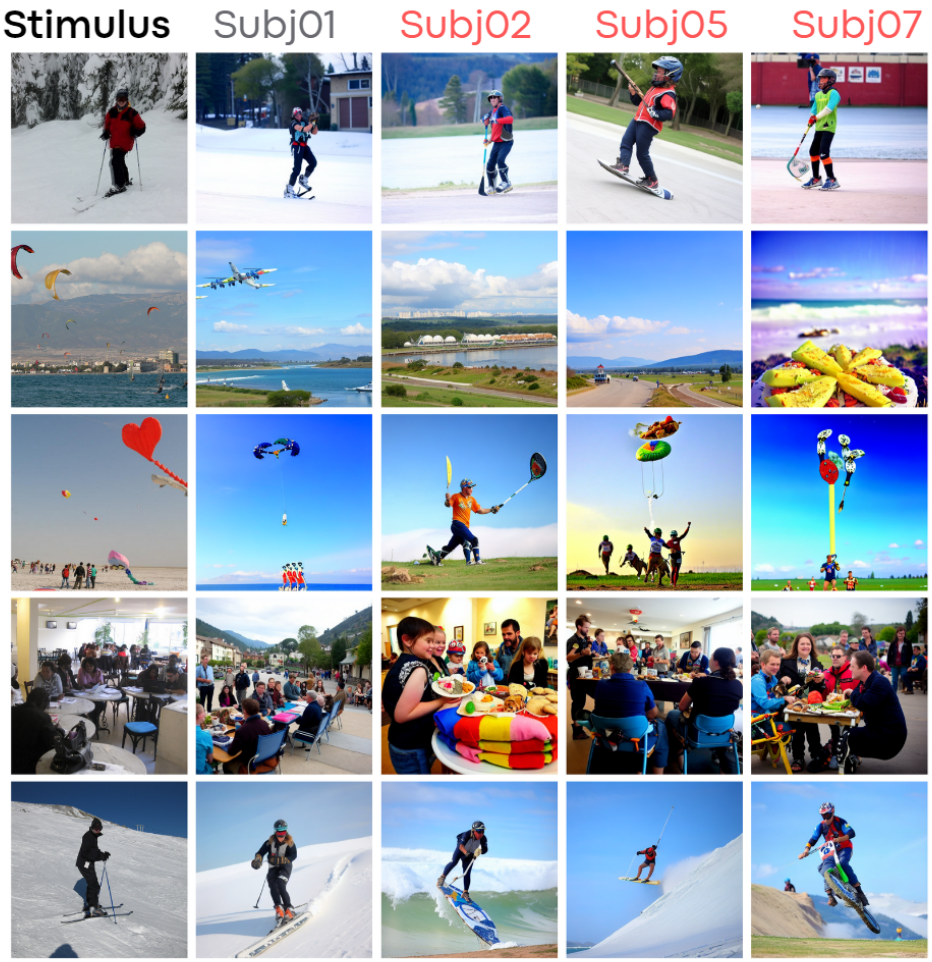


Figure A2: More example results. Format and conventions as in Figure 2

4 Multimodal Brain decoding via Contrastive Learning

This Chapter[1] presents a contrastive learning based approach towards creating a foundational model for aligning neural data and visual stimuli representations by leveraging contrastive learning. We used electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI) data. Our framework's capabilities are demonstrated through three key experiments: decoding visual information from neural data, encoding images into neural representations, and converting between neural modalities. The results highlight the model's ability to accurately capture semantic information across different brain imaging techniques, illustrating its potential in decoding, encoding, and modality conversion tasks.

4.1 Introduction

The non-invasive measurement of neural activity is crucial to understanding the human brain. The advent of artificial intelligence has propelled the field of neuroscience into using novel paradigms, including a wide array of encoding and decoding models. These models have shown remarkable proficiency in interpreting various sensory inputs, encompassing vision, auditory processing, and motor imagery, among others. Key to this endeavor are non-invasive measurements of neural activity such as electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI). Each of these modalities offers a unique window into brain activity, capturing

[1]The work presented in this chapter has been presented at ICLR 2024 Re-align Workshop. Full manuscript is currently under submission to a peer reviewed journal and a preprint version is available [81].

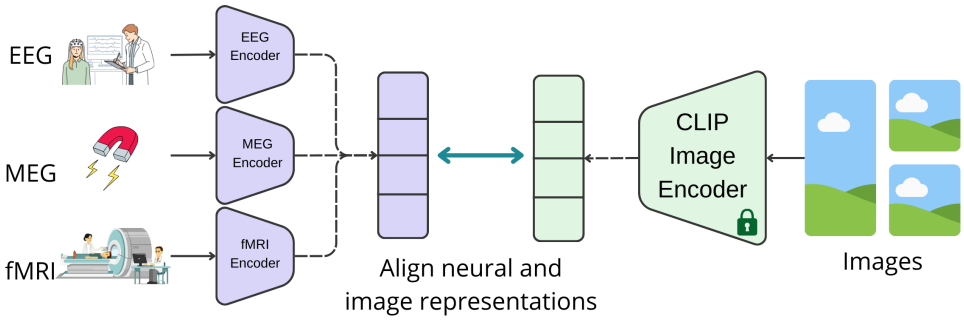


Figure 4.1: Schematic representation of our proposed model, illustrating the alignment of various neural datasets from different modalities into a unified representation space utilizing a frozen CLIP Image encoder.

complementary aspects of its response to external stimuli and providing insights into the perceptual and representational processes within.

In this context, our study introduces a step forward in the realm of neural foundation models for vision. We aim to harmonize disparate modalities drawn from diverse EEG, MEG, and fMRI datasets acquired during a vision task, creating a unified representation that transcends the limitations of individual modalities. Our approach leverages the power of contrastive learning to align representations across these non-invasive measurements of neural correlates and is anchored in the image representations provided by the CLIP (Contrastive Language-Image Pretraining) model [209] as depicted in Fig 4.1.

We assess the capabilities of our framework in performing an array of tasks through information retrieval, as depicted in Fig 4.2. We demonstrate the model's capability in: a) decoding, wherein it can discern and retrieve images corresponding to neural activity recorded during experiments; b) encoding, where it exhibits its potential to predict neural activity patterns from visual stimuli; c) modality conversion, demonstrating the model's ability to translate semantic content across different neural measurement modalities.

This approach not only bridges the gap between neural activity and visual perception but also paves the way for a deeper understanding of how the brain processes and internalizes the visual world. Our work stands at the intersection of neuroscience and artificial intelligence, offering a novel lens through which we can view and interpret the complex narrative of neural activity. It represents a step toward the development of a foundation model for the neuroscience of vision, providing a framework for exploring and understanding the ways in which our brains engage with and make sense of the visual stimuli that permeate our environment.

4.1.a Related Work

Encoding and decoding in vision neuroscience have evolved from classical methods to advanced neural network-based models. fMRI has emerged as a promising tool to extract information with deep learning, connecting biological hypotheses and computational models. Classical encoding predicts brain activity from stimuli, while decoding reconstructs stimuli from brain activity, and it has been shown that both tasks can benefit from a combined approach [183]. Several models have been used for a wide variety of encoding and decoding approaches [283] to analyze fMRI time series obtained in conjunction with visual stimuli, aiming to reconstruct the images linked to observed fMRI patterns or brain activity. Approaches like VAE-GAN have been applied to map fMRI activity to latent representations of human faces using linear models [262]. Additionally, sparse linear regression has been able to predict CNN features for natural images from fMRI data [116]. Recently, diffusion models, noted for their excellent image generation abilities, have become integral to decoding, often employing semantic techniques and multi-step decoding processes [250, 43, 79, 190, 78].

In general, recent advances leverage deep neural networks and large datasets to model complex visual and language representations, enhancing the accuracy of both encoding and decoding models [148, 251, 7, 34, 33, 54, 188, 49]. The Algonauts challenge and subsequent studies emphasize the effectiveness of both pretrained multimodal transformers and tailored models for specific brain regions [86, 2, 184, 279, 46]. Tools like MindEye, which maps brain activity to multimodal latent spaces for precise image retrieval and reconstruction using a contrastive method [236], and DREAM, which replicates image reconstruction from fMRI data, emulating the human visual system [277], are also noteworthy.

Advancements in high-temporal resolution modalities have significantly contributed to the progress in encoding and decoding research within neuroscience. Previous efforts have leveraged linear models for tasks such as image classification from brain activity, prediction of brain activity based on image representations, and inter-modal comparison using representational similarity analysis [222, 87]. Progressing beyond linear approaches, deep neural networks have been employed to classify diverse data types, including speech, mental load, and images, from EEG signals ([54, 192]. EEG, MEG, and fMRI are important tools in non-invasive brain research, each characterized by distinct advantages and limitations. EEG provides the highest temporal resolution, enabling precise temporal tracking of the brain's electrical activity and the observation of rapid neural dynamics. However, its spatial resolution is limited, posing challenges to accurately localizing neural activity. MEG, while offering temporal resolution comparable to EEG, provides slightly better spatial resolution by measuring the

magnetic fields generated by neuronal currents. Nonetheless, the higher costs associated with MEG as well as its lower accessibility limit its widespread use. In contrast, fMRI offers superior spatial resolution, allowing for detailed mapping of brain activity through blood flow changes, albeit with a temporal resolution inferior to that of EEG and MEG, whcih restricts its ability to capture quick neural changes.

Our model extends these ideas by using the principles of contrastive learning, a technique that has yielded promising outcomes in recent fMRI [236] and MEG decoding investigations [54]. These studies concentrate on decoding within a single modality, employing contrastive learning for data retrieval only and in conjunction with other generative methods for stimulus reconstruction. Contrastive learning differentiates between analogous (positive) and non-analogous (negative) data pairs, facilitating the discernment and alignment of semantically congruent representations. In our approach, this methodology is applied not solely to decoding but also to encoding. Through the application of contrastive learning, our model establishes a bidirectional linkage between the visual and neural domains. Moreover, we introduce the concept of neural modality conversion, enabling the translation of semantic content from one neural measurement modality, such as EEG, into another, such as fMRI or MEG. This innovation opens new pathways for comprehensive neural analysis, fostering a more integrated understanding of brain functionality by capitalizing on the synergistic strengths of each modality. The primary contribution of our work is the development of a unified framework adept at managing decoding, encoding, and neural modality conversion, representing a significant advancement beyond existing models that are limited to a single task and modality. By aligning EEG, MEG, and fMRI data through contrastive learning, our model surmounts the inherent limitations of these modalities, offering a versatile infrastructure for the interpretation of neural signals.

4.2 Materials and Methods

Our goal is to make a step towards a shared representation of neural data, i.e., a sort of "foundation model of neural representation of vision". To achieve this, we leveraged a powerful and well-established pretrained model for obtaining image representations—the CLIP Image encoder. We focused on vision processing, selecting a set of human vision datasets where neural activity is measured with different techniques like EEG, MEG, and fMRI.

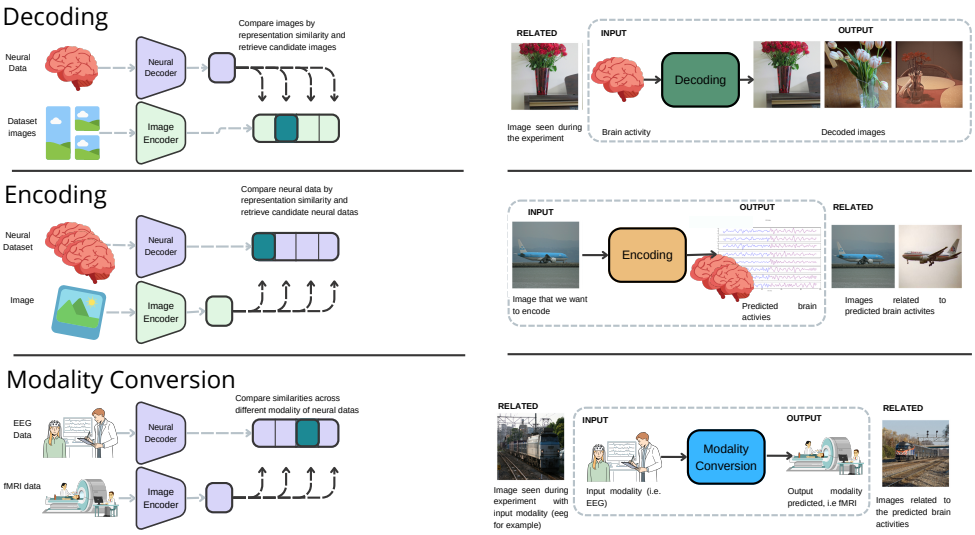


Figure 4.2: The top panel illustrates the 'Decoding' experiment, where neural data is processed to 'decode' and retrieve visually related images from a dataset. The middle panel depicts the 'Encoding' experiment, where an image is used to predict and retrieve neural data that could be associated with the visual perception of that image. The bottom panel shows the 'Modality Conversion' experiment, demonstrating the translation of neural data from one modality, such as EEG, into another, such as fMRI, aiming to find semantically similar brain activity across modalities.

4.2.a Data

EEG: The EEG data for this study were sourced from the ImageNetEEG dataset [245], whcih involves six participants and 40 ImageNet categories [56], totaling 2,000 images recorded at 1000 Hz. The recording protocol involved multiple sessions and sequences, resulting in 11,466 EEG sequences after quality filtering. Preprocessing included notch and band-pass filtering, normalization, and segmentation into 40 ms windows for time-frequency decomposition, producing EEG spectrogram images for model training. To avoid overestimated performances highlighted in [156], a conservative data splitting approach was adopted as described in [192], ensuring more accurate performance assessments.

MEG: in this case, our methodology was evaluated using the "THINGS-MEG" dataset [104]. This dataset involved four participants (two female and two male, average age 23.25 years) who participated in 12 MEG sessions. During these sessions, participants were shown 22,448 distinct images from the THINGS database [105] spanning 720 different categories. Out of this extensive collection, a smaller group of 200 images (each from a unique category) was repeatedly presented to the subjects. Each image was displayed for 500 milliseconds, followed by a

variable fixation period ranging from 800 to 1200 milliseconds. To enhance our retrieval set and demonstrate our method's robustness, we also incorporated an additional 3,659 images from the THINGS dataset that were not shown to the participants. For MEG data preprocessing, the initial step involved downsampling the raw data from the 272 MEG radial gradiometer channels from 1,200 Hz to 120 Hz, followed by centering and standardization. The MEG data was then segmented into epochs extending from 500 ms before to 1000 ms after the onset of each stimulus. The final preprocessing step involved baseline correction, achieved by deducting the average signal value, recorded from the beginning of each epoch to the stimulus onset, for every channel.

fMRI: Here we used the Natural Scenes Dataset (NSD) [5]. This extensive fMRI resource includes data from eight individuals who were shown images from the COCO dataset. Our focus was on four of these subjects (consistent with subjects used in comparable decoding studies), resulting in a dataset comprising 8,859 images and 24,980 fMRI trials for training, and 982 images and 2,770 fMRI trials for testing per participant. To enhance the signal-to-noise ratio, images shown up to three times had their trials averaged. The spatial dimensionality of the fMRI data, recorded at a resolution of 1.8mm, was reduced to approximately 15,000 voxels. This reduction was achieved by applying the NSDGeneral ROI mask, which encompasses several visual areas. Selecting this ROI was crucial for improving the signal-to-noise ratio and reducing the complexity of the data, allowing for a more focused analysis of both low-level and high-level visual features. Temporally, dimensionality reduction was accomplished by using precalculated coefficients (commonly known as "betas") from a General Linear Model (GLM) with a fitted Hemodynamic Response Function (HRF), as detailed in the NSD paper [5], and further denoised as described therein.

4.2.b Neural Vision Alignment

In this section, we focus on aligning the neural representations of different modalities with image representations derived from the CLIP model, specifically aiming to approximate the CLS (Classification) embeddings of images using the pretrained CLIP Image encoder, denoted as h.

For each neural modality, we designed a distinct neural module, represented as f_n. This module is essentially a composite function, $f_n = g_n \circ a_n$, consisting of two primary components. The first component, a_n, is an alignment layer tasked with harmonizing the neural data from various subjects into a unified representation space. Once aligned, these representations are processed by g_n, a shared network that further refines them to closely match the visual representations produced by the image encoder.

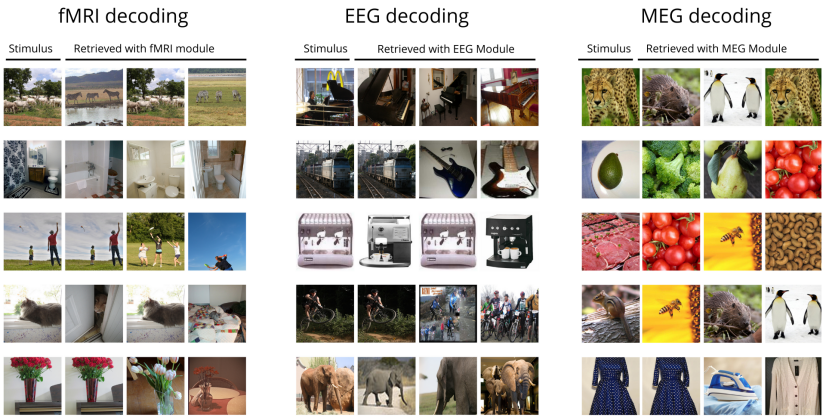


Figure 4.3: Comparative Results of Multimodal Neural Decoding. The figure shows the original visual stimuli and the images retrieved using decoding modules for fMRI, EEG, and MEG data. Each block corresponds to a different modality, illustrating the model's ability to identify and retrieve images that closely resemble or are semantically related to the original stimulus.

To illustrate, consider a subject s who observes an image img while their neural activity n is being recorded. We generate a representation $z_i = f(n, s) = g(a(n, s))$ and, concurrently, we derive the corresponding image representation z_j through the image encoder: $z_j = h(img)$.

Following their generation, these representations are normalized, and the contrastive CLIP loss is calculated, forming the basis of our training regimen. The neural networks for MEG and EEG data are structured as convolutional neural networks (CNNs) resulting in an architecture capable of processing both spatial and temporal patterns in the data. In contrast, the network for fMRI data is configured as a Multilayer Perceptron (MLP), suited for handling the high-dimensional and spatially complex nature of fMRI data.

All networks were implemented using the PyTorch framework and trained using the AdamW optimizer. The training parameters were a learning rate of 3×10^{-4}, weight decay set at 1×10^{-3}, and a batch size of 256. The training process spanned over 30 epochs, ensuring adequate learning while preventing overfitting, leaving room for performance improvement through hyperparameter and neural architecture search in future works.

The contrastive loss function in our model aligns the representations of different modalities with the CLIP model's image representations. Given two normalized representation vectors z_i and z_j, the function proceeds as follows: First, both vectors are normalized using the L2 norm to ensure they lie on a unit hypersphere $z_i = \frac{z_i}{\|z_i\|_2}$ and $z_j = \frac{z_j}{\|z_j\|_2}$.

Then, the similarities (logits) are computed by taking the dot product of z_i and the transpose of z_j, scaled by a temperature parameter τ, so logits $= \frac{z_i z_j^T}{\tau}$.

Targets are defined as a sequence of indices, representing the matching pairs in the batch: *targets* $= [0, 1, 2, \ldots, N-1]$ (where N is the batch size), then transformed into one-hot encoded vectors.

The loss is then calculated using the cross-entropy between the logits and the targets in their one-hot encoded version, so that their value is 1 at index i and 0 everywhere else, thus:

$$\mathcal{L} = -\sum_{i=1}^{N} \log \left(\frac{\exp(\text{logits}_{ii}/\tau)}{\sum_{j=1}^{N} \exp(\text{logits}_{ij}/\tau)} \right) \tag{4.1}$$

This formulation of the contrastive loss function efficiently aligns the neural representations z_i with the corresponding image representations z_j by maximizing cosine similarity for positive pairs and minimizing it for negative pairs. It is modulated by the temperature hyperparameter τ, which is set as 1 in our experiments but can also be learned or modulated during training.

4.2.c Experiments

We demonstrate the versatility of our model through three distinct experiments, each focused on a unique application in neural data analysis.

Decoding Visual Information: This involves starting with neural data (EEG, MEG, or fMRI) and projecting it into a common representation space using the corresponding neural model. Concurrently, we process all images in the test set specific to each modality (comprising 337 for EEG, 2400 for MEG, and 982 for fMRI) and compute their similarities. The goal is to retrieve the top-n images that most closely match the neural signal representation. This process effectively allows us to "decode" the neural data into potential visual images that the subject might have been perceiving.

Encoding: Here, we begin with an image, pass it through the image encoder, and simultaneously process all neural data of a particular modality through the neural encoder. We then search for the top-n neural representations in the shared space and retrieve the corresponding neural data in the test set. This approach enables us to obtain "encoded" neural representations corresponding to the given image.

Modality Conversion: Here we capitalize on the alignment of all modalities in the same representation space. For instance, given the fMRI representation of a subject who has viewed a specific image, we might ask: what could be the EEG or MEG activity resulting in viewing a semantically similar image? To answer

this question, we encode the sample from our input modality and the target search set from the desired output modality, selecting the top-n matches based on cosine similarity. To validate the effectiveness of this modality conversion, we compare the images associated with the source modality during data acquisition with those linked to the target modality.

These experiments collectively illustrate the robustness and multifaceted capabilities of our model, offering significant advancements in the fields of neural encoding, decoding, and inter-modality translation.

4.2.d Evaluation

To assess the performance of our model, we employed various metrics that gauge its proficiency in extracting relevant semantic information from neural data.

Decoding Performance: For decoding, the evaluation methodology is straightforward. The ImageNetEEG and THINGS MEG datasets have distinct classes (40 and 720, respectively), allowing us to calculate and compare top1 and top5 accuracy directly against chance levels and established baselines like [192]. In contrast, the Natural Scene Dataset (NSD) used for fMRI comprises complex scenes from the COCO dataset without distinct classes. To evaluate decoding performance in this context, we employed the CLIP 2-way accuracy metric, comparing with the state of the art [236]. For consistency and ease of comparison, we extended this measure to the EEG and MEG settings as well.

Encoding Performance: In the encoding scenario, where images are input to retrieve neural data, we faced the challenge of the latter being inherently difficult to interpret and visualize. To ascertain whether relevant information was captured, we relied on an indirect metric. Each sample of neural data is associated with a specific image viewed during the vision experiments. By encoding these images, we get candidate neural representations of stimuli. We then compute the CLIP 2-way accuracy between the neural activity relative to the encoded image (ground truth) and the activities retrieved with the encoding model. This process involves starting with an image, obtaining candidate neural data, and then visually comparing the representation of images related to those candidates with the original image. A successful encoding process would typically result in the selection of neural data associated with the observation of semantically similar images.

Modality Conversion Performance: The approach for evaluating modality conversion is similar to previous tasks in the pipeline. We initiate the process with neural data from one modality and obtain representations in another. Naturally, since the datasets come from different subjects, we also obtain activity from another subject, which ideally has the closest representation in the shared space.

Table 4.1: Model performance in decoding experiment. "Baselines" columns refer to top1 and top5 chance level, while "retrieval dataset size" is useful to put CLIP 2-way accuracy in context with other works.

Neural Module	top1 accuracy	top5 accuracy	CLIP 2 way	baseline accuracy (%)	baseline top5 (%)	Retrieval dataset size	Number of classes
EEG (ImageNet)	40.0	54.3	79.4	2.5	12.5	332	40
MEG (THINGS)	1.2	6.1	60.1	0.13	0.65	2400	720
fMRI (NSD)	-	-	90.3	-	-	982	-

The performance of this aspect of our model is gauged using the CLIP 2-way accuracy between the images related to the source and target modalities. This metric effectively measures how well our model translates semantic information across different neural modalities.

4.3 Results

The model's performance in decoding neural data into corresponding visual stimuli is quantified and presented in Table 4.1. The EEG module achieved a top1 accuracy of 40.0% and a top5 accuracy of 54.3%, which is a substantial improvement over the baseline chance level accuracy figures of 2.5% for top1 and 12.5% for top5 accuracies. Notably, this module's CLIP 2-way accuracy reached 79.4%, indicating the model's capability to decode EEG data with high reliability.

Unfortunately, directly comparing these performances with literature could be difficult, since recent work which at first sight delivers impressive performances [11, 193, 131] on this dataset have been seen to rely on contamination between train and test data due to an incorrect use of preprocessing choices [156]. When comparing results with work that explicitly preprocesses data in order to avoid this confounding factor, we found performances that are on par with the state of teh art [192, 77], i.e. top1 accuracy within range (39-45%) for a multisubject network trained for classification.

For the MEG module, which faced the greater complexity of the THINGS dataset (720 classes), the model achieved a top1 accuracy of 1.2% and a top5 accuracy of 6.1%, which corresponds to 10 times the baseline chance level accuracies of 0.13% for top1 and 0.65% for top5. The CLIP 2-way accuracy stood at 60.1%. This demonstrates the model's adeptness in handling a more extensive

and varied set of images, comprising a total of 2,400 images. Again, these results are comparable with recent work on the same dataset which focused only on decoding of MEG images [17] where top5 accuracy was tested in several settings with performances in the range [1-8%].

The fMRI module's decoding performance was particularly notable, achieving a CLIP 2-way accuracy of 90.3% with the NSD dataset. Our results align well with findings from recent literature [250, 78, 190, 43, 236], though direct comparisons are challenging due to the varied focus and methodologies of these studies. Many of these works concentrate on the detailed reconstruction of stimuli using complex pipelines that involve regressing fMRI data to a latent space, generating images, and then computing CLIP 2-way accuracy between the generated and actual images. These processes typically involve training individual models for each subject. Despite these differences, the performance metrics reported in these studies, which generally fall within the range of [77-95%], are comparable to our results. This situates our approach within the high-performance spectrum for multisubject settings. This high accuracy, obtained with a retrieval set of 982 images, underscores the model's proficiency in decoding complex scene representations from fMRI data.

Fig 4.3 provides details on the model's decoding capabilities with fMRI, EEG, and MEG data. In fMRI Decoding, the model's quality is evident. For example, when presented with a stimulus depicting grazing animals, the fMRI module accurately retrieves other images of animals in similar pastoral settings. This level of detail suggests the model's strong semantic grasp, as it not only identifies the subject of the image but also the context in which it is situated.

The EEG Decoding column showcases a broad understanding of the visual stimulus categories. Good examples are showcased by the piano and elephant cases (first and last row of EEG panel in Fig 4.3), indicating its capacity to capture the broader concepts of objects and animals.

MEG Decoding presents a blend of moderately related and thematically similar images. A notable example is the retrieval of images of leopards and penguins in response to a stimulus of the jaguar (first row), demonstrating the model's nuanced semantic retrieval. However, the module also retrieves images that are thematically related but not identical, such as different fruits and vegetables in response to a stimulus of a specific type, suggesting a wider semantic reach of the MEG module compared to fMRI.

In the encoding experiment, the model's efficacy in mapping visual stimuli to neural activities was reflected in the figure 4.4 for each modality. The EEG encoding results showed a high semantic correlation with the original encoded image and achieved a CLIP 2-way accuracy of 85.5%. The MEG encoding results

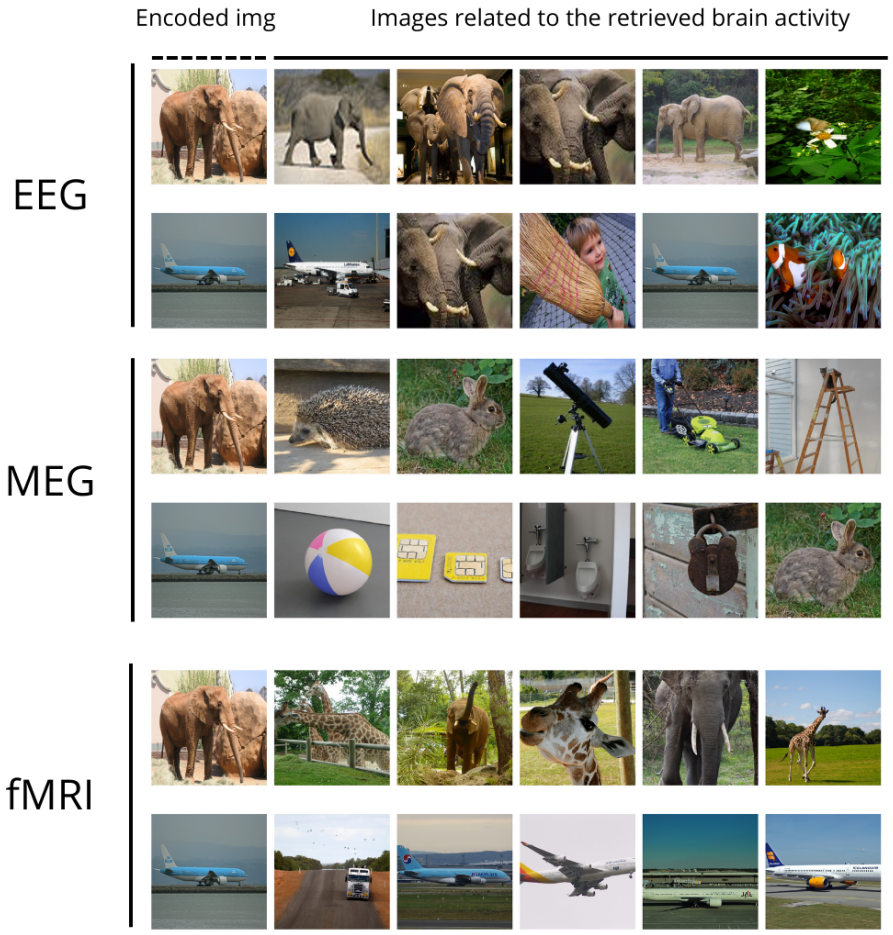


Figure 4.4: Encoding Experiment Results Displaying Image-to-Brain Activity Correlation. Rows illustrate the results for EEG, MEG, and fMRI modalities. The leftmost images are the encoded stimuli, and the subsequent images represent images related to the brain activities retrieved by the model.

depicted a somewhat broader semantic range, achieving a CLIP 2-way accuracy of 58.8%. The fMRI encoding, with yielded the highest precision in matching the semantic content of the original image, reached a CLIP 2-way accuracy of 87.8%. The figure 4.4 effectively showcases the encoding proficiency of our model in transforming visual stimuli into corresponding neural activity representations across various modalities.

For example, in the EEG Encoding section (first two rows of the image), the model skillfully associates an elephant image with EEG activity that reflects similar scenes, indicating a nuanced understanding of the visual to neural translation.

The MEG Encoding section displays a variety of images associated with the encoded images. Although some matches in broader semantic categories were found (e.g., things with airplanes and some animals with elephants), the limited information content of the signal likely causes frequent retrieval failures.

Lastly, fMRI Encoding stands out with its precise retrieval of images to fMRI activity, confirming the model's high precision in encoding complex visual information.

Overall, the figure illustrates the model's capability to accurately encode visual stimuli into neural representations, suggesting its potential application in predicting brain activity from visual inputs across EEG, MEG, and fMRI data modalities.

Lastly, the modality conversion experiment results, as detailed in Table 4.2, indicated the model's capability to accurately transform neural information across modalities. The normalized accuracies for conversions like fMRI to MEG were exceptionally high (95.40%), highlighting the model's ability to preserve semantic content through the conversion process. These conversions demonstrate the model's potential to provide a harmonized representation of brain activity, bridging the gap between different brain imaging techniques. Collectively, these results underscore the capabilities of our model that to solve several tasks. It delivers performances whcih are on par with recent literature in neural decoding tasks, and also is able to encode visual content into neural patterns. Finally, it is successful in converting neural information across modalities, setting a foundation for future breakthroughs in multimodal neural data analysis and interpretation.

4.4 Discussion

The development of our neural foundation model stands as a pioneering stride towards an integrated understanding of the brain's mechanisms through neural

Table 4.2: Conversion modality CLIP-2way accuracies and their normalized values with respect to the decoding performances.

Conversion	Clip 2-way decoding accuracy	Normalized Clip 2-way decoding accuracy
fMRI to EEG	0.6710	0.8370
MEG to EEG	0.6790	0.8470
fMRI to MEG	0.5679	0.9540
EEG to MEG	0.5594	0.9396
EEG to fMRI	0.7648	0.8598
MEG to fMRI	0.7928	0.8912

data. This work signifies the first step in creating a foundational framework akin to what has been seen with large language models in the field of natural language processing. It encapsulates a multi-modal approach that not only decodes but also aligns neural representations from a variety of datasets and modalities, bringing us closer to a shared neural representational space.

A pivotal aspect of this model is its capacity for multi-modal (and subject) representation alignment, effectively creating a shared representation space that harmonizes individual variability. This is particularly reminiscent of the convergence of different languages and dialects into a singular, coherent narrative—where the model serves as an interpreter of the brain's complex 'dialects' of activity.

However, aligning data from disparate neural recording modalities comes with several challenges, ranging from technical discrepancies to differences in spatiotemporal resolution.

This work has navigated some of these complexities, yet the integration process remains a sophisticated and elaborate task, and our model presented is not without limitations. Its current non-generative nature and reliance on diverse, pre-existing datasets indicate that it remains a proof-of-concept. Looking to the future, our goal is to turn this model into a generative tool that could aid in data augmentation and facilitate virtual experiments. The addition of further modalities such as language and audio, alongside more extensive fMRI, EEG, and MEG data, could pave the way for a comprehensive "latent brain representation." This representation would transcend individual modalities, offering a more holistic view of brain activity.

In addition to technical advancements in the field, one must not forget the critical issue of neural data privacy. As our models become increasingly capable of decoding detailed information from neural signals, the imperative for privacy

safeguards grows. All datasets used in this study are in the public domain, and participant consent was obtained in all cases. Nevertheless, future developments could lead to models that decode personal data from minimal scanning, necessitating minimal cooperation from participants. This raises the specter of privacy concerns, as well as the issue of ethics in the use of such technology. It is crucial that we engage in proactive discussions on these topics to avoid potential misuse of neural decoding technologies and ensure that model-generated content can be distinguished from true subject experiences, preventing the propagation of harmful material.

In sum, while this first step towards a neural foundation model marks a significant advancement in our approach to understanding and interpreting neural data, it also beckons us to contemplate the ethical framework within which such technology should operate. As we enhance the model's capabilities and expand its applications, we must concurrently fortify the ethical boundaries that will preserve the privacy and integrity of individual neural data.

4.5 Conclusion

This paper introduces a new step towards a neural foundation model that aligns representations of multi-modal neural datasets using contrastive learning, marking a significant advance in the field of neuroscience. Our model has demonstrated considerable success in decoding, encoding, and converting neural signals, showing its potential to unravel the complex semantics of brain activity. While promising, the model's current non-generative nature and reliance on diverse datasets indicate areas for future enhancement. The next steps involve expanding the model's capabilities to include generative abilities and additional modalities, moving towards a comprehensive representation of brain activity. Crucially, this research also highlights the emergent issue of neural data privacy, necessitating a collaborative effort to establish ethical guidelines for future advancements. As we continue to explore the depths of neural data interpretation, we remain committed to advancing scientific understanding while upholding the utmost respect for individual privacy and data integrity.

Part III

Language and Brain

5 Contrastive learning to identify sentences

In this Chapter[1] a novel contrastive learning approach is proposed to decode brain activity into sentences by mapping fMRI recordings and text embeddings into a shared representational space. Using data from three subjects, we trained a cross-subject fMRI encoder and demonstrated effective sentence identification with a retrieval module. Our model shows strong alignment between brain activity and linguistic features, with top-1 accuracy up to 49.2% and top-10 accuracy up to 84%, significantly outperforming chance levels. Our results highlight the potential of contrastive learning for cross-subject language decoding,

5.1 Introduction

Language is one of the most ubiquitous ways we experience the flow of information in our daily lives. We read, speak, communicate, and even think through language. It is a complex phenomenon that allows us to understand and convey information to others. To do this, our brain generates semantic representations of everything we encounter, taking context into account. Recent research has demonstrated a convergence between brain activity [37, 17, 93] during language tasks, such as listening or reading, and the way large language models process sentences. This has been shown through brain recordings using both non-invasive methods, such as fMRI, EEG, and MEG, as well as invasive techniques like LFP and ECoG. Powerful encoding and decoding models have been developed to link external stimuli, such as acoustically perceived sentences, images, videos, music with neural representations recorded during these tasks

[1]The work presented in this chapter has been presented at UniReps 2024 NeurIPS workshop. [76].

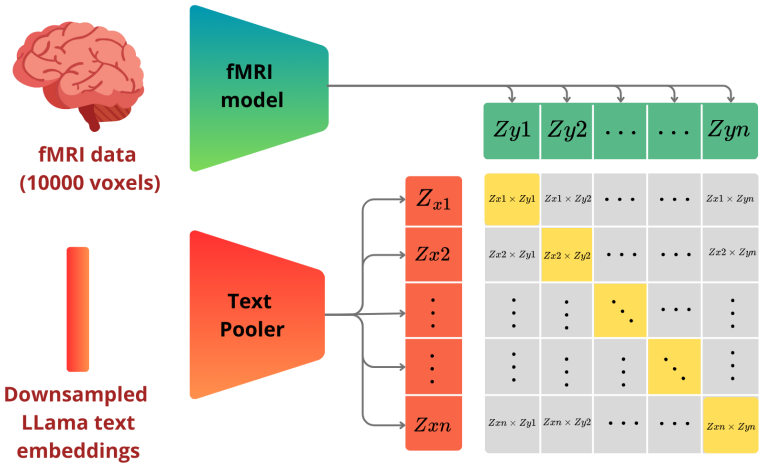


Figure 5.1: Schematic representation of the contrastive learning pipeline. The fMRI model encodes brain activity from 10,000 voxels located in the cortex into a shared latent space, while the text pooler processes downsampled LLama text embeddings. Pairwise dot products between fMRI and text embeddings are computed to create a similarity matrix, which is used for aligning brain activity with corresponding linguistic features. Yellow cells highlight correct matches between fMRI and text embeddings.

[251, 7, 78, 236, 58, 248, 158, 190, 80, 74]. These models have shown remarkable results across both invasive and non-invasive brain recording techniques. Here, we propose a novel method based on contrastive learning to learn a cross-subject fMRI encoder that projects fMRI recordings and text embeddings into a shared space, enabling sentence identification with a retrieval module. Our model takes as input fMRI activity and computes fMRI embeddings that can be compared with pre-computed sentence embeddings belonging to a set of candidates. By selecting the closest sentences in this space we can effectively decode the brain activity and obtain clues about semantic representation in the brain. Fig. 5.1 show a scheme of the pipeline proposed here.

5.2 Material and Methods

5.2.a Data

In our analysis, we used the publicly available dataset from [148]. We focused on subjects S1, S2, and S3, each of whom underwent fMRI recordings for approximately 16 hours while listening to 83 stories from The Moth and Modern Love podcasts. The fMRI data was collected using a 3T Siemens Skyra scanner with a repetition time (TR) of 2.00 seconds and an isotropic voxel size of 2.6 mm.

This allowed for the measurement of BOLD signals, which reflect neural activity with an inherent delay due to the hemodynamic response. Preprocessing steps included motion correction, cross-run alignment, standardization and detrending of low-frequency. For more details, we refer the reader to the original paper [148]. Data from the first 70 stories of each subject were used as training set, while the other 12 stories were used as validation set. Also, the story *"wheretheressmoke"*, which was listened to 10 times to ensure better signal to noise ratio at test set, aligning with recent literature on language encoding and decoding [251, 7].

5.2.b Encoding

The first step of our pipeline involves reducing the complexity of the signal we need to process. While much of the brain is active during language and semantic tasks [123, 122], certain regions are more directly involved in language processing. Therefore, we begin by identifying cortical regions that, at the voxels level, are more easily modeled by a language model, allowing us to focus our analysis on these regions. We trained an encoding model of brain activity using a large language model (LLM) as the foundation. For each word in the training stories, we computed embeddings from LLama3-8B [68], specifically from the 13th layer, using a context window of the previous five words. This choice of layer is supported by previous studies [7] that found early layers in LLama models exhibit stronger correlations with brain activity. To align the word embeddings with the fMRI temporal resolution, we downsampled them using a Lanczos filter, creating matching sentence-fMRI pairs. Finally, we calculated the Pearson correlation between the predicted and true brain activity on held-out validation data, selecting the top 10,000 cortical voxels as our target regions. The activity in these voxels are the targets we aim to embed and decode.

5.2.c Contrastive fMRI model

The core contribution of this work lies in the proposed representation learning pipeline. Since a single TR of fMRI recording can be influenced by several preceding words due to the hemodynamic response (HRF), we must account for this effect. To address the unique properties of fMRI data, we developed a model consisting of two neural networks: an fMRI encoder and a text embedding pooler. The fMRI model is a cross-subject neural network. The first layer is subject-specific and projects the top 10,000 acivity corresponding to the top 10000 voxels of each subject into a shared representation space of dimensionality *common_dim*. The remaining layers of the network are shared across subjects, consisting of a

multi-layer perceptron (MLP) that transforms the data from *common_dim* to the final latent dimension *latent_dim*. The text embedding pooler is an MLP that processes four downsampled word embeddings corresponding to the TRs that might affect the measured brain activity (i.e. those from 1, 2, 3, and 4 timepoints before the corresponding BOLD response). Each word embedding is projected into the final latent space by a linear layer that reduces the dimensionality. These projections are then concatenated, resulting in a final sentence representation of dimensionality *latent_dim*. In appendix, we detail all the hyperparameter search and architectures. Let x represent the fMRI activity, y the downsampled word embeddings, f the fMRI model, and g the text pooler. We define the projections as: $z_x = f(x), \quad z_y = g(y)$. Let $z_x \in \mathbb{R}^{n \times d}$ represent the encoded fMRI features and $z_y \in \mathbb{R}^{n \times d}$ represent the encoded text embeddings, where n is the batch size and d is the latent dimensionality. The logits matrix, which contains the similarity scores between each pair of fMRI and text embeddings, is computed as:

$$\text{logits}_{ij} = \frac{z_{x_i} \cdot z_{y_j}^{\top}}{\tau}$$

where τ is the temperature parameter that controls the sharpness of the distribution. The pairwise similarities are computed using the dot product between the fMRI and text embeddings: $\text{similarities}_{ij} = z_{x_i} \cdot z_{y_j}^{\top}$. Let the target labels be the identity mapping, where each input is matched with itself in the contrastive learning task. The targets vector $t \in \mathbb{N}^n$ is defined as: $t = \{0, 1, \ldots, n-1\}$ This implies that $t_i = i$, ensuring each fMRI sample is matched with the corresponding text sample. The contrastive loss $\mathcal{L}$ is computed as a combination of two cross-entropy losses, one for the alignment of fMRI to text and the other for text to fMRI:

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{\text{CE}}(\text{logits}, t) + \mathcal{L}_{\text{CE}}(\text{logits}^{\top}, t)\right)$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss function. This formulation optimizes both directions in the contrastive learning objective, ensuring that fMRI features are closely aligned with the corresponding text embeddings and vice versa. This contrastive learning approach ensures that both fMRI activity and the corresponding text embeddings are projected into the same latent space, aligning brain activity with linguistic features.

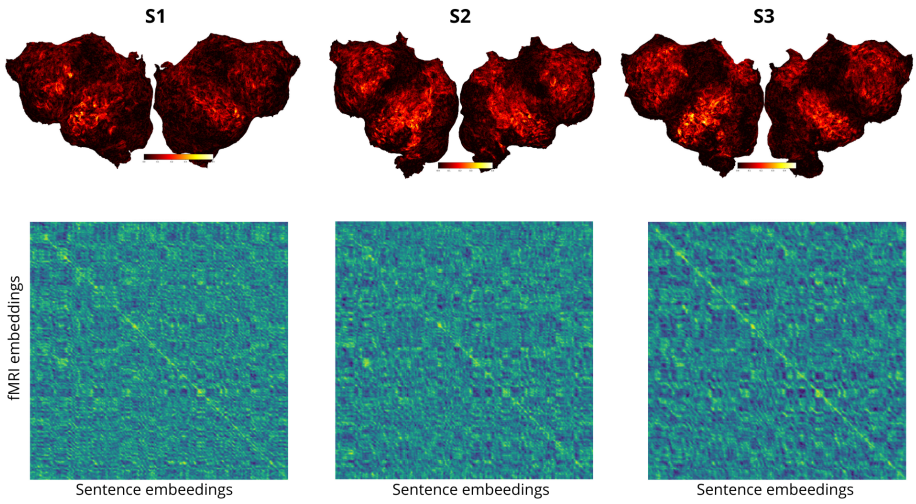


Figure 5.2: Top: Pearson correlation maps of predicted vs. actual brain embeddings for subjects S1, S2, and S3 shown on flattened cortical surfaces. The regions we identified are typically associated with language processing, such as the superior temporal gyrus and parts of the inferior frontal cortex. Bottom: Cosine similarity matrices between sentence embeddings and brain embeddings for each subject.

5.2.d Retrieval and Evaluation

The retrieval module compares the L2 distance between fMRI and text embeddings across the test set to find the closest sentences to each fMRI sample. We evaluate the decoded sentences using three metrics: identification accuracy, top-1 accuracy, and top-10 accuracy. For top-10 retrieval, we select the 10 closest sentences to each fMRI TR based on L2 distance, with the target sentence defined as the previous 4 TRs (8 seconds) plus 5 preceding context words. Identification accuracy, adapted from vision and music decoding literature, measures how well the model identifies the correct sentence by comparing self-correlations in the latent space with other correlations. We compute Pearson correlations between the predicted vectors and targets, storing the results in a correlation matrix, and successively calculate identification accuracy by comparing the self-correlation with others in the same row and normalizing the result.

5.3 Results

The results from our contrastive learning-based language decoding model are shown in Figure 5.2 and Table 5.1. The top panel of Figure 5.2 illustrates the model's encoding performance, with Pearson correlations between predicted and actual brain embeddings across three subjects (S1, S2, S3). The red regions

indicate strong correlations, suggesting the model captures language-related cortical activity patterns, likely in language-processing areas of the brain. The consistent performance across subjects highlights the model's robustness in identifying key neural features associated with sentence comprehension. In the bottom panel, the cosine similarity matrices between sentence embeddings and brain embeddings show a strong alignment, indicated by bright diagonal lines, reflecting the model's ability to map linguistic structures to brain representations effectively. Table 5.1 provides quantitative metrics for decoding accuracy. The Top-1 Accuracy, which ranges from 0.313 (S2) to 0.498 (S3), significantly outperforms chance levels, confirming the model's ability to predict precise sentences. Top-10 Accuracy further validates this, with values as high as 0.838 (S3), indicating that the correct sentence is frequently among the top 10 predictions. Identification Accuracy is also high for all subjects, ranging from 0.910 (S2) to 0.962 (S3), reinforcing the model's strong performance in decoding brain representations of sentences. Overall, both Figure 5.2 and Table 5.1 demonstrate the model's effectiveness in linking sentence embeddings to brain activity, with strong performance across subjects. Subject S3 consistently shows the best results, suggesting individual differences in brain activity may influence decoding accuracy, offering avenues for future investigation.

Subject	Top-1 Acc	Top-10 Acc	Chance Level Top-1	Chance Level Top-10	Identification Acc
s1	0.3780	0.786	0.0114	0.0894	0.9571
s2	0.3127	0.6666	0.0116	0.0855	0.9100
s3	0.4982	0.8381	0.0118	0.0814	0.9624

Table 5.1: Performance metrics for Top-1, Top-10, and Identification Accuracy across different datasets.

5.4 Discussion and Conclusions

Contrastive learning has proven to be a robust method for learning cross-subject mappings between brain activity and sentence-level embeddings. However, a key limitation of our approach is that decoding is performed through a retrieval module (i.e., sentence identification). This requires access to candidate sentences beforehand, limiting the model's ability to generalize to brain activity related to sentences that differ significantly from those in the training dataset. Another important limitation pertains to future applications of this work. The learned fMRI embeddings could potentially be used in conjunction with Bayesian de-

coding techniques or as inputs to modified large language models (LLMs) for open vocabulary decoding. However, these approaches raise concerns about privacy and bias. It will be crucial to address how to distinguish between actual thoughts and brain representations, and how to prevent biases in both models and human interpretations from influencing the results. Future research should explore the concept of neural privacy and develop strategies to disentangle model biases from genuine cognitive processes. In conclusion, this work presents a cross-subject architecture that decodes brain activity into sentences using contrastive learning and sentence identification, laying the groundwork for future advancements in brain-to-language decoding.

Appendix

5.4.a Neural Network Architectures

This appendix describes the architecture of the neural networks used in this work. The models consist of an `Encoder` for processing the input data, an `Embedding Pooler` for projecting embeddings to a common latent space, and a `Contrastive Model` for learning via contrastive losses.

Encoder

The `Encoder` network is designed to map the input data to a lower-dimensional latent space. The architecture is summarized in Table A1. $input\_dim$ is set to be the number of cortical voxels (10000) while $common\_dim$ is chose to be 4096.

Table A1: Encoder Network Architecture

Layer Type	Dimensions	Activation Function
Input Layer	$input\_dim \times hidden\_dim$	-
Alignment Layer (key = 1)	$input\_dim \times common\_dim$	-
Alignment Layer (key = 2)	$input\_dim \times common\_dim$	-
Alignment Layer (key = 3)	$input\_dim \times common\_dim$	-
Layer Normalization	$common\_dim$	-
Linear Layer	$common\_dim \times hidden\_dim$	Identity
Linear Layer	$hidden\_dim \times latent\_dim$	-
Layer Normalization	$latent\_dim$	-

The alignment layers apply different transformations to subsets of the data depending on key values provided at runtime. These layers are followed by sequential linear layers with ReLU activation.

Embedding Pooler

The `Embedding Pooler` projects embeddings, such as fMRI data, into a lower-dimensional latent space. The architecture is summarized in Table A2.
The input is first normalized, followed by a sequence of linear layers and GELU activation. The final step reshapes the output into a pooled embedding vector.

Contrastive Model

The `Contrastive Model` combines the `Encoder` and `Embedding Pooler` for multimodal learning using contrastive loss functions. The model supports various

Table A2: Embedding Pooler Architecture

Layer Type	Dimensions	Activation Function
Input Layer	$input\_dim \times hidden\_dim$	-
Layer Normalization	$input\_dim$	-
Linear Layer	$input\_dim \times hidden\_dim$	GELU
Linear Layer	$hidden\_dim \times (latent\_dim/4)$	GELU
Layer Normalization	$(latent\_dim/4)$	-
Reshaping	$(batch\_size, -1)$	-

loss functions, such as contrastive loss, mean squared error (MSE), and cosine similarity loss. The overall architecture is shown in Table A3.

Table A3: Contrastive Model Components

Component	Description
Encoder	See Table A1
Embedding Pooler	See Table A2
Loss Function	Contrastive

During training, the model minimizes the distance between fMRI projections and embeddings from other modalities. The training process is logged and monitored using the mean squared error, cosine similarity, and the primary loss function.

5.4.b Training Setup

The models are trained using an AdamW optimizer with a learning rate of $1e-4$ and weight decay of $1e-4$. The learning rate is adjusted dynamically using a scheduler that reduces the rate by a factor of 0.1 when the validation loss plateaus for 50 epochs.

Table A4: Training Parameters

Parameter	Value
Optimizer	AdamW
Learning Rate	$1e-4$
Weight Decay	$1e-4$
Learning Rate Scheduler	ReduceLROnPlateau (patience: 50)
Epochs	3

The models are trained for a maximum of 3 epochs using PyTorch Lightning's `Trainer`, with all computations performed on a single GPU device. The data are

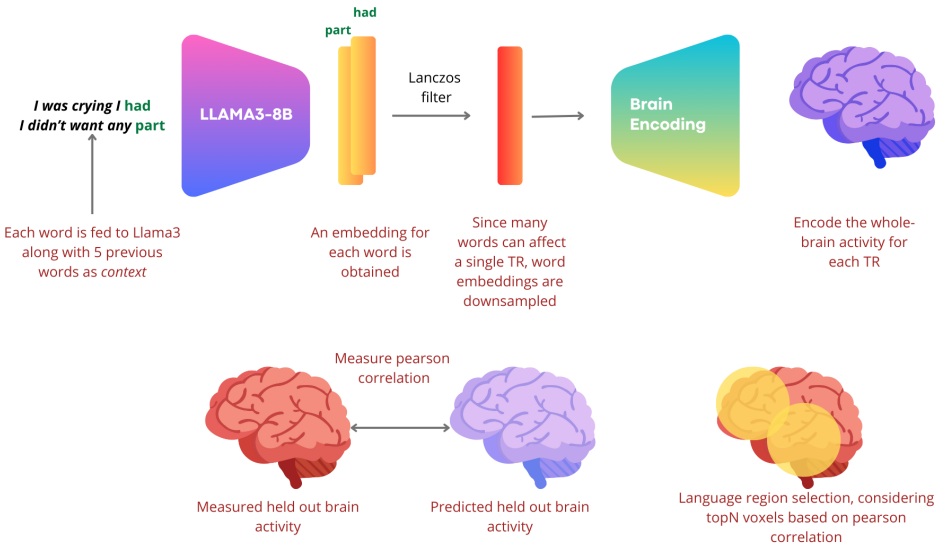


Figure A1: Encoding scheme

publicly available and can be requested at https://openneuro.org/datasets/ds003020/. All experiments and models were trained on a server equipped with four NVIDIA A100 GPU cards (80GB RAM each connected through NVLINK) and 2 TB of System RAM.

5.5 Comparative Analysis of Generative Decoding and Contrastive Learning Approaches

In this section, we detail the connections and differences between the generative decoding approach described in [251] and our proposed contrastive learning decoder. The aforementioned work employs a Bayesian encoding technique, learning subject-specific encoding and noise models from data, and estimates likelihood probabilities using a large language model (LLM) as a generator of candidate sentences. This enables open-vocabulary text generation guided by fMRI data. In contrast, our work adopts a more modern and flexible approach based on contrastive learning, aiming to learn cross-subject latent representations of both text and fMRI data.

We will delve into the mathematical foundations of both methods and highlight their similarities and differences.

Let y denote the fMRI data with shape (t, v), where t represents time and v represents the number of voxels. Let x be the downsampled text features extracted by a pretrained language model (GPT-1 for Bayesian decoding and

LLaMA3-8B for contrastive decoding), with shape (t, f), where f denotes the number of text features.

The encoding model is defined by a set of learned weights W that maps x to y, with W having shape (f, v):

$$\hat{y} = xW$$

In the Bayesian approach, y is modeled as a multivariate Gaussian distribution:

$$y \sim \mathcal{N}(xW, \Sigma)$$

where Σ is the covariance matrix of shape (v, v).

Here the goal is to model the posterior distribution $p(x \mid y)$. Applying Bayes' theorem, we can write:

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

Since $p(y)$ is constant with respect to x, we focus on the numerator $p(y \mid x)p(x)$. The problem thus reduces to estimating a good encoding model (i.e., accurate estimates of $p(y \mid x)$) and utilizing a pretrained language model to estimate the prior probabilities $p(x)$.

The likelihood $p(y \mid x)$ can be expressed as:

$$p(y \mid x) \propto \exp\left(-\frac{1}{2}(y - xW)^\top \Sigma^{-1}(y - xW)\right)$$

Taking the negative logarithm yields the loss function:

$$\mathcal{L} = (y - xW)^\top \Sigma^{-1}(y - xW)$$

Here, the quadratic form represents the residuals between the model predictions and the measurements, re-weighted by the inverse covariance matrix.

Expanding the terms, we obtain:

$$\mathcal{L} = y^\top \Sigma^{-1} y - 2xW\Sigma^{-1}y^\top + xW\Sigma^{-1}W^\top x^\top$$

This results in a (t, t) matrix of residuals. By taking the trace, we obtain a scalar loss function that we can minimize to learn the encoding model weights.

The most significant term is the interaction term:

$$xW\Sigma^{-1}y^\top$$

which involves matrices of shapes:

$$(t, f) \times (\mathbf{f}, \mathbf{v}) \times (\mathbf{v}, \mathbf{v}) \times (v, t)$$

The learnable parameters here are the encoding weights W and potentially the inverse covariance matrix Σ^{-1}.

In our contrastive model, we employ two learned functions, approximated by neural networks, to map x and y into a shared latent space z of dimensionality d:

$$z_x = f(x), \quad z_y = g(y)$$

where $f : \mathbb{R}^f \to \mathbb{R}^d$ and $g : \mathbb{R}^v \to \mathbb{R}^d$.

To simplify the mathematical analysis and highlight the differences and similarities between the two methods, let's assume that both f and g are linear functions:

$$z_x = xA, \quad z_y = yB$$

with A being a matrix of shape (f, d) and B a matrix of shape (v, d), so that z_x and z_y both have shape (t, d).

The objective of the contrastive loss is to make the cosine similarity matrix between z_x and z_y as close as possible to the identity matrix I. This can be achieved by computing:

$$S = \frac{z_x}{\|z_x\|} \left(\frac{z_y}{\|z_y\|} \right)^\top$$

where S has shape (t, t). We can then use a cross-entropy loss for each element along the diagonal or compute the mean squared error between S and I.

By expanding the calculation and ignoring the normalization terms for simplicity, we obtain:

$$S = z_x z_y^\top = xAB^\top y^\top$$

where the matrix multiplications involve shapes:

$$(t, f) \times (\mathbf{f}, \mathbf{d}) \times (\mathbf{d}, \mathbf{v}) \times (v, t)$$

We observe a key connection between the two models: in both cases, we have a similarity (or dissimilarity) matrix of shape (t, t) where the interaction between x and y plays a crucial role.

In the Bayesian model, the encoding projects text features into the brain space and re-weights them based on the noise model, whereas in the contrastive model,

this process occurs implicitly through the interaction of the functions f and g. In the linear case, the product $AB^\top$ takes on a role analogous to $W\Sigma^{-1}$.

However, in the contrastive approach, text features are projected into a latent space (typically with $d \ll v$), resulting in a less descriptive model. This acts as an implicit regularization, but if d is less than the rank of Σ^{-1}, some information might be lost. This is a suggestion that higher dimensionality plays an important role in brain decoding of language, guiding us in our hyperparameter search.

Thus, while the contrastive model offers greater flexibility—such as training cross-subject models, incorporating nonlinearities, and utilizing compressed latent spaces—it may sacrifice some information about the relationship between x and y. A potential future direction could involve modeling z_y directly as a multivariate Gaussian:

$$z_y \sim \mathcal{N}(f(x), Z_\sigma)$$

where Z_σ could capture the noise properties of the latent space, possibly modeled by another neural network.

5.5.a Hyperparameter Search

To optimize the performance of our contrastive learning-based model for decoding brain activity, we conducted a hyperparameter search using a random sampling methodology. Our search focused on minimizing the validation loss across a set of 100 randomly sampled configurations from a predefined search space. The hyperparameter sweep was configured as follows:

- **Batch Size (BS)**: We experimented with batch sizes of {512, 1024, 2048}.
- **Learning Rate (lr)**: We explored two learning rates: {1e-4, 1e-5}.
- **Alpha (α)**: The weight parameter for the contrastive loss was sampled from {0.5, 0.8}.
- **Temperature (τ)**: We used a fixed value of 0.1 to control the sharpness of the similarity distribution.
- **Loss Function**: Three loss types were tested: {contrastive, mean squared error (MSE), mean contrastive}.
- **Weight Decay (wd)**: Regularization was applied with values {1e-4, 1e-5, 1e-2, 0}.
- **Latent Dimension**: The dimensionality of the shared latent space was varied across {512, 1024, 2048, 4096, 8192, 10000, 16384}.
- **Activation Function**: We tested both {ReLU, Identity} activation functions in the hidden layers.

- **Base Channel Size**: We considered different channel size configurations across layers: {[4096, 2048, 1024], [2048, 1024], [2048]}.
- **Hidden Dimensions**: The hidden layers were configured with varying sizes, including {[2048, 1024, 512], [1024, 512], [1024]}.

The metric used to evaluate the model's performance was the validation loss, which we aimed to minimize. The random sampling method allowed us to explore a diverse set of hyperparameter combinations without performing an exhaustive grid search, which would be computationally expensive.

By systematically varying these key hyperparameters and evaluating each sampled configuration, we were able to identify an optimal combination that balanced both accuracy and generalization across subjects. This hyperparameter search played a crucial role in achieving the strong performance metrics reported in our results.

Accordingly to insights given from comparative analysis we found that linear models with high common and latent dim (equal to the input dimensionality) performed better.

5.5.b Bias, Privacy, and Ethical Considerations

The development of models that decode brain activity into language raises important ethical concerns, particularly regarding bias, privacy, and the responsible use of such technology. One significant concern is the potential for bias in both the models and the data. fMRI datasets, as well as the pre-trained language models used in this study, can inherit biases from the populations they are trained on, which may lead to biased or inaccurate decoding, especially across diverse groups of individuals. This could have serious implications when applying these models in clinical or social contexts.

Another critical issue is privacy. Brain-to-text decoding systems pose a unique risk to neural privacy, as they may allow for the reconstruction of internal thoughts and mental states. This raises questions about consent, data security, and the misuse of brain data in contexts where individuals may not have full control over how their neural activity is used or interpreted. It is essential to develop safeguards to ensure that brain data cannot be used to decode private thoughts without explicit consent, and to explore ways to mitigate any unintended consequences of decoding technologies, such as the risk of surveillance or the exploitation of individuals' cognitive data.

As the field progresses, it will be crucial to establish ethical guidelines that prioritize transparency, fairness, and respect for individual autonomy. Future research should also focus on developing methods to disentangle model biases

from genuine cognitive processes and explore the concept of "neural privacy" as a framework for protecting individuals in this emerging area of brain-computer interface technology.

6 Direct decoding

In this Chapter[1], we introduce BrainLLama, a novel language decoding framework leveraging multimodal large language models (LLMs) to decode semantic content from non-invasive neural signals acquired via functional MRI. Our approach integrates an encoding model to map brain activity to language embeddings extracted from LLama3-8B model, data augmentation using synthetic brain patterns, and a multimodal language backbone adapted to reconstruct text from these embeddings. This system generates multiple candidate sequences and refines output by selecting the best alignment with brain data, thus optimizing accuracy and computational efficiency. Experiments with three subjects demonstrate that BrainLLama captures meaningful neural representations and reconstructs coherent text in a fraction of time needed by previous approaches, advancing the capabilities of non-invasive brain-computer interfaces for applications in assistive communication and neurorehabilitation.

6.1 Introduction

Language is one of the most natural forms of communication and is fundamental to human interaction, information exchange, and cognitive processing [202]. The neural basis of language comprehension, generation, and processing has long been a focus of neuroscientific research, with the ultimate goal of deciphering how the brain manages these complex tasks [83, 123, 122, 36, 35, 33, 254, 171]. A detailed understanding of these processes is crucial not only for theoretical advancements but also for practical applications, such as assisting individuals with language impairments. In particular, Brain-Computer Interfaces (BCIs) fall

[1] Full manuscript in preparation for submission.

under the field of "brain decoding", a family of methods that can reconstruct or identify stimuli from neural activity. Decoding language represent a promising avenue to better understand his processing in the brain and toward the development of assistive communication technologies. Recent technological and methodological advances are opening new opportunities for non-invasive BCIs capable of interpreting language-related brain signals with increasing precision [206, 241].

There are numerous ways in which we interact with the world via the langauge medium:: reading, listening, thinking, and speaking, often also combinations of these activities within conversations between people.

This variety complicates brain decoding, as each mode of language engagement activates unique and overlapping brain regions. Invasive neural recording methods, such as electrocorticography (ECoG), have achieved impressive decoding results, reaching accuracies that approach natural spoken rates of up to 60 words per minute [106, 206, 273]. However, non-invasive systems, though less precise, offer safer, more widely applicable alternatives. In this context, recent advancements in deep learning, combined with the availability of richer datasets with an extended range of stimuli per subject, are reshaping the landscape of non-invasive language decoding research [251].

High-temporal-resolution techniques like EEG and MEG have been explored extensively to decode words from brain signals, albeit with mixed results. MEG, for instance, has shown promising correlations with language processing, yet similar findings for EEG remain debateable[53, 67, 125, 280]. Scaling emerges as a critical factor, as initial evidences suggest that achieving above chance performances on language decoding with EEG may require datasets containing 100+ hours of subject data, hinting at a logarithmic scaling law [230]. Recently, functional MRI has emerged as a powerful tool for non-invasively capturing language-related brain activity, offering spatially resolved insights into brain signatures of language processing. Studies have shown that encoding models, which map language embeddings derived from large language models onto brain activity, can be trained to achieve high correlations on held-out data in particular brain regions, including the temporal and frontal lobes. These models have unveiled key neuroscientific insights, such as semantic maps across cortical surfaces, hierarchical processing of language, and mechanisms of language comprehension and contextualization [122, 199] Furthermore, scaling laws in these encoding models indicate that increasing the size and quality of language embeddings enhances prediction in a logarithmically way, measured as correlation between predicted and ground truth held-out brain activity[7]. A recent Bayesian approach to brain-to-text decoding introduced a factorized framework

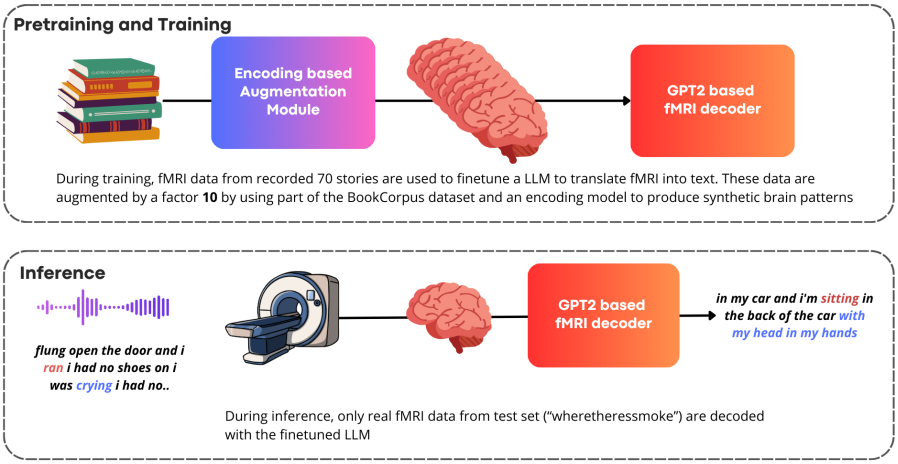


Figure 6.1: Overview of the pretraining, training, and inference workflow for direct language decoding from fMRI data using a GPT2-based model. In the pretraining and training phase, fMRI data from 70 recorded stories is used to fine-tune a large language model (LLM) to translate brain activity into text. This data is augmented tenfold through synthetic brain patterns generated from part of the BookCorpus dataset [285] using an encoding-based augmentation module. In the inference phase, real fMRI data from a test story (e.g., "wheretheressmoke") is processed by the fine-tuned LLM to reconstruct text corresponding to the participant's brain activity, with example output shown in bold and colored text. This process demonstrates the model's ability to decode semantic content directly from fMRI recordings.

for language interpretation from brain signals. This method generates candidate text sequences, project them into brain activity with an encoding model and compare them against corresponding target brain activity, selecting those sentences that best match the intended neural states. However, the need to produce and refine a large pool of candidate sequences for narrative decoding makes it computationally intensive, thereby limiting its suitability for real-time applications..[251].

In this study, we introduce *BrainLLama*, a novel decoding framework that advances non-invasive language decoding from brain activity. Our key contributions include:

1. **LLM-Based Encoding Model:** We leverage sentence embeddings from the large language model Llama3-8B to train an encoding model that predicts brain activity from language embeddings. These embeddings are downsampled to match the fMRI sampling rate using a Lanczos filter, and a measure of correlation between actual and predicted brain activity on held-out data can be used to identify relevant brain regions for language processing and supporting data augmentation through synthetic brain signals. Moreover the same encoding model can be used to produce synthetic brain activity patterns from external text sources to provide data augmentations.

2. **Multimodal Language Backbone:** We fine-tune a LLM to accept downsampled LLama embeddings as input and reconstruct the corresponding text as output with Low-Rank adaptation to minimize the number of learnable parameters. This step creates a multimodal, autoencoder-like structure that efficiently captures and retains semantic information derived from the language embeddings.

3. **Direct Brain-to-Embedding Mapping:** We implement a linear mapping from brain data to downsampled LLama embeddings, allowing for direct estimation of linguistic representations from fMRI measurements. These estimated embeddings are then passed to the language model as input to generate reconstructed sentences.

4. **Candidate Selection for Robust Decoding:** To mitigate the inherent noise in fMRI signals and ensure coherent text reconstruction, we generate multiple candidate text sequences ($N = 10$) per time point. A secondary sentence encoding model is then used to select the best candidate sequence by minimizing the mean squared distance between the estimated and target brain activities.

By integrating these components, *BrainLLama* capitalizes on the spatial specificity of fMRI measurements and the representational power of LLMs, thereby enabling direc and efficient brain-to-text decoding. This decoding task faces multiple challenges due to the inherent noise in fMRI data and the limited number of

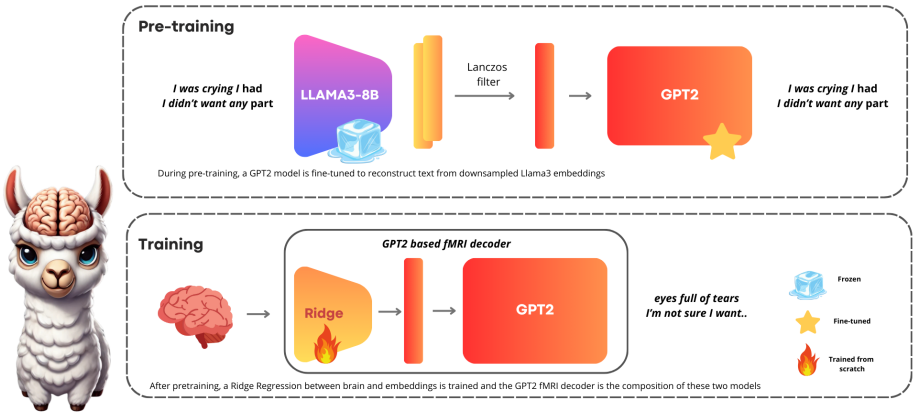


Figure 6.2: Illustration of the pre-training and training stages for the GPT2-based fMRI decoder. In the pre-training phase, the GPT2 model is fine-tuned to reconstruct text from downsampled embeddings produced by the LLAMA3-8B model. The Lanczos filter is applied to reduce embedding dimensionality, preparing the data for GPT2-based reconstruction of linguistic content (example output shown in italics). In the training phase, a Ridge Regression model is introduced to map brain (fMRI) data to the LLama embeddings. The final GPT2-based fMRI decoder combines the Ridge Regression model with the fine-tuned GPT2, enabling it to decode semantic content from fMRI signals. Icons indicate stages of the model: frozen, fine-tuned, and trained from scratch.

data samples obtainable from such experiments. To address these issues, we employ data augmentation strategies and apply regularization methods—such as LoRA during LLM training, ridge regression for the brain-to-embedding mapping, and a modular architecture separating the brain to embedding and the embedding to sentence stages—to reduce the risk of overfitting.

Our model produces candidate text sequences that are already guided by neural signals, and then ranks them by comparing their brain-projected embeddings against the target brain activity. In doing so, *BrainLLama* provides a scalable solution for reconstructing coherent and meaningful text directly from fMRI data in a time that could be compatible with real time feedback experiments.

By advancing the capabilities of non-invasive BCIs, our method opens new pathways for understanding and decoding language representation in the human brain.

6.2 Method

In this section, we describe each step of our pipeline, covering data preparation, data augmentation, and the two-stage training process. Our approach begins by adapting a large language model (LLM) to accept embeddings extracted from the LLama model (downsampled in time with a lanczos filter to match fMRI sampling rate), enabling it to decode text from these embeddings. Following this pretraining phase, we train a linear model to predict LLama embeddings directly from brain activity. During inference, the brain decoding model and language model are combined, forming a pipeline that converts fMRI data directly into text. See Fig 6.2 for a scheme of pretraining and training stage of our LLM decoder and Fig 6.1 to have a scheme of training and inference phases.

Due to the inherent noise in fMRI data and variability in text generation from the LLM, we implement a candidate selection process to improve decoding accuracy. For each fMRI sample, multiple candidate text sequences are generated using beam search. These candidates are then scored using a sentence encoder (CLAP Text Model [70] followed by a Ridge Regression), which maps each candidate back to brain activity and calculates the Euclidean distance to the original brain data. The sequence with the lowest distance is selected as the final decoded text, also providing a measure of alignment between generated language and brain activity.

All experiments and models were trained on a server equipped with eight NVIDIA A100 GPU cards, (80GB RAM each connected through NVLINK) 256 GPU threads, and 2 TB of system RAM.

6.2.a Data

For our analysis, we used the publicly available dataset from [148], composed by subjects S1, S2, and S3. This kind of data, with few subjects and lot of stimuli per subject fall under the umbrella of recent "deep" or "intensive" fMRI framework, where the focus shifts from wide paradigm where many subjects are investigated with few common stimuli, to few subject that are exposed to a large number of different stimuli [144].

Each subject participated in approximately 16 hours of fMRI recordings while listening to a collection of 83 stories from The Moth and Modern Love podcasts. The fMRI data was acquired on a 3T Siemens Skyra scanner, with a repetition time (TR) of 2.0 seconds and an isotropic voxel size of 2.6 mm. Preprocessing included motion correction, cross-run alignment, standardization, and detrending of low-frequency signals. For further details, please refer to the original paper [148].

In our study, the first 70 stories heard by each subject were designated as the

training set, while the remaining 12 stories comprised the validation set. Additionally, the story titled "wheretheressmoke" was repeated 10 times to enhance signal-to-noise ratio in the test set by averaging responses to the 10 different listens of the story, in alignment with recent research on language encoding and decoding [251, 7].

6.2.b Encoding

The initial phase of our pipeline focuses on reducing the complexity of the brain signal to target the regions most relevant to language processing. While language and semantic tasks engage widespread brain regions [123, 122], specific cortical areas play a more central role in language processing. Thus, we began by identifying the cortical regions at the voxel level that are more readily modeled by language models.

We developed an encoding model of brain activity based on a large language model (LLM) as its core architecture. For each word in the training narratives, embeddings were extracted from the final layer of LLama3-8B [68], using a context window spanning the five preceding words.

In spoken language, the temporal resolution of word occurrences significantly exceeds the sampling frequency of fMRI data, necessitating synchronization of these modalities. To align the temporal resolution of the word embeddings with the fMRI data, we applied a Lanczos filter to downsample the embeddings, ensuring synchronized embeddings-fMRI pairs. We then trained a Ridge regression model with 5-fold cross-validation to predict brain activity from the concatenatied embeddings. The regularization parameter (α) was optimized over a logarithmic scale in base 10, ranging from 10^{-2} to 10^{5}, using the himalaya library [69]. The input to the model consisted of the concatenation of four consecutive downsampled LLama embeddings, resulting in a 16,384-dimensional input (4 embeddings $\times$ 4096 dimensions). This formulation represents a linear finite impulse response (FIR) model, which accounts for the temporal delay introduced by the hemodynamic response function.

Pearson correlations between predicted and ground truth brain activity were then calculated on held-out validation data, allowing us to select the top 10,000 cortical voxels with highest correlation as our target regions. These target voxels represent the neural activity we aim to decode. Fig 6.3 shows a scheme of our encoding pipeline.

6.2.c Data Augmentation

Training or fine-tuning transformer models, the core architecture for modern large language models (LLMs), demands vast amounts of data to reach optimal

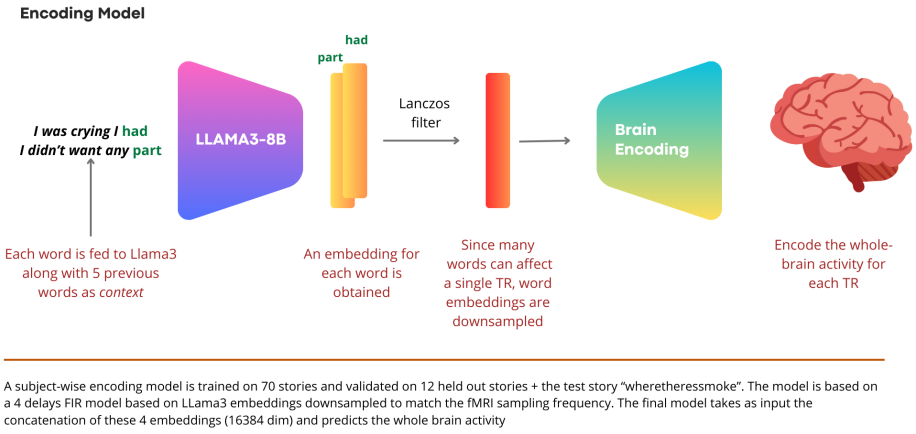


Figure 6.3: Overview of the Encoding Model for predicting brain activity from language. Each word is input into LLAMA3-8B, along with the previous five words as contextual input, to generate embeddings. The embeddings are downsampled via a Lanczos filter to align with the fMRI sampling frequency, allowing multiple words to influence a single temporal resolution (TR) of brain activity. For each TR, 4 previous related embeddings are concatenated resulting in a 16,384-dimensional representation used as input for a brain encoding linear model to predict whole-brain activity for each TR. The encoding model is trained on 70 stories and validated on 12 held-out stories.

performance. In recent neuroscience research, there has been a shift toward enhancing the availability of stimuli for individual subjects, emphasizing dataset depth over the number of participants [144]. This shift allows for a more refined focus on subject-specific neural data acquisition, enabling the construction of high-fidelity encoding and decoding models [148, 5, 38, 47]. Despite these efforts, the current data regime—comprising thousands of stimuli per subject—represents a significant advancement for neuroscience but remains modest in comparison to the data requirements of deep learning architectures, which are notoriously data-intensive.

To overcome this issue, we implemented data augmentation, a widely adopted technique in fields like computer vision, natural language processing, and reinforcement learning, yet relatively underutilized in machine learning applications involving neural data. Our approach represents a first effort to apply data augmentation within the context of neural encoding. We employed our encoding model alongside the external BookCorpus dataset [286], leveraging it as an additional source of textual data.

To simulate word onset timings that would align with the original fMRI experiment structure, we sampled from the distribution of word onsets observed in the training stories, creating synthetic stories with similar temporal characteristics. We generated approximately 700 synthetic stories by extracting text segments comparable in length to the original training stories, thus expanding the dataset by a factor of 10. The augmented dataset was processed using the same embedding extraction approach described in Section 6.2.b, enabling the model to generalize more effectively to variations in language and enhancing its robustness in encoding and decoding tasks.

This augmentation strategy not only increased the data available for model training but also maintained alignment with the temporal dynamics of the original fMRI recordings. See fig 6.4 for a scheme of the augmentation procedure.

6.2.d Multimodal Language Model pretraining

The goal of this component in our pipeline is to decode downsampled language embeddings into coherent text sequences, initially pretraining it on training text embedding sequences extracted from LLama, and during inference replacing these embeddings with fMRI the ones estimated by fMRI data.

Since in our framework, each fMRI sample (time repetition - TR) is modelled by a set of four previous LLama embeddings that were downsampled to match the temporal resolution of the fMRI recordings, we need a model able to decode these 4 downsampled language embeddings into text. To reconstruct the original sentences from these embeddings, we implemented a Brain-Encoder-Text-

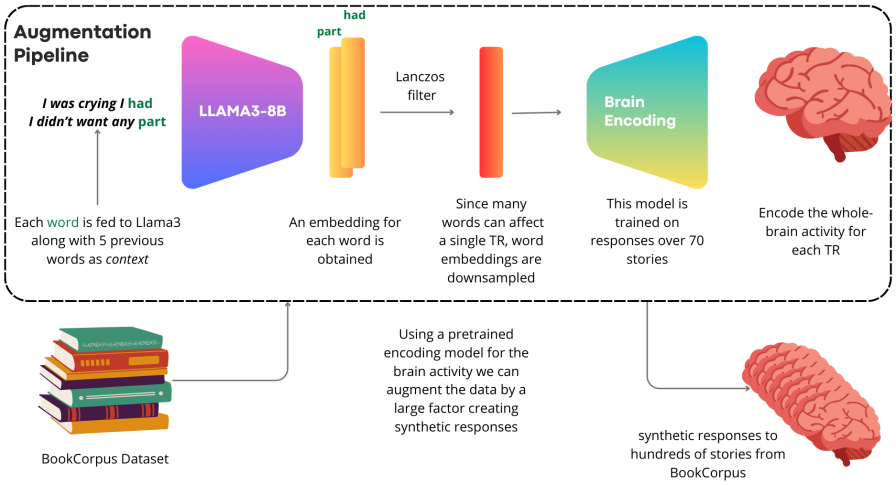


Figure 6.4: Data Augmentation Pipeline for Brain Encoding. Each word from a story is input into LLAMA3-8B along with the previous five words as context to generate embeddings. These embeddings are downsampled using a Lanczos filter to match the fMRI sampling frequency, allowing each temporal resolution (TR) of brain activity to be influenced by multiple words. The resulting embeddings are used to train a brain encoding model on responses to 70 stories. To augment the dataset, the pretrained encoding model generates synthetic brain responses for additional text samples from the BookCorpus dataset, creating a significantly larger corpus of synthetic brain activity responses to enrich the training set.

Decoder architecture, based on GPT-2 [208] or LLama3-8B as the text generation backbone. This approach allows the model to process text embeddings as input and generate corresponding text sequences.

The architecture is a multimodal version of the LLM that integrates both an embedding encoder and an autoregressive language decoder. The encoder was implemented as a linear layer with positional embeddings, trained to produce representations from LLama embeddings data that serve as input to the GPT-2 (or Llama based) decoder. The decoder was adapted to accept these embeddings via cross-attention layers, which were added to facilitate the integration of embedding-derived features into the text generation process. To prepare the embeddings for decoding, we implemented a token-shifting function that aligns the sequence by shifting tokens four positions to the right.

Given the complexity of finetuning an LLM and the need for robust generalization, we employed LoRA (Low-Rank Adaptation) [118] to fine-tune the model with parameter-efficient adaptation. LoRA optimizes the training process by introducing low-rank modifications to only a subset of parameters, which minimizes the risk of overfitting and ensures that the model remains well-suited to processing neural embeddings without overly modifying the pre-trained language model.

To enhance training data diversity, we applied data augmentation using synthetic text sequences similar in structure and length to the original dataset. This augmented data, combined with a random search over 100 hyperparameter configurations, allowed us to test various combinations of learning rate (values tested $1e{-}5, 3e{-}5, 1e{-}4$), weight decay (values tested $1{-}5, 1e{-}4, 1e{-}3, 1e{-}2, 0$), batch size, whether or not using Lora and Lora scaling factors ($r = 8, 16, 32, 64$) and data agumentations. The training objective was to minimize cross-entropy loss between the generated text and actual target sentences, using supervised learning to optimize the model for accurate brain-to-text decoding. Each model was trained using the Adam optimizer up to 20 epochs with early stopping (patience=3) monitoring the validation loss to stop training once signals of overfitting are detected.

6.2.e Brain decoding model

To accurately decode text from brain activity using a fine-tuned LLM capable of reconstructing text from downsampled language embeddings, an fMRI-to-language embedding converter is needed. Based on pilot experiments and prior research, directly fine-tuning an LLM with fMRI data often leads to overfitting. Additionally, studies indicate that language embeddings can be more effectively decoded using linear or nonlinear models [76]. Consequently, we

adopted a two-stage approach: we used Ridge regression to predict each of the four downsampled LLama embeddings from brain activity. The regularization hyperparameter α was optimized within a log-scaled range from 10^{-2} to 10^{6}, with performance evaluated by calculating the Pearson correlation between the predicted embeddings and ground truth embeddings on validation data. Brain-predicted language embeddings are fed to the modified LLM during inference to perform language decoding from fMRI data.

6.2.f Sentence Scoring model

To select the most accurate text representation from multiple candidate sentences generated during brain-to-text decoding, we implemented a Sentence Scoring Model that assesses each candidate by comparing predicted neural responses to actual brain activity, ranking candidates based on alignment with the neural data. This scoring model employs CLAP (Contrastive Language-Audio Pretraining) [71] embeddings to map each candidate sentence to a neural representation in brain space. We initialized the scoring model with a CLAP-based sentence encoder, which translates sentences into feature embeddings projected to brain activity space using a Ridge regression model. This projection is trained to predict brain responses from candidate sentence embeddings by minimizing the difference between predicted and actual neural activity. First, candidate sentences are tokenized and encoded into embeddings using the CLAPText model, leveraging language embeddings to capture contextual information. Each sentence is converted to an embedding vector representing the predicted brain response to that sentence. Then, a Ridge model, with an optimized regularization parameter α selected via cross-validation on the training set, predicts brain activity from these CLAP embeddings. For each batch of candidate sentences, the Ridge model estimates corresponding brain activity responses, producing a set of predicted neural patterns. Finally, to evaluate how well each candidate aligns with the actual fMRI response, we calculate the Euclidean distance between the predicted neural response for each sentence and the actual brain activity recorded in the experiment. For each candidate set, sentences are ranked by their Euclidean distance scores, with the closest match selected as the final decoded text. This approach ensures that selected sentences closely reflect the original neural representation, reducing variability and improving decoding accuracy. By integrating a CLAP-based sentence encoder with a Ridge regression to brain activity, this model effectively identifies sentence candidates that best represent the subject's neural responses, refining the decoding pipeline.

6.2.g Evaluation

Each model in our pipeline was trained on a dedicated training set and optimized using a specific objective on the validation set. For the final assessment of the entire pipeline, we reserved the test story, *"wheretheressmoke"*. The pretraining of the multimodal adaptation of the LLM was evaluated by monitoring the cross-entropy loss on validation stories, optimizing the model for next-word prediction. During optimization of the GPT2 decoder, we observed that using LoRA and data augmentation significantly improved the final validation loss, with a 45% reduction (from 2.59 to 1.45). Consequently, we applied the same configuration to the larger LLama model to reduce computational demands for this research.

The brain-to-embedding model was trained using a regularized least squares loss function, with the regularization parameter α optimized by maximizing the Pearson correlation between the actual and predicted embeddings on the validation set.

Similarly, the sentence scoring model was trained using Ridge regression to map CLAP sentence embeddings to fMRI data. The parameter α was optimized based on correlation with validation data, achieving an average correlation of 0.256 across subjects. To assess the quality of reconstructed text, we employed a range of standard language metrics. For each fMRI sample, the best-matching candidate sentence was selected by measuring the distance in brain space between candidate sequences and actual brain data. The top-ranked sentences were evaluated using Word Error Rate (WER), BLEU-1, METEOR, ROUGE-1, ROUGE-L, and BERT Score, providing a comprehensive assessment of linguistic accuracy.

The evaluation framework was designed to capture both semantic meaning and syntactic fidelity. Identification accuracy was first computed to evaluate how well each candidate sentence reflected the neural data. WER was used to measure word-level accuracy, while BLEU-1 and METEOR captured semantic overlap. ROUGE-1 and ROUGE-L assessed n-gram overlap and longest common subsequences, offering a robust measure of textual similarity. Additionally, BERT Score was calculated using embeddings from a Sentence Transformer model (all-MiniLM-L6-v2) to quantify semantic similarity, comparing candidate and ground truth embeddings based on Recall or F1 scores.

Finally, we assessed identification accuracy by predicting brain activity using the sentence encoder and calculating the frequency with which the synthetic brain activity embeddings correlated more strongly with the ground truth than with random, unrelated patterns. This pairwise-based metric (baseline is 50%) is widely recognized in the decoding literature.

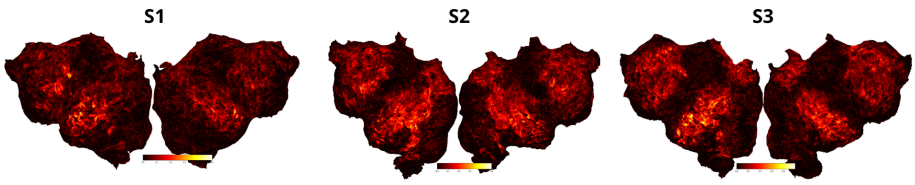


Figure 6.5: Correlation maps for three subjects (S1, S2, S3), showing the correlation between predicted brain activity from the encoding model and actual recorded brain activity on validation stories. Warmer colors indicate higher correlation values, demonstrating regions where the model's predictions align closely with true brain responses. Consistent patterns across subjects reflect the model's robustness in capturing neural encoding, while variations highlight individual differences in brain representation.

6.3 Results

6.3.a Encoding results

To evaluate the effectiveness of our encoding model, we analyzed the correlation between the predicted and ground truth brain activity on the validation data for each subject. Figure 6.5 presents heatmaps of the cortical surface for each subject (S1, S2, and S3), illustrating the correlation values across the cortex. The color bar indicates correlation strength, with warmer colors representing higher correlation values and colder colors indicating lower correlations.

As shown in the figure, our model achieves substantial correlations in specific cortical regions across all three subjects. These high-correlation areas are primarily located in regions typically associated with language processing, such as the superior temporal and frontal regions, consistent with known neural correlates of language comprehension and production. The patterns of activation also suggest subject-specific variability in the distribution and intensity of correlations, likely reflecting individual differences in neural responses to language stimuli. These results are in line with previous research on encoding models [7, 251] showing good encoding performances from LLama embeddings, observing a correlation greater than 0.1 for all the top-10000 voxels selected for subsequent analyses. The consistent presence of high-correlation areas across subjects underscores the model's ability to capture meaningful language-related neural patterns, while subject-specific differences highlight the model's adaptability to individual brain structures. This validation demonstrates that our encoding model can effectively map language embeddings to brain activity, providing a solid foundation for subsequent decoding and text generation tasks in our pipeline.

6.3.b Brain to text performances

Subject	Decoder	Id Accuracy	BERT Score	WER	BLEU-1	METEOR	ROUGE-1	ROUGE-L	Decoding Time (s)
S1	GPT	0.692	0.816	0.955	0.136	0.110	0.184	0.141	16
S2	GPT	0.752	0.819	0.956	0.137	0.106	0.183	0.138	16
S3	GPT	0.772	0.818	0.959	0.141	0.115	0.186	0.142	16
S1	LLama	0.641	0.825	2.090	0.109	0.110	0.141	0.111	667
S2	LLama	0.691	0.824	1.986	0.116	0.120	0.150	0.120	667
S3	LLama	0.693	0.825	2.038	0.112	0.116	0.146	0.117	667
S1	Tang et al. 2023	-	0.8077	0.9407	0.233	0.162	-	-	5100
S2	Tang et al. 2023	-	0.810	0.935	0.243	0.168	-	-	5200
S3	Tang et al. 2023	-	0.812	0.924	0.247	0.170	-	-	5025

Table 6.1: Performance Metrics for Brain-to-Text Decoding Pipeline Across Subjects

Table 6.1 presents a comparative analysis of the performance metrics for the brain-to-text decoding pipeline, evaluated across three subjects (S1, S2, and S3) and using three different decoders: GPT, LLama, and Tang et al. (2023). Each decoder exhibits distinct strengths, highlighting varying levels of effectiveness in capturing subject-specific neural patterns and linguistic fidelity.

The Identification Accuracy metric, applicable only to GPT and LLama decoders, reveals that GPT achieves the highest performance for subject S3 (0.772), followed by S2 (0.752), and S1 (0.692). LLama, on the other hand, shows lower Identification Accuracy scores, with the highest being 0.693 for S3. Given the pairwise nature of this metric with a chance levle of 50% these results suggest that both GPT and LLama-based decoders are able to decode meaningful semantic information.

The BERT Score, which assesses semantic similarity between reconstructed and target sentences, is highest with the LLama decoder across all subjects (0.825 for S1 and S3, 0.824 for S2). While GPT also maintains high BERT Scores, these results indicate LLama's relative strength in capturing the core meaning of language from brain activity. The Tang et al. (2023) decoder achieves comparable BERT Scores, indicating strong semantic alignment, though Identification Accuracy and ROUGE metrics were not reported for this decoder.

Word Error Rate (WER) reflects the lexical accuracy of each model. The Tang et al. (2023) decoder excels in this metric, with the lowest WER values across all subjects (0.940 for S1, 0.935 for S2, and 0.924 for S3), indicating its precision in word-level matching. GPT exhibits moderate WER scores close to 0.95 across subjects, whereas LLama shows significantly higher WER values (2.09 for S1), which may point to challenges in maintaining lexical precision at the word level for this decoder.

BLEU-1, METEOR, ROUGE-1, and ROUGE-L metrics provide insights into syntactic and lexical fidelity. Tang et al. (2023) achieve the highest BLEU-1 scores (0.247 for S3) and superior METEOR scores, suggesting a strong match in word choice and phrase structure with the ground-truth sentences. The GPT decoder performs well with S3 (BLEU-1 of 0.141 and ROUGE-1 of 0.186), highlighting its ability to maintain syntactic alignment. However, LLama demonstrates comparatively lower scores on these metrics, further indicating that it may struggle with exact lexical and syntactic fidelity.

In terms of decoding Time, the GPT decoder completes processing consistently in 16 seconds, making it the most efficient. LLama, however, requires 667 seconds, and the Tang et al. (2023) decoder has the highest time cost, taking over 5000 seconds per subject. The differences in time reflect the varying computational demands of each model, with potential trade-offs between performance and processing speed.

In summary, the results demonstrate that while all decoders can capture meaningful semantic content from brain activity, each has unique strengths. The GPT decoder is efficient and achieves high Identification Accuracy, particularly with subject S3. LLama excels in semantic similarity but has challenges with lexical accuracy and decoding speed. Tang et al. (2023) offers strong lexical fidelity but is the most computationally intensive.

The qualitative analysis of reconstructed samples shown in Table 6.2, 6.3 and 6.4 highlights both strengths and limitations in capturing the semantic content and syntactic structure of the ground truth sentences. Reconstructed sentences often reflect the general theme or emotional tone of the original text, preserving elements of personal interaction, expressions of agreement or disagreement, and questions, even when specific words differ. However, deviations in vocabulary and sentence structure frequently arise, altering the subtlety or intended nuance of the ground truth. Lexical choices tend to diverge, with words and phrases substituted for similar but non-identical terms, impacting coherence and sometimes shifting the tone. Additionally, pronoun and perspective usage are not always consistent, which can disrupt the flow, especially in multi-participant dialogues where the clarity of speaker roles is essential. Although high-predictive words,

highlighted in green, tend to align well with the ground truth, partially matched words, shown in purple, reflect an attempt to maintain semantic closeness, albeit with vocabulary variation. Overall, the reconstructed outputs display a degree of language comprehension and contextual awareness but require further refinement to achieve precise alignment with the original content and structure of the ground truth sentences.

6.4 Discussion

In this study, we have developed and evaluated a novel approach for direct language decoding from fMRI data, advancing the field of non-invasive brain-computer interfaces (BCIs), showing that it is possible to decode semantic language information from brain activity with a reduction in time of two orders of magnitude compared with previous approaches. By leveraging large language models, we successfully reconstructed meaningful phrases and sentences from brain activity with a qualitative focus on semantic coherence and fidelity. By applying augmentation techniques typically reserved for other machine learning domains, we aimed to explore new possibilities for improving model performance in neuroscience applications, potentially setting a precedent for more widespread adoption of data augmentation in neural data modeling. Our approach demonstrates a step forward in decoding language directly from neural representations, offering promising applications in neurorehabilitation, assistive technologies, and advancing our fundamental understanding of brain-language relationships.

Direct decoding of language from brain activity is crucial for several reasons. First, it offers a faster, more intuitive approach than bayesian methods [251]. By bypassing large unoriented beam searches, direct decoding enables a faster translation of thought into communicable language. Second, direct decoding allows for richer semantic modeling, capturing not just phonetic or syntactic information but also deeper semantic constructs encoded within neural activity.

Our work builds upon previous studies in both invasive and non-invasive brain-language decoding. Invasive methods, such as those involving electrocorticography (ECoG), have made considerable strides in capturing high-resolution language-related signals, which facilitate accurate decoding of speech from motor cortex activity. However, these methods are limited in clinical application due to their surgical requirements, making them less accessible for broader populations. Non-invasive approaches, including those using fMRI and EEG, have emerged as viable alternatives, with studies by Tang et al. [251] demonstrating the feasibility of reconstructing coherent semantic content from fMRI data and

Ground Truth	Baseline	Reconstructed (S1)	Reconstructed (S2)	Reconstructed (S3)
out right when i needed to and she says well why don't you come back to my house and i'll give you a ride i say ok great and	i just jumped out right when i needed to because i didn't even have my driver's license yet	and i said okay we can load her into the car and i said go check out the house	her to come on and say hi i took my cue and went to say hi she retrieved my purse and	she was ready to pull out but i told my dad to go ahead and i spent the next hour with him
the man says again you alright and she says yeah i'm just gonna go out and smoke with her and so we go outside and	and the woman then kisses the man and says the same man again you say okay and i'm gonna	she looked up at me and he said well i m not sure she d like this so much he said i m just going to have to go back	and she said hey i m going to help you he said	come over and say thank you he asked her to stay home and i said well we
you and she says never mind i'm back and he says you alright and she says yeah i'm alright and then she turns to me and	he said you and me were he said congratulations and she said yeah i'm never back then and he	asked her son for a moment i told her you never said yes i m so sorry he said i didn t want to say anything to him but	he says she s good friend and i said hey mommy he said and you know i m just	sorry she said again he said you
the fuck is that and she pulls me over and he sees me and he says oh hey i'm not a threat	guy who says who the fk is this and she pulls me over and sees me and he says	she said i ll see you and he said alright so i look up and he says	he said and she just looked up at him and i said hey you can t see him	asked her friend to look at him and he said hey she said i m not sure

Table 6.2: Examples of decoded text from fMRI using GPT2 as decoder

Ground Truth	Baseline	Reconstructed (S1)	Reconstructed (S2)	Reconstructed (S3)
i'm back and he says you alright and she says yeah i'm alright and then she turns to me and says you want a beer and he says who the fuck	and she says you want me to go back and he said you expect	dad she said and he said i m sorry but i said you never want to say anything about it and he was like you re going to have to	asked her to come on and he s like mommy you know and i said hey she came over and he	good she said again he said i m ready to see him
he says who the fuck is that and she pulls me over and he sees me and he says oh hey i'm not a	and i pull her over and he sees me and says who the fk is this guy and she	hear it again and she stepped forward to get his attention	her face and he said hey and she called him out and he looked at	and i said hey she asked him to look after me he said hey
my seatbelt and when his foot hit the brake at the red light i flung open the door and i ran i	she was nt sure what to do but she	to my right the photographer i had just met broke into a grin when i looked up the photographer s name and i was	to the car and pull out of her doorway i crowed when my friend took a right	and i swung open the door at her laughing i pulled out my cell phone and flew over to the edge of my seat jake murmured and i
and he says you alright and she says yeah i'm alright and then she turns to me and says you want a beer and he says who the fuck is that and	and says who the fk is you saying he asks the bartender who says you want a drink and	he s not going to say thank you and she said i m so glad to see you he said no and i was like oh my god he s	she said hey he asked her to come on and he s like i don t have a daughter	he said she s ready to see him he said look up and she said come on he asked her

Table 6.3: Examples of decoded text from fMRI using GPT2 as decoder

Ground Truth	Baseline	Reconstructed (S1)	Reconstructed (S2)	Reconstructed (S3)
a cabana and he left and i had my cigarettes and uh i started to walk in this beautiful neighborhood	and he had his cigarettes and i had uh i went to this place and collapsed and he left his	was in the car and um i remember this guy	took a walk downstairs to get rid of her truck i looked around but he	hallway and i got ready to go home afternoon i stopped at the
flung open the door and i ran i had no shoes on i was crying i had no	had no idea i was crying on stage i had no money	i was walking down the hall and i found my father sitting in the driveway and i wrote my name to him and he said	the door she did nt want to explain but when she looked down at her suitcase i had to say something	to the car and i said oh my god i m not wearing a car anymore and i looked up at him and i was like well i don t
he says where were you and she says never mind i'm back and he says you alright	and i hear the man say where were you and he says you re back	i say hey mommy he says and i m like you know i never said that	he said and she ll call me sono he said and you know i m just going to tell him	come on he said i m not going to tell him but she said i was going to
saying what was the fight about and i say wha what are they all about and she said i know what you mean she said was it a bad	what i mean she said what are they thinking about the fight was what the question was	that she said i don t know what he said but i m not sure why you should have	i said you know what is he supposed to say i m not sure what he said	what she said you re not sure what he said i was

Table 6.4: Examples of decoded text from fMRI using GPT2 as decoder

Antonello et al. [7, 8] exploring the bounds of information encoding capacity in the brain. Our research builds on these foundations by employing direct decoding techniques that harness recent advances in LLMs, offering a scalable, non-invasive solution with impressive semantic modeling capabilities.

Despite these achievements, our study also has several limitations. The temporal resolution of fMRI is comparatively low, potentially restricting the accuracy of dynamic language decoding. Additionally, individual variability in brain activation patterns poses challenges in generalizing our models across subjects, a limitation that could be mitigated by further advances in cross-subject alignment methods [80]. Our qualitative analysis also points to occasional ambiguities in reconstructed language, suggesting that model refinement is necessary to improve specificity and reduce noise in decoded output. Future research should explore richer use of data augmentation and refined techniques to decode language embeddings from brain activity to improve both temporal and semantical fidelity in language decoding. Further, enhancing cross-subject and cross-language decoding models would expand the generalizability and applicability of these methods to diverse populations.

Finally, our work raises important ethical considerations. Direct decoding of language from brain activity coudl enable access to an individuals' private thoughts and intentions, raising concerns regarding privacy and consent. While our current work focuses on voluntary participation and controlled contexts, it is essential to consider robust safeguards for data privacy, informed consent, and secure data handling as this technology progresses. Responsible development will also require establishing clear guidelines to prevent misuse, ensuring that these advancements in brain decoding serve to empower individuals rather than infringe on their autonomy. By maintaining a focus on ethical considerations [90, 282], we aim to guide this field toward beneficial applications while mitigating potential risks associated with direct language decoding.

6.5 Conclusions

In conclusion, our work advances direct language decoding from non-invasive fMRI using large language models to achieve rapid, contextually accurate language reconstruction from neural activity. By removing motor-based translation steps, this approach offers promising improvements in speed and semantic fidelity, enhancing non-invasive brain-computer interface (BCI) capabilities.

Compared to prior methods, our model shows potential for high-quality decoding while remaining accessible and non-invasive. Challenges remain in improving temporal resolution, cross-subject adaptability, and accuracy. Future

work should explore multi-modal integrations and refine models for broader application.

Appendix

Ground Truth	Baseline	Reconstructed (S1)	Reconstructed (S2)	Reconstructed (S3)
about and she said i know what you mean she said was it a bad one and and i said you know like medium	what was the argument about i do nt know she said who do you think she was talking about sorry i said that was a bad question she did	myself no you should nt i said knowing that no matter what i did she was right i should have been there to protect her to	he said what do you mean she said i know that much but he said it s obvious you do nt know what i mean helll-laaaaaaaaaaos	said something wrong he said like you know what i mean no she said except that i know something about you that youllll
yeah i'm alright and then she turns to me and says you want a beer and he says who the fuck	back now he said you re okay and she says yeah i m fine and then she turns to me and says you want to go home	said to her dad hey baby no she said i know but i figured you do nt want me to it i said and was my life thank for that	her he asked and she laughed no i m fine she said and i know you should nt have a drink she told him	you okay she asked and i said no okay i answered because i wanted to tell herlllllll ll l a i uuauu
what are they all about and she said i know what you mean she said was it a bad	he said the question was what was the fight about i do nt know she said and what do you think they were fighting about he asked	her words because she was saying no i mean not that i wanted her to lol i u o oh she	look at her she said i know what you mean no i said it was hard to know if she was	him what she said oh he thought to himself well i did nt know that she was a virgin he said to m o u a f s t i
alright and then she turns to me and says you want a beer and he says who the fuck is that and	he says you re okay and she says yeah i do and then he turns to me and says who the hell wants a beer he says and i	her he said but then i should go she said and turned to jake well i guess you should be off he replied and and then he i you me	he asked and then she looked at me i m sorry babe he said and i know you should nt havellll	to talk to me so you okay i said and i want to talk with you too dad i replied knowing that he was trying to get me to

Table 6.5: Examples of decoded text from fMRI using LLama as decoder

Ground Truth	Baseline	Reconstructed (S1)	Reconstructed (S2)	Reconstructed (S3)
mean she said was it a bad one and and i said you know like medium she said oh	what do you think she said i know what she was saying and it was a bad choice i said sternly you mean like you said	you say no i mean it dad i know you were saying no good because i do nt want to know the rest lol haha ah i you me we	said i mean what do you know about him i said that was very obvious do nt you think she said i u a an f r t w n	asked i thought you knew something like that eh no i said thinking that he meant something else well i did nt know that you were going to ü laa
can tell her the really ugly stuff and she still understands how	to tell them anything at all i can tell this young woman the ugly truth and she can help me figure out what to do about it	be around others that means she s not his problem though i watched her shake her head and the door i my and my heart it was a no not	him and the things she d shared with him but the fact that they were wrong she and was but u o oo la a s m	knowing the things they shared with her that i would tell them how much i hated knowing their secrets the next day
says you alright and she says yeah i'm alright and then she turns to me and	he says and i m back sorry he says you re okay and she nods glancing down at her hands as if they re on fire	her i said come on babe she said no i mean i m llllll olll ooooooooooo i uuuu	her dad please she says to me nope she said she s fine so you you can u i my your our their and for s only	and he says please go she says nope i m good and she and i my her she said it again the of he said and he

Table 6.6: Examples of decoded text from fMRI using LLama as decoder

Ground Truth	Baseline	Reconstructed (S1)	Reconstructed (S2)	Reconstructed (S3)
and on the way up she kisses the little boy and then she kisses the man and the man says again you	out the remaining pack of cigarettes in his pocket the next thing she knew he was kissing the man and she kissed the back of his head	the men set up around her and i leave the room with his kisses and look at theå åüüt the door and utut uu	a few days later jake bought the women cookies and we headed for his room the girls sat on thegh and the bed	moved to purchase a second set of slippers and i moved toward the kitchen with his feet i called out and watched as he pulled the covers up to his
the man says again you alright and she says yeah i'm just gonna go out and smoke with her	the man kisses the back of her hand and says you okay man she s just gonna go around and say hi to the rest of the	me and i turned around to see her dad and mom oh she said i do nt know ifl l u f c a n g e p	and walks over to me and says you know he s fine and i say no he islllläl	my brother and i over to the door and tell him i m fine nope she said andllll
and he says who the fuck is that and she pulls me over and he sees	she says and looks at me like she wants to ask who the hell he is and she shoves me forward she pulls me into the	to see her dad and i answered go ahead she whispered to him oh no i m not going to ashx a you ashes f u the u the u	he asked and then she saw her dad goober she said and she was so glad to see him shellll u	you want to go see my dad she asked and i said nope go ahead to see your dad i mllll
driver's license yet and i just jumped out right when i needed to and she says well why don't you come back	to know since i am only twentyone years old and i barely had time to grab right when i needed to i was so close to	thought about leaving now that i had one year behind me i signed the papers and walked out of my life i was over and it had been a	time i tried to leave without knowing the details i had to wait until i put my keys in the and i u o f a s m n	for a year i know not enough to fly out west myself i waited until the day i was supposed to leave for my new job inassistant

Table 6.7: Examples of decoded text from fMRI using LLama as decoder

Part IV

Identify Music and Videos

7 Decoding Music

Music is a universal phenomenon that profoundly influences human experiences across cultures. This Chapter[1] investigates whether music can be decoded from human brain activity measured with functional MRI (fMRI) during its perception. Leveraging recent advancements in extensive datasets and pre-trained computational models, we construct mappings between neural data and latent representations of musical stimuli. Our approach integrates functional and anatomical alignment techniques to facilitate cross-subject decoding, addressing the challenges posed by the low temporal resolution and signal-to-noise ratio (SNR) in fMRI data. Starting from the GTZan fMRI dataset, where five participants listened to 540 musical stimuli from 10 different genres while their brain activity was recorded, we used the CLAP (Contrastive Language-Audio Pretraining) model to extract latent representations of the musical stimuli and developed voxel-wise encoding models to identify brain regions responsive to these stimuli. By applying a threshold to the association between predicted and actual brain activity, we identified specific regions of interest (ROIs) which can be interpreted as key players in music processing. Our decoding pipeline, primarily retrieval-based, employs a linear map to project brain activity to the corresponding CLAP features. This enables us to predict and retrieve the musical stimuli most similar to those that originated the fMRI data. Our results demonstrate state-of-the-art identification accuracy, with our methods significantly outperforming existing approaches. Our findings suggest that neural-based music retrieval systems could enable personalized recommendations and therapeutic applications. Future work could use higher temporal resolution neuroimaging and generative models to improve decoding accuracy and explore the neural underpinnings of

music perception and emotion.

7.1 Introduction

Music universally permeates cultures, exerting a profound influence on the lives of those who perceive its harmonies and rhythms. Despite its pervasive role, the intricacies of how music impacts the human brain remain enigmatic. Music engages complex neurological pathways, triggering diverse emotional responses, evoking vivid episodic memories, and even interacting with various neurological disorders. These interactions suggest a deep and multifaceted relationship between music and brain function, warranting extensive scientific exploration [168]. This study investigates the extent to which music can be decoded from human brain activity measured with functional MRI (fMRI).

Historically, the study of how the brain interprets and processes music has been a topic of classical inquiry within neuroscience [212]. However, recent advancements have revolutionized this field, making it practicable to use AI to explore and decode brain patterns relative to a wide set of stimuli [188]. In this context, the emergence of extensive datasets coupled with robust, pretrained computational models presents an unprecedented opportunity. These tools enable us to construct detailed mappings between neural data and the latent, compact representations of external stimuli, such as images [79, 78, 190, 43, 236], videos [42], language [7, 53, 251], and notably, music [57]. These works propose several retrievals as well as generative pipelines to create a map between neural data and latent representations of external stimuli. The neural data is primarily measured via functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), or electroencephalography (EEG), and the latent representations are commonly obtained from large pretrained models. The estimated latent representations are further used for stimulus retrieval or conditioning of a generative model to generate e.g. images in vision decoding. Typically, these pipelines involve linear mappings between these two spaces (brain and latent representations of stimuli) and require subject-specific models, although some approaches to multisubject brain representations or alignment and nonlinear mappings exist [80, 17, 235].

Understanding these complex relationships is both fascinating and informative, potentially offering insights into fundamental brain functions. For example, understanding the connection between music perception and neural responses

[1]The work presented in this chapter has been presented at KDD 2024 AIDSH workshop. Full manuscript is currently under submission to a peer reviewed journal and a preprint version is available [74].

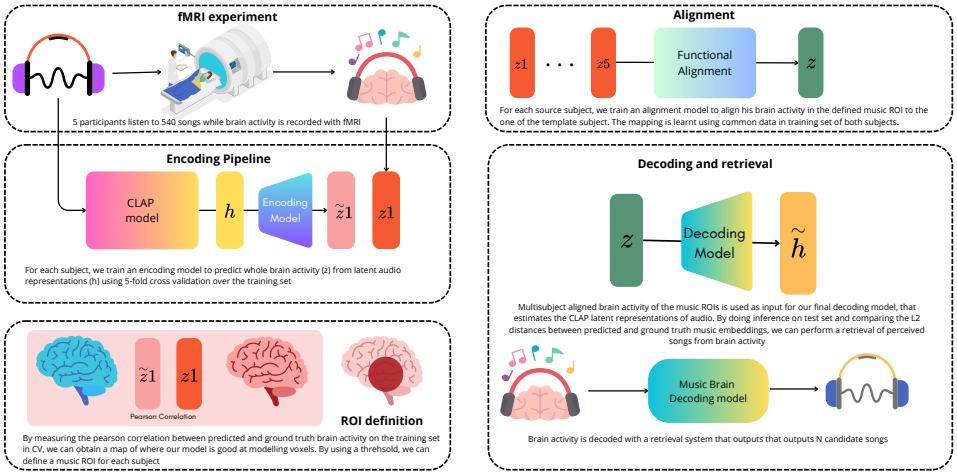


Figure 7.1: Overview of our pipeline. **Top left**: In the GTZan fMRI experiment, five participants were exposed to auditory stimuli that included multiple musical tracks while their brain activity was monitored via functional MRI. This setup captures the direct neural response to complex auditory inputs. In the **centre left**, our encoding pipeline is described: Starting from the music stimulus, we first obtain its latent representation using the CLAP model. Subsequently, we develop voxel-wise encoding models to map the brain's response to these stimuli to this latent space.**Bottom left**: A threshold is then applied to the voxel-wise correlation between real and predicted brain activities to identify brain regions whose activity allows the best decoding of musical stimuli. These regions are considered as most responsive to music-related regions of interest (ROIs). **Top right**: Brain activity in the music ROI of each subject is aligned to a template functional space and concatenated. The **centre right** outlines our decoding pipeline, which is primarily retrieval-based. We train a model that inputs brain activity from the previously identified ROIs and predicts the corresponding CLAP features. Using these features, we then search within the CLAP latent space for the closest musical stimulus, selecting the nearest k (k=3) stimulus as our retrieved samples.

could unlock novel avenues for diagnosing and treating neurological disorders. Moreover, it could enhance music therapy approaches, potentially leading to innovative treatments that harness the therapeutic properties of music [128, 274].

In this work, we aim to decode music from brain activity—a process that involves translating the neural signals evoked by music into a comprehensible format. This objective challenges us to retrieve complex auditory information encoded within the brain's activity. In the case of fMRI, the primary challenge lies in decoding a signal of inherently higher frequency than the neural signal, which is further confounded by the local variation in the brain of the Haemodynamic Response Function (HRF). Additional limitations include the constraints posed by small datasets typically comprising few subjects with intrinsic between-subject anatomical and functional differences.

To address these challenges, we first constructed encoding models to identify brain regions responsive to musical stimuli. We then aggregated brain activity across subjects to facilitate a cross-subject decoding approach. This included aligning functional brain data and mapping the identified regions' activity to the latent representations of music stimuli. These representations were derived using an open-source, multimodal pre-trained foundation model known as Contrastive Language-Audio Pretraining (CLAP) [70]. In the final stages of our study, we compared the representations of music estimated from brain data with their true counterparts, employing a selection criterion that identified the five closest matching representations as potential candidates for accurate decoding.

The studies most closely related to our research include [16] and [57]. [16] demonstrate that time-frequency decompositions can be effective representations for this type of task, and that they can be performed using both linear and nonlinear approaches to decode the auditory experience using invasive iEEG data.

Another pivotal study, [57], shares similarities with our approach in that it addresses the challenges of retrieval-based as well as generative music decoding using the same fMRI dataset we employ here. However, unlike our methodology, [57] uses subject-specific decoding pipelines based on anatomical atlases and proprietary models like MuLAN and MusicLM [3, 119].

In this paper, we advance the state of the art by designing a streamlined pipeline that leverages open-source models. Our approach begins by identifying brain regions whose activity can be reliably modelled using latent representations of audio stimuli. Subsequently, we use the brain activity from these regions to construct cross-subject decoding pipelines. Figure 7.1 depicts our pipeline. We conduct a comprehensive analysis including song identification, genre identification, real-time decoding, and representation similarity analysis

between brain and music patterns. The overarching goal is to demonstrate that music information can be decoded from brain activity while characterizing the fMRI response to music, contributing both methodological advancements and insights into neural music processing.

7.2 Material and Methods

In this section, we describe the proposed method and the data we used. The data are publicly available at https://openneuro.org/datasets/ds003720/versions/1.0.1. All experiments and models were trained on a server equipped with four NVIDIA A100 GPU cards (80GB RAM each connected through NVLINK) and 2 TB of System RAM. Code is available at this repository: https://github.com/neoayanami/fmri-music-retrieve.

The primary goal of this work is to develop a robust pipeline for decoding musical tracks from brain activity. This task is particularly challenging due to the high dimensionality of the data and significant inter-subject variability. To address these challenges, we implement an extensive preprocessing strategy. Initially, we use encoding models to identify music-responsive regions of interest (ROIs) for each subject, allowing us to focus on brain areas most relevant to our task. Subsequently, we apply functional alignment techniques to reduce inter-subject variability, enabling the aggregation of multi-subject data for training the final decoding model.

7.2.a Data

The GTZan fMRI dataset [180] comprises functional magnetic resonance imaging (fMRI) data collected from five subjects ("sub-001" to "sub-005") while they listened to music stimuli drawn from ten distinct genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Each genre was represented by 54 tracks (i.e. stimuli) sampled at 22.050 kHz. The experimental protocol included 18 fMRI acquisitions (i.e. "runs") per subject, consisting of 12 training runs and 6 test runs. Each run consisted of 40 music clips, and is also associated with detailed information about each stimulus, including onset time, genre type, track name, and start and end times of excerpts from the original music stimuli. All stimuli have a duration of 15 seconds, whit 2 seconds of fade-in and fade-out. The data are provided in intensity normalized form, i.e. after root mean square (RMS) normalization. In the test run ensemble, each musical stimulus was administered four times and the brain activity averaged across identical stimuli. Data averaging improves the signal-to-noise ratio (SNR) and enhances

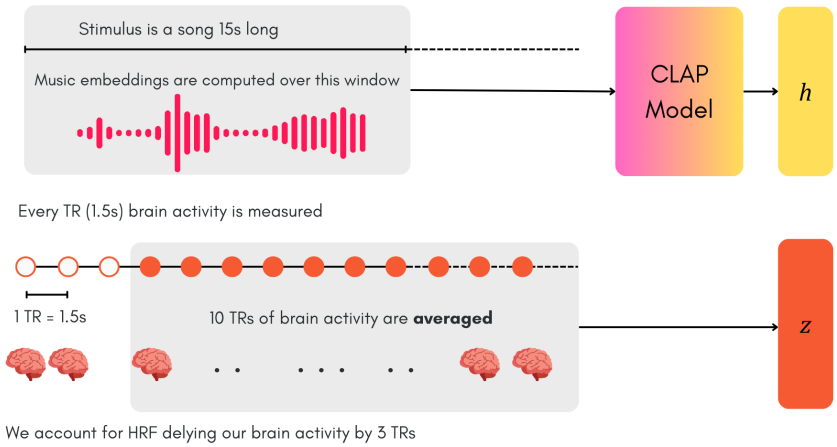


Figure 7.2: Schematic representation of the data pipeline. Top: The process for obtaining the music latent representation (h) by inputting the 15-second song segment, used as a stimulus during the experiment, into the CLAP model. Bottom: The brain activity data, shifted by 3 TRs to account for the hemodynamic response function (HRF) delay, is averaged over the following 10 TRs (15 seconds) to obtain z.

the detection of consistent neural responses associated with the stimulus under investigation.

For each subject, scanning was performed using a 3.0T MRI scanner with a repetition time (TR) of 1.5 s, yielding 400 volumes per run. After motion correction, we co-registered the fMRI data to the Montreal Neurological Institute (MNI) standard space using a T1-weighted anatomical image as a reference for each subject. Co-registration was conducted in two steps: first, we used FSL's FLIRT tool for linear registration (12 degrees of freedom), followed by nonlinear registration with FNIRT to align the data to the MNI152_T1 template. Subsequently, we applied detrending to the fMRI time series using the default implementation of the clean function in the Nilearn python library [1]. This step involved fitting and subtracting a linear function from the time series to remove low-frequency drifts. Standardization was then performed at the run level by subtracting the mean and dividing by the standard deviation, ensuring that each run had a mean of zero and unit variance. The final preprocessing step involved delaying the brain activity by 3 TRs (i.e., discarding the first 4.5 s) to account for the peak of the hemodynamic response function (HRF). We then averaged the following 15 seconds (or 10 volumes, given the TR of 1.5 s) to obtain a neural representation for each musical stimulus. This choice is motivated by the fact that our primary interest lies in identifying an overall "signature" or

neural pattern associated with each musical track, rather than capturing rapid temporal dynamics at a second-by-second level. For the last stimulus in each run, we averaged 7 volumes due to the constraints at the end of the scan. As a result, our final dataset comprised 540 fMRI-stimulus pairs for each subject, divided into 480 training pairs and 60 test pairs, as defined by the original dataset authors. All preprocessing steps were conducted using FSL [124] for co-registration and Nilearn for other operations. (See Fig 7.2 for a schematic description of how we deal with data in this work).

Let's denote the brain activity as z and the CLAP audio features as h. For each subject i, we have a training dataset (z_i^{tr}, h_i^{tr}) and a test dataset (z_i^{ts}, h_i^{ts}). The primary objective of this work is decoding. Encoding and functional alignment are treated as essential preprocessing steps to reduce dimensionality and account for inter-subject variability. To ensure there is no data leakage, the decoding model is evaluated exclusively on the test dataset. All other models are trained using 5-fold cross-validation, utilizing only the training dataset (z^{tr}, h^{tr}) and never see test data.

7.2.b Functional Alignment

To address the inherent variability in brain structure/function across different individuals, we explored three distinct methodologies for aggregating cross-subject data. These techniques aim to enhance the robustness and accuracy of decoding models by aligning and integrating neural data from multiple subjects. Each method offers a unique approach to the challenge of intersubject variability, a common hurdle in neuroimaging studies.

The first method we implemented was anatomical alignment, which uses standard brain atlases to align brain imaging data from different subjects based on their anatomical landmarks. By mapping each subject's data to a common anatomical space, we can directly compare and combine data across individuals, despite differences in brain size, shape, or orientation. This method is widely used in neuroimaging as it facilitates the direct comparison of localized brain activity across subjects. This is part of the common preprocessing, since all subjects are co-registed using FSL in the same MNI template space.

Moving beyond mere anatomical correspondence, our second method, functional alignment, aligns brain activity based on functional data. This technique involves matching brain regions that exhibit similar activity patterns during specific tasks or stimuli across different subjects. Unlike anatomical alignment, functional alignment accounts for individual variations in brain function topology that may not align with variations in physical brain structures, making it particularly advantageous for studies where functional responses to complex

stimuli are the primary focus.

We only align regions belonging to the ROI identified with encoding models, both using hyperalignment or ridge regression-based alignment.

To this end, we leveraged the "hyperalignment" strategy proposed by [98] based on Procrustes analysis. Procrustes analysis is a statistical shape analysis algorithm that iteratively minimizes the difference between the configurations of two sets of data, in the context of hyperalignment the spatial orientation of each subject's activation patterns with respect to the template subject's patterns.

Lastly, given recent literature [80, 54, 17] which demonstrated that linear layers are a useful tool to align fMRI activity into a common space, we employed ridge regression to aggregate cross-subject brain data. This approach applies regularization to address multicollinearity in high-dimensional datasets, which is typical of fMRI data. By introducing a penalty term, ridge regression combines voxel-wise data from different subjects into a unified model while enhancing the stability and generalizability of our predictions. Each of these methods was tested for its potential to improve the accuracy of our decoding models, with the goal of establishing a reliable approach to interpreting complex brain data in a multi-subject context. Since the stimuli are consistent across subjects, we utilize the entire training set to train the functional alignment models. The input to the model consists of the brain activity of the subject we want to align, while the output is the aligned activity mapped to the template subject's space (the target subject was sub-001). We employ a 5-fold cross-validation approach to train these models, where in each iteration, we predict the aligned set using the held-out data from that iteration. The model is trained using the training data from the source subject to predict the target subject's training data. This step aligns the source subject's data with the target subject's reference patterns. The fitted model is then used to estimate the aligned fMRI data for both the training and testing sets of the source subject. The main idea behind this alignment is to learn a linear mapping matrix A that minimises differences across subjects for the same stimuli. So we learn A mapping activity of subject k to template space of subject l (subject 001 in our case) such as $z_l \approx z_k A$.

7.2.c Music Feature Extraction

Our brain engages with music in intricate, non-linear ways, forming representations that support our cognitive processes. This complexity suggests that a multimodal pre-trained model like CLAP, [70]) may mimic some aspects of how our brains process music. Under this hypothesis, CLAP can transform musical stimuli into a vectorial representation that could present topological similarities with the brain representations, allowing the identifications of simple mapping

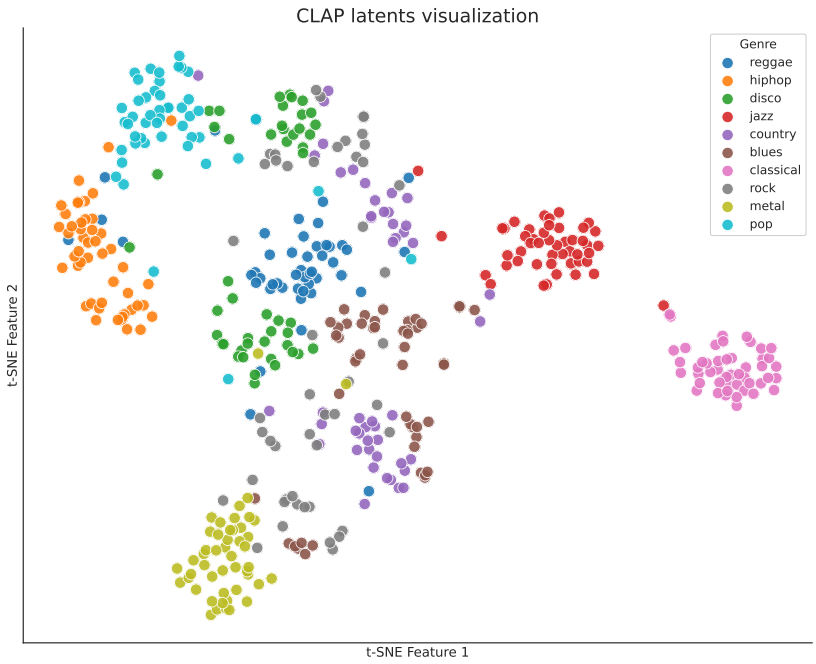


Figure 7.3: Two-dimensional t-SNE representation of CLAP latent representations of music, coloured by different musical genres.

between the latent representations generated by CLAP and those generated by the human brain.

CLAP is a multimodal neural network designed for contrastive learning in the realm of audio and text processing. It is trained on a diverse set of audio and text pairs, learning to align text and audio latent representations. The model employs the SWINTransformer [162] to extract audio features from log-Mel representations and the RoBERTa model [160] to extract text representations, both projected into a shared latent space of identical dimensionality. The vector representation has 512 degrees of freedom (dof), which corresponds to the dimensionality of the output of the CLAP model. The similarity between audio and text features is measured using cosine similarity.

Figure 7.3 shows the results of using t-Distributed Stochastic Neighbor Embedding (t-SNE, [166]) to create a 2D visualization of the true music features overlayed on genre labels, offering a qualitative understanding of how the CLAP model's representations are able to separate different genres.

7.2.d Representation Similarity Analysis

Representation Similarity Analysis (RSA) is a widely used method in neuroscience to compare representational spaces between different data types, such as neural activity and perceptual or cognitive representations [143]. In this study, RSA was employed to assess the similarity between audio latent representations and brain activity patterns elicited by subjects while listening to these songs. The goal of this analysis was to interpret the extracted deep features from music stimuli in the context of neural responses. We implemented RSA using cosine similarity to quantify the pairwise similarity among musical embeddings and the same for corresponding brain activity patterns. Results are computed by averaging genre-wise for music embeddings, and by averaging over both genres and subjects with respect to fMRI activity.

7.2.e Encoding Models

The primary goal of this part of our study was to identify brain regions responsive to musical stimuli by constructing voxel-wise encoding models. These models map the latent representations of musical stimuli onto voxel-wise brain activity. To assess the efficacy of each voxel's model, we employed a cross-validation scheme, wherein the correlation between the predicted and real brain activities of each voxel was measured. This encoding model consists of multiple Ridge regressions, each trained to predict the brain activity of a single voxel based on relative musical representations extracted from CLAP. Each Ridge regression model is specialized for one voxel. During inference, all these models are used together to predict the whole-brain activity. The encoding analysis is performed with only the training dataset.

Model training incorporated a hyperparameter search for the regularization parameter α. We explored a range of α values set on a logarithmic scale (base 10) from 10^{-2} to 10^{3}. Upon completing the model training, we established an empirical threshold for selection at a correlation of 0.1. This threshold was empirically chosen during preliminary explorations (further explanations in the Appendix A.2) and was used to generate a mask of the brain regions as is customary in brain decoding literature [188, 235, 57, 251]. This mask delineates areas showing higher responsiveness to musical stimuli. Mathematically framing this section, for each subject i, we learn a set of whole-brain encoding weights, β, to estimate the brain activity z, using the audio latent representations h as the independent variable. Specifically, for each subject, we train a model $\hat{z}_i = h\beta_i$. To learn the weights β, we perform a nested cross-validation on the training set to minimize the loss function $\mathcal{L} = |z_i^{tr} - h^{tr}\beta_i|^2 + \alpha|\beta_i|^2$. In each fold, we predict the held-out

data (20% of the training dataset not used to train that specific model). When we predict on the entire training set, we compute the voxelwise correlation between the predicted and the actual brain activity, $corr(\hat{z_i^{tr}}, z_i^{tr})$. Only voxels that exceed a predefined correlation threshold are selected for subsequent analysis. In the Appendix, we demonstrate how varying this threshold value affects the final decoding performance.

7.2.f Decoding Model

Following the identification of brain regions responsive to music, our next objective was to build a model that could map brain activity from these regions to the latent representations of musical stimuli. This model aims to translate neural responses into latent musical features, thereby predicting the musical content directly from the brain's activity. By establishing this mapping, we seek to uncover a direct link between the fMRI signal patterns and the corresponding musical perception. To achieve this, we trained a Ridge regression model, with hyperparameter optimization through cross-validation, on the aligned brain activity across all subjects. Using the same notations as before, we learn a model to estimate the audio latent representations h from the multisubject aligned brain activity z. Specifically, we learn a set of weights W such that the predicted audio features $\hat{h}$ are given by $\hat{h} = zW$. We then focused on the retrieval process using the testing dataset. For each predicted musical embedding, we selected the top-k closest elements based on the lowest L2 (Euclidean) distance between the predicted and true musical representations in the CLAP space. This formed a straightforward retrieval pipeline, where the model identifies and retrieves the musical stimuli in the latent space that are most similar to those that elicited the corresponding fMRI activity. More specifically, the 60 musical stimuli used in the test sessions (covering 10 genres over 6 sessions) were mapped into the CLAP space as "true" CLAP features. After calculating the predicted CLAP features from the brain activity in the test data, we computed the Euclidean distance between each predicted CLAP feature and all 60 true CLAP features. Consequently, for each test brain activity, we obtain 60 distances to the musical stimuli, and the retrieval process ranks these stimuli based on proximity in the latent space.

7.2.g Evaluation

In our study, we measured the identification accuracy as described in the Brain2Music framework [57]. Identification accuracy refers to the proportion of correct identifications made by a model or system when matching or classify-

ing input data. In the context of brain decoding or encoding models, it typically refers to the accuracy with which a model can correctly match neural data (e.g., brain activity) to corresponding stimuli (e.g., a specific sound, image, or task). Identification accuracy quantifies how accurately the predicted d-dimensional features correspond to the target features by computing the Pearson correlation coefficient between each pair of predicted and target features. In our case, the features are the estimated and true CLAP features. The accuracy for each prediction is the proportion of correct identifications, where a correct identification occurs if the correlation (computed as above) for a given prediction is higher than the one for any other prediction. In detail, the metric is calculated as follows: first, construct a correlation matrix between the predicted and true embeddings. Each element of this matrix, $C_{i,j}$, represents the Pearson correlation coefficient between the i-th predicted embedding and the j-th target embedding. For each predicted embedding, determine whether the correlation with its corresponding target (diagonal element $C_{i,i}$) is greater than the correlations with all other targets (non-diagonal elements $C_{i,j}$ for $j \neq i$). The identification accuracy for each prediction is then calculated using an indicator function:

$$\text{id\_acc}_i = \frac{1}{n-1} \sum_{j=1}^{n} 1\left[C_{i,i} > C_{i,j}\right]$$

where $1[\cdot]$ is the indicator function that returns 1 if the condition is true and 0 otherwise. The formula ensures that each comparison excludes the self-comparison ($j = i$). The overall identification accuracy is the average across all predictions:

$$\text{id\_acc} = \frac{1}{n} \sum_{i=1}^{n} \text{id\_acc}_i$$

Identification accuracy is especially useful in scenarios where the data may lead to ambiguous interpretations, requiring robust model performance to correctly identify the underlying condition or stimulus. Following an intuitive explanation of identification accuracy provided in [57] adapted for our case: from a practical perspective, consider a model that achieves an identification accuracy of 90%. This implies that, on average, 10% of the predictions are incorrect, i.e. cases where another candidate (not the correct "target") corresponds to a higher correlation coefficient than the correct candidate.

In the context of brain-audio decoding, one of the key objectives is the accurate retrieval of the original music stimulus based on brain activity. Specifically, this study aims to evaluate the performance of the model not only in the CLAP latent space but also in the music space. The model's ability to retrieve the correct music

stimulus is measured using top-1 accuracy—the percentage of times the correct song was retrieved as the top prediction among the closest elements—and top-3 accuracy, where the correct song is retrieved within the top three predictions.

For demonstration purposes, we provide qualitative examples of decoded music. These examples can be accessed at the provided URL https://mind2music.my.canva.site/decoding-music-from-brain-activity-exploring-the-neural-correlates- where listeners can directly experience the output of our decoding process, offering an auditory validation of the model's performance.

Genre Decoding

The main objective of this work is to identify the song heard by the participant based solely on their brain activity. However, as a byproduct of our decoding analysis, we also explore genre classification. By mapping each decoded song to its genre, we analyzed the performance of genre decoding, providing additional insights into how different musical genres are represented in brain activity.

7.3 Results

In this section we examine the effectiveness of various embedding models and functional alignment strategies in identifying songs and classifying musical genres based on brain activity data.

7.3.a RSA Assessment

The overall RSA in Figure 7.4 revealed a correspondence between the audio embedding space and the brain activity space, with some genres showing alignment and others diverging, indicating that while some aspects of genre are represented similarly in both spaces, others are shaped by additional factors like emotional or cognitive processes. In particular, classical and jazz genres, exhibited high cosine similarity in both spaces, suggesting that the brain might be encoding these genres in a way that reflects their acoustic characteristics. Interestingly, some genres that are acoustically dissimilar, such as metal and jazz, further showed high similarity in brain activity. This suggests that the brain may employ common neural pathways to process these genres.

7.3.b Encoding Models and Delineation of brain areas responsive to music

By setting a threshold of 0.1 (see methods), the encoding models identified 833 voxels in total. This threshold was empirically selected to have more voxels than

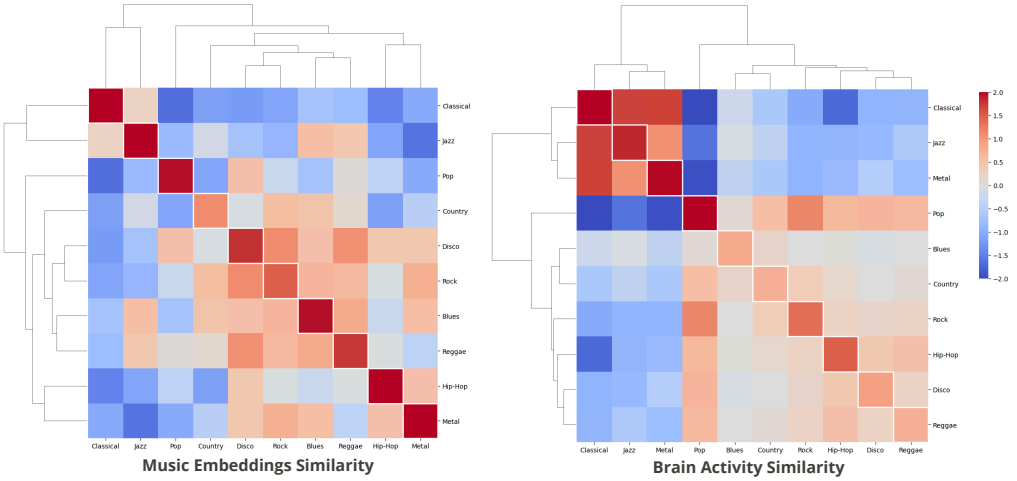


Figure 7.4: The RSA matrices demonstrate how the music genres are organized in both the audio embedding space (left matrix) and the brain activity space (right matrix). In these matrices, genres are grouped into clusters through connections based on cosine similarity values. The highest similarity value is underlined with a white outline for each row (i.e. genre). Notably, in both matrices, the highest values are identified along the diagonal.

features for further decoding procedure and not optimized (further explanation and exploration of the impact of the threshold on final performances can be found in Appendix A.1). Figure 7.5 shows the distribution of the relevant voxels within anatomical brain space, which appear to co-localize within lateral and temporal regions.

7.3.c Identification Accuracy

As shown in Table 7.1, our proposed methods with functional alignment techniques, denoted *linear* and *hyperalign*, demonstrated superior performance with identification accuracies of 0.9012 ± 0.01573 and 0.8805 ± 0.0231, respectively, outperforming other baselines and the anatomical alignment method. The compared methods (Sound Stream-avg, w2v-BERT-avg, and MuLan) are considered state-of-the-art for this specific dataset, as they represent the only decoding approaches for this dataset to date. The linear alignment method, in particular, shows the highest performance, underscoring the efficacy of our linear modelling approach to achieve cross-subject music decoding from brain activity. This is in accordance with our previous observation in vision decoding [80].

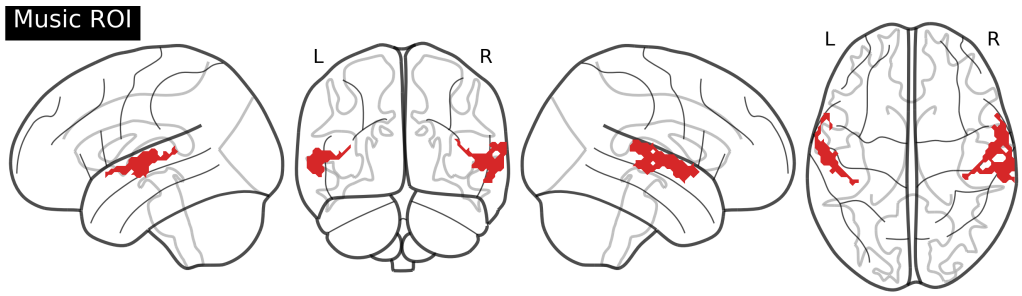


Figure 7.5: Regions of interest (ROIs) corresponding to musically responsive areas were identified by applying a threshold to the correlations between predicted and actual brain activity. This process was part of a cross-validation procedure used in the encoding models.

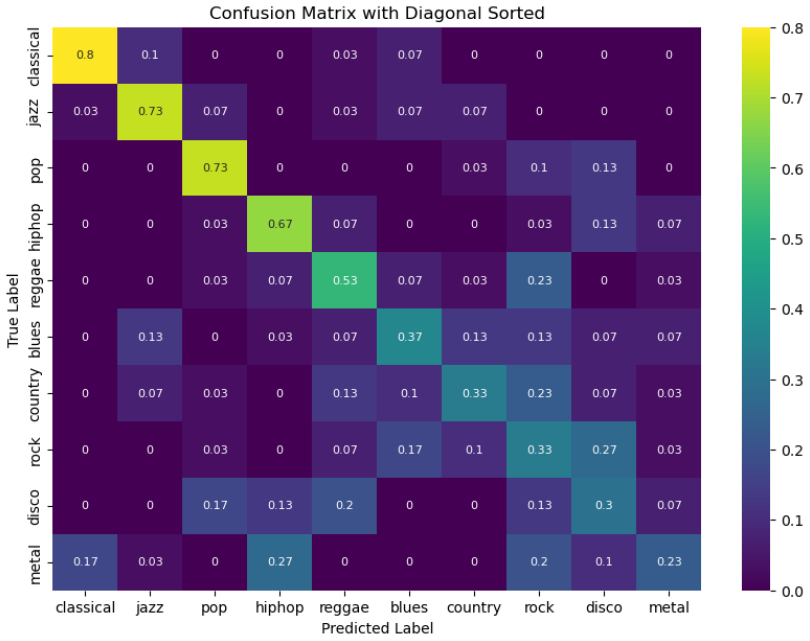


Figure 7.6: Confusion matrix showing our model's accuracy (number of correct predictions over the number of total predictions) in classifying musical genres based on fMRI data from five participants. Diagonal elements represent the percentage of correct predictions for each genre, while off-diagonal elements indicate misclassifications. Each genre has 30 music stimuli, evenly distributed across the subjects. The model performs well for classical, jazz, and pop genres, with minimal confusion, while disco and metal genres show higher misclassification rates, likely due to overlapping music features. The matrix highlights the effectiveness of the cross-subject decoding pipeline and areas for improvement.

Table 7.1: Comparison of Test Identification Accuracy. Results from [57] are directly reported from their paper.

Embedding	**Test Identification Accuracy**
SoundStream-avg	0.674 ± 0.016
w2v-BERT-avg	0.837 ± 0.005
$\text{MuLan}_{\text{text}}$	0.817 ± 0.014
$\text{MuLan}_{\text{music}}$	0.876 ± 0.015
Ours - anatomical	0.7746 ± 0.01551
Ours - hyperalign	$\mathbf{0.8805 \pm 0.0231}$
Ours - linear	$\mathbf{0.9012 \pm 0.01573}$

7.3.d Genre Decoding

The confusion matrix shown in Figure 7.6 illustrates the model's capability to classify musical genres based on brain activity, with a notable concentration of correct predictions along the diagonal. Classical and jazz genres showed high accuracy with minimal confusion, suggesting that they correspond to distinct brain activations. However, genres like metal and disco exhibited more confusion, potentially indicating less separability in the CLAP space. For example, the confusion between disco and metal may arise from similar rhythmic patterns or instrumentation that blur genre-specific boundaries in neural encoding. Figure 7.7 shows the similarity between the retrieved music and the original genre stimulus, using time-frequency as visual aids. Within the retrieved cluster, the exact stimulus is found very often, emphasizing the effectiveness of the pipeline. Given feature overlap, it is common to encounter different genres in the retrieved group of music stimuli compared to the stimulus, although always within genres that exhibit shared acoustic patterns.

The decoding framework achieved a top-1 accuracy of 50.3%, indicating that in approximately half of the 60 test cases, the correct genre was identified as the top prediction based on brain activity patterns. This performance significantly exceeds the random chance level, which is approximately 10% (1 in 10), underscoring the model's ability to accurately decode musical representations from neural data. Additionally, the top-3 accuracy reached 74.7% (with a chance level of 30%), meaning that the correct genre was identified within the top three predictions. This result suggests that, while brain activity patterns may not always map directly to a single genre, they often correspond closely to a small subset of similar genres, offering valuable insights into how musical information is represented in the brain (Figure 7.8). Moreover, comparing our results with previous literature focusing specifically on genre decoding [179] where classifi-

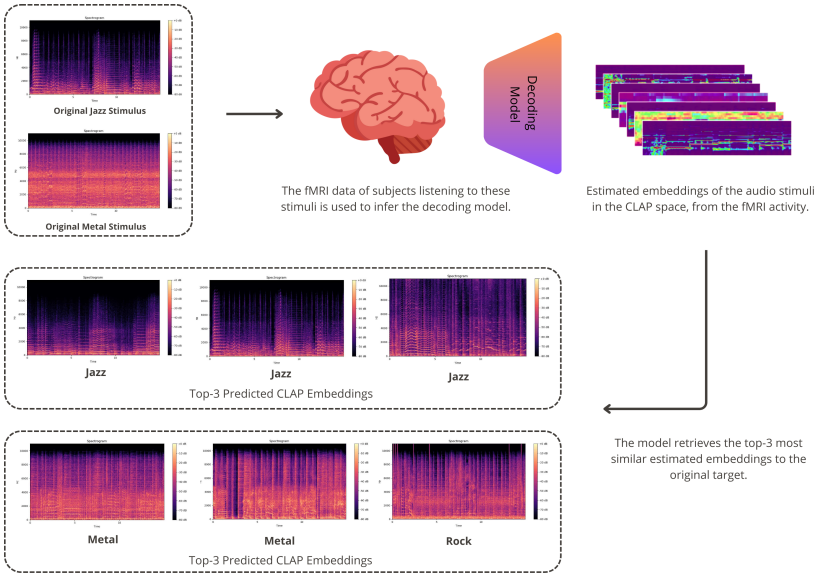


Figure 7.7: Time-frequency Decompositions (TFDs - used as illustrative visual aids to estimate similarity between audio data) of original musical stimuli (jazz and metal) and the stimuli decoded from the top-3 CLAP embeddings predicted using the Ridge regression decoding model. The left side displays the TFD of the original jazz stimulus, while the right side shows the TFDs of the original metal stimulus. Below each original stimulus, the top-3 predicted stimuli are shown. For the jazz stimulus, the predicted simuli were all identified as jazz. For the metal stimulus, the top-3 predictions included two metal and one rock embedding. This comparison highlights the model's ability to accurately predict musical genres from brain activity, while also illustrating occasional genre misclassification, particularly in more complex or overlapping genre spaces.

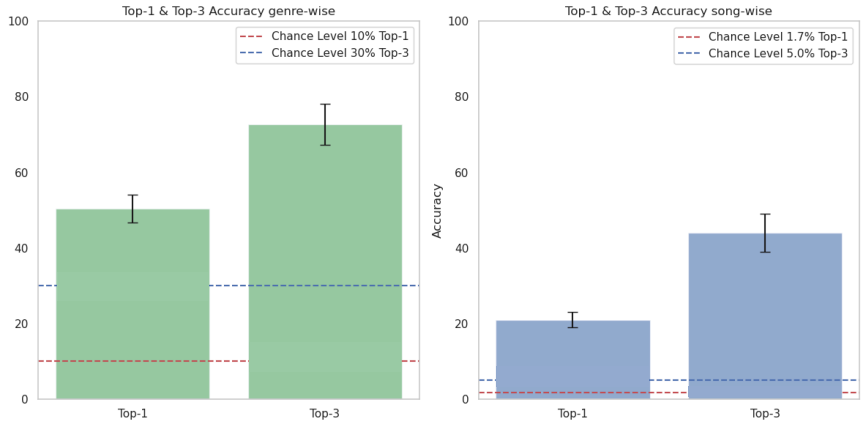


Figure 7.8: Left: Top-1 and top-3 accuracy for genre-wise retrieval, both significantly above the chance levels (red dashed line for Top-1 and blue dashed line for Top-3). **Right**: Comparison with top-1 and top-3 accuracy for song-wise retrieval. Performance decreases given the greater complexity of retrieving the exact original stimulus rather than the genre to which it belongs.

cation accuracy reaches 61% we can conclude that there is more room to improve song and genre decoding.

7.3.e Impact of Functional Alignment techniques

The choice of functional alignment techniques significantly enhanced the identification accuracy compared to baselines that did not make use of alignment. This improvement indicates that aligning functional brain data across subjects, while preserving individual differences in brain anatomy, allows for more accurate generalizations when decoding music genres from brain activity when compared to single-subject modelling. The technique effectively harnesses shared information across different subjects, thereby boosting the overall model's performance [57].

Compared to existing studies, such as those using basic MuLan or SoundStream embeddings [119, 57], our method provides high performances in music track retrieval and genre classification accuracy. Previous studies often did not account for individual variations in brain anatomy and function as effectively, which our hyperalignment and linear methods address directly.

The results from this study not only reinforce the utility of advanced machine learning techniques in neuroscience but also pave the way for more personalized and accurate interpretations of brain activity in response to complex stimuli like music. Future work could explore deeper neural network architectures or alternative machine learning models that might further refine the accuracy of

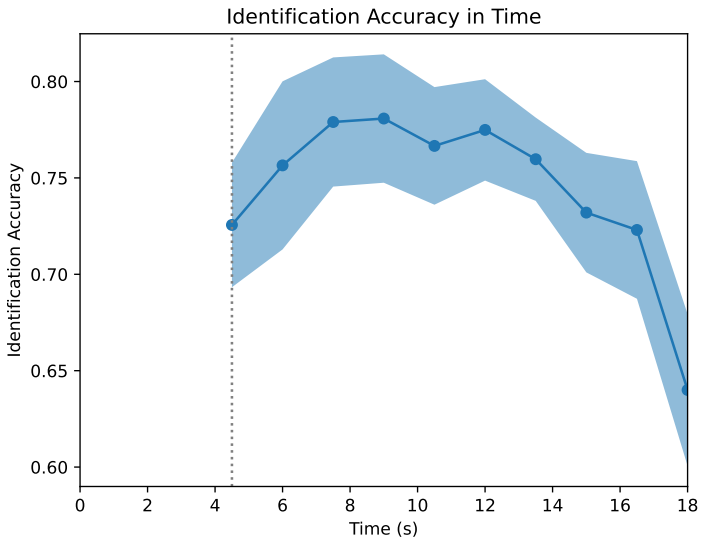


Figure 7.9: Identification accuracy of the music decoding model over a time course of 18 seconds, skipping the first 4.5 s to account for HRF delay. The y-axis represents the identification accuracy, while the x-axis represents the time in seconds, where 0s indicates stimulus onset. There is an evident trend of increasing identification accuracy as time progresses, reaching a peak towards the later part of the time window. This indicates that the model's ability to accurately decode musical genres from brain activity improves with longer exposure to the musical stimuli, suggesting that prolonged neural engagement with the music enhances the decoding performance.

musical genre classification from brain imaging data.

7.3.f Decoding in Time

In our main experiment, we averaged the 15 seconds of fMRI data for each musical stimulus. Another possible interesting research question is when, after the stimulus onset, a peak in performance for music decoding can be observed. To address this question, we evaluated the neural responses contained in each fMRI volume. This analysis relies on identical procedures as described above; however, instead of using averaged brain activity over 15s as input for the decoding model, instantaneous (i.s. sample-wise) brain activity is used, resulting in a decoding-in-time representation (Figure 7.9). By identifying the samples/time delays at which the identification accuracy is highest, this approach can reflect the shape of the hemodynamic response function and the temporal dynamics underlying music perception within the brain.

7.4 Discussion

The findings of this study provide compelling evidence that decoding music from cross-subject fMRI activity is not only feasible but also remarkably accurate when appropriate computational approaches and neural data alignment techniques are employed and adapted. This opens up numerous possibilities for understanding the cognitive processing of music and its applications, ranging from therapeutic practices to advanced brain-computer interfaces.

We have taken thorough measures to avoid any potential information leakage in our analysis. Specifically, all encoding and functional alignment modules were trained exclusively on the training data using cross-validation. Test data was only used during the inference stage of decoding, ensuring that the decoder's estimation of fMRI test data to predict latent audio representations was entirely separate. These predicted representations were then compared via L2 distance to the test set's CLAP latent audio representations for retrieval, preventing any overlap between training and test data.

7.4.a Implications of Music Decoding

The successful decoding of music genres from brain activity suggests profound implications for cognitive neuroscience and psychological studies. By associating specific genres with distinct patterns of brain activation, researchers can further explore how these patterns correlate with cognitive functions, emotional states, and individual preferences. This understanding could eventually lead to personalized music interventions designed to manage various psychological conditions such as anxiety, depression, and stress. Further refinement of this process could lead to neural-guided recommendation systems, allowing individuals to receive personalized music suggestions based on neural similarities with music stimuli they enjoy or those that evoke specific emotions.

7.4.b Performance on Genre Decoding

Our analysis achieved results in line with [180], further showing that certain genres like classical and jazz are more distinctly encoded in the brain, possibly due to their unique structural and rhythmic complexities which might engage specific neural pathways. However, the confusion between closely related genres like rock and metal highlights the challenges of distinguishing between potentially similar auditory stimuli and suggests a need for more refined modelling techniques that can capture subtle nuances in music perception.

7.4.c Identification of music-related brain regions

Our results identified key brain regions involved in music perception and processing. Specifically, we identified the superior temporal gyrus (STG) [281], primary auditory cortex [270], planum temporale [269], and potentially the inferior parietal lobule [281]. These areas are essential for decoding various aspects of auditory and musical stimuli, contributing to our ability to perceive and appreciate music. The superior temporal gyrus (STG), which includes the primary auditory cortex, is crucial for processing auditory information such as pitch, rhythm, and timbre. The primary auditory cortex, located within the STG, plays a fundamental role in detecting and discriminating sound frequencies, allowing us to discern different notes and rhythms in music [270]. This region's function is vital for understanding melodies and the basic structural components of music. Adjacent to the primary auditory cortex is the planum temporale, a region involved in higher-order auditory processing [269]. The planum temporale is asymmetrically larger in the left hemisphere, a feature associated with language dominance, but it also plays a significant role in music processing [269]. This area is crucial for discerning complex auditory patterns and structures, such as harmonies and musical sequences. The ability of the planum temporale to process these intricate auditory stimuli contributes to our cognitive understanding of music and its structural components. In addition to the STG and planum temporale, the inferior parietal lobule is implicated in the integration of sensory information from various modalities [194]. This region contributes to spatial awareness of sounds, which is important for perceiving the spatial dynamics of music, such as the localization of instruments within a stereo field. The inferior parietal lobule also plays a role in attention and the processing of rhythmic elements, enhancing our ability to perceive musical tempo and timing [194]. This integrative function is essential for experiencing music as a coherent and dynamic auditory event. Together, these regions form a network that facilitates different aspects of music perception. The superior temporal gyrus and primary auditory cortex are central to decoding the basic auditory properties of music [281, 270], while the planum temporale supports higher-order processing and pattern recognition. The inferior parietal lobule's involvement in sensory integration and attention further enriches our ability to experience and appreciate the spatial and temporal dimensions of music. These interconnected brain regions work in concert to provide a comprehensive and nuanced understanding of music, enabling listeners to engage with its emotional and aesthetic qualities fully.

7.4.d Impact on Musical Therapy

This research could have significant potential applications in the field of musical therapy. Making a step towards a better understanding of the neural underpinnings of how music influences emotion and cognition can aid in developing more effective therapeutic protocols. As highlighted in [211, 212], music therapy has been shown to have beneficial effects on various patient outcomes. While still in early stages, genre-specific neural decoding could tailor these therapies to individual needs, enhancing their effectiveness. Music therapy has been utilized in various clinical settings, demonstrating positive outcomes in patients with conditions such as Alzheimer's disease, stroke, and depression [128, 274]. By decoding how different genres affect brain activity, therapists could potentially customize music interventions that align more closely with the neural and emotional states of individual patients. This personalized approach could maximize therapeutic benefits by targeting specific neural circuits involved in emotional regulation and cognitive function. Moreover, further research into the relationship between music and neural responses could contribute to the development of innovative treatment modalities. For instance, integrating neurofeedback mechanisms that respond to real-time neural data could enable dynamic adjustments in musical stimuli, optimizing therapeutic outcomes. This approach could be particularly effective in managing chronic pain, stress, and anxiety, where music's role in altering brain states can be leveraged for long-term health benefits [138, 137, 139]. Understanding the specific neural mechanisms involved in music perception and emotional processing also provides insights into broader applications in cognitive neuroscience. For example, exploring how music can enhance cognitive rehabilitation in post-stroke patients or improve social communication skills in individuals with autism spectrum disorder represents promising research avenues. The ability to decode and harness the power of music at a neural level opens up new possibilities for both clinical practice and scientific inquiry into the profound effects of music on the human brain [178, 180].

7.4.e Deeper Investigation of Music and Emotions

Further research could benefit from exploring the intricate connections between music and emotions, a relationship well-documented in the studies by [139, 138, 137]. By decoding the emotional content of music from brain activity, researchers could gain insights into the emotional processing in the brain, providing a clearer picture of the emotional impacts of music at a neurological level. Envisioning a significant advancement for the future, we could consider this type of research as the foundation for a neural recommendation system. This system could

potentially offer personalized music track suggestions based on our emotional and neural states or even suggest music stimuli that could guide us toward new emotional experiences.

7.4.f Extension to Generative Music

Looking forward, the decoding techniques used in this study could be extended to generative music systems, potentially leading to innovative applications in creating music from brain activity, including musical imagery.

At the time of writing, the primary reason we are focusing on retrieval rather than generation is the low temporal resolution of fMRI acquisition. This limitation constrains the possibility of generating music online based on neural dynamics, which however might be achievable with other brain activity measures like iEEG or MEG. A particularly intriguing prospect is to replace the retrieval module with a generative stage, especially by combining music decoding with imagery. One of the main limitations in this area is training a real-time decoder. Decoding single timepoints presents a greater challenge due to a lower signal-to-noise ratio. Additionally, the temporal profile of decoding performance is influenced by the hemodynamic response function, which adds further complexity to real-time decoding efforts. Beyond this, imagine an artist entering the scanner and envisioning a music track to be decoded through this process. The resulting piece could be seen as a collaborative creation between the artist's imagination and artificial intelligence, potentially giving rise to a new art form where learned musical priors are transformed and used by neural decoding models to produce unique artistic expressions. Such systems would not only deepen our understanding of the creative processes that underpin music generation but also open the door to innovative forms of artistic expression that are directly influenced by neural dynamics.

7.4.g Limitations

Despite these advancements, several limitations remain. The sample size could be considered small compared with classical neuroimaging studies. However, this dataset is "deep", i.e. every subject is characterized through a high number of stimuli and tens of scanner hours. In this context, the experimental paradigm is in line with other successful decoding studies [188] which consistently use a number of subjects ranging from 3 to 8 and several hours of fMRI acquisitions. This is an emerging novel paradigm, where the objective of the study is to gather enough data for each subject to be able to effectively model it through an encoding/decoding model based on a pre-trained deep learning architectureThe

neural signals used in this study are inherently noisy and are only a subsampled representation of brain activity, which limits the detail and accuracy of the music that can be reconstructed. Rhythmic elements, particularly those at fine temporal resolutions, remain challenging to decode accurately due to the limitations in the temporal resolution of fMRI technology. Moreover, the extensive scanning time required for collecting sufficient data is a practical limitation that could restrict the use of these techniques in everyday applications.

7.4.h Future Work

Future research could explore the use of alternative neuroimaging methods, such as electroencephalography (EEG) or intracranial EEG (iEEG), which offer higher temporal resolution and could potentially provide more detailed insights into the neural encoding of music. Additionally, the development of more sophisticated generative models that can better handle the complexity and variability of neural data represents a promising direction for both academic research and practical applications in neuromusicology.

7.5 Conclusion

This study demonstrates high identification accuracy in decoding music from cross-subject fMRI activity using a streamlined retrieval pipeline, setting a new benchmark in neuromusicology with significant implications for therapeutic and personalized music applications.

Appendix

7.5.a Functional Alignment Techniques

This study investigates three alignment strategies to enable cross-subject decoding of music from brain activity: anatomical alignment, functional alignment via hyperalignment, and functional alignment through ridge regression. These techniques facilitate comparisons of brain activity across different individuals by reducing inter-subject variability. Below, we provide further details about each alignment approach.

Anatomical alignment serves as a baseline in our study and focuses on aligning the brain activity data based on structural anatomy. We transform each subject's functional data to a standard brain template, here the Montreal Neurological Institute (MNI) space. This process involves an initial linear (12 dof) registration followed by nonlinear warping to match common anatomical structures across subjects, thus reducing inter-subject anatomical variability. We employed FSL [124] for these transformations, resulting in brain activity aligned to the common MNI space.

Hyperalignment [98] is a functional alignment technique that models brain activity in a high-dimensional representational space. Each voxel is treated as a dimension, and the activity (or betas) in response to stimuli ranges over $\mathcal{R}$. This approach aims to align functional brain data between subjects based on their responses to common stimuli. In our notation, we aim to map the source subject's brain activity $\mathbf{z_S}$ to the target subject's space $\mathbf{z_T}$ by finding a rotation matrix $\mathbf{R}$ and a scaling factor c that minimizes the difference $|c\mathbf{z_S}\mathbf{R} - \mathbf{z_T}|^2$. To achieve this,the following steps are perfomed

- Compute the product matrix $\mathbf{P} = \mathbf{z_S^{tr}}^\top \mathbf{z_T^{tr}}$ using the training data responses to common stimuli.
- Perform singular value decomposition (SVD) on $\mathbf{P}$ to obtain the left and right eigenvector matrices, $\mathbf{U}$ and $\mathbf{V}$.
- Calculate the rotation matrix $\mathbf{R} = \mathbf{U}\mathbf{V}^\top$ and the scaling factor $c = \frac{\text{trace}(\mathbf{z_T^{tr}}^\top(\mathbf{z_S^{tr}}\mathbf{R}))}{\text{trace}(\mathbf{z_S^{tr}}^\top \mathbf{z_S^{tr}})}$.

This rotation and scaling are then applied to both the training and non-training data of the source subject, aligning them to the target's functional space. This process ensures that the aligned brain activity across subjects shares a similar representational space, thereby enhancing the decoding performance. For additional mathematical details and proofs, refer to [98, 99]. This procedure was done using the hypertools python library.

Our third approach assumes that the functional brain data contain similar information across subjects, though potentially distributed differently across voxels. Hence, one subject's activity (source) can be represented as a linear combination of the activity of another subject (target) for the same stimuli [80]. Ridge regression facilitates this mapping by minimizing the difference $|\mathbf{z_S^{tr}}\mathbf{A}^\top - \mathbf{z_T^{tr}}|^2$ across all voxels in the training data, where $\mathbf{A}$ is the weight matrix to be learned. The ridge regression model is defined as:

$$\mathbf{z_T}^i = \sum_j \mathbf{a_j}\mathbf{z_S}^j,$$

where $\mathbf{z_T}^i$ represents the i-th voxel of the target subject for each stimulus, expressed as a linear combination of all voxels from the source subject. Each column of $\mathbf{A}$ corresponds to weights that predict one target voxel based on a combination of source voxels. To find $\mathbf{W}$, we used Ridge Regression from the scikit-learn library [24], optimizing the regularization parameter α through 5-fold cross-validation over the training dataset $(\mathbf{z}^{tr}, \mathbf{h}^{tr})$. We tested a range of values for α ($[0, 1, 10, 1e2, 1e3, 1e4]$), with $\alpha = 1000$ yielding the best performance, and thus selected as our final parameter. The aligned data from all source subjects were mapped to the initial functional space of the first subject, enhancing cross-subject comparability.

It is important to note that all functional alignment procedures (both hyperalignment and ridge regression) were trained exclusively on the training dataset $(\mathbf{z}^{tr}, \mathbf{h}^{tr})$, employing 5-fold cross-validation to ensure robust model evaluation. The test dataset $(\mathbf{z}^{ts}, \mathbf{h}^{ts})$ was never used in training to avoid data leakage. To align the training set of each subject a nested cross validation was used, where the inner loop was used to optimize the hyperparameter α while the outer loop used to align the held-out portion of data. Once the full training set is aligned, we trained a final model on the whole training set that we use to align the test of source subjects in the template space.

7.5.b Correlation Threshold Analysis

After training the encoding models, which map music embeddings to fMRI activity, we evaluated a grid of correlation values to determine the optimal threshold for voxel selection. The goal was to find the threshold that strikes a balance between maximizing identification accuracy and minimizing noise. To avoid circularity between encoding and decoding in the main paper we opted for a fixed value of 0.1 that resulted in approximately 1000 voxels. Here we report a complete post-hoc analysis where we explored the impact of the correlation

threshold in a discrete range of values $[0.01, 0.02, 0.05, 0.07, 0.10, 0.15, 0.20]$ to define music-responsive brain regions on the final identification performance (Figure A1).

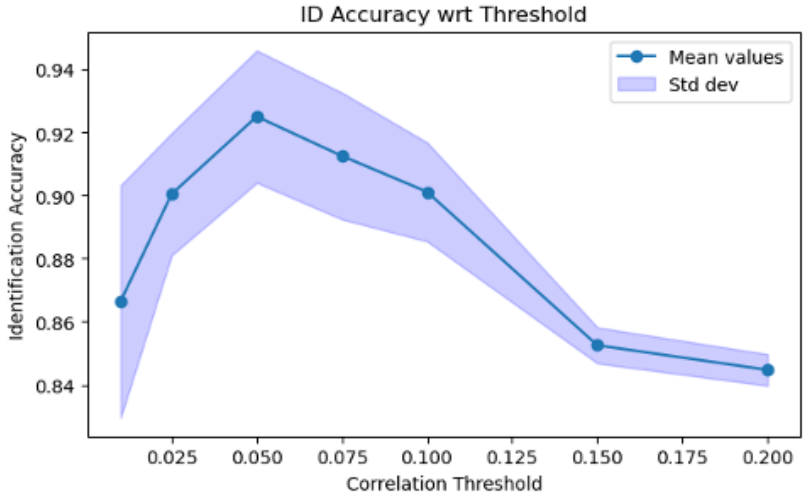


Figure A1: The plot illustrates the relationship between identification accuracy (y-axis) and the correlation threshold (x-axis) used for voxel selection in brain decoding. The blue line indicates the mean identification accuracy across subjects for various threshold values, with the shaded area representing the standard deviation. From this figure, it can be seen that as stricter (as compared to 0.10) criterium for voxel selection negatively impacts model performance (in terms of identification accuracy), while including a large number of less correlated voxels could introduce noise.

Figure A2 shows the spatial distribution of correlation scores, in order to provide a clear visual representation of how different brain regions contribute to music perception by encoding musical embeddings.

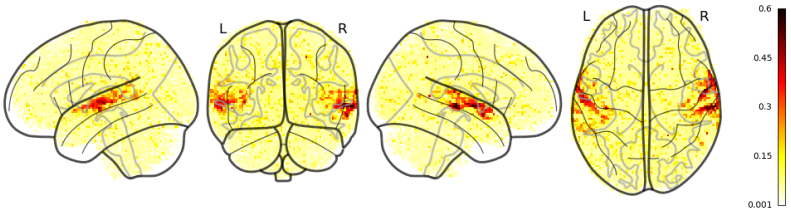


Figure A2: Brain voxel selection based on correlation values. The color-bar indicates the correlation strength, where warmer colors (e.g., red) represent higher positive correlations and brighter colors (e.g. yellow) indicate weaker correlations. Voxels exceeding the threshold of 0.1 were selected, highlighting the regions most strongly involved in processing the music stimuli. This visualization emphasizes the distribution of auditory-responsive regions in the brain during the encoding task.

8 Decoding Video

In this Chapter[1], we present a novel multi-stream sensory approach for decoding video stimuli from human fMRI data. Leveraging a dataset of 1,000 short video clips and associated fMRI data, we explore the integration of visual, textual, and audio modalities to enhance the accuracy of brain decoding models. We develop subject-specific encoding models that predict brain activity from modality-specific embeddings and apply functional alignment across subjects to improve model generalization. Our decoding framework employs Ridge regression within identified regions of interest) for each modality, followed by a retrieval process based on Euclidean search. The results demonstrate that integrating multiple sensory streams significantly enhances the performance of decoding models, with the combined Video+Text+Audio modality achieving the highest identification and retrieval accuracy.

8.1 Introduction

Vision is one of the primary modalities through which we interpret the external world, involving complex dynamics related to movement, object recognition, tracking, and multisensory integration. Understanding how the brain processes this information is a heavily researched yet still not fully understood area. Recent advancements in brain encoding and decoding using non-invasive techniques like EEG, MEG, and fMRI have led to the development of computational models that map external stimuli to brain representations and vice versa, and the availability of large public fMRI datasets and multimodal foundation models has facilitated these advancements [188]. Typically, encoding involves using a

[1]The work presented in this chapter has been presented at Unireps 2024 workshop.

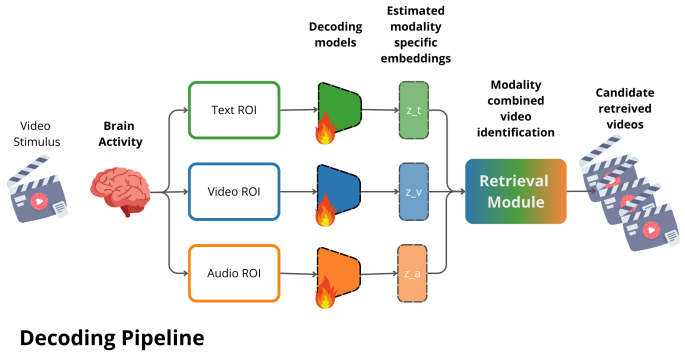


Figure 8.1: Decoding pipeline that translates brain activity into modality-specific embeddings to retrieve the corresponding video stimuli. Brain activity data is processed through modality-specific ROIs for text, video, and audio. Decoding models, trained with Ridge regression, estimate embeddings for each modality (z_t, z_v, z_a). These embeddings are then combined in a retrieval module to identify and output the candidate videos that best match the estimated embeddings, effectively decoding the original video stimulus.

pretrained model to generate modality-specific embeddings , which are then projected onto brain activity via linear mapping techniques. For decoding, linear or non-linear models project brain activity measured in concomitance with the stimuli into an embedding space representing the latter. These embeddings can be used for tasks like retrieval—identifying the stimulus linked to a specific brain activity pattern —or reconstruction of the stimuli themselves using generative models. Recent literature has shown significant progress in fMRI-based image decoding [236, 43, 78, 79, 250, 190], particularly for image retrieval and reconstruction. Most methods are variations of the following concept: the brain creates a representation of external stimuli, while we can obtain external stimulus representation (i.e. image embeddings) using a computational model, then learning a mapping between these two representations completes the decoding pipeline. These methods differ e.g. in the generative models used, in how the models are conditioned, and in the techniques used to compute the mapping between brain activity and embeddings—ranging from linear layers to neural networks. However, they all revolve around the central idea that the brain computes something analogous to image embeddings (ideally a representation in a manifold homomorphic the the model's one) which can be captured via fMRI measurements, and that with sufficient data, a mapping between brain activity and computational model-derived embeddings can effectively link brain activity to external stimuli. This concept also underpins language encoding and decoding [7, 251, 123], where large language model embeddings serve as surrogates

to approximate, through linear layers, the language processing which occurs in the brain during listening and comprehension. Similar approaches have shown promising results in decoding music from brain activity [58, 74]. The closest related work in this domain is represented by [42, 248], which directly addresses the problem of video reconstruction from fMRI data using a different dataset from the on e,ployed here [271]. This latter dataset comprises 18 training videos (each 8 minutes long) and 5 test videos of the same length, collected from 3 subjects. The approach in [42, 248] is based on subject-specific semantic mapping between fMRI data and Contrastive Language Image Pretraining (CLIP) embeddings, along with attention based modules to condition generative models for generating temporally coherent images to reconstruct videos. In this work, we approach video decoding from a different perspective. We use a rich dataset [47] of 1000 short video stimuli and concomitant fMRI data to build cross-subject models [80] for decoding through video retrieval from fMRI data. We hypothesize that video processing in the brain can be decomposed into three distinct streams: a visual stream (recognizing shapes, patterns, and objects in images), a semantic stream (understanding what is happening in the video), and an audio stream, which provides multisensory integration that aids in video comprehension. Based on this hypothesis, our pipeline consists of three main components: First, we construct subject-specific encoding models that predict brain activity from modality-specific embeddings (video, text, and audio) to identify responsive brain regions (regions of interest, ROIs) for each modality. Next, we perform functional alignment [98] to create a robust training and testing set across subjects. Finally, we develop a set of modality-specific decoding models that estimate embeddings from brain activity, which can be used for video retrieval. We demonstrate how multistream integration enhances decoding performance and provide examples at `https://mind2music.my.canva.site/decoding-video-nips-sito`. Figures 8.1 and 8.2 depict our decoding and encoding frameworks, respectively.

8.2 Material and Methods

We analyzed a public available fMRI dataset acquired while ten subjects watched a video, originally released as part of the Algonauts 2021 challenge [47, 145]. In the main experiment, participants viewed 1,000 training videos and 102 testing videos multiple times, all presented without audio. MRI data were collected using a 3T Siemens Trio scanner, and preprocessing included standard fMRI procedures such as slice time correction and normalization. The data were part of the Algonauts 2021 challenge, focusing on 1,000 fMRI-video pairs. For more detailed information about the data collection and preprocessing, please refer to the appendix and the original article [47]. Our primary objective was to

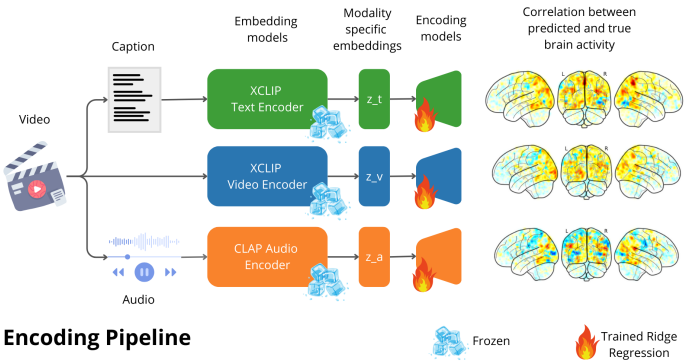


Figure 8.2: This figure shows the encoding pipeline that processes video data into modality-specific embeddings and maps them onto brain activity. Text captions, video frames, and audio are extracted and processed using pretrained models: XCLIP for text and video, and CLAP for audio, generating embeddings (z_t, z_v, z_a). These embeddings are then used in Ridge regression models to predict brain activity. The final step correlates the predicted brain activity with the actual brain data, visualized as brain maps. The frozen symbols indicate that encoder parameters are fixed during training, while Ridge regression models are trained for each modality.

identify brain regions responsive to identify potentially distinct brain regions responsive to each specific modality. We extracted captions from videos using a video captioning model [266], along with video frames and audio. These stimuli representations (semantic, visual, and audio) were processed using pretrained computational models to obtain modality-specific embeddings. Transformer-based models, such as XCLIP [165, 210], were used to extract video and text embeddings, while CLAP [71] was employed to obtain audio embeddings. We then modeled the mapping between brain activity and embeddings using Ridge regression. For each modality, the encoding models took as input embeddings with a dimensionality of 512 and projected them onto estimated brain activity, which had a dimensionality corresponding to the number of voxels (ranging from 10,836 to 21,573, depending on the subject). All models were subject-specific and trained using nested 5-fold cross-validation to prevent any potential circularity with the decoding models. The inner loop was used for hyperparameter optimization, while the outer loop predicted held-out data from the training set. Once the full training set was predicted, Pearson correlation was computed along the samples dimension, resulting in a voxel-wise map of correlations between predicted and actual brain activity for each modality. A threshold of 0.15, determined empirically, was then applied to create a ROI for each modality, which was used in subsequent analyses.We found "activations" in visual, auditory, language, and multimodal integration brain areas corresponds

to video, audio, and text inputs, indicating distributed processing across both unimodal and multimodal regions. To account for variability in brain structure and function across individuals, we used a modality-specific functional alignment strategy within a 5-fold nested cross-validation framework. This approach aligns brain activity patterns across subjects for each modality (semantic, visual, and audio), maximizing the similarity of functional responses while retaining modality-specific information. Following recent literature [80, 54, 17], we employed ridge regression as a regularization technique to improve robustness and generalizability by addressing multicollinearity in the high-dimensional fMRI data. This alignment method aimed to enhance the performance of our decoding models across multiple subjects. The above procedures resulted, for each modality, in a set of 9,000 training video-fMRI pairs (900 pairs per subject) and 1,000 testing pairs (100 pairs per subject). We then developed modality-specific decoding models by training cross-validated Ridge regression models within each identified ROI. These models were used to estimate the modality embeddings from brain activity (e.g., estimating text embeddings from regions responsive to text stimuli, and similarly for visual and audio modalities). For video retrieval, we employed a Euclidean search strategy, selecting the top-N closest test videos based on the L2 distance between the estimated and true embeddings for each modality. Additionally, we implemented a modality integration-based search by concatenating embeddings from different modalities, thereby allowing for a more comprehensive retrieval process that leverages the combined information from multiple sensory streams. As evaluation we report three metric useful to identity quality of retrieval and decoding. The first two are Top-1 and Top-5 accuracy, which simply count how many times the first retrieved videos is exactly the stimulus or when the the stimulus is correctly retrieved among the first 5 five retrieved videos. To complete the analysis and evaluate the quality of the decoded embeddings, we also report the identification accuracy, originally defined in [250]. This metric is a pairwise measure based on correlation, where a value of 0.5 indicates random predictions and a value of 1 signifies perfect predictions. The identification accuracy counts the number of times the estimated embeddings correlate more strongly with the true embedding than with other embeddings. This metric provides a robust evaluation of the quality of the estimated embeddings, which is crucial for more complex tasks such as generation.

8.3 Results

The results, summarized in Fig. 8.3 and Tables 8.2, 8.1, suggest that multimodal integration significantly enhances decoding and retrieval performance.

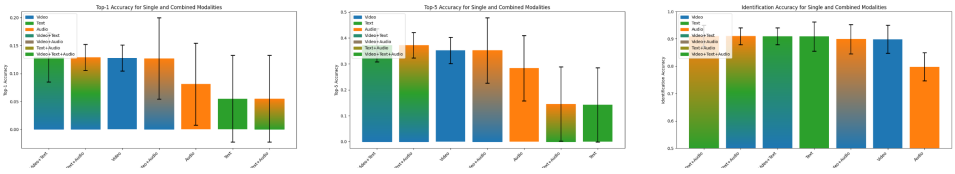


Figure 8.3: This figure presents three key performance metrics evaluating the effectiveness of decoding models across different single and combined modalities (Video, Text, Audio, Video+Text, Video+Audio, Text+Audio, and Video+Text+Audio). From left to right: top1, top5 and identification accuracy.

On average, the Video+Text+Audio combination achieved the highest identification accuracy, around 0.94, consistently outperforming other combinations. Video+Text and Video+Audio combinations also performed well, with average accuracies of 0.92 and 0.91, respectively. All combinations performed above chance level, highlighting the effectiveness of multimodal integration. In terms of video retrieval, the Video+Text+Audio combination again provided the highest average Top-1 and Top-5 accuracies, confirming that combining multiple modalities provides best results for decoding and retrieval tasks.

8.4 Discussion and Conclusions

In this study, we explored the advantages of a multi-stream approach for decoding brain activity, focusing on integrating Video, Text, and Audio modalities. Combining multiple modalities enhances decoding and retrieval performance compared to using individual modalities. The Video+Text+Audio combination consistently outperformed other combinations, indicating that this multi-stream integration captures more comprehensive and complementary information. Interestingly, the audio modality, despite the fact that videos were presented without sound, still performed above chance level in both identification and retrieval tasks. This suggests that the brain may integrate audio-visual information even in the absence of one modality, aligning with findings from hearing and optical illusions where the brain uses available sensory data to construct a complete perceptual experience. This observation underlines the complex and interconnected nature of sensory processing in the brain, where information from one modality can influence the processing of another, even when it is not directly presented. This study in not without certain limitations., such as the reliance on pre-trained models for generating embeddings, which may not fully capture the specific neural representations relevant to each participant. Looking ahead, a promising direction for future work involves the development of generative

models that can not only decode brain activity into embeddings but also reconstruct the original stimuli, such as generating audio-visual content from brain data. This would allow for a more direct and intuitive understanding of how the brain encodes and processes complex sensory information. Additionally, exploring the integration of more sophisticated multi-modal models, potentially leveraging advances in foundation models and deep learning, could further improve the accuracy and applicability of brain decoding techniques in real-world scenarios. As we advance in the field of brain decoding, it is crucial to consider the implications of neural privacy. While decoding and reconstructing stimuli from brain activity can provide valuable insights into cognitive processes, it also raises concerns about the potential misuse of such technology. It is essential to distinguish between genuine brain responses and the reconstructions generated by models, which may contain biases introduced by the algorithms themselves. These biases could lead to misinterpretations or unintended consequences when decoding sensitive or personal neural data. Therefore, developing ethical guidelines and robust safeguards to protect neural privacy is imperative as this technology continues to evolve. This includes ensuring that decoding models are transparent, interpretable, and used responsibly, with a clear understanding of their limitations and the potential risks associated with their application.

Appendix

8.4.a Data details

The study involved ten participants, each undergoing five scanning sessions. The first session was a localizer experiment where participants passively viewed short videos (unrelated to the main experiment) to define visual ROIs. The remaining four sessions constituted the main experiment, where participants focused on a fixation cross while viewing 3-second training and testing videos without audio. Videos were presented in 13 runs per session, each lasting about 7 minutes. By the end, participants had viewed 1,000 training videos three times and another set of 102 videos ten times. Using data from the localizer experiment, nine non-overlapping ROIs were defined for each participant, covering regions from early visual cortex (V1, V2, V3, V4) to higher-level areas responding to objects and categories (EBA, FFA, STS, LOC, PPA). MRI data were collected on a 3T Siemens Trio scanner with consistent acquisition parameters across all sessions (TR = 1750 ms, resolution = 2.5 mm^3, 54 slices, multi-band factor = 2). Preprocessing was done using fMRIprep, including slice time correction, realignment, co-registration, and normalization to MNI space. Data were interpolated from TR = 1750 ms to 1000 ms using the pchip method. FIR basis functions modeled the BOLD signal for each voxel, extracting beta values from 5 to 9 seconds post-video onset. These were averaged across time and used in subsequent analyses. Our study focused on data from the Algonauts 2021 challenge, specifically 1,000 fMRI-video pairs, with the first 900 used for training and the last 100 for testing.

8.4.b Subject performances detail

Subject	Video Identification Accuracy	Text Identification Accuracy	Audio Identification Accuracy	Video Top-1	Video Top-5	Text Top-1	Text Top-5	Audio Top-1	Audio Top-5
sub01	0.952	0.953	0.881	0.300	0.570	0.110	0.260	0.190	0.450
sub02	0.934	0.899	0.805	0.160	0.440	0.050	0.160	0.040	0.240
sub03	0.949	0.922	0.805	0.160	0.460	0.070	0.130	0.070	0.280
sub04	0.916	0.941	0.751	0.100	0.340	0.050	0.120	0.070	0.250
sub05	0.885	0.891	0.785	0.080	0.270	0.050	0.130	0.050	0.260
sub06	0.915	0.893	0.835	0.150	0.370	0.030	0.130	0.100	0.300
sub07	0.822	0.870	0.810	0.040	0.170	0.030	0.070	0.100	0.300
sub08	0.796	0.869	0.745	0.070	0.200	0.060	0.110	0.070	0.240
sub09	0.940	0.952	0.781	0.130	0.430	0.060	0.160	0.080	0.290
sub10	0.878	0.900	0.780	0.090	0.270	0.040	0.150	0.040	0.230

Table 8.1: Single Modality Results

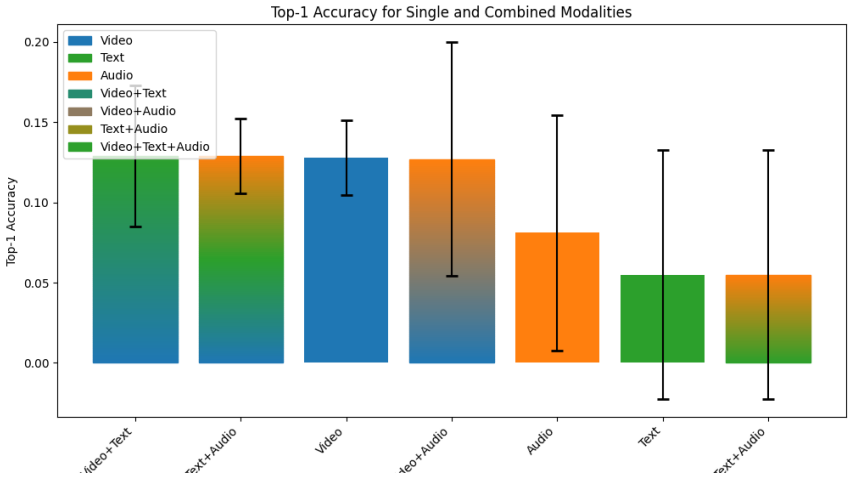


Figure 8.4: Identification accuracy, zoom of Right panel of fig 8.3

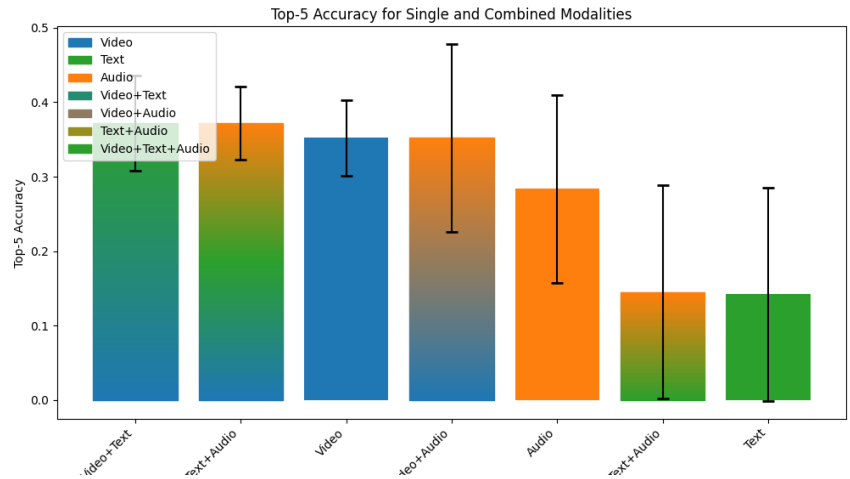


Figure 8.5: Top1 Accuracy, zoom of Left panel of fig 8.3

Subject	Video+Text Identification Accuracy	Video+Audio Identification Accuracy	Text+Audio Identification Accuracy	Video+Text+Audio Identification Accuracy	Video+Text Top-1	Video+Text Top-5	Video+Audio Top-1	Video+Audio Top-5	Text+Audio Top-1	Text+Audio Top-5	Video+Text+Audio Top-1	Video+Text+Audio Top-5
sub01	0.962	0.951	0.953	0.962	0.320	0.600	0.300	0.570	0.110	0.270	0.320	0.600
sub02	0.943	0.934	0.901	0.943	0.100	0.470	0.160	0.440	0.050	0.160	0.100	0.470
sub03	0.956	0.950	0.924	0.956	0.170	0.520	0.160	0.460	0.070	0.140	0.170	0.520
sub04	0.931	0.916	0.942	0931	0.100	0.370	0.090	0.340	0.050	0.130	0.100	0.370
sub05	0.895	0.885	0.893	0.895	0.110	0.300	0.080	0.270	0.050	0.130	0.110	0.300
sub06	0.923	0.916	0.896	0.923	0.130	0.370	0.150	0.370	0.030	0.140	0.130	0.370
sub07	0.834	0.823	0.873	0.834	0.040	0.140	0.040	0.170	0.030	0.070	0.040	0.140
sub08	0.813	0.796	0.870	0.813	0.070	0.210	0.070	0.200	0.060	0.110	0.070	0.210
sub09	0.950	0.940	0.953	0.950	0.160	0.460	0.130	0.430	0.060	0.150	0.160	0.460
sub10	0.889	0.879	0.901	0.890	0.090	0.280	0.090	0.270	0.040	0.150	0.090	0.280

Table 8.2: Mixed Modality Results

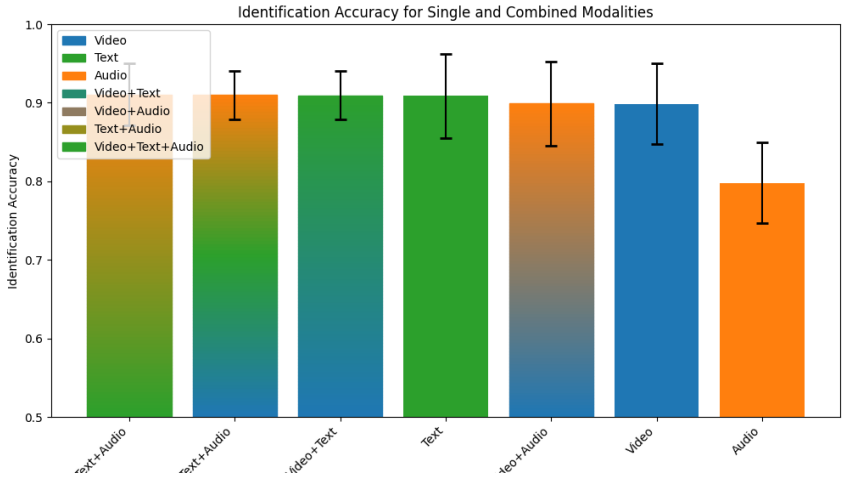


Figure 8.6: Top5 Accuracy, zoom of Center panel of fig 8.3

Figure 8.7: Average Cosine similarity matrix between true and predicted test embeddings.

Figure 8.8: Some examples video stimuli (first frame) and retrieved pool of candidates from mixed modality Video+Audio+Text, subject01

Figure 8.9: Some examples video (first frame) stimuli and retrieved pool of candidates from mixed modality Video+Audio+Text, subject01

Figure 8.10: Some examples video stimuli (first frame) and retrieved pool of candidates from mixed modality Video+Audio+Text, subject01

Part V

Decoding models as probe for neuroscientific experiments

9 Brain Algebra

Electrophysiological and neuroimaging studies have revealed how the brain encodes various visual categories and concepts. An open question is how combinations of multiple visual concepts are represented in terms of the component brain patterns: are brain responses to individual concepts composed according to algebraic rules? This chapter[1] is dedicated to the exploration of this question. To this end, we generated "conceptual perturbations" in neural space by averaging fMRI responses to images with a shared concept (e.g., "winter" or "summer"). After thresholding to ensure specificity, we applied these perturbations to the neural pattern associated with a base image, forming new brain patterns that incorporate the added concept. These modified brain patterns were then decoded into images using a pretrained fMRI-to-image decoding model.

Qualitative and quantitative inspection of the resulting images provides insight into how the brain might combine visual concepts. For example, adding a "winter" perturbation to the brain pattern of a man on a skateboard yields a new pattern representing a man on a snowboard in a winter scene—even when the perturbation modifies only a small subset of voxels. Our findings reveal that compositional processes in neural representations may lead to predictable perceptual outcomes, as interpreted by our decoding model. This suggests that the brain's combinatory encoding of concepts may follow a systematic, algebraic-like process—what we term "brain algebra." Although our study is model-driven, it opens avenues for future empirical work into the mechanisms of compositionality in the brain.

[1]The work presented in this chapter has been accepted for presentation at Cosyne 2025. Full manuscript is currently under submission to a peer reviewed journal.

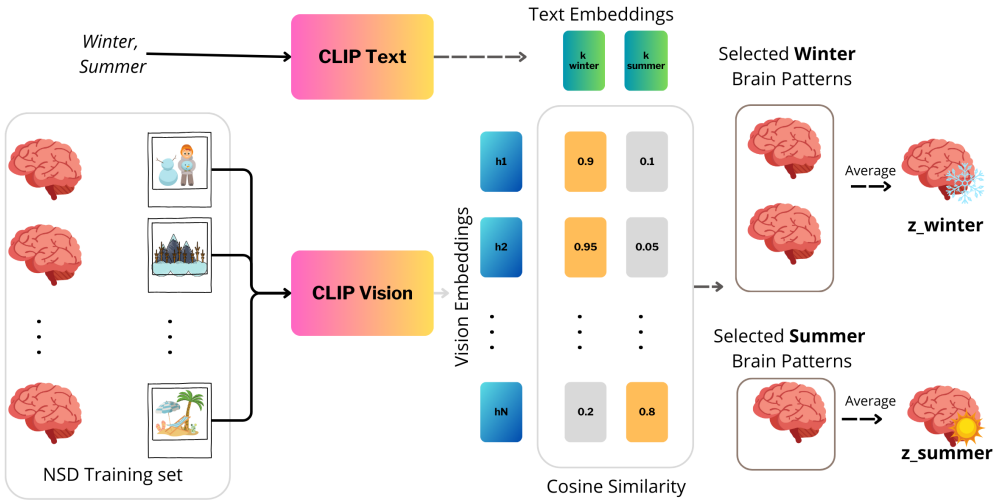


Figure 9.1: A visual overview of the process used to define conceptual perturbation vectors based on "winter" and "summer" concepts. First, the textual representations of the concepts are encoded using the CLIP Text model, generating 512-dimensional embeddings. Simultaneously, images from the NSD training set are processed through the CLIP Vision model to obtain vision embeddings. Cosine similarity is calculated between the vision embeddings and the textual concept embeddings (e.g., "winter" and "summer"), allowing us to select the top-matching images that best represent each concept. The fMRI patterns corresponding to these selected images are then averaged to generate concept-specific perturbation patterns in brain space, such as z_{winter} for winter and z_{summer} for summer. These perturbation vectors are later thresholded, and combined with base fMRI patterns in the brain algebra framework to modulate visual representations (see Figure 9.2)

9.1 Introduction

The compositionality of latent representations in artificial intelligence (AI) systems has contributed to recent advancements in deep learning. Model-based techniques like word embeddings have demonstrated that semantic relationships between concepts can be captured through vector arithmetic—for example, "king" minus "man" plus "woman" yields a vector close to "queen" [173]. Similarly, image and text representations in AI models exhibit compositional properties that allow for the manipulation and combination of visual and semantic concepts [153, 146, 229, 238, 209, 91].

In neuroscience, machine learning has spurred significant progress through the development of encoding and decoding models. These models have established bidirectional mappings between visual or linguistic inputs and corresponding brain activity [188, 236, 190, 34, 251, 123, 79, 78, 80, 7, 33, 250, 43, 158, 74]. Notably, the use of larger, multimodal, and more complex models—which

often exhibit some amount of compositionality—has led to significant improvements in predicting brain activity with encoding models. This suggests that better embeddings, enriched with compositional properties, capture more nuanced information that aligns more closely with the brain's representations [46, 6, 86]. This raises a fundamental question: **does the human brain employ a similar compositional structure in its neural code for vision** [201]? Recent evidence suggests that brain-pattern compositionality may indeed occur in specific linguistic contexts [276]: information regarding analogy questions can be effectively retrieved through the addition and subtraction of functional Magnetic Resonance Imaging (fMRI) patterns. In their study, participants were presented with sequences of related concepts, such as professions, tools, and places (e.g., "doctor", "stethoscope", "hospital"). The researchers demonstrated that the algebraic combination of fMRI activation patterns could reflect analogical reasoning, akin to vector operations in word embeddings (e.g. "mechanic-doctor+stethoscope=wrench").Moreover, the vector space representations utilized by AI models appear to exhibit key properties essential for supporting cognition, such as high-dimensional representations, compositionality, concept distances, and similarity measures [201].

Building on these findings, we investigate whether **brain-pattern compositionality holds for visual representations**. Specifically, we aim to determine whether the neural activity elicited by viewing a composite image can be approximated by algebraically combining the neural patterns associated with its constituent parts—a concept we refer to as **"brain algebra."** In this framework, adding the neural pattern associated with a particular concept to the neural pattern of a base image should yield a new neural pattern corresponding to the perception of the base image modified by the added concept. This idea mirrors the compositional operations observed in AI models, where vector arithmetic in latent spaces captures semantic relationships between concepts. By testing this hypothesis in the context of visual perception, we aim to uncover whether the brain employs a similar mechanism for combining visual information.

To address this question, we use the Natural Scenes Dataset (NSD) [5], a large-scale fMRI dataset where participants viewed approximately 10,000 natural images while their brain activity was recorded using a 7T fMRI scanner. We define "base" brain patterns as the fMRI responses to individual test images from this dataset. We also define "concept" brain patterns, where we average the fMRI responses to multiple training set images that share a specific concept. The presence of these concepts in training images is identified using semantic embedding models (i.e., CLIP; [210]), ensuring that the selected images are strongly related, from a semantic point of view, to the target concept. By algebraically combining

these patterns in brain space, we create a perturbed brain pattern that hypothetically represents the base image with the added concept. This is done by adding the thresholded concept pattern to the base pattern. For example, starting with a base brain pattern corresponding to *an image of a man on a skateboard*, we might add the concept pattern for "*winter*" to generate a perturbed pattern that should represent *a man on a snowboard during winter*.

A critical challenge is evaluating whether this perturbed brain pattern truly corresponds to the brain representation of the base image with the added concept. One approach would be to create corresponding composite images—for instance, using generative AI models to synthesize images that combine the base image with the added concept—and then present them to participants in an fMRI experiment. By recording the brain activity elicited by these composite images, we could compare the observed neural patterns with the predicted ones derived from our "brain algebra" operations (analogous to the approach used in [276]). However, this method presents several difficulties. First, the number of possible combinations of base images and concepts leads to a combinatorial explosion in the number of stimuli required. Testing a wide range of concepts and their combinations would necessitate a prohibitive number of experimental trials. fMRI experiments are inherently long and expensive, with limitations on how long participants can be scanned and associated financial costs with data acquisition. These logistical constraints make it impractical to collect sufficient data to robustly evaluate compositionality across diverse concepts. Second, even if such extensive experiments were feasible, interpreting the results would remain challenging. Differences between the expected and observed brain patterns could arise from various sources, including fMRI noise, individual variability in neural responses, or limitations in the quality and realism of the generated stimuli. These confounding factors would challenge any definitive conclusions about compositionality in the brain's neural code.

Additionally, existing neurostimulation technologies do not currently permit to directly manipulate specific voxel activations in the brain to test compositionality. We cannot selectively stimulate or alter precise patterns of neural activity at the voxel level to create the exact perturbed brain patterns hypothesized by our "brain algebra" model. This limitation means that we cannot empirically test the predicted neural patterns by artificially inducing them in the brain.

Given these challenges, we employ an alternative approach that evaluates the "brain algebra" results using a decoding model. By transforming the perturbed brain patterns into reconstructed images, we can indirectly assess whether the algebraic combination of neural patterns corresponds to meaningful composite perceptions. This method leverages existing fMRI data and advanced decoding

algorithms to infer the perceptual content associated with the combined neural patterns, providing a practical lens to explore our research question within the constraints of current technology.

Thus, instead of attempting to collect new fMRI data or manipulating brain activity directly, we employ Brain-Diffuser [190], a well-established decoding model that maps brain activity into the latent space of a generative model. By decoding the perturbed brain patterns, we can reconstruct images that represent the hypothetical perception resulting from our brain algebra operations. The resulting images can be assessed qualitatively—through visual inspection—or quantitatively using automated semantic analysis tools based on AI systems like CLIP [209].

In summary, our study explores whether the compositionality observed in AI systems and linguistic brain representations extends to visual processing in the human brain. We introduce a novel method to assess "brain algebra," combining base and concept brain patterns derived from actual fMRI data, and employ the Brain-Diffuser model to decode these patterns into images. Through this method, we aim to provide evidence for or against the existence of compositional neural codes in visual cognition. In the following sections, we detail our methodology for defining and combining the base and concept brain patterns, describe how we employ the Brain-Diffuser model for decoding, and present our qualitative and quantitative analyses of the reconstructed images to evaluate compositionality. See Fig 9.1 for a visual explanation of perturbation definition and Fig 9.2 for a visual overview of our approach.

9.2 Results

We explored 12 different semantic concepts encompassing themes such as season (winter, summer), gender (man, woman), lighting (night, day), numerosity (empty, crowded), location (indoor, outdoor) and emotions (happy, sad). Each corresponding perturbation vector was thresholded by a variable amount (retaining between 5% and 100% of the voxels, the rest being set to zero), and scaled by various factors α (from 1 to 4) before being summed with base fMRI patterns corresponding to a random subset composed of 100 test images. We begin this section by discussing the qualitative results, focusing on the exemplary figures 9.3 and 9.4, which were generated using a scaling factor of $\alpha = 2$ and a 50% threshold to visually highlight the key outcomes. Additional figures with varying scaling values and thresholds are provided in the supplementary materials for a more comprehensive evaluation.

In the first set of images (top of Fig 9.3), showing horses in a field and an indoor

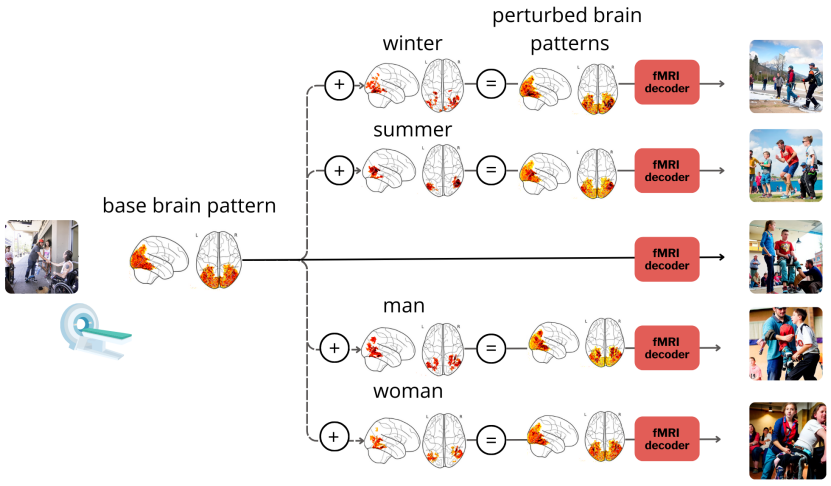


Figure 9.2: Illustration of the "brain algebra" approach used in our study. The leftmost image represents the initial visual stimulus presented to the participant, with corresponding fMRI activations shown as heatmaps across different brain regions. Perturbations are introduced by summing the base brain pattern with a concept-specific perturbation vector, such as "summer," "winter," "man," or "woman." The perturbation vector is computed as a thresholded average of brain patterns evoked by visual perception of images with that content (see Figure 9.1). The perturbed brain patterns (center) are subsequently decoded using a pretrained fMRI decoder, producing modified images that reflect the added conceptual information (right). The results demonstrate how small changes in neural patterns can lead to predictable and meaningful changes in visual perception, supporting the hypothesis of compositionality in neural representations.

Figure 9.3: Qualitative evaluation of brain algebra perturbations applied to base images (images best viewed digitally). Starting from the central base images, decoded from a (non-perturbed) base fMRI pattern, perturbations corresponding to various concepts—such as "summer," "winter," "day," "night," "man," "woman," and more—are applied to the brain patterns. The resulting decoded images show how the base visual perception is altered by the addition of each conceptual perturbation, reflecting changes in environmental conditions, the presence or absence of people, and other context-specific details. This demonstrates the ability of brain algebra to generate compositional modifications in visual representations based on abstract conceptual inputs.

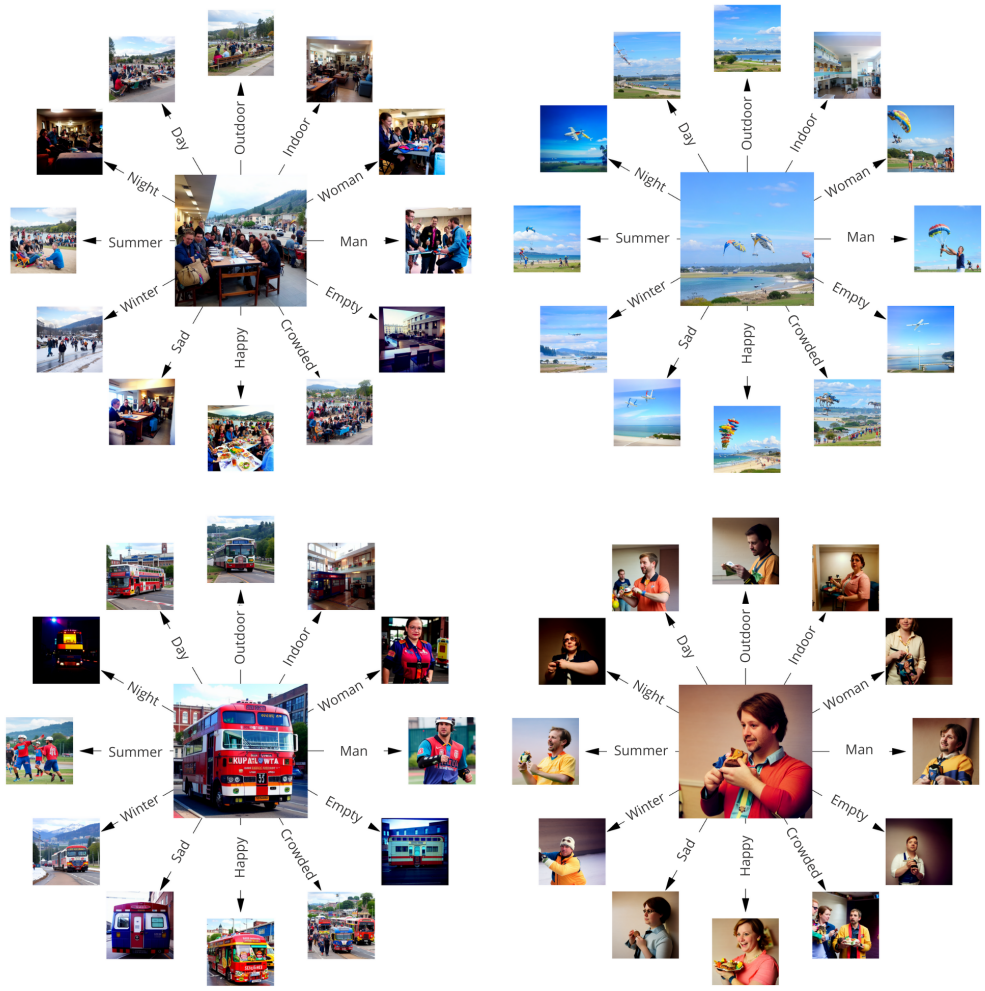


Figure 9.4: More examples as in Fig 9.3

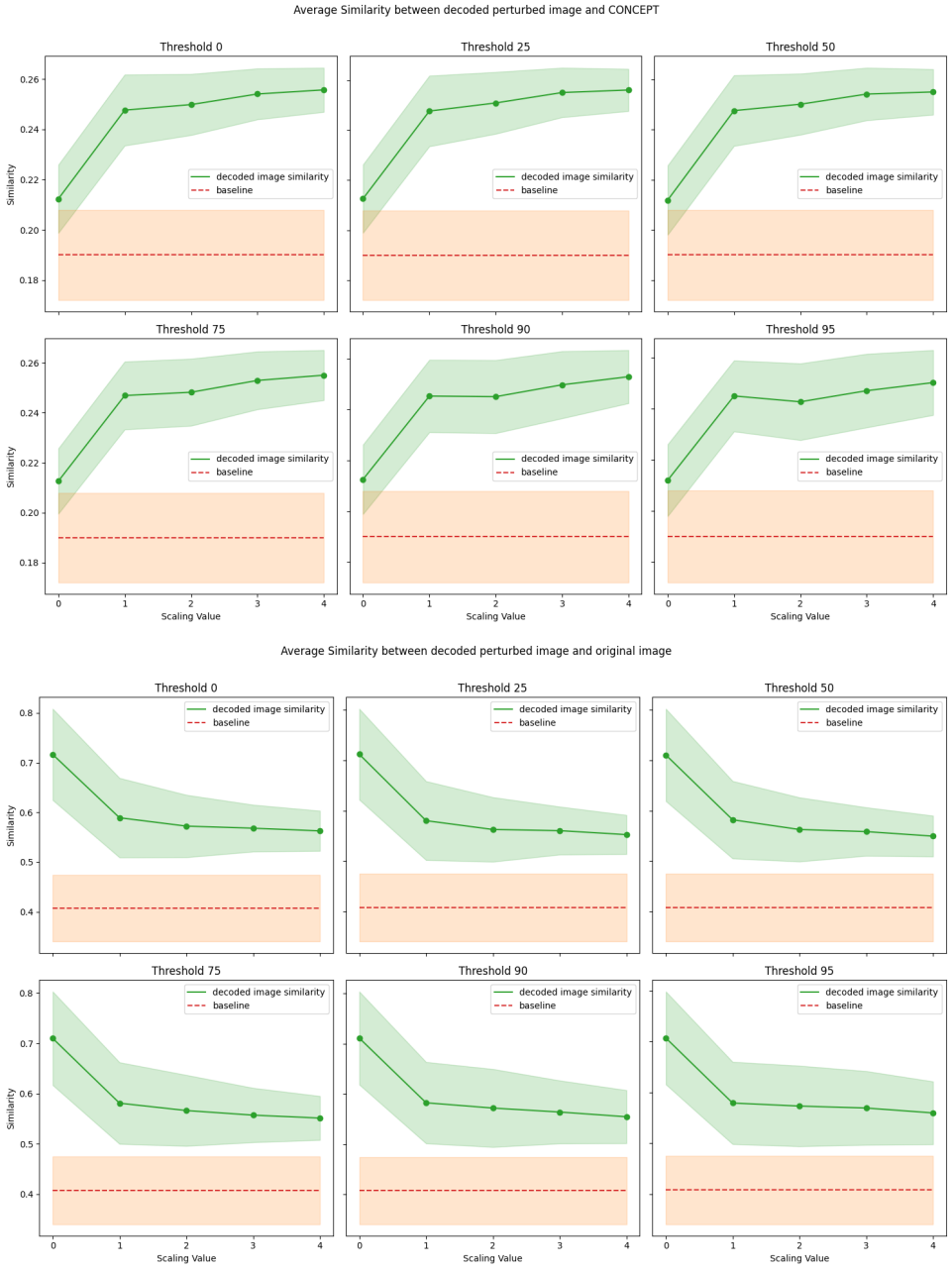


Figure 9.5: Top: Average similarity between decoded perturbed images and the target concept (across CLIP-vision and CLIP-text latent spaces, respectively) across different thresholds (0, 25, 50, 75, 90, 95th percentiles of the voxel distribution) and scaling values (0 to 4, with zero corresponding to the base pattern without perturbation) . The similarity increases with higher scaling values, reflecting that larger perturbations align the decoded image more closely with the target concept. The green shaded region represents variability (standard deviations across 4 subjects, 100 images per concept, 12 concepts) in similarity, while the red dashed line represents the baseline similarity between random images and the target concept. **Bottom**: Average similarity (in the CLIP-vision latent space) between decoded perturbed images and the original images across the same thresholds and scaling values. The similarity decreases as the scaling value increases, indicating that larger perturbations deviate more from the original image. The red dashed line shows the baseline similarity between the original image and random images.These results indicate a trade-off between maintaining original image features and introducing conceptual modifications, depending on the scaling value and threshold.

bathroom, compositionality is evident. Concepts like "summer," "winter," and "night" alter the landscape and lighting in the horse scene, while "woman" and "man" introduce an additional person. In the bathroom scene, "crowded" and "empty" adjust the number of objects or people, and "summer" and "winter" change the mood. The bottom row, showing a skater and a social gathering, also demonstrates compositional transformations. "Summer" and "winter" modify the environment, "woman" changes the skater's gender, and "crowded" and "happy" alter social dynamics and expressions.

Similarly, perturbations in Figure 9.4 introduce clear modifications based on the perturbation concepts. For the bus scene, "summer" brings brighter environments with outdoor activities (however, the bus is no longer present), while "night" darkens the scene with illuminated elements. The "indoor scene" concept transforms the bus into a more enclosed space, like a terminal. Perturbations applied to the man holding a sandwich show clear changes—"woman" alters the subject's gender, "empty" reduces the person's size within the scene, and "crowded" adds individuals to the scene. In the outdoor table scene and paragliding activity, compositional adjustments are also evident. These results suggest that the model effectively generates coherent and predictable changes in response to targeted brain perturbations. Overall, these results support the hypothesis that compositionality in brain patterns can be decoded into visually meaningful images. The perturbations introduced lead to expected modifications in both human-related and environmental aspects, while generally maintaining the integrity of the original base scenes. This indicates that the brain perturbation patterns successfully capture abstract conceptual information and translate it into visual content, providing strong evidence for compositionality in the brain's visual representations.

While these qualitative examples generally support our hypothesis of visual concept compositionality, they also include some perturbations for which the desired concept did not obviously appear or was difficult to evaluate (e.g. paragliding+day appears similar to the base image stimulus), and perturbations that replaced the initial content rather than complementing it (such as the disappearing bus in the bus+summer perturbation). To provide a more systematic evaluation of compositionality, we quantitatively measured the presence of both the perturbation concept and the base image concepts in the decoded images from perturbed brain patterns, by leveraging cosine similarity in the CLIP latent space as a measure of semantic content (since this metric is shown to be aligned with human judgments on image similarity [108, 177]).

This quantitative evaluation of the decoded perturbed images reveals clear trends in the similarity between the images and two targets: the original image

and the concept used for perturbation. In the bottom panel of Fig 9.5, which assesses the similarity (in the CLIP-vision latent space) between the decoded perturbed images and the original images, we observe that similarity decreases as the scaling value increases. This is expected, since larger scaling values apply a stronger perturbation, causing more deviation from the original image. The similarity between the decoded perturbed images and the original images remains significantly above the baseline for all conditions (calculated by contrasting decoded images against randomly chosen base images), indicating that elements of the original image are still retained even as the perturbation intensifies.

The top panel of Fig 9.5, which compares the decoded perturbed images to the target concept (across CLIP-vision and CLIP-text latent spaces, respectively), shows the opposite trend: similarity increases as the scaling value rises. This suggests that larger scaling values make the images more representative of the added concept. Lower thresholds allow the perturbation to have a broader effect, leading to faster increases in similarity to the target concept, while higher thresholds limit the impact of the perturbation, resulting in slightly more modest increases. In all cases, the similarity between the decoded images and the target concept remains consistently above the baseline (calculated by contrasting random images with the target concept), demonstrating that the perturbation effectively introduces the desired conceptual information into the images. Interestingly, even at high thresholds, where only a small portion of the brain's activity is perturbed, we still observe notable changes in similarity to the target concept. The fact that meaningful conceptual shifts occur even when the perturbation is restricted to higher thresholds indicates that the brain might encode these abstract concepts in localized areas, and only subtle changes in activity within these regions are required to reflect concept-driven modifications in the decoded images. Overall, these results illustrate a trade-off between maintaining similarity to the original image and introducing conceptual modifications. As scaling increases, the perturbed images deviate more from the original content but become more aligned with the target concept. The thresholding mechanism provides a way to control the extent of the perturbation, with higher thresholds preserving more of the original image and lower thresholds allowing for greater conceptual compositionality.

9.3 Discussion

The findings from this study provide a promising indication of compositionality in neural representations. By manipulating brain patterns in a "brain algebra" framework—combining a base neural state with a thresholded and scaled

perturbation vector—we observed distinct, meaningful changes in the decoded images. This suggests that visual processing in the brain may follow a compositional structure, much like language, where basic elements can be combined to create more complex representations. The ability to successfully decode these perturbations aligns with broader theories on compositionality in cognition, such as in language, where concepts are combined to produce new meanings (e.g., "queen" = "king" - "man" + "woman"). This parallel between vision and language highlights how the brain may generalize compositional principles across different domains of cognition, supporting flexible and dynamic perceptual and cognitive processes [276, 229, 153].

The use of natural images as stimuli is integral to our approach, as it enables the study of brain patterns in conditions that closely mimic real-world visual experiences. These images contain a variety of visual and semantic elements that reflect everyday interactions with the environment. By leveraging these stimuli, we can investigate how the brain processes complex compositional patterns that are more representative of natural vision, compared to more controlled or artificial stimuli.

There are, however, important limitations to consider. While our results suggest compositionality in neural representations, our evaluation relies on a decoding model to interpret the perturbed brain patterns. Although the two brain patterns we are combining are derived from actual fMRI data, the interpretation of their combination is model-driven because it depends on the decoder's ability to accurately reconstruct images from brain activity. This means that our conclusions are contingent upon the performance and limitations of the decoding model, and we are not directly observing brain processes in real-time but interpreting them through the lens of the model. This limits the extent to which we can confirm that similar compositional operations happen naturally in the brain. Additionally, we are constrained by the need for sufficient training data—only concepts with ample representation in the training set allow us to generate reliable perturbation vectors. As a result, our exploration of neural compositionality is bounded by the availability of data, limiting the range of concepts we can examine. Furthermore, some concepts are not orthogonal, and their representations in the training set can lead to biased perturbations in brain patterns. For instance, visual inspection shows that adding the concept of "happiness" occasionally introduces food elements, likely because in datasets like COCO, "happiness" is often associated with, or co-occurs alongside, images of food. Similarly, concepts like numerosity or emotion might also be biased due to their frequent co-occurrence with humans. As a result, applying a "crowded" perturbation may add humans to scenes with animals, even when the intended

effect is only to increase the number of animals.

One intriguing aspect of our results is that meaningful changes were observed even when the perturbation involved only a small subset of voxels (e.g., the top 5% of voxels). This suggests that relatively small, localized regions of the brain can significantly influence the representation of specific concepts. This finding contributes to the ongoing debate between distributed and localized cortical representations in visual processing [257, 217, 96]. On the one hand, proponents of distributed representations, such as Haxby and colleagues, argue that visual information is encoded across widespread patterns of neural activity, with object and category information represented in distributed and overlapping voxel patterns [101]. On the other hand, researchers like Kanwisher propose that certain visual categories are processed in specialized, localized cortical regions—for example, the fusiform face area (FFA) for face perception [130]. Our observation that concept perturbations are effective even when modifying only a small portion of voxels aligns with the idea that specific cortical areas play a crucial role in representing certain concepts. However, the fact that these perturbations can alter perception suggests that these localized regions are integrated within broader distributed networks. This highlights the complexity of neural coding in the brain, suggesting that both localized and distributed representations contribute to the richness of visual perception.

In summary, our findings provide evidence suggesting compositionality in the brain's processing of visual stimuli. This compositional structure could be key to understanding how the brain flexibly combines sensory information to form different percepts, furthering our understanding of neural coding, perception, and learning.

9.4 Conclusions

In this study, we explored the compositionality of neural representations through the novel framework of "brain algebra," combining base fMRI patterns with conceptual perturbations to decode visual representations. Our findings provide evidence that the brain may employ compositional mechanisms similar to those seen in language and cognition, where smaller elements are combined to form more complex representations. The results demonstrate that neural patterns can be manipulated to create distinct and predictable changes in decoded images, aligning with the target concepts, even when only small portions of brain activity are perturbed.

The findings suggest that neural representations of visual concepts involve both localized and distributed processing. The ability to change perceived im-

ages by modifying a small number of voxels indicates that certain brain regions are specialized for processing specific visual information, supporting the idea of localized specialization. However, these localized changes also integrate with broader neural networks, aligning with the view that visual perception involves distributed representations. This dual role of localized and distributed coding contributes to the debate in neuroscience and underscores the complexity of how the brain processes and combines visual information.

Overall, this work contributes to a deeper understanding of neural compositionality in vision and highlights the brain's capacity to integrate conceptual information. By offering new perspectives on how perturbations in neural space correspond to changes in perception, this research provides a foundation for future studies on neural coding, perceptual constancy, and cognitive flexibility. Expanding this line of inquiry could also reveal insights into how the brain composes and generalizes across other cognitive domains, such as language and broader semantic representations. This suggests that the principles underlying "brain algebra" in vision might extend to other brain functions, hence providing a more comprehensive framework for understanding compositional processes in neural representations.

9.5 Material and methods

In this section, we describe the proposed method and the data we used. The data are publicly available and can be requested at https://naturalscenesdataset.org/. All experiments and models were trained on a server equipped with eight NVIDIA A100 GPU cards (80GB RAM each connected through NVLINK) and 2 TB of System RAM.

9.5.a Data

The study employs the Natural Scenes Dataset (NSD) [5], an extensive fMRI dataset gathered from eight participants who were shown images from the COCO21 dataset. Our analysis focused on four subjects, resulting in a specialized training set containing 8,859 images and 24,980 fMRI trials per subject, as well as a shared dataset consisting of 982 images and 2,770 trials per subject. To reduce the spatial dimensionality of the fMRI signals (with a resolution of 1.8mm isotropic), we applied a mask using the provided NSDGeneral ROI, targeting multiple visual areas. This deliberate selection of ROIs improved the signal-to-noise ratio and reduced data complexity, enabling the investigation of both low- and high-level visual features. Temporal dimensionality was further minimized by leveraging precomputed betas derived from a general linear model (GLM; [205,

133]) with an adjusted hemodynamic response function (HRF) and a denoising process as detailed in the NSD publication.

9.5.b BrainDiffuser

The "Brain-Diffuser" model [190] is a two-stage framework designed to reconstruct natural scenes from fMRI signals. In the first stage, a Very Deep Variational Autoencoder (VDVAE) generates an "initial guess" of the reconstruction, capturing low-level details. This guess is then refined using high-level semantic features from CLIP-Text and CLIP-Vision models, and a latent diffusion model (Versatile Diffusion; [278]) is used for the final image generation. The model takes fMRI signals as input and produces reconstructed images that reflect both low-level properties and the overall scene layout. Brain-Diffuser, was trained subject-wise with data from Subj01, Subj02, Subj05, Subj07). More information about the decoding model is detailed in the original paper. While this specific pipeline is used in our study, our proposed method is universally applicable and can enhance any single-subject decoding pipeline. It offers a versatile, adaptable tool that can seamlessly integrate with novel, advanced pipelines. By focusing on preprocessing input data, our approach enables the underlying pipeline—regardless of its unique aspects—to effectively work with single-subject fMRI data to generate images, without requiring direct modifications to the pipeline itself.

9.5.c Main experiment

Here we outline our main experiment, which aims to decode synthetic brain patterns derived from the algebraic sum of real brain patterns, and examine compositionality in the decoded images. The NSD dataset provides paired fMRI data and corresponding images. Let's define fMRI data as z and images as x, giving us a training set of pairs (z_{tr}, x_{tr}) and a test set (z_{ts}, x_{ts}). Additionally, we have a decoder d, a function that maps z to x, such that $x \approx d(z)$.

The essence of our work is as follows: we explore the outcome when decoding z', defined as $z' = z_{base} + \alpha z_{perturb}$, where α is a scaling factor, z_{base} is a brain pattern drawn from the test set, and $z_{perturb}$ is a perturbation pattern computed by averaging training set brain patterns associated with a specific concept. One way to investigate compositionality in the brain is to hypothesize that the resulting image, $x' = d(z')$, represents a combination of the base image x_{base} and an additional concept.

For instance, if the base pattern z_{base} corresponds to an image of an indoor scene x_{base}, and we add a perturbation brain pattern $z_{perturb}$ related to the concept

of a *man*, then decoding z' might produce an image of a *man in an indoor scene*. Similarly, if the perturbation pattern corresponds to a *woman*, we might expect the decoded image to depict a *woman* in the scene, and so on. We tested our framework for a random subset (identical for all subjects) of 100 images drawn from the test set.

Pattern definition and thresholding

A natural question arises: how do we define the perturbation pattern relative to a specific concept? We adopted a straightforward approach by filtering the image-fMRI pairs in the training dataset based on their similarity to the concept, which is defined in natural language and measured using a CLIP-based cosine similarity.

We explored 12 different semantic concepts: ["man", "woman", "indoor", "outdoor", "summer", "winter", "day", "night", "crowded", "empty", "happy", "sad"]. These concepts encompass themes such as season, gender, lighting, numerosity, emotions.

Each concept was represented as a word and encoded using the CLIP Text model [210], producing a 512-dimensional representation (using version *clip-vit-base-patch32*). We then used the CLIP Vision encoder to encode all the images in the training set and calculated the cosine similarity between each image and all the concepts. This resulted in a similarity matrix of shape (8859, 12). For each concept, we selected the top 100 pairs with the highest similarity scores to extract the corresponding fMRI and image indices. The perturbation patterns were then defined by averaging the fMRI patterns associated with these top-100 pairs, thereby establishing their representation in brain space.

We applied various threshold values to the perturbation patterns based on their percentile values. Specifically, in our experiments, we evaluated the outcomes by thresholding $z_{perturb}$ and retaining only the values above the 0th, 25th, 50th, 75th, 90th, or 95th percentiles. This allows us to assess whether the representation of the chosen broad semantic concepts is distributed across the entire visual cortex or if only a small region is responsible for encoding changes, with the composition of values in these small regions being sufficient to produce compositional images when decoded.

9.5.d Evaluation

The first part of our evaluation is qualitative, focusing on visually assessing decoded images from perturbed brain patterns to examine the compositionality of the original stimulus image and the perturbation concept. However, a

quantitative measure is necessary to rigorously evaluate this compositionality. Compositionality can be loosely defined as the co-occurrence of two concepts within an image. Thus, we need to quantify how closely the decoded image resembles the target perturbation concept while retaining similarity to the original content. If the scaling factor α is too large, the perturbation may replace the original content entirely, leading to misleading results if we only measure the similarity between the decoded images and the perturbation concept.

To address this, we calculated the CLIP cosine similarity between the decoded perturbed images and the original stimuli to ensure that the original content was not entirely replaced. Simultaneously, we measured the CLIP cosine similarity between the decoded perturbed images and the perturbation concept. In the first case, we measured cosine similarity between images, while in the second case, the similarity was measured between images and text. As these two metrics may have different baselines, we also computed a random baseline by measuring the cosine similarity between each base image and 100 randomly selected images from the training set. Similarly, we established a baseline for each concept using 100 random images.

Finally, we averaged the results as a function of the scaling factor α for each threshold across all decoded images and subjects.

Part VI

Discussion

Discussion and conclusions

This thesis aims to address the question: *"What kind of information can be decoded from brain activity?"* by focusing on noninvasive measurements using functional MRI (fMRI) across various cognitive tasks, including vision, language, and music processing. Although each chapter explores specific questions in detail, a common theme throughout this work involves relating brain activity measured with fMRI during cognitive tasks to vector representations of the corresponding stimuli derived from computational models. One of the primary objectives of this thesis is to demonstrate that, with sufficient data and computational resources, it is possible to link these two representations, brain activity and model-derived features, using linear or non-linear models. Establishing this mapping not only enables effective decoding, but also provides a lens through which the brain's processing mechanisms can be explored more deeply.

The most widely analyzed modality in brain decoding studies is functional magnetic resonance imaging (fMRI). Its popularity stems from its noninvasive nature, high spatial resolution, which enables detailed analysis of brain activity patterns at a millimeter scale, and the recent availability of large-scale openly accessible fMRI datasets. However, fMRI comes with several challenges that require careful consideration. fMRI data capture whole brain activity with high spatial and temporal complexity, often necessitating feature engineering to enhance signal quality and remove noise. For example, identifying regions of interest (ROIs) through encoding models or predefined anatomical atlases can help focus the analysis on tasks-relevant brain areas. fMRI analysis also faces limitations in temporal resolution, which typically ranges from 0.5 to 1 Hz for cognitive tasks. This makes it difficult to capture the rapid dynamics of neural processing. In addition, fMRI measures the BOLD signal, which is an indirect, delayed, and temporally dispersed reflection of the underlying neural

activity rather than a direct measurement of electrical brain activity. To address these challenges, data from multiple acquisitions can be averaged to reduce noise, and computational approaches, such as general linear models (GLM), are frequently employed. These models account for different time points and delays introduced by the hemodynamic response function, allowing for a more accurate modeling of task-related brain activity. Together, these methods help mitigate the inherent limitations of fMRI and enhance its utility in decoding studies. Adding to these challenges, fMRI is prone to significant noise from various sources, including acquisition protocols, scanner hardware, subject head motion, and other experimental parameters. These factors complicate the task of isolating meaningful signals from random noise fluctuations. In the context of encoding and decoding, an additional layer of complexity arises from subject-specific anatomical and functional differences. These interindividual variations further increase the difficulty of developing robust models and accurately interpreting neural signals.

In an effort to tackle all these challenges, this thesis provides a comprehensive exploration of brain decoding, focusing on how neural activity represents cognitive processes across various sensory modalities and tasks, including vision, language, music, and video. The primary contributions of this work span the development of advanced computational frameworks that leverage noninvasive functional MRI data to decode meaningful information, address intersubject variability, and integrate multimodal representations to perform stimuli decoding from brain activity.

In Part II, visual decoding served as the foundation for this exploration, demonstrating the ability to classify unseen semantic categories and reconstruct images using semantic-based approaches. Using features derived from convolutional neural networks (CNNs) and multimodal vision transformers, these methods highlighted the hierarchical encoding of visual information in the brain and its alignment with computational representations. Chapter 1 introduced robust pipelines for semantic decoding using the GOD dataset, while Chapter 2 extended this to cross-modal decoding using the NSD dataset to generate linguistic conditioning diffusion models to generate context-rich images, finding that cross-modal decoding is possible probably due to visual and liguistic feature alignment in part of the brain [203]. One of the most significant contributions is presented in Part II, Chapter 3, where a Ridge Regression-based functional alignment approach was introduced to address intersubject variability, a critical challenge in decoding pipelines. This method enabled the alignment of neural representations between individuals, allowing robust decoding of one subject's brain activity using data from another. Remarkably, this approach required only

10% of the data typically needed for training, achieving cross-subject generalization without compromising decoding performance or image reconstruction quality. This advance represents a pivotal step toward scalable and generalizable decoding frameworks.

In Chapter 4, the thesis expanded beyond vision to include multi-modal decoding, integrating data from fMRI, EEG, and MEG. By employing a unified framework based on contrastive learning, this work demonstrated the feasibility of encoding, decoding, and modality conversion within a single model, highlighting its adaptability and effectiveness across diverse data types and cognitive tasks. The integration of multi-modal data further emphasizes the versatility of the proposed approaches.

Language decoding is explored in Part III, where both retrieval-based (Chapter 5) and generative (Chapter 6) approaches were developed to decode linguistic representations from brain activity. The retrieval-based approach utilized contrastive learning to identify sentence fragments, while the generative approach adapted large language models (LLMs) to fMRI data, enabling the reconstruction of context-aware sentences. These advancements represent a step toward developing non-invasive, direct brain-computer interfaces (BCIs) for language decoding, offering potential applications in assistive technologies.

In Part IV, the methods were extended to music and video decoding, demonstrating their applicability across diverse sensory modalities. Techniques such as functional alignment, region-of-interest (ROI) selection, direct Ridge mapping, and retrieval-based decoding were employed to decode auditory and video stimuli, further showcasing the generalizability of the proposed frameworks. Chapter 7 highlights the decoding of musical tracks, while Chapter 8 addresses video retrieval and reconstruction, integrating multi-stream approaches with audio, visual, and semantic information.

Finally, Part V moves beyond decoding for practical applications and introduces the concept of brain algebra to probe the mechanisms of neural computation and representation. By examining perturbations and transformations of brain activity, this work provides insights into how the brain processes, combines, and manipulates information at a macroscale level, offering novel perspectives on the brain's functional architecture.

Why semantic concepts as vectors?

In all of the works presented, the underlying hypothesis is that brain activity, or brain representations, can be mapped onto a vectorial representation of external stimuli. This assumption is aligned with the idea that the brain processes information about the world in a way that mirrors computational vector spaces.

Although the exact mechanisms by which the brain performs these computations is only partially understood, vector spaces provide a useful framework for modeling cognition. They possess critical properties, such as compositionality and continuity, that are necessary to support cognitive processes and reproduce human behavioral judgments [201].

The suitability of vectorial representations for modeling brain activity is further supported by their ability to encode complex semantic relationships. For example, in computational neuroscience and artificial intelligence, vectors can represent not only individual concepts but also their interactions, facilitating reasoning and semantic similarity tasks. This aligns with evidence that the brain employs a high-dimensional semantic space to organize and process information efficiently. As highlighted in recent discussions on why concepts are (probably) vectors [201], vector spaces exhibit properties like compositionality, allowing the representation of complex ideas as combinations of simpler elements, and flexibility in modeling graded or approximate similarities. These characteristics are essential for the brain's ability to generalize from limited data, infer relationships, and make predictions about unseen stimuli. This perspective complements the work in this thesis by suggesting that the brain's use of vector-like representations could reflect an evolutionarily optimized mechanism for managing the complexity of the external world. In the context of this work, vectorial representations serve as a bridge between computational models and brain activity. By aligning brain representations with stimulus-derived vectors, this thesis explores the assumption that the brain's processing can be mapped into a high-dimensional space where meaningful relationships are preserved. This approach not only facilitates decoding tasks, but also provides a framework to investigate how cognitive processes, such as language understanding and visual recognition, are grounded in neural computations of the brain. Specifically in Part V we strongly relied on this assumption of the brain that operates as a vector space in his semantic space and explored what happens when we perturb brain representations and move in this space to explore the complexity of visual concepts.

How is the information structured in the brain? Why do simple models such as Ridge regression appear to be so effective?

One of the key tools extensively used for brain decoding in this thesis is Ridge Regression, which may initially seem counterintuitive given the inherent complexity of the brain. The human brain comprises approximately 10^{11} neurons, each a non-linear computational unit. Scaling this complexity from the microscale (single neurons) to the mesoscale (populations of neurons) and finally

to the macroscale (whole brain activity), one might expect an increase in overall non-linearity. However, linear models such as Ridge Regression have shown remarkable efficacy in modeling macroscale brain activity, raising the question: *Why are such simple models effective for such a complex system?* The linearity observed at the macroscale is not a direct reflection of neuron-level dynamics but rather a result of spatiotemporal aggregation. As demonstrated in the study by Nozari et al. (2024) [186], the process of spatial and temporal averaging inherent in macroscopic brain recordings, such as functional magnetic resonance imaging (fMRI) and intracranial EEG (iEEG), plays a critical role in linearizing otherwise non-linear interactions at smaller scales. Spatial averaging, for instance, integrates signals across numerous neurons with diverse and nonlinear individual dynamics, effectively canceling out many of the nonlinear components. Similarly, temporal averaging, often implemented as low-pass filtering in preprocessing pipelines (or in the effect of the hemodynamic response function), smooths rapid fluctuations, further linearizing the observed signal. Both mechanisms act together to produce a dataset in which linear models often outperform or match the predictive performance of more complex nonlinear models. This linearizing effect has been empirically verified. For example, Nozari et al. (2024) found that linear autoregressive models consistently outperformed non-linear models in predicting the dynamics of the resting state in modalities and subjects [186]. Complementary findings by Popov et al. (2024) [204] further reinforce the utility of simple models (even models not aware of time dynamics such as recurrent neural networks of transformers), showing that a mean-based MLP architecture could rival or exceed the performance of more sophisticated machine learning models for classification tasks involving fMRI data. This suggests that much of the discriminative information in macroscale signals is embedded in spatially and temporally aggregated features rather than in dynamic temporal sequences. The implications of these findings are significant. They challenge the prevailing assumption that brain activity must be modeled with highly nonlinear systems and emphasize the importance of rigorously evaluating the suitability of simpler models for specific tasks. Moreover, they highlight the potential trade-offs of macroscopic linearity: While it facilitates interpretability and computational efficiency, it also obscures the rich nonlinear dynamics present at smaller scales, potentially losing valuable information. These results underscore the need for further investigation into the conditions and scales at which linear models are effective. Both studies are based on data from the resting state. My research is based on similar principles and approaches, focusing on task-driven data to investigate mappings between brain activity and semantic spaces. We consistently found that linear or relatively simple models outperformed more complex

models in capturing these relationships. Additionally, our findings revealed that the predictive performance of brain activity models improves significantly when paired with robust computational models—particularly those that are large-scale and multimodal [46, 49]. This observation suggests that high-dimensional vector spaces trained to be multimodal (and therefore more general) serve as effective approximations of macroscale brain representations. These representations appear to underlie cognition and behavior by encoding the external world in a structured, interpretable manner [201].

Limitations and ethical concerns

The field of brain decoding has made remarkable advancements during the last years; however, several limitations remain that challenge its applicability and generalizability. One significant limitation is the variability in brain anatomy and function between individuals. Subject-specific differences require personalized model training, which limits scalability and the development of universally applicable systems. Functional alignment techniques have been proposed to address this issue, but they are not always effective for fine-grained decoding, particularly when working with small datasets or noisy signals.

A critical concern in the recent decoding literature is the disentangling of the contributions of the decoding model from those of the prior information embedded in computational models. As noted recently [240], biases introduced by both the research methodology and the computational models used in decoding pipelines can distort interpretations. For example, using priors from pre-trained networks for reconstruction may result in outputs that reflect the model's internal biases rather than genuine neural representations. These biases are further compounded by the inherent complexity of defining what constitutes "ground truth" in the context of brain activity, where cognitive processes and representations often overlap and interact in non-linear ways. Generality and robustness also pose challenges. Decoding pipelines often struggle to generalize beyond the datasets on which they were trained. This is particularly evident when moving between datasets collected on different MRI machines, at varying field strengths, or with distinct experimental paradigms. Addressing these limitations requires improved methods to disentangle the contributions of task, subject, and experimental settings to the recorded neural data.

From an ethical perspective, the development of brain decoding technologies raises profound questions about privacy, identity, and agency. Yuste et al. [282] emphasize four key areas of concern: privacy, consent, agency, and equality. Privacy is particularly critical, as decoding neural data involves accessing deeply personal information, including thoughts, intentions, and emotions. Without

stringent regulations, there is potential for misuse, such as unauthorized access to neural data by corporations, governments, or malicious actors. To mitigate these risks, strong data protection frameworks must be established, ensuring that individuals maintain control over their neural information.

The implications of brain decoding for personal identity and agency are equally significant. Technologies capable of decoding or writing to the brain could blur the boundaries between human decision-making and machine influence, challenging notions of free will and selfhood. For example, systems designed to improve cognitive or motor abilities might unintentionally alter an individual's sense of self, as highlighted by Yuste [282]. Furthermore, the commercialization of these technologies could exacerbate social inequalities, creating divisions between those who have access to cognitive enhancements and those who do not. Finally, ethical challenges extend to the broader implications of brain-computer interfaces (BCIs) for society. Although these technologies hold promise for the treatment of neurological conditions and improving quality of life, they also risk being weaponized or exploited for purposes contrary to the societal good. To address these concerns, Yuste et al. propose the establishment of "neurorights" to safeguard fundamental human rights in the era of neurotechnology. These include the rights to mental privacy, personal identity, and equitable access to cognitive enhancement. Although brain decoding offers exciting possibilities for understanding cognition and developing transformative technologies, it is essential to address its inherent limitations and ethical challenges. By fostering interdisciplinary collaboration and implementing robust ethical guidelines, the field can advance in a way that respects and protects human dignity.

Future directions

The field of brain decoding has made remarkable strides in recent years, but several exciting avenues for future research are unexplored. A key area for advancement lies in improving generative brain decoding. Although current models have shown promise, future work should aim to refine the granularity and fidelity of reconstructed data. An interesting avenue is the development of fMRI foundation models, trained on huge-scale human datasets capable of converting high-dimensional spatio-temporal data into well-organized and useful embeddings (latent brain representations). Future decoding work could be extended to mapping between latent brain representations and latent model ones, potentially improving performance. However, to date, such models are still in their infancy and focus on ROI-wise brain activity, reducing the dimensionality from hundreds of thousands of voxel-wise time series to two hundred ROI time

series, making the whole problem more manageable but less informative for fine-grained decoding. For instance, transitioning from region-of-interest (ROI)-based decoding frameworks, such as BrainLM [28], to voxel-level foundation models could provide a more precise understanding of brain representations, enabling richer and more nuanced reconstructions.

Another promising direction involves the paradigm of intensive fMRI. As described by Kupers et al. [144], intensive fMRI emphasizes the collection of large, high-quality datasets from a small number of participants, allowing detailed exploration of cognitive phenomena at the single voxel level that can support multiple neuroscientific hypotheses. This approach offers opportunities to uncover fine-grained neural representations and could be instrumental in addressing variability between subjects. Future studies should consider integrating datasets from intensive fMRI with those from wide sampling paradigms to create unified frameworks capable of robustly generalizing across individuals and experimental contexts.

The aggregation of diverse datasets is another critical step forward. Combining data sets with varying acquisition parameters, experimental paradigms, and stimulus types would facilitate the creation of general-purpose decoding models. Advances in transfer learning and domain adaptation techniques could help bridge gaps between datasets, allowing a wider applicability of decoding frameworks across different neuroimaging modalities and experimental settings.

All of these steps could contribute to the development of foundation models for brain data, which represents an exciting frontier. Inspired by the success of foundation models in natural language processing and vision, such models for neuroimaging could leverage large-scale datasets and multimodal inputs to provide a unified representation of brain activity.

Conclusion

This thesis has explored brain decoding bridging computational neuroscience and machine learning to unravel how the brain represents and processes information across vision, language, music, and video. By leveraging non-invasive neuroimaging, advanced AI frameworks, and multimodal foundation models, this work demonstrated how brain activity can be mapped to semantic representations of external stimuli. Key contributions include novel methodologies for cross-modal decoding, cross-subject generalization, and unified frameworks for encoding and decoding across diverse modalities. While significant challenges remain, such as addressing dataset variability and integration, enhancing temporal resolution, and dealing with brain data dynamics, the findings underscore the transformative potential of decoding models. These models not

only hold promise for advancing neuroscience but also lay the groundwork for future applications in brain-computer interfaces, ultimately deepening our understanding of the brain's functional architecture and its connection to cognition and behavior.

A Appendix on EEG decoding

Decoding visual representations from human brain activity has emerged as a thriving research domain, particularly in the context of brain-computer interfaces. Most of the research shown in this thesis is focused on fMRI as neuroimaging modalities, while here we show that some of the same ideas can be translated to other modalities such as EEG. This chapter[1] presents an innovative method that employs knowledge distillation to train an EEG classifier and reconstruct images from the ImageNet and THINGS-EEG 2 datasets using only electroencephalography (EEG) data from participants who have viewed the images themselves (i.e. "brain decoding"). We analyzed EEG recordings from 6 participants for the ImageNet dataset and 10 for the THINGS-EEG 2 dataset, exposed to images spanning unique semantic categories. These EEG readings were converted into spectrograms, which were then used to train a convolutional neural network (CNN), integrated with a knowledge distillation procedure based on a pre-trained Contrastive Language-Image Pre-Training (CLIP)-based image classification teacher network. This strategy allowed our model to attain a top-5 accuracy of 87%, significantly outperforming a standard CNN and various RNN-based benchmarks. Additionally, we incorporated an image reconstruction mechanism based on pre-trained latent diffusion models, which allowed us to generate an estimate of the images that had elicited EEG activity. Therefore, our architecture not only decodes images from neural activity but also offers a credible image reconstruction from EEG only, paving the way for, e.g., swift, individualized feedback experiments.

[1]The work presented in this chapter has been presented at ICLR 2024 Time-series for health workshop. Full manuscript has been published at "Computers in Biology and Medicine" [77].

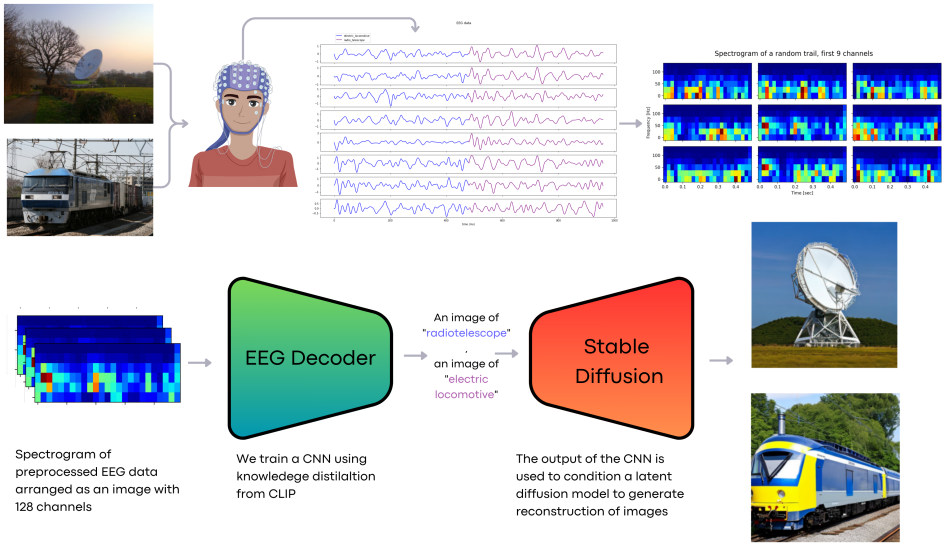


Figure A.1: Our pipeline can be described as follows: First, we record EEG data while the subject is viewing natural images. This data is then preprocessed and converted into spectrograms, which serve as the input for our neural network. Our EEG decoder is trained using a knowledge distillation method based on the CLIP model. The outputs from the EEG decoder, which are predictions of the image that elicited the EEG data, are then combined with an image generation pipeline. This end-to-end approach allows us to reconstruct images from the neural activity data captured by the EEG.

A.1 Introduction

Electroencephalography (EEG) is increasingly recognized as a valuable instrument for decoding visual representations within the human brain. The primary advantage of EEG lies in its non-invasive nature and its ability to provide real-time insights into human brain function via electrical activity recordings from the scalp. Despite its spatial resolution constraints, its unparalleled temporal resolution renders it ideal for real-time applications.

Recent technological advancements have facilitated the decoding of intricate visual stimuli from EEG signals, notably from expansive datasets such as ImageNet [193, 11]. Both convolutional (CNN) and recurrent neural networks (RNN) have demonstrated efficacy in classifying EEG signals into distinct image categories with appreciable accuracy. The successful decoding of complex visual stimuli from EEG signals can pave the way for innovative neural prosthetics and biofeedback systems. Translating brain activity patterns into decoded image categories or reconstructions could potentially offer visually impaired individuals a semblance of artificial vision. Additionally, EEG decoding can revolutionize brain-centric image searches, communication platforms, and augmented real-

ity interfaces. Real-time visualizations of decoded brain activity can also usher in novel neurofeedback paradigms, facilitating self-regulation of brain states through integrated EEG decoding and external visual feedback mechanisms [72].

However, a predominant focus in current research is on multisubject models, which involve averaging EEG signals across multiple participants. This methodology may overlook the nuances of individual-specific neural representations. Models tailored to individual participants could offer a more granular decoding and introduce an added dimension of data privacy, as each model is uniquely calibrated for a specific individual, precluding its application to others. Also, in spite of recent progress, the task of reconstructing visual stimuli based on the EEG signals they elicit remains a formidable challenge. The inherent low spatial resolution of EEG poses difficulties in reconstructing detailed visual nuances. Presently, image reconstructions predominantly capture broader features, such as shapes, colors, and textures, thereby constraining the depth of visual feature decoding and image reconstructions. To overcome this obstacle, instead of attempting pixel-precise reproductions, a more pragmatic approach might be semantic image reconstructions. For example, approaches like generative adversarial networks (GANs) [131] show promise for creating semantically meaningful reconstructions from EEG. EEG provides a useful macro-level window into visual processing in the brain. Multimodal approaches that combine EEG with imaging modalities like fMRI could help overcome the limitations of EEG alone. Using fMRI, the higher spatial resolution, is possible to reconstruct images with low-level agreement [78, 189, 190, 250]. Nevertheless, could be interesting reconstructing in real-time images from EEG data and show this reconstruction to the subject during the experiment, enabling a feedback loop [72], so the subject can learn how to focus on images to improve classifier performances. This research aims to improve existing methodologies for translating perceptual experiences from EEG patterns, with a focus on real-time applications. We present a methodology that advances this field, outlining a pipeline (as shown in Fig A.1) that facilitates the training of a single-subject model within a limited experimental timeframe, leading to near-real-time brain decoding. video The central innovation of this work lies in the proposed methodology for addressing the challenge at hand. Our goal is to achieve semantic decoding of visual content from electroencephalogram (EEG) activity, which is commonly very noisy and comes with reduced spatial resolution, limiting the chance of achieving fine-grained decoding of image details. To this end, we initially train an EEG classifier using an asymmetric student-teacher knowledge distillation approach [109] In this context, the 'teacher' model is the pre-trained Contrastive Language-

Image Pre-training (CLIP) model [209], which generates class probabilities from images. Unlike traditional frameworks, where the "student" model learns to replicate the "teacher's" outputs with either reduced capacity or on a corrupted version of the same stimuli, our approach involves feeding EEG activity into the "student" model. This compels the student model to learn how to predict class probabilities based on neural signals. Following the training phase, we retain only the EEG decoder, which we then integrate with a generative model based on latent diffusion. This combination is employed to produce novel images that possess semantic content derived from EEG signals

A.2 Related Works

EEG are widely processed in the context of brain-computer interfaces (BCI) to perform brain decoding for a wide variety of tasks [283, 239, 264, 26, 176]. A number of prior works have explored decoding visual representations from EEG signals using deep learning models. Kavasidis et al. [132] were among the first to propose generating images from EEG data. They recorded EEG while participants viewed ImageNet images, and used an Long Short Term Memory (LSTM) model combined with variational autoencoders or GANs to reconstruct images. The key difference is they aimed for class-level image generation rather than detailed reconstruction and focuses on processing data in the time domain. Spampinato et al. [244] also analyzed EEG responses to ImageNet stimuli. They trained an LSTM encoder to classify EEG signals into image categories. For reconstruction, they trained a separate CNN regressor to predict EEG features from images and replaced the EEG signal with this encoder model. Palazzo et al. [191] extended [244] using contrastive learning to align EEG and visual image features. However, their goal was improving image classification rather than reconstruction, and various challenges emerged [157]. Singh et al. [242] proposed an EEG-to-image GAN framework but focused on smaller (i.e. with fewer images) datasets of characters and shapes. In this work, we propose a modularized pipeline for reconstructing detailed photorealistic visual stimuli (i.e. images) directly from EEG brain signals, using a CLIP based knowledge distillation of a convolutional neural network trained on time-frequency decomposition (TFD) and generative diffusion synthesis, generating semantically plausible and visually similar images reconstruction to the original stimuli.

A.3 Material and Methods

This section delineates the methodology adopted and the dataset utilized. The datasets, sourced from ImageNet EEG [131] and THINGS-EEG2 [88], are publicly accessible. All computational experiments and model training were conducted on a server outfitted with four NVIDIA A100 GPU cards (each with 80GB RAM connected via NVLINK) and 2 TB of system RAM. The codebase was developed using Python 3.9, leveraging libraries such as Pytorch, Pytorch Lightning, and scikit-learn for model implementation. Code is freely accessible at `https://github.com/matteoferrante/EEG_decoding`.

A.3.a Data

The ImageNet-EEG recordings employed in this study were sourced from [245]. These recordings were obtained from six participants who were exposed to images from 40 distinct ImageNet [56] classes, with each class comprising 50 images. The sampling rate for these recordings was 1000 Hz. The image presentation protocol involved sequential display in 25-second intervals, succeeded by a 10-second intermission. In each display interval, images are shown sequentially for 0.5 seconds each. This protocol yielded a total of 2,000 images spanning 1,400 seconds (or 23 minutes and 20 seconds) of recording time. Each subject underwent four recording sessions, each lasting 350 seconds. The experiments utilized a 128-channel cap with active, low-impedance electrodes (actiCAP 128Ch, Brainproducts) for EEG data collection. Brainvision amplifiers and data acquisition systems were used to record the EEG signals at a sampling rate of 1000 Hz with 16-bit resolution. The EEG data resulted in 11,466 sequences post the exclusion of recordings of suboptimal quality. The comprehensive nature of this experimental design facilitated the examination of EEG responses to a diverse array of visual stimuli from ImageNet. The multi-channel EEG recordings, captured during the viewing of thousands of stimuli, furnish a rich dataset conducive for training decoding models. For further detail about acquisition protocol please see the original article [245]. To show the generalization ability of our method, we also included another dataset, from the THINGS initiative collection, named THINGS-EEG2 [88] This dataset comprises a collection of EEG readings taken at high temporal precision, recording reactions to pictures of objects against a natural backdrop. It encompasses data from 10 participants, covering 82,160 instances across 16,740 different image scenarios. Image stimuli belong to the THINGS Image dataset, spanning across 1854 different classes. In this work, the EEG activity is recorded while 1654 categories were shown as part of the training set and the other 200 categories were shown as test set. Since the very fine granu-

larity of concepts of this dataset makes the problem more complex, we obtained pseudo-labels for the entire dataset using a K-Means over the CLIP embeddings of all images. We used a k-Elbow approach to identify the optimal number of clusters (that turned out to be 8 in this case) and trained a K-Means clustering to predict cluster labels to re-label this dataset. EEG data was processed as described before. Since the cluster labels cannot be used as conditioning for the generative part because they are obtained via an unsupervised method and hence would require a human re-labelling, we adopted a simpler approach for the second dataset, restricting the analysis to the classification part.

A.3.b Preprocessing

Prior to utilizing the EEG signals for training our decoding models, a series of preprocessing steps were executed. Initially, a notch filter in the 49-51 Hz range was applied to mitigate power line interference. Subsequently, a second-order band-pass Butterworth filter, ranging between 14 and 70 Hz, was employed to focus on frequency bands pertinent to visual attention and object recognition. The signals were then standardized across channels. For the purpose of neural network input generation, the filtered EEG signals were segmented into 40 ms windows moving each time 20 ms. Time-frequency decompositions (TFD) were computed for these segments using the short-time Fourier transform (STFT), converting each trial into a 128-channel image that depicted the spectrum across both time and frequency dimensions. For the ImageNet-EEG this process yielded 2,000 EEG spectrogram images, each with 128 channels, for every subject. For the THINGS-EEG2 dataset, we applied the same preprocessing steps, resulting in 16,540 spectrograms for and 200 spectrograms for testing. Given the dataset's highly detailed categorization into 1,854 distinct classes, with no overlap between training and testing categories, traditional classification methods were deemed inappropriate for handling the data. To address this challenge, we re-run the classification using pseudo-labels generated by a clustering algorithm (K-means). This algorithm was applied to the image embeddings derived from the pretrained CLIP model, leading to the formation of 8 classes. The decision to use 8 clusters was based on the K-Elbow method, which searched within a range of 2 to 20, ensuring these classes were consistently represented across both training and testing datasets. Spectrograms were then used for training and evaluation of our convolutional neural network tailored for EEG decoding. This multi-channel spectral representation encapsulates the spatial and temporal intricacies of the EEG, allowing our model to extract features essential for visual stimuli classification. It is worth noting that the preprocessing described herein is specific to the architecture proposed in this study. Alternative baselines

adopted slightly varied preprocessing techniques, such as direct time domain data analysis, starting from the same filtered data in the time domain. These variant preprocessing methodologies are elaborated upon in A.3.f.

A.3.c Decoding pipeline

Our approach employs a CNN with integrated residual connections to classify EEG TFDs. The architecture begins with a series of convolutional layers, progressively increasing the number of filters to effectively extract both spatial and temporal features. Subsequent to this, global average pooling and fully-connected layers are utilized for classification tasks. For the training of the CNN, we adopt a knowledge distillation methodology [109]. Initially, an image classifier is pretrained using CLIP (Contrastive Language-Image Pre-Training) [209] features to anticipate the stimulus classes, achieving a commendable accuracy of 99%. This pretrained classifier furnishes "soft targets" to guide our EEG model. During the training phase, EEG spectrograms are fed into the CNN, while CLIP image features are directed to the teacher classifier. The objective is to train the CNN such that it aligns with the class probability distributions produced by the teacher. This distillation approach not only stabilizes the training process but also enhances the model's performance in comparison to direct training on class labels. For inference, only the EEG-based CNN is deployed to predict classes from novel time-frequency decompositions. Through the distillation of knowledge from the image model, our CNN is equipped to derive robust representations, enabling the decoding of visual stimuli solely from EEG signals.

Post the training of our EEG decoding model, it becomes capable of predicting ImageNet classes from fresh EEG TFDs. To validate these predictions and reconstruct images that could potentially induce analogous neural responses, we employ the Stable Diffusion generative model [214]. For every EEG prediction, a text prompt such as "an image of a predicted class" is formulated. This prompt, in conjunction with random noise vectors, is input into Stable Diffusion to generate images congruent with the predicted class. This methodology facilitates the reconstruction of visual stimuli exclusively from neural activity patterns. The EEG decoder identifies the class, while Stable Diffusion fabricates a semantically coherent image. A comprehensive diagram of the decoding pipeline is depicted in Fig A.1, and the knowledge distillation procedure is illustrated in Fig. A.2.

A.3.d Reconstruction Pipeline

Diffusion models are generative frameworks trained to invert a noise diffusion process, facilitating image synthesis. Stable Diffusion operates as a latent diffusion model, proficient in generating lifelike images from random noise vectors, conditioned by textual descriptions. The model's strategy involves the iterative addition of noise to genuine images, followed by the learning of a parametric denoising function to eradicate the noise over multiple timesteps. By repetitively applying the denoising function, the model can produce lifelike images, conditioned on textual descriptions. This iterative denoising offers tight control over image generation, guided by text at every iteration. In the sampling phase, Stable Diffusion accepts a text prompt and progressively diffuses noise vectors until they converge into an image that aligns semantically with the provided description. For the task of reconstructing images from EEG signals, Stable Diffusion's text conditioning capability proves invaluable. The EEG decoder outputs a label indicative of the visual stimulus class. This discrete label is then employed to generate corresponding images via Stable Diffusion, bypassing the need for direct pixel reconstruction. This approach facilitates the synthesis of plausible image reconstructions based on the decoded semantic category from neural activity patterns. This model-centric strategy also addresses the inherent resolution constraints of EEG for high-fidelity decoding. The guided diffusion modeling ensures the generation of visualizations that are both realistic and interpretable to human observers.

A.3.e Knowledge Distillation

Knowledge distillation facilitates the transfer of insights from a comprehensive, pretrained teacher model to a more compact student model [109]. This process empowers the student model to attain performance metrics that are typically associated with larger models. Consider $f_t(x)$ as the output vector of class probabilities produced by the teacher model for a given input x, representing the stimulus image. Similarly, let $f_s(e;\theta)$ denote the student model, characterized by parameters θ, where e represents the EEG recordings obtained during the presentation of stimulus x. The student model is trained through knowledge distillation by minimizing:

$$\mathcal{L}(\theta) = \alpha\mathcal{L}_{CE}(fs(e;\theta), y) + (1-\alpha)\mathcal{L}_{KD}(fs(x;\theta), f_t(x)) \tag{A.1}$$

Here, $\mathcal{L}_{CE}$ represents the cross-entropy loss between the predictions of the student model and the actual ground truth labels y. In contrast, $\mathcal{L}_{KD}$ denotes the distillation loss, capturing the difference between the outputs of the student

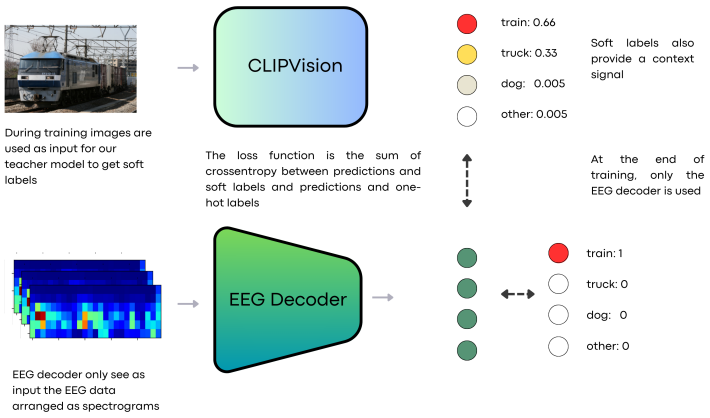


Figure A.2: Illustration of the training procedure. Knowledge distillation facilitates the training of a compact "student" model to emulate the outputs of a more extensive "teacher" model. This enables the student to achieve performance levels akin to larger models, even when initiated from distinct yet related inputs.

and teacher models. The temperature parameter T is employed to modulate the probability distribution of the teacher:

$$\mathcal{L}_{KD}(fs, ft) = -\sum_{c} \frac{\exp(f_{t,c}/T)}{\sum_{c'} \exp(f_{t,c'}/T)} \log \frac{\exp(f_{s,c}/T)}{\sum_{c'} \exp(f_{s,c'}/T)} \tag{A.2}$$

Training the student model to replicate the comprehensive probability distribution of the teacher facilitates the transfer of insights regarding inter-class relationships, offering a richer supervisory signal than mere ground truth labels. In our implementation, we set $\alpha = 0.5$ and $T = 1$ after initial empirical experimentation. For EEG decoding, a linear classifier was trained atop the CLIP [209] CLS tokens. CLIP, an acronym for Contrastive Language-Image Pre-Training, is a neural architecture trained to correlate images and text through contrastive learning. Comprising an image encoder and a text encoder, CLIP is trained to discern whether an image-text pairing is congruent or not. The image encoder in CLIP, a vision transformer (VIT), embeds images into latent representations. Throughout its training, CLIP cultivates an embedding space where semantically congruent images and texts are proximate. A pivotal element of the image encoder is the CLS token, an auxiliary token introduced to the network's input, enabling the encoder to generate a holistic representation of the entire image. A linear classifier was trained atop this CLS token for every image in the training dataset to predict the appropriate class. This amalgamation of CLIP and the classifier served as the teacher model, functioning as a bridge between EEG

spectrograms and image classes. The student CNN, when exposed solely to EEG data, derives insights from both the teacher's distributions and the true labels. This distillation process accentuates the student's focus on neural patterns pertinent to visual recognition, enhancing convergence, accuracy, and generalization. By assimilating insights from a domain expert in image processing, the streamlined student decoder becomes adept at extracting visual representations from EEG signals.

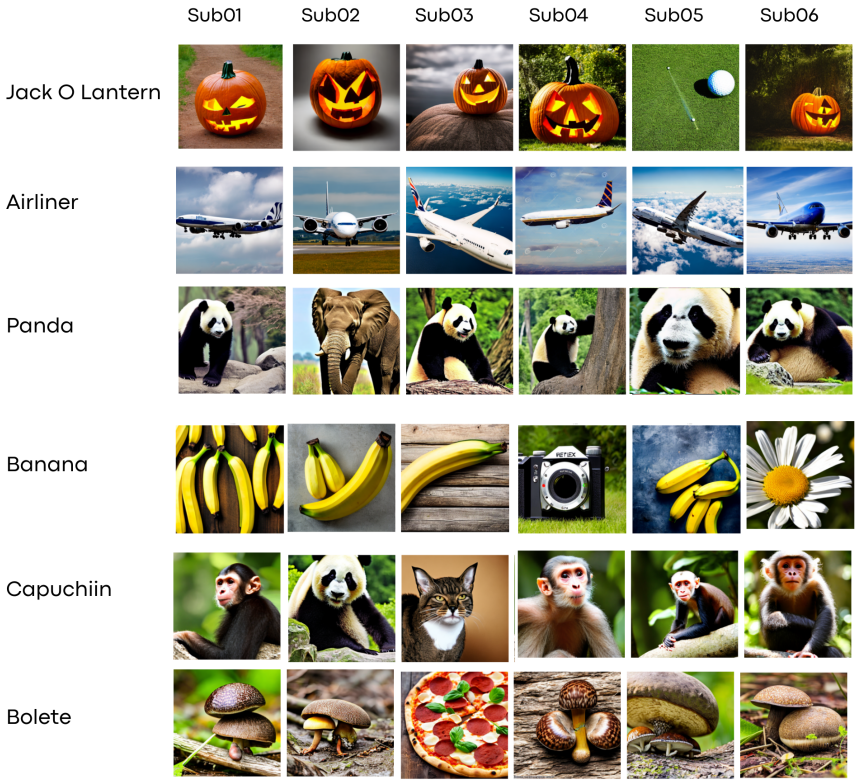


Figure A.3: Reconstructed images. Left column: target classes; subsequent columns: results from individual participants.

A.3.f Baselines

In order to underscore the efficacy of employing computer vision techniques for EEG signal decoding, we assessed a spectrum of baseline methodologies, spanning from conventional machine learning paradigms to contemporary neural network architectures.

Recently, several studies with remarkable results have been published on this dataset [131, 12, 245]. However, a subsequent analysis [156] revealed that, de-

spite the methodological advancements being valid and innovative, the reported performance metrics are significantly inflated. This inflation is attributed to erroneus data preprocessing. Specifically, some preprocessing filters can induce temporal correlations between data points before splitting them into training and test sets, leading to information leakage. In response to these findings, follow-up counter-analyses [192] have demonstrated that, when eliminating this effect, the results remain valuable, albeit lower than initially reported. Therefore, in situating our work within the broader context of existing literature, in order to maintain best practices and avoid leakage, we have opted for the most conservative approach as outlined in the above-mentioned papers [192, 156]. For similar reasons, in this paper we also provide an extensive set of baselines for performance comparison.

Initially, we employed a basic baseline wherein the raw EEG signals were standardized, squared, and subsequently averaged across channels. Following this, a Logistic Regression classifier was trained on the resultant data. An extension of this approach involved applying the Logistic Regression classifier to EEG signals that were averaged over an 80-point sliding window. In another variant we executed PCA on the windowed average EEG, preserving 29 components that accounted for 95% of the variance, prior to classifier training. Notably, these methodologies overlook the inherent spatial and temporal intricacies of the EEG signal. The main advantage of using the PCA is providing orthogonal features to the model that already integrate relevant spatiotemporal relationships. In this context, a recent proposition by CEBRA [232] demonstrated a deep learning technique that employs contrastive learning to project neural data onto lower-dimensional manifolds conducive for decoding. In alignment with this, we projected our EEG data onto a 32-dimensional manifold, utilizing CLIP features as a guiding mechanism. The value was chosen to be close to the number of features used in the PCA, picking the closest power of 2. This offers a robust nonlinear neural baseline that effectively harnesses both spatial and temporal patterns.

In terms of neural network architectures that directly process EEG time series data, we examined both a LSTM model and a 1D convolutional network (CNN) equipped with temporal convolutions. Both architectures incorporated 4 layers and were regularized using dropout, ensuring a consistent parameter count across models.

Further, we explored CNNs that operate on 2D representations of the EEG, thereby leveraging computer vision methodologies. One such model treated the raw EEG traces as a 2D image. Another model employed a wavelet decomposition utilizing the Daubechies db4 wavelet from PyWavelets [2] [152], which

has been recognized as an efficient time-frequency representation for EEG [247]. Our final CNN baseline ingested the short-time Fourier transform (STFT) of the EEG, processed with a 40 ms window.

This ensemble of baselines, ranging from classical signal processing to avant-garde deep learning, offers a holistic comparative framework and accentuates the significance of spatiotemporal neural network modeling in the realm of EEG decoding. The computer vision-oriented strategies adeptly harness the structural nuances present in the multi-channel EEG.

For consistency, all neural networks were evaluated with a similar parameter count range (1.1-1.2 M). Each was trained using the Adam optimizer at a learning rate of $3e-4$. Additional training specifications included an early stopping callback with a 10-epoch patience based on validation loss variations, a batch size of 64, gradient clipping at a magnitude of 1.0, and a maximum epoch count set to 50.

A.4 Results

In this section, we present the outcomes for both datasets. For the initial dataset, ImageNet-EEG, we provide a comprehensive overview of the entire process, including classification outcomes and qualitative image reconstruction. This approach is feasible due to the relatively small (40) number of categories, allowing us to condition the generative model directly using the class labels. For the THINGS-EEG2 dataset, we focus exclusively on the classification results derived from the pseudo-labels assigned by the clustering algorithm, with the main objective of demonstrating the decoding of semantic information from EEG data."

A.4.a Performance Evaluation

Several metrics are available to evaluate the performances of a classification model [19, 20, 30, 114]. In our case, the efficacy of our model is evaluated using a comprehensive set of metrics: top-5, top-3, top-1 accuracy, F1 score, and the normalized kappa score to evaluate performances. Figure A.5 demonstrates that our knowledge distillation CNN consistently outperforms both the standard CNN baseline and a random classifier. Notably, the proposed approach—employing a CNN on TFD with CLIP-based knowledge distillation—exhibits superior performance compared to the same network without the distillation technique. This superiority is further evident when juxtaposed with other baselines detailed in Table A.1.

Table A.1 provides a summarized view of the decoding performance across various methods applied to EEG data. Clear trends in accuracy emerge across

Figure A.4: On the left, the target classes are presented and each column show result from a single subject.

model types. Classical machine learning baselines, which utilize averaged or PCA-reduced EEG, yield near chance-level accuracy, underscoring the inadequacy of hand-engineered features for decoding intricate visual stimuli. An exception is the Logistic Regression model trained on squared data averages.

Conversely, deep learning models that harness spatiotemporal EEG TFDs patterns consistently achieve superior accuracy. Both convolutional and recurrent neural networks processing raw EEG time series deliver satisfactory results. Yet, the best performance is reached by models using 2D representations of multi-channel EEG. Specifically, CNNs fed with TFD computed using wavelet-transformed or spectrogram images both surpass 85% in top-5 accuracy, underscoring the benefits of computer vision techniques that learn directly from 2D structures in signal processing. Both wavelet and spectrogram decompositions seem to encapsulate pertinent time-frequency domain information for decoding. A closer examination of the top-3 and top-5 accuracy metrics reveals a consistent trend: deep learning models outclass classical baselines. The elite CNNs achieve over 75% in top-3 accuracy, implying that in approximately 3 out of 4 trials, the true label ranks within the top three predictions. The performance gap relative

Method	Metrics [Mean (Std)]				
	Accuracy	Top3 Accuracy	Top5 Accuracy	F1	Kappa
LR on average square signal	0.3600 (0.1313)	0.6619 (0.1758)	0.8156 (0.1619)	0.3493 (0.1375)	0.3435 (0.1345)
LR on windowed signal	0.0205 (0.0058)	0.0636 (0.0083)	0.1092 (0.0110)	0.0156 (0.0054)	0.0009 (0.0061)
LR on PCA windowed signal	0.0175 (0.0040)	0.0536 (0.0084)	0.0961 (0.0063)	0.0097 (0.0047)	0.0020 (0.0039)
CEBRA + kNN	0.0240 (0.0050)	0.0831 (0.0116)	0.1402 (0.0136)	0.0223 (0.0061)	-0.0012 (0.0056)
LSTM	0.3605 (0.0938)	0.7376 (0.1226)	0.8868 (0.1030)	0.3392 (0.0894)	0.3437 (0.0960)
Conv1d	0.2623 (0.0511)	0.6013 (0.0826)	0.7971 (0.0851)	0.2582 (0.0520)	0.2432 (0.0524)
Knowledge distillation on eeg (img)	0.2819 (0.0836)	0.5773 (0.1379)	0.7295 (0.1339)	0.2742 (0.0794)	0.2632 (0.0857)
Knowledge distillation on wavelet	0.4060 (0.1154)	0.7490 (0.1282)	0.8787 (0.1007)	0.3889 (0.1148)	0.3905 (0.1183)
plain CNN on spectrograms	0.2819 (0.0836)	0.5773 (0.1379)	0.7295 (0.1339)	0.2742 (0.0794)	0.2632 (0.0857)
Knowledge distillation on STFT	**0.4120 (0.1131)**	**0.7530 (0.1068)**	**0.8782 (0.0806)**	**0.4027 (0.1133)**	**0.3966 (0.1160)**
Palazzo et al [192]	0.3350 (0.089)	-	-	-	-
Knowledge distillation (THINGS-EEG2)	**0.58 (0.04)**	-	-	**0.52 (0.036)**	-
Plain CNN (THINGS-EEG2)	0.52 (0.03)	-	-	0.48 (0.032)	-

Table A.1: Performance comparison of decoding baselines. The table presents the mean values accompanied by the standard deviation (enclosed in parentheses) for each evaluation metric across all participants. Results from [192] are reported from the original paper in the same setting used here. The first part of the table reports results for ImageNet-EEG dataset, while the second part report comparison between our method and plain CNN on the THINGS-EEG2 dataset.

to the LSTM network is also noteworthy. This accentuates the efficacy of 2D convolutions in discerning the pertinent semantic categories from EEG patterns. The consistency of the top-5 accuracy across deep learning models suggests potential inherent challenges in precisely mapping EEG to granular image labels. However, the models adeptly identify the overarching category within their top predictions, underscoring the viability of EEG-based visual concept decoding. Finally, using our 8 clustering-derived pseudo-labels, we also verified that our approach outperforms a plain CNN baseline on the THINGS-EEG2 dataset.Our final model trained with knowledge distillation was able to achieve a top-1 prediction of 58 %, hence discovering semantic content of the seen image from the neural data and confirming previously results.”

From a qualitative perspective, Figures A.3 and A.4 showcase examples of predicted and reconstructed images. While the model predominantly identifies the correct visual concept from EEG patterns, minor category confusions do arise. For instance, "bolete" might be misinterpreted as "pizza," or "banana" as "Margherita". Nevertheless, the model’s ability to accurately discern the overarching semantic category and produce corresponding reconstructions is noteworthy.

In conclusion, our findings underscore the pivotal role of neural networks and image-centric representations in harnessing the rich multidimensional EEG representation. Directly classifying TFD inputs using a computer vision approach emerges as the potent strategy for EEG-based decoding.

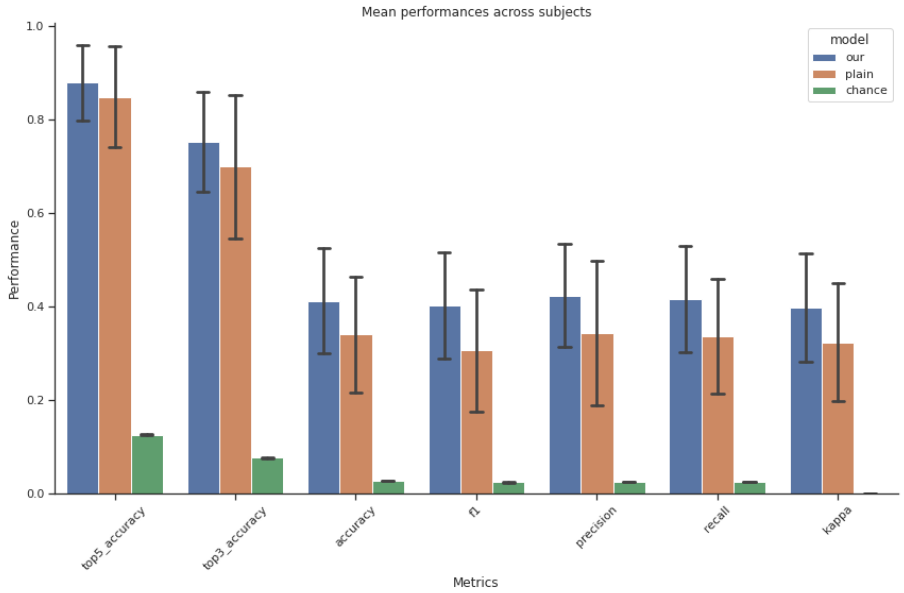


Figure A.5: Results for EEG decoder. **Ours** is the CLIP-based approach, **plain** is a vanilla CNN with the same architecture trained for classification and **chance** serves as a comparison with chance level. Bars are average across participants and error bars are standard deviations.

A.5 Discussion

The primary objective of this study was to decode and reconstruct visual representations from EEG-recorded human brain activity. By employing deep convolutional neural networks trained on EEG TFD and guided by the CLIP-based knowledge distillation technique, we managed to predict image classes from the ImageNet dataset with an accuracy of 87% in the top-5 category. This knowledge distillation approach yielded a marked improvement in performance when compared to a baseline model and other data processing methodologies. While the model's predictions were generally reliable for the majority of participants, it did exhibit some confusion between closely related classes. The capability to extract the semantic content of image stimuli from non-invasive EEG recordings presents significant implications for the future of brain-computer interfaces. The methodology we developed for image reconstruction could potentially pave the way for a form of artificial vision, where decoded contents from a user's neural activity are visualized in real-time. Furthermore, our model introduces the possibility of innovative neurofeedback experiments, wherein participants could receive instantaneous visual feedback of decoded EEG patterns, facilitating the voluntary self-regulation of brain states [72]. However, the study is not

without limitations. EEG serves as a macroscopic lens into the brain's visual processing mechanisms. To address the limitations of EEG's spatial resolution, integrating it with other imaging techniques, such as fMRI, which boasts superior spatial resolution, is a promising avenue. Such multimodal strategies have shown potential in reconstructing images with a higher degree of detail [78, 189, 190, 250]. Also, the model in its current configuration has not been optimized for decoding images outside the 40 categories or the 8 clusters used in the experiments, suggesting a need for further refinement. The variability in EEG decoding abilities across different participants or sessions, influenced by cognitive and neural factors, remains a topic that warrants deeper exploration. One of the significant concerns in EEG decoding revolves around the inadvertent extraction of personal perceptual data, which must be rigorously addressed. Our methodology places a strong emphasis on the creation of subject-specific models. This ensures that the decoding process is both consensual and uniquely tailored to the individual, mitigating potential ethical concerns. This approach not only necessitates voluntary participation but also minimizes the risk of misinterpretations due to the model's specificity to individual neural patterns. The rapid training methodology we have introduced also holds promise for real-time feedback paradigms using models tailored to individual participants, with a couple of seconds in inference time needed to predict class and generate the image on an A100 GPU. As the field of deep learning and generative models continues to evolve, we anticipate parallel advancements in EEG decoding and reconstruction capabilities.

A.6 Conclusions

In conclusion, our research demonstrates the capability of an integrated EEG decoding system using a novel knowledge distillation technique paired with latent diffusion models. This approach not only advances theoretical understanding but also holds significant promise for practical applications. The potential real-world applications of this technology are vast, including the development of assistive technologies for individuals with disabilities, enhancing communication for those unable to speak or use traditional input devices, and improving neurorehabilitation methods. One immediate application could be in the realm of augmented and virtual reality, where users could manipulate environments directly through neural inputs, creating a more immersive and intuitive user experience. Moreover, the integration of our decoding approach with existing technologies could lead to more responsive and adaptive systems, tailored to individual neurological profiles for personalized user interfaces. Future work

will focus on refining the decoding accuracy and efficiency of the system, exploring the integration with other modalities like fMRI for improved spatial resolution and incorporating real-time feedback mechanisms to enhance learning and adaptation in the brain-computer interface. Additionally, further research into the ethical implications and the security of neural data in such applications is paramount to ensure privacy and consent in the use of this technology. The methodologies and findings from this study could significantly influence the development of next-generation brain-computer interfaces by providing a framework that employs advanced machine learning techniques to interpret and translate complex neural signals into actionable outputs. This could eventually lead to breakthroughs where brain-computer interfaces may offer seamless integration between human cognitive states and machine operations, heralding a new era of interaction between humans and technology In this study, we demonstrated the potential of deep neural networks, coupled with generative diffusion models, to reconstruct visual experiences directly from non-invasive EEG recordings from two independent datasets achieving a top-1 accuracy in prediction of 40 classes of 45% (and a top-5 accuracy of 87%) on the ImageNet-EEG dataset and a top-1 accuracy of 58 % in prediction of 8 semantic clusters on the THINGS-EEG2 dataset. The application of knowledge distillation from language-image pretraining enabled our convolutional decoder to effectively extract semantic information from brain activity patterns. This capability significantly surpassed the performance of classical signal processing baselines. By generating images based on the predicted labels, we were able to produce visualizations that closely align with the decoded neural activity. Our emphasis on creating subject-specific models not only ensures a certain degree of privacy but also underscores the unique capabilities of EEG data in decoding individual mental representations. These techniques, which focus on translating neural signals into their corresponding images, can kickstart significant advancements in the domains of brain-computer interfaces and neural prosthetics, as well as human-computer interaction research. Overall, our findings highlight the potential of non-invasive brain imaging as a tool to provide insights into the human cognitive experience.

Architecture details

Here we report some additional details, such as the network's architecture structure. Each CNN model has the following structure and was trained using the Adam optimizer at a learning rate of $3e-4$. Additional training specifications included an early stopping callback with a 10-epoch patience based on validation loss variations, a batch size of 64, gradient clipping at a magnitude of 1.0, and a maximum epoch count set to 50.

Table A.2: CNN Classifier Network Structure

Layer	Type of Operation	Details
cnn_model	Classifier	-
net	Sequential	-
layer_0	ResidualUnit	-
- conv	Sequential	-
– unit0	Convolution	Conv2d(17, 64, kernel_size=(3,), stride=(2,), padding=(1,))
– adn	ADN	BatchNorm2d(64), Dropout(p=0.2), GELU
- residual	Conv2d	Conv2d(17, 64, kernel_size=(3,), stride=(2,), padding=(1,))
layer_1	ResidualUnit	-
- conv	Sequential	-
– unit0	Convolution	Conv2d(64, 64, kernel_size=(3,), stride=(1,), padding=(1,))
– adn	ADN	BatchNorm2d(64), Dropout(p=0.2), GELU
- residual	Identity	-
layer_2	ResidualUnit	-
- conv	Sequential	-
– unit0	Convolution	Conv2d(64, 128, kernel_size=(3,), stride=(2,), padding=(1,))
– adn	ADN	BatchNorm2d(128), Dropout(p=0.2), GELU
- residual	Conv2d	Conv2d(64, 128, kernel_size=(3,), stride=(2,), padding=(1,))
layer_3	ResidualUnit	-
- conv	Sequential	-
– unit0	Convolution	Conv2d(128, 128, kernel_size=(3,), stride=(1,), padding=(1,))
– adn	ADN	BatchNorm2d(128), Dropout(p=0.2), GELU
- residual	Identity	-
layer_4	ResidualUnit	-
- conv	Sequential	-
– unit0	Convolution	Conv2d(128, 128, kernel_size=(3,), stride=(2,), padding=(1,))
– adn	ADN	BatchNorm2d(128), Dropout(p=0.2), GELU
- residual	Conv1d	Conv2d(128, 128, kernel_size=(3,), stride=(2,), padding=(1,))
layer_5	ResidualUnit	-
- conv	Sequential	-
– unit0	Convolution	Conv2d(128, 128, kernel_size=(3,), stride=(2,), padding=(1,))
- residual	Conv2d	Conv2d(128, 128, kernel_size=(3,), stride=(2,), padding=(1,))
reshape	Reshape	-
final	Sequential	-
- 0	Flatten	Flatten(start_dim=1, end_dim=-1)
- 1	Linear	in_features=896, out_features=*num classes*, bias=True

Detailed performances

Table A.3: Detail of average classification performances for ImageNet-EEG for each class

Class	Precision	Recall	F1-Score
sorrel	0.71	1.00	0.83
parachute	.90	0.90	0.95
iron	0.50	0.67	0.57
anemone_fish	0.20	0.33	0.25
espresso_maker	0.33	0.33	0.33
coffee_mug	0.14	0.40	0.21
mountain_bike	0.50	0.33	0.40
revolver	0.20	0.14	0.17
giant_panda	0.67	0.20	0.31
daisy	0.36	0.44	0.40
canoe	0.50	0.56	0.53
lycaenid	0.43	0.38	0.40
German_shepherd	1.00	0.71	0.83
running_shoe	0.17	0.17	0.17
jack-o′-lantern	0.44	0.36	0.40
cellular_telephone	0.33	0.25	0.29
golf_ball	1.00	0.67	0.80
desktop_computer	0.50	0.46	0.48
broom	0.18	0.40	0.25
pizza	0.22	0.33	0.27
missile	0.50	0.25	0.33
capuchin	0.50	0.33	0.40
pool_table	0.50	0.71	0.59
mailbag	0.09	1.00	0.17
convertible	0.22	0.20	0.21
folding_chair	0.38	0.60	0.46
pajama	0.56	0.62	0.59
mitten	0.55	0.50	0.52
electric_guitar	0.44	0.33	0.38
reflex_camera	0.25	0.25	0.25
grand_piano	0.40	0.50	0.44
mountain_tent	0.88	0.78	0.82
banana	0.43	0.50	0.46

Table A.3: Detail of average classification performances for ImageNet-EEG for each class

Class	Precision	Recall	F1-Score
bolete	0.62	0.45	0.53
digital_watch	0.12	0.12	0.12
African_elephant	0.60	0.50	0.55
airliner	0.36	0.44	0.40
electric_locomotive	0.50	0.33	0.40
radio_telescope	0.75	0.60	0.67
Egyptian_cat	0.50	0.80	0.62

Table A.3 shows details of classification of ImageNet-EEG dataset for all classes. In examining performance metrics across various classes, we note significant variances in precision, recall, and F1-scores, indicating the model's strengths and weaknesses in classifying diverse items. High-performing classes like "parachute" and "sorrel" demonstrate the model's efficacy with high precision and recall, suggesting these classes have unique, easily distinguishable features. In contrast, lower-performing classes such as "anemone fish" and "revolver" exhibit challenges in accurate identification, likely due to feature overlaps with other classes or insufficient training data. The impact of semantic similarity is evident, where high-performance classes typically show little resemblance to others, aiding their classification. For instance, "parachute" has distinct features unlike any other class. However, low-performing classes like "espresso maker" and "coffee mug" may share commonalities, leading to confusion and incorrect class prediction even if this could be considered a good semantic approximation of the context if derived from neural activity. For instance, consider a scenario where the model incorrectly identifies a "coffee mug" as an "espresso maker." While this constitutes an error, it's noteworthy that the misclassification still falls within the same semantic realm of concepts related to coffee. The employment of a knowledge distillation approach is deliberately designed to nurture such similarities, ensuring that when errors occur, they remain semantically closer to the original concept. This strategy aims to mitigate the impact of mistakes by aligning them more closely with the underlying theme or category of the target object.

Fig A.6 shows examples of the clusters pseudo-labels obtained using K-Means on the CLIP CLS embeddings of training images. Since it is based on pseudo-labels, there is not an exact match with a specific class, however we can qualita-

tively infer the following groupings:

- Outdoor and Tactical Equipment: The first row (pseudo-label 0) seems to contain items related to outdoor activities or tactical equipment
- Sports Equipment: The second row (pseudo-label 1) appears to be sports equipment, including balls and a bicycle.
- Clothing and Accessories: The third row (pseudo-label 2) includes various items of clothing and personal accessories.
- Mixed food and small objects: The fourth row (pseudo-label 3) includes small objects, cakes and ice.
- Home and Living: The fifth row (pseudo-label 4) has items commonly found in a home or associated with living spaces, like furniture and appliances.
- Plants and Nature: The sixth row (pseudo-label 5) seems to focus on plants or elements commonly found in gardens.
- Animals: The sixth row shows animals, both wild and domestic.
- Food and Kitchen: The seventh row (pseudo-label 7) has images of food and items related to the kitchen or food preparation.

Table A.4 summarizes details of performances on pseudolabels for THINGS-EEG2 dataset.

Table A.4: Detail of average classification performances for THINGS-EEG2 for each classs

Class	Precision	Recall	F1-Score
0	0.375	0.387	0.381
1	0.421	0.571	0.485
2	0.778	0.438	0.560
3	0.300	0.273	0.286
4	0.500	0.190	0.276
5	0.167	0.231	0.194
6	0.689	0.861	0.765
7	0.625	0.568	0.595

Based on the table A.4 and the image clusters (Fig A.6, we can comment on the performance of each pseudo-label cluster as follows:

Outdoor and Tactical Equipment (Class 0): This class has moderate precision and recall, suggesting the model has a reasonable ability to identify items within this cluster, but there's still room for improvement in recognizing and distinguishing these objects with greater accuracy.

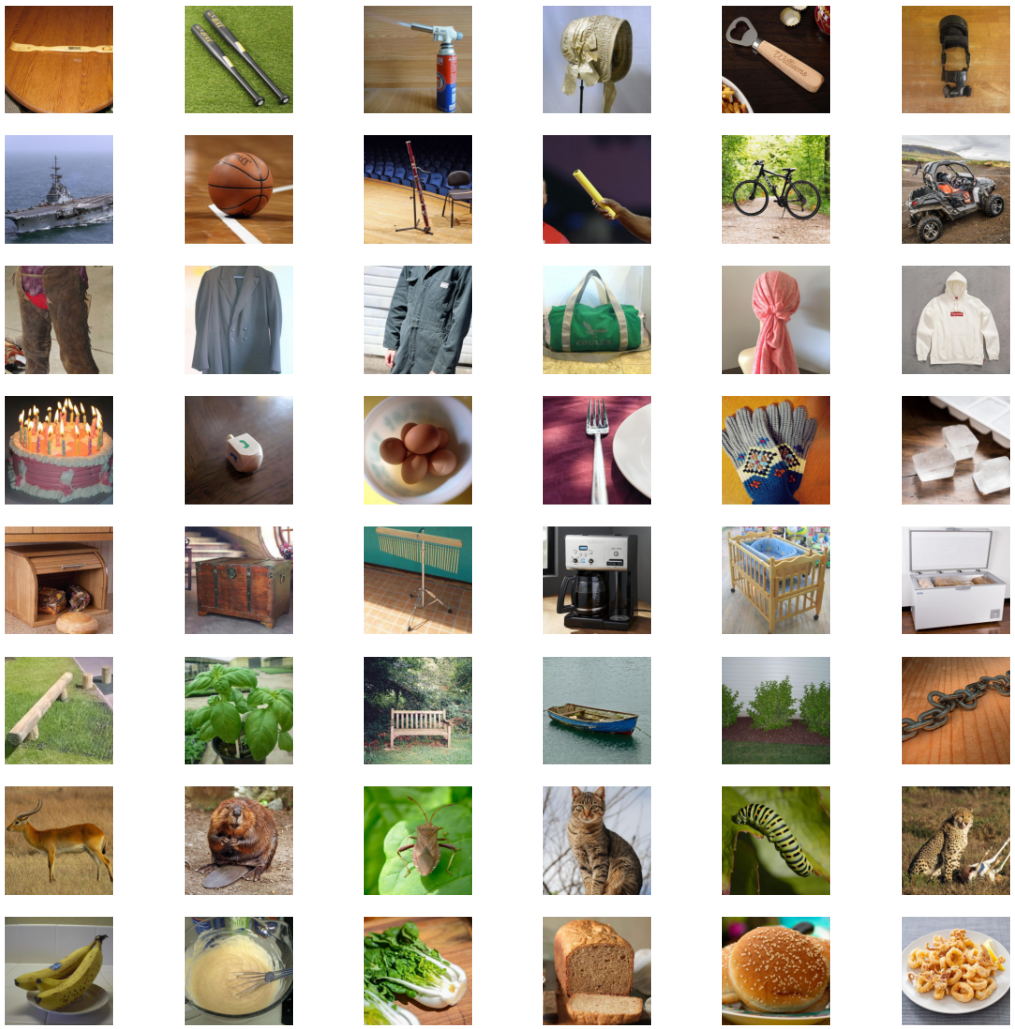


Figure A.6: Example of clusters for THINGS-EEG2. Each row is a different cluster with some examples to have a qualitative idea of the semantic content.

Sports Equipment (Class 1): With a precision slightly above average and a relatively high recall, this cluster is better identified by the model, indicating that the distinctive features of sports equipment are more easily recognized.

Clothing and Accessories (Class 2): This class has the highest precision but lower recall, which could mean that while the items classified as clothing and accessories are often correct, the model is missing quite a few actual instances of this class.

Mixed Food and Small Objects (Class 3): The low precision and recall in this cluster imply that the model struggles significantly with this category. The heterogeneity of the group may contribute to this difficulty, as it combines various unrelated items.

Home and Living (Class 4): This class also has low precision and recall scores. Similar to the mixed food and small objects class, the diversity of items in home and living could be leading to challenges in accurate classification.

Plants and Nature (Class 5): This class has the lowest performance metrics across all clusters, with both precision and recall below 0.2. It suggests that the model has substantial difficulty in recognizing and categorizing these images accurately.

Animals (Class 6): The model performs best in this class, showing high precision and recall. This indicates that the model can effectively identify and categorize animal images, which might be due to more distinctive and recognizable features in these images compared to other classes.

Food and Kitchen (Class 7): The performance here is quite good, with both precision and recall above 0.5. The model is reasonably competent at identifying items related to food and kitchen, which might be due to their specific shapes and contexts that are easier to learn.

The variations in performance across the clusters may be influenced by the intrinsic properties of the items within them. Clusters with more visually distinct and less varied items (like Animals and Food and Kitchen) are classified more accurately. In contrast, clusters containing a wide range of heterogeneous items (like Mixed Food and Small Objects and Home and Living) tend to have lower performance metrics, indicating a need for model improvements in these areas, perhaps through better feature extraction methods or more representative training data.

Bibliography

[1] Alexandre Abraham et al. "Machine learning for neuroimaging with scikit-learn". In: *Frontiers in Neuroinformatics* 8 (2014). ISSN: 1662-5196. DOI: 10.3389/fninf.2014.00014. URL: https://www.frontiersin.org/articles/10.3389/fninf.2014.00014.

[2] Hossein Adeli, Sun Minni, and Nikolaus Kriegeskorte. "Predicting brain activity using Transformers". In: *bioRxiv* (2023). DOI: 10.1101/2023.08.02.551743. eprint: https://www.biorxiv.org/content/early/2023/08/05/2023.08.02.551743.full.pdf. URL: https://www.biorxiv.org/content/early/2023/08/05/2023.08.02.551743.

[3] Andrea Agostinelli et al. *MusicLM: Generating Music From Text*. 2023. arXiv: 2301.11325 [cs.SD].

[4] Emily J. Allen et al. "A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence". In: *Nature Neuroscience* 25.1 (Jan. 2022), pp. 116–126. ISSN: 1546-1726. DOI: 10.1038/s41593-021-00962-x. URL: https://doi.org/10.1038/s41593-021-00962-x.

[5] Emily J. Allen et al. "A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence". In: *Nature Neuroscience* 25.1 (Jan. 2022), pp. 116–126. ISSN: 1546-1726. DOI: 10.1038/s41593-021-00962-x. URL: https://doi.org/10.1038/s41593-021-00962-x.

[6] R. Antonello and A. Huth. "Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data". In: *Neurobiology of Language* 5.1 (Apr. 2024), pp. 64–79. DOI: 10.1162/nol_a_00087.

[7] Richard Antonello, Aditya Vaidya, and Alexander G. Huth. *Scaling laws for language encoding models in fMRI*. 2023. arXiv: 2305.11863 [cs.CL].

[8] Richard Antonello et al. *How Many Bytes Can You Take Out Of Brain-To-Text Decoding?* 2024. arXiv: 2405.14055 [cs.CL]. URL: https://arxiv.org/abs/2405.14055.

[9] R M Awangga, T L R Mengko, and N P Utama. "A literature review of brain decoding research". en. In: *IOP Conference Series: Materials Science and Engineering* 830.3 (Apr. 2020), p. 032049. ISSN: 1757-8981, 1757-899X. DOI: 10.1088/1757-899X/830/3/032049. URL: https://iopscience.iop.org/article/10.1088/1757-899X/830/3/032049 (visited on 11/28/2022).

[10] Fakhirah Badrulhisham et al. "Machine learning and artificial intelligence in neuroscience: A primer for researchers". In: *Brain, Behavior, and Immunity* 115 (Jan. 2024), pp. 470–479. ISSN: 0889-1591. DOI: 10.1016/j.bbi.2023.11.005. URL: https://www.sciencedirect.com/science/article/pii/S0889159123003380 (visited on 02/01/2024).

[11] Yunpeng Bai et al. *DreamDiffusion: Generating High-Quality Images from Brain EEG Signals*. 2023. arXiv: 2306.16934 [cs.CV].

[12] Yunpeng Bai et al. *DreamDiffusion: Generating High-Quality Images from Brain EEG Signals*. June 30, 2023. arXiv: 2306.16934[cs]. URL: http://arxiv.org/abs/2306.16934 (visited on 07/03/2023).

[13] Moshe Bar. "Visual objects in context". en. In: *Nature Reviews Neuroscience* 5.8 (Aug. 2004). Number: 8 Publisher: Nature Publishing Group, pp. 617–629. ISSN: 1471-0048. DOI: 10.1038/nrn1476. URL: https://www.nature.com/articles/nrn1476 (visited on 11/22/2022).

[14] Thomas Bazeille et al. "An empirical evaluation of functional alignment using inter-subject decoding". In: *NeuroImage* 245 (2021), p. 118683. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2021.118683. URL: https://www.sciencedirect.com/science/article/pii/S1053811921009563.

[15] Thomas Bazeille et al. "Local Optimal Transport for Functional Brain Template Estimation". In: *IPMI 2019 - 26th International Conference on Information Processing in Medical Imaging*. hal-02278663. Hong Kong, China, June 2019. DOI: 10.1007/978-3-030-20351-1_18.

[16] L. Bellier et al. "Music can be reconstructed from human auditory cortex activity using nonlinear decoding models". In: *PLoS Biology* 21.8 (2023), e3002176. DOI: 10.1371/journal.pbio.3002176. URL: https://doi.org/10.1371/journal.pbio.3002176.

[17] Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. *Brain decoding: toward real-time reconstruction of visual perception*. 2023. arXiv: 2310.19812 [eess.IV].

[18] Yoshua Bengio et al. *Generalized Denoising Auto-Encoders as Generative Models*. 2013. arXiv: 1305.6663 [cs.LG]. URL: https://arxiv.org/abs/1305.6663.

[19] Minakshi Boruah and Ranjita Das. "CaDenseNet: a novel deep learning approach using capsule network with attention for the identification of HIV-1 integration site". In: *Neural Comput. Appl.* 35.23 (Apr. 2023), pp. 17113–17128. ISSN: 0941-0643. DOI: 10.1007/s00521-023-08585-y. URL: https://doi.org/10.1007/s00521-023-08585-y.

[20] Minakshi Boruah and Ranjita Das. "Identification of DNA motif using likelihood and attention based pooling method in the GRU framework". In: *2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*. Vol. 6. 2021, pp. 1–5. DOI: 10.1109/ICRAIE52900.2021.9704010.

[21] Hervé Bourlard and Yves Kamp. "Auto-association by multilayer perceptrons and singular value decomposition". In: *Biological Cybernetics* 59.4-5 (1988), pp. 291–294.

[22] Paul Broca. "Remarks on the seat of the faculty of articulated language, followed by an observation of aphemia". In: *Bulletin de la Société Anatomique* 6 (1861), pp. 330–357.

[23] Tom Brown, Benjamin Mann, et al. "Language Models Are Few-Shot Learners". In: *Advances in Neural Information Processing Systems* (2020), pp. 1877–1901.

[24] Lars Buitinck et al. "API design for machine learning software: experiences from the scikit-learn project". In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 2013, pp. 108–122.

[25] Erica L. Busch et al. "Hybrid hyperalignment: A single high-dimensional model of shared information embedded in cortical patterns of response and functional connectivity". In: *NeuroImage* 233 (2021), p. 117975. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2021.

117975. URL: https://www.sciencedirect.com/science/article/pii/S1053811921002524.

[26] Eglė Butkevičiūtė et al. "Removal of Movement Artefact for Mobile EEG Analysis in Sports Exercises". In: *IEEE Access* 7 (2019), pp. 7206–7217. DOI: 10.1109/ACCESS.2018.2890335.

[27] Vince D. Calhoun, Jingyu Liu, and Tülay Adalı. "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data". In: *NeuroImage* 45.1, Supplement 1 (2009). Mathematics in Brain Imaging, S163–S172. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2008.10.057. URL: https://www.sciencedirect.com/science/article/pii/S1053811908012032.

[28] Josue Ortega Caro et al. "BrainLM: A foundation model for brain activity recordings". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=RwI7ZEfR27.

[29] Mathilde Caron, Hugo Touvron, Ishan Misra, et al. "Emerging Properties in Self-Supervised Vision Transformers". In: *arXiv preprint arXiv:2104.14294* (2021).

[30] André M Carrington et al. "A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms". en. In: *BMC Med. Inform. Decis. Mak.* 20.1 (Jan. 2020), p. 4.

[31] Arantxa Casanova et al. *Instance-Conditioned GAN*. 2021. DOI: 10.48550/ARXIV.2109.05070. URL: https://arxiv.org/abs/2109.05070.

[32] B. J. Casey et al. "The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites". In: *Developmental Cognitive Neuroscience* 32 (2018), pp. 43–54. DOI: 10.1016/j.dcn.2018.03.001.

[33] C. Caucheteux, A. Gramfort, and JR. King. "Evidence of a predictive coding hierarchy in the human brain listening to speech". In: *Nature Human Behaviour* 7 (2023), pp. 430–441. DOI: 10.1038/s41562-022-01516-2.

[34] C. Caucheteux and JR. King. "Brains and algorithms partially converge in natural language processing". In: *Communications Biology* 5 (2022), p. 134. DOI: 10.1038/s42003-022-03036-1.

[35] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. "Deep language algorithms predict semantic comprehension from brain activity". en. In: *Scientific Reports* 12.1 (Sept. 2022), p. 16327. ISSN: 2045-2322. DOI: 10.1038/s41598-022-20460-9. URL: https://www.nature.com/articles/s41598-022-20460-9 (visited on 11/14/2022).

[36] Charlotte Caucheteux and Jean-Rémi King. "Brains and algorithms partially converge in natural language processing". en. In: *Communications Biology* 5.1 (Dec. 2022), p. 134. ISSN: 2399-3642. DOI: 10.1038/s42003-022-03036-1. URL: https://www.nature.com/articles/s42003-022-03036-1 (visited on 11/14/2022).

[37] Charlotte Caucheteux and Jean-Rémi King. "Brains and algorithms partially converge in natural language processing". In: *Communications Biology* 5.1 (Dec. 2022), p. 134. ISSN: 2399-3642. DOI: 10.1038/s42003-022-03036-1. URL: https://www.nature.com/articles/s42003-022-03036-1 (visited on 11/14/2022).

[38] Nadine Chang et al. "BOLD5000, a public fMRI dataset while viewing 5000 visual images". In: *Scientific Data* 6.1 (May 2019), p. 49. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0052-3. URL: https://doi.org/10.1038/s41597-019-0052-3.

[39] Po-Hsuan (Cameron) Chen et al. "A Reduced-Dimension fMRI Shared Response Model". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015.

[40] Po-Hsuan (Cameron) Chen et al. "A Reduced-Dimension fMRI Shared Response Model". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/b3967a0e938dc2a6340e2586
Paper.pdf.

[41] Shaohui Chen et al. "BEATs: Audio Pre-training with Acoustic Tokenizers". In: *arXiv preprint arXiv:2212.09058* (2022).

[42] Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. *Cinematic Mindscapes: High-quality Video Reconstruction from Brain Activity*. 2023. arXiv: 2305.11675 [cs.CV].

[43] Zijiao Chen et al. *Seeing Beyond the Brain: Conditional Diffusion Model with Sparse Masked Modeling for Vision Decoding*. 2022. arXiv: 2211.06956 [cs.CV].

[44] Zijiao Chen et al. *Seeing Beyond the Brain: Conditional Diffusion Model with Sparse Masked Modeling for Vision Decoding*. 2023. arXiv: 2211.06956 [cs.CV].

[45] Rewon Child. *Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images*. 2021. arXiv: 2011.10650 [cs.LG].

[46] Bhavin Choksi et al. "Multimodal neural networks better explain multivoxel patterns in the hippocampus". In: *Neural Networks* 154 (2022), pp. 538–542. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2022.07.033. URL: https://www.sciencedirect.com/science/article/pii/S0893608022002982.

[47] R. M. Cichy et al. *The Algonauts Project 2021 Challenge: How the Human Brain Makes Sense of a World in Motion*. 2021. arXiv: 2104.13714 [cs.CV]. URL: https://arxiv.org/abs/2104.13714.

[48] Mike X Cohen. *Analyzing Neural Time Series Data: Theory and Practice*. MIT Press, 2014.

[49] Colin Conwell et al. "What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines?" In: *bioRxiv* (2023). DOI: 10.1101/2022.03.28.485868. eprint: https://www.biorxiv.org/content/early/2023/07/01/2022.03.28.485868.full.pdf. URL: https://www.biorxiv.org/content/early/2023/07/01/2022.03.28.485868.

[50] Tolga Çukur et al. "Attention during natural vision warps semantic representation across the human brain". en. In: *Nature Neuroscience* 16.6 (June 2013), pp. 763–770. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn.3381. URL: http://www.nature.com/articles/nn.3381 (visited on 11/14/2022).

[51] Hanna Damasio and Antonio R. Damasio. "Cortical systems for retrieval of concrete knowledge: The convergence zone framework". In: *Large-scale theories of the brain*. Ed. by Christof Koch and Joel L. Davis. MIT Press, 1994, pp. 61–74.

[52] Hiroto Date et al. "Deep Learning for Natural Image Reconstruction from Electrocorticography Signals". In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2019, pp. 2331–2336. DOI: 10.1109/BIBM47256.2019.8983029.

[53] A. Défossez, C. Caucheteux, J. Rapin, et al. "Decoding speech perception from non-invasive brain recordings". In: *Nature Machine Intelligence* 5 (2023), pp. 1097–1107. DOI: 10.1038/s42256-023-00714-5. URL: https://doi.org/10.1038/s42256-023-00714-5.

[54] A. Défossez, C. Caucheteux, J. Rapin, et al. "Decoding speech perception from non-invasive brain recordings". In: *Nature Machine Intelligence* 5 (2023), pp. 1097–1107. DOI: 10.1038/s42256-023-00714-5.

[55] André Y. Denault et al. "Chapter 7 - Near-Infrared Spectroscopy". In: *Neuromonitoring Techniques*. Ed. by Hemanshu Prabhakar. Academic Press, 2018, pp. 179–233. ISBN: 978-0-12-809915-5. DOI: https://doi.org/10.1016/B978-0-12-809915-5.00007-3. URL: https://www.sciencedirect.com/science/article/pii/B9780128099155000073.

[56] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[57] Timo I. Denk et al. *Brain2Music: Reconstructing Music from Human Brain Activity*. 2023. arXiv: 2307.11078 [q-bio.NC].

[58] Timo I. Denk et al. *Brain2Music: Reconstructing Music from Human Brain Activity*. 2023. arXiv: 2307.11078 [q-bio.NC]. URL: https://arxiv.org/abs/2307.11078.

[59] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[60] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

[61] Prafulla Dhariwal and Alex Nichol. "Diffusion Models Beat GANs on Image Synthesis". In: *CoRR* abs/2105.05233 (2021). arXiv: 2105.05233. URL: https://arxiv.org/abs/2105.05233.

[62] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. "How does the brain solve visual object recognition?" In: *Neuron* 73.3 (Feb. 2012), pp. 415–434. DOI: 10.1016/j.neuron.2012.01.010.

[63] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. *Unsupervised Visual Representation Learning by Context Prediction*. 2016. arXiv: 1505.05192 [cs.CV]. URL: https://arxiv.org/abs/1505.05192.

[64] Jeff Donahue and Karen Simonyan. *Large Scale Adversarial Representation Learning*. 2019. DOI: 10.48550/ARXIV.1907.02544. URL: https://arxiv.org/abs/1907.02544.

[65] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.

[66] Bing Du et al. "fMRI Brain Decoding and Its Applications in Brain and Computer Interface: A Survey". In: *Brain Sciences* 12.2 (2022). ISSN: 2076-3425. DOI: 10.3390/brainsci12020228. URL: https://www.mdpi.com/2076-3425/12/2/228.

[67] Yiqun Duan et al. "DeWave: Discrete Encoding of EEG Waves for EEG to Text Translation". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[68] Abhimanyu Dubey et al. *The Llama 3 Herd of Models*. 2024. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

[69] Thibault Dupré La Tour et al. "Feature-space selection with banded ridge regression". In: *NeuroImage* (2022). DOI: 10.1016/j.neuroimage.2022.118928.

[70] Benjamin Elizalde et al. *CLAP: Learning Audio Concepts From Natural Language Supervision*. 2022. arXiv: 2206.04769 [cs.SD].

[71] Benjamin Elizalde et al. *CLAP: Learning Audio Concepts From Natural Language Supervision*. 2022. arXiv: 2206.04769 [cs.SD]. URL: https://arxiv.org/abs/2206.04769.

[72] Stefanie Enriquez-Geppert, René J. Huster, and Christoph S. Herrmann. "EEG-Neurofeedback as a Tool to Modulate Cognition and Behavior: A Review Tutorial". In: *Frontiers in Human Neuroscience* 11 (2017). ISSN: 1662-5161. DOI: 10.3389/fnhum.2017.00051. URL: https://www.frontiersin.org/articles/10.3389/fnhum.2017.00051.

[73] Gustav Theodor Fechner. *Elemente der Psychophysik*. Leipzig: Breitkopf und Härtel, 1860.

[74] Matteo Ferrante, Matteo Ciferri, and Nicola Toschi. *R&B – Rhythm and Brain: Cross-subject Decoding of Music from Human Brain Activity*. 2024. arXiv: 2406.15537 [q-bio.NC]. URL: https://arxiv.org/abs/2406.15537.

[75] Matteo Ferrante, Matteo Ciferri, and Nicola Toschi. "Video decoding from human fMRI data with a multi-stream sensory approach". In: *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*. 2024. URL: https://openreview.net/forum?id=ZzBMz4tYPd.

[76] Matteo Ferrante, Nicola Toschi, and Alexander Huth. "Language decoding from human brain activity via contrastive learning". In: *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*. 2024. URL: https://openreview.net/forum?id=rBQPJSPNsw.

[77] Matteo Ferrante et al. *Decoding visual brain representations from electroencephalography through Knowledge Distillation and latent diffusion models*. 2023. arXiv: 2309.07149 [eess.SP].

[78] Matteo Ferrante et al. "Multimodal decoding of human brain activity into images and text". In: *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*. Ed. by Marco Fumero et al. Vol. 243. Proceedings of Machine Learning Research. PMLR, 15 Dec 2024, pp. 87–101. URL: https://proceedings.mlr.press/v243/ferrante24a.html.

[79] Matteo Ferrante et al. "Retrieving and reconstructing conceptually similar images from fMRI with latent diffusion models and a neuro-inspired brain decoding model". In: *Journal of Neural Engineering* 21.4 (2024), p. 046001. DOI: 10.1088/1741-2552/ad593c.

[80] Matteo Ferrante et al. "Through their eyes: Multi-subject brain decoding with simple alignment techniques". In: *Imaging Neuroscience* 2 (May 2024), pp. 1–21. ISSN: 2837-6056. DOI: 10.1162/imag_a_00170. URL: https://doi.org/10.1162/imag%5C_a%5C_00170.

[81] Matteo Ferrante et al. *Towards Neural Foundation Models for Vision: Aligning EEG, MEG, and fMRI Representations for Decoding, Encoding, and Modality Conversion*. 2024. arXiv: 2411.09723 [cs.CV]. URL: https://arxiv.org/abs/2411.09723.

[82] Steven M. Frankland and Joshua D. Greene. "Concepts and Compositionality: In Search of the Brain's Language of Thought". In: *Annual Review of Psychology* 71 (2020). First published as a Review in Advance on September 24, 2019, pp. 273–303. DOI: 10.1146/annurev-psych-122216-011829.

[83] Angela D. Friederici. "The cortical language circuit: from auditory perception to sentence comprehension". In: *Trends in Cognitive Sciences* 16.5 (May 2012). Epub 2012 Apr 18, pp. 262–268. DOI: 10.1016/j.tics.2012.04.001. URL: https://doi.org/10.1016/j.tics.2012.04.001.

[84] Jack L. Gallant, Shinji Nishimoto, and Thomas Naselaris. "The brain's eye: Decoding mental images from the human brain". In: *Frontiers in Human Neuroscience* 6 (2012), p. 68. DOI: 10.3389/fnhum.2012.00068.

[85] Guy Gaziv et al. "Self-supervised Natural Image Reconstruction and Large-scale Semantic Classification from Brain Activity". en. In: *NeuroImage* 254 (July 2022), p. 119121. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2022.119121. URL: https://linkinghub.elsevier.com/retrieve/pii/S105381192200249X (visited on 11/13/2022).

[86] A. T. Gifford et al. *The Algonauts Project 2023 Challenge: How the Human Brain Makes Sense of Natural Scenes*. 2023. arXiv: 2301.03198 [cs.CV].

[87] Alessandro T Gifford et al. "A large and rich EEG dataset for modeling human visual object recognition". en. In: *Neuroimage* 264.119754 (Dec. 2022), p. 119754.

[88] Alessandro T. Gifford et al. "A large and rich EEG dataset for modeling human visual object recognition". In: *NeuroImage* 264 (Dec. 2022), p. 119754. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2022.119754. URL: https://www.sciencedirect.com/science/article/pii/S1053811922008758 (visited on 03/19/2024).

[89] Charles D. Gilbert and Mariano Sigman. "Brain States: Top-Down Influences in Sensory Processing". en. In: *Neuron* 54.5 (June 2007), pp. 677–696. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2007.05.019. URL: https://www.sciencedirect.com/science/article/pii/S0896627307003765 (visited on 11/22/2022).

[90] Sara Goering et al. "Recommendations for Responsible Development and Application of Neurotechnologies". In: *Neuroethics* 14.3 (2021). Epub 2021 Apr 29, pp. 365–386. DOI: 10.1007/s12152-021-09468-6.

[91] Gabriel Goh et al. "Multimodal Neurons in Artificial Neural Networks". In: *Distill* (2021). https://distill.pub/2021/multimodal-neurons. DOI: 10.23915/distill.00030.

[92] Ariel Goldstein et al. "Shared computational principles for language processing in humans and deep language models". en. In: *Nature Neuroscience* 25.3 (Mar. 2022), pp. 369–380. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/s41593-022-01026-4. URL: https://www.nature.com/articles/s41593-022-01026-4 (visited on 11/14/2022).

[93] Ariel Goldstein et al. "Shared computational principles for language processing in humans and deep language models". In: *Nature Neuroscience* 25.3 (Mar. 2022), pp. 369–380. ISSN: 1546-1726. DOI: 10.1038/s41593-022-01026-4. URL: https://doi.org/10.1038/s41593-022-01026-4.

[94] M. A. Goodale and A. D. Milner. "Separate visual pathways for perception and action". eng. In: *Trends in Neurosciences* 15.1 (Jan. 1992), pp. 20–25. ISSN: 0166-2236. DOI: 10.1016/0166-2236(92)90344-8.

[95] J. C. Gower. "Generalized procrustes analysis". In: *Psychometrika* 40.1 (Mar. 1975), pp. 33–51. ISSN: 1860-0980. DOI: 10.1007/BF02291478. URL: https://doi.org/10.1007/BF02291478.

[96] Kalanit Grill-Spector and Kevin Weiner. "The functional architecture of the ventral temporal cortex and its role in categorization". In: *Nature Reviews Neuroscience* 15 (2014), pp. 536–548. DOI: 10.1038/nrn3747.

[97] C. G. Gross, C. E. Rocha-Miranda, and D. B. Bender. "Visual properties of neurons in inferotemporal cortex of the Macaque". eng. In: *Journal of Neurophysiology* 35.1 (Jan. 1972), pp. 96–111. ISSN: 0022-3077. DOI: 10.1152/jn.1972.35.1.96.

[98] James V Haxby et al. "A common, high-dimensional model of the representational space in human ventral temporal cortex". en. In: *Neuron* 72.2 (Oct. 2011), pp. 404–416.

[99] James V Haxby et al. "Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies". In: *eLife* 9 (June 2020). Ed. by Chris I Baker and Floris P de Lange, e56601. ISSN: 2050-084X. DOI: 10.7554/eLife.56601. URL: https://doi.org/10.7554/eLife.56601.

[100] James V. Haxby, Andrew C. Connolly, and J. Swaroop Guntupalli. "Decoding neural representational spaces using multivariate pattern analysis". In: *Annual Review of Neuroscience* 37 (2014), pp. 435–456. DOI: 10.1146/annurev-neuro-062012-170325.

[101] James V. Haxby et al. "Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex". In: *Science* 293.5539 (2001), pp. 2425–2430. DOI: 10.1126/science.1063736.

[102] Kaiming He et al. *Masked Autoencoders Are Scalable Vision Learners*. 2021. arXiv: 2111.06377 [cs.CV]. URL: https://arxiv.org/abs/2111.06377.

[103] Kaiming He et al. "Masked Autoencoders Are Scalable Vision Learners". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 16000–16009.

[104] Martin N Hebart et al. "THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior". In: *eLife* 12 (Feb. 2023). Ed. by Morgan Barense et al., e82580. ISSN: 2050-084X. DOI: 10.7554/eLife.82580. URL: https://doi.org/10.7554/eLife.82580.

[105] Martin N Hebart et al. "THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior". In: *eLife* 12 (Feb. 2023). Ed. by Morgan Barense et al., e82580. ISSN: 2050-084X. DOI: 10.7554/eLife.82580. URL: https://doi.org/10.7554/eLife.82580.

[106] Christian Herff et al. "Brain-to-text: decoding spoken phrases from phone representations in the brain". In: *Frontiers in Neuroscience* 9 (2015), p. 217. DOI: 10.3389/fnins.2015.00217. URL: https://doi.org/10.3389/fnins.2015.00217.

[107] Katherine L. Hermann, Ting Chen, and Simon Kornblith. *The Origins and Prevalence of Texture Bias in Convolutional Neural Networks*. 2020. arXiv: 1911.09071 [cs.CV].

[108] Pablo Hernández-Cámara et al. "Measuring Human-CLIP Alignment at Different Abstraction Levels". In: *ICLR 2024 Workshop on Representational Alignment*. 2024. URL: https://openreview.net/forum?id=xQyhHjLGmj.

[109] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: 1503.02531 [stat.ML].

[110] Geoffrey E. Hinton and Richard S. Zemel. "Autoencoders, minimum description length and Helmholtz free energy". In: *Advances in Neural Information Processing Systems*. Vol. 6. Morgan Kaufmann Publishers Inc., 1994, pp. 3–10.

[111] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG]. URL: https://arxiv.org/abs/2006.11239.

[112] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. DOI: 10.48550/ARXIV.2006.11239. URL: https://arxiv.org/abs/2006.11239.

[113] Jun Kai Ho et al. "Inter-individual deep image reconstruction via hierarchical neural code conversion". In: *NeuroImage* 271 (2023), p. 120007. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2023.120007. URL: https://www.sciencedirect.com/science/article/pii/S1053811923001532.

[114] Andreas Holzinger, André Carrington, and Heimo Müller. "Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations". In: *KI - Künstliche Intelligenz* 34.2 (Jan. 2020), pp. 193–198. ISSN: 1610-1987. DOI: 10.1007/s13218-020-00636-z. URL: http://dx.doi.org/10.1007/s13218-020-00636-z.

[115] John J. Hopfield. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities". In: *Proceedings of the National Academy of Sciences* 79.8 (1982), pp. 2554–2558. DOI: 10.1073/pnas.79.8.2554. URL: https://doi.org/10.1073/pnas.79.8.2554.

[116] Tomoyasu Horikawa and Yukiyasu Kamitani. "Generic decoding of seen and imagined objects using hierarchical visual features". en. In: *Nature Communications* 8.1 (Aug. 2017), p. 15037. ISSN: 2041-1723. DOI: 10.1038/ncomms15037. URL: http://www.nature.com/articles/ncomms15037 (visited on 11/13/2022).

[117] Harold Hotelling. "Analysis of a complex of statistical variables into principal components". In: *Journal of Educational Psychology* 24.6 (1933), pp. 417–441. DOI: 10.1037/h0071325. URL: https://doi.org/10.1037/h0071325.

[118] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL]. URL: https://arxiv.org/abs/2106.09685.

[119] Qingqing Huang et al. *MuLan: A Joint Embedding of Music Audio and Natural Language*. 2022. arXiv: 2208.12415 [eess.AS].

[120] Zhizheng Huang et al. "Masked Autoencoders for Large-Scale Audio Pretraining". In: *arXiv preprint arXiv:2205.02556* (2022).

[121] Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional Magnetic Resonance Imaging*. Sinauer Associates, 2004.

[122] Alexander G. Huth, Willem A. de Heer, Thomas L. Griffiths, et al. "Natural speech reveals the semantic maps that tile human cerebral cortex". In: *Nature* 532.7600 (2016), pp. 453–458. DOI: 10.1038/nature17637. URL: https://doi.org/10.1038/nature17637.

[123] Alexander G. Huth et al. "A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain". en. In: *Neuron* 76.6 (Dec. 2012), pp. 1210–1224. ISSN: 08966273. DOI: 10.1016/j.neuron.2012.10.014. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627312009348 (visited on 11/14/2022).

[124] M. Jenkinson et al. "FSL". In: *NeuroImage* 62.2 (2012), pp. 782–790. DOI: 10.1016/j.neuroimage.2011.09.015. URL: https://doi.org/10.1016/j.neuroimage.2011.09.015.

[125] Hyejeong Jo et al. *Are EEG-to-Text Models Working?* 2024. arXiv: 2405.06459 [cs.CL]. URL: https://arxiv.org/abs/2405.06459.

[126] Ian T. Jolliffe. *Principal Component Analysis*. 2nd. New York: Springer-Verlag New York, Inc., 2002. ISBN: 978-0387954424. DOI: 10.1007/b98835. URL: https://doi.org/10.1007/b98835.

[127] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus, and Giroux, 2011.

[128] Hiroharu Kamioka et al. "Effectiveness of music therapy: a summary of systematic reviews based on randomized controlled trials of music interventions". en. In: *Patient Prefer. Adherence* 8 (May 2014), pp. 727–754.

[129] Nancy Kanwisher. "Functional specificity in the human brain: A window into the functional architecture of the mind". In: *Proceedings of the National Academy of Sciences* 107.25 (2010), pp. 11163–11170. DOI: 10.1073/pnas.1005062107.

[130] Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. "The fusiform face area: A module in human extrastriate cortex specialized for face perception". In: *Journal of Neuroscience* 17.11 (1997), pp. 4302–4311. DOI: 10.1523/JNEUROSCI.17-11-04302.1997.

[131] Isaak Kavasidis et al. "Brain2Image: Converting Brain Signals into Images". In: *Proceedings of the 25th ACM International Conference on Multimedia*. MM '17. Mountain View, California, USA: Association for Computing Machinery, 2017, pp. 1809–1817. ISBN: 9781450349062. DOI: 10.1145/3123266.3127907. URL: https://doi.org/10.1145/3123266.3127907.

[132] Isaak Kavasidis et al. "*Brain2Image*: Converting Brain Signals into Images". In: *Proceedings of the 25th ACM international conference on Multimedia*. MM '17: ACM Multimedia Conference. Mountain View California USA: ACM, Oct. 23, 2017, pp. 1809–1817. ISBN: 978-1-4503-4906-2. DOI: 10.1145/3123266.3127907. URL: https://dl.acm.org/doi/10.1145/3123266.3127907 (visited on 08/09/2023).

[133] Kendrick Kay et al. "GLMdenoise: a fast, automated technique for denoising task-based fMRI data". In: *Frontiers in Neuroscience* 7 (2013). ISSN: 1662-453X. DOI: 10.3389/fnins.2013.00247. URL: https://www.frontiersin.org/articles/10.3389/fnins.2013.00247.

[134] Kendrick N. Kay and Jack L. Gallant. "I can see what you see". In: *Nature Neuroscience* 12.3 (2009), pp. 245–246. DOI: 10.1038/nn0309-245.

[135] Alexander J. Kell and Joshua H. McDermott. "Deep neural network models of sensory systems: Windows onto the role of task constraints". In: *Current Opinion in Neurobiology* 55 (Apr. 2019). Epub 2019 Mar 15, pp. 121–132. DOI: 10.1016/j.conb.2019.02.003.

[136] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML]. URL: https://arxiv.org/abs/1312.6114.

[137] S. Koelsch. "Brain correlates of music-evoked emotions". In: *Nature Reviews Neuroscience* 15 (2014), pp. 170–180. DOI: 10.1038/nrn3666. URL: https://doi.org/10.1038/nrn3666.

[138] Stefan Koelsch. "Toward a neural basis of music perception - a review and updated model". en. In: *Front. Psychol.* 2 (June 2011), p. 110.

[139] Stefan Koelsch et al. “Investigating emotion with music: an fMRI study”. eng. In: *Human Brain Mapping* 27.3 (Mar. 2006), pp. 239–250. ISSN: 1065-9471. DOI: 10.1002/hbm.20180.

[140] Naotaka Koide-Majima, Shinji Nishimoto, and Kazuyuki Majima. “Mental image reconstruction from human brain activity: Neural decoding of mental imagery via deep neural network-based Bayesian estimation”. In: *Neural Networks* 170 (2024), pp. 349–363. DOI: 10.1016/j.neunet.2023.11.024. URL: https://doi.org/10.1016/j.neunet.2023.11.024.

[141] Naoya Koide-Majima, Shinji Nishimoto, and Kazuaki Majima. “How Can We Reconstruct Mental Imagery from Brain Activities?” Japanese. In: *Brain Nerve* 76.11 (Nov. 2024), pp. 1256–1261. DOI: 10.11477/mf.1416202768.

[142] Nikolaus Kriegeskorte and Pamela K. Douglas. “Cognitive computational neuroscience”. In: *Nature Neuroscience* 21.9 (Sept. 2018). Epub 2018 Aug 20, pp. 1148–1160. DOI: 10.1038/s41593-018-0210-5.

[143] Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. “Representational similarity analysis - connecting the branches of systems neuroscience”. In: *Frontiers in Systems Neuroscience* 2 (2008). ISSN: 1662-5137. DOI: 10.3389/neuro.06.004.2008. URL: https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008.

[144] Eline R. Kupers et al. “Principles of intensive human neuroimaging”. In: *Trends in Neurosciences* 47.11 (2024), pp. 856–864. ISSN: 0166-2236. DOI: 10.1016/j.tins.2024.09.011. URL: https://www.sciencedirect.com/science/article/pii/S0166223624001838.

[145] B. Lahner, K. Dwivedi, P. Iamshchinina, et al. “Modeling short visual events through the BOLD moments video fMRI dataset and metadata”. In: *Nature Communications* 15 (2024), p. 6241. DOI: 10.1038/s41467-024-50310-3.

[146] Brenden M. Lake and Marco Baroni. “Human-like systematic generalization through a meta-learning neural network”. In: *Nature* 623 (2023), pp. 115–121. DOI: 10.1038/s41586-023-06668-3. URL: https://doi.org/10.1038/s41586-023-06668-3.

[147] Richard D. Lange et al. “Bayesian encoding and decoding as distinct perspectives on neural coding”. en. In: *Nature Neuroscience* 26.12 (Dec. 2023). Number: 12 Publisher: Nature Publishing Group, pp. 2063–2072.

ISSN: 1546-1726. DOI: 10.1038/s41593-023-01458-6. URL: https://www.nature.com/articles/s41593-023-01458-6 (visited on 02/01/2024).

[148] A. LeBel, L. Wagner, S. Jain, et al. "A natural language fMRI dataset for voxelwise encoding models". In: *Scientific Data* 10 (2023), p. 555. DOI: 10.1038/s41597-023-02437-z.

[149] Yann LeCun et al. "A Path Towards Autonomous Machine Intelligence". In: *arXiv preprint arXiv:2205.12887* (2022).

[150] Yann LeCun. "Modèles connexionnistes de l'apprentissage". In: *Proceedings of Cognitiva 87*. Paris, France, 1987, pp. 305–313.

[151] Yann LeCun et al. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1.4 (1989), pp. 541–551. DOI: 10.1162/neco.1989.1.4.541. URL: https://doi.org/10.1162/neco.1989.1.4.541.

[152] Gregory R. Lee et al. "PyWavelets: A Python package for wavelet analysis". In: *Journal of Open Source Software* 4.36 (2019), p. 1237. DOI: 10.21105/joss.01237. URL: https://doi.org/10.21105/joss.01237.

[153] Michael A. Lepori, Thomas Serre, and Ellie Pavlick. *Break It Down: Evidence for Structural Compositionality in Neural Networks*. 2023. arXiv: 2301.10884 [cs.CL]. URL: https://arxiv.org/abs/2301.10884.

[154] Mike Lewis, Yinhan Liu, Naman Goyal, et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)* (2020).

[155] Ren Li et al. "The Perils and Pitfalls of Block Design for EEG Classification Experiments". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (2021), pp. 316–333. DOI: 10.1109/TPAMI.2020.2973153.

[156] Ren Li et al. *Training on the test set? An analysis of Spampinato et al. [31]*. 2018. arXiv: 1812.07697 [cs.CV].

[157] Ren Li et al. *Training on the test set? An analysis of Spampinato et al. [31]*. 2018. arXiv: 1812.07697 [cs.CV].

[158] Sikun Lin, Thomas Sprague, and Ambuj K Singh. *Mind Reader: Reconstructing complex images from brain activities*. 2022. arXiv: 2210.01769 [q-bio.NC].

[159] Grace W Lindsay. "Convolutional neural networks as a model of the visual system: Past, present, and future". en. In: *J. Cogn. Neurosci.* 33.10 (Sept. 2021), pp. 2017–2031.

[160] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL].

[161] Yulong Liu et al. *BrainCLIP: Bridging Brain and Visual-Linguistic Representation Via CLIP for Generic Natural Visual Stimulus Decoding*. 2023. arXiv: 2302.12971 [cs.CV].

[162] Ze Liu et al. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. 2021. arXiv: 2103.14030 [cs.CV].

[163] Stuart P. Lloyd. "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

[164] Andrew F. Luo et al. *Brain Diffusion for Visual Exploration: Cortical Discovery using Large Scale Generative Models*. 2023. arXiv: 2306.03089 [cs.CV].

[165] Yiwei Ma et al. *X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval*. 2022. arXiv: 2207.07285 [cs.CV]. URL: https://arxiv.org/abs/2207.07285.

[166] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[167] Weijian Mai and Zhijun Zhang. *UniBrain: Unify Image Reconstruction and Captioning All in One Diffusion Model from Human Brain Activity*. 2023. arXiv: 2308.07428 [cs.CV].

[168] Elizabeth Hellmuth Margulis et al. "What the music said: narrative listening across cultures". In: *Palgrave Communications* 5.1 (Nov. 2019), p. 146. ISSN: 2055-1045. DOI: 10.1057/s41599-019-0363-1. URL: https://doi.org/10.1057/s41599-019-0363-1.

[169] Eri Matsuo et al. "Generating Natural Language Descriptions for Semantic Representations of Human Brain Activity". en. In: *Proceedings of the ACL 2016 Student Research Workshop*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 22–29. DOI: 10.18653/v1/P16-3004. URL: http://aclweb.org/anthology/P16-3004 (visited on 11/13/2022).

[170] Warren S. McCulloch and Walter H. Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity". In: *Bulletin of Mathematical Biophysics* 5 (1943), pp. 115–133. DOI: 10.1007/BF02478259. URL: https://doi.org/10.1007/BF02478259.

[171] Gabriele Merlin and Mariya Toneva. *Language models and brains align due to more than next-word prediction and word-level information*. 2024. arXiv: 2212.00596 [cs.CL]. URL: https://arxiv.org/abs/2212.00596.

[172] M-Marsel Mesulam. "Large-scale neurocognitive networks and distributed processing for attention, language, and memory". In: *Annals of Neurology* 28.5 (Nov. 1990), pp. 597–613. DOI: 10.1002/ana.410280502.

[173] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. "Linguistic Regularities in Continuous Space Word Representations". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 746–751. URL: https://aclanthology.org/N13-1090.

[174] Karla L. Miller et al. "Multimodal population brain imaging in the UK Biobank prospective epidemiological study". In: *Nature Neuroscience* 19.11 (2016), pp. 1523–1536. DOI: 10.1038/nn.4393.

[175] Milad Mozafari, Leila Reddy, and Rufin VanRullen. "Reconstructing Natural Scenes from fMRI Patterns using BigBiGAN". en. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. arXiv:2001.11761 [cs, eess, q-bio]. July 2020, pp. 1–8. DOI: 10.1109/IJCNN48605.2020.9206960. URL: http://arxiv.org/abs/2001.11761 (visited on 11/13/2022).

[176] N. Murali Krishna et al. "An Efficient Mixture Model Approach in Brain-Machine Interface Systems for Extracting the Psychological Status of Mentally Impaired Persons Using EEG Signals". In: *IEEE Access* 7 (2019), pp. 77905–77914. DOI: 10.1109/ACCESS.2019.2922047.

[177] Lukas Muttenthaler et al. *Aligning Machine and Human Visual Representations across Abstraction Levels*. 2024. arXiv: 2409.06509 [cs.CV]. URL: https://arxiv.org/abs/2409.06509.

[178] Tomoya Nakai, Naoko Koide-Majima, and Shinji Nishimoto. "Correspondence of categorical and feature-based representations of music in the human brain". en. In: *Brain and Behavior* 11.1 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/brb3.1936, e01936. ISSN: 2162-3279. DOI: 10.1002/brb3.1936. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/brb3.1936 (visited on 04/27/2024).

[179] Tomoya Nakai, Naoko Koide-Majima, and Shinji Nishimoto. "Encoding and Decoding of Music-Genre Representations in the Human Brain". In: *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2018, pp. 584–589. DOI: 10.1109/SMC.2018.00108.

[180] Tomoya Nakai, Naoko Koide-Majima, and Shinji Nishimoto. "Music genre neuroimaging dataset". In: *Data in Brief* 40 (2022), p. 107675. ISSN: 2352-3409. DOI: https://doi.org/10.1016/j.dib.2021.107675. URL: https://www.sciencedirect.com/science/article/pii/S2352340921009501.

[181] Thomas Naselaris et al. "Bayesian Reconstruction of Natural Images from Human Brain Activity". en. In: *Neuron* 63.6 (Sept. 2009), pp. 902–915. ISSN: 08966273. DOI: 10.1016/j.neuron.2009.09.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627309006850 (visited on 11/14/2022).

[182] Thomas Naselaris et al. "Encoding and decoding in fMRI". In: *NeuroImage* 56.2 (2011), pp. 400–410. DOI: 10.1016/j.neuroimage.2010.07.073.

[183] Thomas Naselaris et al. "Encoding and decoding in fMRI". en. In: *NeuroImage* 56.2 (May 2011), pp. 400–410. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2010.07.073. URL: https://linkinghub.elsevier.com/retrieve/pii/S1053811910010657 (visited on 11/13/2022).

[184] Xuan-Bac Nguyen et al. *The Algonauts Project 2023 Challenge: UARK-UAlbany Team Solution*. 2023. arXiv: 2308.00262 [cs.CV].

[185] Shinji Nishimoto et al. "Reconstructing visual experiences from brain activity evoked by natural movies". In: *Current Biology* 21.19 (2011), pp. 1641–1646. DOI: 10.1016/j.cub.2011.08.031.

[186] Erfan Nozari, Dani S. Bassett, et al. "Macroscopic resting-state brain dynamics are best described by linear models". In: *Nature Biomedical Engineering* 8 (2024), pp. 7–8. DOI: 10.1038/s41551-023-01117-y.

[187] Seiji Ogawa et al. "Brain magnetic resonance imaging with contrast dependent on blood oxygenation". In: *Proceedings of the National Academy of Sciences* 87.24 (1990), pp. 9868–9872. DOI: 10.1073/pnas.87.24.9868. URL: https://www.pnas.org/content/87/24/9868.

[188] Subba Reddy Oota et al. *Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey)*. arXiv:2307.10246 [cs, q-bio]. July 2023. URL: http://arxiv.org/abs/2307.10246 (visited on 08/25/2023).

[189] F. Ozcelik and R. VanRullen. "Natural scene reconstruction from fMRI signals using generative latent diffusion". In: *Scientific Reports* 13 (2023), p. 15666. DOI: 10.1038/s41598-023-42891-8.

[190] Furkan Ozcelik and Rufin VanRullen. *Brain-Diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion*. 2023. arXiv: 2303.05334 [cs.CV].

[191] S. Palazzo et al. "Generative Adversarial Networks Conditioned by Brain Signals". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, Oct. 2017, pp. 3430–3438. ISBN: 978-1-5386-1032-9. DOI: 10.1109/ICCV.2017.369. URL: http://ieeexplore.ieee.org/document/8237631/ (visited on 08/09/2023).

[192] Simone Palazzo et al. *Correct block-design experiments mitigate temporal correlation bias in EEG classification*. 2020. arXiv: 2012.03849 [cs.CV].

[193] Simone Palazzo et al. "Generative Adversarial Networks Conditioned by Brain Signals". In: Oct. 2017, pp. 3430–3438. DOI: 10.1109/ICCV.2017.369.

[194] V. Pando-Naude, A. Patyczek, L. Bonetti, et al. "An ALE meta-analytic review of top-down and bottom-up processing of music in the brain". In: *Scientific Reports* 11 (2021), p. 20813. DOI: 10.1038/s41598-021-00139-3.

[195] Stefano Panzeri et al. "Cracking the Neural Code for Sensory Perception by Combining Statistics, Intervention, and Behavior". In: *Neuron* 93.3 (Feb. 2017), pp. 491–507. DOI: 10.1016/j.neuron.2016.12.036.

[196] Karalyn Patterson and Matthew A. Lambon Ralph. "Chapter 61 - The Hub-and-Spoke Hypothesis of Semantic Memory". In: *Neurobiology of Language*. Ed. by Gregory Hickok and Steven L. Small. San Diego: Academic Press, 2016, pp. 765–775. ISBN: 978-0-12-407794-2. DOI: https://doi.org/10.1016/B978-0-12-407794-2.00061-4. URL: https://www.sciencedirect.com/science/article/pii/B9780124077942000614.

[197] Karl Pearson. "On Lines and Planes of Closest Fit to Systems of Points in Space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. DOI: 10.1080/14786440109462720. URL: https://doi.org/10.1080/14786440109462720.

[198] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. "WordNet::Similarity: Measuring the Relatedness of Concepts". In: *Demonstration Papers at HLT-NAACL 2004*. HLT-NAACL–Demonstrations '04. Boston, Massachusetts: Association for Computational Linguistics, 2004, pp. 38–41.

[199] Francisco Pereira et al. "Toward a universal decoder of linguistic meaning from brain activation". In: *Nature Communications* 9 (2018), p. 963. DOI: 10.1038/s41467-018-03068-4. URL: https://doi.org/10.1038/s41467-018-03068-4.

[200] Ronald C. Petersen et al. "Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization". In: *Neurology* 74.3 (Jan. 2010). Epub 2009 Dec 30, pp. 201–209. DOI: 10.1212/WNL.0b013e3181cb3e25.

[201] Steven T. Piantadosi and et al. "Why concepts are (probably) vectors". In: *Trends in Cognitive Sciences* 28.9 (2024). Trends in Cognitive Sciences, Volume 28, Issue 9, pp. 844–856. DOI: 10.1016/j.tics.2024.07.004.

[202] S. Pinker. *The Language Instinct: How the Mind Creates Language*. Harper-Perennial modern classics. HarperCollins, 2000. ISBN: 9780060958336. URL: https://books.google.it/books?id=ednWUqVRFpgC.

[203] Sean F. Popham et al. "Visual and linguistic semantic representations are aligned at the border of human visual cortex". In: *Nature Neuroscience* 24.11 (Nov. 2021). Epub 2021 Oct 28, pp. 1628–1636. DOI: 10.1038/s41593-021-00921-6. URL: https://doi.org/10.1038/s41593-021-00921-6.

[204] Pavel Popov et al. "A simple but tough-to-beat baseline for fMRI time-series classification". In: *NeuroImage* 303 (2024), p. 120909. DOI: 10.1016/j.neuroimage.2024.120909.

[205] Jacob S Prince et al. "Improving the accuracy of single-trial fMRI response estimates using GLMsingle". In: *eLife* 11 (Nov. 2022). Ed. by Peter Kok et al., e77599. ISSN: 2050-084X. DOI: 10.7554/eLife.77599. URL: https://doi.org/10.7554/eLife.77599.

[206] Timothée Proix et al. "Imagined speech can be decoded from low- and cross-frequency intracranial EEG features". In: *Nature Communications* 13 (2022), p. 48. DOI: 10.1038/s41467-021-27725-3. URL: https://doi.org/10.1038/s41467-021-27725-3.

[207] Kai Qiao et al. "Accurate Reconstruction of Image Stimuli From Human Functional Magnetic Resonance Imaging Based on the Decoding Model With Capsule Network Architecture". en. In: *Frontiers in Neuroinformatics* 12 (Sept. 2018), p. 62. ISSN: 1662-5196. DOI: 10.3389/fninf.2018.00062. URL: https://www.frontiersin.org/article/10.3389/fninf.2018.00062/full (visited on 11/13/2022).

[208] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2019).

[209] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].

[210] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].

[211] Alfredo Raglio et al. "Effects of active music therapy on the normal brain: fMRI based evidence". eng. In: *Brain Imaging and Behavior* 10.1 (Mar. 2016), pp. 182–186. ISSN: 1931-7565. DOI: 10.1007/s11682-015-9380-x.

[212] Alfredo Raglio et al. "Music in the workplace: A narrative literature review of intervention studies". eng. In: *Journal of Complementary & Integrative Medicine* (Oct. 2019), /j/jcim.ahead–of–print/jcim–2017–0046/jcim–2017–0046.xml. ISSN: 1553-3840. DOI: 10.1515/jcim-2017-0046.

[213] Matthew A Lambon Ralph et al. "The neural and computational bases of semantic cognition". en. In: *Nat. Rev. Neurosci.* 18.1 (Jan. 2017), pp. 42–55.

[214] Aditya Ramesh et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. arXiv: 2204.06125 [cs.CV].

[215] Aditya Ramesh et al. *Zero-Shot Text-to-Image Generation*. 2021. DOI: 10.48550/ARXIV.2102.12092. URL: https://arxiv.org/abs/2102.12092.

[216] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. *Vision Transformers for Dense Prediction*. 2021. arXiv: 2103.13413 [cs.CV].

[217] Leila Reddy and Nancy Kanwisher. "Coding of visual objects in the ventral stream". In: *Current Opinion in Neurobiology* 16.4 (2006), pp. 408–414. DOI: 10.1016/j.conb.2006.06.004.

[218] Ziqi Ren et al. *Reconstructing Perceived Images from Brain Activity by Visually-guided Cognitive Representation and Adversarial Learning*. en. arXiv:1906.12181 [cs]. Oct. 2019. URL: http://arxiv.org/abs/1906.12181 (visited on 11/13/2022).

[219] Douglas Reynolds. "Gaussian Mixture Models". In: *Encyclopedia of Biometrics*. Ed. by Stan Z. Li and Anil K. Jain. Boston, MA: Springer, 2009, pp. 659–663. DOI: 10.1007/978-0-387-73003-5_196. URL: https://doi.org/10.1007/978-0-387-73003-5_196.

[220] Hugo Richard et al. *Fast shared response model for fMRI data*. 2019. arXiv: 1909.12537 [cs.CV].

[221] Blake A. Richards et al. "A deep learning framework for neuroscience". en. In: *Nature Neuroscience* 22.11 (Nov. 2019). Number: 11 Publisher: Nature Publishing Group, pp. 1761–1770. ISSN: 1546-1726. DOI: 10.1038/s41593-019-0520-2. URL: https://www.nature.com/articles/s41593-019-0520-2 (visited on 02/01/2024).

[222] Amanda K Robinson, Tijl Grootswagers, and Thomas A Carlson. "The influence of image masking on object representations during rapid serial visual presentation". en. In: *Neuroimage* 197 (Aug. 2019), pp. 224–231.

[223] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021. DOI: 10.48550/ARXIV.2112.10752. URL: https://arxiv.org/abs/2112.10752.

[224] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: 2112.10752 [cs.CV]. URL: https://arxiv.org/abs/2112.10752.

[225] Frank Rosenblatt. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain". In: *Psychological Review* 65.6 (1958), pp. 386–408. DOI: 10.1037/h0042519. URL: https://doi.org/10.1037/h0042519.

[226] Tiasha Saha Roy et al. "Do vision and imagery share common principal signal components?" English. In: *Conference on Cognitive Computational Neuroscience (CCN)* (2024).

[227] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning internal representations by error propagation". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Ed. by David E. Rumelhart and James L. McClelland. Cambridge, MA: MIT Press, 1986, pp. 318–362.

[228] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning Representations by Back-Propagating Errors". In: *Nature* 323 (1986), pp. 533–536. DOI: 10.1038/323533a0. URL: https://doi.org/10.1038/323533a0.

[229] Jacob Russin et al. *From Frege to chatGPT: Compositionality in language, cognition, and deep neural networks*. 2024. arXiv: 2405.15164 [cs.NE]. URL: https://arxiv.org/abs/2405.15164.

[230] Motoshige Sato et al. *Scaling Law in Neural Data: Non-Invasive Speech Decoding with 175 Hours of EEG Data*. 2024. arXiv: 2407.07595 [q-bio.NC]. URL: https://arxiv.org/abs/2407.07595.

[231] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. "Learnable latent embeddings for joint behavioural and neural analysis". In: *Nature* (May 3, 2023). Publisher: Nature Publishing Group, pp. 1–9. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06031-6. URL: https://www.nature.com/articles/s41586-023-06031-6 (visited on 05/07/2023).

[232] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. "Learnable latent embeddings for joint behavioural and neural analysis". In: *Nature* 617.7960 (May 2023), pp. 360–368. ISSN: 1476-4687. DOI: 10.1038/

s41586-023-06031-6. URL: https://doi.org/10.1038/s41586-023-06031-6.

[233] Martin Schrimpf et al. "The neural architecture of language: Integrative modeling converges on predictive processing". In: *Proceedings of the National Academy of Sciences* 118.45 (2021), e2105646118. DOI: 10.1073/pnas.2105646118. URL: https://doi.org/10.1073/pnas.2105646118.

[234] Gunter Schumann et al. "The IMAGEN study: Reinforcement-related behaviour in normal brain function and psychopathology". In: *Molecular Psychiatry* 15.12 (2010), pp. 1128–1139. DOI: 10.1038/mp.2010.4.

[235] Paul S. Scotti et al. *MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data*. 2024. arXiv: 2403.11207 [cs.CV].

[236] Paul S. Scotti et al. *Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors*. 2023. arXiv: 2305.18274 [cs.CV].

[237] Guohua Shen et al. "End-to-end deep image reconstruction from human brain activity". en. In: *Front. Comput. Neurosci.* 13 (Apr. 2019), p. 21.

[238] Changhao Shi et al. *Exploring Compositional Visual Generation with Latent Classifier Guidance*. 2023. arXiv: 2304.12536 [cs.CV]. URL: https://arxiv.org/abs/2304.12536.

[239] Tianwei Shi et al. "Brain Computer Interface Based on Motor Imagery for Mechanical Arm Grasp Control". In: *Inf. Technol. Control.* 52 (2023), pp. 358–366. URL: https://api.semanticscholar.org/CorpusID:259927745.

[240] Ken Shirakawa et al. *Spurious reconstruction from brain activity*. 2024. arXiv: 2405.10078 [q-bio.NC]. URL: https://arxiv.org/abs/2405.10078.

[241] A. B. Silva, K. T. Littlejohn, J. R. Liu, et al. "The speech neuroprosthesis". In: *Nature Reviews Neuroscience* 25 (2024), pp. 473–492. DOI: 10.1038/s41583-024-00819-9. URL: https://doi.org/10.1038/s41583-024-00819-9.

[242] Prajwal Singh et al. *EEG2IMAGE: Image Reconstruction from EEG Brain Signals*. Mar. 18, 2023. arXiv: 2302.10121[cs,q-bio]. URL: http://arxiv.org/abs/2302.10121 (visited on 08/09/2023).

[243] Burrhus Frederic Skinner. *The behavior of organisms: An experimental analysis*. New York: Appleton-Century-Crofts, 1938.

[244] C. Spampinato et al. "Deep Learning Human Mind for Automated Visual Classification". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, July 2017, pp. 4503–4511. ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.2017.479. URL: http://ieeexplore.ieee.org/document/8099962/ (visited on 08/09/2023).

[245] Concetto Spampinato et al. *Deep Learning Human Mind for Automated Visual Classification*. 2019. arXiv: 1609.00344 [cs.CV].

[246] Stanley S. Stevens. "On the psychophysical law". In: *Psychological Review* 64.3 (1957), pp. 153–181. DOI: 10.1037/h0046162.

[247] Abdulhamit Subasi. "EEG signal classification using wavelet feature extraction and a mixture of expert model". In: *Expert Systems with Applications* 32.4 (2007), pp. 1084–1093. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2006.02.005. URL: https://www.sciencedirect.com/science/article/pii/S0957417406000844.

[248] Jingyuan Sun et al. *NeuroCine: Decoding Vivid Video Sequences from Human Brain Activties*. 2024. arXiv: 2402.01590 [cs.CV]. URL: https://arxiv.org/abs/2402.01590.

[249] Saya Takada et al. "Generation of Viewed Image Captions From Human Brain Activity Via Unsupervised Text Latent Space". In: *2020 IEEE International Conference on Image Processing (ICIP)*. 2020 IEEE International Conference on Image Processing (ICIP). ISSN: 2381-8549. Oct. 2020, pp. 2521–2525. DOI: 10.1109/ICIP40778.2020.9191262.

[250] Yu Takagi and Shinji Nishimoto. "High-resolution image reconstruction with latent diffusion models from human brain activity". In: *bioRxiv* (2023). DOI: 10.1101/2022.11.18.517004. eprint: https://www.biorxiv.org/content/early/2023/03/11/2022.11.18.517004.full.pdf. URL: https://www.biorxiv.org/content/early/2023/03/11/2022.11.18.517004.

[251] J. Tang, A. LeBel, S. Jain, et al. "Semantic reconstruction of continuous language from non-invasive brain recordings". In: *Nature Neuroscience* 26 (2023), pp. 858–866. DOI: 10.1038/s41593-023-01304-9.

[252] Paul M. Thompson et al. "The ENIGMA Consortium: Large-scale collaborative analyses of neuroimaging and genetic data". In: *Brain Imaging and Behavior* 8.2 (2014), pp. 153–182. DOI: 10.1007/s11682-013-9269-5.

[253] Edward C. Tolman. "Cognitive maps in rats and men". In: *Psychological Review* 55.4 (1948), pp. 189–208. DOI: 10.1037/h0061626.

[254] Mariya Toneva and Leila Wehbe. "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/749a8e6c231831ef7756db230b4359c8-Paper.pdf.

[255] Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. "LLaMA: Open and Efficient Foundation Language Models". In: *arXiv preprint arXiv:2302.13971* (2023).

[256] Anne M. Treisman and Garry Gelade. "A feature-integration theory of attention". In: *Cognitive Psychology* 12.1 (1980), pp. 97–136. DOI: 10.1016/0010-0285(80)90005-5.

[257] Doris Y. Tsao et al. "Faces and objects in macaque cerebral cortex". In: *Nature Neuroscience* 6.9 (Sept. 2003), pp. 989–995. DOI: 10.1038/nn1111.

[258] L. G. Ungerleider and J. V. Haxby. "'What' and 'where' in the human brain". eng. In: *Current Opinion in Neurobiology* 4.2 (Apr. 1994), pp. 157–165. ISSN: 0959-4388. DOI: 10.1016/0959-4388(94)90066-3.

[259] Laurens J. P. Van der Maaten and Geoffrey E. Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605. URL: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

[260] David C Van Essen et al. "The WU-Minn Human Connectome Project: an overview". en. In: *Neuroimage* 80 (Oct. 2013), pp. 62–79.

[261] Marcel A. J. Van Gerven. "A primer on encoding models in sensory neuroscience". In: *Journal of Mathematical Psychology* 76 (2017), pp. 172–183. DOI: 10.1016/j.jmp.2016.06.004.

[262] Rufin VanRullen and Leila Reddy. "Reconstructing faces from fMRI patterns using deep generative neural networks". In: *Communications Biology* 2.1 (May 2019), p. 193. ISSN: 2399-3642. DOI: 10.1038/s42003-019-0438-y. URL: https://doi.org/10.1038/s42003-019-0438-y.

[263] Ashish Vaswani et al. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[264] Kola Venu and P. Natesan. "Optimized Deep Learning Model Using Modified Whale's Optimization Algorithm for EEG Signal Classification". In: *Information Technology and Control* 52 (Sept. 2023), pp. 744–760. DOI: 10.5755/j01.itc.52.3.33320.

[265] Mai-Anh T. Vu et al. "A Shared Vision for Machine Learning in Neuroscience". In: *The Journal of Neuroscience* 38.7 (Feb. 2018), pp. 1601–1607. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.0508-17.2018. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5815449/ (visited on 02/01/2024).

[266] Jianfeng Wang et al. *GIT: A Generative Image-to-text Transformer for Vision and Language*. Dec. 15, 2022. arXiv: 2205.14100[cs]. URL: http://arxiv.org/abs/2205.14100 (visited on 04/23/2023).

[267] Ran Wang and Zhe Sage Chen. *Large-scale Foundation Models and Generative AI for BigData Neuroscience*. 2023. arXiv: 2310.18377 [q-bio.NC].

[268] Zhenhailong Wang and Heng Ji. "Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 5. 2022, pp. 5350–5358.

[269] J D Warren and T D Griffiths. "Distinct mechanisms for processing spatial sequences and pitch sequences in the human auditory brain". en. In: *J. Neurosci.* 23.13 (July 2003), pp. 5799–5804.

[270] Jason Warren. "How does the brain process music?" en. In: *Clin. Med.* 8.1 (Feb. 2008), pp. 32–36.

[271] Haiguang Wen et al. "Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision". In: *Cerebral Cortex* 28.12 (Oct. 2017), pp. 4136–4160. ISSN: 1047-3211. DOI: 10.1093/cercor/bhx268. eprint: https://academic.oup.com/cercor/article-pdf/28/12/4136/26338870/bhx268.pdf. URL: https://doi.org/10.1093/cercor/bhx268.

[272] Carl Wernicke. *Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis*. Breslau: Cohn & Weigert, 1874.

[273] Francis R. Willett et al. "A high-performance speech neuroprosthesis". In: *Nature* 620 (2023), pp. 1031–1036. DOI: 10.1038/s41586-023-06377-x. URL: https://doi.org/10.1038/s41586-023-06377-x.

[274] Martina de Witte et al. "Music therapy for stress reduction: a systematic review and meta-analysis". en. In: *Health Psychol. Rev.* 16.1 (Mar. 2022), pp. 134–159.

[275] Thomas Wolf et al. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. 2019. DOI: 10.48550/ARXIV.1910.03771. URL: https://arxiv.org/abs/1910.03771.

[276] Meng-Huan Wu et al. "Analogy-Related Information Can Be Accessed by Simple Addition and Subtraction of fMRI Activation Patterns, Without Participants Performing any Analogy Task". In: *Neurobiology of Language* 3.1 (Feb. 2022), pp. 1–17. ISSN: 2641-4368. DOI: 10.1162/nol_a_00045. eprint: https://direct.mit.edu/nol/article-pdf/3/1/1/1986844/nol_a_00045.pdf. URL: https://doi.org/10.1162/nol%5C_a%5C_00045.

[277] Weihao Xia et al. *DREAM: Visual Decoding from Reversing Human Visual System*. 2023. arXiv: 2310.02265 [cs.CV].

[278] Xingqian Xu et al. *Versatile Diffusion: Text, Images and Variations All in One Diffusion Model*. 2024. arXiv: 2211.08332 [cs.CV]. URL: https://arxiv.org/abs/2211.08332.

[279] Huzheng Yang, James Gee, and Jianbo Shi. *Memory Encoding Model*. arXiv:2308.01175 [cs]. Aug. 2023. DOI: 10.48550/arXiv.2308.01175. URL: http://arxiv.org/abs/2308.01175 (visited on 09/19/2023).

[280] Yiqian Yang et al. *NeuSpeech: Decode Neural signal as Speech*. 2024. arXiv: 2403.01748 [cs.CL]. URL: https://arxiv.org/abs/2403.01748.

[281] Hyun-Joon Yoo, Hyun Im Moon, and Sung-Bom Pyun. "Amusia after right temporoparietal lobe infarction: A case report". en. In: *Ann. Rehabil. Med.* 40.5 (Oct. 2016), pp. 933–937.

[282] Rafael Yuste, Sara Goering, and et al. "Four ethical priorities for neurotechnologies and AI". In: *Nature* 551 (2017), pp. 159–163. DOI: 10.1038/551159a.

[283] Raheel Zafar et al. "Decoding of visual information from human brain activity: A review of fMRI and EEG studies". en. In: *Journal of Integrative Neuroscience* 14.02 (June 2015), pp. 155–168. ISSN: 0219-6352, 1757-448X. DOI: 10.1142/S0219635215500089. URL: http://www.worldscientific.com/doi/abs/10.1142/S0219635215500089 (visited on 11/13/2022).

[284] Lvmin Zhang and Maneesh Agrawala. *Adding Conditional Control to Text-to-Image Diffusion Models*. 2023. arXiv: 2302.05543 [cs.CV].

[285] Yukun Zhu et al. "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books". In: *The IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015.

[286] Yukun Zhu et al. "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books". In: *The IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015.

Acknowledgments

Completing this PhD has been one of the most challenging and rewarding journeys of my life, and it would not have been possible without the support, guidance, and encouragement of many people to whom I am deeply grateful.

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Nicola Toschi, for his unwavering support, invaluable guidance, and belief in my abilities throughout this process. His insights and expertise have not only shaped this research but have also inspired my growth as a scientist. Thank you for challenging me to think critically and for providing the freedom to explore ideas while always offering a steady hand when needed.

I am deeply grateful to all my lab members and collaborators for their stimulating discussions, constructive feedback, and shared enthusiasm for science. I really loved the environment we created where we're not only collegues but often also friends and nice people. I really felt the unique and rare privilege of going at work with joy, sure to find a good place, full of friendship, ideas and energy. Keep going this way!

A big and special thank you to Tommaso for his technical expertise, countless brainstorming sessions, and lot of encouragement during challenging moments. Sharing this PhD journey with you has been an incredible experience.

To Marianna and Matteo, working alongside you and exchanging ideas has been truly inspiring. Your insights and conversations brought both intellectual stimulation and much-needed humor to the lab, making this journey all the more memorable.

I would like to extend my heartfelt gratitude to the external professors who hosted me at various stages throughout my PhD. Marco L. Loggia, Rufin VanRullen, and Alexander Huth—thank you for welcoming me into your labs and for the time, energy, and mentorship you generously offered. The opportunity

to live in different countries, immerse myself in new research environments, and explore novel ideas was an invaluable experience that deeply enriched both my personal and academic growth. I am truly grateful for your support and guidance.

I would also like to extend my gratitude to my examiners and committee members for their time, thoughtful comments, and valuable insights that helped refine my work. Their constructive feedback has significantly improved the quality of this thesis.

To my friends, thank you for providing balance, laughter, and perspective throughout this journey. Your belief in me kept my spirits high even during the most stressful times.

Finally, I am profoundly grateful to my family. To my parents, whose endless love and sacrifices have always supported my dreams—thank you for teaching me honesty and dedication. To my partner, Sofia, your patience, understanding, and constant encouragement have been my anchor through every challenge.

This thesis is dedicated to all of you. Thank you for making this journey possible.