Information Systems Frontiers
https://doi.org/10.1007/s10796-026-10741-3

Concept-Enhanced Multimodal RAG: Towards Interpretable and Accurate Radiology Report Generation

Marco Salmè[1] · Federico Siciliano[5] · Fabrizio Silvestri[2] · Paolo Soda[1,3] · Rosa Sicilia[4] · Valerio Guarrasi[1]

Received: 16 December 2025 / Accepted: 31 March 2026

Abstract

Radiology Report Generation (RRG) through Vision-Language Models (VLMs) promises to reduce documentation burden, improve reporting consistency, and accelerate clinical workflows. However, their clinical adoption remains limited by the lack of interpretability and the tendency to hallucinate findings misaligned with imaging evidence. Existing research typically treats interpretability and accuracy as separate objectives, with concept-based explainability techniques focusing primarily on transparency, while Retrieval-Augmented Generation (RAG) methods targeting factual grounding through external retrieval. We present Concept-Enhanced Multimodal RAG (CEMRAG), a unified framework that decomposes visual representations into interpretable clinical concepts and integrates them with multimodal RAG. This approach exploits enriched contextual prompts for RRG, improving both interpretability and factual accuracy. Experiments on MIMIC-CXR and IU X-Ray across multiple VLM architectures, training regimes, and retrieval configurations demonstrate consistent improvements over both conventional RAG and concept-only baselines on clinical accuracy metrics and standard NLP measures. These results challenge the assumed trade-off between interpretability and performance, showing that transparent visual concepts can enhance rather than compromise diagnostic accuracy in medical VLMs. Our modular design decomposes interpretability into visual transparency and structured language model conditioning, providing a principled pathway toward clinically trustworthy AI-assisted radiology. The project page is available at https://github.com/marcosal30/cemrag-rrg.

Keywords Radiology report generation · Vision-Language Models · Medical imaging · Interpretability · Retrieval-Augmented Generation · Multimodal AI

Rosa Sicilia and Valerio Guarrasi contributed equally to this work.

✉ Paolo Soda
paolo.soda@umu.se

Marco Salmè
marco.salme@unicampus.it

Federico Siciliano
federico.siciliano@ingv.it

Fabrizio Silvestri
fsilvestri@diag.uniroma1.it

Rosa Sicilia
rosa.sicilia@unicamillus.org

Valerio Guarrasi
valerio.guarrasi@unicampus.it

[1] Department of Engineering, Research Unit of Artificial Intelligence and Computer Systems, Università Campus Bio-Medico of Roma, Rome, Italy

[2] Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy

[3] Department of Diagnostics and Intervention, Radiation Physics, Biomedical Engineering, Umeå University, Umeå, Sweden

[4] UniCamillus-Saint Camillus International University of Health Sciences, Rome, Italy

[5] National Institute of Geophysics and Volcanology, Rome, Italy

Published online: 11 May 2026

1 Introduction

Vision-Language Models (VLMs) (Zhang et al., 2024) have emerged as a breakthrough technology in medical imaging. By jointly modeling images and textual data, they have demonstrated remarkable capabilities across several clinical applications, including visual question answering, image classification, disease diagnosis, and automated report generation (Van et al., 2024; Hartsock & Rasool, 2024). Among these applications, Radiology Report Generation (RRG) presents a particularly challenging task: generating comprehensive textual reports from medical images that accurately describe imaging findings and identify potential pathologies. In this domain, VLMs offer the potential to streamline radiological workflows by automating the initial drafting of reports. However, despite these advantages, the adoption of VLMs in clinical settings is limited by two critical factors. Firstly, VLMs lack interpretability (Zhao et al., 2024b), operating as black boxes that do not reveal how visual evidence observed in medical images translates into diagnostic statements within generated reports. Without visibility into the anatomical structures or radiological patterns that support specific diagnostic statements, clinicians cannot verify the model's reasoning, undermining both clinical trust and patient safety. Secondly, VLMs are prone to hallucinations (Huang et al., 2025), producing medically inaccurate statements that are misaligned with imaging findings, such as reporting non-existent pathologies, incorrect anatomical localizations, or impressions inconsistent with observed abnormalities. Such inaccuracies pose particular concerns in radiology, where the recognition of subtle findings and the precise alignment between visual evidence and domain-specific terminology are essential for reliable diagnostic support.

Current research efforts have predominantly tackled these issues independently. Existing interpretability approaches (Vatsa et al., 2024) provide useful insights but often operate as post-hoc explanations that do not meaningfully influence the model's predictions. Recently, techniques such as Sparse Linear Concept Embeddings (SpLiCE) (Bhalla et al., 2024) have shown that visual embeddings can be decomposed into sparse, human-interpretable concepts without requiring manual annotations, offering a scalable pathway toward transparency in VLMs. In parallel, recent work aimed at improving factual grounding has increasingly relied on Retrieval-Augmented Generation (RAG) (Gao et al., 2023). By retrieving similar cases and reports from a database, RAG provides external context to ground outputs in existing knowledge, reducing hallucinations and improving clinical relevance in RRG (Bernardi et al., 2024; Sun et al., 2024). Yet, RAG-based models remain limited by retrieval errors, as the context may be insufficient, noisy, redundant or irrelevant, leading to diluting the model's focus, potential misattribution, and even factual inconsistencies (Yu et al., 2024).

These two research areas, interpretability and factual grounding, have evolved largely in isolation. This separation is reinforced by the widespread assumption that transparency and performance trade off against each other, by the empirical tendency of deeper, more accurate models to be harder to interpret, and by the practical difficulty of building systems that are both highly accurate and transparently reasoned (Ennab & Mcheick, 2024). We challenge this assumption by asking: *Can interpretable visual concepts be integrated into retrieval-augmented report generation to jointly improve transparency and factual accuracy in medical VLMs?*

To answer this question, we present Concept-Enhanced Multimodal RAG (CEMRAG), a unified framework that reconciles interpretability and factual accuracy in RRG by integrating interpretable visual concept extraction with multimodal RAG. The core innovation of our approach lies in transforming interpretable visual concepts from passive post-hoc explanations into active components of the generation pipeline, using them to prioritize clinically pertinent portions of retrieved content and direct the model toward information supported by visual evidence. Our experimental framework spans two established radiology benchmarks, MIMIC-CXR and IU-Xray, which differ substantially in scale and report characteristics. We examine both in-domain retrieval scenarios, where similar cases are retrieved from the same dataset, and cross-domain retrieval settings, where reports are retrieved from an external database, reflecting realistic clinical deployments where knowledge bases may originate from different institutional sources. For each retrieval configuration, we evaluate two distinct VLM architectures under two training paradigms: a Zero-Shot prompting setting that assesses the immediate effectiveness of CEMRAG without model adaptation, and a Supervised Fine-Tuning (SFT) regime that examines how interpretable concepts interact with task-specific optimization.

Our main contributions are summarized as follows:

- We propose CEMRAG, a framework that integrates interpretable visual decomposition with retrieval-based grounding to enhance transparency and factual accuracy in RRG.
- We provide the first systematic comparison of RAG and SFT paradigms in RRG, establishing a comprehensive benchmark that evaluates their individual and combined effectiveness across two VLM architectures, two retrieval configurations, and two radiology datasets.
- We demonstrate that CEMRAG consistently outperforms traditional RAG and concept-based approaches

on both Natural Language Processing (NLP) and clinical accuracy metrics across diverse experimental conditions.
- We provide empirical evidence that interpretable visual concepts can enhance rather than compromise factual accuracy, challenging the assumption of a trade-off between interpretability and performance in medical AI systems.

The remainder of this paper is organized as follows. We begin by reviewing related work on interpretability for VLMs and multimodal RAG in Section 2, then present our proposed framework in Section 3. Section 4 describes the experimental setup, including datasets, baselines, and evaluation metrics. We present both quantitative and qualitative results in Section 5, and conclude with discussion and future research directions in Section 6.

2 Related Work

Our work builds upon two complementary research areas that we review in turn. First, we discuss interpretability methods for VLMs, then we examine the application of multimodal RAG to medical domains.

2.1 Interpretability for Vision-Language Models

The interpretability of VLMs has emerged as a critical requirement for clinical deployment, particularly in medical imaging where diagnostic decisions directly impact patient outcomes. However, most interpretability research focuses on classification tasks, leaving the integration of transparency mechanisms into generative frameworks like RRG largely unexplored.

Current interpretability methods can be distinguished by whether they rely on implicit explanation mechanisms or explicit concept representations. In the first category, post-hoc textual explanation techniques, exemplified by rationale generation (Park et al., 2018; Nguyen et al., 2024) and Chain-of-Thought reasoning (Wei et al., 2022; Zheng et al., 2023), train models to articulate their reasoning processes after making predictions through natural language. While these methods have demonstrated improvements in perceived transparency and, in some cases, task performance, they often act as plausible rationalizations rather than faithful reflections of the underlying computational mechanisms (Chen et al., 2024). This limitation is particularly concerning in medical settings, where clinicians require insight into the actual diagnostic cues driving the model's predictions.

Mechanistic Interpretability (Bereska & Gavves, 2024; Sharkey et al., 2025; Jiang et al., 2025) represents a more ambitious approach, seeking to reverse-engineer neural network computations through detailed analysis of attention patterns, information flow, and component functionality (Huo et al., 2024; Conmy et al., 2023). Despite its theoretical appeal in pursuing causal rather than correlational understanding, the architectural complexity of modern large-scale VLMs renders this approach extremely challenging to implement systematically, especially in real-world medical AI systems.

An alternative paradigm achieves interpretability through explicit concept representations. Concept Bottleneck Models (Koh et al., 2020; Rao et al., 2024) exemplify this approach by forcing all information to flow through an intermediate layer of predefined human-interpretable concepts, ensuring that predictions depend exclusively on explicit semantic attributes. This architectural constraint offers genuine transparency but comes at a substantial cost: it demands extensive manual concept annotation and restricts model expressiveness, limiting the ability to capture visual patterns that fall outside the predefined space.

Recent work on representation decomposition offers a more flexible concept-based approach by decomposing dense embeddings into interpretable components without requiring manual annotations or architectural constraints. Methods such as those proposed by Gandelsman et al. (2023); Balasubramanian et al. (2024) leverage internal model components such as attention mechanisms to translate high-dimensional representations into natural language descriptions, revealing how models encode visual features and their relationships to semantic concepts. However, methods relying on attention-based decomposition remain dependent on the learned representations of these internal mechanisms, which may not align with domain-specific interpretability requirements in specialized contexts where transparency demands explicit grounding in established vocabulary. Recently, an emerging line of work has pursued interpretability by explicitly leveraging predefined, domain-specific vocabularies of clinically meaningful concepts, enabling explanations to be expressed directly in human-readable terms rather than inferred post hoc from latent attention patterns (Parekh et al., 2024; Bhalla et al., 2024). Among these, SpLiCE (Bhalla et al., 2024) operationalizes this paradigm by factorizing visual representations into sparse linear combinations of interpretable concepts drawn from a domain-specific vocabulary, achieving scalable and transparent explanations without sacrificing representational flexibility.

2.2 Multimodal Retrieval-Augmented Generation in Medicine

Multimodal RAG has emerged as a promising approach to mitigating factual hallucinations in medical VLMs by grounding generation in existing clinical knowledge. These systems extend traditional text-based retrieval by employing specialized encoders to extract features from multiple modalities, retrieving relevant cases from curated medical databases, and conditioning generation on both the input data and the retrieved context (He et al., 2025; Abootorabi et al., 2025; Molino et al., 2026). This shift from purely parametric knowledge to retrieval-augmented pipelines has yielded substantial gains across applications such as orthopedic diagnosis (Jin & Zhang, 2024), lung cancer staging (Tozuka et al., 2025), and prescription interpretation for medication management (Thetbanthad et al., 2025). In the specific domain of RRG, multimodal RAG approaches have shown particularly promising results. The MMed-RAG system reports substantial improvements in factual accuracy for both medical VQA and report generation across radiology, pathology, and ophthalmology (Xia et al., 2024a). Similarly, the RULE framework achieves notable gains by combining calibrated retrieval strategies with preference-based fine-tuning to balance reliance on parametric knowledge versus retrieved context (Xia et al., 2024b). These results suggest that providing models with concrete clinical examples can substantially reduce hallucinations and improve diagnostic consistency.

Although multimodal RAG provides indirect interpretability by exposing which cases inform generation, this transparency remains fundamentally passive, revealing available information without guaranteeing how it is utilized during generation. Retrieval operates through global similarity matching in learned embedding spaces, capturing overall visual resemblance without explicit guidance on which anatomical structures or pathological patterns should be prioritized (Wang et al., 2025). This absence of semantic grounding creates a fundamental dilemma: insufficient retrieval fails to capture necessary clinical information, while excessive retrieval introduces irrelevant details that interfere with coherent generation (Xia et al., 2024b). Consequently, models may inappropriately incorporate findings from retrieved cases that do not correspond to visual evidence in the input image, leading to factual hallucinations even in retrieval-augmented systems (Chu et al., 2025; Sloan et al., 2024; Salmè et al., 2025).

2.3 Limitations and Motivations

The limitations identified across interpretability and RAG approaches reveal a fundamental gap: existing methods treat transparency and factual accuracy as separate objectives rather than mutually reinforcing components. Interpretability techniques provide insight into model representations without actively constraining generation toward factually grounded outputs (Caragliano et al., 2025), while RAG systems improve factual grounding without semantic control to align retrieved information with specific visual evidence. This separation motivates our central hypothesis that interpretable visual concepts can serve as semantic guidance mechanisms, simultaneously enhancing transparency and factual accuracy by directing retrieval and generation toward clinically relevant content present in the input image.

3 Methodology

To jointly improve transparency and factual accuracy in RRG, we propose CEMRAG, a framework that integrates interpretable visual concepts with multimodal RAG through a structured prompting strategy. We extract clinically meaningful concepts from visual embeddings and retrieve similar cases from a database, both derived from the same image representation. Rather than treating these as independent augmentations, we structure them hierarchically in a unified prompt that guides the language component to focus on relevant portions of retrieved context.

Figure 1 illustrates the complete architecture of our proposed framework. Given as input a medical image $\mathrm{I} \in \mathbb{R}^{H \times W}$, where H and W denote the height and width in pixels, our objective is to generate a radiology report $\hat{R} = \{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_{\hat{n}}\}$ that approximates the ground truth report $R = \{r_1, r_2, \ldots, r_n\}$, with r_i and $\hat{r}_i$ representing individual tokens, and n and $\hat{n}$ denoting the number of tokens in each report. To achieve this, CEMRAG generates reports by coordinating four key components, each represented by distinct pathways in Fig. 1: (a) a visual encoding branch (yellow line), comprising a medical VLM encoder and a projector, extracts dense visual features from the input image; (b) a concept-extraction module (purple dashed line) maps image embeddings into interpretable clinical keywords; (c) a multimodal retrieval module (teal dash-dotted line) identifies similar training reports based on visual similarity in the embedding space; and (d) an LLM (orange dotted line) conditioned on hierarchically structured prompts integrates visual tokens, concept keywords, and retrieved reports to generate the final radiology report.

3.1 Visual Encoding and Projection

The foundation of our framework is the extraction of dense visual features to condition the generation process. The

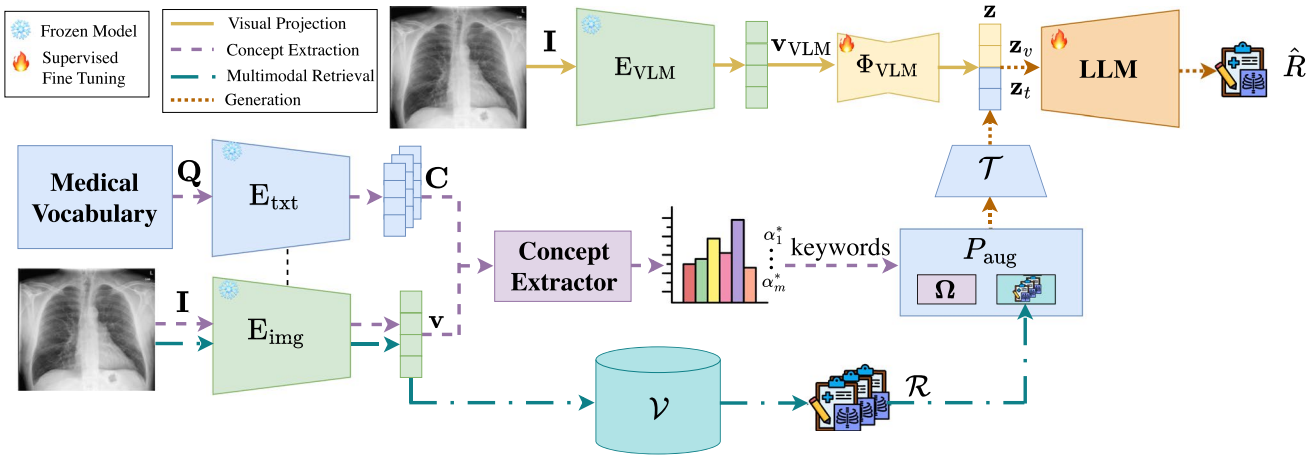


Fig. 1 CEMRAG overall framework, combining interpretable concept extraction with RAG for transparent and accurate radiology reporting

input image I is processed by a pretrained medical VLM encoder $\mathrm{E}_{\mathrm{VLM}}$ to produce a sequence of visual features $\mathrm{v}_{\mathrm{VLM}} = \mathrm{E}_{\mathrm{VLM}}(\mathrm{I}) \in \mathbb{R}^{\ell_v \times d_{\mathrm{VLM}}}$, where ℓ_v is the number of visual tokens and d_{VLM} is the dimensionality of the VLM feature space. A projection module Φ_{VLM}, implemented as a token-wise multi-layer perceptron, maps the visual features into the LLM token embedding space, producing visual token embeddings $\mathrm{z}_v = \Phi_{\mathrm{VLM}}(\mathrm{v}_{\mathrm{VLM}}) \in \mathbb{R}^{\ell_v \times d_{\mathrm{LLM}}}$. While z_v provides the primary visual features for generation, we employ additional aligned vision and text encoders ($\mathrm{E}_{\mathrm{img}}$, $\mathrm{E}_{\mathrm{txt}}$) to supply structured contextual information in a shared vision-language space, enabling both the decomposition of visual content into explicit clinical concepts and the retrieval of similar cases.

3.2 Concept Extraction

The input image I is processed by the vision encoder $\mathrm{E}_{\mathrm{img}}$ to produce a visual embedding $\mathrm{v} = \mathrm{E}_{\mathrm{img}}(\mathrm{I}) \in \mathbb{R}^d$, where d is the dimensionality of a shared vision-language embedding space. To enable concept extraction, we define a medical vocabulary $\mathrm{Q} = \{q_1, q_2, \ldots, q_m\}$ composed of m domain-specific concept terms derived from the training corpus. Each concept $q_j \in \mathrm{Q}$ is encoded with the text encoder $\mathrm{E}_{\mathrm{txt}}$ to obtain its embedding $\mathrm{c}_j = \mathrm{E}_{\mathrm{txt}}(q_j) \in \mathbb{R}^d$, and we collect these embeddings into the concept matrix $\mathrm{C} = [\mathrm{c}_1, \mathrm{c}_2, \ldots, \mathrm{c}_m] \in \mathbb{R}^{d \times m}$, where d denotes the dimensionality of the multimodal latent space. Then, the concept extraction module decomposes the visual embedding v as a non-negative linear combination of concept embeddings, as detailed in Appendix A. This decomposition yields a coefficient vector $\boldsymbol{\alpha}^* \in \mathbb{R}^m_{\geq 0}$ that quantifies the contribution of each vocabulary concept to the image representation. We select the top-τ concepts according to their coefficient magnitudes to form the interpretable keyword set $\boldsymbol{\Omega} = [\, q_1, \ldots, q_\tau \,] \subseteq \mathrm{Q}$.

In our implementation, we instantiate $\mathrm{E}_{\mathrm{img}}$ and $\mathrm{E}_{\mathrm{txt}}$ using CLIP encoders (Radford et al., 2021), which provide aligned vision-language embeddings through contrastive pretraining. The same embedding v also enables retrieval of similar documented cases, as described next.

3.3 Multimodal Retrieval Augmented Generation

While $\boldsymbol{\Omega}$ provides explicit clinical concepts, complete documented cases are required to ground generation in established clinical patterns and linguistic structure. We use $\mathrm{E}_{\mathrm{img}}$ to construct a vector database $\mathcal{V} = \{(\mathrm{v}_i^{\mathrm{train}}, R_i^{\mathrm{train}})\}_{i=1}^N$ that indexes all N training images through their visual embeddings $\mathrm{v}_i^{\mathrm{train}}$ and associates each with its corresponding radiology report R_i^{train}.

Using v as query, the retrieval mechanism identifies the top-k most similar cases by computing the cosine similarity $\mathcal{S}$ in the image embedding space:

$$\mathcal{S}(\mathrm{I}) = \text{top-}k \left\{ \frac{\mathrm{v} \cdot \mathrm{v}_i^{\mathrm{train}}}{\|\mathrm{v}\| \|\mathrm{v}_i^{\mathrm{train}}\|} \right\}_{i=1}^{N}$$

yielding retrieved reports $\mathcal{R} = \{R_{i_1}^{\mathrm{train}}, \ldots, R_{i_k}^{\mathrm{train}}\}$.

By operating in the image embedding space, this retrieval strategy prioritizes cases with similar radiological appearances, capturing visual patterns that are indicative of comparable clinical findings (You et al., 2023).

3.4 Hierarchical Prompt Construction and Report Generation

To synthesize these information sources effectively, we employ a hierarchical prompting strategy designed to mitigate the limitations of each component. While retrieved reports $\mathcal{R}$ provide rich clinical context, they

may inadvertently introduce findings absent in the query image. Conversely, concept keywords $\mathbf{\Omega}$ act as precise visual anchors but lack narrative structure. Therefore, we construct a structured prompt P_{aug} where $\mathbf{\Omega}$ serves as a priority filter, guiding the LLM to selectively leverage the linguistic patterns in $\mathcal{R}$ that align with the observed features.

The prompt structure consists of four components: (i) a coordination instruction that establishes the task of generating a report while prioritizing concept-related content from retrieved examples, (ii) an explicit list of extracted concept keywords $\mathbf{\Omega}$ framed as visual findings identified in the image, and (iii) the retrieved radiology reports $\mathcal{R}$ presented as reference examples from similar cases, and (iv) a final instruction reinforcing the generation objective. This structured presentation provides the LLM with both explicit concept annotations and implicit contextual knowledge, while maintaining a clear separation between the different information sources.

The augmented prompt P_{aug} is then processed by the LLM tokenizer $\mathcal{T}$ to produce a sequence of textual token embeddings $\mathrm{z}_t = \mathcal{T}(P_{\text{aug}}) \in \mathbb{R}^{\ell_t \times d_{\text{LLM}}}$, where ℓ_t denotes the number of text tokens and d_{LLM} is the LLM embedding dimension. The visual and textual token sequences are then concatenated along the sequence dimension to form the full multimodal input $\mathrm{z} = [\mathrm{z}_v; \mathrm{z}_t] \in \mathbb{R}^{(\ell_v+\ell_t)\times d_{\text{LLM}}}$, which is fed to the LLM. We denote the generated report as $\hat{R} = \text{LLM}(\mathrm{z})$, and model generation in an autoregressive formulation, where the LLM predicts each token $\hat{r}_i$ conditioned on all previous tokens and on the multimodal context encoded in z:

$$p(\hat{R} \mid \mathrm{I}, P_{\text{aug}}) = \prod_{i=1}^{\hat{n}} p(\hat{r}_i \mid \hat{r}_{<i}, \mathrm{z}).$$

To preserve robust medical pretrained representations, we consistently keep all the encoders frozen. In contrast, the training strategy for the LLM and the projection module Φ_{VLM} varies according to the model configuration, as detailed in Section 4.2.

4 Experimental Setup

This section outlines the experimental setup, detailing the datasets, model configurations, experimental conditions, and evaluation metrics used to assess report generation quality.

4.1 Datasets

We conduct experiments on two well-established benchmark datasets for RRG: MIMIC-CXR (Johnson et al., 2019) and IU X-ray (Demner-Fushman et al., 2015). Both datasets provide chest radiographs paired with corresponding clinical reports, enabling comprehensive evaluation of our proposed approach across different data scales and clinical contexts. The MIMIC-CXR dataset (Johnson et al., 2019) is a large-scale publicly available collection comprising over 370,000 chest radiographs from more than 65,000 patients. We adopt the official training, validation, and test split, restricting our experiments to the 156,344 frontal views (posteroanterior and anteroposterior projections) due to computational limitations. The IU X-ray dataset (Demner-Fushman et al., 2015) consists of 7,470 chest radiographs paired with 3,955 radiological reports. This dataset serves as a complementary evaluation benchmark to assess model performance under more constrained data conditions. The original IU X-ray dataset includes both frontal and lateral radiological images for most reports. However, to maintain consistency with our experimental protocol on MIMIC-CXR, we only consider the 3,307 frontal projections in our experiments. Following established conventions in the literature (Chen et al., 2021; Liu et al., 2021b), we exclude samples lacking a findings section, as this section provides the essential ground truth for supervised learning in RRG tasks. We employ a dataset split allocating 80% of the data for training, 10% for validation, and 10% for testing, with strict enforcement of patient-level separation to prevent data leakage across splits.

4.2 Model Configurations and Experimental Conditions

Our framework requires a model with aligned multimodal embeddings $(\mathrm{E}_{\text{img}}, \mathrm{E}_{\text{txt}})$ capable of supporting both concept extraction and similarity-based retrieval. To this end, we employ CXR-CLIP (You et al., 2023) across all experiments, which integrates a SwinTransformer (Liu et al., 2021a) as visual encoder E_{img} and a BioClinicalBERT (Alsentzer et al., 2019) as text encoder E_{txt}, both pretrained on the MIMIC-CXR dataset. We instantiate the concept-extraction module with Sparse Linear Concept Embeddings (SpLiCE) (Bhalla et al., 2024), which performs explicit sparse factorization of visual embeddings. Unlike attention-based approaches, which derive concepts from learned internal representations, SpLiCE operates on a predefined medical vocabulary. This ensures that the extracted concepts are clinically relevant, and allows each

term's contribution to be quantified directly through optimized sparse coefficients. For experiments on the IU X-ray dataset, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) to CXR-CLIP for obtaining refined embeddings tailored to this specific data distribution (more details are given in Appendix B).

We evaluate CEMRAG under two architectural configurations to assess whether the hierarchical prompting strategy remains effective across different levels of medical domain adaptation and vision-to-language alignment. Both configurations adopt a LLaVA-style architecture (Liu et al., 2023) with Mistral-7B (Jiang et al., 2023) as the LLM backbone. The first configuration employs LLaVA-Med (Li et al., 2023), where both the vision encoder $\mathrm{E}_{\mathrm{VLM}}$ and the LLM incorporate medical domain-specific pretraining. The second configuration uses CXR-CLIP as a unified encoder: $\mathrm{E}_{\mathrm{img}}$ extracts embeddings v that enable concept extraction $\boldsymbol{\Omega}$, retrieval of reports $\mathcal{R}$, and generation of visual tokens $\mathrm{v}_{\mathrm{VLM}}$. This configuration pairs medically-pretrained CXR-CLIP with base Mistral-7B without medical language pretraining. Unlike LLaVA-Med, which provides a pretrained projection layer, this configuration requires the introduction of a projection module Φ_{CLIP} to map CLIP visual features into the LLM embedding space. Images are preprocessed according to each encoder's specifications: 224×224 pixels with normalization to $[-1, 1]$ for CXR-CLIP, and 336×336 center crops with encoder-specific normalization for LLaVA-Med. Implementation details, including the initial alignment phase for Φ_{CLIP}, are provided in Appendix B.

We evaluate four distinct prompting strategies that progressively incorporate retrieval and concept information. To ensure reproducibility, we provide detailed prompt templates and configuration details for each strategy in Table 1.

Table 1 Experimental configurations for radiology report generation, comparing input modalities, prompt structures, interpretability mechanisms, and factual grounding strategies. Notation: v denotes visual embeddings, $\boldsymbol{\Omega}$ extracted concepts and $\mathcal{R}$ retrieved reports. `Instruction`: "Provide a description of the findings in the radiology image"; `Task`: "Write the report of the radiology image taking information from similar FINDINGS. Consider as more relevant sentences that contain any of the KEYWORDS in the FINDINGS"; `Final Instruction`: "Write a paragraph with only the report relying in detail on the FINDINGS"

Strategy	Inputs	Prompt Structure	Interpretability	Factual Grounding
Image-Only	v	<Instruction>	× black-box	× Visual encoder and LLM priors only
Concepts	$\mathrm{v}, \boldsymbol{\Omega}$	<Instruction> + <Keywords>	✓ Concept-level via visual-semantic alignment	∼ Sparse visual concept decomposition
RAG	$\mathrm{v}, \mathcal{R}$	<Instruction> + <Similar Findings>	∼ Indirect via retrieved context	✓ Visual similarity-based nearest neighbours
CEMRAG	$\mathrm{v}, \boldsymbol{\Omega}, \mathcal{R}$	<Task> + <Keywords> + <Similar Findings> + <Final Instruction>	✓✓ Dual: concept annotations and similar cases	✓✓ Concept-guided focus to salient findings

Image-Only This approach provides the LLM with visual features alongside a minimal prompt instructing the model to describe the radiological image. This provides a reference point for the level of performance achievable from visual information alone, without any external context.

Concepts This strategy augments the prompt with interpretable clinical concepts extracted from the input image using SpLiCE. For each dataset, we construct a domain-specific vocabulary of medical concepts by selecting the 200 most frequent bigrams from the training reports. Each bigram is encoded with the CLIP text encoder $\mathrm{E}_{\mathrm{txt}}$ to obtain its embedding. SpLiCE is then applied to derive a sparse, non-negative decomposition of the image embedding over this concept vocabulary, yielding a coefficient vector that quantifies the contribution of each concept. For each image, we rank concepts according to their coefficients, select the top five, and inject these bigrams into the prompt as explicit keywords. Additional details regarding vocabulary construction, normalization procedures, and sensitivity analyses for different vocabulary sizes and sparsity levels are provided in the Appendix A.1.

Multimodal RAG This strategy provides contextual grounding by retrieving reports from visually similar cases. We use FAISS (Douze et al., 2024) to compute cosine similarity between the query image embedding v derived from CXR-CLIP and embeddings in database $\mathcal{V}$, retrieving the top-$k = 3$ most similar cases whose reports are then incorporated into the prompt as reference examples. The choice of $k = 3$ follows established practice in retrieval-augmented RRG (Garcia-Carmona et al., 2025; Li et al., 2023; Zhang et al., 2025), balancing the provision of sufficient contextual information against the risk of introducing irrelevant or contradictory findings from less similar cases, while remaining within the context window constraints of the LLM. The retrieval configuration varies by dataset. For the MIMIC-CXR dataset, retrieval is performed within the training set of the same dataset to ensure that similar cases are drawn from the same distribution. For IU X-ray, which has a substantially smaller training set, in-domain retrieval would provide insufficient

contextual variety and limit the benefits of the RAG strategy. We therefore implement cross-domain retrieval from the MIMIC-CXR training set, simulating a realistic scenario where a smaller institutional dataset leverages a larger external knowledge base to enhance report generation quality.

CEMRAG This method combines the two previous approaches to construct an enriched prompt that presents the top five concept bigrams as priority keywords alongside three retrieved reports. We design the prompt structure following Mistral's prompt engineering guidelines (Mistral, 2024), which emphasize the importance of clear task definition, hierarchical organization, explicit formatting, and concrete examples for effective instruction-based generation. Unlike previous conditions with simple generation instructions, CEMRAG employs a coordination directive that establishes the task objective of generating a report from similar findings while prioritizing content related to the extracted concepts. Following this directive, the extracted keywords are presented in a dedicated formatted section. The three retrieved reports are then incorporated with separators. The prompt concludes with a final instruction that explicitly directs the model to produce a radiology report relying on the provided findings.

This design ensures clear attribution of information sources and guides the LLM toward clinically relevant content by orienting the model toward portions of retrieved reports that align with observed visual features.

Each of the four prompting strategies is evaluated under two LLM training paradigms. In the *Zero-Shot* setting, the entire model (visual encoders, projection layers, and the LLM) is kept frozen, assessing transfer capability with only prompt content varying across conditions. In the *Supervised Fine-Tuning (SFT)* setting, we adapt the LLM using LoRA (Hu et al., 2022) and jointly fine-tune the corresponding projection layer, while keeping all visual encoders frozen. This design follows the LLaVA training paradigm, isolating adaptation to the language and projection components. This enables a controlled investigation into how prompting strategies influence RRG performance without confounding changes in visual feature extractors. Additional training details for both the *Zero-Shot* and *SFT* settings are provided in Appendix B.

For all experiments, RRG is performed with greedy decoding (temperature $T = 0$) to ensure deterministic and reproducible outputs, thereby eliminating variability arising from stochastic sampling. All experiments are performed on four NVIDIA A100 GPUs. On MIMIC-CXR, *SFT* training time varies by strategy: Image-Only requires approximately 7 hours, *Concepts* strategy 8 hours, RAG 13 hours, and CEMRAG 14 hours.

4.3 Evaluation Metrics

Our evaluation framework incorporates two complementary categories of metrics: NLP metrics, which measure lexical similarity to reference reports, and clinical metrics, which assess the factual correctness of medical content. To evaluate lexical similarity, we report three standard NLP metrics that are commonly used in text generation tasks. ROUGE-L (Lin et al., 2004) measures the longest common subsequence between the generated and reference texts, capturing sentence-level structural similarity. BLEU-1 and BLEU-4 (Papineni et al., 2002) evaluate n-gram overlap at the unigram and 4-gram levels, respectively, with BLEU-4 providing assessment of longer phrasal matches that better reflect fluency and coherence. While these metrics effectively evaluate linguistic properties such as fluency and phrasal coherence, they capture surface-level textual similarity rather than clinical accuracy.

To address this limitation, we emphasize clinical evaluation through two established factual correctness metrics: F1-CheXbert (Zhang et al., 2020) and F1-RadGraph (Delbrouck et al., 2022). The former computes the F1 score between disease labels extracted from the generated and reference reports using the CheXbert labeler (Smit et al., 2020), which is a BERT-based model trained to identify the presence, absence, or uncertainty of pathological findings. Following standard practice, we report F1-CheXbert across all 14 CheXbert classes, encompassing the comprehensive range of conditions identifiable in chest radiographs. Additionally, we report performance on the five most prevalent clinical findings in real-world chest radiograph reports: atelectasis, cardiomegaly, consolidation, edema and pleural effusion. This focused evaluation on common pathologies provides insight into the model's performance in frequently encountered diagnostic scenarios. The second clinical metric, F1-RadGraph, quantifies factual correctness by measuring the overlap between the semantic graphs extracted from the generated and reference reports. This captures not only the presence of clinical findings, but also their anatomical locations and relationships between entities. Together, these two clinical metrics provide a complementary perspective on factual correctness: F1-CheXbert focuses on diagnostic label accuracy, while F1-RadGraph assesses the representation of structured clinical content.

5 Results

We organize our findings into two complementary parts. Section 5.1 presents a quantitative evaluation across datasets, model configurations, and training paradigms, using

both NLP and clinical accuracy metrics. Section 5.2 provides qualitative analyses that illustrate how concept extraction and retrieval affect the generated reports and their interpretability.

5.1 Quantitative Results

We structure our quantitative analysis by first examining performance on the MIMIC-CXR dataset in Section 5.1.1, followed by results on the IU X-ray dataset in Section 5.1.2. For each dataset, we evaluate two model configurations: LLaVA-Med with its standard visual encoder, and LLaVA with CXR-CLIP as the visual encoder. Each configuration is assessed under both *Zero-Shot* and *Supervised Fine-Tuning (SFT)* settings to determine how domain adaptation influences the effectiveness of our approach.

5.1.1 MIMIC-CXR

Table 2 presents comprehensive quantitative results on the MIMIC-CXR dataset, where retrieval is performed from the MIMIC-CXR training set, establishing an in-domain retrieval scenario.

Zero-Shot In the *Zero-Shot* setting, the LLaVA-Med baseline exhibits clear limitations for RRG. Despite leveraging general medical pretraining, it attains an F1-Radgraph (F1-RG) of only 0.052, indicating poor alignment of clinical entities and relations, and very low NLP metrics, reflecting limited lexical overlap with reference reports. CheXbert Micro-$F1_{14}$ reaches 0.255, suggesting some ability to recognize common pathologies but insufficient overall clinical reliability. Augmenting the prompt with concepts or retrieval leads to a consistent progression in performance. Adding concepts improves both clinical and textual metrics: F1-RG rises to 0.073, CheXbert Micro-$F1_{14}$ to 0.401, and ROUGE-L to 0.146, indicating that even in a *Zero-Shot* regime, visual concepts provide useful structured cues about relevant findings. RAG yields substantially larger gains, with F1-RG increasing to 0.184 and CheXbert Micro-$F1_{14}$ to 0.498. These gains demonstrate that retrieved similar cases provide rich contextual and linguistic information that enhances both clinical accuracy and report fluency. CEMRAG achieves the best overall *Zero-Shot* performance for LLaVA-Med, significantly outperforming both *Image-Only* and *Concepts* baselines across all metrics ($p < 0.05$, paired bootstrap). Relative to RAG alone, the gains are more targeted: in this configuration, the medically pretrained LLM already captures the dominant clinical content from retrieved reports, so concept augmentation primarily contributes improved lexical coverage and broader pathology recognition, yielding significant gains on BLEU-1, ROUGE-L, and Macro-$F1_{14}$. Crucially, beyond these quantitative improvements, the explicit concept decomposition provides an interpretability layer that RAG alone cannot offer, making the

Table 2 Quantitative results on the MIMIC-CXR test set for two model configurations (LLaVA-Med and LLaVA with CXR-CLIP) and two training regimes (*Zero-Shot* and *Supervised Fine-Tuning*). We report F1-RadGraph (F1-RG), and NLP metrics (BLEU-1 as B-1, BLEU-4 as B-4, ROUGE-L as R-L). CheXbert-based label metrics are reported over 14 labels (Micro-$F1_{14}$, Macro-$F1_{14}$) and over the 5 most prevalent findings (Micro-$F1_5$, Macro-$F1_5$). Retrieval is performed from the MIMIC-CXR training set. Bold indicates the best result within each model-regime block. * denotes a statistically significant difference from CEMRAG, based on a paired bootstrap test with 1,000 resamples ($p < 0.05$)

Model	Method	Micro-$F1_{14}$	Micro-$F1_5$	Macro-$F1_{14}$	Macro-$F1_5$	F1-RG	B-1	B-4	R-L
Zero-Shot									
LLaVA-Med	Image-Only	0.255*	0.237*	0.113*	0.134*	0.052*	18.81*	0.71*	0.131*
	+ Concepts	0.401*	0.412*	0.238*	0.355*	0.073*	21.47*	1.29*	0.146*
	+ RAG	0.498	0.528	0.314*	0.447	0.184	27.08*	4.60	0.187*
	+ CEMRAG	**0.502**	**0.529**	**0.319**	**0.449**	**0.185**	**29.34**	**4.64**	**0.189**
LLaVA with CXR-CLIP	Image-Only	0.213*	0.174*	0.101*	0.123*	0.151*	19.20*	4.69*	0.185*
	+ Concepts	0.328*	0.330*	0.161*	0.230*	0.172*	19.62*	4.87*	0.192*
	+ RAG	0.489*	0.518*	0.309*	0.442	0.181*	25.53*	5.58*	0.198*
	+ CEMRAG	**0.498**	**0.526**	**0.314**	**0.443**	**0.187**	**29.78**	**6.08**	**0.201**
Supervised Fine-Tuning									
LLaVA-Med	Image-Only	0.470*	0.486*	0.266*	0.392*	0.174	27.81*	7.43	0.211*
	+ Concepts	0.476*	0.502*	0.284*	0.412*	**0.188***	28.83*	**7.87***	**0.220***
	+ RAG	0.477*	0.499*	0.287*	0.412*	0.176	30.50*	7.26*	0.212
	+ CEMRAG	**0.488**	**0.510**	**0.301**	**0.424**	0.177	**30.81**	7.47	0.213
LLaVA with CXR-CLIP	Image-Only	0.393*	0.448*	0.215*	0.343*	0.161*	22.38*	5.60*	0.207
	+ Concepts	0.486	0.510	0.283*	0.413*	**0.184***	28.17*	**7.57***	**0.225***
	+ RAG	0.477*	0.501*	0.293*	0.414*	0.168*	30.16*	6.84*	0.204
	+ CEMRAG	**0.488**	**0.512**	**0.300**	**0.423**	0.180	**30.49**	6.98	0.206

model's reliance on specific visual findings transparent to the clinician.

A similar trend is observed for the configuration using CXR-CLIP as the visual encoder, which starts from a stronger *Zero-Shot* baseline in terms of fluency as a result of the one-epoch alignment of its projection layer on MIMIC-CXR. The baseline achieves an F1-RG of 0.151 and BLEU-4 of 4.69, considerably higher than LLaVA-Med, whilst CheXbert Micro-$F1_{14}$ remains lower at 0.213. This pattern suggests that the alignment primarily improves generic report structure and phrasing, but does not immediately translate into superior pathology coverage. As in the previous case, both concept and retrieval augmentation are beneficial: *Concepts* strategy increases CheXbert Micro-$F1_{14}$ to 0.328 and F1-RG to 0.172, whereas RAG raises these values to 0.489 and 0.181, respectively. CEMRAG configuration attains again the best overall *Zero-Shot* performance in this setting, with F1-RG of 0.187, CheXbert Micro-$F1_{14}$ of 0.498, and the highest BLEU-1 (29.78), BLEU-4 (6.08), and ROUGE-L (0.201). In this configuration, CEMRAG significantly outperforms RAG seven out of eight metrics, confirming that concept augmentation provides a stronger complementary signal when paired with a domain-aligned visual encoder that lacks the built-in medical language priors of LLaVA-Med. Across both architectures, statistical testing consistently supports the superiority of CEMRAG over *Image-Only* and *Concepts* baselines, as well as over RAG on the majority of metrics, with particularly robust significance in the CXR-CLIP configuration.

SFT In the *SFT* regime, the two augmentation components contribute differently depending on the metric type: concept augmentation drives gains on sequence-level measures, while CEMRAG achieves superior performance on clinical accuracy metrics. The *Image-Only* baselines of both architectures exhibit substantial improvements over their *Zero-Shot* counterparts, confirming the importance of task-specific adaptation for RRG. The LLaVA-Med baseline shows substantial gains across all metrics, with F1-RG increasing from 0.052 to 0.174, CheXbert Micro-$F1_{14}$ from 0.255 to 0.470, and BLEU-4 from 0.71 to 7.43, with ROUGE-L reaching 0.211. On top of this stronger baseline, the three augmentation strategies remain beneficial, but their role shifts compared to the *Zero-Shot* setting. Adding only concepts yields the highest F1-RG in this configuration (0.188) and the best BLEU-4 and ROUGE-L scores (7.87 and 0.220), along with moderate gains in CheXbert Micro- and Macro-F1 (0.476 and 0.284). In this setting, *Concepts* augmentation yields moderate improvements on CheXbert metrics while achieving the highest F1-RG, BLEU-4, and ROUGE-L scores across all configurations. These differences are statistically significant also with respect to CEMRAG, suggesting that the benefit of concept augmentation concentrates on sequence-level report quality, where explicit visual cues help the model produce more coherent and clinically detailed descriptions. RAG shows a different pattern: F1-RG remains close to the baseline (0.176 vs. 0.174), whereas BLEU-1 increases to 30.50. This pattern suggests that, once the model has been adapted on MIMIC-CXR using retrieval-augmented prompts, the additional in-domain retrieved reports mainly enrich lexical diversity and increase report length, without yielding proportional gains on metrics that are more sensitive to clinical structure and relational correctness. Finally, CEMRAG achieves statistically significant improvements over all other configurations on CheXbert metrics, including Micro-$F1_{14}$ (0.488) and Macro-$F1_{14}$ (0.301), as well as on BLEU-1 (30.81). These results indicate that combining interpretable concepts with retrieval augmentation improves clinical accuracy beyond either component alone, while preserving competitive overall generation quality.

For the CXR-CLIP configuration, the *SFT* baseline remains weaker than the corresponding LLaVA-Med baseline on MIMIC-CXR (e.g., CheXbert Micro-$F1_{14}$ of 0.393 vs. 0.470), reflecting the advantage of LLaVA-Med's medically pretrained language component. The augmentation strategies follow the same progressive improvement pattern observed in other configurations. *Concepts* augmentation produces dramatic gains, with F1-RG jumping to 0.184 and CheXbert Micro-$F1_{14}$ reaching 0.488, substantially exceeding the improvements observed with LLaVA-Med and nearly recovering the baseline performance deficit. Notably, this configuration achieves the highest BLEU-4 score of 7.57 across all *SFT* CXR-CLIP settings and the highest overall ROUGE-L of 0.225, mirroring the pattern observed with LLaVA-Med. Consistent with LLaVA-Med, the *Concepts* configuration concentrates its gains on sequence-level measures, significantly exceeding CEMRAG on F1-RG, BLEU-4, and ROUGE-L, while CEMRAG retains a significant advantage on CheXbert Macro-$F1_{14}$ and Macro-$F1_5$. These results further demonstrate that interpretable visual concepts offer structured information that effectively complements visual encoder features. RAG-only yields mixed results, with F1-RG at 0.168 and CheXbert Macro-$F1_{14}$ reaching 0.293, while CheXbert Micro-$F1_{14}$ of 0.477 remains below the *Concepts* configuration. The combined CEMRAG approach achieves balanced performance with the highest CheXbert Micro-$F1_{14}$ at 0.488 and Macro-$F1_{14}$ at 0.300, alongside the highest BLEU-1 of 30.49.

MIMIC-CXR Summary Taken together, the MIMIC-CXR experiments reveal two main patterns regarding the

interaction between augmentation strategies and training regimes.

First, the effectiveness of RAG is strongly regime-dependent. In the *Zero-Shot* setting, adding in-domain retrieved reports substantially improves clinical and linguistic metrics, consistent with models that rely heavily on external context to compensate for limited task-specific adaptation. In the *SFT* regime, where models are trained with retrieved reports in the prompt, the role of retrieval changes. While unigram coverage (BLEU-1) improves substantially, metrics sensitive to longer-range structure (BLEU-4, ROUGE-L) underperform relative to *Concepts* augmentation. Furthermore, clinical metrics exhibit modest yet consistent degradation compared to *Zero-Shot* for both RAG and CEMRAG. This behaviour suggests that when supervision and retrieval are drawn from the same in-domain distribution, retrieval provides diminishing returns as the model internalizes distributional patterns present in the retrieved context.

Secondly, SpLiCE-derived concepts contribute consistently across both regimes. In *Zero-Shot* configurations, *Concepts* alone clearly improve pathology identification and F1-RG. In the *SFT* setting, SpLiCE contributes more evidently on metrics that capture the quality of clinically complex sequences, while the gain on CheXbert F1 is comparatively smaller.

CEMRAG leverages the complementary strengths of concept and retrieval augmentation: in *Zero-Shot*, it amplifies the benefits of RAG, with statistically significant gains over all individual baselines on the large majority of metrics; in *SFT*, it maintains significant superiority on CheXbert pathology metrics, while *Concepts* alone achieves higher scores on sequence-level measures. This trade-off between clinical accuracy and sequence-level quality reflects the distinct roles of each component, and makes CEMRAG the only configuration able to jointly deliver competitive performance, factual grounding, and interpretability.

Finally, the systematic gap between CheXbert Micro-F1 and Macro-F1 across all methods reflects the pronounced class imbalance in chest X-ray reports. The improvements in Macro-F1 observed with CEMRAG indicate that the method contributes to a better coverage of underrepresented findings, supporting its suitability for clinically realistic reporting scenarios.

5.1.2 IU X-Ray

Table 3 presents quantitative results on the IU X-ray dataset. This dataset, differently from MIMIC-CXR, features considerably shorter and more concise reports, and employs a more limited medical vocabulary. In this case, our experimental setup implements cross-domain retrieval, where similar cases are retrieved from the MIMIC-CXR database rather than from IU X-ray's own training set. This configuration puts our framework's ability to generalize across different datasets and reporting styles to the test, since the

Table 3 Quantitative results on the IU X-Ray test set for two model configurations (LLaVA-Med and LLaVA with CXR-CLIP) and two training regimes (*Zero-Shot* and *Supervised Fine-Tuning*). We report F1-RadGraph (F1-RG), and NLP metrics (BLEU-1 as B-1, BLEU-4 as B-4, ROUGE-L as R-L). CheXbert-based label metrics are reported over 14 labels (Micro-$F1_{14}$, Macro-$F1_{14}$) and over the 5 most prevalent findings (Micro-$F1_5$, Macro-$F1_5$). Retrieval is performed cross-domain from the MIMIC-CXR training set. Bold indicates the best result within each model-regime block. * denotes a statistically significant difference from CEMRAG, based on a paired bootstrap test with 1,000 resamples ($p < 0.05$)

Model	Method	Micro-$F1_{14}$	Micro-$F1_5$	Macro-$F1_{14}$	Macro-$F1_5$	F1-RG	B-1	B-4	R-L
Zero-shot									
LLaVA-Med	Image-Only	0.063*	0.042*	0.047*	0.019*	0.074*	17.36*	1.08*	0.125*
	+ Concepts	0.122*	0.162*	0.092*	0.134*	0.064*	13.7*	0.85*	0.111*
	+ RAG	0.377*	0.344*	0.220*	0.271*	0.228*	21.04*	3.60*	0.177*
	+ CEMRAG	**0.387**	**0.397**	**0.252**	**0.315**	**0.234**	**24.34**	**4.10**	**0.191**
LLaVA with CXR-CLIP	Image-Only	0.305*	0.082*	0.038*	0.043*	0.188*	5.56*	1.26*	0.168*
	+ Concepts	0.307*	0.182*	0.085*	0.087*	0.199*	8.15*	1.98*	0.199*
	+ RAG	0.367*	0.343*	0.212*	0.247*	0.213*	25.35*	6.13*	0.203*
	+ CEMRAG	**0.378**	**0.361**	**0.232**	**0.298**	**0.247**	**27.84**	**6.75**	**0.221**
Supervised Fine-Tuning									
LLaVA-Med	Image-Only	0.326*	0.115*	0.031*	0.059*	0.175*	14.25*	4.50*	0.179*
	+ Concepts	0.336*	0.185*	0.082*	0.081*	0.178*	23.21*	5.67*	0.181*
	+ RAG	0.468*	0.356*	0.183*	0.205*	0.249	28.15	7.73*	0.251
	+ CEMRAG	**0.501**	**0.526**	**0.244**	**0.355**	**0.252**	**28.37**	**8.00**	**0.252**
LLaVA with CXR-CLIP	Image-Only	0.376*	0.102*	0.037*	0.053*	0.235*	14.25*	4.50*	0.235*
	+ Concepts	0.427*	0.362*	0.118*	0.178*	0.244	18.83*	6.06*	0.242*
	+ RAG	0.468*	0.395*	0.172	0.245*	0.244	22.58	6.91	0.243*
	+ CEMRAG	**0.486**	**0.439**	**0.174**	**0.256**	**0.248**	**22.90**	**7.10**	**0.249**

retrieved context originates from a different institutional source with different imaging protocols and documentation practices.

Zero-Shot In the *Zero-Shot* setting, the LLaVA-Med baseline performs very poorly on IU X-Ray, with a CheXbert Micro-F1$_{14}$ of 0.063 and an F1-RG of 0.074, confirming that general medical pretraining does not directly translate into effective RRG for this dataset. The *Concepts* condition produces a mixed effect: CheXbert Micro-F1$_{14}$ roughly doubles to 0.122, indicating improved detection of some pathologies, yet F1-RG decreases slightly to 0.064 and all NLP metrics deteriorate (e.g., BLEU-1 drops from 17.36 to 13.79). This behaviour is consistent with a style mismatch: when concepts are injected without further constraints, LLaVA-Med tends to expand each keyword into lengthy explanatory sentences. This contrasts with the highly concise IU X-ray references, reducing n-gram overlap despite potentially correct clinical content. In contrast, cross-domain retrieval yields substantial improvements. RAG on MIMIC-CXR increases CheXbert Micro-F1$_{14}$ to 0.377, F1-RG to 0.228, and BLEU-4 to 3.60, indicating that retrieved examples provide useful templates despite originating from a different institution and reporting style. Finally, CEMRAG achieves the best overall *Zero-Shot* performance for LLaVA-Med (Micro-F1$_{14}$ 0.387, F1-RG 0.234, BLEU-4 4.10), significantly outperforming all three baselines on every evaluated metric ($p < 0.05$, paired bootstrap). This suggests that in this scenario explicit concept decomposition and retrieved context address complementary aspects of RRG, jointly improving clinical accuracy and lexical quality beyond what either component achieves independently.

For LLaVA with CXR-CLIP configuration, the *Zero-Shot* baseline attains CheXbert Micro-F1$_{14}$ = 0.305 and F1-RG = 0.188, substantially higher than LLaVA-Med. This reflects the one-epoch projector alignment on IU X-ray, which provides better dataset-specific adaptation compared to LLaVA-Med's general medical pretraining. On the other hand, performance on the 5-label subset is extremely weak (Micro-F1$_5$ = 0.082) and several Macro-F1 scores are low, indicating uneven pathology recognition across label subsets. NLP metrics are also poor (BLEU-1 = 5.56, BLEU-4 = 1.26), showing that this short alignment phase only partially adapts the model to the concise IU X-ray style. *Concepts* augmentation provides modest yet consistent gains (F1-RG = 0.199, slight improvements in CheXbert metrics), though effectiveness is limited by the projector alignment on IU X-ray's small training set. RAG produces much larger benefits: Micro-F1$_{14}$ rises to 0.367, F1-RG to 0.213, and BLEU-1 jumps to 25.35, confirming that cross-domain retrieval from MIMIC-CXR supplies useful clinical templates and linguistic structure. CEMRAG yields the strongest *Zero-Shot* performance for this configuration, with Micro-F1$_{14}$ = 0.378, F1-RG = 0.247, BLEU-1 = 27.84, BLEU-4 = 6.75, and ROUGE-L = 0.221. As with LLaVA-Med, all differences from CEMRAG are statistically significant across all eight metrics for every baseline, confirming that concept-level signals derived from SpLiCE help the model exploit cross-domain retrieved context more selectively, improving both clinical accuracy and report quality.

SFT In the *SFT* setting, LLaVA-Med exhibits substantial gains over its *Zero-Shot* performance on IU X-Ray. The baseline Micro-F1$_{14}$ increases from 0.063 to 0.326 and F1-RG from 0.074 to 0.175, confirming that even a relatively small amount of supervision is sufficient to substantially improve task-specific behaviour on a new dataset. On top of this, the augmentation strategies continue to provide consistent benefits. *Concepts* augmentation yields modest improvements, whereas RAG leads to more pronounced gains, reaching 0.468 Micro-F1$_{14}$ and 0.249 F1-RG. CEMRAG achieves the strongest overall performance, with 0.501 Micro-F1$_{14}$, 0.252 F1-RG, and BLEU-4 of 8.00. *Image-Only* and *Concepts* baselines remain significantly below CEMRAG on all metrics. Compared to RAG alone, CEMRAG yields significant improvements on all CheXbert metrics and BLEU-4, while F1-RG, BLEU-1, and ROUGE-L show consistent but non-significant gains. Unlike the in-domain setting on MIMIC-CXR, where RAG becomes partially redundant after *SFT*, here cross-domain retrieval from MIMIC-CXR continues to supply complementary information that is not fully captured by supervised training on the much smaller IU X-Ray corpus, and CEMRAG is able to exploit this additional signal more effectively.

For LLaVA with CXR-CLIP configuration, *SFT* on IU X-ray leads to analogous trends. The *SFT* baseline reaches CheXbert Micro-F1$_{14}$ = 0.376 and F1-RG = 0.235; *Concepts* augmentation improves sequence-level metrics, with F1-RG = 0.244, BLEU-4 = 6.06, and ROUGE-L = 0.242; RAG further increases label-based scores (Micro-F1$_{14}$ = 0.468, F1-RG = 0.244). The full CEMRAG configuration yields a balanced improvement, with Micro-F1$_{14}$ = 0.486, F1-RG = 0.248, and BLEU-4 = 7.10, combining the benefits of concept guidance and retrieval-based context. CEMRAG significantly outperforms both *Image-Only* and *Concepts* on nearly all metrics. The comparison with RAG is more nuanced: significant gains are observed on CheXbert Micro-F1$_{14}$, Micro-F1$_5$, Macro-F1$_5$, and ROUGE-L, while the remaining metrics show consistent improvements that do not reach statistical significance, reflecting the smaller test set size of IU X-ray. Despite the comparable clinical metrics

under CEMRAG, a clear discrepancy persists in lexical quality: LLaVA-Med attains substantially higher BLEU-1 (28.37 vs. 22.90). This is plausibly due to its medically pretrained language component, which starts from a richer medical vocabulary and a broader set of reporting patterns. In the CXR-CLIP variant, the Mistral-7B backbone is adapted only on the small IU X-ray corpus, limiting its ability to acquire diverse and specialised radiology phrasing. As a result, report generations remain less varied and exhibit lower n-gram overlap, even when clinical content is comparable.

IU X-Ray Summary Overall, the IU X-ray experiments show that the proposed augmentation strategies remain effective in a low-resource, cross-domain scenario. Statistical testing provides particularly strong support in *Zero-Shot* configurations, where CEMRAG significantly outperforms every baseline on all eight metrics across both architectures. Under *SFT*, significance is maintained on the majority of metrics, with the narrower gaps against RAG attributable to the smaller test set rather than diminished effect sizes. *SFT* substantially improves both backbones, cross-domain RAG continues to offer clear benefits rather than becoming redundant, and CEMRAG consistently matches or exceeds the single augmentations. At the same time, the persistent gap in BLEU-1 between LLaVA-Med and CXR-CLIP configurations underscores the practical value of a medically pretrained language module for producing lexically rich reports when task-specific supervision is scarce.

5.1.3 Computational Cost Analysis

A central concern for clinical deployment is whether the gains in accuracy and interpretability justify the additional computational cost. We evaluate this trade-off on the MIMIC-CXR test set by measuring the average end-to-end inference time per image, using LLaVA-Med with *SFT*, which is the most computationally demanding configuration, on a single NVIDIA A100 GPU.

Figure 2 illustrates the relationship between average total inference time and clinical accuracy across the four experimental conditions. The *Image-Only* baseline requires 1.99 s per image and achieves a Micro-F1$_{14}$ of 0.470, serving as the reference point. Concept augmentation increases inference time to 2.45 s, with the overhead attributable to a CXR-CLIP forward pass (10.1 ms), the SpLiCE decomposition (7.3 ms), and a modestly longer input sequence (625 tokens on average vs. 600). This configuration improves Micro-F1$_{14}$ to 0.476 while introducing concept-level interpretability entirely absent from the baseline. RAG augmentation reaches a comparable inference time of 2.45 s

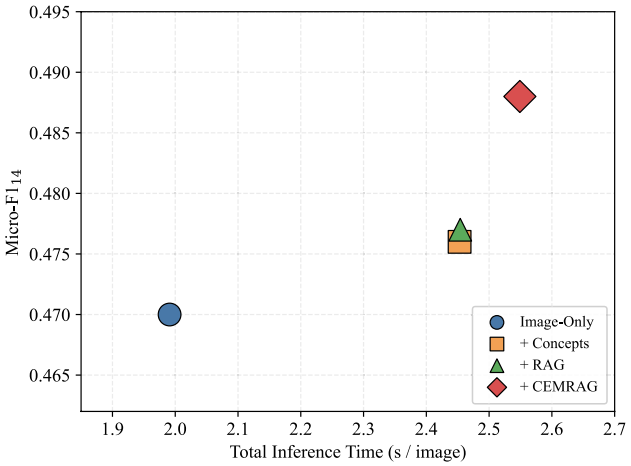


Fig. 2 Average per-image inference time versus clinical accuracy (Micro-F1$_{14}$) for each augmentation strategy, evaluated on the MIMIC-CXR test set with LLaVA-Med under *SFT*

despite a substantially longer input sequence (885 tokens on average), as the pipeline overhead is dominated by FAISS retrieval (57.2 ms) rather than LLM computation. RAG improves Micro-F1$_{14}$ to 0.477, though it provides only indirect interpretability through exemplar-based context. The similar inference time of the two strategies is coincidental: concept augmentation involves a lighter pipeline but longer generation time, consistent with its superior performance on sequence-level metrics such as BLEU-4 and F1-RadGraph (Table 2). RAG, conversely, incurs heavier retrieval cost but produces more constrained outputs grounded in retrieved report structure. This results in nearly identical average total inference times, despite fundamentally different cost profiles. These outcomes demonstrate a favorable cost-benefit profile: with a modest increase in inference time, CEMRAG simultaneously improves diagnostic accuracy and provides dual-layer interpretability, where SpLiCE concepts expose which visual features drive generation and retrieved reports offer grounding in documented clinical evidence. In RRG, where factual correctness and interpretability are both prerequisites for clinical trust, this computational overhead represents a practical and justified investment. Further details on the GFLOPs measurement methodology and cost analysis are provided in Section C.

5.2 Qualitative Results

While quantitative metrics capture global trends across datasets and configurations, qualitative analysis helps clarify how concept extraction and retrieval affect individual predictions. In this section, we analyse two representative chest X-ray cases from MIMIC-CXR using the LLaVA-Med configuration. The first illustrates typical *Zero-Shot* failure

modes and their mitigation with CEMRAG, while the second examines how the same mechanisms behave after *SFT*. For each case, we compare reports produced by the *Image-Only*, *Concepts*, RAG, and CEMRAG conditions, and highlight patterns of hallucination, omission, superfluity, as well as clinically accurate description.

5.2.1 Generated Report Comparison

Zero-Shot Table 4 shows a *Zero-Shot* example on MIMIC-CXR that reflects the error patterns observed in Table 2. The ground truth report describes multiple devices (endotracheal tube, orogastric tube, right internal jugular catheter),

Table 4 Qualitative Comparison in *Zero-Shot* Setting for LLaVA-Med. Highlighting indicates: incorrect/hallucinated findings, incomplete descriptions, superfluous details, accurate content

Zero-Shot Setting	
Chest X-Ray Image	**Ground Truth Report**
	Endotracheal tube terminates 2 cm above the carina. Orogastric tube terminates in the stomach. Right internal jugular catheter terminates in the mid SVC. Lungs are low in volume with stable right upper lung opacities which are better assessed on the recent chest CT but suspicious for pneumonia. There is no pneumothorax or pleural effusion. Heart is normal in size. Normal cardiomediastinal silhouette.
	SpLiCE Concepts: *pulmonary vascularity, overt pulmonary, right upper, endotracheal tube, internal jugular*
Image-Only	The chest X-ray image shows bilateral diffuse infiltrates, which are abnormal substances or collections of cells that have entered the lung tissue. These infiltrates can be seen in both lungs and are distributed throughout the lung fields.
+ Concepts	The chest X-ray shows overt pulmonary edema, which is a condition where there is an abnormal accumulation of fluid in the lungs. Additionally, there is a right upper lobe opacity, which could be due to various reasons such as infection, inflammation, or other lung abnormalities. The internal jugular vein is visible, and an endotracheal tube is in place, which is a tube that is inserted through the mouth or nose and into the trachea to help maintain an open airway and assist with breathing. Lastly, the pulmonary vascularity appears to be normal.
+ RAG	The chest X-ray findings are similar to those in the earlier study. The main differences observed are the placement of a left IJ catheter and the presence of an indistinct area of opacification in the right upper zone medially, which could represent a developing consolidation. Additionally, there is a right IJ central venous catheter with its tip at the lower SVC near the cavoatrial junction, and the endotracheal tube has been slightly advanced. No pneumothorax is noted. Pulmonary vascular congestion is observed.
+ CEMRAG	In the right upper zone, there is an indistinct area of opacification, which could represent a developing consolidation. Additionally, there is a right internal jugular central venous catheter with its tip at the lower superior vena cava near the cavoatrial junction. The endotracheal tube has been slightly advanced, with its tip 2.2 cm above the carina. No pneumothorax is noted, and the pulmonary vascular congestion is present. The cardiomediastinal silhouette remains unchanged.

low lung volumes, and a right upper lobe opacity suspicious for pneumonia, explicitly ruling out pleural effusion and pneumothorax.

The baseline LLaVA-Med model produces a markedly incorrect description, hallucinating bilateral diffuse infiltrates and failing to mention any of the indwelling devices, in line with the very low *Zero-Shot* F1-RG observed quantitatively. This behaviour indicates that, without additional guidance, the model tends to generate generic patterns of abnormality that are not well grounded in the specific image. With *Concepts* augmentation, all extracted keywords (e.g., "endotracheal tube", "internal jugular", "right upper") are explicitly mentioned in the generated text, and the report correctly identifies both the endotracheal tube and a right upper lobe opacity. However, the model now over-interprets the concept set by asserting overt pulmonary edema, a finding not present in the reference report. This illustrates how SpLiCE decomposition can substantially increase CheXbert Micro-F1 by encouraging the mention of clinically salient terms, while still lagging behind retrieval-based strategies on metrics that reward correct entity–relation structure. Notably, the model produces verbose explanations of medical concepts (e.g., describing what an endotracheal tube does), as LLaVA-Med's general medical pretraining has not been adapted to the concise technical language conventions of radiology reports. RAG-only reduces some of these concept-driven hallucinations but introduces others. Conditioning on retrieved in-domain reports leads the model to blend details from similar but non-identical cases, such as mentioning a left internal jugular catheter and pulmonary vascular congestion that are not supported by the target image. This exemplifies a typical retrieval-induced failure mode in which information from nearest neighbours leaks into the generated report. The combined CEMRAG condition achieves the most clinically plausible description. It correctly identifies the right internal jugular catheter, the endotracheal tube position, and the right upper zone opacity, and avoids the more severe hallucinations observed in the RAG-only report, although it still mentions mild vascular congestion. This example mirrors the quantitative trends: concept cues alone can act as over-strong priors, and retrieval alone can import spurious details, whereas their combination yields better grounded descriptions that more closely match the reference report.

SFT The second example (Table 5) illustrates behaviour in the *SFT* setting on MIMIC-CXR. The reference report describes low lung volumes and bibasilar opacities most likely representing atelectasis, and explicitly states the absence of pleural effusion or pneumothorax, with a stable and unremarkable cardiomediastinal silhouette.

The Image-Only baseline captures part of this picture: it correctly rules out pleural effusion and pneumothorax and describes normal cardiac and mediastinal contours, but introduces a clinically relevant error by emphasising a *new* left lower lobe opacity concerning for pneumonia and omitting any mention of low lung volumes. This combination of partial correctness and misplaced emphasis is consistent with the moderate CheXbert Micro-F1 and F1-RG achieved by the *SFT* baseline in Table 2. With *Concepts* augmentation, the model more faithfully reflects the ground truth: it explicitly recognises lower lung volumes and describes left basal opacification as atelectatic change, while conditionally mentioning possible pneumonia "in the appropriate clinical setting", closely echoing the phrasing of the reference report. The main discrepancy is the mention of a possible small effusion, which is explicitly ruled out in the ground truth. Overall, this aligns with the observed improvements in F1-RG and sequence-level metrics under SpLiCE. In the RAG-only condition, the report fails to mention the bibasilar opacities and instead states that the lungs are clear without focal consolidation. This represents a clinically concerning omission and is consistent with a failure mode in which the model appears to rely too heavily on retrieved studies and does not fully ground its description in the current image. The combined CEMRAG report strikes a better balance: it correctly identifies low lung volumes and patchy basal opacities likely reflecting atelectasis, preserves the correct absence of pleural effusion and pneumothorax, and maintains a coherent description of the cardiomediastinal silhouette. This progression from baseline to CEMRAG mirrors the quantitative gains reported for MIMIC-CXR and qualitatively illustrates how concept-level signals can prevent complete omission of pathologies when retrieval alone is unreliable.

5.2.2 Interpretability Through Concept Visualization

Beyond improvements in clinical accuracy, a central motivation of CEMRAG is to make the visual evidence underlying generated reports explicitly inspectable. In this section, we show how SpLiCE concepts and gradient-based explanations can be combined to reveal where the model "looks" in the image when producing specific medical terms in the report.

For each case, we select SpLiCE concepts that appear in the generated report (e.g., *endotracheal tube, right upper, median sternotomy, bilateral pleural*). We apply GradECLIP (Zhao et al., 2024a) to compute the gradient of the

Table 5 Qualitative Comparison in *Supervised Fine-Tuning* Setting or LLaVA-Med. Highlighting indicates: incorrect/hallucinated findings, incomplete descriptions, superfluous details, accurate content

Supervised Fine-Tuned Setting	
Chest X-Ray Image	**Ground Truth Report**
	Frontal and lateral views of the chest were obtained. There are low lung volumes which accentuate the bronchovascular markings. Bibasilar opacities are seen, which most likely represent atelectasis, although aspiration or infection are not excluded in the appropriate clinical setting. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are stable and unremarkable.
	SpLiCE Concepts: *lateral view, lung base, volume low, left hemidiaphragm, appropriate clinical*
Image-Only	The heart size is normal. The hilar and mediastinal contours are normal. There is a new left lower lobe opacity, concerning for pneumonia. No pleural effusion or pneumothorax. The visualized osseous structures are unremarkable.
+ Concepts	In comparison with the study of ___, there are lower lung volumes. Cardiac silhouette is within normal limits. Opacification at the left base is consistent with atelectatic changes and possible small effusion. In the appropriate clinical setting, superimposed pneumonia would have to be considered.
+ RAG	The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable.
+ CEMRAG	Lung volumes are low. Mediastinal and hilar contours are unremarkable. Pulmonary vasculature is not engorged. Patchy opacities in the lung bases likely reflect areas of atelectasis. No pleural effusion or pneumothorax is present. No acute osseous abnormality is visualized.

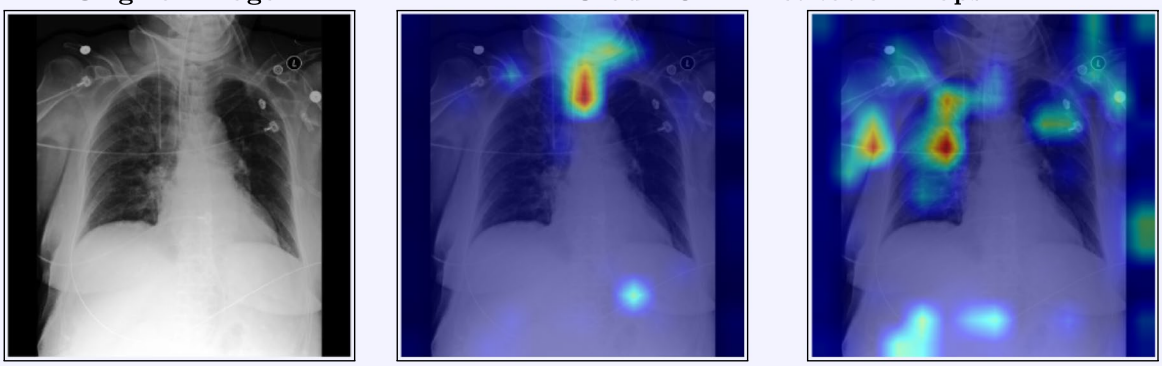
image–text similarity score with respect to the visual features used by the LLaVA-Med vision encoder. The resulting relevance map is then upsampled and overlaid on the chest X-ray as a heat map, yielding a concept-specific visualization of which regions support that term. CEMRAG-generated reports and their corresponding concept heat maps are illustrated in Table 6 for two representative MIMIC-CXR cases, one in the *Zero-shot* regime and one after *SFT*. In each example, the left column shows the reference report, while the right column shows the CEMRAG output with the relevant concept tokens highlighted. Below these reports, we display the original image together with Grad-ECLIP maps for two selected concepts. In the *Zero-shot* example, concepts such as endotracheal tube and right upper are both present in the generated report and exhibit Grad-ECLIP activations that concentrate along the tracheal tube and in the right upper lung zone, respectively. In the *SFT* example, the concept median sternotomy yields strong activation along the midline sternal wires and retrosternal region, while bilateral pleural produces strongest activation in the lower lung zones near the right and left costophrenic angles, where pleural effusions typically accumulate. Taken together, these patterns are consistent with standard radiological practice and suggest that, when a concept appears in the report, it is usually backed by visually plausible evidence in the corresponding image regions.

These concept-specific maps serve two complementary purposes. First, they allow clinicians to verify that SpLiCE-derived concepts, when injected into the prompt, are grounded in image regions that are consistent with their radiological meaning, rather than reflecting arbitrary directions in embedding space. Second, they expose failure modes in a transparent way: if a concept is mentioned in the report but its Grad-ECLIP map does not align with plausible anatomy, this discrepancy can be inspected, questioned, and potentially used to flag low-trust outputs. At the same time, this analysis remains focused on the vision component: Grad-ECLIP explains how visual encoders support individual concepts, but does not by itself reveal how the LLM composes multiple concepts into full sentences. Extending

Table 6 Qualitative Examples with Concept-Specific Visual Explanations. Grad-ECLIP heat maps highlight image regions that influence the LLaVA-Med encoder's matching between input images and SpLiCE concepts (highlighted in blue). The visualizations show which spatial features contribute to concept identification in the generated reports. Top row: Zero-Shot setting; bottom row: supervised fine-tuning setting

Ground Truth	Generated Report
Endotracheal tube terminates 2 cm above the carina. Orogastric tube terminates in the stomach. Right internal jugular catheter terminates in the mid SVC. Lungs are low in volume with stable right upper lung opacities which are better assessed on the recent chest CT but suspicious for pneumonia. There is no pneumothorax or pleural effusion. Heart is normal in size. Normal cardiomediastinal silhouette.	The chest X-ray shows several findings. In the **right upper** zone, there is an indistinct area of opacification, which could represent a developing consolidation. Additionally, there is a right internal jugular central venous catheter with its tip at the lower superior vena cava near the cavoatrial junction. The **endotracheal tube** has been slightly advanced, with its tip 2.2 cm above the carina. No pneumothorax is noted, and the pulmonary vascular congestion is present. The cardiomediastinal silhouette remains unchanged.

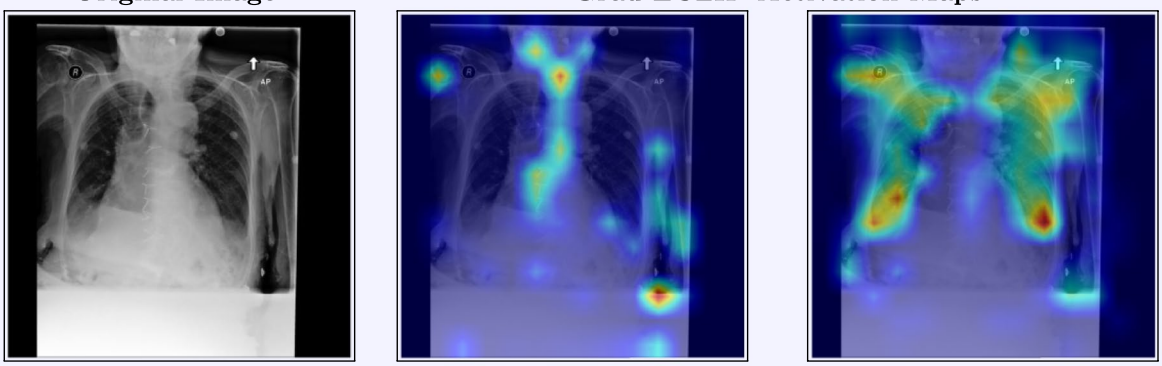
Original Image | **Grad-ECLIP Activation Maps**



Endotracheal Tube | *Right Upper*

Ground Truth	Generated Report
The patient is status post median sternotomy and aortic valve replacement. Heart size is moderately enlarged but unchanged. The aorta is tortuous and calcified. There is mild interstitial pulmonary edema, relatively unchanged. At least small bilateral pleural effusions are present. Bibasilar airspace opacities may reflect compressive atelectasis. There is no pneumothorax. There is evidence of prior vertebroplasty at the thoracolumbar junction.	The patient is status post **median sternotomy** and CABG. The heart size remains mildly enlarged. The aorta is tortuous and diffusely calcified. There is mild pulmonary edema, worse in the interval, with small **bilateral pleural** effusions. Bibasilar airspace opacities likely reflect atelectasis. No pneumothorax is identified. There are no acute osseous abnormalities.

Original Image | **Grad-ECLIP Activation Maps**



Median Sternotomy | *Bilateral Pleural*

concept-level tracing to the language component remains an important direction for future work toward fully interpretable medical VLMs.

6 Conclusion

This work addressed two major barriers to the clinical deployment of vision–language models in radiology: limited interpretability and susceptibility to hallucinations. We introduced CEMRAG, a unified framework that integrates concept decomposition with multimodal RAG to jointly enhance transparency and factual accuracy in radiology report generation. We further established a comprehensive evaluation protocol that compares multiple prompting strategies across different VLM architectures, retrieval configurations, and datasets, using both lexical similarity and clinically oriented correctness metrics. Across experiments on MIMIC-CXR and IU X-ray, our results indicate that interpretable visual concepts can improve factual grounding and concept-level transparency simultaneously, challenging the commonly assumed trade-off between interpretability and performance in medical AI. From a clinical perspective, the proposed framework offers a practical way to present AI-generated draft reports together with explicit visual concepts and retrieved reference cases, potentially facilitating more efficient review while preserving radiologists' ability to verify how findings in the image relate to the generated text.

Despite these advances, several limitations warrant further investigation. The overall effectiveness of the framework depends critically on the quality of the SpLiCE decomposition: if the underlying CLIP encoders are not sufficiently aligned with domain-specific semantics, the extracted concepts may be noisy or incomplete. Improving this alignment, for example through domain-adaptive pretraining or architectural refinements, is an important direction for future work. Moreover, interpretability in the current pipeline is primarily concentrated in the vision encoder and concept layer, while the language model is influenced only indirectly via prompt conditioning. Future research should explore mechanisms that more directly constrain or regularize token probabilities during generation, extending interpretability to the full model. Finally, our experiments focused on an LLM backbone such as Mistral-7B; however, recent progress on smaller language models with competitive performance and lower computational cost suggests a promising avenue for achieving finer-grained control over interpretability and deployment in resource-constrained clinical settings.

By showing that concept-level interpretability can enhance rather than undermine factual accuracy, this work provides empirical support for the development of transparent and accurate VLMs for radiology. The modular design of CEMRAG enables targeted optimization of individual components and offers a general methodological template that can be extended beyond chest X-ray analysis to other medical imaging domains where visual interpretation and textual reporting are required, provided that suitable domain-specific concept vocabularies and retrieval corpora are available.

Appendix A Visual Concept Extraction

For extracting interpretable visual concepts from CLIP embeddings, we employ Sparse Linear Concept Embeddings (SpLiCE) (Bhalla et al., 2024). Given the CLIP visual embedding $\mathrm{v} \in \mathbb{R}^d$ extracted from the input image I, SpLiCE approximates this embedding as a sparse linear combination of concept embeddings drawn from a learned vocabulary.

The concept vocabulary construction begins with a set of m medical concepts $\mathrm{Q} = [q_1, q_2, \ldots, q_m]$ encompassing radiological findings, anatomical structures, and pathological patterns relevant to chest radiograph interpretation. Each concept $q_j \in \mathrm{Q}$ is encoded through the CLIP text encoder $\mathrm{E}_{\mathrm{txt}}$ to obtain its embedding $\mathrm{c}_j = \mathrm{E}_{\mathrm{txt}}(q_j) \in \mathbb{R}^d$, and we collect these embeddings into the matrix $\mathrm{C} = [\mathrm{c}_1, \mathrm{c}_2, \ldots, \mathrm{c}_m] \in \mathbb{R}^{d \times m}$. We denote by $\sigma(\mathrm{x}) = \mathrm{x}/\|\mathrm{x}\|_2$ the normalization operator. Let $\boldsymbol{\mu}_{\mathrm{c}} \in \mathbb{R}^d$ be the mean concept embedding computed over C. The centered and normalized vocabulary is then given by $\tilde{\mathrm{c}}_j = \sigma(\mathrm{c}_j - \boldsymbol{\mu}_{\mathrm{c}})$, and $\tilde{\mathrm{C}} = [\tilde{\mathrm{c}}_1, \tilde{\mathrm{c}}_2, \ldots, \tilde{\mathrm{c}}_m] \in \mathbb{R}^{d \times m}$, where each element represents a centered and normalized concept embedding. Analogously, we center and normalize the image embedding. Let $\boldsymbol{\mu}_{\mathrm{img}} \in \mathbb{R}^d$ denote the mean CLIP image embedding computed over the training corpus. The centered and normalized image embedding is then given by $\tilde{\mathrm{v}} = \sigma(\mathrm{v} - \boldsymbol{\mu}_{\mathrm{img}})$. The sparse decomposition is formulated as an optimization problem that balances reconstruction accuracy against sparsity:

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha} \geq 0} \|\tilde{\mathrm{C}}\boldsymbol{\alpha} - \tilde{\mathrm{v}}\|_2^2 + 2\lambda\|\boldsymbol{\alpha}\|_1 \tag{A1}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_m] \in \mathbb{R}^m_{\geq 0}$ represents the coefficient vector encoding the contribution of each concept. The first term enforces reconstruction fidelity in the centered embedding space, while the second term promotes sparsity with the regularization parameter $\lambda > 0$ controlling the trade-off. The solution to this optimization yields a sparse coefficient vector $\boldsymbol{\alpha}^* = [\alpha_1^*, \ldots, \alpha_m^*] \in \mathbb{R}^m_{\geq 0}$, where only a small subset of entries are non-zero. To obtain the final set of concept keywords, we rank concepts by their corresponding coefficient magnitudes and select the top-τ concepts. These selected

concepts correspond to interpretable keywords $\mathbf{\Omega}$ that represent clinically relevant visual features present in the input image. The resulting keyword set $\mathbf{\Omega}$ provides transparency into which aspects of the image inform the subsequent RRG process and serves as the concept component for the prompt augmentation.

A.1 Vocabulary Construction and Hyperparameter Choice

The vocabulary was constructed by extracting the most frequent bigrams from the training corpus, with systematic exclusion of English stopwords and common medical acronyms. The decision to focus on bigrams rather than unigrams is motivated by the inherent compositional nature of radiological terminology. Many clinically meaningful concepts emerge only through the combination of terms. For instance, "pleural effusion" and "cardiomediastinal silhouette" convey precise diagnostic information that cannot be adequately represented by their constituent words in isolation. All extracted terms underwent lemmatization to ensure morphological normalization and direct compatibility with downstream textual prompts fed to the LLM.

Having established the vocabulary construction methodology, we conducted systematic ablation studies to determine optimal parameters for the SpLiCE decomposition. The optimization process involved joint exploration of three critical hyperparameters: vocabulary size, L1 regularization strength (λ), and the number of top-ranked concepts (τ) selected for each image. Our objective was to identify a configuration that simultaneously satisfies three competing criteria: (i) precision, (ii) cosine similarity and (iii) sparsity. Precision, defined as the fraction of extracted SpLiCE concepts that appear in the corresponding ground truth radiology report, serves as our primary quality metric. This measure directly quantifies the clinical relevance and factual accuracy of the extracted concepts. High precision is essential in our framework, as these concepts are incorporated into retrieval-augmented prompts that guide report generation. Cosine similarity between the original CLIP image embedding and its sparse reconstruction quantifies the fidelity of the decomposition. This metric ensures that the dimensionality reduction and sparsification process preserves the essential visual information encoded in the original representation. Sparsity, measured as the average number of non-zero coefficients per image, reflects the interpretability-informativeness trade-off. Excessive sparsity may omit clinically relevant concepts, while insufficient sparsity produces verbose, representations that overwhelm the downstream LLM with redundant information.

Figure 3 presents a comprehensive analysis across five vocabulary sizes, corresponding to the top 100, 200, 500, 700, and 1000 most frequent bigrams from the MIMIC-CXR training corpus, and three L1 penalty values ($\lambda \in \{0.1, 0.3, 0.5\}$). It should be noted that for vocabulary sizes below 500, results at $\lambda = 0.5$ are unavailable due to over-regularization: the strong sparsity penalty caused certain images to yield entirely zero-valued weight vectors, indicating that no concept exceeded the activation threshold. This phenomenon confirms the theoretical prediction that excessively high λ values can suppress all activations when the vocabulary-embedding alignment is insufficient. Precision trends (left panel of Fig. 3) reveal a pronounced inverse relationship with vocabulary size. This systematic pattern reflects a fundamental trade-off between lexical coverage and terminological precision: smaller vocabularies, comprising only the most frequently occurring clinical terms, naturally align with the standardized terminology that radiologists consistently employ in their reports, whereas larger vocabularies introduce lower-frequency terms that, despite exhibiting visual correlation with CLIP embeddings, often represent rare synonyms, anatomical descriptors, or overly specific variants absent from actual ground truth reports. Within each vocabulary configuration, increasing λ consistently enhances precision by enforcing greater sparsity, thereby selecting only the most strongly activated concepts while reducing false positive extractions. Reconstruction

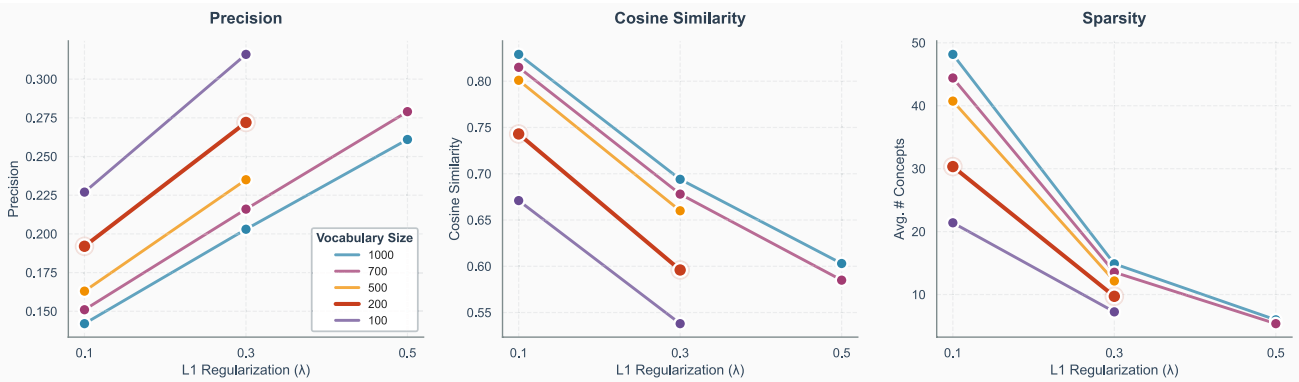


Fig. 3 SpLiCE performance across vocabulary sizes and λ values. The figure reports three complementary metrics: **Left**: precision of extracted concepts; **Center**: cosine similarity between the original CLIP embedding and its sparse reconstruction; **Right**: average number of active concepts (sparsity). The results highlight the trade-off between fidelity, interpretability, and terminological precision

fidelity (central panel of Fig. 3) exhibits the opposite trend: cosine similarity increases monotonically with vocabulary size, ranging from 0.671 for the 100-bigram vocabulary at $\lambda = 0.1$ to 0.829 for the 1000-bigram vocabulary at the same regularization strength. This observation is expected, as larger dictionaries provide greater representational capacity to approximate the original high-dimensional CLIP embedding through linear combinations. However, this improved reconstruction comes at the documented cost of reduced precision. Sparsity characteristics (right panel of Fig. 3) demonstrate that the average number of active concepts decreases both with vocabulary size reduction and λ increment. Large vocabularies with minimal regularization, such as the 1000-bigram vocabulary at $\lambda = 0.1$, produce approximately 48 non-zero coefficients per image, a density incompatible with interpretable prompting and efficient retrieval. Conversely, aggressive sparsification achieved through the 100-bigram vocabulary at $\lambda = 0.3$ yields only 6-7 active concepts, representing a focused subset that provides sufficient contextual information to guide generation while avoiding overwhelming the language model with excessive detail. In selecting the final configuration, we assigned priority to precision maximization, recognizing that extracted concepts directly influence the factual accuracy of generated reports, thereby identifying a vocabulary size of 100 or 200 as the most promising candidates, both achieving precision exceeding 0.27.

To further refine this selection and ensure consistency across the dataset, we conducted an additional experiment evaluating the impact of selecting a fixed number τ of top-ranked concepts per image, with $\tau \in [3, 5, 7]$. This analysis served two purposes: first, to assess whether further reduction in concept count would significantly degrade precision; and second, to standardize the input representation such that all images contribute exactly the same number of keywords to the retrieval-augmented prompts, facilitating uniform processing by the downstream LLM. Results are presented in Fig. 4, where precision decreases monotonically with increasing τ for both vocabularies, consistent with the progressive inclusion of concepts exhibiting lower activation coefficients. Notably, the precision advantage of 100-bigram vocabulary diminishes as τ increases, from a 3.9% margin at $\tau = 3$ to 2.8% at $\tau = 7$, suggesting that the highest-ranked concepts exhibit comparable accuracy across both vocabularies, with divergence manifesting primarily in lower-ranked selections. We adopt $\tau = 5$ with 200-bigram vocabulary as our definitive configuration (precision=0.313). While the vocabulary size of 100 offers marginally superior precision, the 200-bigram vocabulary provides substantially broader lexical coverage, encompassing twice the vocabulary size, with only modest precision degradation. This configuration embodies

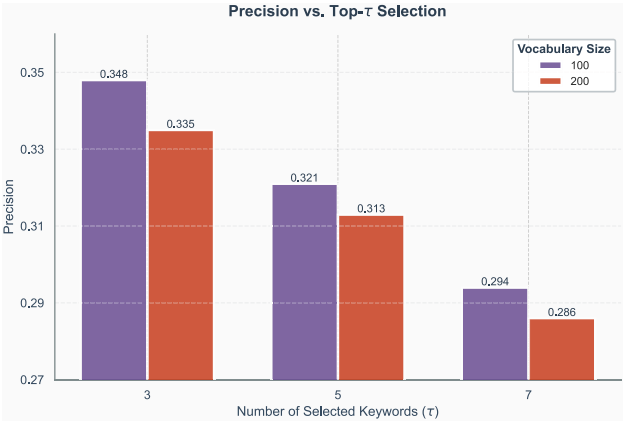


Fig. 4 Precision as a function of the number of selected concepts $\tau \in \{3, 5, 7\}$ for both the 100- and 200-bigram vocabularies. Precision decreases monotonically with increasing τ, reflecting the progressive inclusion of lower-ranked concepts with weaker activation strengths. The 100-bigram vocabulary maintains a consistent advantage across all τ values, indicating that the highest-confidence concepts are highly informative for guiding retrieval-augmented report generation

a principled balance between factual grounding through high precision and expressive capacity through adequate lexical diversity, optimizing the extracted concepts for their ultimate role in guiding accurate and comprehensive RRG via retrieval-augmented prompting.

Appendix B Training Details

CXR-CLIP Adaptation For the IU X-ray experiments, we adapt CXR-CLIP using LoRA (Hu et al., 2022). The LoRA modules are applied to the last stage of the Swin vision encoder and to the final BERT text encoder layer, with rank $r = 8$, scaling parameter $\alpha = 16$, and dropout rate 0.1. Training is performed for ten epochs using the AdamW optimizer, with a learning rate of 5×10^{-5} and a weight decay of 0.01.

Projection Module $\Phi_{\mathbf{CLIP}}$ The projection module Φ_{CLIP} is implemented as a single linear layer that maps the 768-dimensional CLIP visual representation from the last vision transformer block (before the standard CLIP projection head) to the LLM embedding dimension. As an alignment step, Φ_{CLIP} is initially trained for one epoch with both the CLIP encoder and the LLM frozen, using a learning rate of 1×10^{-3}, a warmup ratio of 0.03, cosine learning-rate scheduling, and a batch size of 16.

Supervised Fine-Tuning (SFT) For SFT of the LLM, we again employ LoRA with rank $r = 64$, scaling parameter $\alpha = 16$, and dropout rate 0.05. Fine-tuning is carried out

for three epochs with a batch size of 16, a learning rate of 1×10^{-4}, a warmup ratio of 0.03, and cosine learning-rate scheduling, while keeping all visual encoders frozen.

Appendix C Computational Cost Details

Table 7 reports the per-image computational cost for each augmentation strategy, averaged over the full MIMIC-CXR test set using LLaVA-Med with *SFT* on a single NVIDIA A100 GPU. All GFLOPs are measured using the `fvcore` library (Research, 2021), which traces the actual computation graph of each module.

As shown in Table 7, CEMRAG requires 7,540 GFLOPs per image compared to 4,876 for the Image-Only baseline. This difference is not attributable to the pre-generation pipeline, which comprises CXR-CLIP encoding (10.1 ms), the SpLiCE Lasso solver (7.3 ms), and FAISS retrieval (57.2 ms), accounting for at most 74.6 ms with no neural-network FLOPs. Rather, the dominant source of computational variation is the LLM input processing stage, which scales linearly with prompt length: from 600 average tokens for the *Image-Only* baseline to 966 for CEMRAG. The vision encoder and projection MLP contribute identical costs across all conditions, as they process the same input image regardless of the augmentation strategy. Regarding training cost, *SFT* for CEMRAG requires 14 GPU-hours on four NVIDIA A100 GPUs, only one hour more than RAG and seven more than the *Image-Only* baseline. This one-time overhead does not affect inference and is justified by the simultaneous gains in clinical accuracy and concept-level interpretability.

Table 7 Per-image computational cost and clinical accuracy for each augmentation strategy, evaluated on the MIMIC-CXR test set with LLaVA-Med under SFT. GFLOPs and inference time are averaged over test images on a single NVIDIA A100 GPU. Training time refers to SFT on four NVIDIA A100 GPUs

Method	GFLOPs	Time (s)	Training (GPU-h)	Micro-F1$_{14}$
Image-Only	4,876	1.99	7	0.470
+ Concepts	5,160	2.45	8	0.476
+ RAG	6,959	2.45	13	0.477
+ CEMRAG	7,540	2.55	14	**0.488**

Appendix D Failure Analysis

Our modular design enables principled diagnosis of failure cases by localizing errors to specific pipeline components. We identify two distinct failure modes, illustrated in Tables 8 and in 9.

Failure Mode 1: Incorrect SpLiCE Concept Extraction In these cases, SpLiCE extracts concepts that do not correspond to actual findings in the image. In the *Zero-Shot* example, the incorrect concepts "monitoring support" and "moderate cardiomegaly" lead the model to hallucinate an orogastric tube and an enlarged cardiac silhouette, neither of which appear in the ground truth. Similarly, in the *SFT* example, three out of five extracted concepts ("since prior", "low lung", and "left base") are absent from the ground truth. However, only "left base" visibly propagates into the generated report, producing "left basal atelectasis" instead of the bilateral pattern documented in the reference. The remaining incorrect concepts are effectively filtered out, suggesting that retrieved reports may serve as an implicit corrective signal, allowing the LLM to discount spurious keywords that lack support in similar cases.

Failure Mode 2: Correct Concepts Ignored by the LLM In these cases, SpLiCE correctly identifies clinically relevant concepts present in the ground truth, but the LLM fails to incorporate them during generation. In the *Zero-Shot* example, the concepts "vascular congestion" and "lateral view" are both present in the reference report, yet the model omits them entirely. In the *SFT* example, the concept "central venous" is correctly extracted and confirmed by the ground truth, which describes a "right internal jugular central venous catheter". However, the model reduces this to a generic "right jugular line", losing the precise clinical terminology. Similarly, "small right" is present in the reference as "small right pleural effusion", but the generated report describes it as "right mild pleural effusion", altering the severity characterization. These cases indicate that the LLM does not consistently adhere to the concept keywords injected in the prompt, suggesting that its learned generation priors can override explicit prompt guidance.

Table 8 Failure Mode 1: Incorrect SpLiCE concept extraction. SpLiCE extracts concepts absent from the ground truth, potentially inducing hallucinations in the generated report. Concept highlighting: correct and used, wrong (absent from ground truth)

	Zero-Shot Setting
SpLiCE Concepts	retrocardiac opacity, venous catheter, et tube, monitoring support, moderate cardiomegaly
Ground Truth	There is an endotracheal tube whose distal tip is 6.2 cm above the carina appropriately sited. There is a left-sided IJ line with distal lead tip in the mid SVC. There is a nasogastric tube whose tip and sideport are below the GE junction. There is a persistent left retrocardiac opacity. There is some atelectasis at the left lung base. There is improved aeration at the right lung base. No pneumothoraces are seen.
CEMRAG	The chest X-ray shows a persistent left retrocardiac opacity with air bronchograms, reflecting atelectasis or consolidation. This finding is unchanged since the prior examination. Additionally, there is a new atelectasis at the level of the left lower lobe. The monitoring and support devices, such as the endotracheal tube, left IJ catheter, and orogastric tube, are constant in size. The cardiac silhouette is enlarged, and there is evidence of pulmonary vascular congestion.
	Supervised Fine-Tuning Setting
SpLiCE Concepts	median sternotomy, low lung, since prior, basilar atelectasis, left base
Ground Truth	PA and lateral chest radiographs were provided. Lung volumes are significantly low. There is no focal consolidation, pleural effusion or pneumothorax. There is bibasilar atelectasis. The cardiomediastinal silhouette is unchanged. Median sternotomy wires are intact. A right chest wall Port-A-Cath terminates at the cavoatrial junction. There is no free air under the hemidiaphragms. Osseous structures are intact.
CEMRAG	PA and lateral views of the chest provided. Port-A-Cath resides over the right chest wall with catheter tip in the region of the mid SVC. Midline sternotomy wires and mediastinal clips are again noted. Lung volumes are low limiting assessment. There is mild left basal atelectasis. No convincing signs of pneumonia or edema. No large effusion or pneumothorax. Cardiomediastinal silhouette is stable. Bony structures are intact.

These failure modes suggest two complementary directions for future improvement. On the vision side, the precision of SpLiCE concept extraction could be improved by adopting CLIP models trained to produce more decomposable and concept-aligned representations. Recent work on adversarial fine-tuning for visual interpretability (Gong et al., 2025) and disease-aware contrastive pretraining for medical imaging (Phan et al., 2024) shows that training objectives can be designed to yield sparser, more semantically faithful embeddings, directly benefiting downstream concept decomposition. On the language side, contrastive decoding techniques (Leng et al., 2024; Mahdavi et al., 2026) offer a promising direction to enforce stronger adherence to the provided concepts during generation, steering the LLM toward incorporating visual evidence without requiring architectural modifications or additional training.

Table 9 Failure Mode 2: Correct concepts ignored by the LLM. SpLiCE correctly identifies concepts present in the ground truth, but the LLM ignores them during generation. Concept highlighting: correct and used, wrong (absent from groundtruth), correct but ignored by LLM

Zero-Shot Setting	
SpLiCE Concepts	sternotomy wire, median sternotomy, moderate cardiomegaly, vascular congestion, lateral view
Ground Truth	AP and lateral views of the chest. The lungs are clear of focal consolidation, effusion, or pulmonary vascular congestion. Cardiac silhouette is mildly enlarged, similar to priors. Hypertrophic changes noted in the spine. Median sternotomy wires are again noted.
CEMRAG	The chest radiographs show clear lungs, with no focal consolidation, effusion, or overt pulmonary edema. The cardiomediastinal silhouette is mildly enlarged. Median sternotomy wires are identified in the images.
Supervised Fine-Tuning Setting	
SpLiCE Concepts	right pleural, right ij, central venous, lung otherwise, small right
Ground Truth	A right internal jugular central venous catheter tip terminates in the mid SVC. No pneumothorax is identified. Moderate to severe cardiomegaly persists. Mediastinal and hilar contours are unchanged. A septal closure device is noted again. There is a small right pleural effusion with atelectatic changes in the right lung base. Left lung remains clear.
CEMRAG	The right jugular line has been pulled back and is now in the lower portion of the superior vena cava. There is no pneumothorax. Stability of the right mild pleural effusion. Mild bibasilar atelectasis. The mediastinal and cardiac contours are unchanged. There is a left carotid stent. Stability of the calcified round density at the right lung apex measuring 1.5 cm.

Acknowledgements Marco Salmè is a Ph.D. student enrolled in the National Ph.D. in Artificial Intelligence, XXXIX cycle, course on Health and Life Sciences, organized by Università Campus Bio-Medico di Roma. This work was partially funded by: i) Università Campus Bio-Medico di Roma under the program "University Strategic Projects" within the project "AI-powered Digital Twin for next-generation lung cancEr cAre (IDEA)"; ii) PNRR MUR project PE0000013-FAIR. iii) Cancerforskningsfonden Norrland project MP23-1122; iv) Kempe Foundation project JCSMK24-0094. Resources are provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at Alvis @ C3SE, partially funded by the Swedish Research Council through grant agreements no. 2022-06725 and no. 2018-05973.

Author Contributions 1. Marco Salmè - Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft 2. Federico Siciliano - Conceptualization, Methodology, Writing – review & editing 3. Fabrizio Silvestri - Funding acquisition, Project administration 4. Paolo Soda - Funding acquisition, Project administration, Resources, Supervision 5. Rosa Sicilia - Data curation, Project administration, Writing – review & editing 6. Valerio Guarrasi - Data curation, Methodology, Project administration, Supervision, Writing – review & editing

Funding Open access funding provided by Umea University. This work was partially funded by: i) Università Campus Bio-Medico di Roma under the program "University Strategic Projects" within the project "AI-powered Digital Twin for next-generation lung cancEr cAre (IDEA)"; ii) PNRR MUR project PE0000013-FAIR. iii) Cancerforskningsfonden Norrland project MP23-1122; iv) Kempe Foundation project JCSMK24-0094. Resources are provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at Alvis @ C3SE, partially funded by the Swedish Research Council through grant agreements no. 2022-06725 and no. 2018-05973.

Data Availability The datasets used in this study are publicly available. Specifically, we use the MIMIC-CXR dataset (Johnson et al., 2019) and the IU X-ray dataset (Demner-Fushman et al., 2015), which can be accessed subject to their respective data use agreements. Additional details required to reproduce the experiments are provided in the Appendix.

Code Availability Not mentioned.

Declarations

Competing Interests The authors declare that they have no competing financial or non-financial interests.

Consent for Publication All authors have reviewed and approved the final version of the manuscript and consent to its submission and potential publication in this journal.

Ethical Approval This study does not involve any new data collection from human participants or animals. All experiments were conducted using publicly available, de-identified datasets. Therefore, ethical approval was not required.

Informed Consent Not applicable. The study uses publicly available datasets that were collected with informed consent under the original data collection protocols.

Animal Welfare Not applicable. This research did not involve experiments on animals.

References

Abootorabi, M.M., Zobeiri, A., Dehghani, M., Mohammadkhani, M., Mohammadi, B., Ghahroodi, O., Baghshah, M.S., & Asgari, E. (2025). Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. arXiv:2502.08826

Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical bert embeddings. arXiv:1904.03323

Balasubramanian, S., Basu, S., & Feizi, S. (2024). Decomposing and interpreting image representations via text in vits beyond clip. *Advances in Neural Information Processing Systems, 37*, 81046–81076.

Bereska, L., & Gavves, S. (2024). Mechanistic interpretability for ai safety-a review. *Transactions on Machine Learning Research*

Bernardi, M.L., Cimitile, M., et al. (2024). Report generation from x-ray imaging by retrieval-augmented generation and improved image-text matching. In *2024 International Joint Conference on Neural Networks (IJCNN)*, (pp. 1–8). IEEE.

Bhalla, U., Oesterling, A., Srinivas, S., Calmon, F., & Lakkaraju, H. (2024). Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems, 37*, 84298–84328.

Caragliano, A. N., Tacconi, C., Greco, C., Nibid, L., Ippolito, E., Fiore, M., ... & Guarrasi, V. (2025). Multimodal doctor-in-the-loop: A clinically-guided explainable framework for predicting pathological response in non-small cell lung cancer. In *2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN64981.2025.11228780

Chen, Y., Sikka, K., Cogswell, M., Ji, H., Divakaran, A., et al. (2024). Measuring and improving chain-of-thought reasoning in vision-language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 192–210).

Chen, Z., Shen, Y., Song, Y., Wan, X., et al. (2021). Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 5904–5914).

Chu, Y.-W., Zhang, K., Malon, C., & Min, M. R. (2025). Reducing hallucinations of medical multimodal large language models with visual retrieval-augmented generation. In *Workshop on Large Language Models and Generative AI for Health at AAAI*.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems, 36*, 16318–16352.

Delbrouck, J.-B., Chambon, P., Bluethgen, C., Tsai, E., Almusa, O., Langlotz, C., et al. (2022). Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 4348–4360).

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., & McDonald, C. J. (2015). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association, 23*(2), 304–310.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., & Jégou, H. (2024). The faiss library. arXiv:2401.08281.

Ennab, M., & Mcheick, H. (2024). Enhancing interpretability and accuracy of ai models in healthcare: a comprehensive review on challenges and future directions. *Frontiers in Robotics and AI, 11*, 1444763.

Gandelsman, Y., Efros, A.A., & Steinhardt, J. (2023). Interpreting clip's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey.

Garcia-Carmona, A. M., Prieto, M.-L., Puertas, E., & Beunza, J.-J. (2025). Leveraging large language models for accurate retrieval of patient information from medical reports: Systematic evaluation study. *JMIR AI, 4*(1), 68776.

Gong, S., Haoyu, L., Dou, Q., Farnia, F., et al. (2025). Boosting the visual interpretability of clip via adversarial fine-tuning. In *The Thirteenth International Conference on Learning Representations*.

Hartsock, I., & Rasool, G. (2024). Vision-language models for medical report generation and visual question answering: A review. *Frontiers in Artificial Intelligence, 7*, 1430984.

He, J., Zhang, B., Rouhizadeh, H., Chen, Y., Yang, R., Lu, J., Chen, X., Liu, N., Li, I., & Teodoro, D. (2025). Retrieval-augmented generation in biomedicine: A survey of technologies, datasets, and clinical applications. arXiv:2505.01146.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), 1–55.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR, 1*(2), 3.

Huo, J., Yan, Y., Hu, B., Yue, Y., Hu, X., et al. (2024). Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 6801–6816).

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.-A., Stock, P., Scao, T.L., ... & Sayed, W.E. (2023). Mistral 7b.

Jiang, Z., Chen, J., Zhu, B., Luo, T., Shen, Y., Yang, X., et al. (2025). Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 25004–25014).

Jin, Y., & Zhang, Y. (2024). Orthodoc: Multimodal large language model for assisting diagnosis in computed tomography. arXiv:2409.09052.

Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., & Horng, S. (2019). Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv:1901.07042.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., Liang, P., et al. (2020). Concept bottleneck models. In *International Conference on Machine Learning* (pp. 5338–5348). PMLR.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L., et al. (2024). Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13872–13882).

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., & Gao, J. (2023). Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems, 36*, 28541–28564.

Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X., et al. (2023). Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 3334–3343).

Lin, C.-Y., et al. (2004). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, (pp. 74–81).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., et al. (2021a). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022).

Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y., et al. (2021b). Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13753–13762).

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems, 36*, 34892–34916.

Mahdavi, Z., Khodakaramimaghsoud, Z., Khaloo, H., Taleshani, S. B., Hashemi, E., Kaleybar, J. M., & Manzari, O. N. (2026). Medvcd: Mitigating hallucination for medical large vision language models through visual contrastive decoding. *Computers in Biology and Medicine, 200*, 111347.

Mistral AI (2024). Mistral AI documentation: Prompt engineering. https://docs.mistral.ai/. Accessed Apr 2025.

Molino, D., Caruso, C. M., Soda, P., & Guarrasi, V. (2026). Retrieval-augmented anatomical guidance for text-to-ct generation. https://arxiv.org/abs/2603.08305

Nguyen, T.T.H., Clement, T., Nguyen, P.T.L., Kemmerzell, N., Truong, V.B., Nguyen, V.T.K., Abdelaal, M., & Cao, H. (2024). Langxai: Integrating large vision models for generating textual explanations to enhance explainability in visual perception tasks. CoRR

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., et al. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, (pp. 311–318).

Parekh, J., Khayatan, P., Shukor, M., Newson, A., & Cord, M. (2024). A concept-based explainability framework for large multimodal models. *Advances in Neural Information Processing Systems, 37*, 135783–135818.

Park, D.H., Hendricks, L.A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M., et al. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 8779–8788).

Phan, V.M.H., Xie, Y., Qi, Y., Liu, L., Liu, L., Zhang, B., Liao, Z., Wu, Q., To, M.-S., Verjans, J.W., et al. (2024). Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 11492–11501).

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, (pp. 8748–8763). PMLR

Rao, S., Mahajan, S., Böhle, M., Schiele, B., et al. (2024). Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*, (pp. 444–461). Springer

Research, M. (2021). fvcore: FAIR Computer vision library. https://github.com/facebookresearch/fvcore. Accessed: 2025

Salmè, M., Sicilia, R., Soda, P., & Guarrasi, V. (2025). Evaluating vision language model adaptations for radiology report generation in low-resource languages. In *2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN64981.2025.11227552

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J. I., et al. (2025). Open problems in mechanistic interpretability. CoRR.

Sloan, P., Clatworthy, P., Simpson, E., & Mirmehdi, M. (2024). Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering, 18*, 368–387.

Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M., et al. (2020). Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1500–1519).

Sun, L., Zhao, J., Han, M., & Xiong, C. (2024). Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. arXiv:2407.15268.

Thetbanthad, P., Sathanarugsawait, B., & Praneetpolgrang, P. (2025). Application of generative artificial intelligence models for accurate prescription label identification and information retrieval for the elderly in northern east of thailand. *Journal of Imaging, 11*(1), 11.

Tozuka, R., Johno, H., Amakawa, A., Sato, J., Muto, M., Seki, S., Komaba, A., & Onishi, H. (2025). Application of notebooklm, a large language model with retrieval-augmented generation, for lung cancer staging. *Japanese Journal of Radiology, 43*(4), 706–712.

Van, M.-H., et al. (2024). On large visual language models for medical imaging analysis: An empirical study. In *2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)* (pp. 172–176). IEEE.

Vatsa, M., Jain, A., Singh, R., et al. (2024). Adventures of trustworthy vision-language models: A survey. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, pp. 22650–22658).

Wang, J., Ashraf, T., Han, Z., Laaksonen, J., & Anwer, R. M. (2025). Mira: A novel framework for fusing modalities in medical rag. arXiv:2507.07902.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.

Xia, P., Zhu, K., Li, H., Wang, T., Shi, W., Wang, S., Zhang, L., Zou, J., & Yao, H. (2024a). Mmed-rag: Versatile multimodal rag system for medical vision language models. In *The Thirteenth International Conference on Learning Representations.*

Xia, P., Zhu, K., Li, H., Zhu, H., Li, Y., Li, G., Zhang, L., & Yao, H. (2024b). Rule: Reliable multimodal rag for factuality in medical vision language models. CoRR.

You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E. K., Baek, W., Roh, B., et al. (2023). Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 101–111). Springer.

Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., Liu, Z., et al. (2024). Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data* (pp. 102–120). Springer.

Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(8), 5625–5644.

Zhang, K., Barrett, C.D., Kim, J., Sun, L., Taghavi, T., & Kenthapadi, K. (2025). Radagents: Multimodal agentic reasoning for chest x-ray interpretation with radiologist-like workflows. arXiv:2509.20490.

Zhang, Y., Merck, D., Tsai, E., Manning, C. D., Langlotz, C., et al. (2020). Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5108–5120).

Zhao, C., Wang, K., Zeng, X., Zhao, R., Chan, A.B., et al. (2024a). Gradient-based visual explanation for transformer-based CLIP. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (Eds.), *Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research* (Vol. 235, pp. 61072–61091). PMLR.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024b). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology, 15*(2), 1–38.

Zheng, G., Yang, B., Tang, J., Zhou, H.-Y., & Yang, S. (2023). Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems, 36*, 5168–5191.

Marco Salmè is a Ph.D. student in the Italian National Ph.D. Program in Artificial Intelligence for Healthcare and Life Sciences at University Campus Bio-Medico di Roma, where he works in the Unit of Artificial Intelligence and Computer Systems within the Department of Engineering. His current research focuses on trustworthiness in deep generative models applied to the medical domain.

Federico Siciliano is a researcher at the National Institute of Geophysics and Volcanology. He holds a PhD in Data Science with a thesis on Architectural Components of Trustworthy Artificial Intelligence with Prof. Fabrizio Silvestri. His current research projects include Climate Change, Information Retrieval, Recommender Systems and Explainable Artificial Intelligence. He had the privilege of partnering with numerous world-renowned institutions, including Cambridge University, Meta, Amazon, and UniPi.

Fabrizio Silvestri is a Full Professor and the coordinator of the Ph.D. in Data Science, at Dipartimento di Ingegneria informatica, automatica e gestionale (DIAG) of Sapienza University of Rome. His research interests lie in Artificial Intelligence, and in particular, machine learning applied to web search problems and natural language processing. He is the author of more than 150 papers in international journals and conference proceedings. He holds nine industrial patents. He is the holder of the ‘‘test-of-time’’ award at the ECIR 2018 conference for an article published in 2007. He is the holder of three best paper awards and other international awards. Fabrizio Silvestri spent eight years abroad in industrial research laboratories (Yahoo! and Facebook). Fabrizio Silvestri has a Ph.D. in computer science awarded by the University of Pisa.

Paolo Soda graduated with honours in Biomedical Engineering at University Campus Bio-Medico (UCBM), Rome, in 2004 and received a Ph.D Biomedical Engineering (Computer Science area) in 2008 from the same University. Currently he is Full Professor of Computer Science and Computer Engineering at UCBM, and visiting professor in AI at Umea University, Sweden. His research interests include artificial intelligence, pattern recognition, machine learning, with a focus on multimodal learning, generative AI and sequential data processing. Applications are directed towards data, signal, 2D and 3D image processing and analysis.

Rosa Sicilia is an Associate Professor in Information Processing Systems at UniCamillus – Saint Camillus International University of Health Sciences. Her research focuses on artificial intelligence, with emphasis on machine learning, deep learning, and multimodal data mining for biomedical and social data analysis. Her main interests include multivariate time series analysis, radiomics and radiopathomics, explainable AI (XAI), and multimodal models for precision medicine. She has received the Best Paper Award at the 34th IEEE CBMS 2021, served as General Chair of IEEE CBMS 2023 and Co-General Chair for the 2025 edition, and since 2025 chairs the IEEE CBMS Steering Committee. She is also a reviewer for high-impact journals and Guest Editor for Artificial Intelligence in Medicine and Health Information Science and Systems.

Valerio Guarrasi is an Assitant Professor at the Università Campus Bio-Medico di Roma, where he works in the Unit of Artificial Intelligence and Computer Systems within the Department of Engineering. He earned a PhD in Data Science from Sapienza University of Rome, following a Master’s Degree in Data Science and a Bachelor’s Degree in Management Engineering. His research focuses on multimodal deep learning and generative artificial intelligence (AI), with a strong emphasis on developing innovative methodologies for biomedical applications. He is particularly interested in how AI can enhance healthcare by improving diagnostics and patient outcomes. Dr. Guarrasi has contributed to several high-impact projects at the intersection of AI and medicine, covering topics such as AI-driven medical imaging, synthetic data generation, and multimodal fusion strategies.