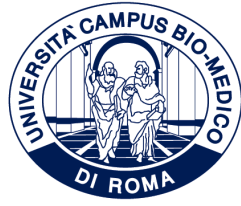


ID N. 8



UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA

DEPARTMENT OF ENGINEERING

UNIVERSITÀ DI CATANIA

DIPARTIMENTO DI INGEGNERIA ELETTRICA

ELETTRONICA E INFORMATICA

Italian National Ph.D. in Artificial Intelligence

Health and Life Sciences

XXXVIII BIS Cycle

Trustworthy Learning in Medical AI

Supervisors

Concetto Spampinato

Federica Proietto Salanitri

Ulas Bagci

Candidate

Rutger Hendrix

May, 2026

Abstract

Clinical AI should meet trustworthiness demands now set out in the GDPR and the EU AI Act: a deployed system should explain its decisions, indicate when those decisions are unreliable, avoid unnecessary centralization of patient data, and permit the removal of an individual’s data influence on request. These demands are usually addressed as an afterthought, recovered from a model once it is already trained, through explanation, repeated inference, or retraining applied after the fact. This thesis investigates a different commitment: that trustworthiness, operationally framed by these four demands, is built into the structure of the system from the start.

Three chapters develop the argument. *Knowing* applies the commitment to interpretability through uncertainty-aware risk classification and gaze-driven prompting of a frozen segmentation model. *Doubting* applies it to uncertainty through federated learning where uncertainty-filtered attention maps replace parameter aggregation as the unit of inter-client exchange. *Forgetting* applies it to unlearning through prompt-removal and passport-embedded LoRA adapters with authority-mediated audit. Across this work, where two demands are met through a shared structural commitment, the combination yields value that neither delivers separately.

Contents

1	Introduction	8
2	Background	12
2.1	Architectures and Adaptation Strategies	12
2.1.1	Vision Transformers	12
2.1.2	Foundation Models and the Segment Anything Model	14
2.1.3	Parameter-Efficient Fine-Tuning	15
2.2	Explainability in Medical Image Analysis	16
2.3	Uncertainty Quantification	19
2.4	Federated Learning	22
2.5	Machine Unlearning	25
2.6	Datasets	27
3	<i>Knowing</i>: Interpretable Medical	
	Image Analysis	29
3.1	Uncertainty-aware Explainable IPMN risk classification	30
3.1.1	Method	31
3.1.2	Experimental Results	37
3.1.3	Discussion	45
3.1.4	Publications	50
3.2	Gaze-Guided Medical Image Segmentation: A	
	Training-Free Approach using SAM Foundation Model	51
3.2.1	Related Work	52
3.2.2	Method	53
3.2.3	Experimental Results	56
3.2.4	Discussion	61
3.2.5	Publications	62

3.3	Predictive Gaze-Driven Prompts for Medical Image Segmentation using Frozen SAM	63
3.3.1	Method	64
3.3.2	Experimental Results	68
3.3.3	Discussion	73
3.3.4	Publications	74
4	<i>Doubling: Uncertainty-Aware Federated Learning</i>	75
4.1	Evidential Federated Learning for Skin Lesion Image Classification	76
4.1.1	Method	77
4.1.2	Experimental Results	81
4.1.3	Discussion	83
4.1.4	Publications	85
5	<i>Forgetting: Instant Machine Unlearning</i>	86
5.1	Pre-Forgettable Models: Prompt Learning as a Native Mechanism for Unlearning	87
5.1.1	Related Work	88
5.1.2	Method	89
5.1.3	Experimental Results	94
5.1.4	Discussion	104
5.1.5	Publications	104
5.2	TrustErase: Auditable Instant Machine Unlearning with Passport-Embedded Representations	105
5.2.1	Related Work	105
5.2.2	Method	107
5.2.3	Experimental Results	113
5.2.4	Discussion	118
5.2.5	Publications	120
6	Conclusion	121
	Appendices	148
A	PreForgettable Models	148
A.1	Inference-Time Uncertainty Estimation: Additional Results	148
A.2	Sensitivity to the LoRA dropout rate	152

B TrustErase	153
B.1 Experimental Setup	153
B.1.1 Training details and configurations	153
B.1.2 Hyperparameter Search for Competitors	154
B.1.3 Hypernetwork Architecture	154
B.2 Parameter Dimensionality and Scaling Behavior	155
B.2.1 Lemma	155
B.2.2 Implications and Scaling Limits	155
B.3 Extensions to Already-Trained Models	156
B.4 Class Unlearning	157
B.5 Instance-Level Unlearning	158
B.6 Backbone Statistics Unlearning	161
B.7 Inference and Training Time Analysis	161
B.8 Reconstruction Attack Analysis	162
B.9 Audit Protocol and Use Cases	163

List of Figures

3.1	Cohort selection flowchart	32
3.2	Representative IPMN categorization cases	33
3.3	Radiomics–deep learning fusion pipeline	35
3.4	IPMN characterization examples	39
3.5	Classification performance by cyst type	41
3.6	Uncertainty thresholding and weighting	44
3.7	Gaze-based prompting pipeline for SAM	54
3.8	Gaze-based segmentation examples	59
3.9	Saliency-prediction pipeline overview	65
3.10	Predicted versus real clinician gaze	71
3.11	Prompt-noise robustness	72
4.1	Overview of FedEvPrompt	78
4.2	Fed-ISIC2019 client distribution	81
5.1	Prompt-based learning and unlearning framework	90
5.2	Accuracy during sequential unlearning	98
5.3	Entropy distribution on DermaMNIST	102
5.4	Accuracy versus coverage on DermaMNIST	103
5.5	Overview of TrustErase	108
5.6	Hypernetwork-based passport composition	111
5.7	Passport perturbation sensitivity	118
A.1	Entropy distribution on BloodMNIST	148
A.2	Accuracy versus coverage on BloodMNIST	149
A.3	Entropy distribution on OrganSMNIST	150
A.4	Accuracy versus coverage on OrganSMNIST	151
B.1	Verification protocol for auditable unlearning	165

List of Tables

3.1	Average cyst size by type and risk	38
3.2	IPMN distribution by region and extent	38
3.3	Accuracy and uncertainty by IPMN characteristics	42
3.4	Gaze-prompt segmentation on Kvasir-SEG	57
3.5	Gaze-prompt segmentation on NCI-ISBI 2013	58
3.6	Comparison with SAM-based models on Kvasir-SEG	60
3.7	Comparison with SAM-based models on NCI-ISBI 2013	60
3.8	Dataset usage overview	68
3.9	Quantitative comparison on NCI-ISBI 2013	69
3.10	Quantitative comparison on CVC-300	70
4.1	Comparison on the skin lesion dataset	83
4.2	Ablation on shared prompts and KD	84
5.1	Unlearning results for $f = 1$	96
5.2	Unlearning results for $f = C/2$	97
5.3	Unlearning results for $f = C - 1$	97
5.4	Ablation on BloodMNIST	99
5.5	AST unlearning on UrbanSound8K	100
5.6	Resistance to membership inference	101
5.7	Impact of LoRA removal	101
5.8	Capability comparison of unlearning methods	106
5.9	Single-class unlearning on CIFAR-10/100	114
5.10	Multi-class unlearning on CIFAR-100	115
5.11	Comparison with pre-forgettable models	116
5.12	Compositional unlearning scalability	117
A.1	AURC at two LoRA dropout rates	152

B.1	Two-class unlearning on CIFAR-20	157
B.2	Pre-forgettable comparison on CIFAR-20	158
B.3	Single-class unlearning with VGG-16	158
B.4	Instance-level unlearning with canaries, CIFAR-10, 4,500 forgotten images (10%), C-LoRA rank 64 (3 seeds; dist mean $\pm$ std).	160
B.5	Unlearning with backbone statistics	161

Chapter 1

Introduction

In the second half of the past decade, Artificial Intelligence (AI) attained or exceeded expert-level performance across a broad range of medical imaging benchmarks, including diabetic retinopathy grading and disease classification in radiology and dermatology [36, 111, 129]. Clinical adoption, however, did not keep pace with these advances. Reliability, interpretability, and regulatory oversight were articulated as central barriers to deployment [159, 79, 171], and these concerns remain operative to this day. Meanwhile, the regulatory landscape has matured substantially: by 2025, the FDA’s published registry of AI-enabled medical devices comprised more than 1,000 entries [161, 170]. Nevertheless, a substantial gap persists between regulatory authorisation and routine clinical integration. Recent survey evidence indicates that use of AI by physician has increased substantially over the past three years, yet such adoption remains concentrated in workflow-support and administrative functions rather than core diagnostic reasoning. Patient privacy stands out as the only area where more physicians expect AI to cause harm than to help [5].

The legislative response to this gap has crystallised over the last decade into two instruments that together define operational requirements for any AI system processing personal data in the European Union: the General Data Protection Regulation (GDPR, 2016) [37] and the EU AI Act (2024) [38]. It is tempting to read these primarily as data-protection instruments, but for medical AI they function more broadly as an *accountability framework* [74]. The GDPR’s framework for automated decision-making (Art. 22), data minimization and privacy-by-design principles (Art. 5 and Art. 25), and the right to erasure (Art. 17), together with the EU AI Act’s classification of many medical AI systems as high-risk, and its requirement for human oversight (Art. 14), define four demands a deployed medical AI system should meet: it should explain its decisions, indicate when those decisions are unreliable, avoid unnecessary centralization of patient data, and permit the removal of an individual’s data influence on request. Each of these demands corresponds to an active research problem

in machine learning. The contributions of this thesis address all four.

Explainable AI (XAI) has predominantly developed through *post-hoc* attribution methods such as SHAP, LIME, and Grad-CAM [103, 132, 142], which generate explanations as a secondary estimation rather than as part of the model’s own computation, and therefore may not faithfully reflect the underlying model’s reasoning. Furthermore, the additional inference-time cost constrains their use in clinical workflows [136]. An alternative line of work designs interpretability into the model itself, so that the basis of a decision is a structural property of the architecture rather than a downstream reconstruction. For example, concept-bottleneck networks route predictions through human-interpretable concepts at an intermediate layer [85]. Another approach supervises or prompts the model with clinician attention, so that its reasoning is, by construction, grounded in regions a clinician would inspect. This thesis adapts the latter.

Data minimization and privacy-by-design provide the principled foundation for federated learning [112, 133], a collaborative training paradigm in which institutions exchange model updates rather than data. The third demand is therefore a design consequence rather than a statutory obligation: the GDPR does not mandate decentralized training, it mandates minimization (Art. 5) and data protection by design and by default (Art. 25). Centralized training moves identifiable records out of the controlling institution and holds them in one place, which increases the personal data processed relative to the purpose, concentrates the impact of a single breach, and complicates Art. 17 erasure because the aggregator holds a second copy whose influence on the trained model must be traced separately. Keeping raw data at the site and sharing only derived signals avoids all three. Yet, what is shared still matters: gradients and parameters leak information about training data to an adversary [43], while offering nothing a clinician could meaningfully inspect. This thesis explores client-side representations beyond parameter sharing, using visual representations as a structurally different knowledge transfer unit.

Meaningful human oversight requires that the model can signal when its predictions are unreliable. A model that is always equally confident provides no basis for deferral to an expert, and a confident wrong answer is more dangerous than no answer at all. Uncertainty quantification addresses this need [9].

Finally, the right to erasure has motivated the field of machine unlearning [16, 116], which seeks to remove specific data influence from a model without full retraining. Existing methods are typically iterative, depend on access to training data, and provide no mechanism by which an external party can confirm that erasure has occurred. This thesis takes up these three gaps directly: the speed, data-independence, and auditability of forgetting.

Crucially, the four demands are not independent. As a thought experiment, one can trace

a chain of relationships between them. Interpretability can inform uncertainty: a clinician can only judge whether a prediction should be trusted if both the reasoning process and the model’s confidence are visible, enabling appropriate deferral when the model is uncertain. Uncertainty, in turn, supports safe collaboration by identifying when local knowledge is insufficient and when external information should contribute to learning. Collaborative learning then raises questions of data ownership and control, particularly when model behavior continues to depend on information contributed by individual institutions or patients. This naturally leads to the problem of machine unlearning and the right to be forgotten.

Knowing, Doubting, Forgetting

The four demands (interpretability, uncertainty, federated learning and unlearning) provide the regulatory and clinical motivation for this thesis. The contribution is the investigation of a design commitment: trustworthiness, operationally framed by these four regulatory demands, is implemented as a structural property of the system from the start, rather than recovered as an afterthought. This commitment operates at two scales. At the pipeline level, the relevant signals are produced together within the deployed clinical workflow, so that explanation and uncertainty arrive alongside the prediction rather than being assembled from separate analyses. At the model level, the same signals are a structural feature of the architecture itself, produced as part of the same forward pass as the prediction. The thesis does not claim that post-hoc methods cannot deliver these demands in principle; it claims that committing to them by design unlocks capabilities that the post-hoc route does not, and each chapter exhibits one such capability.

A consequence of this commitment is that the same architectural surface can serve more than one demand at once, as demonstrated in Sections 3.1, 4.1, and 5.1. This recurrence also shapes how the thesis is structured: the chapters are organized by the dominant role each contribution plays rather than by the mechanism it employs. The three chapters correspond to three operations a clinical system should be able to do with its own output: *knowing* what it sees, *doubting* what it concludes, and *forgetting* on demand.

Chapter 2. Background. This chapter consolidates the technical background shared across the contributions, providing architecture descriptions, problem formulations for relevant research areas, a taxonomy of methodologies within the field, metrics and datasets used throughout the thesis.

Chapter 3. *Knowing: Interpretable Medical Image Analysis.* This chapter applies the design commitment to interpretability at both scales identified above. Section 3.1 contributes an uncertainty-aware framework for Intraductal papillary mucinous neoplasms (IPMN) risk classification in which XAI and uncertainty quantification are produced together in the deployed pipeline, supporting clinician deferral on cases where the model is unreliable. Sections 3.2 and 3.3 move the commitment into the model itself, contributing a gaze-driven prompting framework for a frozen segmentation foundation model in which the prompt is the explanation by construction. Section 3.2 uses recorded fixations directly without training, while Section 3.3 trains a saliency predictor on recorded clinician gaze, yielding a fully automated pipeline that runs without eye-tracking hardware at inference.

Chapter 4. *Doubting: Uncertainty-Aware Federated Learning.* This chapter applies the design commitment to uncertainty. Section 4.1 contributes FedEvPrompt, an evidential federated learning framework in which uncertainty governs inter-client contribution, and where attention maps replace parameter aggregation as the unit of knowledge transfer.

Chapter 5. *Forgetting: Instant Machine Unlearning.* This chapter applies the design commitment to forgetting, where knowledge to be forgotten is pre-committed and held in a small, modular component on top of a frozen pre-trained backbone, so that unlearning reduces to removing or swapping that component. Section 5.1 contributes Pre-Forgettable Models, in which class-wise unlearning is reduced to the removal of a learnable prompt, making forgetting instant and data-free. Section 5.2 contributes TrustErase, which holds class-wise unlearning in passport-embedded LoRA adapters, where passport deactivation performs the unlearning, and an authority-mediated audit protocol allows a trusted verifier to confirm that the deployed adapter implements the claimed unlearning configuration.

Chapter 2

Background

This chapter consolidates the technical foundations on which the contributions of this thesis are built. It is organized into two parts. Section 2.1 covers the architectural and adaptation strategy that recur across chapters. Sections 2.2–2.5 then introduce the four demands around which the thesis pillars are organized: explainability, uncertainty quantification, federated learning, and machine unlearning. Each presents its clinical motivation, the foundational methods and families of state-of-the-art approaches, and the evaluation metrics used throughout the corresponding chapters. Paper-specific positioning is deferred to the Related Work section of each contribution. The datasets used in this work are described in the experimental settings at the end of the chapter (Section 2.6)

2.1 Architectures and Adaptation Strategies

The predominant architectural choice in this thesis is a transformer-based backbone, adapted to a downstream (medical) task without modifying its core weights. This section introduces the three components on which that strategy depends: the Vision Transformer (Section 2.1.1) as the underlying architecture, the Segment Anything Model (Section 2.1.2) as the foundation model used for segmentation, and parameter-efficient fine-tuning (Section 2.1.3) as the family of adaptation methods through which task knowledge is added on top of a frozen backbone.

2.1.1 Vision Transformers

For over a decade, Convolutional Neural Networks (CNNs) such as ResNet [59] and U-Net [135] have been the default backbone for classification and segmentation in medical image analysis. Their inductive biases of locality and translation equivariance capture spatial structure efficiently but confine each layer’s receptive field to a local neighborhood. Long-

range spatial dependencies are therefore recovered only through deep stacks of convolutions, introducing depth-related optimization difficulties. The Vision Transformer (ViT) [34] addresses this limitation by replacing convolution with self-attention [163], which computes pairwise relationships across the entire input in a single operation.

A ViT processes an image $I \in \mathbb{R}^{H \times W \times 3}$ by partitioning it into a sequence of N non-overlapping patches of size $P \times P$, each linearly projected into a d -dimensional embedding. A learnable classification token [CLS] is prepended to the sequence, and positional encodings are added to preserve spatial order. The resulting token sequence $\mathbf{z}_0 \in \mathbb{R}^{(N+1) \times d}$ is passed through L identical transformer blocks, each composed of a multi-head self-attention (MSA) layer and a feed-forward network with residual connections and layer normalization.

At the core of each block, self-attention computes, for every token, a weighted sum of all other tokens, where the weights are derived from learned query, key, and value projections:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V. \quad (2.1)$$

Multi-head attention runs h such operations in parallel on different projections and concatenates the results. Throughout this thesis, d denotes the embedding dimension, h the number of heads, and L the number of transformer blocks. At every block, the [CLS] token attends to all patch tokens and aggregates their information into a single embedding; after the final block, this embedding acts as a global summary of the input and is used as the image-level representation for classification.

The ViT architecture supports two mechanisms not naturally available in CNNs. First, attention rollout [1] combines the per-layer attention scores into a single spatial map indicating which input regions contribute most to a given output token. This map serves both as an interpretability signal (Section 2.2) and as a compact, parameter-free representation that can be shared across clients in federated learning (Chapter 4). Second, the input is a token sequence of variable length rather than a fixed grid, so that small numbers of learnable tokens can be prepended and trained while the backbone remains frozen; this is a particular instance of the broader family of parameter-efficient fine-tuning strategies introduced in Section 2.1.3. Combined with the strong transfer behavior of ViTs under large-scale pre-training [128, 21, 83], these two properties motivate the use of pretrained ViTs as a frozen foundation-model backbone in the low-data medical imaging settings considered in this thesis.

2.1.2 Foundation Models and the Segment Anything Model

The term foundation model [15] refers to a model pretrained on data at sufficient scale that it can be adapted to a wide range of downstream tasks through mechanisms ranging from full fine-tuning to lightweight alternatives such as prompting, linear probing, or small adapter modules, requiring progressively less task-specific supervision and, in the limit, enabling zero-shot or few-shot generalization to new tasks. This is particularly attractive in medical imaging, where labeled data is scarce and expert annotation expensive.

The Segment Anything Model (SAM) [83] is the foundation model used in the segmentation contributions of this thesis. It was trained on SA-1B, a dataset of over one billion masks across eleven million natural images, and provides *promptable segmentation*: given an image and a user-specified prompt, the model returns a binary mask for the indicated region. Its architecture consists of three components:

- **Image encoder.** A Vision Transformer pretrained with masked autoencoding [58], available in base, large, and huge variants, that maps the input image to a dense embedding. The encoder is computed once per image and reused across all prompts for that image.
- **Prompt encoder.** A module that maps user-specified prompts to embeddings in the same space as the image features. Sparse prompts (points or bounding boxes) are represented as positional encodings combined with learned type embeddings; dense prompts (low-resolution masks) are encoded by convolutional layers and summed into the image embedding.
- **Mask decoder.** A lightweight transformer decoder that applies bidirectional cross-attention between the image and prompt embeddings, followed by an MLP head that produces a per-pixel mask probability.

Despite its strong zero-shot performance on natural images, SAM exhibits a substantial domain gap when applied directly to medical images [110]: low contrast, ambiguous boundaries, and unfamiliar anatomical semantics degrade segmentation quality relative to natural-image benchmarks. However, the clean separation of SAM into encoder, prompt encoder, and decoder allows for many modular improvements, as reflected in the literature: full fine-tuning of the encoder and decoder on curated medical corpora [105]; parameter-efficient adapters and trainable modules inserted into the frozen encoder [27, 99, 186], including volumetric variants for 3D data [25, 50]; replacement of the prompt encoder by a learned module that produces prompts in latent space [146]; and prompt engineering, in which all weights are frozen and adaptation is performed on the input side [166, 167]. Chapter 3 takes the

prompt-engineering route, with prompts derived from clinicians’ eye-gaze, recorded directly and approximated by a learned predictor. The interpretability of this signal is discussed in Section 2.2.

2.1.3 Parameter-Efficient Fine-Tuning

As foundation-model backbones approach the billion-parameter scale, full fine-tuning on downstream tasks has become impractical: it is costly in compute and memory, and demands a separate model per task. Parameter-efficient fine-tuning (PEFT) keeps the backbone frozen and adapts it through a small number of additional learnable parameters, two to three orders of magnitude fewer than the backbone itself. PEFT methods can be classified by the type of operation they apply to the frozen backbone [56, 168]. *Additive* methods introduce new trainable modules or parameters into the network, the most common being adapter layers [64] and prompt-based methods that prepend learnable tokens to the input sequence [90, 76]. *Subtractive* methods make a subset of the existing backbone parameters trainable while keeping the rest frozen, such as the bias terms [183]. *Reparameterization* methods construct a low-dimensional reparameterization of the original weights for training and transform it back for inference, with the well-known instance LoRA [66]. Hybrid methods combine elements from the other three families. In this thesis, prompt tuning and LoRA are used in Sections 4.1, 5.1, and 5.2, and the following sections describe both in detail.

Prompt tuning. In image classification, prompt tuning trains only a small set of prepended token embeddings and a lightweight classifier head, keeping the backbone frozen. The prompts provide task-specific context that the frozen self-attention layers attend to at every forward pass, aligning the input distribution with the pretrained backbone and biasing the attention patterns through an additional learned signal, rather than modifying the backbone itself [76, 125]. The application is divided into *shallow* and *deep* placement [76], where deeper insertion provides finer granularity at the cost of additional learnable parameters. Under deep placement, each multi-head self-attention block operates on the augmented key and value sequences:

$$\text{MSA}\left(h_Q, [pr_k^{(i)}; h_K], [pr_v^{(i)}; h_V]\right), \quad (2.2)$$

where h_Q, h_K, h_V are the queries, keys, and values produced by the previous layer and $pr_k^{(i)}, pr_v^{(i)}$ are the learnable prompts inserted at layer i . Because distinct prompts can be associated with distinct semantic units, the mechanism pairs naturally with class- or task-conditional learning: Section 4.1 exploits this to share task-conditioned attention maps across federated clients, and Section 5.1 makes the prompt itself the unit of forgetting.

Low-Rank Adaptation (LoRA). LoRA adapts a frozen weight matrix $W_0 \in \mathbb{R}^{d \times k}$ by representing its task-specific update as a low-rank product:

$$\Delta W = BA, \quad B \in \mathbb{R}^{d \times r}, \quad A \in \mathbb{R}^{r \times k}, \quad r \ll \min(d, k). \quad (2.3)$$

Only A and B are trained; W_0 remains frozen. At inference, the adapted weight is $W_0 + BA$, which can be kept as two factors (preserving the frozen base) or merged into a single matrix. Confining the task-specific knowledge to a low-rank subspace drops the number of trainable parameters by orders of magnitude. The update is also structurally separated from the backbone, so it can be inspected, modulated, or removed without altering W_0 . Section 5.2 builds on this separation to construct verifiable unlearning mechanisms by inserting fixed, task-specific matrices between the LoRA factors.

2.2 Explainability in Medical Image Analysis

Problem and clinical motivation

Adoption of deep learning in clinical workflows depends on the transparency of the reasoning behind a diagnostic AI model’s predictions. Explainable AI (XAI) addresses this by exposing the features, regions, or examples that drive a prediction. A common distinction in explainable AI is between *global* explanations, which describe patterns learned across an entire population, and *local* explanations, which explain an individual prediction for a specific input. Because medical image analysis often requires case-level justification for a particular diagnosis or segmentation, this thesis primarily focuses on local explanation methods.

Families of Local Explainability Approaches

Attribution methods estimate the contribution of each input feature to the output by perturbing inputs and measuring the change. LIME [132] does this locally: for a given prediction, it perturbs the input, observes the resulting changes in the model output, and fits a simple linear surrogate that approximates the model’s behavior in a small neighborhood of the input; the coefficients of this surrogate are reported as the explanation. LIME is model-agnostic and case-specific, but because each explanation is computed from independent perturbations, it offers no consistency guarantee across samples. SHAP [103] grounds attribution in cooperative game theory, assigning each feature a Shapley value that represents its average marginal contribution to the prediction across all possible feature subsets; the values are globally consistent across samples and admit both local and dataset-level

views. A complementary line of attribution methods propagates gradients back to pixel space rather than perturbing the input, including Integrated Gradients [150] and layer-wise relevance propagation [6].

Activation-based methods produce spatial heatmaps from intermediate feature maps. The dominant instance is class-activation mapping in its various forms; Grad-CAM [142] computes the gradient of the predicted class score with respect to the activations of a late feature map, uses these gradients as channel-wise importance weights, and aggregates the weighted feature maps into a coarse heatmap over the input. The map indicates *where* in the image the model attended most strongly, and is the workhorse of CNN-based medical-imaging explainability. Attention rollout [1] extends the same idea to transformers. Instead of reading attention weights from a single layer, attention rollout recursively composes the attention matrices across layers by multiplying the residual-augmented attention matrices, typically after averaging attention weights across heads. The resulting matrix captures the cumulative flow of information from input tokens to deeper representations; the row corresponding to the [CLS] token is then interpreted as a spatial attribution map indicating which input regions contribute most strongly to the final global representation. Attention rollout can be used both as an interpretability signal and as a compact shareable representation across federated clients (Section 4.1).

Example-based methods explain a prediction by retrieving training instances the model treats as similar to the input, providing case-based reasoning that domain experts find intuitive [24].

A persistent concern with these post-hoc methods is faithfulness: Adebayo et al. [3] showed that several widely used methods fail basic sanity checks, producing similar explanations for trained and randomly initialised networks. A saliency map that looks anatomically plausible is therefore not evidence that the model is reasoning correctly, which matters in clinical settings where the explanation may be trusted as such.

These limitations motivate a second paradigm: rather than recovering an explanation after training, one builds interpretability directly into the prediction process itself. Concept Bottleneck Models [85] force every prediction to pass through a layer of human-interpretable concepts, exposing a fixed set of clinically meaningful features as the model’s reasoning. Prototype-based classifiers [24] learn a dictionary of training-image fragments and classify by similarity to them, making the comparison itself the explanation. A more recent line of work incorporates human attention into the model’s decision process. A clinician’s gaze, recorded by an eye-tracker during image interpretation, captures the fixation points and times at which the expert concentrates. These gaze distributions can be incorporated either during training, by encouraging model saliency to align with expert fixations [17, 162], or during

inference, where gaze acts as an external prompt that guides the model toward clinically relevant regions. In both settings, human attention becomes part of the prediction process itself, either by constraining model attention toward clinically meaningful regions or by making the reasoning process externally controllable and observable.

Although gaze is not a transcript of expert reasoning, it is indicative of regions where clinical features tend to live, and what it does for interpretability in this case is counterfactual. Human attention is rarely drawn to the spurious features a model would otherwise exploit [104], so constraining the model to attend within gaze regions excludes the regions where shortcut learning typically arises. What matters for clinical interpretability is that the model uses the right information for its predictions, and constraining attention to expert gaze regions makes this a structural property rather than something to be verified post hoc. The remaining trade-off is hardware. Physical eye-trackers must be present at acquisition and require calibration, which limits scalability and confines gaze-guided methods to settings where the hardware is already in place.

The field has known unresolved tensions around faithfulness, method disagreement, and the cost of interpretability [136]. The chapters that follow do not claim to settle them; they take positions on which trade-offs are acceptable in the clinical settings considered.

Metrics

Quantitative evaluation of post-hoc attribution methods is intrinsically difficult: there is no ground truth for the right explanation, and the standard faithfulness metrics that have been proposed in the literature are unstable across hyperparameters and datasets [158, 63]. This thesis evaluates explanations indirectly, through their relationship to other measurable properties. In Section 3.1, the agreement between explanation strength and prediction confidence is quantified by Spearman and Pearson correlations between total absolute SHAP and LIME values and predictive uncertainty. In Sections 3.2–3.3, the interpretability signal (clinician gaze, recorded or predicted) is converted into a SAM prompt and analyzed qualitatively and in relation to the segmentation it produces.

The two segmentation metrics reported in both sections are the **Dice Similarity Coefficient (DSC)**,

$$\text{DSC}(P, G) = \frac{2|P \cap G|}{|P| + |G|}, \quad (2.4)$$

and the **Intersection over Union (IoU)**,

$$\text{IoU}(P, G) = \frac{|P \cap G|}{|P \cup G|}, \quad (2.5)$$

where P is the predicted mask and G the ground-truth mask. Both range from 0 (no overlap) to 1 (perfect overlap); DSC is more sensitive at low overlap and IoU at high overlap, and both are reported for comparability with prior work. For volumetric data, metrics are computed slice-by-slice and averaged at the volume level.

2.3 Uncertainty Quantification

Problem and clinical motivation

A clinical decision system that is equally confident in all predictions is of limited practical value. Uncertainty quantification (UQ) addresses this by estimating the reliability of each prediction, which can then be used to defer uncertain cases to a clinician.

Predictive uncertainty is typically decomposed into *aleatoric* uncertainty, arising from inherent ambiguity or noise in the data, and *epistemic* uncertainty, arising from limited model knowledge or insufficient training coverage. Unlike aleatoric uncertainty, epistemic uncertainty can in principle be reduced with additional data [80].

Foundational methods

Three families of UQ methods are used in this thesis: Monte Carlo dropout, ensembles, and evidential deep learning. The first two are post-hoc wrappers around standard training; the third makes uncertainty a direct output of the model.

Monte Carlo (MC) dropout. MC dropout [42] reinterprets dropout from a regularization technique applied only during training, to an approximate Bayesian inference procedure. By keeping dropout active at inference and performing T stochastic forward passes for the same input, the resulting distribution of predictions approximates a posterior over the network’s outputs. Aggregating the softmax outputs $p_t \in \mathbb{R}^C$ across the T passes yields the mean predictive probability

$$\hat{p} = \frac{1}{T} \sum_{t=1}^T p_t. \quad (2.6)$$

The total predictive uncertainty is the Shannon entropy of this mean prediction,

$$H(\hat{p}) = - \sum_{c=1}^C \hat{p}_c \log(\hat{p}_c + \varepsilon), \quad (2.7)$$

with ε a small constant for numerical stability. The aleatoric component is estimated as the expected entropy across the T samples,

$$\mathbb{E}[H(p)] = \frac{1}{T} \sum_{t=1}^T \left(- \sum_{c=1}^C p_{t,c} \log(p_{t,c} + \varepsilon) \right), \quad (2.8)$$

and the *epistemic* component as the mutual information between the predictions and the model posterior,

$$I = H(\hat{p}) - \mathbb{E}[H(p)]. \quad (2.9)$$

To compare uncertainties across tasks with different numbers of classes, each measure is normalized by $\log(C)$. MC dropout is widely used in medical imaging because it is conceptually simple and requires no architectural change, but it incurs T -fold inference cost and the dropout-as-Bayesian-posterior interpretation has been shown to rely on assumptions that hold only approximately in practice [119, 65].

Ensembles. Ensembles train M independently initialised models on the same task and aggregate their predictions, and their disagreement serves as an estimate of epistemic uncertainty. Deep ensembles [88] often outperform single-network Bayesian methods on out-of-distribution and calibration benchmarks, but they multiply training and memory costs by M , which is prohibitive for resource-constrained clinical hardware and federated clients. Classical tree ensembles such as random forests [18] produce the same disagreement signal at no additional cost, since a single forward pass already involves M trees whose per-tree predictions are aggregated to form the output. The uncertainty decomposition introduced for MC dropout applies in both cases, with the T samples replaced by M ensemble members.

Evidential Deep Learning (EDL). EDL [144] departs from MC dropout and ensemble approaches by making uncertainty a direct output of a single deterministic network, computed in one forward pass. The motivation is that softmax outputs, the standard classification head, can produce overconfident predictions even on samples far from the training distribution [149]. EDL replaces the softmax with parameters of a Dirichlet distribution over class probabilities, a second-order distribution that represents uncertainty about the predicted probability itself.

The framework draws on Dempster–Shafer evidence theory, which relaxes the additivity principle of standard probability and allows the sum of class beliefs to be less than one, with the remainder attributed to uncertainty. For a frame of K mutually exclusive class labels, each class k is assigned a belief mass $b_k \geq 0$ and an overall uncertainty mass $u \geq 0$,

with $u + \sum_{k=1}^K b_k = 1$. The belief mass for class k is determined by the amount of *evidence* $e_k \geq 0$ supporting that class, and uncertainty is inversely proportional to the total evidence, taking the value 1 in the limit of no evidence. A belief assignment corresponds to a Dirichlet distribution with parameters $\alpha_k = e_k + 1$, motivated by the role of the Dirichlet as conjugate prior to the categorical distribution:

$$\text{Dir}(\mathbf{p}, \boldsymbol{\alpha}) = \frac{\Gamma(S)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K p_k^{\alpha_k-1}, \quad \alpha_k > 0, \quad (2.10)$$

where $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_K]$ and $S = \sum_{k=1}^K \alpha_k$ is termed the Dirichlet strength. From these parameters the per-class belief and uncertainty are recovered as

$$b_k = \frac{\alpha_k - 1}{S}, \quad u = \frac{K}{S}, \quad (2.11)$$

and the expected class probability, the mean of the Dirichlet, is

$$\hat{p}_k = \frac{\alpha_k}{S}. \quad (2.12)$$

In practice, the network is trained to output non-negative evidences e_k via a ReLU (or similar) activation on the logits, from which the Dirichlet parameters and uncertainty follow analytically. Training proceeds by minimizing an evidential mean-squared-error loss that integrates over the Dirichlet's probability mass. For a sample i with one-hot label $\mathbf{y}_i$ and predicted probabilities $\mathbf{p}_i$, the loss begins as

$$\mathcal{L}_i(\boldsymbol{\Theta}) = \int [(\mathbf{y}_i - \mathbf{p}_i)^\top (\mathbf{y}_i - \mathbf{p}_i)] \text{Dir}(\mathbf{p}_i, \boldsymbol{\alpha}_i) d\mathbf{p}_i, \quad (2.13)$$

and using the identity $\mathbb{E}[p_{i,k}^2] = \mathbb{E}[p_{i,k}]^2 + \text{Var}(p_{i,k})$ admits the closed form [144]

$$\mathcal{L}_i(\boldsymbol{\Theta}) = \sum_{k=1}^K \left(y_{i,k} - \frac{\alpha_{i,k}}{S_i} \right)^2 + \frac{\alpha_{i,k}(S_i - \alpha_{i,k})}{S_i^2(S_i + 1)}. \quad (2.14)$$

The two terms penalize the squared error of the expected probability and the variance of the prediction respectively.

To prevent the network from accumulating misleading evidence for incorrect classes, a Kullback–Leibler regularization term shrinks the evidence for non-target classes toward a uniform Dirichlet prior:

$$\mathcal{L}_{KL} = \text{KL} \left[\text{Dir}(\mathbf{p}_i \mid \tilde{\boldsymbol{\alpha}}_i) \parallel \text{Dir}(\mathbf{p}_i \mid \mathbf{1}) \right], \quad (2.15)$$

where $\tilde{\alpha}_i = \mathbf{y}_i + (\mathbf{1} - \mathbf{y}_i) \odot \alpha_i$ are the Dirichlet parameters after the removal of non-misleading evidence. The total evidential loss used throughout Section 4.1 is

$$\mathcal{L}_\varepsilon = \mathcal{L}_i + \mathcal{L}_{KL}. \quad (2.16)$$

EDL is cheap at inference and provides uncertainty as an explicit network output rather than a derived statistic, which makes it well suited to resource-constrained federated clients. However, the evidential loss is more difficult to optimize than standard cross-entropy and more sensitive to training hyperparameters, which can be observed as reduced predictive performance relative to a cross-entropy baseline.

Metrics

Uncertainty quality is assessed both intrinsically and through downstream performance. *Calibration* is measured by the Brier score, the squared error between predicted probabilities and observed outcomes (lower is better). *Clinical utility* is captured by the coverage–accuracy curve from selective prediction: given an uncertainty threshold, the model commits to a prediction whenever its uncertainty falls below that threshold and defers otherwise; the trade-off between coverage (the fraction of cases committed) and selective accuracy (the accuracy on committed cases) characterizes how reliability scales with autonomy. The area under this curve, AURC, summarizes selective-prediction quality across thresholds, with higher values indicating a more useful uncertainty signal. *Agreement with explanation strength*, the extent to which higher uncertainty corresponds to weaker post-hoc attribution, is quantified by Spearman and Pearson correlation in Section 3.1.

Across this thesis, uncertainty is treated not as a diagnostic afterthought but as an operational signal: the basis for routing cases to a clinician (Section 3.1), for filtering which representations are worth sharing in a federation (Section 4.1), and as both a training-time mechanism and an inference-time signal in machine unlearning (Section 5.1).

2.4 Federated Learning

Problem and structural motivation

Medical data is structurally distributed over hospital archives across countries and jurisdictions; patient records are protected by national data-protection law and inter-institutional transfer of identifiable data often is legally impermissible. Centralizing raw patient images for model training is therefore an action of last resort. Federated learning (FL) [112] ad-

addresses this by training a shared global model across a set of clients (typically hospitals or imaging centers), exchanging only model updates rather than raw records, with each client keeping its data local.

However, this locality alone does not eliminate the challenges of distributed training. Medical data is heterogeneous across sites along nearly every axis: scanner vendors, patient demographics, class distributions tied to referral patterns, and label conventions all differ between institutions. The resulting violation of the independent identical distribution assumption causes *client drift*, where local updates pull the global model in conflicting directions and degrade the performance of the aggregated model on each client. Communication is similarly nontrivial: transmitting full backbone updates across rounds dominates the bandwidth budget of standard FL, and the cost scales with model size, which becomes binding as foundation-model backbones approach the billion-parameter scale (Section 2.1.3). Privacy, finally, is not guaranteed by data locality alone: gradient inversion attacks [194] show that the parameter updates exchanged in standard FL can, under specific conditions, be used to approximately reconstruct the training images that produced them. Sharing parameters is therefore not a neutral choice; it leaks information about the underlying data, and offers nothing a clinician can meaningfully inspect.

Foundational protocol and aggregation taxonomy

The standard FL protocol is FedAvg [112]. At each communication round t , the server broadcasts the current global model parameters $\theta^{(t)}$ to a subset of clients. Each selected client k performs a fixed number of local training epochs on its private dataset $\mathcal{D}_k$, producing updated parameters $\theta_k^{(t+1)}$ which then in turn are sent back to the server. The server aggregates the client updates by a weighted average proportional to local dataset size,

$$\theta^{(t+1)} = \sum_k \frac{|\mathcal{D}_k|}{\sum_j |\mathcal{D}_j|} \theta_k^{(t+1)}, \quad (2.17)$$

and broadcasts the result for the next round.

Subsequent work has diverged along what clients exchange and how the server combines it. The simplest extensions retain FedAvg’s parameter-exchange structure but constrain how far local updates can move from the global model, either through proximal regularization in the local objective [95] or through control-variate corrections that estimate and subtract the drift direction at each round [78]. A second family partitions the network into a globally shared component and a client-specific component, typically aggregating the feature extractor globally while training the classification head locally [94, 30]; this re-

covers most of the global representational benefit while letting each client specialize to its own data. A third family changes how the server weights contributions: instead of weighting by dataset size, contributions are weighted by client similarity, validation performance, or attention over client embeddings, producing per-client aggregates rather than a single shared model [71, 184]. Beyond parameter sharing, knowledge-distillation aggregation replaces weights with soft predictions over a public reference set [91, 98]: the shared object is a low-dimensional probability distribution that any client architecture can produce, which shrinks the parameter-leakage surface. Smaller still, prototype-based aggregation exchanges only class-mean feature embeddings [154], with each client training to match the federated prototype rather than reconstruct a full output distribution. Section 4.1 pushes this direction further by sharing attention maps over each client’s local data: each map serves simultaneously as a task representation, a distillation target, and an interpretable spatial output.

Metrics

Federated systems are evaluated along several axes that no single metric captures. *Per-client and average classification performance* (typically balanced accuracy for class-imbalanced medical data) measures whether the federation actually helps each participant. *Communication cost* (number of rounds, payload size per round) constrains feasibility in real deployments. *Privacy*, if at all, is most commonly evaluated empirically through membership inference attacks (Section 2.5), which estimate whether an adversary observing the shared signal can determine which samples were used to train a given client. What such an attack establishes as harmful is worth stating. One position is that recovering a medical image matters little if it cannot be linked to an identity. Unlinkability, however, is not a stable property of a release. The Netflix Prize de-anonymization [114] showed that a dataset considered safe in isolation could be re-identified by cross-referencing a second, public source, so the threat model was wrong about what the adversary already held rather than about the release itself. Medical images admit the same structure through acquisition metadata, rare-phenotype uniqueness, surface reconstruction from head CT or MR, and repeated releases from the same site. Since a controller cannot enumerate the auxiliary information that will exist later, this thesis treats recovery of the training sample itself as the harm rather than arguing that a recovered sample is harmless while it remains unlinked.

2.5 Machine Unlearning

Problem and regulatory motivation

The ‘right to erasure’ mandates that an individual can request to remove their personal data. In extension, the argument holds that the compressed representation of that data in a AI system is covered as well, terming any ‘influence’ of that data on any trained model as in-scope. Machine unlearning [20] is the field that develops mechanisms by which a trained model can be made to behave as if those records had never been seen.

The gold standard is to retrain the model on the data with the forget set removed. However, it requires that the full training set be retained indefinitely, which itself runs counter to GDPR’s data-minimization principle. Retraining times also run in the order of days for modern backbones, making it an unrealistic response to an erasure request in settings that must remain continuously available, such as diagnostic tools and triage queues. This gap motivates approximate unlearning, methods that aim to match retraining behavior at a fraction of the cost. Formally, given a model f_θ trained on a dataset $\mathcal{D}$ and a forget set $\mathcal{D}_f \subset \mathcal{D}$, unlearning produces a model $f_{\theta'}$ that is, by some criterion, indistinguishable from a model retrained on $\mathcal{D} \setminus \mathcal{D}_f$, ideally without access to $\mathcal{D}_f$ itself. The unit of forgetting can vary, ranging from individual samples to semantic classes, distributional shifts, or arbitrary subsets specified at request time.

These units are not equivalent from a regulatory standpoint. Art. 17 grants the right to the data subject, so the compliance-relevant unit is per-subject: every sample belonging to one patient, which may span classes, sites, and modalities. Per-sample forgetting is finer than this unit, class-level forgetting is coarser and generally not aligned with it, and distributional or subset-level forgetting corresponds to governance actions rather than to individual rights. Sections 5.1 and 5.2 operate at the class level, with instance-level forgetting supported through bucketed partitions rather than per-subject allocation.

Unlearning strategies

Exact methods structure training to make deletion cheap, partitioning the dataset and training an independent module on each partition. An erasure request then reduces to retraining the corresponding module [16], lowering the per-request cost. However, the architectural constraint of separate modules with global aggregation falls short of the representational capacity that standard non-partitioned models have, resulting in reduced predictive performance. *Approximate* methods give up parameter equivalence with the retrained model and instead target behavioral indistinguishability on the forget set.

Approximate methods can be organized by what the unlearning operation modifies. *Data modification* methods retrain the model on a perturbed version of the dataset in which the forget set has been relabeled, so that the model’s response to the forget set is decoupled from their original class [52]. *Loss-surface manipulation* methods instead reverse the optimization direction on the forget set, by gradient ascent on the forget set loss or by pushing the forget set across decision boundaries through targeted perturbation [155, 26]. *Weight-targeted* methods first identify the parameters carrying the forget set’s information, using Fisher-information [41], gradient saliency [39], or influence functions [84, 53] as the identification mechanism, and then dampen or update them accordingly. *Distillation-based* methods supervise the unlearned model against teacher signals matching retained classes and deviating on forgotten ones [28, 87].

Metrics

Unlearning is evaluated jointly along forgetting effectiveness and retention. The dominant empirical test is the membership inference attack (MIA) [148], in which a binary classifier is trained on the target model’s outputs to distinguish samples that were in the training set from samples that were not. To evaluate unlearning, the classifier is trained on outputs from the retain set (members) and the test set (non-members), and then applied to the forget set. Successful unlearning produces a model on which the classifier finds no features in the forget set’s outputs that indicate training-set membership, so the forget set is classified as non-member. MIA is a behavioral test under a specific adversary and probe set, and a negative result is consistent with successful forgetting without proving it.

The metrics reported throughout Sections 5.1–5.2 are:

- **Retain accuracy** Acc_r : classification accuracy on retained classes, higher is better.
- **Forget accuracy** Acc_f : classification accuracy on forgotten classes, ideally at 0 or chance level ($1/K$ for K classes), depending on the objective.
- **MIA success rate**: the fraction of forget-set samples the attack classifier still identifies as members, with lower values indicating a weaker membership signature.
- **Computational cost** of the unlearning operation, measured in epochs, wall-clock time, or whether retraining is required at all. Instant unlearning corresponds to zero retraining cost.
- **Data dependence**: whether the unlearning procedure requires access to the forget set, the retain set, both, or neither.

2.6 Datasets

The datasets used throughout this thesis are predominantly medical-imaging benchmarks spanning endoscopy, prostate MRI, dermoscopy, and pancreatic CT, complemented by standard non-medical benchmarks for ablation and comparison against the broader literature. Medical-imaging datasets in this scale remain scarcer and smaller than their natural-image counterparts, in part because expert annotation is the bottleneck rather than data collection.

- **Kvasir-SEG** [75]: 1,000 colonoscopy images of gastrointestinal polyps with manual segmentation masks verified by an experienced gastroenterologist; gaze annotations on the same images were subsequently collected by Zhong et al. [189] from medically-experienced annotators under a 9-point calibration protocol.
- **NCI-ISBI 2013** [14]: 80 prostate T2-weighted MRI cases from the Cancer Imaging Archive, acquired under two protocols across two institutions (Boston Medical Center at 1.5T and Radboud University Nijmegen Medical Center at 3T), with annotations of the central gland and peripheral zone and matching gaze annotations from Zhong et al. [189].
- **Cyst-X cohort** [120]: a multicenter dataset of 746 patients with pancreatic cystic lesions from seven academic medical centers, with histopathologically confirmed IPMN labels and paired MRI. Section 3.1 uses a 359-patient subset of histopathologically confirmed cases, with an independent test set of 170 patients matching the radiomics-DL fusion model evaluation cohort.
- **PROMISE12** [100]: prostate MRI segmentation challenge, 50 training cases acquired across four centers on different scanners.
- **CVC-ClinicDB** [11] and **CVC-300** [164]: colonoscopy polyp segmentation datasets used as held-out evaluation sets for cross-dataset transfer from Kvasir-SEG.
- **ISIC 2019** [61, 29, 160]: 23,247 dermoscopic images of skin lesions aggregated from three publicly released sources (BCN20000, HAM10000, and the MSK dataset). Section 4.1 partitions the dataset by source institution into six federated clients with heterogeneous class distributions, following the Fed-ISIC2019 setup.
- **MedMNIST v2** [175]: a collection of standardized lightweight medical image classification benchmarks across modalities. This thesis uses three of its 2D subsets for prompt-based unlearning experiments: BloodMNIST (blood cell microscopy, 8 classes),

DermaMNIST (dermatoscopic images of pigmented skin lesions from HAM10000, 7 classes), and OrganSMNIST (sagittal abdominal CT slices, 11 organ classes).

- **UrbanSound8K** [138]: 8,732 labeled urban sound clips across ten classes, used for prompt-based unlearning beyond the image domain.
- **CIFAR-10/100** [86]: standard natural-image classification benchmarks used for verifiable unlearning experiments.

Chapter 3

Knowing: Interpretable Medical Image Analysis

This chapter argues that pre-committing to interpretability structurally unlocks capabilities that post-hoc explanation cannot deliver. The chapter develops this argument at two scales. Section 3.1 works at the pipeline scale, with a conventional model and post-hoc explanation and uncertainty methods, but with a structural commitment to producing the signals together at the decision moment rather than as separate analyses. Sections 3.2 and 3.3 move the commitment into the architecture itself, constraining the model’s spatial attention by construction to regions a clinician fixates on, through prompts derived from gaze and supplied to a frozen Segment Anything Model. The prompt is the explanation.

3.1 Uncertainty-aware Explainable IPMN risk classification

Intraductal papillary mucinous neoplasms (IPMN) account for up to 80% of the roughly three million pancreatic cysts diagnosed annually in the United States and impose lifetime surveillance costs that reach approximately \$171,000 per patient [147, 134]. These lesions present a clinical dilemma: although at risk of malignant transformation, the majority remain benign, resulting in unnecessary surveillance and interventions [118, 49], and unreliable guidelines demonstrate the need for better clinical management systems [60, 179, 33, 173].

Malignancy potential varies from 1–38% for branch-duct IPMNs to 33–85% for main-duct and mixed-type IPMNs [118, 49, 153, 32], and inter-user variability in guideline interpretation leaves a substantial fraction of patients either overtreated or undertreated. The latest standardization attempt ‘2024 Kyoto Criteria’ [118] integrates MRI, endoscopic ultrasound with fine-needle aspiration (EUS-FNA), and molecular analysis, but each imaging modality has known limitations. MRI and magnetic resonance cholangiopancreatography (MRCP) demonstrate inadequate sensitivity for high-grade dysplasia, particularly in small lesions [137, 68]. EUS-FNA provides tissue-specific information but involves procedural risks, operator-dependent variability, and cannot confidently exclude malignancy [35, 126, 151]

Artificial intelligence offers a promising alternative. Radiomics and deep learning extract quantitative imaging features beyond human visual perception [45], and recent studies report encouraging performance for IPMN classification [31, 89, 140, 177]. However, a critical barrier impedes clinical adoption: conventional classification architectures provide no indication of which cases they are likely to misclassify, exposing the radiologist to silent failure. Despite the recognized importance of this problem, no prior study has comprehensively applied explainability methods to IPMN risk stratification. Existing pancreatic AI research has focused predominantly on performance metrics, leaving clinicians without insight into when and why models succeed or fail, and without a systematic understanding of how IPMN type, size, and anatomical location interact with model confidence. This study, to our knowledge the first of its kind, deliberately challenges our previously established radiomics-deep learning fusion model [176] with a histopathologically confirmed cohort representing the most diagnostically difficult cases encountered in clinical practice. Rather than presenting AI as an infallible diagnostic replacement, we introduce a comprehensive XAI and uncertainty quantification framework designed to act as a ‘safety net’ by identifying AI vulnerabilities across specific Kyoto Criteria subgroups and enabling a human-in-the-loop triage system

3.1.1 Method

Study Design and Patient Population

The imaging dataset was sourced from the Cyst-X dataset, previously established by our collaborative research consortium and made publicly available. This dataset comprises 746 patients from seven academic medical centers [120]. Of these patients, we included cases with only histopathologically confirmed IPMNs, yielding a cohort of 359 patients that was also applied to our earlier work [10]. For this in-depth explainability analysis, we evaluated a subset of 170 cases from the 359-patient cohort, corresponding to the independent test set of our radiomics–DL fusion model (Fig. 3.1). The following sections outline the categorization framework and analytical methods used in this study.

IPMN Categorization for Performance Analysis

To enable the first systematic analysis of the influence of IPMN characteristics on AI performance, we developed a comprehensive categorization framework encompassing four clinically relevant dimensions: risk group, cyst type, anatomical location, and size based on Kyoto guideline thresholds. These dimensions reflect established clinical risk markers whose relationship to AI performance has never been systematically investigated.

Cyst Type and Malignancy Risk Categorization

Dysplasia grades of Branch-Duct (BD) IPMNs were determined from pathology results. All Main-Duct (MD) IPMN and mixed-type IPMN cases underwent surgical resection with pathological evaluation. Subjects were stratified based on dysplasia grade: lesions exhibiting low-grade dysplasia (LGD) were classified as Low-Risk, while lesions with high-grade dysplasia (HGD) and/or invasive carcinoma (IC) were classified as High-Risk.

Cyst Location Categorization

Lesions were analyzed using both region and extent categorizations. For the region analysis, lesions were classified into six groups: (1) head, (2) body, (3) tail, (4) head + body, (5) body + tail, or (6) whole-pancreas (head + body + tail). For the extent analysis, the same cysts were recategorized based on the extent of pancreatic involvement: confined to a single region, spanning two regions, or involving the whole-pancreas.

This dual framework enabled evaluation of both focal and diffuse cyst involvement patterns, as extensive pancreatic involvement may influence both surgical planning and model prediction accuracy (Table 3.3).

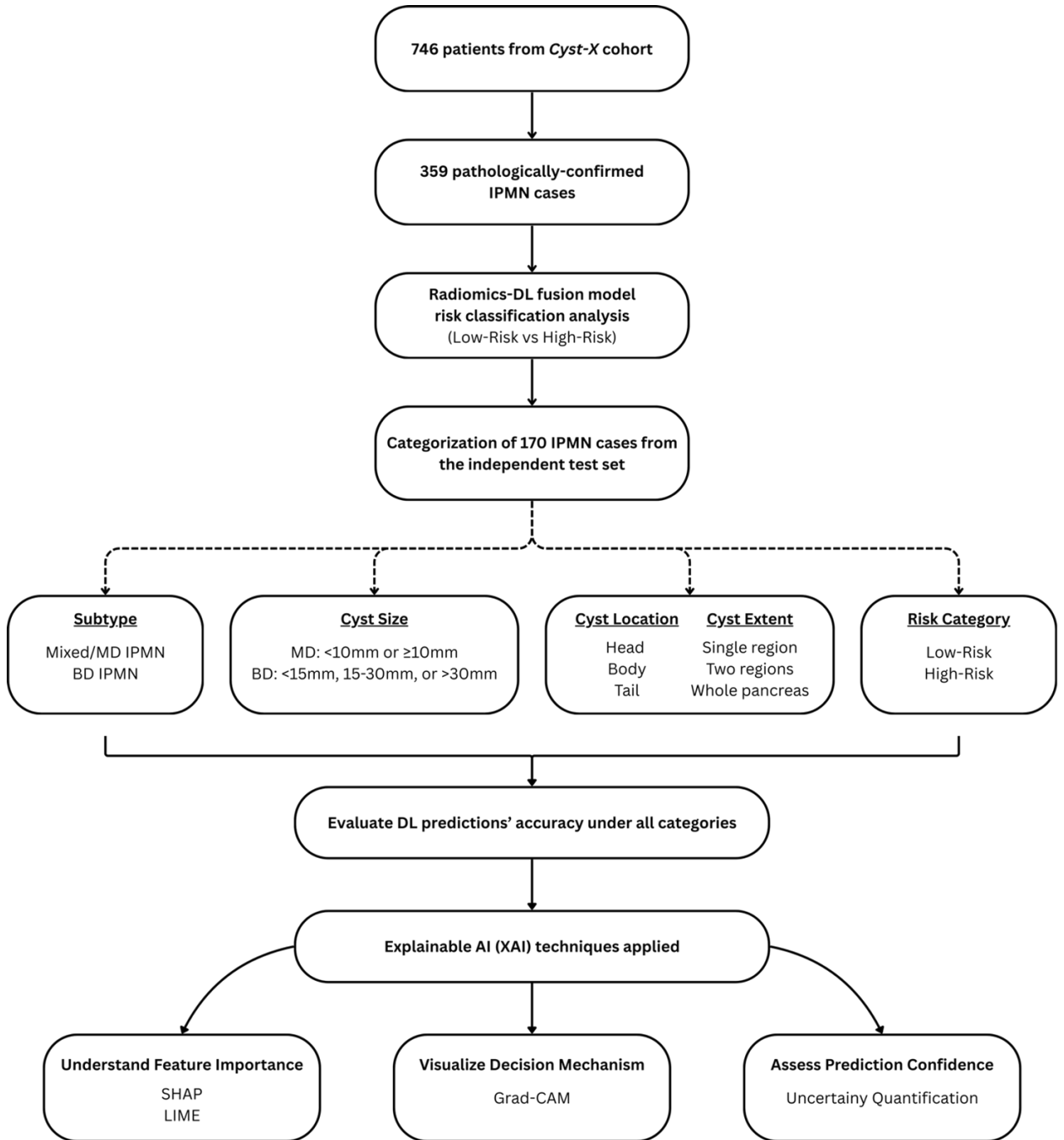


Figure 3.1: Cohort selection and methodology flowchart.

Cyst Size Categorization

Cyst size categorization was performed using type-specific thresholds informed by clinical guidelines and prior evidence [118]. For BD-IPMNs, maximum cyst diameter was categorized as <15 mm, 15–30 mm, or >30 mm. The ≥ 30 mm threshold represents a worrisome feature per Kyoto guidelines, while the 15 mm cutoff was supported by a multicenter surveillance study demonstrating that cysts ≥ 15 mm carried significantly increased risk of developing worrisome features or high-risk stigmata compared with cysts <15 mm, even when initially classified as low-risk BD-IPMNs [118, 153].

For MD-IPMNs, categorization was instead based on main pancreatic duct (MPD) diameter, stratified as <10 mm or ≥ 10 mm, with the latter representing high-risk stigmata according to the Kyoto guidelines [118] (Fig. 3.2).

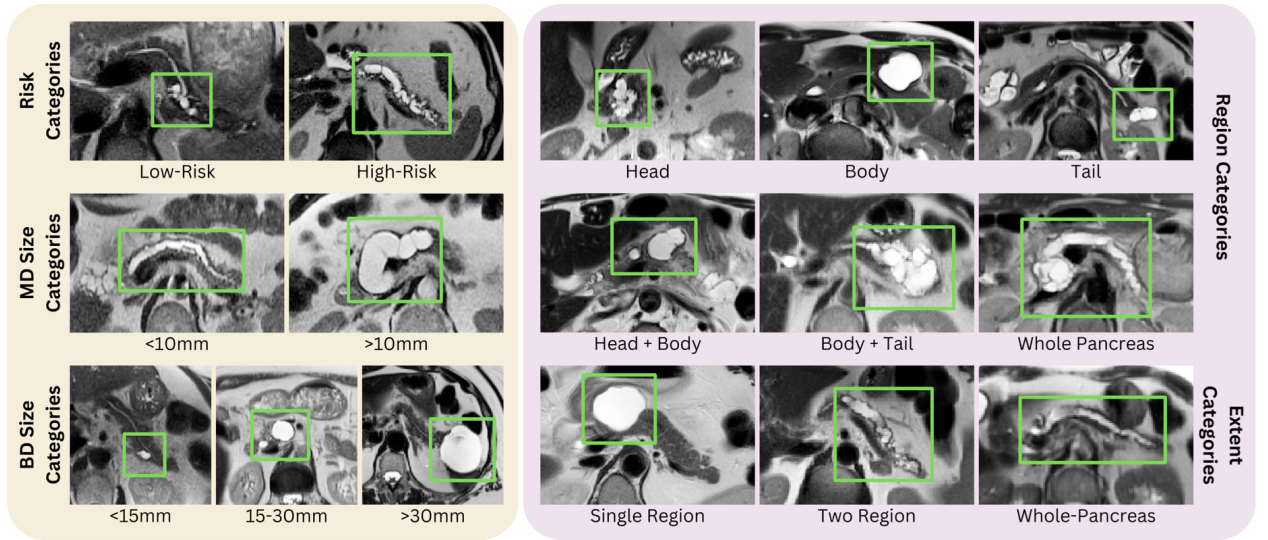


Figure 3.2: Representative cases of in-depth categorization from our cohort. Left panel (yellow) — Risk Categories: (Upper row) Low-risk and High-risk IPMNs; (Middle row) Main Duct (MD) Size Categories: IPMNs with MPD diameter <10 mm versus ≥ 10 mm; (Bottom row) Branch Duct (BD) Size Categories: IPMNs stratified by maximum cyst diameter <15 mm, 15–30 mm, and >30 mm. Right panel (pink) — (Upper row) Region Categories: IPMNs categorized by location, including head, body, tail, head + body, body + tail, and whole-pancreas; (Bottom row) Extent Categories: Classified by cyst distribution as single region, two regions, or whole-pancreas involvement.

AI Framework for IPMN Classification

The radiomics–deep learning (DL) fusion model evaluated in this study corresponds to our previously published IPMN risk stratification framework [10], in which model development, performance optimization, and validation were previously performed. The present study

focuses on explainability analysis of this established model rather than model development or optimization. Therefore, the model was applied without architectural or training modifications to enable systematic explainability and uncertainty analysis across clinically relevant IPMN subgroups.

Manual Segmentation and Segmentation Quality Assessment

Manual segmentation of each IPMN lesion was performed using ITK-Snap (Version 4.2.0) by an interdisciplinary team with expert radiologists’ review to ensure accuracy and consistency (Fig. S1) [182]. For BD-IPMN subjects, the index lesion was defined as the cyst sampled in EUS-FNA or surgically resected. Approximated MD cyst boundaries were discussed among abdominal radiologists and decided by consensus prior to segmentation due to the complex nature of their anatomy. Segmentation quality and reproducibility were validated through interobserver and intraobserver agreement analyses, with complete methodological details provided in our prior publication [10].

Radiomics–Deep Learning Fusion Algorithm for IPMN Risk Stratification

The data from individual centers were organized into four distinct trial configurations for testing purposes, while the remaining data served as a basis for cross-validation procedures. To replicate real-world clinical scenarios, each trial’s test set comprised complete datasets from one or two centers, resulting in variable test set sizes across trials. Four test groups were divided as Trial 1 ($n = 29$), Trial 2 ($n = 35$), Trial 3 ($n = 35$), and Trial 4 ($n = 75$). Trial groups were strategically selected to ensure balanced representation of low-risk and high-risk cases across both training and testing datasets.

Our hybrid radiomics–DL fusion approach was constructed by integrating probabilities from a radiomics-based Random Forest classifier [18] and a DL-based DenseNet121 model [69] (Fig. 3.7). Radiomic feature selection used 5-fold cross-validation, retaining features with Spearman correlation coefficients below 0.6 to reduce redundancy. Both the Random Forest radiomics model and the DL architecture were retrained on identical training datasets, with consistent data partitioning to ensure valid comparisons. After model training, the probability outputs from both approaches were combined to form an ensemble, which was assessed on the validation and test datasets.

The decision-level fusion strategy combined probabilities from both DenseNet121 and Random Forest models. Building upon our previous methodology [176], we implemented an exact sampling approach for the fusion technique and identified optimal hyperparameters through grid search within a five-fold cross-validation framework. The fusion algorithm

incorporated two key hyperparameters: threshold t and weight k . When radiomics predictions surpassed threshold t , the final output relied exclusively on the radiomics prediction, effectively discarding the DL prediction. Conversely, when the threshold is not exceeded, the fusion output represents a weighted combination of both model predictions, with the radiomics model weight set to $1 - k$ and the DL model weight assigned as k .

This fusion methodology was executed twice, utilizing either 2D or 3D radiomic feature sets. Weighted averaging was implemented to account for varying subject representation across participating centers. Our complete fusion workflow is illustrated in Figure 3.7.

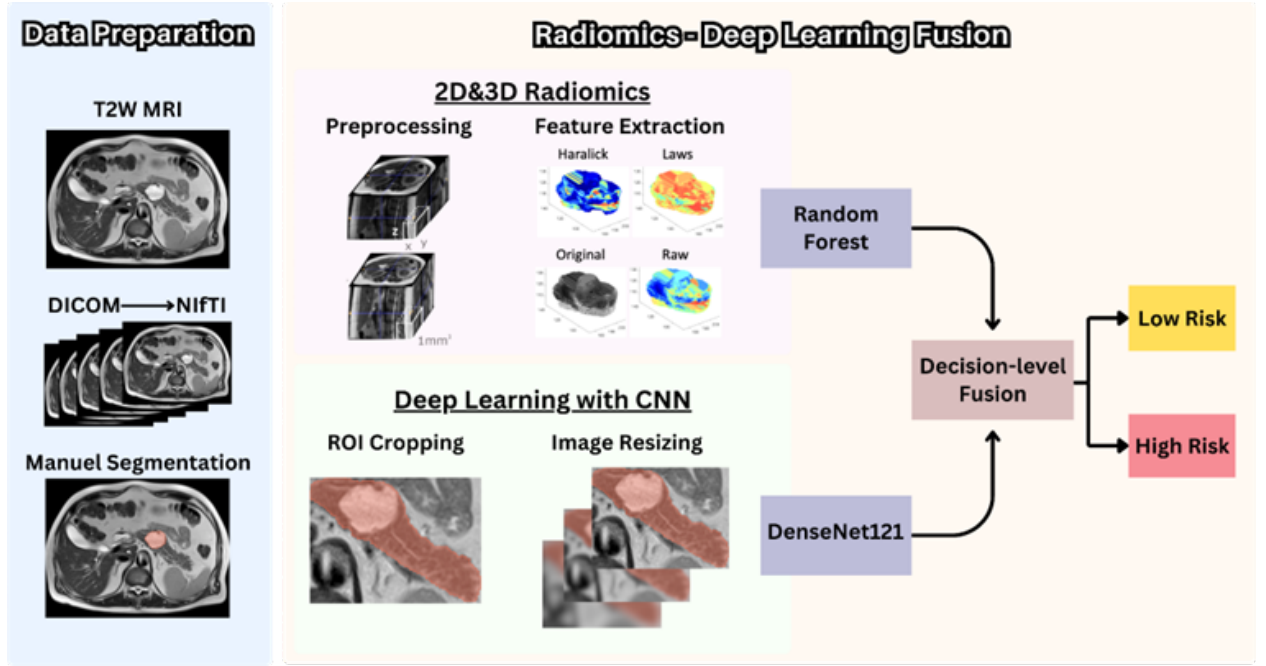


Figure 3.3: Our radiomics–deep learning fusion algorithm pipeline.

Explainability Frameworks

To elucidate the model’s decision-making process and identify key imaging biomarkers, we performed comprehensive explainability analyses using SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) for feature attribution, Grad-CAM for spatial visualization, and uncertainty quantification for confidence estimation across the entire cohort.

SHAP and LIME Analysis

We implemented two complementary explainability methods—SHAP [103] and LIME [132]—to quantify how individual radiomic features contributed to model predictions.

SHAP assigns to each radiomic feature a contribution score by computing its influence across all possible feature combinations. This provides both global feature importance (identifying which features are most influential across all cases) and local interpretability (showing how features contribute to individual predictions). SHAP values were computed for all samples in each test set using the TreeExplainer, with the model output set to probability mode to explain predicted class probabilities. Global feature importance was derived by aggregating SHAP values across all samples, while local SHAP values indicated the direction and magnitude of each feature’s contribution to individual high-risk predictions. To assess explanation strength, we computed the total absolute SHAP value for each prediction by summing the absolute values of all feature contributions.

LIME generates case-specific explanations by systematically perturbing radiomic features and observing how predictions change, then fitting a simplified linear model to approximate the complex model’s behavior locally. We used LimeTabularExplainer with approximately 5,000 perturbations per case. Total absolute LIME values were similarly computed by summing absolute feature contributions. While SHAP provides globally consistent attributions across the entire dataset, LIME is optimized for individual case interpretation. These methods enabled us to identify which imaging characteristics (e.g., texture patterns, shape descriptors) drove the model’s risk predictions and to evaluate whether explanation strength correlates with prediction confidence.

Grad-CAM

Grad-CAM [142] was used to visualize the deep learning model’s inference process. Grad-CAM computes the gradient of the network’s output with respect to the feature maps in the last convolutional layer. These gradients are then used to produce a heatmap that highlights the image regions most relevant to the model’s decision. In our study, Grad-CAM enabled us to generate attention maps that revealed whether the model focused on the actual cyst or on surrounding tissues during classification (Fig. S2).

Uncertainty Quantification

For the radiomics-fused DL method, we implemented and trained our models using Monte Carlo (MC) dropout with a dropout probability of 0.3. At inference time, we performed T stochastic forward passes ($T_{\text{pass}} = 30$) to approximate the predictive distribution and estimate uncertainty [42]. Formal definitions follow Section 2.3.

In our two-branch uncertainty estimation architecture, we further leveraged the principle that the radiomic features random forest classifier inherently quantifies uncertainty through

ensemble voting [18]. Each of the T forward passes from the deep learning branch was paired with predictions from all 47 decision trees of the random forest, resulting in a combined set of $T = 30 \times 47$ predictive samples. This extended distribution enabled a robust estimation of both data- and model-driven uncertainties.

Statistical Analysis

Model performance was evaluated using standard classification metrics. The area under the receiver operating characteristic curve (AUC) was calculated to assess discrimination ability. Sensitivity and specificity were computed at the optimal operating point determined by Youden’s index. The Brier score was calculated to evaluate overall prediction accuracy and calibration, with lower values indicating better performance.

Comparisons between categories and predictions (correct and incorrect) were evaluated using Fisher’s exact tests. Associations between categories and prediction uncertainty (a continuous variable) were assessed using Analysis of Variance (ANOVA) models, with pair-wise p -values adjusted using the Tukey method for feature variables with more than two categories.

To evaluate the relationship between explanation strength and prediction confidence, associations between total absolute SHAP/LIME values and model uncertainty were analyzed using Spearman rank and Pearson correlation analyses. Scatter plots with linear regression lines and 95% confidence intervals were used to visualize these relationships on a per-case basis. Statistical significance is defined as $p < 0.05$. Analyses were conducted using R, SciPy, Statsmodels, Scikit-learn, SHAP, and LIME libraries.

3.1.2 Experimental Results

Cohort Characteristics

The cohort was stratified by pathologically confirmed dysplasia grade into Low-Risk (low-grade dysplasia) and High-Risk (high-grade dysplasia or invasive carcinoma) groups. The Low-Risk group consisted of 107 scans, comprising 74 BD-IPMN cases (79.6% of 93 total BD-IPMN cases) and 33 MD or mixed-type IPMN cases (42.9% of 77 total MD/mixed-type cases). The High-Risk group included 63 subjects: 19 with BD-IPMN (20.4% of all BD-IPMN cases) and 44 with MD/mixed-type IPMN (57.1% of all MD/mixed-type cases) (Table 3.1).

In the BD-IPMN group, mean maximum cyst diameter was 21.9 mm (SD = 11.7, median = 18.2 mm, IQR 14.0–29.0 mm) in Low-Risk cases and 28.8 mm (SD = 15.4, median = 25.1 mm, IQR 15.5–37.5 mm) in High-Risk cases. In MD-IPMNs, the mean MPD

Table 3.1: Average cyst size by IPMN types and risk groups in our cohort.

IPMN Type	Risk Group	n (%)	Mean Size $\pm$ SD (mm)	Median (IQR) (mm)
BD-IPMN	Low-Risk	74 (79.6)	21.9 ± 11.7	18.2 (14.0–29.0)
	High-Risk	19 (20.4)	28.8 ± 15.4	25.1 (15.5–37.5)
MD-IPMN	Low-Risk	33 (42.9)	9.7 ± 5.0	8.6 (6.8–9.9)
	High-Risk	44 (57.1)	11.8 ± 5.8	10.3 (7.9–14.1)

Table 3.2: Distribution of Low-Risk and High-Risk IPMN lesions by region and extent categorization in our cohort.

Region	Low-Risk n (%)	High-Risk n (%)
Head	40 (37.4)	12 (19.0)
Body	15 (14.0)	3 (4.8)
Tail	15 (14.0)	8 (12.7)
Multi-regional	37 (34.6)	40 (63.5)
Head + Body	12 (11.2)	6 (9.5)
Body + Tail	4 (3.7)	8 (12.7)
Whole-pancreas	21 (19.6)	26 (41.3)
Extent	Low-Risk n (%)	High-Risk n (%)
Single region	70 (65.4)	23 (36.5)
Two regions	16 (15.0)	14 (22.2)
Whole-pancreas	21 (19.6)	26 (41.3)
Total	107 (100)	63 (100)

diameter was 9.7 mm (SD = 5.0, median = 8.6 mm, IQR 6.8–9.9 mm) for Low-Risk cases and 11.8 mm (SD = 5.8, median = 10.3 mm, IQR 7.9–14.1 mm) for High-Risk cases. These size categories were used to evaluate model performance across different cyst dimensions, as reported in Table 3.1 (Fig. 3.4).

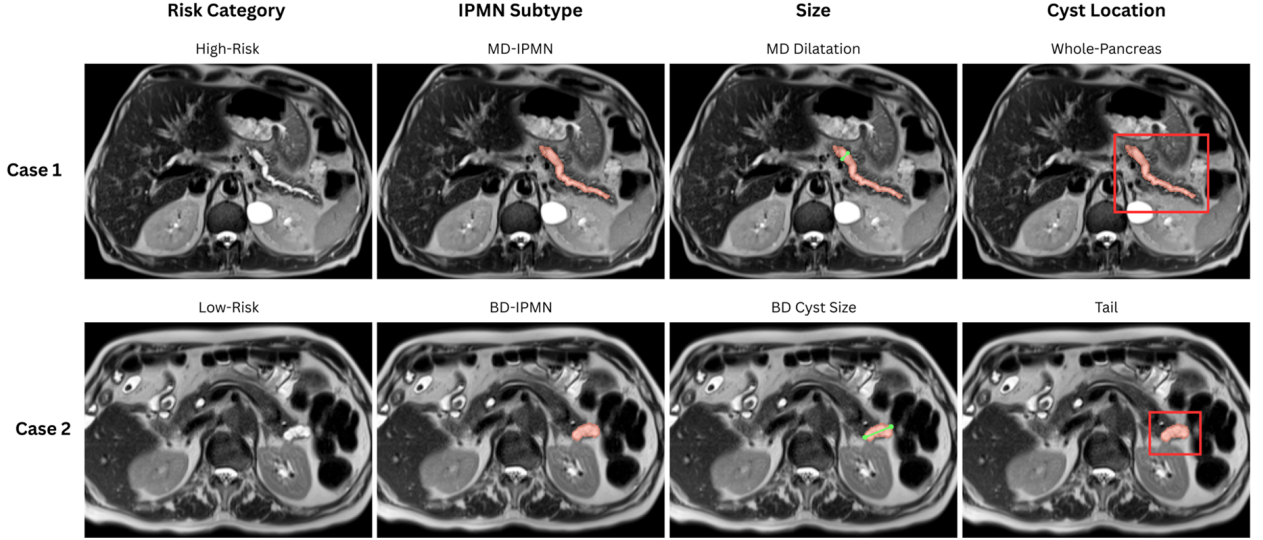


Figure 3.4: Representative IPMN characterization by risk category, type, size, and location. Axial T2-weighted MRI images demonstrating the classification framework applied to IPMN cases. Case 1 shows a High-Risk MD-IPMN with main duct dilatation and whole-pancreas involvement. Case 2 demonstrates a Low-Risk BD-IPMN with a cyst in the pancreatic tail. Red overlays indicate segmented regions of interest. Each case illustrates the four-dimensional characterization approach: (1) risk stratification based on dysplasia grade, (2) type classification (MD-IPMN vs BD-IPMN), (3) size measurements (main duct diameter vs cyst diameter), and (4) anatomical location mapping.

Region analysis showed Low-Risk lesions ($n = 107$) most frequently isolated to the head (37.4%), while High-Risk lesions ($n = 63$) predominantly showed whole-pancreas involvement (41.3%). Extent analysis showed Low-Risk lesions were more often confined to a single region (65.4%), while High-Risk lesions more frequently involved the whole-pancreas (41.3%). Detailed distributions are provided in Table 3.2.

Model Performance Overview

We evaluated the performance of the radiomics–deep learning fusion model for malignancy risk prediction across IPMN characteristics, examining both prediction accuracy and uncertainty as complementary metrics (Table 3.3). The model achieved an overall accuracy of 67.1% (114/170 cases) in classifying IPMN risk. Performance and uncertainty varied substantially across IPMN subgroups when analyzed by cyst type, size, risk category, and

anatomical location.

Model uncertainty was strongly correlated with prediction accuracy. Correct predictions exhibited significantly lower uncertainty (0.72 ± 0.01 s.e.m.) compared to incorrect predictions (0.78 ± 0.01 s.e.m., $p < 0.0001$).

Cyst Types and Risk Analysis

BD-IPMNs demonstrated higher accuracy than MD-IPMNs (71.0% vs. 62.3%, $p = 0.254$), though the difference was not statistically significant. However, the model exhibited significantly lower uncertainty for BD lesions compared to MD lesions (0.72 ± 0.01 s.e.m. vs. 0.76 ± 0.01 s.e.m., $p = 0.012$).

The model’s ability to correctly classify cases differed significantly between risk groups. Low-Risk lesions classified more accurately (81.3%, 87/107) compared to High-Risk lesions (42.9%, 27/63; $p < 0.001$). Model uncertainty was significantly lower in Low-Risk lesions compared to High-Risk lesions (0.73 ± 0.01 vs. 0.76 ± 0.02 , $p = 0.034$).

To evaluate classification performance across cyst subtypes, ROC analysis was performed for the entire cohort and stratified by cyst type, showing moderate overall performance with an AUC of 0.685. Performance differs by subtype, with better discrimination and calibration in BD cysts (AUC = 0.694, lower Brier score) compared with MD cysts (AUC = 0.594), while maintaining high sensitivity across all groups (Fig. 3.5).

Cyst Size Analysis

For BD-IPMNs, model accuracy declined progressively with increasing cyst size: <15 mm (81.3%), 15–30 mm (63.2%), and >30 mm (61.5%); $p = 0.164$. Furthermore, uncertainty increased significantly with large (>30 mm) cysts compared to small (<15 mm) cysts (0.75 ± 0.02 vs. 0.69 ± 0.02 , $p = 0.048$).

In MD-IPMNs, model performance remained stable across ductal diameter categories (<10 mm: 65.9%, ≥ 10 mm: 63.6%; $p = 1.000$), and uncertainty did not differ significantly between the two groups ($p = 0.461$).

Cyst Region and Extent Analysis

Model performance varied both by region of the cyst and extent of pancreatic involvement. In the region analysis, accuracy was highest for cysts in the pancreatic body (77.8%), followed by head (69.2%), and lowest in the tail (47.8%). Cases with whole-pancreas involvement achieved 76.6% accuracy, though overall differences among anatomical sites were not statistically significant ($p = 0.074$). Furthermore, uncertainty showed no significant variation

Metrics	Overall	BD-IPMN	MD-IPMN
AUC (%)	68.5	69.4	59.4
Specificity (%)	34.6	43.2	15.2
Sensitivity (%)	93.7	94.7	93.2
Brier Score	0.2307	0.2188	0.2451

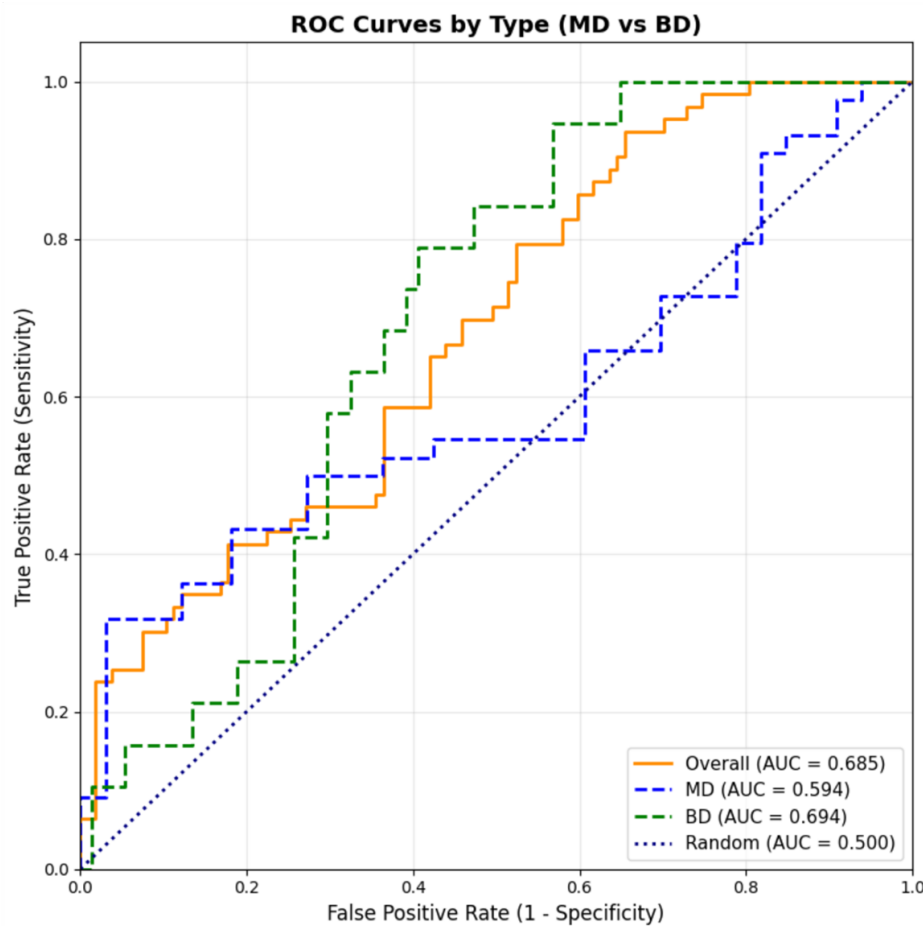


Figure 3.5: Classification performance metrics and ROC curves stratified by IPMN cyst type. The table summarizes AUC, sensitivity, specificity, and Brier Score for the overall cohort, BD-IPMNs, and MD-IPMNs. ROC curves demonstrate risk stratification performance for BD-IPMNs (green dashed, AUC = 0.694), MD-IPMNs (blue dashed, AUC = 0.594), and the overall cohort (orange solid, AUC = 0.685). The diagonal dotted line indicates random classification (AUC = 0.500).

among specific anatomical locations ($p = 0.152$).

In the extent analysis, accuracy varied across involvement patterns: single region (65.6%), two regions (56.7%), and whole-pancreas (76.6%; $p = 0.168$). Model uncertainty showed a trend toward higher values with greater pancreatic involvement (single region: 0.73 ± 0.01 , two regions: 0.75 ± 0.02 , whole-pancreas: 0.76 ± 0.01), though differences were not statistically significant ($p = 0.121$).

Table 3.3: Model accuracy and uncertainty across IPMN characteristics. Uncertainty values represent estimated marginal means $\pm$ standard error (s.e.m.) from linear regression models. Reported p -values represent the overall p -value from the ANOVA test.

Parameter	Group	Acc. (%)	p -value	Uncertainty (mean $\pm$ s.e.m.)	p -value
Overall	–	67.1	–	–	–
Prediction outcome	Correct	–	–	0.72 (0.01)	< 0.0001
	Incorrect	–	–	0.78 (0.01)	
Risk	Low-Risk	81.3	< 0.001	0.73 (0.01)	0.034
	High-Risk	42.9		0.76 (0.01)	
Type	BD-IPMN	71.0	0.254	0.72 (0.01)	0.012
	MD-IPMN	62.3		0.76 (0.01)	
BD size	<15 mm	81.3	0.164	0.69 (0.02)	0.048
	15–30 mm	63.2		0.73 (0.02)	
	>30 mm	61.5		0.75 (0.02)	
MD size	<10 mm	65.9	1.000	0.77 (0.01)	0.461
	≥ 10 mm	63.6		0.76 (0.01)	
Region	Head	69.2	0.074	0.73 (0.01)	0.152
	Body	77.8		0.70 (0.03)	
	Tail	47.8		0.75 (0.02)	
	Multi-regional	68.8		–	
	Head + Body	66.7		0.73 (0.03)	
	Body + Tail	41.7		0.77 (0.03)	
	Whole-pancreas	76.6		0.76 (0.02)	
Extent	Single region	65.6	0.168	0.73 (0.01)	0.121
	Two regions	56.7		0.75 (0.02)	
	Whole-pancreas	76.6		0.76 (0.01)	

Explainability Analysis

SHAP and LIME Analysis

Across the cohort, model predictions were predominantly influenced by radiomic texture features, particularly entropy- and correlation-based descriptors, which consistently exhibited larger attribution magnitudes across independent test sets. SHAP summary plots demon-

strated stable feature importance patterns, with several texture-derived features ranking among the most influential predictors across all evaluation groups (Fig. S3).

Complementary LIME analyses were performed to evaluate local feature contributions using both signed and absolute aggregation strategies. While signed aggregation reflected the average directional influence of features, absolute aggregation captured overall contribution magnitude irrespective of direction. These analyses revealed that several features exerted substantial influence on model predictions but with variable directional effects across samples, suggesting local instability in feature contributions (Fig. S4).

We further examined the relationship between explanation strength and predictive uncertainty. Total absolute SHAP values demonstrated a moderate inverse correlation with uncertainty (Spearman $\rho = -0.41$, $p < 0.001$; Pearson $r = -0.38$, $p < 0.001$), indicating that stronger feature attributions were associated with lower uncertainty. In contrast, total absolute LIME values showed no meaningful association with uncertainty (Spearman $\rho \approx 0$, $p > 0.9$; Pearson $r \approx 0$, $p > 0.6$) (Fig. S5). Linear regression with 95% confidence intervals confirmed a statistically significant negative relationship between SHAP magnitude and uncertainty, whereas LIME explanations remained largely uncorrelated across the full range of prediction confidence.

Uncertainty Quantification Analysis

In our cohort, uncertainty differed significantly across key clinical comparisons. Higher uncertainty was consistently associated with incorrect predictions, high-risk lesions, and MD-IPMNs, whereas no significant differences were observed across cyst location or extent of pancreatic involvement. Detailed subgroup analyses are presented in the corresponding sections, with summary statistics provided in Table 3.3.

To evaluate the clinical utility of uncertainty estimation, we analyzed how prediction accuracy changes when the model defers high-uncertainty cases to human review. We examined three uncertainty types across all cases: combined uncertainty (sum of aleatoric and epistemic components), aleatoric uncertainty (reflecting inherent ambiguity in imaging data), and epistemic uncertainty (reflecting model knowledge limitations).

The relationship between uncertainty thresholds and prediction accuracy when implementing selective prediction strategies is demonstrated in Figure 3.6. When the model restricts predictions to only its most confident cases (low uncertainty threshold), accuracy approaches 100%, but coverage—the proportion of cases receiving automated predictions—decreases substantially. Combined uncertainty (total uncertainty) provided the most reliable and consistent tradeoff between accuracy and coverage across all threshold levels. Epistemic uncertainty performed well at identifying the most confident predictions but exhibited in-

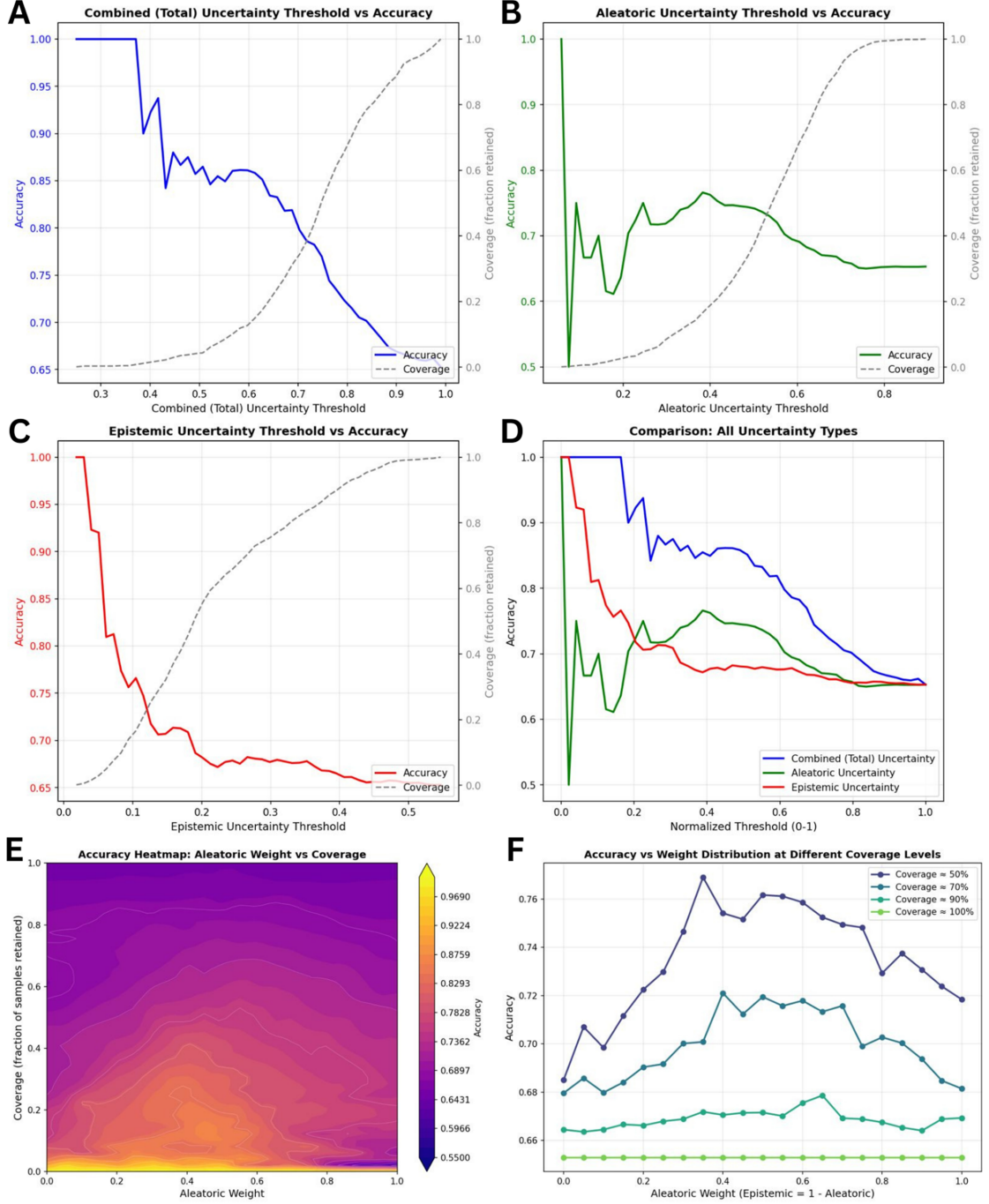


Figure 3.6: Uncertainty thresholding and optimal weighting of aleatoric and epistemic uncertainty components. (A–C) Relationship between uncertainty thresholds and prediction accuracy for combined (blue), aleatoric (green), and epistemic (red) uncertainty, respectively. Accuracy is evaluated as predictions exceeding a given uncertainty threshold are excluded, while dashed curves indicate the corresponding coverage (fraction of retained samples). (D) Comparison of all uncertainty types on a normalized threshold scale. (E) Accuracy heatmap showing model performance across different weighting schemes between aleatoric and epistemic uncertainty components and varying coverage levels, with brighter colors indicating higher predictive accuracy. (F) Accuracy curves at fixed coverage levels illustrating how optimal weighting between aleatoric and epistemic uncertainty varies depending on the desired coverage target.

stability at higher coverage levels. Aleatoric uncertainty showed the most variable behavior, suggesting limited utility as a standalone criterion for clinical deferral decisions.

To determine whether a weighted combination of aleatoric and epistemic uncertainty could outperform their simple sum, we systematically evaluated different weighting schemes across multiple coverage levels (Fig. 3.6). At 50% coverage, where half of cases are deferred to radiologists, the optimal weighting assigned 45% to aleatoric and 55% to epistemic uncertainty, achieving 78.3% accuracy compared to 77.3% with equal weighting—a 1.0 percentage point improvement. At 70% coverage, shifting the balance to 60% aleatoric and 40% epistemic weighting yielded 73.0% accuracy. These findings indicate that epistemic uncertainty is more informative when selecting the most reliable predictions (lower coverage), while aleatoric uncertainty becomes relatively more important when accepting a larger proportion of cases (higher coverage). At near-complete coverage ($> 90\%$), the choice of weighting had minimal impact on accuracy, as most cases were included regardless of their uncertainty profiles.

Crucially, the model’s lowest accuracy was observed in High-Risk lesions (42.9%). However, this critical performance gap was successfully captured by our uncertainty metrics. Epistemic uncertainty was significantly higher for these incorrect, high-risk predictions compared to low-risk predictions ($p = 0.034$). This demonstrates that the algorithm ‘knows what it does not know’. By implementing the combined uncertainty thresholding strategy shown in Figure 3.6, we simulated a clinical triage scenario. When the model deferred the top 30–40% of highly uncertain cases to a human radiologist, the automated accuracy on the remaining autonomous predictions increased significantly to approximately 80–85%. This proves that uncertainty quantification can transform a moderately accurate standalone model into a highly reliable triage partner.

3.1.3 Discussion

This section evaluates whether interpretability and uncertainty, treated as joint properties of the same model, are informative enough to govern when an AI prediction should be acted on. The behavior of the previously published radiomics–DL fusion model [10] was analyzed across 170 histopathologically confirmed IPMN cases from seven international centres, using SHAP, LIME, Grad-CAM, and MC-dropout uncertainty. The model achieved an overall accuracy of 67.1% on the independent test set. Model uncertainty was significantly lower for correct predictions than for incorrect predictions (0.72 vs. 0.78, $p < 0.001$), which is the precondition for any deferral policy: an uncertainty signal that does not track accuracy cannot support clinical triage. Detailed radiological and explainability results follow.

It is important to highlight that sensitivity is much more critical for MRI-based diagnosis, as this is usually followed by a second-line EUS-FNA for confirmation. Since EUS-FNA is considerably more specific, especially when combined with molecular classification, it is probable that the combination of the MRI-Cyst XAI algorithm and EUS will be more accurate. Furthermore, our model’s 67.1% accuracy should be interpreted within the context of our study population. All cases underwent pathological confirmation by surgery or EUS-FNA, representing the challenging subset of patients in whom clinical suspicion warranted invasive procedures. This cohort deliberately excludes the substantial proportion of small, low-risk cysts, which are typically managed with surveillance alone, and lesions with straightforward imaging characteristics that would likely be classified more accurately. Our dataset, therefore, reflects the diagnostically difficult cases in which even clinical guidelines and expert radiologists face challenges, rather than the easier cases that dominate routine practice. In an unselected population including conservatively managed cysts, model performance would likely be higher, as the model would encounter a more favorable case mix reflecting real-world prevalence.

Our algorithm achieved higher classification accuracy for Low-Risk cases (81.3%) compared to High-Risk cases (42.9%; $p < 0.001$), along with significantly lower prediction uncertainty in Low-Risk cases ($p = 0.034$). A similar pattern was seen for cyst types, with lower uncertainty for BD-IPMNs than for MD-IPMNs ($p = 0.012$). This may be because low-risk and BD-IPMN cases tend to have simpler morphology and imaging features, such as uniformity and well-defined borders, making them easier to evaluate. Conversely, high-risk and MD-IPMN cases are more heterogeneous and complex on imaging, which can be more difficult even for expert radiologists. Moreover, even the current reference guidelines demonstrate suboptimal performance in accurately identifying malignancy potential in High-Risk lesions. Despite these, the reduced uncertainty and improved reliability observed in Low-Risk lesions and BD-IPMNs offer clinical value, as confident identification of low-risk conditions can help avoid unnecessary invasive procedures.

Our SHAP analysis consistently identified entropy and correlation-based texture descriptors as the primary drivers of model predictions. From a radiological perspective, high entropy corresponds to increased intra-lesional heterogeneity on T2W MRI. Clinically, this aligns with the biological progression of IPMNs; as low-grade dysplasia transitions to high-grade dysplasia or invasive carcinoma, the cyst fluid and architecture become increasingly complex due to the presence of solid components, mucin density and plugging, or mural nodules. The AI model is effectively mathematically quantifying this visual heterogeneity. However, as our uncertainty analysis shows, when this heterogeneity reaches the extreme levels seen in advanced MD-IPMNs, the imaging data becomes inherently ambiguous (aleatoric

uncertainty), prompting the model to correctly signal for expert human intervention.

The model evaluated in this study reflects the intrinsic behavior of our previously published framework, enabling explainability and uncertainty analyses to identify where predictions are reliable and where limitations arise. Building on this, size-based analysis revealed that MPD dilation ≥ 10 mm is a well-established high-risk feature and is cited in the Kyoto guidelines as one of the high-risk stigmata of IPMNs. While analyzing performance by cyst size, we found that algorithm accuracy remained stable across size categories for MD-IPMN, with 65.9% in the < 10 mm category and 63.6% in the ≥ 10 mm category ($p = 1.000$). Similarly, for BD-IPMNs, although a trend toward decreasing accuracy with increasing cyst size was observed: <15 mm (81.3%), 15–30 mm (63.2%), and >30 mm (61.5%), model accuracy showed no statistically significant difference between size groups ($p = 0.164$). Therefore, the algorithm aligns with the threshold specified in the Kyoto guideline and maintains stable performance across both below- and above-threshold groups. In such scenarios, AI has the potential to assist in confirming or refining risk feature thresholds and to support future clinical trials.

Beyond descriptive performance metrics, our uncertainty analysis demonstrates a clinically actionable strategy for improving the reliability of AI-assisted IPMN risk stratification. By applying predefined uncertainty thresholds, the model can selectively restrict automated predictions to cases in which it is most confident, while deferring higher-uncertainty cases to radiologist review. Under this selective prediction framework, model accuracy increased to approximately 80% at moderate coverage levels, illustrating that uncertainty-aware triage can substantially improve predictive reliability without excluding all automated assistance. Importantly, cases associated with higher uncertainty—such as high-risk lesions and MD-IPMNs—are the same scenarios in which clinical decision-making is most challenging. This underscores the value of uncertainty as an objective signal for heightened human oversight. Rather than replacing clinical judgment, this uncertainty-guided approach supports a hybrid decision model that offers a pragmatic pathway toward safer, more trustworthy integration of AI into pancreatic cyst management.

Explainability analysis demonstrated that prediction uncertainty was inversely associated with SHAP explanation strength, whereas no meaningful relationship was observed with LIME’s. Lower-uncertainty predictions exhibited stronger SHAP attributions, consistent with more confident and coherent feature contributions, while LIME explanations remained relatively stable across the uncertainty spectrum. This divergence likely stems from methodological differences: LIME focuses on local linear approximations optimized for individual predictions, potentially missing global relationships, while SHAP provides globally consistent attributions that better reflect how imaging features influence confident predic-

tions. From a clinical perspective, SHAP explanations may therefore provide added value by informing both interpretability and prediction reliability, helping radiologists identify cases that warrant greater scrutiny.

Similar explainability approaches to those applied in our study are utilized to understand AI classification algorithms in diverse medical applications. To our knowledge, only one study has directly employed these XAI methods to classify pancreatic cystic lesions. Huang et al. (2025) [70] developed a Multimodal Deep Forest (MDF) model to classify pancreatic cystic lesions into four categories: benign IPMN, malignant IPMN, serous cystadenoma, or mucinous cystic neoplasm. Their model utilized cyst features (e.g., main PD width, cyst communication with main PD, and cyst location) that were manually measured by radiologists from CT ($n = 568$) and MRI ($n = 449$) scans, in addition to clinical features (e.g., medical history, blood test results, age, and gender). Additionally, they used SHAP to identify the impact of each feature on the model’s performance. The MDF model achieved an average accuracy of 91.2% and identified the following as the most important predictors: complicated with pancreatitis, main pancreatic duct width, age, communication with the pancreatic duct, CA19-9 level, and CA125 level. Their results demonstrate the value of integrating clinical features with analyses of imaging features. Compared to theirs, our study inputs MRI images directly into the algorithm, enabling us to extract radiomic features and use DL to predict IPMN dysplasia grade. Although their MDF model performed well, manually extracting imaging features limits the number and types of usable features, potentially excluding those that cannot be directly observed by the human eye, and limits the model’s ability to be fully explained. Furthermore, our study focused solely on IPMN and classified lesions by type, size, and location. This is particularly valuable when considering type features that can influence model performance, as MD-IPMN is a higher-risk condition; thus, a combined evaluation of MD and BD IPMN could bias algorithm performance. Our study also collected data from seven international centers, thereby increasing the generalizability of our findings compared to Huang (2025).

A few additional studies have used XAI tools to evaluate pancreatic cancer risk in medical imaging. Pan et al. (2025) and Salanitri et al. (2022) both applied Grad-CAM to generate attention maps of their models’ analysis of MRI images [120, 140]. Heatmaps provide valuable insights into the areas of an image that carry more weight in the model’s decision-making, but they do not identify the specific features that attract the model’s focus. Both studies included Grad-CAM heatmaps as part of their analyses, noting that extra-cystic features in the image were identified but not investigating what these features specifically were. Other studies analyzed the features used by their models to diagnose tumor marker expression with SHAP. These studies either used patient clinical data, urine sample information, or

genetic markers as inputs. They successfully identified specific markers that affected their models’ performance and created SHAP-based scatter plots to illustrate the impact of each feature [123, 122]. However, these studies differ in scope from our work.

Several limitations should be considered when interpreting these findings. First, the retrospective study design limits causal inference and may introduce selection bias. Second, data gathered over two decades inherently vary in image quality and acquisition technique. Third, the relationship between radiomic features and clinical outcomes remains only partially understood, limiting the ability to confirm whether the model focuses on clinically relevant imaging biomarkers, and post-hoc explanations may not fully capture the internal workings of deep neural networks.

The contribution of this section to the chapter’s *Knowing* argument is that the model becomes legible from converging directions, through what features it uses, through what regions it attends to, and through when it should be trusted. SHAP identified entropy- and correlation-based texture descriptors as the consistent drivers of prediction across the cohort. These are precisely the features that prior radiomics work has independently associated with IPMN malignancy [62, 57, 157], so the model’s reasoning aligns with established clinical knowledge rather than with a shortcut. Grad-CAM showed spatial attention concentrated on the cyst itself rather than on surrounding tissue, reinforcing this conclusion from the spatial side. Two XAI methods, two different interrogation axes, the same result. The model is doing something a radiologist would recognize as reasonable.

A second XAI finding is methodological. SHAP attribution strength correlated inversely with predictive uncertainty, while LIME showed no comparable association. Not all post-hoc explanation methods carry reliability information. SHAP, with its globally consistent Shapley-value formulation, does. LIME’s local linear approximations do not. For clinical deployment, this implies that explanation strength is informative about reliability only when the underlying method is structured to be globally consistent.

The uncertainty signal supplies the third axis. Uncertainty tracks correctness, epistemic uncertainty rises on the high-risk subgroup where the model performs worst, and deferring the top 30–40% of high-uncertainty cases raises retained-case accuracy to approximately 80–85% on this cohort. The cases the model flags as uncertain are the same cases where clinical guidelines themselves underperform, high-risk lesions and MD-IPMNs.

Taken together, the three signals make the model knowable at the decision moment. A clinician can see what features drove the prediction, where the model looked, and how reliable the prediction is.

3.1.4 Publications

Halil Ertugrul Aktas, Gorkem Durak, Andrea Mia Bejar, Ziliang Hong, Rutger Hendrix, Hongyi Pan, Elif Keles, Fergan Bol, Yavuz B. Taktak, Cagdas Topel, et al. Uncertainty-aware explainable AI for pancreatic cysts: Identifying deep learning vulnerabilities and ensuring safe clinical triage in IPMN management. Submitted to *Nature Scientific Reports*, 2026.

3.2 Gaze-Guided Medical Image Segmentation: A Training-Free Approach using SAM Foundation Model

Medical image segmentation is a fundamental yet time-consuming task in radiology, necessitating automation solutions to alleviate the burden on clinicians. Deep learning models have demonstrated remarkable progress in segmentation tasks, but their widespread adoption is often hindered by the need for extensive labeled datasets and computationally expensive training processes. The emergence of foundation models, such as the Segment Anything Model (SAM, Section 2.1.2), offers a promising alternative. Designed for zero-shot generalization across diverse domains, SAM can generate segmentation masks based on user-provided prompts, such as bounding boxes or points, without requiring task-specific training [83]. This flexibility has made SAM a potential candidate for medical image segmentation, despite being originally trained on natural images.

Yet, adapting SAM to specialized medical applications presents challenges. The primary limitation lies in the domain gap between natural and medical images, which often leads to suboptimal performance in complex cases. One common solution is fine-tuning SAM on medical datasets, as demonstrated by MedSAM, which has been optimized for segmentation tasks in medical imaging [105]. While fine-tuning improves accuracy, it demands substantial computational resources and annotated medical datasets, which may not be readily available to many healthcare institutions.

A more practical alternative is **training-free** adaptation, where SAM is used as-is, guided by suitable prompt designs. However, selecting effective prompts is non-trivial. Simply using points or bounding boxes may not yield optimal performance, and providing multiple inputs remains time-consuming, limiting efficiency gains in clinical workflows. A promising alternative is leveraging **eye-gaze data** from low-cost commercial eye trackers as implicit prompts for segmentation [8]. Gaze constrains the model to attend within regions a clinician fixates on, which counterfactually excludes the regions where shortcut learning typically arises (Section 2.2). The prompt and the explanation become the same object by construction. Eye tracking also offers a natural and intuitive interaction, reducing cognitive and manual effort.

The optimal integration of gaze data into segmentation pipelines remains an open challenge. Previous works have explored the feasibility of combining eye-tracking with SAM for medical image segmentation. In particular, [166] proposed a straightforward usage of eye-gaze data to drive SAM-based segmentation. However, this approach resulted in suboptimal performance, yielding lower accuracy compared to traditional point-based interactions. This highlights the need for more effective strategies to leverage gaze data for segmentation tasks.

This work investigates different strategies for gaze-based prompting, ranging from directly using raw gaze points as input prompts to processing gaze data into structured forms such as heatmaps or fixation-based bounding boxes or combinations thereof. Through this exploration, we identify a simple yet effective approach for integrating eye-gaze data into a training-free pipeline, achieving performance comparable to several SAM-based methods, but without requiring additional model training or substantial computational resources.

We evaluate our training-free strategy on two clinically relevant datasets: polyp segmentation in colonoscopy images from Kvasir-SEG [75] and prostate segmentation using NCI-ISBI 2013 [14]. We compare these strategies against SAM variants fine-tuned on medical data, and examine which properties of the conversion determine whether gaze is usable as a prompt at all.

3.2.1 Related Work

Medical image segmentation has long relied on fully supervised deep learning models, with U-Net [135] and its variants, such as U-Net++ [192] and nnU-Net [73], emerging as dominant architectures. These models employ an encoder-decoder architecture to extract hierarchical features and generate detailed segmentation masks. While highly effective, they require large annotated datasets, which are often difficult to obtain in medical imaging due to the time and expertise required for manual annotation. To address these challenges, self-supervised learning and transformer-based architectures [19] have been explored, offering improvements in generalization and feature extraction. However, these models still require fine-tuning on medical data, making their deployment in real-world applications costly and impractical. The recent emergence of foundation models, such as SAM [83], offers a potential alternative by exploiting large-scale pretraining for zero-shot segmentation.

SAM was designed as a general-purpose segmentation model, capable of generating masks from user-defined prompts such as points or bounding boxes. While its zero-shot capabilities have demonstrated success in natural image segmentation, several studies have shown that SAM struggles when applied directly to medical images [105]. This limitation is primarily due to the significant domain gap between natural and medical images, with challenges such as low contrast, complex anatomical structures, and poorly defined boundaries.

To improve the performance of SAM in medical applications, several adaptations have been proposed. *Fine-tuning-based approaches*, such as MedSAM [105] and SAMed [186], refine SAM’s weights using large annotated medical datasets, significantly enhancing segmentation accuracy.

Another line of research has explored *architectural modifications* to SAM to better capture

medical image features. AutoSAM [67], modify the model’s architecture by introducing alternative prediction heads, reducing reliance on explicit user prompts. SAM-Adapter [27] introduces a lightweight module to inject domain-specific knowledge, while SAMUS [99] integrates a CNN-based encoder alongside SAM’s transformer backbone. Other works focus on *optimizing prompt-based* interactions to improve SAM’s segmentation results. SurgicalSAM [181] replaces manual prompts with prototype-based class embeddings, while SAM-Path [185] and Polyp-SAM++ [13] explore class-based and text-based prompts, respectively, to refine segmentation accuracy. These methods reduce user effort but still require specific task-based tuning, making them less adaptable to general medical applications. Despite these improvements, all the approaches discussed above require either fine-tuning, model modifications, or specialized prompting strategies, introducing complexity and computational overhead.

Eye tracking has been explored as an alternative input modality for segmentation tasks, offering a more intuitive interaction than traditional manual annotations. [8] investigates the use of gaze fixations and heatmaps as segmentation prompts, demonstrating their potential as an alternative to standard input methods. However, this approach involves fine-tuning SAM’s prompt encoders on gaze data, requiring additional training and adaptation.

A different approach is taken by GazeSAM [166], which integrates real-time eye-tracking as an interactive prompt for SAM. Instead of manually placing points or bounding boxes, GazeSAM converts raw gaze fixations into segmentation prompts, allowing users to generate masks simply by looking at the region of interest. However, while this method aligns with our training-free approach, it does not investigate how to best utilize gaze-based prompts for segmentation, leaving open the question of their optimal representation. In contrast, our work fully eliminates the need for fine-tuning while systematically analyzing different gaze-based prompting strategies, including raw gaze points, bounding boxes, and heatmaps, to determine the most effective way to guide SAM in a zero-shot, training-free manner.

3.2.2 Method

The foundational Model SAM consists of three main components (Section 2.1.2): an image encoder, a prompt encoder, and a mask decoder. The image encoder is a pre-trained Vision Transformer (ViT) available in different sizes (base, large, huge), which generates high-dimensional embeddings of the input image. The prompt encoder processes both sparse (points, bounding boxes, text) and dense (masks) inputs.

In this work, we explore the use of **eye-gaze fixations** recorded during image observation to generate both *sparse prompts* and *dense prompts* for guiding SAM zero-shot segmentation. In our case, sparse prompts consist of raw gaze points along with bounding boxes derived

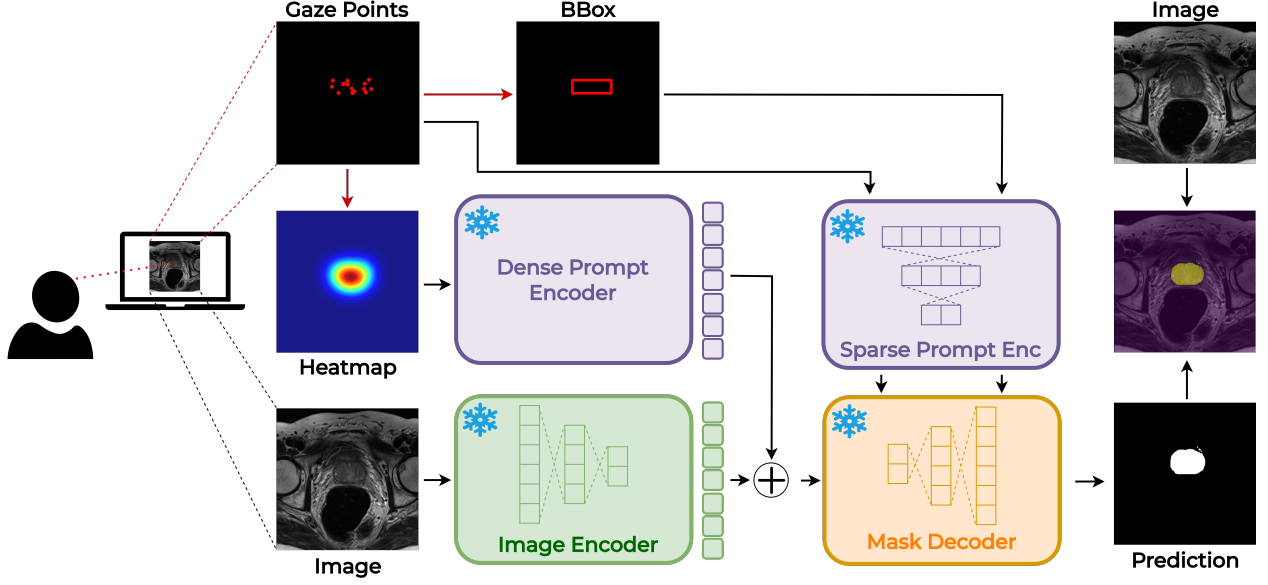


Figure 3.7: **Overview of our gaze-based prompting pipeline for SAM.** Our method converts recorded gaze fixations into bounding boxes (sparse prompts) and heatmaps (dense prompts), providing spatial constraints and attention maps to guide SAM for segmentation.

from the spatial distribution of fixations. Dense prompts, on the other hand, leverage gaze fixations to construct probabilistic heatmaps, capturing spatial attention as a continuous distribution. The entire pipeline, illustrating how SAM is conditioned on gaze-based prompts, is shown in Figure 3.7.

Sparse Prompt Generation from Eye Gaze Data. Given a set of recorded gaze points $\mathcal{G} = \{g_i\}_{i=1}^{N_g}$, where each $g_i = (x_i, y_i)$ represents a fixation coordinate in the image space, we define sparse prompts as the direct usage of these raw gaze points. Sparse prompting follows a point-based interaction scheme, where each gaze point serves as an individual prompt for the segmentation model.

Additionally, we derive bounding box prompts from the gaze points by computing the axis-aligned bounding box:

$$B = [x_{\min}, y_{\min}, x_{\max}, y_{\max}], \quad (3.1)$$

where $x_{\min} = \min_i(x_i)$, $y_{\min} = \min_i(y_i)$, $x_{\max} = \max_i(x_i)$, $y_{\max} = \max_i(y_i)$.

This bounding box serves as a spatial constraint for the segmentation model, encapsulating the region of interest indicated by the gaze fixations.

Dense Prompt Generation via Gaze Heatmaps. To incorporate gaze information as a continuous spatial distribution, we generate a probabilistic heatmap from the gaze points. The heatmap is constructed as a probability density function (PDF) over the image domain $\Omega \subset \mathbb{R}^2$, defined as:

$$H(x, y) = \sum_{i=1}^{N_g} w_i \cdot \mathcal{N}((x, y) \mid g_i, \Sigma), \quad (3.2)$$

where $\mathcal{N}((x, y) \mid g_i, \Sigma)$ represents a bivariate Gaussian function centered at gaze point g_i :

$$\mathcal{N}((x, y) \mid g_i, \Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} e^{-\frac{1}{2}(p-g_i)^\top \Sigma^{-1}(p-g_i)}, \quad (3.3)$$

where $p = (x, y)$, and $\Sigma = \sigma^2 I$ is the isotropic covariance matrix controlling the spatial spread of the heatmap. The weight w_i corresponds to the fixation duration at gaze point g_i , normalizing the influence of each fixation in the overall distribution.

The resulting heatmap is then normalized to:

$$\tilde{H}(x, y) = \frac{H(x, y)}{\max(H(x, y))}, \quad (3.4)$$

ensuring that all values are scaled between 0 and 1, and resized to a standard resolution $\Omega' = (h', w')$ for input to the segmentation model.

The Gaussian kernel variance σ^2 is a key parameter that determines the spatial spread of the heatmap and should be selected based on the expected gaze dispersion. The heatmap resolution (h', w') should be chosen to match the input resolution of the segmentation model.

3.2.3 Experimental Results

Datasets

We evaluate our training-free strategy on two clinically relevant datasets, Kvasir-SEG [75] for polyp segmentation in colonoscopy images, and NCI-ISBI 2013 [14] for prostate segmentation in T2-weighted MRI, with paired gaze annotations collected by Zhong et al. [189] (Section 2.6).

Experimental setup

We process each input slice by first converting it to RGB and resizing it to 1024×1024 , thus aligning with the default input dimensions of the SAM encoder. For Kvasir-SEG, the polyp segmentation masks are binarized at a threshold of 0.5. All images are normalized to the range $[0, 1]$. Whenever volumetric data are involved (e.g., in NCI-ISBI 2013), we apply our pipeline on a slice-by-slice basis.

We employ two variants of SAM, SAM_b (base, ViT with 12 attention layers) and SAM_h (huge, ViT with 32 attention layers), to examine the impact of model capacity on segmentation performance in a training-free setting. Additionally, we compare with MedSAM [105] a specialized version of SAM_{base} fine-tuned on medical images. All methods are provided with the same gaze-derived prompts (bounding boxes, points, heatmaps) to ensure a consistent basis of comparison.

For quantitative assessment, we use two widely recognized segmentation metrics: the Dice score and the Intersection over Union (IoU).

All experiments are performed in a *training-free* manner: no additional fine-tuning or domain adaptation is applied to the pre-trained weights. Specifically, SAM’s image encoder and mask decoder remain fixed; gaze data generate prompts that are processed by the SAM prompt encoder without parameter updates. We use publicly available checkpoints for both SAM and MedSAM.

Evaluation of Eye-Gaze Prompting Strategies

To assess the effectiveness of different eye-gaze-based prompts for medical image segmentation, we evaluate three strategies:

- **Gaze Points (GP)** – using eye-gaze fixations as direct sparse prompts;
- **Bounding Box (BB) Around Gaze Points** – computing the minimal bounding box enclosing all gaze points to serve as a spatial constraint;

- **Bounding Box with Heatmap (BB + HM)** – incorporating a Gaussian-based heatmap derived from gaze points as an additional dense prompt alongside the bounding box.

We also explored various combinations of these strategies, but none led to meaningful improvements in segmentation performance. Therefore, we focus our evaluation on these three approaches, comparing them against existing methods on Kvasir-SEG for polyp segmentation and NCI-ISBI 2013 for prostate segmentation.

For Kvasir-SEG (Table 3.4), using gaze points alone achieves reasonable performance. Using only a bounding box around the gaze points slightly reduces mDICE and mIoU for SAM_b and SAM_h while improving performance for MedSAM. However, the combination of bounding boxes and heatmaps, i.e., **BB + HM**, significantly improves segmentation accuracy across all models, with SAM_h achieving the highest performance (mDICE: 0.881, mIoU: 0.802).

For NCI-ISBI 2013, Table 3.5 shows that direct gaze points yield moderate segmentation quality, (especially with SAM_h), while using a bounding box notably improves performance for MedSAM. The **BB + HM** combination further enhances performance, particularly for SAM_h , which achieves the best results (mDICE: 0.909, mIoU: 0.885).

These findings highlight that while gaze-based prompts alone provide useful segmentation cues, leveraging both spatial constraints and probabilistic attention maps derived from eye-gaze data through the integration of bounding boxes and heatmaps enhances segmentation accuracy.

Table 3.4: **Segmentation performance (mDICE and mIoU) for gaze-based prompts on Kvasir-SEG.** GP: gaze fixations, BB: Bounding box around gaze points, and **BB + HM**: BB with a Gaussian heatmap.

Prompt Type	Metric	SAM_b	SAM_h	MedSAM
GP	mDICE	0.725	0.827	0.516
	mIoU	0.625	0.752	0.424
BB	mDICE	0.709	0.786	0.739
	mIoU	0.599	0.688	0.655
BB + HM	mDICE	0.779	0.881	0.656
	mIoU	0.674	0.802	0.556

One notable observation from the results is that SAM_b consistently outperforms MedSAM (fine-tuning of SAM_b on extensive medical images) across all gaze-based prompting strategies, despite MedSAM being specifically fine-tuned on medical images. This outcome

Table 3.5: **Segmentation performance (mDICE and mIoU) for gaze-based prompts on NCI-ISBI 2013.** GP: gaze fixations, BB: Bounding box around gaze points, and BB + HM: BB with a Gaussian heatmap.

Prompt Type	Metric	SAM_b	SAM_h	MedSAM
GP	mDICE	0.781	0.833	0.512
	mIoU	0.683	0.748	0.374
BB	mDICE	0.814	0.822	0.885
	mIoU	0.703	0.713	0.816
BB + HM	mDICE	0.815	0.909	0.735
	mIoU	0.707	0.885	0.638

may seem counterintuitive, as one would expect a domain-adapted model to perform better in medical segmentation tasks.

These findings suggest that while domain-specific fine-tuning can enhance performance for standard segmentation tasks, foundation models like SAM retain a significant advantage in prompt-driven, training-free settings. Their ability to generalize across different domains, coupled with superior contour identification capabilities, makes them particularly well-suited for integrating gaze-based prompting strategies in medical image segmentation. Finally, as expected, SAM_h outperforms SAM_b on all settings due to its higher number of parameters (about $7\times$ more parameters than SAM_b).

Figure 3.8 presents segmentation examples from the Kvasir-SEG and NCI-ISBI 2013 datasets using gaze-based prompts and SAM_h . Each example includes the ground truth segmentation, recorded gaze points, a heatmap, and a bounding box enclosing the fixations. The combination of heatmaps and bounding boxes effectively guides the segmentation model, yielding accurate results while minimizing manual input. These visualizations confirm that gaze-based prompting provides a practical and efficient alternative for medical image segmentation.

Quantitative Comparison with State-of-the-Art Methods

To evaluate the effectiveness of our approach, we compare our gaze-based prompting strategy with several state-of-the-art SAM-based adaptations. Tables 3.6 and 3.7 summarize the segmentation performance on Kvasir-SEG and NCI-ISBI 2013, respectively, including models fine-tuned on medical datasets, prompt-driven approaches, and fully automated segmentation methods.

Table 3.6 presents the segmentation performance on Kvasir-SEG. Among the fine-tuned models, ASPS (mDICE: 0.920, mIoU: 0.858) achieves the highest performance without struc-

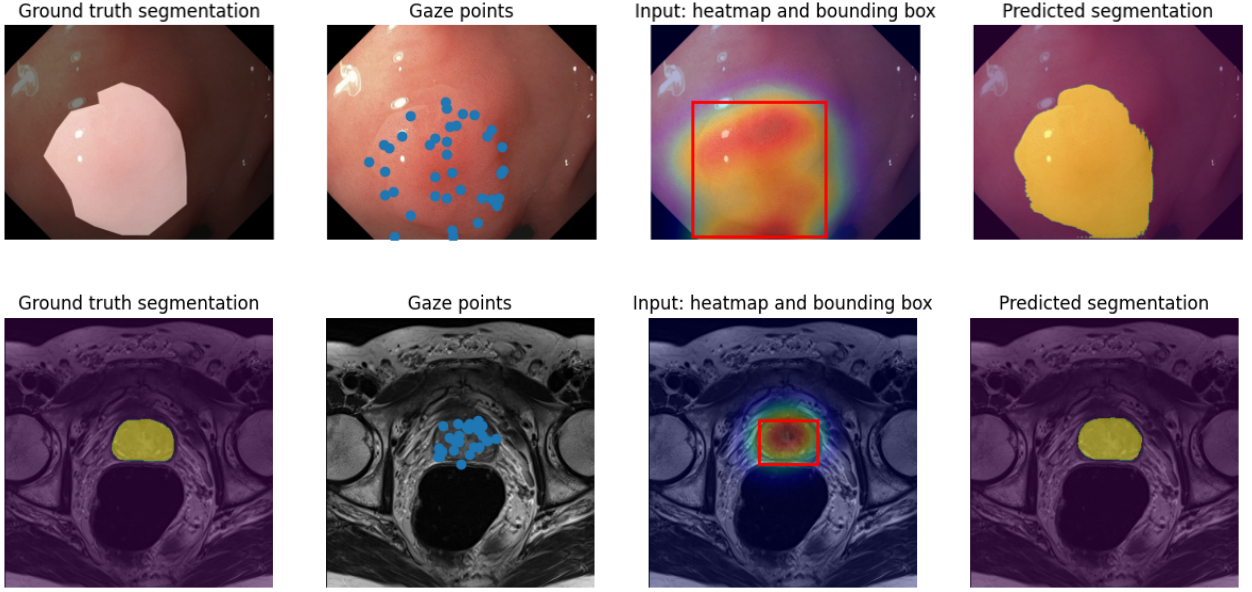


Figure 3.8: **Visualization of gaze-based segmentation inputs and results on Kvasir-SEG and NCI-ISBI 2013.** Each dataset sample includes the ground truth segmentation, gaze points recorded during image observation, a heatmap generated from the gaze distribution, and a bounding box enclosing the fixations. The combination of the heatmap and bounding box as input prompts enables accurate segmentation predictions, which are shown in the last column of the image.

tured prompts, while PolypSAM and PolypSAM++ (mDICE: 0.900, mIoU: 0.860) perform well due to their specific adaptation to polyp segmentation. Other fine-tuned methods, such as MedSAM, SAM-Adapter, and SAMUS, also show good results, benefiting from medical dataset training.

Notably, our gaze-based prompting approach achieves competitive results (mDICE: 0.881, mIoU: 0.802), surpassing most of the fine-tuned models on mDICE, including MedSAM (mDICE: 0.862), SAMUS (mDICE: 0.859), SAM-Adapter (mDICE: 0.847) and SAM-Path (mDICE: 0.828), while obtaining the second best performance in terms of mIoU. This demonstrates the effectiveness of structured prompts compared to model modification or additional training, reinforcing the feasibility of eye-gaze prompting as a scalable, efficient, and low-effort solution for medical image segmentation.

Table 3.7 presents segmentation results on NCI-ISBI 2013, we observed a similar trend as in the previous case, with our method achieving performance on par with state of the art approaches that, however, have benefitted from extensive training on the target data. This confirms the potential of gaze-driven prompting and its generalization capabilities across datasets as an effective alternative for obtaining accurate segmentation without requiring

Table 3.6: **Segmentation performance on Kvasir-SEG**. Comparison of our method with state-of-the-art SAM-based models, differing in fine-tuning on medical data and structured prompt usage. Competitors’ performance are from [109].

Model	Fine-tuning	Prompt-based	mDICE	mIoU
SAM-Adapter [27]	✓	✗	0.847	0.763
AutoSAM [67]	✓	✗	0.784	0.675
SAM-Path [185]	✓	✗	0.828	0.730
SAMed [186]	✓	✗	0.459	0.300
SAMUS [99]	✓	✗	0.859	0.763
SurgicalSAM [181]	✓	Class prompt	0.740	0.597
MedSAM [105]	✓	Bounding box	0.862	0.795
PolypSAM [97]	✓	Bounding box	0.900	0.860
PolypSAM++ [13]	✓	Text prompt	0.900	0.860
ASPS [93]	✓	✗	0.920	0.858
Ours	✗	Eye gaze	0.881	0.802

additional training.

Table 3.7: **Segmentation performance on NCI-ISBI 2013**. Comparison of our method with state-of-the-art SAM-based models. MedSAM results are from our experiments, UnCLE-SAM from [131], and AutoSAM, 3DSAM-adapter, SAMed_h, and MA-SAM from [25].¹

Model	Training	Prompt-based	mDICE
AutoSAM [67]	✓	✗	0.867
3DSAM-adapter [50]	✓	Point	0.879
SAMed_h [186]	✓	✗	0.921
MA-SAM [25]	✓	✗	0.940
UnCLE-SAM [131]	✓	✗	0.960
MedSAM [105]	✓	Bounding box	0.934
Ours	✗	Eye gaze	0.909

Overall, our results demonstrate that gaze-based prompting can achieve segmentation accuracy on par with state-of-the-art fine-tuned models, despite requiring no dataset-specific training. Remarkably, our approach outperforms multiple models explicitly trained on the target datasets, underscoring the potential of effective prompting strategies to compensate for the lack of fine-tuning.

¹MA-SAM scores are averaged across Site A and Site B, corresponding to the dataset used in this work.

3.2.4 Discussion

Gaze-based prompting matched fine-tuned SAM variants on Kvasir-SEG and NCI-ISBI 2013 without any gradient updates or medical training data. The combination of a fixation-derived bounding box and a dense gaze heatmap was the configuration that worked, with raw point prompts performing worst. This pattern is consistent with the prior GazeSAM [166] negative result on raw points, and suggests that the value of gaze as a prompt depends substantially on how it is converted into a SAM-compatible representation.

A specific empirical observation deserves discussion. SAM with gaze-prompting consistently outperformed MedSAM, despite MedSAM being fine-tuned on medical data. A plausible explanation involves the training pipelines of the two models. SAM was trained on the SA-1B dataset, in which the masks themselves were collected through a human-in-the-loop annotation process. Human annotators chose what counted as an object, where to click, which boundaries to correct, and which structures were worth annotating. The training prompts during SAM’s optimization were largely sampled algorithmically from the resulting masks, but the masks themselves embed human attention and objectness priors [83]. An expert gaze prompt is structurally similar in that it reflects where a human directed visual attention. With its generalization capability preserved, SAM responds to gaze-derived prompts naturally because they match the fine-grained input signal seen during training, and by extension a human attention signal it was exposed to indirectly through its training data, despite the domain shift to medical imaging. MedSAM’s fine-tuning would have disrupted this match by only sampling random-jitter bounding boxes from ground-truth masks [105], narrowing the prompt input distribution to coarse boxes that more weakly represent the human-attention signal. This remains a hypothesis rather than a tested explanation.

The method delivers what fine-tuned medical SAM variants deliver – competitive segmentation accuracy – without the gradient updates, the medical training data, or the annotation effort those variants require. Yet, the approach also has clear limitations: eye-tracking hardware must be present at acquisition, and calibration is required per session. Furthermore, a radiologist must be actively looking at the image at inference time, which limits any retrospective use.

The contribution of this section to the chapter’s *Knowing* argument is that gaze-prompting makes spatial attention a structural part of the prediction rather than something reconstructed after the fact. A clinician can inspect the fixation map before accepting the segmentation, and the model is constrained to attend within the regions the clinician fixated on, which by the counterfactual argument in Section 2.2 would exclude the regions where shortcut learning typically arises. The hypothesis raised above adds a layer to this framing.

If SAM’s training data encodes indirect human attention and objectness priors through its annotation pipeline, then gaze-prompting at inference is not a foreign signal but one that aligns with the implicit priors the model already carries. One qualification bounds this framing. Gaze is used here as a proxy for clinical reasoning, justified by the counterfactual observation that on average the fixated regions are the regions whose removal would change the reading. What the method makes inspectable is therefore where the evidence lies, and not the order in which it is gathered, its temporal structure, or its relation to the reported finding. Whether an observer attends a lesion or its margin is likewise not expressible in a purely spatial representation. Spatial evidence is a first-order component of an explanation rather than an explanation itself, and the contribution here is that this component is made structural rather than reconstructed post-hoc.

3.2.5 Publications

Giulia Pötgens, Rutger Hendrix, Cian Scannell, Ulas Bagci, Concetto Spampinato, and Federica Proietto Salanitri. Gaze-guided medical image segmentation: A training-free approach using SAM foundation model. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–7. IEEE, 2025. doi.org/10.1109/EMBC58623.2025.11253157.

3.3 Predictive Gaze-Driven Prompts for Medical Image Segmentation using Frozen SAM

Medical image segmentation is a cornerstone of computer-aided diagnosis, but acquiring the necessary pixel-level expert annotations is prohibitively expensive and time-consuming. Convolutional encoder-decoder architectures such as U-Net [135] and its variants [193, 81] have been the standard for over a decade, with strong domain-specific results on volumetric and dynamic anatomical structures ranging from craniomaxillofacial CT scans [139] to 4D cardiac MRI sequences and endoscopy videos [178, 187]. However, these fully supervised models tend to generalize poorly across modalities and demand extensive labeled data, which is scarce in medical settings due to privacy regulations and expert annotation costs. Foundation models, particularly the Segment Anything Model (SAM) [83], offer zero-shot generalization on natural images but suffer notable performance drops on radiological data due to a substantial domain shift [110]. Classic fine-tuning on medical datasets, as in MedSAM [105] and related variants, mitigates this gap at the cost of substantial computational resources, large annotated medical corpora, and the risk of overfitting on limited samples. Beyond compute, fine-tuning also constrains the model to the training distribution and reduces prompt flexibility. Training-free prompt engineering, where SAM remains frozen and adaptation is performed on the input side, is therefore an attractive alternative when annotated medical data is limited.

A particularly promising line within training-free prompting is guidance from clinician visual attention. Recorded gaze, captured by an eye-tracker during image interpretation, has been shown to rival manual annotations as a SAM prompt and in some settings to outperform fine-tuned medical SAM variants [166, 82]. Section 3.2 develops this direction by converting recorded fixations into structured sparse and dense prompts, achieving segmentation accuracy comparable to medical SAM variants without any gradient updates or medical training data. The bottleneck shared by all recorded-gaze methods is hardware. Eye-tracking equipment must be present at acquisition, calibration is required per session, and a radiologist must be actively looking at the image at inference time. These constraints prevent the approach from scaling to retrospective dataset curation or to clinical workflows in which gaze is not being recorded.

This section investigates whether the human-attention learning signal that recorded gaze provides can be preserved when the hardware is removed. Recorded clinician gaze worked well as a SAM prompt in Section 3.2. One plausible reason discussed there is that SAM’s training data already exposed it to fine-grained point-like prompts and, by extension, to the

human attention and objectness priors that the annotation pipeline indirectly encoded. If a saliency predictor trained on recorded gaze can produce maps close enough to real expert attention, the same alignment may hold without the hardware in the loop. Predicted gaze is by construction an approximation of recorded gaze, and replacing the real signal with a learned one moves the method one step away from the directly human-grounded foundation that motivated Section 3.2. What survives the substitution is the inspectability property: the predicted saliency map is rendered and can be reviewed before accepting the segmentation, so a clinician can verify that the model was directed at the right anatomy even though none was in the loop at inference time.

In this work a saliency predictor is trained on recorded clinician gaze and combined with the hybrid sparse-and-dense prompt-generation pipeline of Section 3.2. The resulting fully automated pipeline is evaluated on the same datasets without fine-tuning SAM, and a Sensitivity analysis quantifies how prompt-location noise propagates into segmentation error. The contributions are:

- A hardware-free saliency predictor that approximates expert visual attention on medical images, trained on recorded clinician gaze.
- A fully automated prompting pipeline that converts the predicted saliency into sparse and dense cues for a frozen SAM.
- Zero-shot segmentation performance on prostate MRI (NCI-ISBI 2013 [14]) and colonoscopy polyp segmentation (Kvasir-SEG [75] and CVC-300 [164]) comparable to medical-domain fine-tuned SAM variants.

3.3.1 Method

Our approach simulates radiologist visual attention to guide SAM without requiring physical eye-tracking hardware or computationally expensive fine-tuning. Starting from a medical image I , our pipeline predicts fixation probabilities that capture where an expert would naturally look. These synthetic fixations are then converted into explicit visual prompts—namely sparse points, bounding boxes, and dense heatmaps—to steer a SAM toward accurate segmentations. An overview of the proposed framework is illustrated in Figure 3.9. To achieve this, we employ a two-phase training strategy for the fixation predictor, followed by a prompt generation module.

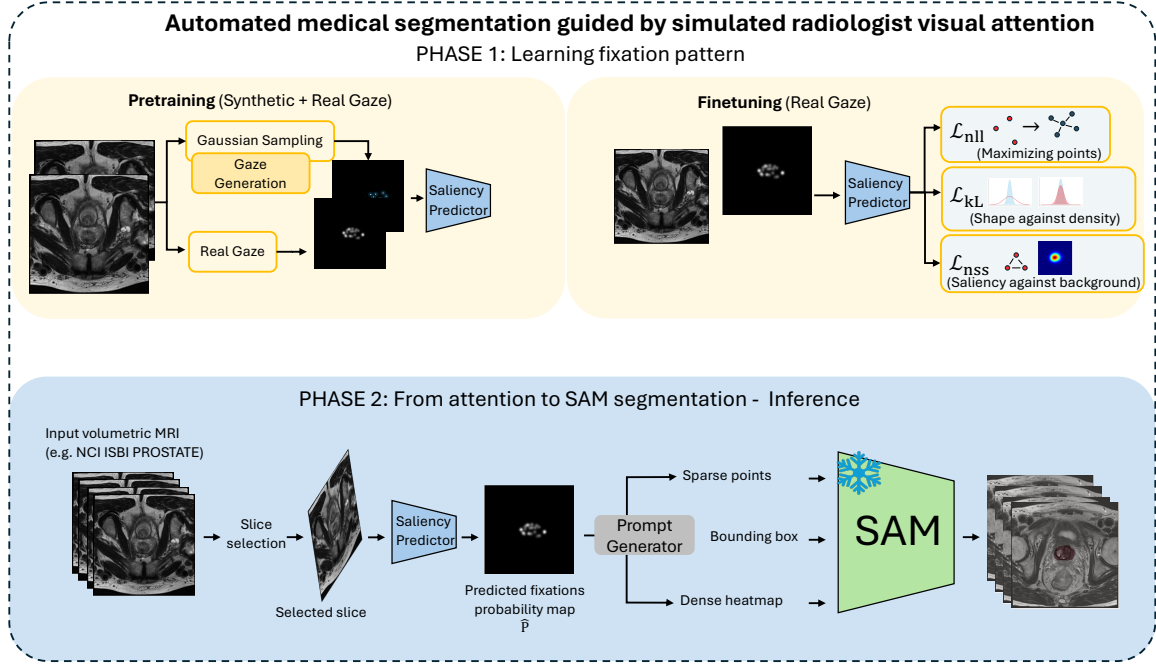


Figure 3.9: Overview of our pipeline. The saliency predictor is first pretrained on real and synthetic gaze from anatomical centroids, then finetuned on real clinician fixations using NLL, KL, and NSS losses. At inference, the predicted fixation map $\hat{P}$ is converted into sparse points, bounding boxes, and dense heatmaps to prompt a frozen SAM without any weight updates or eye-tracking hardware.

Saliency-Driven Fixation Prediction

We represent fixations as a spatial probability map $\hat{P} \in [0, 1]^{H \times W}$, where the intensity at each pixel reflects the accumulated fixation duration, mimicking how radiologists dwell longer on diagnostically relevant regions. Because collecting large-scale human gaze data is impractical, we bootstrap supervision from unlabeled images using a *synthetic gaze* strategy. For each image, we pre-compute anatomical anchor points (e.g., organ centroids or lesion boundaries) and generate synthetic fixation distributions P^{syn} modeled as spatial Gaussian mixtures centered at these anchors. The spread σ_{syn} adapts naturally to the target structure ($\sigma_{\text{syn}} \propto 1/\sqrt{M}$ for M anchors).

As the saliency predictor, we use a U-Net [135]. The network is first pretrained to learn coarse anatomical priors by minimizing the Kullback-Leibler (KL) divergence between

smoothed versions of the synthetic target and the prediction:

$$\mathcal{L}_{\text{kl}}^{\text{pre}} = \text{KL}\left(P_{\text{soft}}^{\text{syn}} \parallel \hat{P}_{\text{soft}}\right), \quad (3.5)$$

with $\hat{P}_{\text{soft}} = G_{\sigma_{\text{blur}}} * \hat{P} / Z_{\text{pred}}$ the smoothed prediction, $P_{\text{soft}}^{\text{syn}} = G_{\sigma_{\text{blur}}} * P^{\text{syn}} / Z_{\text{syn}}$ the smoothed synthetic target, and $G_{\sigma_{\text{blur}}}$ a Gaussian kernel of standard deviation σ_{blur} . Crucially, this synthetic sampling allows the incorporation of non-gaze-annotated data, resulting in a larger pretraining set, and broadening the anatomical and acquisition diversity seen by the saliency predictor.

Once coarse attention is established, we refine the predictions to clinical accuracy using real gaze data. Let $\{p_i\}_{i=1}^N$ denote the ground-truth fixation coordinates. The refinement phase combines three complementary objectives. First, we maximize the probability at true fixation locations using a Negative Log-Likelihood (NLL) loss:

$$\mathcal{L}_{\text{nll}} = -\frac{1}{N} \sum_{i=1}^N \log \hat{P}(p_i). \quad (3.6)$$

Second, the same KL divergence used in pretraining is applied between the smoothed binary fixation mask $P_{\text{soft}}^{\text{gt}} = G_{\sigma_{\text{blur}}} * P^{\text{gt}} / Z_{\text{gt}}$ and the smoothed prediction:

$$\mathcal{L}_{\text{kl}}^{\text{fine}} = \text{KL}\left(P_{\text{soft}}^{\text{gt}} \parallel \hat{P}_{\text{soft}}\right). \quad (3.7)$$

Finally, we employ a Normalized Scanpath Saliency (NSS) objective $\mathcal{L}_{\text{nss}}$ to encourage predicted fixations to strongly stand out against the background:

$$\mathcal{L}_{\text{nss}} = -\frac{1}{N} \sum_{i=1}^N \frac{\hat{P}(p_i) - \mu_{\hat{P}}}{\sigma_{\hat{P}}}, \quad (3.8)$$

where $\mu_{\hat{P}}$ and $\sigma_{\hat{P}}$ are the mean and standard deviation of the prediction across all pixels in the image. The total fine-tuning loss combines these objectives as a weighted sum:

$$\mathcal{L}_{\text{fine}} = \mathcal{L}_{\text{nll}} + 0.01 \mathcal{L}_{\text{kl}}^{\text{fine}} + 0.1 \mathcal{L}_{\text{nss}}. \quad (3.9)$$

The weights are chosen so that the three contributions are of comparable magnitude, with the likelihood term leading. Since the KL term already established the coarse spatial prior during pre-training, it only has to anchor that prior against drift rather than learn it, so its contribution is the smallest in practice, with the NSS term somewhat above it.

Automated Prompt Generation and Segmentation

At inference time, the trained predictor generates a high-quality fixation map $\hat{P}$. To convert this continuous distribution into explicit SAM prompts, we first extract the top- K highest-probability locations where $\hat{P} > \tau$ with $\tau = \frac{1}{H \cdot W}$. These coordinates act directly as **point prompts**. To maximize robustness, we aggregate these point candidates into a **bounding box prompt**. We employ a variance-based box scaling strategy: regions with high predictive focus produce tightly clustered points with low spatial variance $\text{Var}(\mathbf{p})$, yielding compact minimum/maximum bounding boundaries. To account for prediction uncertainty, we expand each bounding box dynamically by adding a padding of $\delta = \frac{\sigma}{2} + 1$ pixels. Thus, confident, tight clusters receive minimal padding, while diffuse predictions yield more conservative boxes. Concurrently, the fixation probabilities are mapped into dense **heatmap prompts** to provide rich spatial context.

Volumetric Slice Selection.

To extend our framework from 2D imaging to volumetric MRI data (e.g., the NCI-ISBI prostate dataset), our pipeline integrates a dedicated slice-selection classifier. When processing an MRI volume containing T slices, forcing a heavy foundation model like SAM to segment every slice—even those where the target organ is absent—would not only introduce false positives but also incur unnecessary computational cost.

To address this, we train a slice-level classifier g_ϕ that reuses the frozen encoder of our U-Net backbone as a feature extractor and predicts the probability $\hat{y}_t \in [0, 1]$ of organ presence for each slice $t \in \{1, \dots, T\}$. We preserve the U-Net’s spatial feature maps and aggregate them with 3D convolutions, so that the prediction for slice t depends jointly on local spatial appearance and axial context rather than on slice t in isolation. The model is optimized using a standard Binary Cross-Entropy (BCE) loss:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{T} \sum_{t=1}^T \left[y_t \log(\hat{y}_t) + (1 - y_t) \log(1 - \hat{y}_t) \right], \quad (3.10)$$

where $y_t \in \{0, 1\}$ is the ground-truth presence label for slice t . To enhance sensitivity at the organ’s upper and lower apices, where the structure is faint and easily confused with adjacent empty slices, the head incorporates a local self-attention module along the slice axis. The module operates over a fixed window of $2w+1$ neighboring slices centered on t , allowing the classifier to compare slice t to its immediate neighbors and respond to appearance/disappearance transitions rather than global volumetric appearance.

At inference time, we threshold the predicted probabilities at $\tau = 0.5$. Slices classified as

organ-absent are immediately skipped and automatically assigned an empty segmentation mask ($M_t = \emptyset$). For valid slices, the generated points, scaled bounding boxes, and heatmaps are passed to the frozen SAM. By discarding irrelevant slices upstream, this module reduces the number of forward passes required by the foundation model while preserving automated segmentation across 3D volumes.

3.3.2 Experimental Results

Training and Evaluation Settings

We evaluate our framework on two clinical segmentation tasks: prostate MRI segmentation and polyp segmentation. Real clinician gaze was collected for two of the five datasets used in this work (NCI-ISBI 2013 and Kvasir-SEG); the remaining three (PROMISE12, CVC-ClinicDB, CVC-300) contain no eye-tracking annotations. Table 3.8 summarizes the role of each dataset across the pipeline.

Dataset	Modality	Real gaze	Pretraining	Finetuning
NCI-ISBI 2013 [14]	Prostate MRI	✓	✓	✓
PROMISE12 [100]	Prostate MRI	–	✓	–
Kvasir-SEG [75]	Endoscopy	✓	✓	✓
CVC-ClinicDB [11]	Endoscopy	–	✓	–
CVC-300 [164]	Endoscopy	–	–	–

Table 3.8: Dataset usage overview

Prostate MRI segmentation. We use the **NCI-ISBI 2013** Challenge dataset [14], consisting of multi-scanner T2-weighted MRI volumes annotated for the central gland and peripheral zone, reflecting clinically relevant inter-scanner variability. We follow the MA-SAM test split, with all splits performed at the *patient level* to prevent leakage between training, validation, and test sets. For pretraining we combine NCI-ISBI with **PROMISE12** [100], which contributes three additional acquisition sites and no gaze annotations; PROMISE12 is used exclusively in the synthetic-gaze pretraining stage.

Polyp segmentation. We finetune on an 80/20 split of **Kvasir-SEG** [75], which provides both segmentation masks and clinician gaze. Pretraining is additionally supported by **CVC-ClinicDB** [11], a colonoscopy dataset of 612 frames extracted from 31 clinical sequences with pixel-level polyp annotations but no gaze data. Evaluation is performed on **CVC-300** [164], a held-out colonoscopy benchmark comprising 300 frames with challenging polyp appearances

(irregular shapes and heterogeneous textures); CVC-300 is never seen during training in any form and serves as a generalization test.

Unless otherwise specified, we use **SAM ViT-B** as the backbone. Performance is reported using the Dice Similarity Coefficient (DSC, %) and Intersection over Union (IoU, %). All experiments were conducted on a workstation equipped with a single NVIDIA A100 GPU. During the offline training phase, models are trained for 75 epochs with batch size B (set to 32 in the final version) and early stopping based on validation performance when applicable.

Quantitative Results

Tables 3.9 and 3.10 report the quantitative comparison with representative baselines on the NCI-ISBI 2013 prostate MRI dataset and the CVC-300 polyp dataset, respectively. On NCI-ISBI 2013, our predictive-fixation prompting achieves a DSC of 0.811, performing competitively while keeping SAM entirely **frozen** (i.e., without fine-tuning the foundation model) and requiring no manual prompts at inference time. Although supervised methods obtain higher scores, our results indicate that automatically generated fixation-based prompts can narrow the gap to adaptation-based approaches (e.g., MA-SAM at 0.816 DSC) without modifying SAM.

Table 3.9: Quantitative comparison on the NCI-ISBI 2013 Prostate Dataset.

Method	Training Required	DSC (%)	IoU (%)
MA-SAM [25]	✓	0.816	0.689
HardNet68 [23]	✓	<u>0.908</u>	<u>0.831</u>
AutoSAM [146]	✓	0.877	0.782
MSSAMLES [141]	✓	0.917	0.846
Ours (Pred. Gaze)	No	0.811	0.774

On CVC-300, our method attains a DSC of 0.839, improving over several SAM adaptation baselines such as SAM-Adapter (0.815) and remaining in the same range as other prompt-based/adapter-based approaches (e.g., AutoSAM at 0.829 and SAMPath at 0.844). Overall, these results support the effectiveness of predicted-fixation prompting as a practical way to deploy a frozen foundation model with competitive segmentation performance and reduced reliance on human interaction.

Table 3.10: Quantitative comparison on the CVC-300 Dataset.

Method	Training SAM	DSC (%)	IoU (%)
SAM-Adapter [27]	✓	0.815	0.725
AutoSAM [146]	✓	0.829	0.739
SAMPath [185]	✓	0.844	0.756
SAMed [186]	✓	0.332	0.202
SAMUS [99]	✓	0.859	0.760
SurgicalSAM [181]	✓	0.623	0.472
MedSAM [105]	✓	0.870	0.798
Polyp-sam [97]	✓	<u>0.905</u>	<u>0.860</u>
ASPS [93]	✓	0.919	0.852
Ours (Pred. Gaze)	No	0.839	0.741

In addition, our predicted fixations provide *intrinsic interpretability*: clinicians can directly inspect the predicted saliency map and fixation points used to construct the prompts, verifying whether the model focuses on clinically relevant anatomy.

Qualitative Results

Figure 3.10 compares the predicted point prompts against the recorded clinician fixations on three Kvasir-SEG and three NCI-ISBI samples. Although the predicted prompts are more sparse than the recorded fixations, their shape and location match the recorded distribution. The predicted points fall within the region the radiologist actually attended to and cluster around the lesion. This qualitative match supports the quantitative result that predicted-gaze prompts produce segmentation accuracy comparable to recorded-gaze prompts.

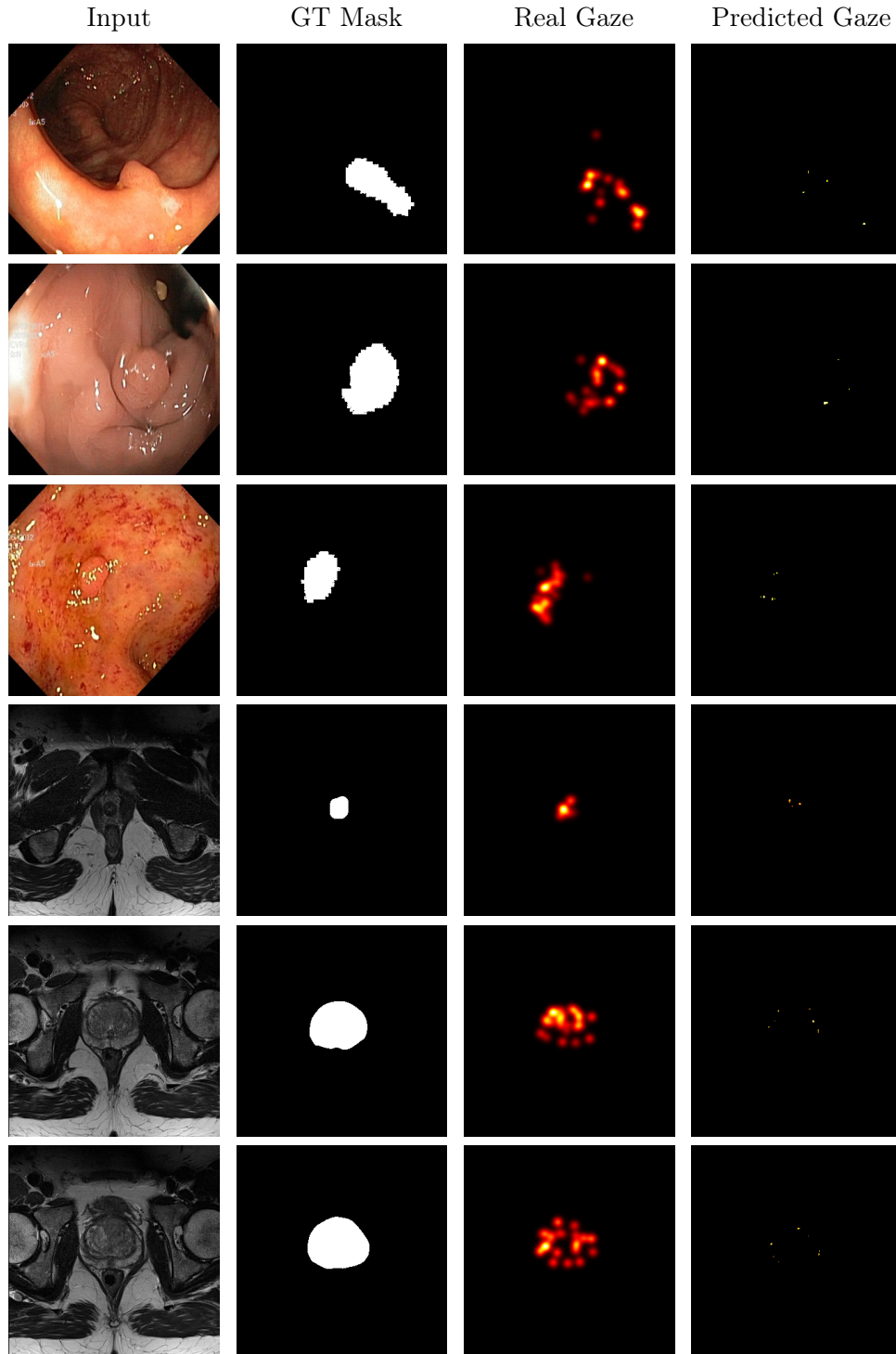


Figure 3.10: Qualitative comparison of predicted versus real clinician gaze on Kvasir-SEG (top three rows) and NCI-ISBI (bottom three rows). Columns show, from left to right: input image, ground-truth segmentation mask, real clinician fixation map (smoothed binary fixations), and the top-25 predicted point prompts after thresholding at $\tau = 1/(H \cdot W)$.

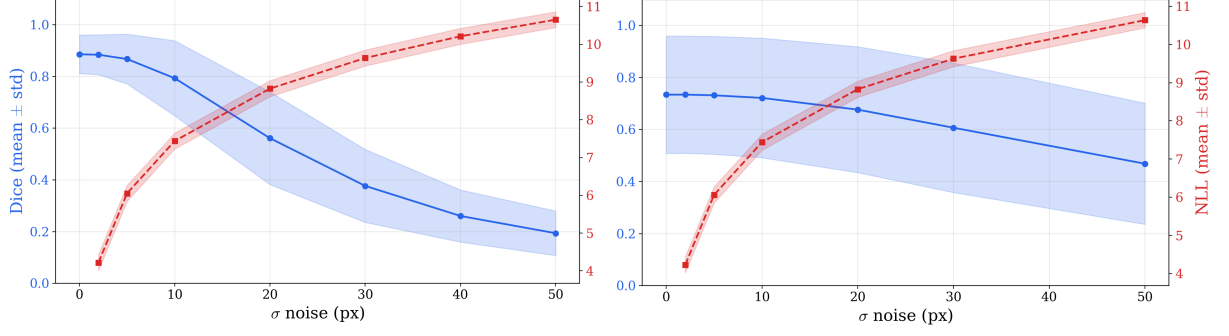


Figure 3.11: Prompt-noise robustness on NCI-ISBI 2013 prostate (left) and KVASIR-SEG/CVC-300 polyp (right). For each test image, we perturb the ground-truth gaze location by adding isotropic Gaussian noise $dx, dy \sim \mathcal{N}(0, \sigma^2)$, generate the corresponding prompts, and measure the resulting Dice score obtained by frozen SAM. We report Dice as a function of the perturbation error, quantified via the Negative Log-Likelihood (NLL) between the original and perturbed gaze under a Gaussian noise model. The predicted-gaze operating point is overlaid on each curve.

Prompt-Noise Robustness and Sensitivity Analysis

Since our framework relies on *predicted* fixations (rather than hardware-recorded gaze), it is important to quantify how spatial inaccuracies in point prompts affect the downstream segmentation produced by a frozen SAM. We therefore conduct a controlled sensitivity analysis by perturbing the ground-truth gaze location with isotropic Gaussian noise and measuring the resulting Dice score over N Monte Carlo trials at each noise level.

For each test sample, let $\mathbf{p}_{\text{gt}} = (x, y)$ denote the ground-truth gaze coordinates. We generate a noisy prompt location by injecting isotropic Gaussian noise:

$$\mathbf{p}_{\text{noisy}} = (x + dx, y + dy), \quad dx, dy \sim \mathcal{N}(0, \sigma^2). \quad (3.11)$$

For each σ , we construct the corresponding point-based prompts from $\mathbf{p}_{\text{noisy}}$ and record the Dice Similarity Coefficient (DSC) achieved by frozen SAM.

To express perturbation magnitude consistently with our training objective, we quantify the prompt displacement through the Negative Log-Likelihood (NLL) of the perturbed point under a Gaussian model centered at $\mathbf{p}_{\text{gt}}$:

$$\text{Error}_{\text{NLL}} = -\log P(\mathbf{p}_{\text{noisy}} \mid \mathbf{p}_{\text{gt}}), \quad (3.12)$$

where P models the spatial spread of fixations. Figure 3.11 plots DSC as a function of this NLL-based error for the NCI-ISBI 2013 prostate dataset and the polyp dataset, and overlays the operating region corresponding to the typical prediction error of our saliency model.

This analysis provides a direct quantitative link between gaze-prediction accuracy and segmentation quality, highlighting the level of prompt precision required to reliably guide a frozen SAM. On average, segmentation accuracy degrades smoothly with prompt displacement. Particularly for the polyp dataset, the spread in Dice at a given noise level is substantially wider, indicating that the segmentation response to prompt perturbation is more variable across the test set. The aggregate operating-point performance remains favorable, but this variance suggests that a per-case confidence signal would be valuable for identifying when the predicted prompt is unlikely to suffice.

3.3.3 Discussion

This section introduced a predicted-fixation prompting strategy that replaces hardware-recorded eye tracking with computationally predicted fixation points, while keeping SAM fully frozen. On NCI-ISBI 2013 prostate MRI, the pipeline achieves DSC 0.811 without any model fine-tuning or clinician interaction, comparable to fine-tuned adaptation methods such as MA-SAM (0.816). On CVC-300, the pipeline achieves DSC 0.839, outperforming several SAM adaptation baselines and remaining in the range of prompt-engineered methods that require training.

The prompt-noise robustness analysis qualifies the aggregate results. The prompt being the explanation rests on the operational assumption that a prompt that approximates where a clinician would attend yields a correct segmentation. The variance observed, especially on polyp, shows that this assumption holds on average but not uniformly: a prompt that approximates clinician attention is necessary for SAM to succeed but is not always sufficient. Variance across the test set does not scale simply with the noise level, and its causes are not isolated by the analysis: polyp morphology, target size relative to perturbation magnitude, and SAM’s handling of multi-point prompts whose centroid remains in-target may all contribute. The structural interpretability property of the method is therefore bounded by SAM’s per-case prompt sensitivity, not by the predictor’s accuracy alone.

The approach has two further limitations. The predictor is trained on gaze data from a single institution covering both datasets, so generalization to other institutions, modalities, or anatomies has not been tested and would require collecting gaze data from a broader clinical population. The pipeline also has no internal mechanism for detecting when the saliency prediction itself is unreliable, so when the predicted saliency fails to highlight the target the resulting prompts may not activate the correct mask and the pipeline produces an incorrect segmentation without flagging the failure. A natural extension would couple the saliency predictor with a predictive-uncertainty signal so that low-confidence predictions

trigger a fallback rather than commit to a single uncertain prompt. Possible fallbacks include multiple candidate prompts, recorded gaze where available, or human review.

This section completes the structural interpretability route of the *Knowing* chapter by removing the hardware dependency that bounded Section 3.2. Replacing recorded gaze with a learned saliency predictor moves the method one step away from the directly human-grounded foundation: the prompt is no longer what a clinician looked at, but what a model trained on clinician gaze predicts a clinician would look at. What survives the substitution is the inspectability property. The robustness analysis bounds what this inspection can deliver: agreement between the predicted map and the clinically expected attention region is necessary for SAM to segment correctly but is not always sufficient.

3.3.4 Publications

Rutger Hendrix, Giovanni Patanè, Giovanni Bellitto, Federica Proietto Salanitri, and Concetto Spampinato. Predictive gaze-driven prompts for medical image segmentation using Frozen SAM. Submitted to *MICCAI 2026*, under revision.

Chapter 4

Doubting: Uncertainty-Aware Federated Learning

This chapter argues that uncertainty, efficiently implemented as a structural property, can serve as a learning signal in the federated setting, governing what is shared between clients in addition to flagging unreliable predictions. The previous chapter treated uncertainty as a deferral cue: a per-prediction estimate that warrants clinician takeover when the model should not commit. That concept remains, but is applied inversely as a signal for whether a client’s contribution will help the global model or harm the rest. An uncertainty estimate that stays inside a single client wastes information that could discipline this exchange. The standard federated protocol exchanges model parameters or gradients (Section 2), and this unit of exchange carries two costs: parameter updates potentially leak information about the data that produced them, and they offer nothing a clinician can inspect.

Section 4.1 builds this into FedEvPrompt, an evidential federated learning framework in which uncertainty-filtered attention maps replace parameter aggregation as the unit of inter-client exchange. Each client learns prompts on a frozen ViT backbone under an evidential loss, and contributes to a shared buffer only those attention maps whose per-sample uncertainty is lowest, outperforming parameter-sharing baselines and prior non-parameter-sharing methods in a non-i.i.d data setting.

4.1 Evidential Federated Learning for Skin Lesion Image Classification

In recent decades, deep learning has played a leading role in medical image analysis, including skin lesion classification. However, most of the existing methods rely on centralized learning, assuming data uniformity and accessibility, which often does not align with the reality of decentralized and privacy-sensitive clinical settings. Federated learning (FL) emerges as a promising solution to this challenge, enabling model training across distributed devices while preserving data privacy. Methods like FedAvg [112] and FedProx [95] have addressed issues such as non-i.i.d. data and system heterogeneity, yet they still face obstacles, particularly in scenarios with class imbalance. Evidential Deep Learning (EDL) [144] has found adoption in FL to handle these limitations in medical data, thereby enhancing model confidence and reliability, crucial for clinical applications. For example, the recent work on uncertainty-aware aggregation of federated models for diabetic retinopathy classification demonstrates its efficacy in improving model performance and reliability [169].

Furthermore, the scarcity of data poses an additional significant challenge, often leading to model overfitting and suboptimal federation performance. Recent techniques like learnable prompting [96], particularly effective in low-data regimes, offer a promising solution by facilitating personalized model tuning across distributed clients [92]. Nonetheless, privacy concerns persist, particularly due to the sharing and aggregation of model parameters, which poses the risk of reconstructing training images, as demonstrated by recent studies [43, 194]. To mitigate these concerns, one strategy involves sharing suitably-constructed synthetic data generated through generative models [124]. Yet, the use of generative models carries its own risks, potentially memorizing and reproducing sensitive training samples, thus exacerbating privacy concerns. Alternatively, sharing visual representations rather than parameters structurally excludes the attack regime that targets shared parameter updates $\Delta W = W^{t+1} - W^t$ (which approximate the accumulated local gradients), including white-box model inversion and the parameter-access variants of membership inference [115] and property inference [113]. Input reconstruction remains a threat in principle, but is not worse for visual-representation sharing than for parameter sharing: attackers exploiting saliency-style explanations [188] achieve reconstruction quality on par with gradient-inversion attackers exploiting shared parameters [194, 43].

We propose FedEvPrompt, which integrates evidential deep learning, prompt tuning, and knowledge distillation of visual representations to address these limitations. FedEvPrompt trains prompts prepended to pre-trained ViT models in an evidential learning setting, max-

imizing class evidence. Knowledge sharing across clients in the federation is achieved only through knowledge distillation on attention maps generated by ViT models, which avoids sharing parameters or pixel-level data. The choice of attention as the exchange unit connects to Chapter 3, where clinician attention guided the model from the input side; here, model attention is shared across clients on the output side. While our approach maintains a high level of abstraction for minimizing privacy leaks, it also provides richer information than average logits, as in FedDistill [145], or prototypes, as in FedProto [154]. Thus, FedEvPrompt represents a principled way to share insights into the decision-making process of local models for enhanced federated performance, as demonstrated by the results achieved on a real-world distributed setting for skin lesion classification.

4.1.1 Method

We introduce FedEvPrompt leverages prompt evidential learning and knowledge distillation on ViT attention maps for enabling effective knowledge aggregation across federated clients. The overall learning strategy is described in Fig. 4.1.

FedEvPrompt is based on a pre-trained ViT model, kept frozen across all clients within the federation. Upon the fixed backbone, prompts are prepended on each client model and optimized using local data. Each client also computes attention maps (through attention rollout mechanism [1]) for each class and shares a subset of them with the federation. The attention maps by all clients form our *uncertainty-aware attention buffer* that is used for knowledge distillation during prompt learning.

Learning is organized in rounds: at each round, federation clients carry out different local training epochs for prompt optimization through a combination of evidential loss for learning class evidence and knowledge distillation loss on the per-class attention maps present within the buffer. At the end of training round, each client identifies its M most informative attention maps for each class and updates the buffer.

More in detail, each client employs a frozen ViT, as the backbone, with two sets of prompts *b-prompts* and *t-prompts*. Each prompt is associated with a specific attention layer, with the *b-prompts* (basic prompts) prepended to layers with low-level feature representation and the *t-prompts* (task-specific prompts) to deeper layers with high-level feature representation. The parameters of the prompts are incorporated through pre-fix tuning. Let's denote the output of the i_{th} attention layer as h_i with $i \in 1 \cdots H$ where H is the number of attention layers. The prepended parameters for the key and value inputs, denoted as pr_k and pr_v respectively, are incorporated into the multi-head self-attention (MSA) operation as follows:

$$MSA(h_Q, [pr_k^{(i)}; h_K], [pr_v^{(i)}; h_V]) \quad (4.1)$$

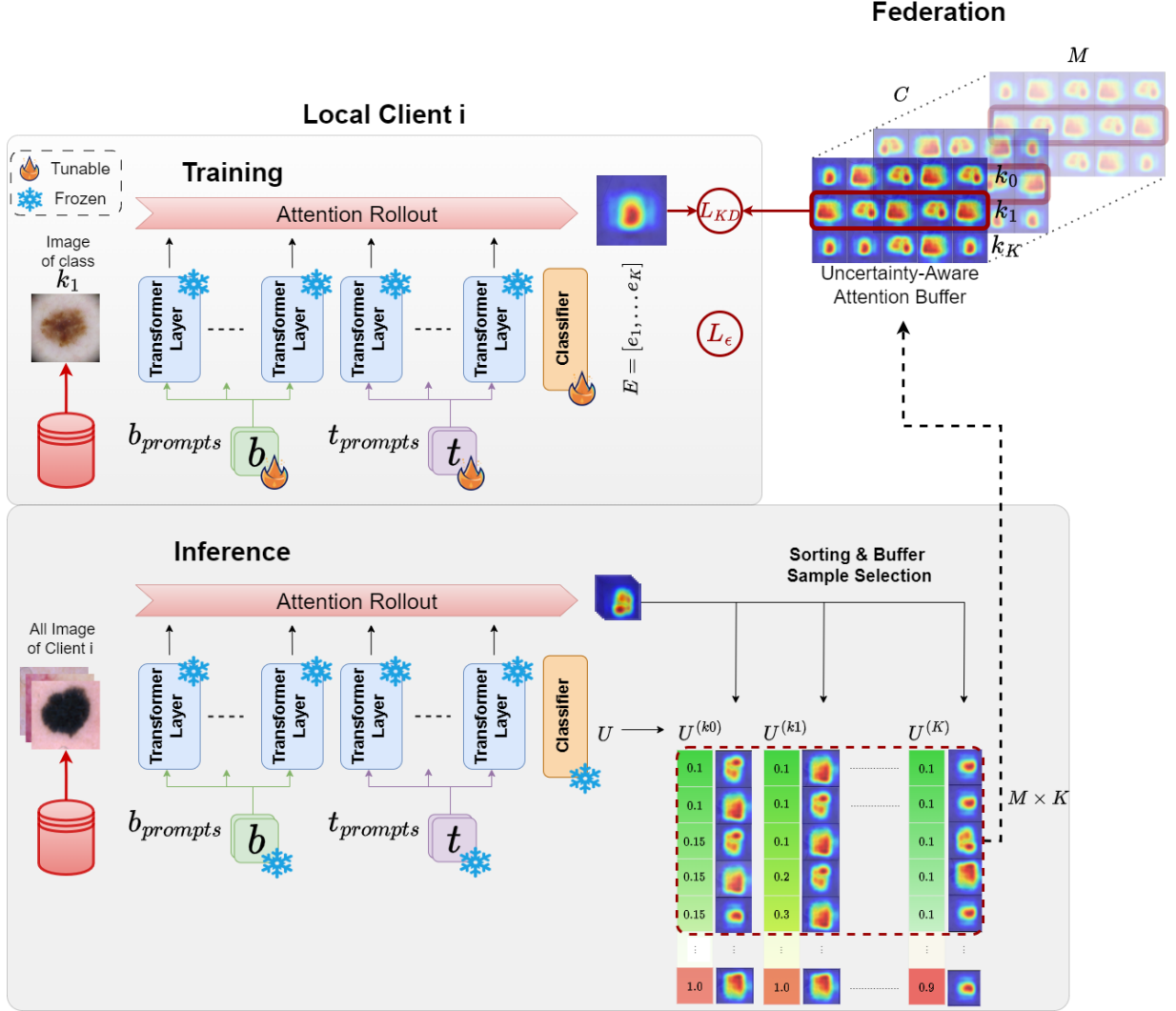


Figure 4.1: **Overview of FedEvPrompt.** During a round of **Training (top)**, local data is used to optimize b -prompts, encoding general visual features, and t -prompts, encoding task-specific information, prepended to a frozen ViT encoder. Optimization is carried out by minimizing evidential loss ($\mathcal{L}_e$) and a knowledge distillation loss ($\mathcal{L}_{KD}$) between local attention maps and those of the federation available in the *uncertainty-aware attention buffer*. After a round of training at **Inference (bottom)**, the client identifies, for each class K , its M most informative attention rollout maps (sorted by lowest uncertainty) to contribute to the federated *uncertainty-aware attention buffer*.

where h_Q , h_K , and h_V represent the query, key, and value outputs from the previous layer, respectively. The prepended prompts $pr_k^{(i)}$ and $pr_v^{(i)}$ are *b-prompts* to the first l layers, and *t-prompts* to the last $H-l$ layers. The two sets of prompts undergo distinct optimization strategies: *b-prompts* require slower adaptation since the frozen backbone (ViT) has already grasped general visual features. Conversely, *t-prompts* necessitate faster adjustments to accommodate varying data distributions. Consequently, $\mu_1 < \mu_2$, where μ_1 and μ_2 denote the learning rates for the *b-prompts* and *g-prompts* optimizers, respectively. Both set of prompts are optimized by minimizing an overall loss $\mathcal{L}_G$ that includes an evidential loss term $\mathcal{L}_\epsilon$ and a knowledge distillation loss term $\mathcal{L}_{KD}$ on the shared attention map buffer A :

$$\mathcal{L}_G = \mathcal{L}_\epsilon + \lambda \mathcal{L}_{KD} \quad (4.2)$$

where $\lambda = 1e^6$ is a parameter controlling the balance between the two terms to have a similar order of magnitude.

Evidential loss

Our method is based on evidential learning, i.e., the classification model outputs evidences $E = [e_1, \dots, e_K]$, with K categorical class elements (number of classes). The Dirichlet distribution characterizes the likelihood of each discrete probability value within a set of possible probabilities. It is parameterized by a vector of K elements (classes), $\alpha = [\alpha_1, \dots, \alpha_K]$, defined as $\alpha_k = e_k + W_k$, with e_k being the model evidence for class k , and W_k the prior weight for that class. Classical EDL assumes a uniform Dirichlet (Dir(1)) distribution as a prior, i.e., $\mathbf{W} = \langle 1, 1, \dots, 1 \rangle$. The uncertainty for the i^{th} input sample is then estimated as $u_i = \frac{K}{S}$, with S being the total Dirichlet strength $S = \sum_{k=1}^K \alpha_k$. Due to the strong class imbalance typical in federated learning settings, we change the uniform evidential prior to a skewed distribution, weighted by class frequency:

$$\alpha_k = e_k + W_k \quad \text{with} \quad W_k = \frac{K}{K-1} \left(1 - \frac{N_k}{N} \right) \quad (4.3)$$

such that $\sum_{k=1}^K W_k = K$.

Prompt parameters are finally optimized by minimizing the evidential loss defined in [144] as a combination of MSE and KL divergence. Given the i^{th} input sample, the one-hot-encoded vector y_i of its class label k , and its expected probabilities p_i , the **evidential loss** $\mathcal{L}_\epsilon$ is computed as:

$$\mathcal{L}_\epsilon(\theta) = \mathbb{E}_{\mathbf{p} \sim \text{Dir}(\alpha)} [(y_i - p_i)^T (y_i - p_i)] + \lambda_{KL} D_{KL}(\text{Dir}(p_i | \tilde{\alpha}_i) || \text{Dir}(p_i | \mathbf{w}_i)) \quad (4.4)$$

with $\lambda_{KL} = \min(1, t/10)$ being an annealing factor applied to gradually increase the regularization impact with the number of epochs t .

In order to let the evidence for incorrect classes shrink to the weighted prior values $\mathbf{W}$, the KL divergence loss term minimizes the distributional difference between $\mathbf{W}$ and misleading evidence $\tilde{\alpha}$, formulated as $\tilde{\alpha} = y \cdot \mathbf{w} + (1 - y) \odot \alpha$. Given the weighted prior distribution $\mathbf{W} = \text{Dir}(\mathbf{p}|\mathbf{w})$ and $P = \text{Dir}(\mathbf{p}|\tilde{\alpha})$, the general KL divergence form for the Dirichlet distribution [127] becomes:

$$\begin{aligned} D_{KL}(\text{Dir}(p|\tilde{\alpha})||\text{Dir}(p|\mathbf{w})) &= \log \left(\frac{\Gamma \left(\sum_{k=1}^K \tilde{\alpha}_k \right)}{\Gamma \left(\sum_{k=1}^K w_k \right)} \right) + \sum_{k=1}^K \log \left(\frac{\Gamma(w_k)}{\Gamma(\tilde{\alpha}_k)} \right) + \\ &\sum_{k=1}^K (\tilde{\alpha}_k - w_k) \cdot \left[\psi(\tilde{\alpha}_k) - \psi \left(\sum_{k=1}^K \tilde{\alpha}_k \right) \right] = \log \left(\frac{\Gamma \left(\sum_{k=1}^K \tilde{\alpha}_k \right) \cdot \prod_{k=1}^K \Gamma(w_k)}{\Gamma(K) \cdot \prod_{k=1}^K \Gamma(\tilde{\alpha}_k)} \right) + \\ &+ \sum_{k=1}^K (\tilde{\alpha}_k - w_k) \cdot \left[\psi(\tilde{\alpha}_k) - \psi \left(\sum_{k=1}^K \tilde{\alpha}_k \right) \right] \end{aligned} \quad (4.5)$$

With Γ being the gamma function, and ψ being the digamma function.

Uncertainty-Aware Attention Buffer for Knowledge Distillation

Prompt optimization involves minimizing a knowledge distillation loss $\mathcal{L}_{KD}$ term between the model attention maps (computed through attention rollout) and the maps available in our *uncertainty-aware attention buffer* A shared within all the C clients of the federation, with each client providing M attention maps for each of the K classes:

$$A = \bigcup_{c=1}^C \bigcup_{k=1}^K \bigcup_{m=1}^M a_{c,k,m} \quad (4.6)$$

here $a_{c,k,i} \in \mathcal{R}^{H \times W}$ represents the i^{th} attention map for the k^{th} class of the c^{th} client of the federation. H and W are the height and width of the attention maps equal to input image dimensions. The **knowledge distillation loss** $\mathcal{L}_{KD}$ for a generic training sample of client c , with class k can be expressed as:

$$\mathcal{L}_{KD} = \frac{1}{M} \sum_{i=1}^{C \setminus c} \sum_{m=1}^M \|a_{c,k,-} - a_{i,k,m}\|^2 \quad (4.7)$$

where, $\|a_{c,k,-} - a_{i,k,m}\|^2$ denotes the squared Euclidean distance between the attention map $a_{c,k,-}$ of the considered training sample and $a_{i,k,m}$ being an item of the buffer A .

The selection of samples for the attention buffer A by each client is based on the assumption that each local model should share its most confident predictions and indicate the image regions it focuses on. We use uncertainty scores from our evidential learning approach to guide the selection of attention maps for sharing within the federation. Specifically, we compute uncertainty scores, $u_j^{(k)}$, for each sample in class k , where j ranges from 1 to $N^{(k)}$, the total number of samples in class k . From these, we select the M samples with the lowest uncertainty scores, denoted as $\{u_1^{(k)}, u_2^{(k)}, \dots, u_M^{(k)}\}$, and corresponding attention maps for inclusion in our uncertainty-aware attention buffer, A , replacing older ones.

4.1.2 Experimental Results

We validate the effectiveness of our proposed method on a multicenter dataset of 23,247 dermoscopic images of nine skin lesions from different populations and medical centers, based on the ISIC2019 dataset [61, 29, 160]. To carry out federated learning, we organized the dataset into six nodes, with each node representing data from a specific source: Client C1

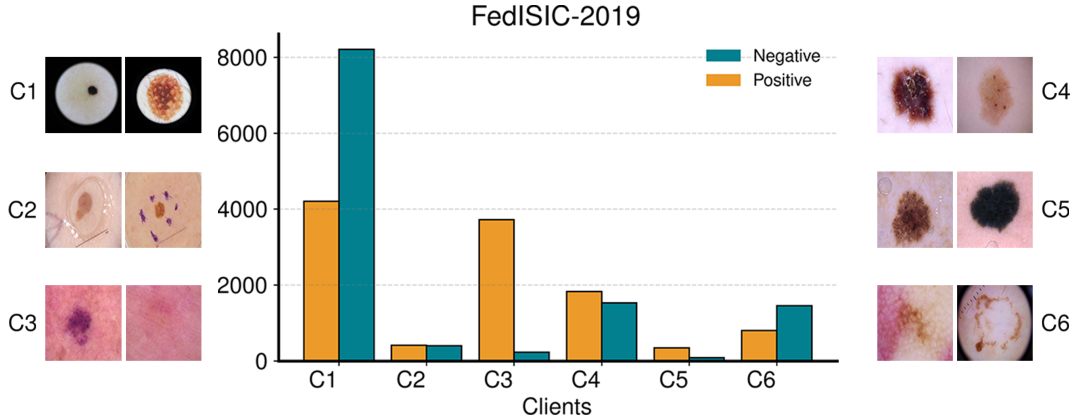


Figure 4.2: Distribution of the Fed-ISIC2019 dataset across clients.

contains the BCN20000 dataset as described by Combalia et al. [61], which includes 19,424 images from the Hospital Clínic in Barcelona; Clients C2, C3 and C4 are from the Austrian portion of the HAM10000 dataset [160], with images from the ViDIR Group at the Department of Dermatology at the Medical University of Vienna; Client C5 is also part of the HAM10000 dataset and contains the Rosendahl image set from the University of Queensland in Australia; while client C6 includes the MSK4 dataset [29]. The overall dataset exhibits heterogeneity in both the number of images contributed by each client as well as in the

distribution of classes, as illustrated in Fig. 4.2, making it a strong real-world use case for testing federated learning methods. For this study, we will focus on the binary classification task of distinguishing *Melanocytic nevus* from other skin lesions.

Training procedure In our setup, for each client, data is divided into a 75% training and 25% test split. Training is executed over 5 communication rounds, with 15 training epochs per round. Our model architecture employs a frozen ViT backbone augmented with additional parameters for b-prompts, t-prompts, and a classification-head. The ViT backbone specifications include an embedding dimension of 384, 6 attention heads, 12 blocks, and an input size of 224x224 pixels. For the b-prompts and t-prompts, the prompt keys (k and v) have a sequence length of 50, while l is 3 out of the 12 attention layers. We set the learning rates μ_1 and μ_2 to $2.5e - 4$ and $5e - 4$ respectively, with a weight decay factor of $1e - 2$. Additionally, each client contributes 5 attention rollout maps per class (i.e., M in Eq. 4.6) to the uncertainty-aware attention buffer. Results are presented as in terms of balanced accuracy on the test set at the conclusion of all rounds.

Results. In Table 4.1, we present a comprehensive performance comparison between FedEvPrompt and existing federated learning methods. FedAvg [112] serves as our baseline. FedAvgPers builds upon FedAvg by integrating a personalization step through local data fine-tuning, aligning with our emphasis on personalized learning via b-prompts and t-prompts tuning. Additionally, we incorporate FedProx [95], specifically tailored to address non-IID data like our skin lesion dataset.

Given that our approach employs knowledge distillation without parameter sharing, we include two analogous methods in our analysis: FedProto [154] and FedDistill [145]. We also evaluate the performance of local training, where client models are trained independently without parameter sharing, using both sets of prompts (i.e., $(b, t)_{prompts}$), and using only one set of prompts ($g_{prompts}$ - general prompts) across all attention layers. We define $g_{prompts} = [b_{prompts}, t_{prompts}]$ with both learning rates set to μ_1 . This evaluation aims to validate our choice to apply different parameters across different attention layers and to demonstrate the advantages provided by our federated learning approach.

Results show that FedEvPrompt outperforms its competitors, including those that share parameters (thus being less privacy-preserving), such as FedAvg [112] and FedProx [95]. Notably, when comparing FedEvPrompt performance with other methods that do not share parameters, namely FedProto [154] and FedDistill [145], we observe higher performance across all clients and a lower standard deviation, indicating better convergence in accuracy among clients.

We finally conducted an ablation study to assess the impact of various prompting options

Table 4.1: **Comparison with state-of-the-art methods** on the Skin Lesion Dataset. In bold, best accuracy values.

	C1	C2	C3	C4	C5	C6	Avg
Local $b, t_{prompts}$	72.56	50.00	89.91	79.94	67.61	50.00	68.34 $\pm$ 10.98
Local $g_{prompts}$	50.00	50.48	84.79	78.91	71.59	50.00	64.30 $\pm$ 10.31
FedAvg [112]	74.74	72.79	67.74	79.89	67.61	77.09	73.31 $\pm$ 3.64
FedAvgPers	77.79	68.01	84.84	78.08	67.61	82.66	76.50 $\pm$ 7.25
FedProx [95]	81.34	68.14	74.69	75.15	77.84	78.65	75.97 $\pm$ 4.55
FedProto [154]	71.88	69.04	66.34	69.71	58.52	73.57	68.18 $\pm$ 5.34
FedDistill [145]	81.08	67.42	60.96	77.13	70.45	80.33	72.90 $\pm$ 7.98
<i>FedEvPrompt</i>	81.02	71.83	84.15	79.02	68.18	79.35	77.26 $\pm$ 4.65

and sharing strategies among nodes within the federation. It’s worth noting that while prompt sharing may potentially compromise privacy guarantees, exploring its effectiveness compared to using private prompts and our proposed knowledge distillation approach is interesting.

To this end, we initially assessed the performance of a variant of FedAvg where only low-level b -prompts are shared, gradually incorporating t -prompts and knowledge distillation on the uncertainty-aware attention buffer. Furthermore, we examined the variant of the proposed prompting strategy using a single set of general prompts g -prompts shared between nodes and coupled with our knowledge distillation method. Our findings, outlined in Table 4.2, underscore the significance of separate prompt learning, as evidenced by the subpar performance of the g -prompts variants. Interestingly, sharing separate sets of b -prompts and t -prompts (first two rows of Tab. 4.2) proved less effective than keeping them private and employing knowledge distillation (best performance observed in the last two rows of the same table). Moreover, we demonstrate that our strategy of incorporating attention maps based on uncertainty scores (as detailed in Sect. 4.1.1) yields superior performance compared to random selection of buffer samples. These two last considerations highlight the informative contribution provided by the attention maps corresponding to the lowest uncertainty samples driving clients’ models towards the most significant regions of skin lesion images.

4.1.3 Discussion

This section introduced FedEvPrompt, a federated learning framework combining evidential deep learning, prompt tuning, and knowledge distillation on Vision Transformer attention maps. On the Fed-ISIC2019 benchmark, which partitions skin lesion classification across six institutional sources with heterogeneous class distributions, FedEvPrompt outper-

Table 4.2: **Ablation study results** showing the impact of shared prompts and knowledge distillation (KD) on federated learning performance.

	C1	C2	C3	C4	C5	C6	Avg
FedAvg $b_{prompts}$	50.00	70.04	70.29	79.61	71.59	79.21	70.18±7.56
+ $t_{prompts}$	74.74	72.79	67.74	79.89	67.61	77.09	73.31±3.64
+ KD	73.02	71.41	76.80	81.74	67.61	79.25	74.97±3.94
FedAvg $g_{prompts}$	50.00	70.88	72.77	50.00	67.61	50.00	60.21±6.81
FedAvg $g_{prompts}$ + KD	81.1	71.32	72.83	78.89	63.64	80.49	74.71±6.18
KD _{random}	78.12	69.40	76.80	80.30	68.75	77.65	75.17±3.66
KD _{uncertainty} (<i>Ours</i>)	81.02	71.83	84.15	79.02	68.18	79.35	77.26±4.65

forms parameter-sharing baselines (FedAvg, FedAvgPers, FedProx) and prior non-parameter-sharing methods (FedDistill, FedProto) on average balanced accuracy, in a setting where the non-i.i.d. structure arises from the data sources themselves rather than from a synthetic partitioning of a centralized dataset. The privacy advantage of attention-map sharing over parameter sharing is a structural property of the exchange unit rather than one measured by a formal privacy analysis. Evidential uncertainty is well-matched to the federated setting because it adds no inference-time cost beyond the standard forward pass: the Dirichlet output that supports the buffer selection is produced as part of the same prediction step the client would run regardless. Approaches that produce uncertainty through stochastic sampling or model ensembling require either multiple forward passes per sample or multiple parallel models, costs that compound when every candidate sample must be scored per round. The buffer selects, per client and per class, the M attention maps with the lowest predictive uncertainty, a scalar derived from the Dirichlet output. Decomposing this scalar into vacuity (insufficient evidence) and dissonance (conflicting evidence) [77, 54] is a natural refinement: the two failure modes have different implications for whether a map is worth sharing, with vacuity indicating a client that lacks evidence to contribute and dissonance indicating genuinely ambiguous cases that may still be informative to peers. A more fundamental bound applies to both decompositions. Evidential uncertainty quantifies the evidence a model has within its own feature space. It flags samples that are poorly supported by the training distribution, but it cannot detect that the feature space itself is inadequate, nor separate correlational from causal support. A model with a systematically biased representation may remain confident and unflagged, and separating little evidence from wrong evidence would require interventional data or distributional assumptions that the federated setting does not provide.

This section’s contribution to the chapter’s *Doubting* argument is that uncertainty, when efficiently implemented as a structural output of the model, serves as a selection criterion on what enters the federation. The choice of attention maps as the exchange unit also connects to the preceding chapter, where clinician attention served as a prompt to the model, and where the same representational surface now becomes the interface between models in a federation.

4.1.4 Publications

Rutger Hendrix, Federica Proietto Salanitri, Concetto Spampinato, Simone Palazzo, and Ulas Bagci. Evidential federated learning for skin lesion image classification. In *International Conference on Pattern Recognition (ICPR)*, pages 354–365. Springer, 2024.

Chapter 5

Forgetting: Instant Machine Unlearning

This chapter argues that pre-committing to forgetting structurally unlocks capabilities that post-hoc unlearning cannot deliver. Standard unlearning methods pay the cost of forgetting at request time, through retraining, access to the original data, or an auxiliary teacher. The contributions here pre-commit that cost at training time: knowledge to be forgotten is localised in a small, modular component on top of a frozen pre-trained backbone, and unlearning reduces to removing or swapping that component. Section 5.1 builds this into Pre-Forgettable Models, in which class-wise knowledge is held in dedicated learnable prompt tokens on a frozen backbone, so that unlearning a class is the removal of its prompt. Section 5.2 extends this to TrustErase, which binds each unlearning configuration to a non-trainable passport embedded inside the model’s LoRA adapters. The passport makes the deployed model’s unlearning configuration auditable, so that an authority can confirm which configuration is in force rather than relying on the developer’s claim.

5.1 Pre-Forgettable Models: Prompt Learning as a Native Mechanism for Unlearning

The advent of foundation models has significantly advanced multimedia understanding, enabling robust and transferable representations across a range of modalities and tasks [34, 130, 58]. Yet alongside this capacity to learn, models are increasingly expected to *unlearn* — to remove specific knowledge or data influence efficiently and on demand. This demand is driven by privacy mandates such as the GDPR’s right to be forgotten [165] and to broader needs for fairness, adaptability, and bias correction [152]. A common approach is approximate unlearning, which sits between full retraining (a gold standard that is rarely affordable) and exact removal (which generalizes poorly to deep networks). Targeted strategies such as data sharding [16], gradient negation [108], parameter perturbation or layer-wise resets [41, 47], and teacher-student or prompt-based distillation [28, 172] reduce the cost relative to retraining, but remain post-hoc corrections: they assume an unlearning stage that runs after training, with access to data, an auxiliary teacher, or a window of computation that is incompatible with continuously deployed systems.

We propose a different starting point: embed the unlearning operation into the architecture, so that it is available without retraining and without data at request time. Our strategy leverages prompt-based learning, but reinterprets its purpose. Where prior work uses prompts to adapt a frozen backbone to new tasks [90, 102, 190], we treat them as containers for class-specific knowledge that can be independently manipulated, and in particular removed. Each class is associated with a dedicated, learnable prompt token and during training only the prompts, adapters, and classifier head are updated while the foundation-model backbone remains frozen. Class-specific knowledge is therefore localised within prompts, and unlearning a class reduces to removing the corresponding prompt — without modifying weights, accessing training data, or triggering any retraining.

We evaluate this design across three medical image classification tasks from the MedMNIST v2 benchmark — BloodMNIST, DermaMNIST, and OrganSMNIST — using a frozen Vision Transformer with shallow prompt tuning and LoRA adapters, and an additional audio benchmark (UrbanSound8K) to test applicability across modalities. Unlearning is measured under three forget-set configurations: a single class, half of the classes, and all but one. The method achieves near-random accuracy on forgotten classes while maintaining high accuracy on retained ones, and resists membership-inference attacks. Beyond the unlearning evaluation itself, we also examine a bonus of the design, where variations across subsets at inference yields a per-sample uncertainty signal, which we compare against MC dropout.

5.1.1 Related Work

Machine unlearning has become an increasingly important area of study, driven by privacy regulations such as the GDPR and the growing demand for dynamic data governance in AI systems. Existing techniques can be broadly classified into two categories: data-centric and model-centric approaches, each presenting distinct trade-offs in terms of scalability, guarantees, and practicality.

Data-centric methods aim to localize data influence within the training set to simplify removal operations. For instance, SISA [16] partitions the data into disjoint shards and trains separate sub-models, allowing efficient retraining upon deletion. While this avoids full model retraining, it requires storage overhead and may degrade performance as the number of shards increases. Other strategies include data obfuscation [52], which corrupts or relabels targeted samples, and summation-based representations [20], which compress training data to support fast recomputation. However, these methods typically assume access to the original dataset and offer limited guarantees of complete forgetting.

Model-centric techniques instead modify the trained model directly, without relying on the training data. Certified removal [53] uses influence functions to estimate and reverse the effect of individual data points—mainly in convex settings. SSD [41], a Fisher-based approach, perturbs parameters according to their importance to selectively erase memorized information. While these methods bypass retraining and offer some theoretical grounding, they often incur high computational costs and face challenges in deep, non-convex architectures. Hybrid techniques like CF- k and EU- k [47] achieve partial unlearning by retraining upper layers while freezing lower ones, but this can leave residual knowledge in the frozen backbone, limiting their effectiveness.

Building on these categories, **class-level unlearning** has emerged as a more structured and task-aligned formulation, especially relevant for classification problems. Instead of targeting individual samples, these methods aim to erase entire semantic categories. PBU [121] treats unlearning as a Bayesian optimization process, selectively degrading performance on the forgotten class while maintaining accuracy elsewhere. More recently, prompt-based techniques [172] associate clustered data regions with learned prompts and classifier heads, enabling post hoc removal through retraining specific clusters. While promising, these methods still require additional unlearning phases and lack true modularity or data-free operation, that is, an unlearning step that at request time needs access to neither the forget set nor the retain set.

Our method extends this line of work by reimagining prompts as self-contained carriers of class-specific knowledge, explicitly localized during training. By binding each class to a

dedicated prompt token and freezing the backbone, our model enables immediate, verifiable unlearning simply by removing the corresponding prompt—without accessing data, modifying model weights, or performing any retraining. This design supports scalable, interpretable, and retraining-free unlearning, directly addressing the limitations of prior approaches in both practicality and deployability.

5.1.2 Method

We propose a prompt-based framework that attaches a *learnable prompt* to each class in order to gate its recognition. In essence, the knowledge of each class is isolated into a small prompt vector (or set of vectors) that conditions the model’s prediction for that class. This design is modality-agnostic: it can be applied to image classification, text classification, or any domain where a pretrained encoder can be conditioned on prompts. Crucially, **unlearning a class is as simple as removing its prompt**—with no retraining and no changes to the encoder’s weights. Once a class’s prompt is removed, the model no longer recognizes that class, as if the class had never been learned, while all other classes remain fully recognizable.

For example, consider a medical image classification scenario in which a model diagnoses diseases from X-ray images. In our framework, each disease class (e.g., pneumonia, nodule, etc.) is associated with its own prompt embedding that “activates” recognition of that condition. During inference, an X-ray image is analyzed in conjunction with each disease’s prompt to produce class-specific confidence scores. If a certain disease must be unlearned (say due to a patient data removal request or a shift in policy), we simply detach the corresponding disease prompt. The model, without that prompt, will cease to predict the removed disease—any input that previously triggered the now-removed prompt will yield no high-confidence prediction for that class. This medical imaging use case illustrates the approach in a high-stakes domain, but the method is general: the prompt-based gating works for any classification task (but it can be extended to all those tasks that can be tackled through prompt tuning), providing a flexible mechanism to add or remove classes on the fly.

Figure 5.5 illustrates the mechanism using a medical image classification case: during training, the model learns to associate each class with its unique prompt (key). During inference, removing the prompt prevents recognition of that class, effectively achieving unlearning.

The architecture of the proposed approach consists of a frozen backbone encoder f_θ (with parameters θ fixed after pre-training) and a set of class-specific prompt embeddings. Each class c in the label set C is assigned a prompt p_c , which is a small learnable vector or collection

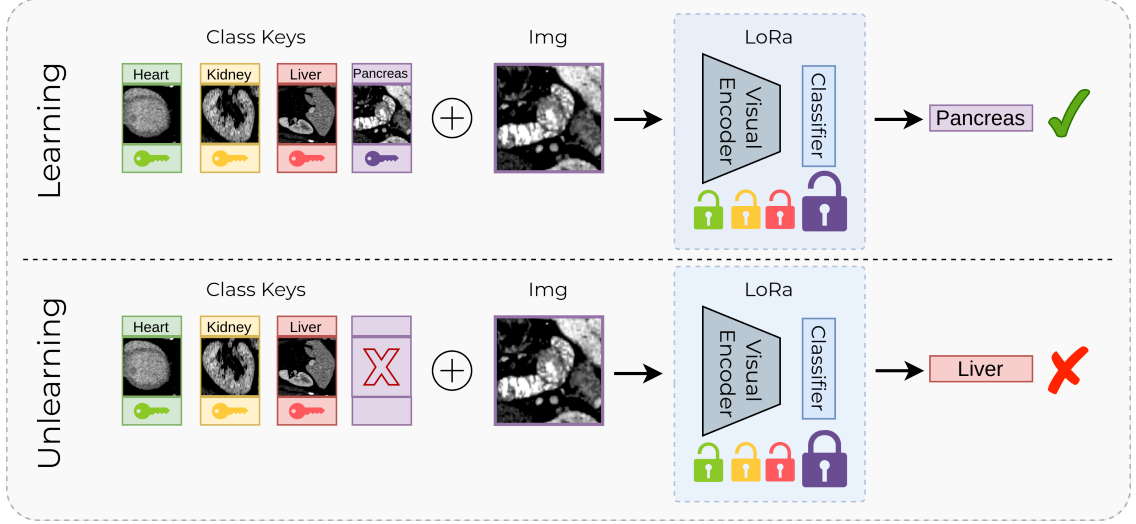


Figure 5.1: **Prompt-based learning and unlearning framework for medical image classification.** Each class is linked to a specific prompt. During training, prompts are used to learn class-specific representations. At inference, removing a prompt prevents the model from recognizing that class.

of vectors. These prompts serve as *dynamic conditioning parameters* that modulate the encoder’s behavior for their respective class. We represent each prompt p_c as, for example, a trainable sequence of M tokens in the same embedding dimension d as the encoder.

The prompts are integrated into the input processing pipeline of the encoder through concatenate to the input features: for vision transformers (e.g., ViT backbone used for medical image classification), we prepend the M prompt tokens to the sequence of image patch embeddings at the input layer. This way, the self-attention layers will attend to both the image patches and the prompt tokens.

Formalisation

Let f_θ be a frozen pretrained foundation model suitable for the task domain, e.g., a vision transformer (ViT) for image classification or a CLIP-like encoder for multimodal inputs—with fixed parameters θ . The backbone is never updated during training, which ensures that learned representations remain stable and that changes to prompt structure are fully traceable.

We define a set of target classes $\mathcal{C} = \{c_1, \dots, c_K\}$. To each class $c_i \in \mathcal{C}$, we associate a unique learnable prompt vector p_{c_i} . These prompt vectors, together with a single set of lightweight low-rank adaptation (LoRA) modules with parameters ϕ , are the only trainable components of our system. The prompts collectively form the prompt set $P = \{p_{c_1}, p_{c_2}, \dots, p_{c_K}\}$. Each prompt is intended to function as a semantic key that unlocks

the model’s ability to recognize the corresponding class.

Let $x \in \mathcal{X}$ denote an input sample and $c \in \mathcal{C}$ its corresponding label, with training pairs drawn from a data distribution $\mathcal{D}$, i.e., $(x, c) \sim \mathcal{D}$. For each training instance, we provide the model with the correct class prompt p_c , along with a set of m distractor prompts, sampled from the remaining class prompts $P \setminus \{p_c\}$, with m drawn from a distribution $q(m)$. This forces the model to learn to discriminate based on the correct prompt, even in the presence of misleading alternatives.

The prediction is thus made by the following formulation:

$$\hat{y} = f_{\theta, \phi}(x, \mathcal{S}(\{p_c\} \cup \sigma(P \setminus \{p_c\}, m))) \quad (5.1)$$

where $\mathcal{S}$ is a permutation function used to randomize prompt order and prevent positional bias, and $\sigma(P \setminus \{p_c\}, m)$ is a sampling function defined below.

To construct the distractor prompt set, we define the sampling operation:

$$\begin{aligned} \sigma(P \setminus \{p_c\}, m) &= \{p_{\tilde{c}_1}, p_{\tilde{c}_2}, \dots, p_{\tilde{c}_m}\}, \\ \text{where } \{p_{\tilde{c}_1}, \dots, p_{\tilde{c}_m}\} &\sim \text{Uniform}(\mathcal{P}_m(P \setminus \{p_c\})) \end{aligned} \quad (5.2)$$

where $\tilde{c}_1, \dots, \tilde{c}_m \in \mathcal{C} \setminus \{c\}$ denote m distinct distractor classes, drawn without replacement, with $m \leq K - 1$.

This ensures that the model is consistently exposed to a diverse mixture of incorrect class prompts during training, which strengthens the coupling between correct prompt and class label, while weakening associations with unrelated prompts.

Thanks to this design, forgetting a class is immediate and straightforward: to unlearn class k , we remove its prompt p_k from the set P . Since the model has been trained to produce a near-uniform output distribution whenever the corresponding class prompt is absent, removing p_k causes the model to no longer confidently predict k : for inputs of class k , the output distribution approaches uniformity, with no class exceeding chance-level confidence.

Training objective

The training objective is designed to simultaneously achieve two complementary goals: (i) accurate classification when the correct class prompt is present, and (ii) effective unlearning when the class prompt is absent. To this end, we define two loss terms—a standard learning objective and a novel unlearning objective—that are jointly optimized.

Learning Objective. The primary objective ensures that the model correctly predicts the class label c when the correct prompt p_c is available among the input prompts. This encourages the model to associate the presence of a specific prompt with the corresponding class semantics. The learning loss is computed as:

$$\mathcal{L}_{\text{learn}} = \mathbb{E}_{(x,c) \sim \mathcal{D}, m \sim q(m)} [\ell(f_{\theta,\phi}(x, \mathcal{S}(\{p_c\} \cup \sigma(P \setminus \{p_c\}, m))), c)] \quad (5.3)$$

where the expectation is taken over training pairs $(x, c) \sim \mathcal{D}$ and the number of distractors $m \sim q(m)$, ℓ is the standard cross-entropy loss, $\mathcal{S}$ denotes the permutation operator applied to the set of prompts (to prevent position encoding bias), and σ is the distractor sampling function described earlier. This loss enforces that classification performance remains high under the correct prompt-conditioning scenario.

Unlearning Objective. To ensure that the model abstains from predicting a class when its corresponding prompt is removed, we introduce an unlearning objective. Specifically, when the input is paired only with incorrect prompts (i.e., the target class prompt is excluded), the model is trained to produce a high-entropy output distribution over the class space. This is achieved by minimizing the divergence between the model’s output and a uniform distribution:

$$\mathcal{L}_{\text{unlearn}} = \mathbb{E}_{(x,c) \sim \mathcal{D}, m \sim q(m)} \left[D_{\text{KL}} \left(f_{\theta,\phi}(x, \mathcal{S}(\sigma(P \setminus \{p_c\}, m))) \parallel \mathbf{u} \right) \right] \quad (5.4)$$

where $\mathbf{u}$ is the uniform distribution over the K classes, and D_{KL} denotes the Kullback-Leibler divergence. This term penalizes confident predictions made in the absence of the correct prompt, promoting uncertainty and thus enabling effective unlearning behavior.

The final training loss combines the two components described above:

$$\mathcal{L} = \mathcal{L}_{\text{learn}} + \lambda \mathcal{L}_{\text{unlearn}} \quad (5.5)$$

where $\lambda \in \mathbb{R}^+$ is a hyperparameter controlling the trade-off between accurate classification and the strength of the unlearning signal.

Inference-Time Uncertainty Estimation

With the permutation operation S and the varying number of distractors m applied during training, $\mathcal{L}$ marginalises over a distribution of prompt-subset configurations. The trained model therefore learns to assign the same target class under different configurations, but the underlying computation need not be invariant, and residual variation across configurations

remains measurable at the output. We use this source of variation at inference, and evaluate this variance usefulness as deferral signal.

Let x be a test image and let $\mathcal{T}(x) = \{(P^{(t)}, S^{(t)})\}_{t=1}^T$ be a set of T prompt subsets sampled from the configuration space described below. For each t we run a forward pass and obtain a softmax output $p_t(x) \in \mathbb{R}^C$. The total uncertainty score is the entropy of the average prediction:

$$\bar{p}(x) = \frac{1}{T} \sum_{t=1}^T p_t(x), \quad H(x) = -\frac{1}{\log C} \sum_{c=1}^C \bar{p}_c(x) \log \bar{p}_c(x) \in [0, 1] \quad (5.6)$$

where $\log C$ normalizes H to the unit interval.

Configuration space. Each test-time sample varies in non-target subset content and ordering. Each subset size $s \in \{1, \dots, C-1\}$ is allocated an equal share of the budget T where the combinatorial space allows; sizes whose configuration space is smaller than this share are covered completely, and the remaining budget is split evenly across the larger sizes.

Unlearning-Driven Design Advantages

This unified optimization scheme ensures that each prompt functions as both a semantic key and a control token: the model learns to predict a class only when its associated key is present and reliably abstains when it is absent. By encoding this behavior during training, we eliminate the need for post hoc unlearning procedures, achieving fast, interpretable, and modular forgetting directly at inference time. This makes the approach particularly appealing for dynamic or privacy-sensitive multimedia systems, where class availability may evolve over time or depend on external constraints (e.g., regulatory requests, user consent, or clinical scope).

Overall, the advantages of the proposed formulation are the following ones:

- **Modularity.** Our framework is highly modular: each class is encapsulated by its prompt, acting as an independent module that can be added or removed at will. Eliminating a class is done by simply dropping its prompt. The rest of the system continues to function with zero change.
- **Inference-time uncertainty estimation.** The output variation across subsets at inference provides a usable uncertainty signal. The entropy of the averaged prediction (5.6) provides a per-sample uncertainty estimate, with no auxiliary component or separate training objective.

- **Interpretability.** The effect of removing knowledge is transparent in this framework. When a prompt is removed, the model explicitly loses the ability to recognize the associated class, which is reflected in its outputs: if an image of the removed class is given, no prompt will trigger a high response, resulting in either an abstention (no class scores high) or a roughly uniform probability across remaining classes. This behavior is intuitively interpretable as “the model doesn’t know that class.” In contrast to opaque weight pruning or obscure fine-tuning adjustments, our method provides a direct correspondence between a unit (the prompt) and a semantic concept (the class). Thus we can understand and control the model’s knowledge at the granularity of classes.
- **Scalability.** Because the backbone encoder is fixed, scaling the model to many classes is efficient. The memory and computational footprint of prompts is minimal – prompt vectors are tiny compared to model weights – so adding dozens of new classes only increases the model size and inference cost marginally. There is no need to retrain or even touch the heavy backbone θ when the class set changes, avoiding the expensive retraining typically required for updating a classifier.

5.1.3 Experimental Results

We evaluate our method through a comprehensive set of experiments designed to assess its effectiveness, robustness, and generalizability in achieving class-level unlearning. Given the increasing demand for privacy-aware machine learning—especially in sensitive domains like healthcare—our primary evaluation focuses on medical image classification, where forgetting disease-specific classes is of practical importance. To demonstrate the modality-agnostic nature of our approach, we extend the evaluation to audio classification. The evaluation protocol is structured to quantify not only how well the model forgets specified classes but also how effectively it retains performance on the remaining ones. In addition, we assess privacy resilience, computational efficiency, and the reliability of inference-time uncertainty estimation, thereby establishing the practicality of our method in real-world dynamic learning environments.

Datasets

For medical image classification, we used three publicly available medical image classification datasets as part of the MedMNIST v2 collection [175]: BloodMNIST, DermaMNIST, and OrganSMNIST. These datasets span diverse medical imaging modalities and diagnostic tasks, allowing us to assess the generalizability of our method across different domains:

- *BloodMNIST* [2]: Comprises 17,092 microscopy images of blood cells annotated for classification into eight categories, including different types of white blood cells, red blood cells, and platelets.
- *DermaMNIST* [160, 29]: A dataset of 10,015 dermatoscopic images representing seven common pigmented skin lesions.
- *OrganSMNIST* [12]: Contains 25,211 abdominal CT scans for organ classification, covering 11 different organ types.

All images have a resolution 224×224 pixels, and we follow the standard dataset splits for training, validation, and testing, as described in the medMNIST v2 collection.

For the audio classification task, we use the UrbanSound8K [138] a widely adopted benchmark for environmental sound classification, consisting of 8,732 labeled audio clips, each up to 4 seconds in duration, covering 10 distinct sound categories.

Experimental setup

For medical image classification, we use a Vision Transformer (ViT-Base) as the frozen backbone and adapt it through parameter-efficient fine-tuning. For audio classification, we, instead, use the Audio Spectrogram Transformer (AST) backbone [51], a convolution-free Transformer model pre-trained on large-scale audio datasets. In both scenarios, only the classifier head, prompts, and adapters are trained while keeping the backbone fixed. Visual Prompt Tuning (VPT) [76] is applied in a shallow configuration. LoRa adapters [66] are integrated into each Query and Value matrix of the transformer layers with a low bottleneck dimension $r = 4$, a scaling factor $s = 4$, and dropout rate 0.1. A learning rate of $1e^{-3}$ is used for all learnable parameters. The model is trained for 10 epochs, where class-specific prompts are optimized simultaneously for learning and unlearning. At inference, the total budget for uncertainty estimation is set to $T = 360$ forward passes per test image, applied identically to the prompt-subset and MC dropout procedures so the number of forward passes is the same for both. Thresholds on H are calibrated on the validation split at three target accuracy levels (95%, 97%, 99.9%) and applied unchanged to the test split.

We also evaluate performance of a *Full Knowledge* model, which follows the same configuration as above, but is trained solely for learning using a standard cross-entropy loss (i.e., we do not apply the unlearning loss defined in Formula 5.4. This provides an upper bound for accuracy retention after unlearning.

To assess the effectiveness of unlearning, we measure accuracy on two sets: the retain set and the forget set. The **retain accuracy**, Acc_r , computed as

$$\text{Acc}_r = \frac{1}{|\mathcal{C}_r|} \sum_{c \in \mathcal{C}_r} \mathbb{E}_{(x,c) \sim \mathcal{D}_r} \mathbb{1}[f_\theta(x) = c] \quad (5.7)$$

quantifies the classification performance on the classes that should be retained, where $\mathcal{C}_r$ represents the set of retained classes, $\mathcal{D}_r$ is the corresponding data distribution and $\mathbb{1}$ is the indicator function.

The **forget accuracy**, Acc_f , computed as

$$\text{Acc}_f = \frac{1}{|\mathcal{C}_f|} \sum_{c \in \mathcal{C}_f} \mathbb{E}_{(x,c) \sim \mathcal{D}_f} \mathbb{1}[f_\theta(x) = c] \quad (5.8)$$

measures how well the model still classifies forgotten classes, where $\mathcal{C}_f$ represents the set of forgotten classes and $\mathcal{D}_f$ is the corresponding data distribution. An effective unlearning method should reduce forget accuracy (Acc_f) to near-random levels while maintaining stable retain accuracy (Acc_r), ensuring that forgotten knowledge is erased without affecting retained classes.

We consider three incremental scenarios based on the number f of classes to be forgotten. In the first scenario, $f = 1$, only a single class is removed, leaving all others intact. In the second scenario, half of the classes are removed, such that $f = C/2$ where C is the total number of classes in the dataset. In the third scenario, all but one class are forgotten, which means $f = C - 1$. Per scenario, we report the average of all possible combinations.

Table 5.1: Unlearning results for $f = 1$, where a single class is forgotten. Notably, all competing methods require 10 additional epochs of unlearning, whereas our method achieves effective forgetting without any retraining—demonstrating its retraining-free capability and deployment efficiency.

Method	BloodMNIST		DermaMNIST		OrganSMNIST	
Full Knowledge	98.60		85.99		79.77	
	Acc_r	Acc_f	Acc_r	Acc_f	Acc_r	Acc_f
SSD [41]	94.06 $\pm$ 02.03	07.68 $\pm$ 01.12	76.47 $\pm$ 07.30	13.40 $\pm$ 05.34	71.65 $\pm$ 04.02	07.14 $\pm$ 03.85
Goel et al. [47]	81.81 $\pm$ 01.90	10.95 $\pm$ 01.75	76.87 $\pm$ 01.75	43.83 $\pm$ 02.29	76.86 $\pm$ 02.41	44.33 $\pm$ 06.35
LMEraser [172]	91.71 $\pm$ 06.00	0.02 $\pm$ 0.06	79.67 $\pm$ 04.46	00.01 $\pm$ 0.03	62.86 $\pm$ 10.03	00.13 $\pm$ 00.23
Ours	97.61 $\pm$ 02.63	10.54 $\pm$ 11.38	85.14 $\pm$ 04.55	18.84 $\pm$ 09.27	82.19 $\pm$ 02.63	07.69 $\pm$ 08.47

Among existing unlearning methods, we selected SSD [41] and Goel et al. [47] and LMEraser [172] due to their ease of adaptation across different datasets, and close resemblance to our instant and prompt unlearning objective. Methods requiring extensive modifications were excluded to prevent potential bias. To ensure a fair comparison, we performed a grid search for hyperparameter selection. As a baseline model for the competitors, we

Table 5.2: Unlearning results for $f = C/2$, where half of the classes are forgotten. Competitors are re-trained for 10 epochs.

Method	BloodMNIST		DermaMNIST		OrganSMNIST	
Full Knowledge	98.60		85.99		79.77	
	Acc_r	Acc_f	Acc_r	Acc_f	Acc_r	Acc_f
SSD [41]	95.45 $\pm$ 01.51	95.45 $\pm$ 03.01	70.82 $\pm$ 00.81	67.01 $\pm$ 01.17	69.52 $\pm$ 04.32	69.12 $\pm$ 04.50
Goel et al. [47]	95.56 $\pm$ 08.30	95.56 $\pm$ 10.40	70.87 $\pm$ 07.30	67.06 $\pm$ 08.15	79.83 $\pm$ 11.27	62.58 $\pm$ 10.06
LMEraser [172]	97.75 $\pm$ 01.38	00.02 $\pm$ 00.04	84.76 $\pm$ 06.62	00.03 $\pm$ 00.04	81.41 $\pm$ 07.56	00.01 $\pm$ 00.00
Ours	98.43 $\pm$ 01.37	12.74 $\pm$ 08.86	87.52 $\pm$ 05.68	16.00 $\pm$ 05.86	86.62 $\pm$ 05.93	07.72 $\pm$ 04.52

Table 5.3: Unlearning results for $f = C - 1$, where all but one class is forgotten. Competitors are re-trained for 10 epochs.

Method	BloodMNIST		DermaMNIST		OrganSMNIST	
Full Knowledge	98.60		85.99		79.77	
	Acc_r	Acc_f	Acc_r	Acc_f	Acc_r	Acc_f
SSD [41]	73.29 $\pm$ 06.58	77.36 $\pm$ 07.15	56.36 $\pm$ 08.87	77.31 $\pm$ 06.33	66.30 $\pm$ 03.59	71.72 $\pm$ 2.46
Goel et al. [47]	77.39 $\pm$ 04.81	68.80 $\pm$ 05.30	59.45 $\pm$ 06.69	74.61 $\pm$ 08.39	80.86 $\pm$ 03.93	68.17 $\pm$ 04.89
LMEraser [172]	99.75 $\pm$ 00.31	01.12 $\pm$ 01.73	97.76 $\pm$ 02.33	00.35 $\pm$ 00.35	99.80 $\pm$ 00.18	01.61 $\pm$ 01.31
Ours	99.27 $\pm$ 01.43	16.22 $\pm$ 04.19	96.09 $\pm$ 07.37	19.82 $\pm$ 09.28	98.51 $\pm$ 01.95	11.80 $\pm$ 04.31

trained the same ViT with LoRa and prompts for 10 epochs without unlearning loss, then fine-tuned it for 10 more epochs using their respective unlearning strategies.

Unlearning performance

Tables 5.1, 5.2 and 5.3 compare the performance of our method against SSD [41] and Goel et al. [47] and LMEraser [172] on BloodMNIST, DermaMNIST, and OrganSMNIST on forgetting a single class ($f = 1$), forgetting half of the classes ($f = C/2$) and forgetting all but one class ($f = C - 1$) scenarios. In the first case ($f = 1$), our method significantly outperforms SSD and Goel et al. in terms of Acc_r , achieving the highest retain accuracy (in some cases performance even improves when unlearning) across all datasets, demonstrating that it does not degrade performance on the retained classes. In terms of forgetting effectiveness, our method shows a moderate drop in Acc_f , with values of 10.54% (BloodMNIST), 18.84% (DermaMNIST) and 7.69% (OrganSMNIST). Although SSD achieve lower Acc_f on BloodMNIST and OrganSMNIST, it does so at the cost of significantly lower Acc_r , suggesting an excessive forgetting that negatively impacts retained classes. In contrast, Goel et al. exhibit weaker forgetting, with Acc_f values above 40% for DermaMNIST and OrganSMNIST, indicating incomplete unlearning.

When half of the classes are forgotten ($f = C/2$), SSD and Goel et al. struggle to suppress forgotten information while preserving useful knowledge, failing to effectively reduce Acc_f or maintain high Acc_r . Our methods consistently has chance-level accuracies for Acc_f , with values of 12.74%, 16.00% and 7.72%, for BloodMNIST, DermaMNIST and OrganSMNIST, respectively.

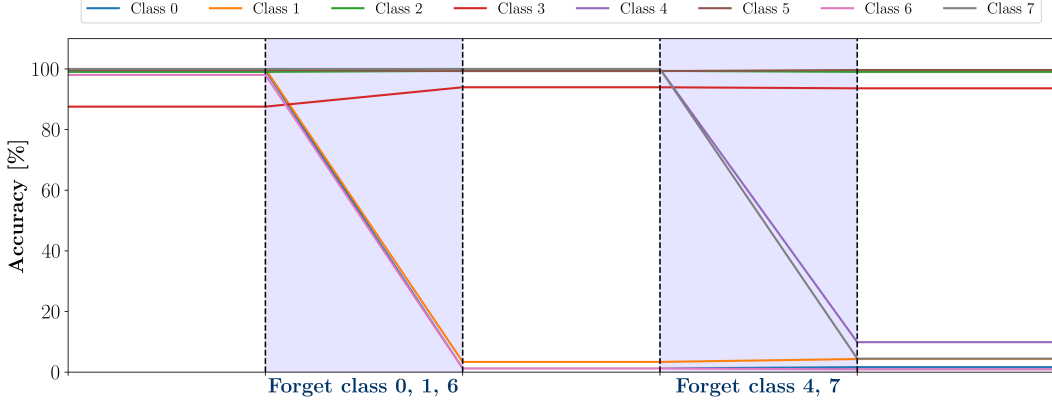


Figure 5.2: **Accuracy evolution during sequential inference.** Violet regions mark unlearning phases, where the corresponding classes (shown beneath) are forgotten by removing their prompts, causing a sharp drop in accuracy. Retained class accuracy, instead, remains stable throughout the entire inference phase.

When nearly all classes are removed ($f = C - 1$), our method continues to show robust retention with Acc_r values of 99.27% on BloodMNIST, 96.09% on DermaMNIST and 98.51% on OrganSMNIST, maintaining a high classification performance on the single retained class. This is significantly higher than SSD and Goel et al., both of whom experience considerable drops in retain accuracy. Regarding forgetting, our approach reduces Acc_f to 16.22% on BloodMNIST, 19.82% on DermaMNIST, and 11.80% on OrganSMNIST, effectively suppressing information about forgotten classes. In contrast, SSD and Goel et al. exhibit significantly higher Acc_f , exceeding 70%, suggesting they fail to erase knowledge of forgotten classes.

Across all experimental settings, LMeraser achieves stronger unlearning performance (i.e., lower Acc_f) and comparable retain accuracy (Acc_r) relative to our method. However, this advantage comes at a substantial cost. LMeraser requires a dedicated post hoc unlearning phase involving 10 full epochs of retraining for each update, which significantly increases training time and memory overhead. Specifically, with 30 centroids, it retrains an average of 14.9, 22.4, and 28.7 clusters across the BloodMNIST, DermaMNIST, and OrganSMNIST datasets for $f = 1$, $f = C/2$, and $f = C - 1$, respectively. As a consequence, the effective model size becomes $6.95\times$ larger when forgetting a single class ($f = 1$) and grows up to

13.34 $\times$ when forgetting all but one class ($f = C-1$).

In contrast, our method achieves effective unlearning with zero retraining—a key distinction that enables substantial gains in computational efficiency and deployment flexibility. Our retraining-free design supports on-the-fly forgetting during inference by simply removing the class-specific prompt tokens. As illustrated in Figure 5.2, this leads to an immediate and predictable drop in classification performance for forgotten classes, while preserving stable accuracy on retained ones. Moreover, as the number of forgotten classes increases, the unlearning performance gap between LMeraser and our method progressively narrows, confirming the robustness and scalability of our strategy under larger deletion scenarios. This, combined with the absence of retraining, minimal memory overhead, and modular structure, positions our method as a better solution for real-time deployments than LMEraser.

We also performed an ablation study on BloodMNIST to evaluate the impact of KL divergence, shuffling $S(\cdot)$, and prompt sampling $\sigma(\cdot)$ on unlearning performance as the number f of forgotten classes varies. When using only KL divergence, retain accuracy is significantly degraded ($\text{Acc}_r = 20.96\%$ for $f = 3$) while the forget accuracy remains high ($\text{Acc}_f = 24.10\%$ for $f = 1$), indicating that KL alone is insufficient for effective unlearning. Adding shuffling improves forgetting ($\text{Acc}_f = 16.04\%$ for $f = 1$) but still limits retention ($\text{Acc}_r = 49.25\%$ mean). The full method, incorporating prompt sampling, achieves the best trade-off, maintaining high retain accuracy while effectively reducing forget accuracy.

Table 5.4: **Ablation study** on the BloodMNIST dataset as the number f of forgotten classes varies

			f=1		f=3		f=7		Mean	
KL	$S(\cdot)$	$\sigma(\cdot)$	Acc _r	Acc _f	Acc _r	Acc _f	Acc _r	Acc _f	Acc _r	Acc _f
✓	✗	✗	71.79	24.10	20.96	18.03	23.43	17.72	38.73	19.95
✓	✓	✗	66.73	16.04	40.67	21.16	40.37	25.38	49.25	20.86
✓	✓	✓	97.61	10.54	98.42	12.73	99.27	16.22	98.43	13.16

Extending our prompt unlearning strategy to other audio modality, Table 5.5 summarizes the performance of our method on an audio classification task. Across various configurations, the method consistently achieves strong retain accuracy while effectively reducing performance on the forgotten classes. These results confirm that class-level unlearning objectives are met, while maintaining state-of-the-art performance across different forgetting scenarios.

Table 5.5: Unlearning performance comparison for AST model on UrbanSound8K dataset across different forget set scenarios ($f = 1$, $f = C/2$, and $f = C - 1$).

	UrbanSound8K	
	Acc_r	Acc_f
$f = 1$	87.98 $\pm$ 00.02	19.90 $\pm$ 00.21
$f = C/2$	91.49 $\pm$ 00.03	14.87 $\pm$ 00.06
$f = C - 1$	96.97 $\pm$ 00.03	15.27 $\pm$ 00.02

Privacy Results

To assess the privacy vulnerabilities of the trained model, we performed confidence-based Membership Inference Attack (MIA) [148] in order to determine whether a given sample was originally part of the model’s training set before class-based forgetting. This evaluation is essential for measuring privacy leaks after the forgetting step. As shown in Table 5.6, our method demonstrates near-perfect resistance across all datasets, indicating strong privacy guarantees.

Finally, we assess whether removing LoRa after training allows the prompts to retain functionality, potentially enabling a jailbreak, i.e., whether an adversary could bypass the intended forgetting mechanism by reusing learned prompts. Table 5.7 evaluates this scenario. The *Full Knowledge* model refers to a baseline trained with access to all classes and using both prompts and LoRa adapters, without any unlearning. In this setting, we observe that classification accuracy remains high even after the LoRa module is removed, indicating that learned knowledge has been partially absorbed by the prompts and the classification head. This suggests a vulnerability: residual information may remain accessible through prompts alone, posing a risk of unauthorized knowledge recovery.

In contrast, our method experiences a sharp accuracy drop under the same LoRa-removal condition, demonstrating that prompts by themselves are ineffective without the LoRa components. This behavior highlights a key security advantage of our design—class knowledge is tightly coupled to both the prompt tokens and their corresponding LoRa paths, and becomes inaccessible once LoRa is detached. As a result, even if prompts are exposed or reused, they do not yield meaningful predictions. This architecture reinforces our method’s robustness against jailbreak attempts and ensures stronger guarantees for secure, verifiable unlearning.

Table 5.6: **Resistance to Membership Inference Attack (MIA)**. Percentage probability of successful MIA in the case $f = C - 1$. Lower is better.

Method	BloodMNIST	DermaMNIST	OrganSMNIST
SSD [41]	12.38	36.71	32.98
Goel et al. [47]	00.00	00.32	00.12
LMEraser [172]	12.67	37.85	10.04
Ours	00.00	00.12	00.02

Table 5.7: **Impact of LoRa removal on classification accuracy**. Our method experiences a sharp accuracy drop, ensuring robustness against jailbreak attempts, while the baseline retains performance, indicating residual knowledge.

Method	BloodMNIST	DermaMNIST	OrganSMNIST
Full Knowledge	75.65	65.14	49.81
Ours	11.74	31.50	08.74

Importantly, while LMeraser achieves lower forget accuracy than our method—as shown in Tables 5.1, 5.2, and 5.3—this advantage comes with critical trade-offs. In addition to requiring access to training data and full retraining of affected clusters, LMeraser exhibits serious privacy vulnerabilities. As reported in Table 5.6, it is significantly more susceptible to membership inference attacks (MIA), with higher success rates across all datasets. In contrast, our method achieves near-perfect resistance to MIA, preventing recovery of forgotten information even under adversarial probing.

Inference-Time Uncertainty Estimation

As shown in Figure 5.3, prompt-subset variation separates correctly classified from incorrectly classified test images on H more sharply than MC dropout: the incorrect distribution is cleanly displaced toward high H , with limited overlap with the correct distribution’s tail. Median H for correct versus incorrect predictions is 0.021 versus 0.759 for prompt-subset, against 0.054 versus 0.676 for MC dropout.

Summarizing selective-prediction quality across thresholds, the area under the risk-coverage curve (AURC) is 0.985 for prompt-subset and 0.969 for MC dropout. At the three validation-calibrated targets defined in the setup, prompt-subset retains substantially more of the test set at matched accuracy (Figure 5.4): 89% coverage at 93.2% accuracy versus 73% at 94.5% (95% target); 84% at 96.5% versus 65% at 96.8% (97% target); 76% at 99.5% versus 36% at 99.9% (99.9% target). Appendix A.1 reports equivalent results for BloodMNIST and OrganSMNIST, and a sensitivity check showing that prompt-subset retains its advantage when MC dropout is run at $p = 0.3$.

Cumulatively, all results highlight the robustness of our approach, which combines efficient, retraining-free unlearning, jailbreak resistance, and a signal suitable for inference-time uncertainty calibratable that arises directly from the prompt structure. As a result, it is well-suited for real-world, regulation-compliant deployments where security and accountability are critical.

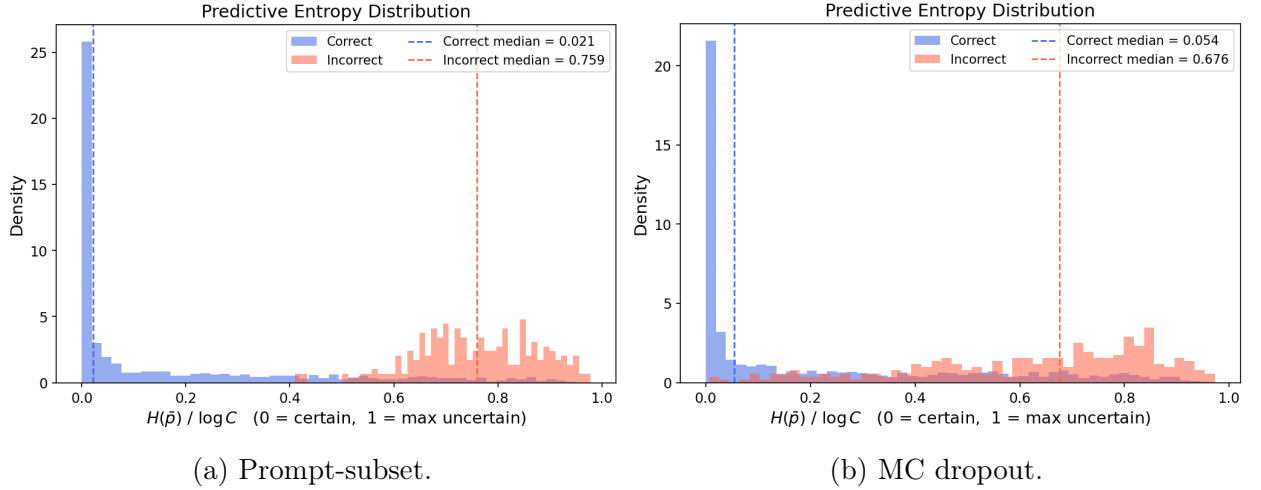
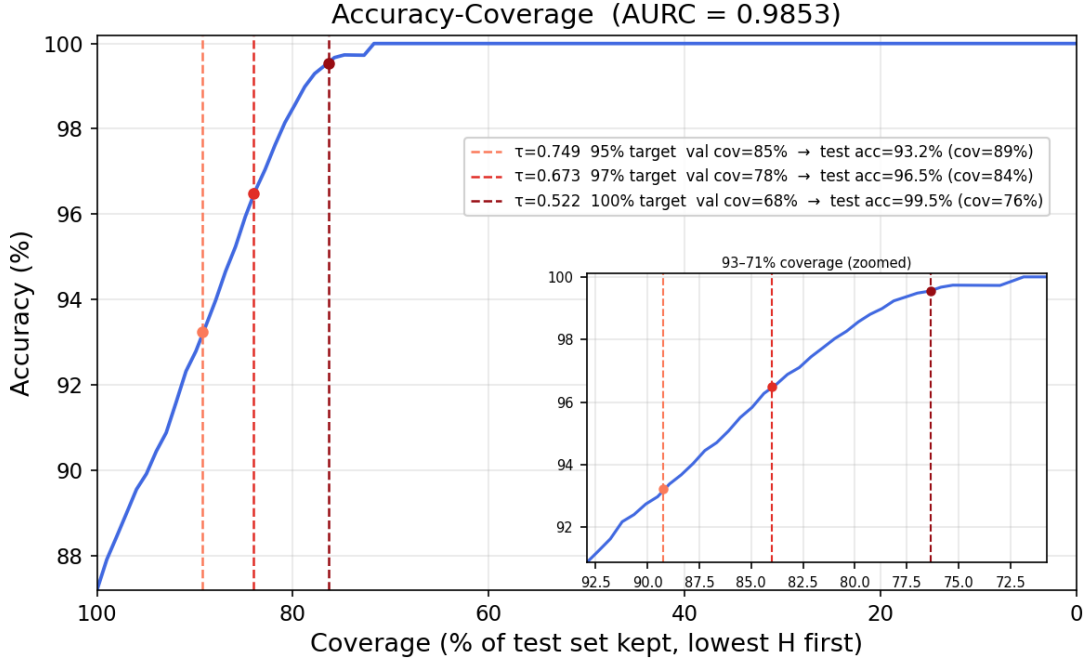
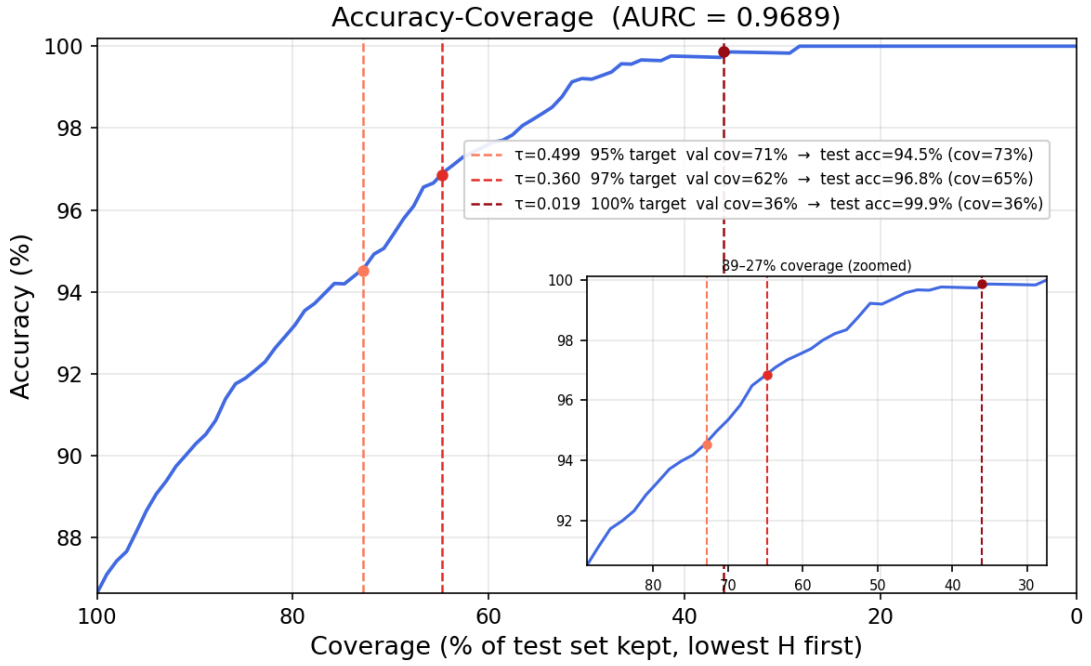


Figure 5.3: Distribution of H on DermaMNIST test set, separated by prediction correctness.



(a) Prompt-subset.



(b) MC dropout.

Figure 5.4: Accuracy versus coverage on DermaMNIST, with val-calibrated thresholds applied to test.

5.1.4 Discussion

Pre-Forgettable Models couple learning and unlearning within a single training phase by associating class-level knowledge with dedicated learnable prompt tokens on a frozen backbone. Unlearning a class reduces to removing the corresponding prompt, requiring no re-training, no access to the original training data, and no modification to the backbone. Across three MedMNIST v2 medical image classification benchmarks and an audio benchmark, the method produces near-random accuracy on forgotten classes while retaining the trained baseline on the others, and shows low membership-inference success rates relative to SSD, Goel et al., and LMEraser. The LMEraser comparison is the informative one: LMEraser reaches lower forget accuracy in some configurations, but does so by retraining the affected cluster on the original data and at substantially larger model size, with no mechanism for unlearning at request time. The same prompt structure also yields a calibratable inference-time uncertainty signal, stronger than the MC dropout baseline on the reported datasets and usable for per-sample clinical deferral. The underlying mechanism can be described as an ensemble of routings, where each forward pass conditions on a different prompt subset, inducing a slightly different attention pattern over the input tokens. This differs from MC dropout, whose ensemble members are produced by random masking rather than by a structured axis.

The limitation that follows from the Pre-Forgettable design is that the prompts must remain on a trusted server. An attacker with access to the deployed modules, inherits the attack regime familiar from large language models: prompt extraction, inversion, and reconstruction attacks that recover or reactivate suppressed behaviors.

The conceptual contribution of this section to the chapter’s *Forgetting* argument is the re-framing of the cost of unlearning to a pre-commitment at training time, resulting in instant unlearning by design. This instant, data-free property is a concrete step toward operationally compliant implementations of the GDPR’s right to be forgotten, for which post-hoc unlearning approaches struggle to provide a workable response time. The same architectural choice that makes unlearning instant also provides a usable per-sample deferral signal at inference, without any additional training or machinery.

5.1.5 Publications

Rutger Hendrix, Giovanni Patanè, Leonardo G. Russo, Simone Carnemolla, Federica Proietto Salanitri, Giovanni Bellitto, Concetto Spampinato, and Matteo Pennisi. Pre-forgettable models: Prompt learning as a native mechanism for unlearning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12446–12454, 2025. doi.org/10.1145/3700000.3700000

[145/3746027.3758171](#).

5.2 TrustErase: Auditable Instant Machine Unlearning with Passport-Embedded Representations

Pre-Forgettable Models (Section 5.1) showed that forgetting can be a native architectural property, achieved by deleting learned prompt tokens without retraining or access to training data. However, the prompt-based design exposes the model to prompt-based attacks, and it does not provide a way to verify which unlearning configuration a deployed model actually implements. This gap matters under the growing regulatory pressure for mechanisms that can guarantee auditable compliance of deployed models [4]. Parallel research on passport-based model verification [40, 117] has shown that embedding constant matrices within network layers can bind a model’s functionality to hidden credentials. Although originally conceived for copyright protection [117, 40], this paradigm suggests a promising direction for unlearning: coupling a model’s behavior with auditable structures.

Building on these insights, we propose ***TrustErase***, a data-free unlearning framework with authority-mediated audit. *TrustErase* extends the activate-at-inference paradigm of Pre-Forgettable Models and draws on two further lines of prior work. From SEAL [117], we adopt the passport-in-LoRA construction with SVD-based hiding and repurpose it for unlearning. From DELETE [191], we adapt the masked-distribution forgetting objective into single-stage multi-task training without a teacher. Our contributions are (i) joint multi-task training of shared LoRA factors against a passport library coexisting within one parameter space; (ii) a hypernetwork that composes atomic passports into credentials for unseen forget-set combinations without retraining or data access; and (iii) an authority-mediated audit binding each released adapter to its claimed unlearning configuration. The result is instant, modular, data-free forgetting with auditable evidence. Experimental results demonstrate that *TrustErase* matches or exceeds state-of-the-art unlearning methods on MNIST, CIFAR-10, and CIFAR-100, while operating in a strictly data-free regime at deployment. By shifting the computational burden to the pre-optimization phase, *TrustErase* establishes a *train-once, forget-anytime* paradigm in which compliance is achieved through passport deactivation rather than post-hoc retraining.

5.2.1 Related Work

Machine unlearning. Machine unlearning aims to remove the influence of specific data or concepts from trained models. Early works addressed this in traditional ML settings, where

Table 5.8: Comparison of core capabilities across representative unlearning approaches. A checkmark (✓) indicates native support for the corresponding feature.

Capability	L2UL [22]	Boundary Shrink [26]	DELETE [191]	Pre-Forgettable 5.1	<i>TrustErase</i> (Ours)
Instant unlearning	✗	✗	✗	✓	✓
Forget-data free	✗	✗	✓	✓	✓
Retain-data free	✗	✗	✗	✓	✓
Multi-setting adaptability	✓	✓	✓	✓	✓
Authority-mediated audit	✗	✗	✗	✗	✓

convex objectives allowed exact removal through analytical or incremental updates [46, 107, 7]. However, such formulations do not generalize to deep networks [20], where nonlinearity and overparameterization prevent efficient inversion. Unlearning for deep neural networks (DNNs) has therefore focused on *post-hoc* interventions. SISA [16] partitions the training data into independent shards, retraining only those affected by deletion requests. Subsequent methods accelerate unlearning via gradient tracking [156], parameter perturbation [41], or Fisher-based importance weighting [48], yet they still depend on stored gradients or access to specific data sets. Other strategies, such as Boundary Shrink [26] and L2UL [22], suppress forgotten information by regularizing latent decision boundaries or selectively constraining parameter updates, but they remain data-dependent. The recent DELETE framework [191] introduced a decoupled distillation process that explicitly separates forgetting and retention terms, enabling strong performance across class-centric tasks without access to remaining data. However, DELETE still requires multiple optimization phases and teacher–student distillation, limiting its use for instant or data-free unlearning. In contrast, Pre-Forgettable Models (Section 5.1) reinterpret forgetting as an architectural property: learned prompt tokens corresponding to target classes can be removed to achieve immediate forgetting. While efficient and data-free, such prompt-based methods, as well as the other existing strategies, lack auditable guarantees: once models are deployed, there is no mechanism to ensure that forgotten knowledge cannot be reactivated.

Parameter merging. Task arithmetic [72] showed that task-specific weight differences can be linearly combined in parameter space to compose or remove model behaviors. The broader model merging literature [174] extends this to handling parameter interference across many tasks. These observations motivate our hypernetwork-based [55, 106] passport composition, which exploits parameter-space additivity in the low-rank subspace to synthesize new unlearning configurations without retraining.

Passport-based verification. Parallel efforts in model watermarking and ownership verification have introduced *passport-based* schemes [40, 117]. These embed non-trainable ma-

trices (passports) between LoRA weights to entangle model functionality with hidden credentials that can later be used for verification. SEAL [117], for instance, integrates constant passports into LoRA adapters and hides them through singular-value decomposition, enabling ownership proof without accuracy degradation.

However, passport methods target intellectual-property protection rather than data removal; they certify ownership but not forgetting. We unify these research threads by embedding passport matrices within LoRA adapters to realize an *auditable instant unlearning* framework. Here, passports serve dual roles: as containers of task-specific knowledge and as hidden credentials enabling authority-mediated audit of the applied forgetting. This results in a unique approach in the unlearning landscape, as shown in Table 5.8. *TrustErase* in particular, is the only method that combines instant unlearning (removal without retraining), independence from both forget and retain data, and a built-in audit mechanism. Moreover, it supports multi-setting adaptability, not only across different forgetting granularities, such as instance-level and class-level unlearning, but also across diverse unlearning objectives. This flexibility makes *TrustErase* applicable in a wide range of operational and regulatory scenarios.

5.2.2 Method

We introduce *TrustErase*, a framework for auditable machine unlearning that leverages passport-modulated LoRA adapters to encode multiple unlearning behaviors within a single model. Each unlearning task is associated with a fixed passport matrix that modulates shared LoRA parameters, enabling selective forgetting without retraining.

As illustrated in Figure 5.5, the framework consists of three stages: multi-task training with task-specific passports, passport hiding through matrix decomposition before model distribution, and an audit protocol that reconstructs and validates the embedded passport.

The remainder of this section details the components of the proposed framework. We first review the passport mechanism and its integration with LoRA adapters, then describe the unlearning formulation and training objective. Next, we introduce a hypernetwork for compositional passport synthesis to enable multi-class forgetting on-the-fly. Finally, we describe the audit procedure.

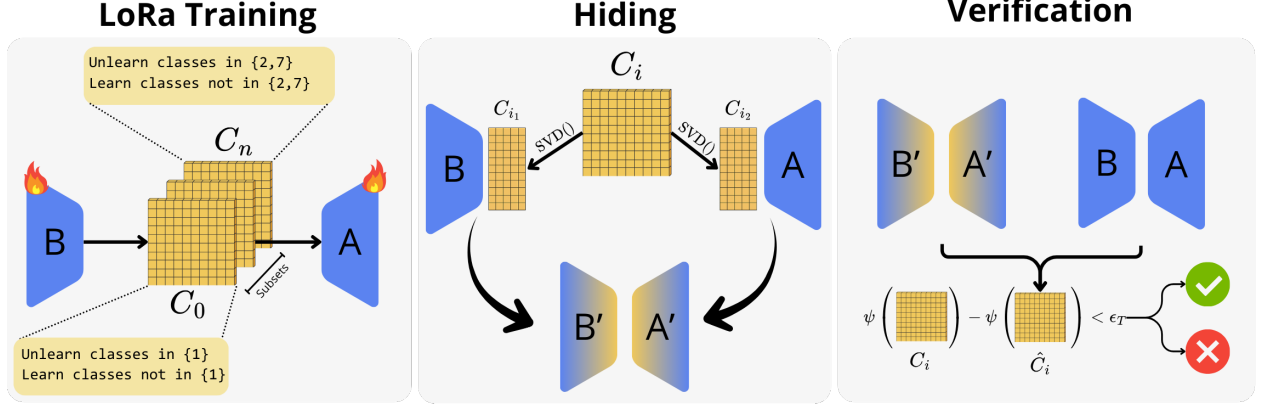


Figure 5.5: **Overview of *TrustErase* for Auditable Machine Unlearning.** *Left:* Multi-task training with task-specific passports $\mathbf{C}_i$ inserted between shared LoRA matrices $\mathbf{A}$ and $\mathbf{B}$, enabling simultaneous optimization across multiple unlearning objectives. *Middle:* Post-training passport hiding via SVD decomposition ($\mathbf{C}_i = \mathbf{C}_{1,i} \mathbf{C}_{2,i}$), where factors are absorbed into public weights $\mathbf{B}'_i = \mathbf{B} \mathbf{C}_{1,i}$ and $\mathbf{A}'_i = \mathbf{C}_{2,i} \mathbf{A}$. *Right:* Audit protocol reconstructs the embedded passport $\hat{\mathbf{C}}_i = \mathbf{B}'^\dagger \mathbf{B}'_i \mathbf{A}'_i \mathbf{A}^\dagger$ from public adapter and private base matrices, certifying compliance when $\|\psi(\mathbf{C}_i) - \psi(\hat{\mathbf{C}}_i)\| < \epsilon_T$.

Background

Passports were first introduced in [40] as constant matrices embedded within neural networks to enable model ownership verification. When integrated into training, a passport acts as a hidden credential that entangles the model parameters with a secret key: the network performs correctly only when the valid passport is present, and its accuracy degrades if the passport is removed or replaced. The SEAL framework [117] extended this concept to parameter-efficient finetuning by embedding passports within the *Low-Rank Adaptation* (LoRA) formulation [66]. In standard LoRA, the task-specific update of a pretrained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d \times k}$ is defined as

$$\Delta \mathbf{W} = \mathbf{B} \mathbf{A}, \quad (5.9)$$

with $\mathbf{B} \in \mathbb{R}^{d \times r}$, $\mathbf{A} \in \mathbb{R}^{r \times k}$, and $r \ll \min(d, k)$. SEAL introduces a non-trainable, constant passport matrix $\mathbf{C}$ between them:

$$\Delta \mathbf{W} = \mathbf{B} \mathbf{C} \mathbf{A}, \quad (5.10)$$

thereby *entangling* $\mathbf{C} \in \mathbb{R}^{r \times r}$ with the trainable parameters $\mathbf{A}$ and $\mathbf{B}$. Before model distribution, the passport is hidden through a decomposition function $f : \mathbf{C} \mapsto (\mathbf{c}_1, \mathbf{c}_2)$ such that $\mathbf{c}_1 \times \mathbf{c}_2 = \mathbf{C}$. A common choice is the singular value decomposition (SVD), $\mathbf{C} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top$, yielding $\mathbf{c}_1 = \mathbf{U} \times \mathbf{\Sigma}^{1/2}$ and $\mathbf{c}_2 = \mathbf{\Sigma}^{1/2} \times \mathbf{V}^\top$. The decomposed components are then absorbed

into the LoRA weights:

$$\mathbf{B}' = \mathbf{B} \times \mathbf{c}_1 \quad \text{and} \quad \mathbf{A}' = \mathbf{c}_2 \times \mathbf{A}, \quad (5.11)$$

so that the distributed model $(\mathbf{B}', \mathbf{A}')$ is indistinguishable from a standard LoRA adapter, while still implicitly containing the hidden passport.

Verification relies on the fact that only the legitimate owner, who possesses the original $\mathbf{A}$ and $\mathbf{B}$, can reconstruct the passport using the Moore–Penrose pseudoinverses $\mathbf{A}^\dagger$ and $\mathbf{B}^\dagger$:

$$\hat{\mathbf{C}} = \mathbf{B}^\dagger \times \mathbf{B}' \times \mathbf{A}' \times \mathbf{A}^\dagger. \quad (5.12)$$

If $\hat{\mathbf{C}}$ statistically matches the claimed passport $\mathbf{C}$, ownership (or authorization) is verified. This mechanism establishes a binding between model performance and the presence of the correct passport: a property that we exploit in this work to achieve *auditable instant unlearning*.

Auditable Unlearning Framework

TrustErase extends the role of passports from ownership verification to *machine unlearning*, proposing a framework where a single Vision Transformer (ViT) expresses multiple unlearning behaviors through task-specific passports modulating the LoRA updates (Eq. 5.10). This modification is applied to the query (Q) and value (V) projection layers within each self-attention block. While $\mathbf{A}$ and $\mathbf{B}$ remain shared and trainable across all tasks, swapping $\mathbf{C}$ allows multiple unlearning configurations to coexist within a single parameter space.

Task Unlearning Definition

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote the complete training set. We define the space of potential forgetting configurations as the power set $\mathcal{P}(\mathcal{D})$. In practice, a finite collection of unlearning tasks is considered during training:

$$\mathcal{S} = \{S_1, S_2, \dots, S_K\} \subseteq \mathcal{P}(\mathcal{D}), \quad (5.13)$$

where S_k represents the data to forget for task k , and $\bar{S}_k = \mathcal{D} \setminus S_k$ the data to retain. Each task k is associated with a fixed passport $\mathbf{C}_k \in \mathbb{R}^{r \times r}$, initialized randomly and kept frozen throughout training, yielding a library $\mathcal{C} = \{\mathbf{C}_1, \dots, \mathbf{C}_K\}$ of task-specific keys. During training, $\mathbf{A}$ and $\mathbf{B}$ are shared and updated jointly across all tasks, while each $\mathbf{C}_k$ remains fixed. The adapted forward computation for task k is:

$$\mathbf{h}_k = (\mathbf{W}_0 + \mathbf{B}\mathbf{C}_k\mathbf{A}) \mathbf{x}. \quad (5.14)$$

By switching the active passport $\mathbf{C}_k$, the same model can forget specific classes, instances, or arbitrary data subsets without altering its base parameters or accessing the original data.

Learning and Unlearning Optimization.

In line with Section 5.1, *TrustErase* jointly optimizes $(\mathbf{A}, \mathbf{B})$ to retain general task performance while embedding the ability to selectively forget specific data subsets through the fixed passports $\{\mathbf{C}_k\}$, enabling instant unlearning without access to the original data. For each passport, the objective combines a *forgetting* and a *retaining* loss.

Forgetting Loss (L_{forget}). Inspired by DELETE [191], we adopt a masking strategy that redistributes probability mass away from forgotten classes onto retained ones. Unlike DELETE, which has post-hoc access to frozen teacher, we apply it in a single training stage using the current model’s logits. Let $\mathbf{z}(x; \theta)$ be the logits from the current model and S_k the indices of classes or samples to forget. We define a masking operator:

$$\text{Mask}_{S_k}(\mathbf{z})_i = \begin{cases} \mathbf{z}_i, & i \notin S_k, \\ -\infty, & i \in S_k. \end{cases} \quad (5.15)$$

The masked distribution is then:

$$p^*(y|x, \mathbf{C}_k) = \text{Softmax}(\text{Mask}_{S_k}(\mathbf{z}(x; \theta))). \quad (5.16)$$

The forgetting loss minimizes the KL divergence between this masked distribution and the current model prediction:

$$L_{\text{forget}} = \text{KL}(p^*(y|x, \mathbf{C}_k) \parallel q_\theta(y|x, \mathbf{C}_k)). \quad (5.17)$$

Retaining Loss (L_{retain}). To preserve discriminative ability on the remaining data, we apply standard cross-entropy over classes not in S_k :

$$L_{\text{retain}} = -\frac{1}{|\bar{S}_k|} \sum_{j \notin S_k} \log p_\theta(y_j|x, \mathbf{C}_k). \quad (5.18)$$

The overall objective for task k combines forgetting and retention:

$$L_k = \lambda L_{\text{forget}} + (1 - \lambda) L_{\text{retain}}, \quad (5.19)$$

where $\lambda \in [0, 1]$ balances the two terms. A single backward pass aggregates per-task objectives across all passports, resulting in:

$$L_{\text{total}} = \frac{1}{K} \sum_{k=1}^K L_k. \quad (5.20)$$

Hypernetwork-based Passport Composition

The framework described so far requires defining each unlearning configuration *a priori*. For example, one passport may be trained to forget class 1 and another to forget class 2. If a new configuration such as forgetting both classes 1 and 2 simultaneously is needed, a dedicated passport must normally be trained for that specific combination. To avoid retraining for every possible combination, we introduce a *hypernetwork-based passport synthesis* mechanism, as depicted in Fig. 5.6 that learns to compose new passports from the existing atomic ones.

Atomic Passports. During base training, *TrustErase* learns a compact set of *atomic passports*, each corresponding to a forgetting task while retaining all others. For a C -class problem, this produces C base passports $\{C_1, \dots, C_C\}$, which act as compositional building blocks.

Hypernetwork Composition. After training all atomic passports, a simple hypernetwork $\mathcal{H}_\phi$ is trained *once* to synthesize composite passports by combining a given selection

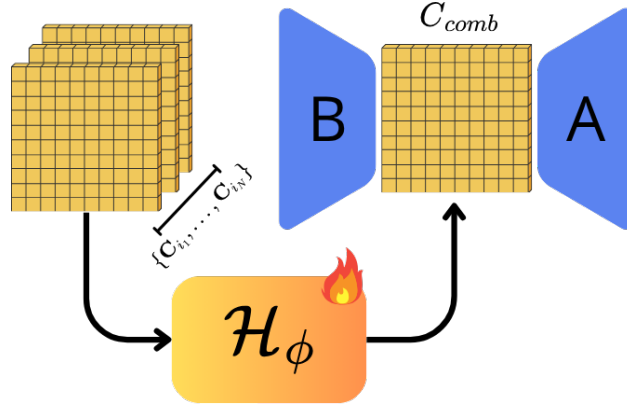


Figure 5.6: **Hypernetwork-based Passport Composition.** Atomic passports are sampled from the power set and fed into a hypernetwork to synthesize composite passports. The hypernetwork is optimized end-to-end with frozen LoRA parameters and base passports.

of atomic ones $\{\mathbf{C}_{i_1}, \dots, \mathbf{C}_{i_N}\}$ associated with the classes or samples to be forgotten:

$$\mathbf{C}_{\text{comb}} = \mathcal{H}_\phi(\mathbf{C}_{i_1}, \dots, \mathbf{C}_{i_N}). \quad (5.21)$$

During this stage, $\mathbf{A}$ and $\mathbf{B}$ remain frozen, and $\mathcal{H}_\phi$ is optimized end-to-end using the same joint objective as base training (Eq. 5.17, Eq. 5.18). The choice of a hypernetwork is motivated by the additivity of behavioral composition in parameter space, here operating in the low-rank passport space, as seen in task arithmetic and model merging literature [72, 174]. To handle variable-sized subsets, $\mathcal{H}_\phi$ leverages a Transformer encoder with padding masks; architectural details are provided in the Appendix B.1.3.

Passport Verification and ϵ -Tolerance

The passports $\{\mathbf{C}_k\}$ encode sensitive information, as each acts as a key controlling which unlearning configuration the model performs. Distributing these matrices directly would compromise both the integrity of the unlearning mechanism and the privacy of the authorization process. Instead, we apply the SVD decomposition introduced in Section 5.2.2 as a privacy mechanism, absorbing the passport factors into $(\mathbf{A}'_k, \mathbf{B}'_k)$ such that the distributed model is indistinguishable from a standard LoRA adapter. This ensures that the true passport cannot be inferred from the distributed weights alone, yet controlled audit by trusted entities remains possible through the original private matrices $(\mathbf{A}, \mathbf{B})$. Given access to the original private matrices, an estimated passport $\hat{\mathbf{C}}_k$ can be reconstructed following equation 5.12. Verification then consists of checking whether $\hat{\mathbf{C}}_k$ faithfully approximates the original passport $\mathbf{C}_k$ under a suitable distance metric $d(\cdot)$:

$$\|\psi(\hat{\mathbf{C}}_k) - \psi(\mathbf{C}_k)\| < \epsilon_T, \quad (5.22)$$

where $\psi(\cdot)$ denotes an arbitrary feature mapping and ϵ_T is a predefined tolerance accounting for numerical imprecision. If this condition holds, the adapter $(\mathbf{A}'_k, \mathbf{B}'_k)$ is certified as correctly implementing the unlearning configuration associated with $\mathbf{C}_k$, providing a lightweight and privacy-preserving form of *structural verification*. To strengthen assurance, the procedure can be complemented by a *functional verification* step. Here, one compares the behavior of the reconstructed model $(\mathbf{W}_0 + \mathbf{B}'_k \mathbf{A}'_k)$ with that of an ideal reference $(\mathbf{W}_0 + \mathbf{B} \mathbf{C}_k \mathbf{A})$ on a held-out set, assessing their agreement in terms of logits, intermediate activations, or task-specific metrics. When both structural and functional checks fall within the ϵ -tolerance, the model is deemed to have the valid unlearning configuration.

5.2.3 Experimental Results

Experimental Setup

We evaluate *TrustErase* along four complementary dimensions: *single-class unlearning*, assessing erasure of one semantic concept while preserving the others; *multi-class unlearning*, testing scalability as the number of forgotten classes increases (forget-set sizes 2, 5, 10, and 20 on CIFAR-100); *compositional generalization*, analyzing the hypernetwork’s ability to synthesize passports for unseen class combinations; and *audit robustness*, evaluating tolerance to passport perturbations. Experiments use MNIST, CIFAR-10, and CIFAR-100 on Vision Transformer backbones (ViT-Tiny, ViT-Small, ViT-Base); the framework is architecture-agnostic, and VGG-16 results are reported in Appendix B.4. We compare against unlearning methods compatible with transformer architectures: *Gradient Ascent*, *Fisher Forgetting* [48], *Boundary Shrink/Expand* [26], *Learn-to-Unlearn (L2UL)* [22], *Saliency Unlearn (SalUn)* [39], *Bad Teacher* [28], *Amnesiac* [52], *UNSIR* [155], and *DELETE* [191], with *Finetune*, *Random Label*, *Retrain* (upper bound), and *Original* (lower bound) as reference points. *TrustErase* and the hypernetwork are trained with AdamW and cosine annealing for 10 epochs each, with LoRA modules applied to the query and value projections. Full training details, hyperparameter search, reconstruction attacks, instance-level forgetting, extended quantitative tables across methods, objectives, architectures, time-performance evaluation, and audit protocol use cases are detailed in the Appendix B.

Metrics. Following prior work [191, 26], we report accuracy on the forgotten test set (Acc_{ft}), accuracy on the retained test set (Acc_{rt}), and *Membership Inference Attack* (MIA) success rate [148], measuring residual memorization; for MIA we use a logistic regressor implementation [41]. Since the formal goal of unlearning is indistinguishability from a model retrained without the forget set we additionally report the distance to a retrained reference model,

$$\text{dist} = |Acc_{rt} - Acc_{rt, \text{oracle}}| + |Acc_{ft} - Acc_{ft, \text{oracle}}| + |MIA - MIA_{\text{oracle}}|, \quad (5.23)$$

where the oracle subscript denotes the value obtained by the retrained reference and accuracies are taken as fractions. The reference is retrained from scratch on the retained data within the same model family as the method under evaluation. Lower dist indicates closer behavioural agreement with retraining. To compare against *Pre-Forgettable Models* 5.1, we additionally report the average forgetting gap (Gap_f), defined as the distance to the target Acc , and its maximum deviation ($Gap_f^{\max}$). Audit robustness is quantified through *Passport*

Table 5.9: **Single-class unlearning** performance on CIFAR-10 (LoRA rank 32) and CIFAR-100 (LoRA rank 64) for ViT-Small (3 seeds, mean $\pm$ std). Dist is the aggregate distance to the retrain reference (Eq. 5.23); lower is better.

Method	CIFAR10				CIFAR100			
	Acc_{ft}	Acc_{rt}	MIA	dist $\downarrow$	Acc_{ft}	Acc_{rt}	MIA	dist $\downarrow$
<i>Retrain (full)</i>	0.0	98.1 $\pm$ 0.1	0.131 $\pm$ 0.01	0.00	0.0	90.7 $\pm$ 0.3	0.244 $\pm$ 0.04	0.00
<i>Retrain (LoRA)</i>	0.0	98.3 $\pm$ 0.1	0.108 $\pm$ 0.02	0.00	0.0	90.5 $\pm$ 0.2	0.257 $\pm$ 0.01	0.00
Finetune	0.0 $\pm$ 0.0	91.1 $\pm$ 3.8	0.300 $\pm$ 0.06	0.24 $\pm$ 0.10	0.0 $\pm$ 0.0	82.0 $\pm$ 0.0	0.261 $\pm$ 0.05	0.12 $\pm$ 0.03
Random Labels	14.0 $\pm$ 0.9	95.3 $\pm$ 0.9	0.333 $\pm$ 0.58	0.55 $\pm$ 0.43	4.0 $\pm$ 3.0	88.9 $\pm$ 0.1	0.007 $\pm$ 0.01	0.30 $\pm$ 0.04
Gradient Ascent	0.5 $\pm$ 0.7	89.2 $\pm$ 5.0	0.936 $\pm$ 0.03	0.90 $\pm$ 0.07	83.3 $\pm$ 2.1	90.5 $\pm$ 0.1	0.797 $\pm$ 0.04	1.39 $\pm$ 0.04
Fisher	32.9 $\pm$ 52.7	7.6 $\pm$ 6.0	0.346 $\pm$ 0.15	1.44 $\pm$ 0.49	0.0 $\pm$ 0.0	0.8 $\pm$ 0.3	0.519 $\pm$ 0.23	1.17 $\pm$ 0.22
Delete	0.0 $\pm$ 0.1	97.8 $\pm$ 0.0	0.000 $\pm$ 0.00	0.14 $\pm$ 0.00	5.0 $\pm$ 2.0	89.4 $\pm$ 0.3	0.081 $\pm$ 0.05	0.23 $\pm$ 0.06
Bad Teacher	68.8 $\pm$ 17.2	97.4 $\pm$ 0.2	0.001 $\pm$ 0.00	0.83 $\pm$ 0.17	82.3 $\pm$ 1.5	90.5 $\pm$ 0.1	0.765 $\pm$ 0.02	1.35 $\pm$ 0.00
Boundary Expand	4.5 $\pm$ 4.0	59.0 $\pm$ 11.0	0.620 $\pm$ 0.52	1.00 $\pm$ 0.29	0.0 $\pm$ 0.0	46.2 $\pm$ 19.6	0.367 $\pm$ 0.48	0.77 $\pm$ 0.47
Boundary Shrink	10.7 $\pm$ 4.8	93.3 $\pm$ 1.0	0.324 $\pm$ 0.52	0.49 $\pm$ 0.35	83.3 $\pm$ 1.2	90.5 $\pm$ 0.2	0.817 $\pm$ 0.03	1.41 $\pm$ 0.03
SalUn	12.2 $\pm$ 3.8	95.7 $\pm$ 1.0	0.333 $\pm$ 0.58	0.52 $\pm$ 0.40	7.7 $\pm$ 3.8	89.3 $\pm$ 0.3	0.016 $\pm$ 0.02	0.32 $\pm$ 0.05
L2UL (adv)	2.8 $\pm$ 2.8	92.5 $\pm$ 4.5	0.629 $\pm$ 0.28	0.58 $\pm$ 0.31	84.7 $\pm$ 1.5	90.5 $\pm$ 0.1	0.847 $\pm$ 0.03	1.45 $\pm$ 0.02
L2UL (imp)	2.8 $\pm$ 2.8	92.5 $\pm$ 4.5	0.631 $\pm$ 0.28	0.58 $\pm$ 0.31	84.3 $\pm$ 1.5	90.5 $\pm$ 0.1	0.847 $\pm$ 0.03	1.45 $\pm$ 0.01
UNSIR	0.1 $\pm$ 0.1	73.4 $\pm$ 1.0	0.266 $\pm$ 0.23	0.47 $\pm$ 0.10	48.3 $\pm$ 11.4	85.0 $\pm$ 1.1	0.421 $\pm$ 0.11	0.72 $\pm$ 0.22
Amnesiac	0.0 $\pm$ 0.0	98.3 $\pm$ 0.0	0.000 $\pm$ 0.00	0.13 $\pm$ 0.00	0.0 $\pm$ 0.0	90.6 $\pm$ 0.0	0.001 $\pm$ 0.00	0.25 $\pm$ 0.00
Ours	0.0 $\pm$ 0.0	98.0 $\pm$ 0.2	0.063 $\pm$ 0.06	0.06$\pm$0.03	0.0 $\pm$ 0.0	88.1 $\pm$ 0.2	0.341 $\pm$ 0.06	0.11$\pm$0.06

Verification, which measures whether perturbed or invalid passports are correctly rejected within an ϵ_T -tolerant bound.

Unlearning performance

We evaluate the ability of *TrustErase* to forget one or multiple semantic classes while preserving accuracy on the remaining ones. A single passport based LoRA model is trained once to encode all unlearning configurations jointly. Each class or subset is associated with a specific passport, and forgetting is achieved by deactivating the corresponding passports, without any retraining or access to data.

Single Class Unlearning. This experiment measures how effectively the model can erase one class while retaining knowledge of the others. Table 5.9 reports results on CIFAR-10 and CIFAR-100. TrustErase reaches complete forgetting ($Acc_{ft} = 0.0 \pm 0.0$) on both datasets, with $98.0 \pm 0.2\%$ retained accuracy on CIFAR-10 and $88.1 \pm 0.2\%$ on CIFAR-100, and obtains the smallest distance to the retrain reference of all methods (dist = 0.06 and 0.11, respectively). Competitors such as Amnesiac and Delete instead drive MIA to zero, far below the retrain level, which Eq. 5.23 penalises as over-forgetting. On CIFAR-100 the advantage is significant (Welch’s t-test, $p < 0.05$) against 12 of the 13 competitors (all but Finetune, $p \leq 0.035$ in every case); on CIFAR-10 against 5 of the 13 (Gradient Ascent, UNSIR, Bad Teacher, Boundary Expand, and Fisher).

Table 5.10: **Multi-class unlearning** performance on CIFAR-100 for varying numbers of forgotten classes using ViT-Small with LoRA rank 64 (3 seeds, mean $\pm$ std). Our method trains once to handle all subset combinations simultaneously. Dist is the aggregate distance to the retrain reference (Eq. 5.23); lower is better.

Method	2 Classes				5 Classes			
	Acc _{ft}	Acc _{rt}	MIA	dist $\downarrow$	Acc _{ft}	Acc _{rt}	MIA	dist $\downarrow$
<i>Retrain (full)</i>	0.0	90.6 $\pm$ 0.0	0.241 $\pm$ 0.03	0.00	0.0	91.2 $\pm$ 0.1	0.178 $\pm$ 0.01	0.00
<i>Retrain (LoRA)</i>	0.0	90.3 $\pm$ 0.1	0.219 $\pm$ 0.01	0.00	0.0	90.5 $\pm$ 0.2	0.203 $\pm$ 0.01	0.00
Finetune	0.0 $\pm$ 0.0	81.6 $\pm$ 0.6	0.390 $\pm$ 0.03	0.24 $\pm$ 0.02	0.0 $\pm$ 0.0	82.2 $\pm$ 0.6	0.264 $\pm$ 0.03	0.18 $\pm$ 0.02
Random Labels	1.7 $\pm$ 0.8	86.9 $\pm$ 0.2	0.001 $\pm$ 0.00	0.29 $\pm$ 0.01	2.7 $\pm$ 1.2	69.8 $\pm$ 9.1	0.001 $\pm$ 0.00	0.42 $\pm$ 0.10
Gradient Ascent	85.7 $\pm$ 2.8	90.4 $\pm$ 0.1	0.787 $\pm$ 0.05	1.41 $\pm$ 0.07	68.8 $\pm$ 1.0	90.0 $\pm$ 0.2	0.757 $\pm$ 0.01	1.28 $\pm$ 0.02
Fisher	0.0 $\pm$ 0.0	0.8 $\pm$ 0.3	0.631 $\pm$ 0.13	1.29 $\pm$ 0.13	0.0 $\pm$ 0.0	0.8 $\pm$ 0.3	0.347 $\pm$ 0.12	1.07 $\pm$ 0.12
Delete	6.2 $\pm$ 2.5	89.1 $\pm$ 0.2	0.018 $\pm$ 0.01	0.30 $\pm$ 0.03	4.7 $\pm$ 1.4	89.6 $\pm$ 0.3	0.009 $\pm$ 0.00	0.23 $\pm$ 0.02
Bad Teacher	85.5 $\pm$ 0.9	90.4 $\pm$ 0.2	0.699 $\pm$ 0.03	1.32 $\pm$ 0.04	64.4 $\pm$ 2.4	90.3 $\pm$ 0.2	0.309 $\pm$ 0.04	0.78 $\pm$ 0.06
Boundary Expand	0.0 $\pm$ 0.0	32.5 $\pm$ 2.4	0.057 $\pm$ 0.02	0.77 $\pm$ 0.04	0.4 $\pm$ 0.4	15.2 $\pm$ 5.7	0.076 $\pm$ 0.07	0.87 $\pm$ 0.12
Boundary Shrink	86.3 $\pm$ 0.8	90.3 $\pm$ 0.1	0.760 $\pm$ 0.05	1.39 $\pm$ 0.06	66.2 $\pm$ 3.1	89.0 $\pm$ 0.9	0.401 $\pm$ 0.04	0.91 $\pm$ 0.06
SalUn	3.7 $\pm$ 0.6	87.9 $\pm$ 0.4	0.001 $\pm$ 0.00	0.30 $\pm$ 0.00	1.7 $\pm$ 0.6	71.5 $\pm$ 7.7	0.001 $\pm$ 0.00	0.39 $\pm$ 0.08
L2UL (adv)	88.7 $\pm$ 1.3	90.4 $\pm$ 0.1	0.852 $\pm$ 0.02	1.50 $\pm$ 0.01	76.1 $\pm$ 0.8	90.4 $\pm$ 0.2	0.641 $\pm$ 0.01	1.23 $\pm$ 0.01
L2UL (imp)	88.7 $\pm$ 1.3	90.4 $\pm$ 0.1	0.851 $\pm$ 0.02	1.50 $\pm$ 0.01	76.0 $\pm$ 0.7	90.4 $\pm$ 0.2	0.640 $\pm$ 0.01	1.23 $\pm$ 0.01
UNSIR	73.3 $\pm$ 15.3	87.2 $\pm$ 2.3	0.704 $\pm$ 0.15	1.23 $\pm$ 0.28	72.4 $\pm$ 2.3	89.2 $\pm$ 0.6	0.582 $\pm$ 0.07	1.15 $\pm$ 0.09
Amnesiac	0.2 $\pm$ 0.3	90.5 $\pm$ 0.3	0.001 $\pm$ 0.00	0.24 $\pm$ 0.00	0.8 $\pm$ 0.2	90.8 $\pm$ 0.1	0.001 $\pm$ 0.00	0.19 $\pm$ 0.00
Ours	0.0 $\pm$ 0.0	88.2 $\pm$ 0.4	0.362 $\pm$ 0.09	0.17$\pm$0.09	0.0 $\pm$ 0.0	88.8 $\pm$ 0.4	0.183 $\pm$ 0.01	0.04$\pm$0.01

Method	10 Classes				20 Classes			
	Acc _{ft}	Acc _{rt}	MIA	dist $\downarrow$	Acc _{ft}	Acc _{rt}	MIA	dist $\downarrow$
<i>Retrain (full)</i>	0.0	91.2 $\pm$ 0.0	0.155 $\pm$ 0.01	0.00	0.0	92.4 $\pm$ 0.2	0.187 $\pm$ 0.01	0.00
<i>Retrain (LoRA)</i>	0.0	90.9 $\pm$ 0.2	0.159 $\pm$ 0.00	0.00	0.0	92.2 $\pm$ 0.1	0.195 $\pm$ 0.01	0.00
Finetune	0.0 $\pm$ 0.0	82.2 $\pm$ 0.6	0.232 $\pm$ 0.02	0.17 $\pm$ 0.02	0.0 $\pm$ 0.0	83.6 $\pm$ 0.4	0.236 $\pm$ 0.03	0.14 $\pm$ 0.03
Random Labels	4.1 $\pm$ 3.2	38.6 $\pm$ 6.2	0.335 $\pm$ 0.56	0.94 $\pm$ 0.43	3.2 $\pm$ 1.1	9.7 $\pm$ 7.7	0.429 $\pm$ 0.51	1.23 $\pm$ 0.32
Gradient Ascent	70.3 $\pm$ 2.3	87.3 $\pm$ 0.9	0.799 $\pm$ 0.05	1.39 $\pm$ 0.06	0.0 $\pm$ 0.0	1.3 $\pm$ 0.0	0.000 $\pm$ 0.00	1.10 $\pm$ 0.00
Fisher	0.3 $\pm$ 0.3	0.8 $\pm$ 0.4	0.577 $\pm$ 0.18	1.33 $\pm$ 0.17	1.9 $\pm$ 2.7	0.5 $\pm$ 0.6	0.402 $\pm$ 0.17	1.15 $\pm$ 0.14
Delete	6.8 $\pm$ 1.4	89.0 $\pm$ 0.1	0.017 $\pm$ 0.01	0.23 $\pm$ 0.02	8.3 $\pm$ 1.1	89.9 $\pm$ 0.3	0.022 $\pm$ 0.00	0.27 $\pm$ 0.02
Bad Teacher	46.0 $\pm$ 13.9	80.7 $\pm$ 9.8	0.166 $\pm$ 0.05	0.61 $\pm$ 0.05	7.0 $\pm$ 3.2	20.9 $\pm$ 7.1	0.346 $\pm$ 0.38	1.02 $\pm$ 0.32
Boundary Expand	2.4 $\pm$ 2.6	6.8 $\pm$ 3.5	0.379 $\pm$ 0.07	1.09 $\pm$ 0.09	3.0 $\pm$ 1.8	2.8 $\pm$ 1.1	0.367 $\pm$ 0.32	1.23 $\pm$ 0.12
Boundary Shrink	18.8 $\pm$ 21.3	48.7 $\pm$ 25.5	0.495 $\pm$ 0.15	0.95 $\pm$ 0.19	6.0 $\pm$ 3.4	14.0 $\pm$ 5.6	0.524 $\pm$ 0.05	1.18 $\pm$ 0.07
SalUn	3.3 $\pm$ 2.4	37.5 $\pm$ 4.3	0.339 $\pm$ 0.57	0.95 $\pm$ 0.42	4.1 $\pm$ 1.6	15.8 $\pm$ 2.3	0.378 $\pm$ 0.46	1.12 $\pm$ 0.33
L2UL (adv)	73.2 $\pm$ 3.9	89.0 $\pm$ 0.7	0.641 $\pm$ 0.07	1.24 $\pm$ 0.10	1.7 $\pm$ 2.9	1.1 $\pm$ 1.0	0.301 $\pm$ 0.51	1.28 $\pm$ 0.34
L2UL (imp)	73.0 $\pm$ 4.0	89.1 $\pm$ 0.7	0.641 $\pm$ 0.07	1.24 $\pm$ 0.10	1.7 $\pm$ 2.9	1.1 $\pm$ 0.9	0.302 $\pm$ 0.51	1.29 $\pm$ 0.34
UNSIR	70.6 $\pm$ 5.0	87.4 $\pm$ 1.5	0.582 $\pm$ 0.08	1.17 $\pm$ 0.12	58.0 $\pm$ 7.6	86.2 $\pm$ 0.8	0.441 $\pm$ 0.06	0.90 $\pm$ 0.13
Amnesiac	3.3 $\pm$ 0.7	91.0 $\pm$ 0.0	0.000 $\pm$ 0.00	0.19 $\pm$ 0.01	9.7 $\pm$ 2.7	91.7 $\pm$ 0.1	0.001 $\pm$ 0.00	0.29 $\pm$ 0.03
Ours	0.0 $\pm$ 0.0	88.7 $\pm$ 0.3	0.148 $\pm$ 0.00	0.03$\pm$0.00	0.0 $\pm$ 0.0	90.1 $\pm$ 0.1	0.195 $\pm$ 0.00	0.03$\pm$0.00

Multi Class Unlearning. We then extend the analysis to forgetting several classes simultaneously. Forgetting is performed by deactivating the passports of the target classes, again with no additional training. As shown in Table 5.10, TrustErase achieves complete forgetting ($\text{Acc}_{ft} = 0.0 \pm 0.0$) across all configurations while maintaining stable retained accuracy: even when removing 20 classes on CIFAR-100, the model preserves $90.1 \pm 0.1\%$ test accuracy. Its distance to the oracle decreases from 0.17 at two classes to 0.03 at ten and twenty and is the smallest of all methods at every size, significantly so (Welch’s t-test, $p < 0.05$) against 8 of the 13 competitors at two classes, 13 of 13 at five and twenty, and 11 of 13 at ten. Unlike Finetune, L2UL, and Random Label, which require multiple opti-

mization steps, TrustErase obtains all configurations from a single trained model through a constant-time passport swap. These results demonstrate that TrustErase enables instant, auditable unlearning in both single and multi-class settings, matching the retrained model more closely than existing approaches while being fully data independent.

We extend our evaluation by comparing *TrustErase* with Pre-Forgettable Models 5.1, which perform modular unlearning by removing class-specific prompts. While our main experiments adopt the masked loss of equation 5.17 to suppress forgotten logits, Pre-Forgettable Models instead enforce a uniform distribution over them. To ensure a fair comparison, we re-train *TrustErase* using the same uniform objective, replacing the masked target distribution in equation 5.17 with a uniform one across all classes. Table 5.11 summarizes single-class unlearning results on MNIST and CIFAR-10 using ViT-Tiny and ViT-Small backbones with two LoRA ranks (R8 and R32), respectively. Unlike Tables 5.9 and 5.10, here we report the mean and standard deviation of all single-class unlearning tasks following the evaluation protocol of Pre-Forgettable Models. Across datasets and architectures, *TrustErase* achieves forgetting gaps comparable to or smaller than those of Pre-Forgettable Models while maintaining high retained accuracy, confirming that its passport mechanism is compatible with uniform-distribution unlearning. Unlike Pre-Forgettable Models, *TrustErase* additionally supports authority-mediated audit of the applied unlearning configuration.

Table 5.11: **Comparison with Pre-Forgettable Models** on MNIST and CIFAR-10 under the uniform-distribution forgetting objective. Both approaches employ LoRA adapters; results are shown for ViT-Tiny and ViT-Small backbones with two LoRA ranks (R8, R32). Reported values are mean and standard deviation averaged over all single-class unlearning tasks.

Method	MNIST				CIFAR10			
	Acc _{rt} ↑	Acc _{ft} ↓	Gap _{ft} ↓	Gap _f ^{max} ↓	Acc _{rt} ↑	Acc _{ft} ↓	Gap _{ft} ↓	Gap _f ^{max} ↓
ViT-Tiny								
Pre-Forgettable-R8	99.54±0.04	11.95±7.34	1.95	18.50	96.17±0.46	05.27±6.24	4.73	12.10
Pre-Forgettable-R32	99.50±0.04	07.46±6.98	2.54	15.58	96.66±0.41	04.95±2.62	5.05	0.80
Ours-R8	98.68±0.16	09.45±2.26	0.55	4.36	88.54±2.89	05.88±1.60	4.12	2.00
Ours-R32	99.28±0.09	09.47±2.71	0.53	3.66	93.32±0.52	09.04±2.10	0.96	2.20
ViT-Small								
Pre-Forgettable-R8	99.63±0.03	11.02±8.68	1.02	14.42	98.04±0.23	03.46±2.13	6.54	7.92
Pre-Forgettable-R32	99.64±0.05	07.51±5.30	2.49	7.92	98.32±0.20	05.40±4.97	4.60	8.60
Ours-R8	99.14±0.11	10.41±3.27	0.41	6.55	94.94±0.90	10.93±4.22	0.93	10.90
Ours-R32	99.44±0.09	10.61±3.11	0.61	8.02	96.32±0.42	11.40±2.24	1.40	5.40

We evaluate the hypernetwork that learns to synthesize composite passports by combining atomic single-class ones, enabling flexible unlearning without retraining or data access afterwards. For compositions of up to 9 unique atomic passports on CIFAR-10, the combination

Table 5.12: **Hypernetwork-based compositional unlearning on CIFAR-10.** Scalability across 100 seen and 100 unseen passport combinations, each composed of up to 9 atomic passports. Unseen combinations are excluded during training to evaluate generalization to novel unlearning configurations.

Setting	$\text{Acc}_{rt} \uparrow$	$\text{Acc}_{ft} \downarrow$	MIA $\downarrow$
Seen	98.92 ± 0.64	0.00 ± 0.00	0.250 ± 0.176
Unseen	96.07 ± 14.85	2.22 ± 11.11	0.240 ± 0.180

space grows combinatorially to over 600 possible combinations, making exhaustive evaluation infeasible. We therefore sample 100 seen and 100 unseen combinations (Table 5.12), where the unseen combinations are deliberately excluded during training to assess generalization. For seen configurations, all 100 achieve perfect forgetting with retained accuracy above 98%. For unseen configurations, 96 out of 100 reach perfect forgetting with retained accuracy above 96%, with only 4 cases showing minimally incomplete unlearning. In practice, all combinations would be included during training. These results demonstrate that the hypernetwork generalizes well beyond its training combinations, confirming that *TrustErase* extends its instant, data-free, and auditable unlearning to unseen tasks without additional training.

The passport decomposition mechanism enables authority-mediated audit of unlearning configurations without exposing private credentials. Verification consists of assessing whether the reconstructed passport obtained from the released adapter remains within an ϵ -tolerance of the original one, both structurally and functionally. Since random passport matrices are almost surely full-rank, honest reconstruction errors are dominated by floating-point precision, yielding relative errors of the order 10^{-6} .

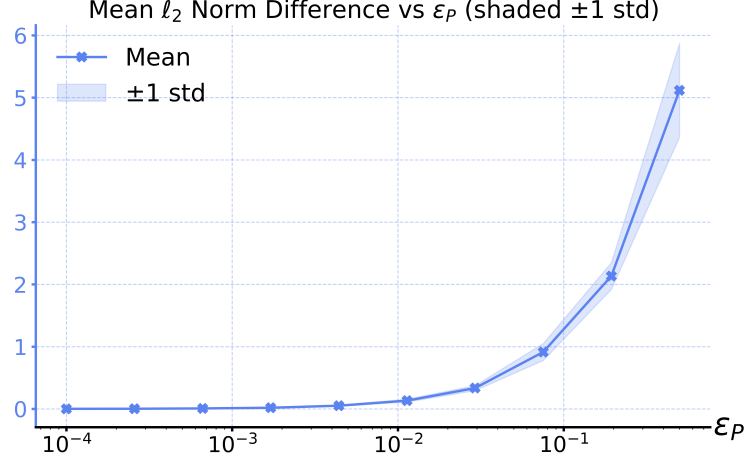


Figure 5.7: Sensitivity to perturbations. The x-axis represents the magnitude of a random perturbation ϵ_P added to the passport, and the y-axis shows the mean ℓ_2 distance between the logits produced using the original and perturbed passports.

To evaluate robustness, we perturb the passport matrix with random noise of magnitude ϵ_P and measure the resulting change in model logits (Fig. 5.7). For perturbations below approximately 10^{-2} , both the mean ℓ_2 and per-logit deviations remain negligible, showing that the functional behavior of the passport-entangled LoRA update is stable within this range. Beyond this threshold, deviations grow rapidly, indicating that even small perturbations outside the tolerance window cause significant changes in model behavior. Based on this stability elbow, we set $\epsilon_T = 0.01$, which is far above numerical noise yet strictly rejects perturbations that would collapse model behavior. These results confirm that passport verification is reliable within a well-defined ϵ_T region, robust to reconstruction noise yet sensitive enough to detect incorrect or tampered passports, thereby providing a practical and auditable basis for unlearning.

5.2.4 Discussion

TrustErase reformulates machine unlearning as a pre-committed architectural property rather than a post-hoc optimization step. Authorized forgetting configurations are encoded directly into LoRA adapters during training, enabling a *train-once, forget-anytime* paradigm where unlearning is performed instantly at deployment without access to user data. Passport decomposition makes unlearning configurations auditable through authority-mediated reconstruction, a property absent from existing approaches. We further demonstrate extensibility through a hypernetwork that composes atomic passports into credentials for previously unseen subset combinations, allowing the framework to scale to combinatorially large forget

spaces without additional training. Experiments on MNIST, CIFAR-10, and CIFAR-100 show effective forgetting with strong retention within a single unified model, positioning *TrustErase* as a practical mechanism for privacy-compliant AI that combines parameter efficiency, instant unlearning, and authority-mediated audit.

In terms of scope and limitations, *TrustErase* pre-commits to atomic forget targets at training time; this is the price of the train-once, forget-anytime regime, and it rules out scenarios where the forget set cannot be defined in advance. Two extensions to already-trained models, post-hoc passport-augmented LoRA finetuning and passport distillation from existing adapters, are sketched in Appendix B.3 but not empirically evaluated. Our evaluation targets the operational regime of ≤ 100 atomic targets; extending to ≥ 1000 targets requires addressing gradient interference between per-task losses $\{L_k\}$ in the shared low-rank subspace, for which gradient-surgery or task-grouping strategies are natural next steps (Appendix B.2). At the instance level, *TrustErase* supports bucketed forgetting via prototype-divergence partitions; per-instance passport allocation is not yet supported. From a privacy standpoint, this is the largest remaining gap. Art. 17 operates on the data subject, so the granularity that matters operationally is per-subject rather than per-class, and bucketing only approximates it. Relatedly, the datasets used here exercise the mechanism more directly than the privacy motivation: a disease class is a less plausible unit for a data-subject request than a per-subject record, although retracting a class that is no longer used does bear indirectly on storage limitation and data minimisation. CIFAR-100 exercises scalability rather than compliance. Per-subject partitions of a medical cohort, and biometric-style datasets in which the class is the individual, would test the privacy motivation more directly.

Finally, the audit protocol is authority-mediated: it assumes an honest verification authority that holds private $(\mathbf{A}, \mathbf{B})$. Cryptographic third-party verification, for example commitment plus zero-knowledge proof of the reconstruction identity, is left to future work.

These limitations bound the scope but not the contribution. Real-world GDPR compliance is fundamentally data-free and auditable: by the time a deletion request arrives, the data subject’s records may no longer be accessible. *TrustErase* (Table 5.8) is the only method in the surveyed landscape designed for this regime.

This section completes the chapter’s *Forgetting* argument that unlearning can be pre-committed as a structural property during training. Where Section 5.1 pre-commits to one prompt per class on a frozen backbone, *TrustErase* pre-commits to one passport per forget configuration on shared LoRA adapters. The two also expose a design-space trade-off: Pre-Forgettable Models keep the per-class unit small and inspectable but grow the input sequence with the number of classes, while *TrustErase* keeps inference at constant cost but

trades the per-class prompt for a per-configuration passport that is not inspectable in the same way. The contribution beyond instant unlearning is the audit layer: by binding the deployed model’s identity to a passport that an authority can reconstruct, *TrustErase* turns the developer’s claim about which unlearning configuration is in force from a behavioral assertion into a verifiable property of the deployed adapter.

5.2.5 Publications

Rutger Hendrix, Leonardo Giuseppe Russo, Matteo Pennisi, Concetto Spampinato, and Giovanni Bellitto. TrustErase: Auditable Instant Machine Unlearning with Passport-Embedded Representations. Submitted to *NeurIPS 2026*.

Chapter 6

Conclusion

This thesis treated four demands of clinical AI – interpretability, uncertainty, federated learning, unlearning – as connected rather than separate objectives, and as design choices to be built into the system from the start rather than recovered after the fact. Three chapters develop this argument, organized by the dominant role each contribution serves, with the understanding that the same mechanism recurs across chapters and several of the contributions could plausibly have sat in more than one place.

Knowing argued that pre-committing to interpretability structurally unlocks capabilities that post-hoc explanation cannot deliver. The chapter spans two scales of structural commitment. Section 3.1 contributed an uncertainty-aware framework for IPMN risk classification, combining Monte Carlo dropout with SHAP attribution to support clinician deferral at the workflow level: the explanation and the uncertainty are produced alongside the prediction in a single clinical pipeline. Sections 3.2 and 3.3 moved the structural commitment into the model itself, contributing a framework for gaze-driven prompting of a frozen segmentation model, so that where the model attends is determined by the clinician’s gaze; by the counterfactual argument of Section 2.2, this excludes the regions where shortcut learning typically arises.

Doubting then argued that uncertainty, efficiently implemented as a structural property, can serve as a learning signal in the federated setting. In FedEvPrompt, attention maps filtered by per-sample uncertainty replaced parameter aggregation as the unit of inter-client exchange, turning uncertainty from a standard deferral signal into a selection criterion on what each client contributes to a federation.

Forgetting closed the chain by arguing that pre-committing to forgetting structurally unlocks capabilities that post-hoc unlearning cannot deliver. Pre-Forgettable Models localised class-level knowledge in dedicated prompt tokens on a frozen backbone, so that unlearning reduced to removing the corresponding prompt, with the same prompt structure also

producing a calibratable inference-time uncertainty signal at no additional training cost. TrustErase bound each unlearning configuration to a non-trainable passport embedded in the model’s LoRA adapters, so that an authority can audit which configuration the deployed model implements. Across both sections of Chapter 5, the cost of unlearning is paid at training time through a structural commitment rather than at request time through retraining or distillation.

These contributions point to a stronger claim than the recurrence of mechanism alone. In each case where two demands were combined through a shared structural commitment, the combination yielded value that neither delivered separately. In Section 3.1 SHAP attribution correlated with predictive uncertainty while LIME did not, showing that the reliability information carried by an explanation method depends on its compatibility with the uncertainty estimate – an insight made possible only by considering them together. In FedEvPrompt, Section 4.1, the uncertainty-filtered attention-map exchange outperformed both parameter-sharing baselines and prior non-parameter-sharing methods on the Fed-ISIC2019 benchmark. In Section 5.1, the same prompt structure built for unlearning produced an inference-time deferral signal stronger than MC dropout. The structural and shared commitment is empirically what makes the combination measurably stronger than the sum of its components. What the thesis does not demonstrate is that a single integrated system implements all four demands together. An integrated system combining the four, evaluated with clinicians in the loop, is the next step before the structural argument can be tested where it matters. Clinical AI should *know*, *doubt*, and *forget*, like a clinician does.

List of Publications

- (1) Rutger Hendrix, Federica Proietto Salanitri, Concetto Spampinato, Simone Palazzo, and Ulas Bagci. Evidential federated learning for skin lesion image classification. In *International Conference on Pattern Recognition (ICPR)*, pages 354–365. Springer, 2024.
- (2) Viacheslav Iarkin, Evelyn E. C. de Jong, Rutger Hendrix, Frank Verhaegen, and Cecile J. A. Wolfs. Turning the attention to time-resolved EPID-images: treatment error classification with transformer multiple instance learning. *Physics in Medicine & Biology*, 69(16):165030, 2024.
- (3) Rutger Hendrix, Giovanni Patanè, Leonardo G. Russo, Simone Carnemolla, Federica Proietto Salanitri, Giovanni Bellitto, Concetto Spampinato, and Matteo Pennisi. Pre-forgettable models: Prompt learning as a native mechanism for unlearning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12446–12454, 2025. doi.org/10.1145/3746027.3758171.
- (4) Giulia Pötgens, Rutger Hendrix, Cian Scannell, Ulas Bagci, Concetto Spampinato, and Federica Proietto Salanitri. Gaze-guided medical image segmentation: A training-free approach using SAM foundation model. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–7. IEEE, 2025. doi.org/10.1109/EMBC58623.2025.11253157.
- (5) Raffaele Mineo, Giovanni Bellitto, Federica Proietto Salanitri, Rutger Hendrix, Ulas Bagci, Daniela Giordano, Concetto Spampinato, and Simone Palazzo. Conditional visuo-textual prompt learning for medical image analysis. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2025. doi.org/10.1109/ISBI60581.2025.10980870.
- (6) Debesh Jha, Gorkem Durak, Abhijit Das, Jasmer Sanjotra, Onkar Susladkar, Suramyaa Sarkar, Ashish Rauniyar, Nikhil Kumar Tomar, Linkai Peng, Sirui Li, Koushik Biswas, Ertugrul Aktas, Elif Keles, Matthew Antalek, Zheyuan Zhang, Bin Wang, Xin Zhu,

- Hongyi Pan, Deniz Seyithanoglu, Alpay Medetalibeyoglu, Vanshali Sharma, Vedat Cicek, Amir A. Rahsepar, Rutger Hendrix, A. Enis Cetin, Bulent Aydogan, Mohamed Abazeed, Frank H. Miller, Rajesh N. Keswani, Hatice Savas, Sachin Jambawalikar, Daniela P. Ladner, Amir A. Borhani, Concetto Spampinato, Michael B. Wallace, and Ulas Bagci. Ethical framework for responsible foundational models in medical imaging. *Frontiers in Medicine*, 12, 2025. doi.org/10.3389/fmed.2025.1544501.
- (7) Raffaele Mineo, Giovanni Bellitto, Federica Proietto Salanitri, Rutger Hendrix, Concetto Spampinato, and Simone Palazzo. Learning joint text and visual tokens in CLIP for medical image analysis. In *Proceedings of the 2nd International Workshop on Multimedia Computing for Health and Medicine*, pages 45–51, 2025.

Bibliography

- [1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. Association for Computational Linguistics.
- [2] Andrea Acevedo, Anna Merino, Santiago Alf  rez,   ngel Molina, Laura Bold  , and Jos   Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30:105474, 2020.
- [3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.
- [4] NIST AI. Artificial intelligence risk management framework (ai rmf 1.0). *URL: <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai>*, pages 100–1, 2023.
- [5] American Medical Association. Physician sentiments around the use of AI in health care: Augmented intelligence research. <https://www.ama-assn.org/system/files/physician-ai-sentiment-report.pdf>, 2025. Accessed: 2026-05-11.
- [6] Sebastian Bach, Alexander Binder, Gr  goire Montavon, Frederick Klauschen, Klaus-Robert M  ller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.
- [7] Thomas Baumhauer, Pascal Sch  ttle, and Matthias Zeppelzauer. Machine unlearning: Linear filtration for logit-based classifiers. *Machine Learning*, 111(9):3203–3226, 2022.
- [8] Daniel Beckmann, Jacqueline Kockwelp, Joerg Gromoll, Friedemann Kiefer, and Benjamin Risse. Sam meets gaze: Passive eye tracking for prompt-based instance segmentation. In *NeurIPS 2023 Workshop on Gaze Meets ML*, 2023.

- [9] Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.
- [10] Andrea M Bejar, Maria Jaramillo Gonzalez, Ziliang Hong, Gorkem Durak, Elif Keles, Halil Ertugrul Aktas, Zheyuan Zhang, Hongyi Pan, Zeynep Sue Jozwiak, Fergan Bol, et al. Multi-center evaluation of radiomics and deep learning to stratify malignancy risk of ipmns. *Abdominal Radiology*, pages 1–16, 2026.
- [11] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.
- [12] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680, 2023.
- [13] Risab Biswas. Polyp-sam++: Can a text guided sam perform better for polyp segmentation?, 2023.
- [14] Nicholas Bloch, Anant Madabhushi, Henkjan Huisman, John Freymann, Justin Kirby, Michael Grauer, Andinet Enquobahrie, Carl Jaffe, Larry Clarke, and Keyvan Farahani. Nci-isbi 2013 challenge: automated segmentation of prostate structures. *The Cancer Imaging Archive*, 370(6):5, 2015.
- [15] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [16] Ludovic Bourtole, Varun Chandrasekaran, Christopher A Choquette-Choo, Huan Jia, Alex Travers, Bitu Zhang, Timon Lie, Nicolas Papernot, and Christopher A Pullin. Machine unlearning. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 141–159, 2021.
- [17] Aidan Boyd, Patrick Tinsley, Kevin W Bowyer, and Adam Czajka. Cyborg: Blending human saliency into the loss improves deep learning-based synthetic face detection. In

- Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6108–6117, 2023.
- [18] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [19] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *ECCV 2022 Workshops*, pages 205–218, Cham, 2023. Springer Nature Switzerland.
- [20] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015.
- [21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.
- [22] Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 11186–11194, 2024.
- [23] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. Hardnet: A low memory traffic network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3552–3561, 2019.
- [24] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.
- [25] Cheng Chen, Juzheng Miao, Dufan Wu, Aoxiao Zhong, Zhiling Yan, Sekeun Kim, Jiang Hu, Zhengliang Liu, Lichao Sun, Xiang Li, et al. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Medical Image Analysis*, 98:103310, 2024.
- [26] Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7766–7775, 2023.

-
- [27] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3359–3367, 2023.
- [28] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7210–7217, 2023.
- [29] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172, 2018.
- [30] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pages 2089–2099. PMLR, 2021.
- [31] Juan E Corral, Sarfaraz Hussein, Pujan Kandel, Candice W Bolan, Ulas Bagci, and Michael B Wallace. Deep learning to classify intraductal papillary mucinous neoplasms using magnetic resonance imaging. *Pancreas*, 48(6):805–810, 2019.
- [32] Stefano Crippa, Francesca Aleotti, Enrico Longo, Giulio Belfiori, Stefano Partelli, Domenico Tamburrino, Francesca Di Salvo, Francesco De Cobelli, Luigi Romano, Paolo Giorgio Arcidiacono, et al. Main duct thresholds for malignancy are different in intraductal papillary mucinous neoplasms of the pancreatic head and body-tail. *Clinical Gastroenterology and Hepatology*, 20(2):390–399, 2022.
- [33] Mohamad Dbouk, Olaya I Brewer Gutierrez, Anne Marie Lennon, Miguel Chuidian, Eun Ji Shin, Ihab R Kamel, Elliot K Fishman, Jin He, Richard A Burkhart, Christopher L Wolfgang, et al. Guidelines on management of pancreatic cysts detected in high-risk individuals: an evaluation of the 2017 fukuoka guidelines and the 2020 international cancer of the pancreas screening (caps) consortium statements. *Pancreatology*, 21(3):613–621, 2021.
- [34] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold,

- Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [35] Mohamad A Eloubeidi, Frank G Gress, Thomas J Savides, Maurits J Wiersema, Michael L Kochman, Nuzhat A Ahmad, Gregory G Ginsberg, Richard A Erickson, John DeWitt, Jacques Van Dam, et al. Acute pancreatitis after eus-guided fna of solid pancreatic masses: a pooled analysis from eus centers in the united states. *Gastrointestinal endoscopy*, 60(3):385–389, 2004.
- [36] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.
- [37] European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Official Journal of the European Union, 2016. OJ L 119, 4.5.2016, pp. 1–88.
- [38] European Parliament and Council of the European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act), 2024. OJ L, 12.7.2024.
- [39] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, volume 2024, pages 53643–53673, 2024.
- [40] Lixin Fan, Kam Woh Ng, and Chee Seng Chan. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks. *Advances in neural information processing systems*, 32, 2019.
- [41] Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 12043–12051, 2024.
- [42] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

-
- [43] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.
 - [44] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for non-convex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.
 - [45] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016.
 - [46] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.
 - [47] Shashwat Goel, Ameya Prabhu, Amartya Sanyal, Ser-Nam Lim, Philip Torr, and Pon-nurangam Kumaraguru. Towards adversarial evaluations for inexact machine unlearning. *arXiv preprint arXiv:2201.06640*, 2022.
 - [48] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *European Conference on Computer Vision*, pages 383–398. Springer, 2020.
 - [49] Tamas A Gonda, Djuna L Cahen, and James J Farrell. Pancreatic cysts. *New England Journal of Medicine*, 391(9):832–843, 2024.
 - [50] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Medical Image Analysis*, 98:103324, December 2024.
 - [51] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.
 - [52] Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11516–11524, 2021.
 - [53] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *International Conference on Machine Learning*, 2020.

- [54] Zhen Guo, Zelin Wan, Qisheng Zhang, Xujiang Zhao, Feng Chen, Jin-Hee Cho, Qi Zhang, Lance M Kaplan, Dong H Jeong, and Audun Jøsang. A survey on uncertainty reasoning and quantification for decision making: Belief theory meets deep learning. *arXiv preprint arXiv:2206.05675*, 2022.
- [55] David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *International Conference on Learning Representations*, 2017.
- [56] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.
- [57] Alexander N Hanania, Leonidas E Bantis, Ziding Feng, Huamin Wang, Eric P Tamm, Matthew H Katz, Anirban Maitra, and Eugene J Koay. Quantitative imaging to evaluate malignant potential of ipmns. *Oncotarget*, 7(52):85776, 2016.
- [58] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- [59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [60] Max Heckler, Christoph W Michalski, Susanne Schaeffle, Jörg Kaiser, Markus W Büchler, and Thilo Hackert. The sendai and fukuoka consensus criteria for the management of branch duct ipmn-a meta-analysis on their accuracy. *Pancreatology*, 17(2):255–262, 2017.
- [61] Carlos Hernández-Pérez, Marc Combalia, Sebastian Podlipnik, Noel CF Codella, Veronica Rotemberg, Allan C Halpern, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Brian Helba, et al. Bcn20000: Dermoscopic lesions in the wild. *Scientific data*, 11(1):641, 2024.
- [62] David H Hoffman, Justin M Ream, Christina H Hajdu, and Andrew B Rosenkrantz. Utility of whole-lesion adc histogram metrics for assessing the malignant potential of pancreatic intraductal papillary mucinous neoplasms (ipmns). *Abdominal radiology*, 42(4):1222–1228, 2017.

-
- [63] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.
- [64] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.
- [65] Jiri Hron, Alex Matthews, and Zoubin Ghahramani. Variational bayesian dropout: pitfalls and fixes. In *International conference on machine learning*, pages 2019–2028. PMLR, 2018.
- [66] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- [67] Xinrong Hu, Xiaowei Xu, and Yiyu Shi. How to efficiently adapt large segmentation model(sam) to medical images, 2023.
- [68] Chenchuan Huang, Diane M Simeone, Lyndon Luk, Elizabeth M Hecht, Gaurav Khatri, Avinash Kambadakone, Hersh Chandarana, Justin M Ream, Jessica N Everett, Alexander Guimaraes, et al. Standardization of mri screening and reporting in individuals with elevated risk of pancreatic ductal adenocarcinoma: consensus statement of the precede consortium. *American Journal of Roentgenology*, 219(6):903–914, 2022.
- [69] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [70] Wei Huang, Yue Xu, Zhao Li, Jun Li, Qing Chen, Qiang Huang, Yaping Wu, and Hongtan Chen. Enhancing noninvasive pancreatic cystic neoplasm diagnosis with multimodal machine learning. *Scientific Reports*, 15(1):16398, 2025.
- [71] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7865–7873, 2021.
- [72] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

- [73] Fabian Isensee, Paul F Jäger, Peter M Full, Philipp Vollmuth, and Klaus H Maier-Hein. nnu-net for brain tumor segmentation. In *BrainLes 2020, MICCAI Workshop*, pages 118–132. Springer, 2021.
- [74] Debesh Jha, Gorkem Durak, Abhijit Das, Jasmer Sanjotra, Onkar Susladkar, Suramyaa Sarkar, Ashish Rauniyar, Nikhil Kumar Tomar, Linkai Peng, Sirui Li, Koushik Biswas, Ertugrul Aktas, Elif Keles, Matthew Antalek, Zheyuan Zhang, Bin Wang, Xin Zhu, Hongyi Pan, Deniz Seyithanoglu, Alpay Medetalibeyoglu, Vanshali Sharma, Vedat Cicek, Amir A. Rahsepar, Rutger Hendrix, A. Enis Cetin, Bulent Aydogan, Mohamed Abazeed, Frank H. Miller, Rajesh N. Keswani, Hatice Savas, Sachin Jambawalikar, Daniela P. Ladner, Amir A. Borhani, Concetto Spampinato, Michael B. Wallace, and Ulas Bagci. Ethical framework for responsible foundational models in medical imaging. *Frontiers in medicine*, 12:1544501, 2025.
- [75] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-seg: A segmented polyp dataset. In *MMM 2020, Daejeon, South Korea*, page 451–462. Springer-Verlag, 2020.
- [76] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.
- [77] Audun Josang, Jin-Hee Cho, and Feng Chen. Uncertainty characteristics of subjective opinions. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 1998–2005. IEEE, 2018.
- [78] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.
- [79] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):195, 2019.
- [80] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.
- [81] Eric Kerfoot, James Clough, Ilkay Oksuz, Jack Lee, Andrew P King, and Julia A Schnabel. Left-ventricle quantification using residual u-net. In *Statistical Atlases and*

- Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9*, pages 371–380. Springer, 2019.
- [82] Leila Khaertdinova, Tatyana Shmykova, Ilya Pershin, Andrey Laryukov, Albert Khanov, Damir Zidikhanov, and Bulat Ibragimov. Gaze assistance for efficient segmentation correction of medical images. *IEEE Access*, 13:14199–14213, 2025.
 - [83] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.
 - [84] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.
 - [85] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.
 - [86] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
 - [87] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987, 2023.
 - [88] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.
 - [89] Rodney LaLonde, Irene Tanner, Katerina Nikiforaki, Georgios Z Papadakis, Pujan Kandel, Candice W Bolan, Michael B Wallace, and Ulas Bagci. Inn: inflated neural networks for ipmn diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–109. Springer, 2019.
 - [90] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and

- Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [91] Daliang Li and Junpu Wang. FedMD: Heterogenous federated learning via model distillation. In *NeurIPS Workshop on Federated Learning*, 2019.
- [92] Guanghai Li, Wansen Wu, Yan Sun, Li Shen, Baoyuan Wu, and Dacheng Tao. Visual prompt based personalized federated learning. *arXiv preprint arXiv:2303.08678*, 2023.
- [93] Huiqian Li, Dingwen Zhang, Jieru Yao, Longfei Han, Zhongyu Li, and Junwei Han. Asps: Augmented segment anything model for polyp segmentation. In *MICCAI*, pages 118–128. Springer, 2024.
- [94] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368. PMLR, 2021.
- [95] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.
- [96] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for 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)*, pages 4582–4597, 2021.
- [97] Yuheng Li, Mingzhe Hu, and Xiaofeng Yang. Polyp-sam: Transfer sam for polyp segmentation. In *Medical imaging 2024: computer-aided diagnosis*, volume 12927, pages 749–754. SPIE, 2024.
- [98] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in neural information processing systems*, 33:2351–2363, 2020.
- [99] Xian Lin, Yangyang Xiang, Li Zhang, Xin Yang, Zengqiang Yan, and Li Yu. Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation, 2023.
- [100] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang

- Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.
- [101] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890, 2021.
- [102] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [103] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
- [104] Chong Ma, Lin Zhao, Yuzhong Chen, Sheng Wang, Lei Guo, Tuo Zhang, Dinggang Shen, Xi Jiang, and Tianming Liu. Eye-gaze-guided vision transformer for rectifying shortcut learning. *IEEE Transactions on Medical Imaging*, 42(11):3384–3394, 2023.
- [105] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1), 1 2024.
- [106] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. 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)*, pages 565–576, 2021.
- [107] Ananth Mahadevan and Michael Mathioudakis. Certifiable machine unlearning for linear models. *arXiv preprint arXiv:2106.15093*, 2021.
- [108] Ananth Mahadevan and Michael Mathioudakis. Certifiable unlearning pipelines for logistic regression: An experimental study. *Machine Learning and Knowledge Extraction*, 4(3):591–620, 2022.
- [109] Mobina Mansoori, Sajjad Shahabodini, Jamshid Abouei, Konstantinos N. Plataniotis, and Arash Mohammadi. Polyp SAM 2: Advancing Zero shot Polyp Segmentation in Colorectal Cancer Detection, 8 2024.

- [110] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023.
- [111] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.
- [112] Brendan McMahan et al. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- [113] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706. IEEE, 2019.
- [114] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (SP)*, pages 111–125. IEEE, 2008.
- [115] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.
- [116] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46, 2025.
- [117] Giyeong Oh, Saejin Kim, Woohyun Cho, Sangkyu Lee, Jiwan Chung, Dokyung Song, and Youngjae Yu. Seal: Entangled white-box watermarks on low-rank adaptation. *arXiv preprint arXiv:2501.09284*, 2025.
- [118] Takao Ohtsuka, Carlos Fernandez-del Castillo, Toru Furukawa, Susumu Hijioka, Jin-Young Jang, Anne Marie Lennon, Yoshihiro Miyasaka, Eizaburo Ohno, Roberto Salvia, Christopher L Wolfgang, et al. International evidence-based kyoto guidelines for the management of intraductal papillary mucinous neoplasm of the pancreas. *Pancreatol-ogy*, 24(2):255–270, 2024.

- [119] Ian Osband. Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. In *NIPS workshop on bayesian deep learning*, volume 192. MIT Press, 2016.
- [120] Hongyi Pan, Gorkem Durak, Elif Keles, Deniz Seyithanoglu, Zheyuan Zhang, Alpay Medetalibeyoglu, Halil Ertugrul Aktas, Andrea Mia Bejar, Ziliang Hong, Yavuz Tak-tak, Ulas Bagci, et al. Cyst-X: A federated ai system outperforms clinical guidelines to detect pancreatic cancer precursors and reduce unnecessary surgery. *arXiv preprint arXiv:2507.22017*, 2025.
- [121] A Panda et al. Partially-blinded unlearning: A bayesian approach. *arXiv preprint arXiv:2401.00001*, 2024.
- [122] Sang Won Park, Na Young Yeo, Tae-Hoon Kim, Myoung Nam Lim, Inhyeok Yim, Oh Beom Kwon, Seung-Joo Nam, Hui-Young Lee, and Woo Jin Kim. Explainable ai for colorectal cancer mortality and risk factor prediction in korea: A nationwide cancer cohort study. *International Journal of Medical Informatics*, page 106125, 2025.
- [123] Leping Peng, Jingjing Niu, Gang Huang, Fan Zhang, Fang Ma, Xiuling Zhang, Yu Wang, Kai Ai, Xiaoyue Zhang, Yuqi He, et al. Noninvasive prediction of lymphovascular invasion in rectal cancer without lymph node metastasis using a shap-interpretable combined model integrating mri radiomics features and clinical immune-inflammatory biomarkers: a bicenter study. *Cancer Imaging*, 25(1):123, 2025.
- [124] Matteo Pennisi, Federica Proietto Salanitri, Giovanni Bellitto, Bruno Casella, Marco Aldinucci, Simone Palazzo, and Concetto Spampinato. Feder: Federated learning through experience replay and privacy-preserving data synthesis. *Computer Vision and Image Understanding*, 238:103882, 2024.
- [125] Aleksandar Petrov, Philip Torr, and Adel Bibi. When do prompting and prefix-tuning work? a theory of capabilities and limitations. In *International Conference on Learning Representations*, volume 2024, pages 6031–6054, 2024.
- [126] Marcin Polkowski, A Larghi, Birgit Weynand, C Boustière, M Giovannini, B Pujol, and J-M Dumonceau. Learning, techniques, and complications of endoscopic ultrasound (eus)-guided sampling in gastroenterology: European society of gastrointestinal endoscopy (esge) technical guideline. *Endoscopy*, 44(02):190–206, 2012.
- [127] JoramSoch(2021). Proof:. Kullback-leibler divergence for the dirichlet distribution. <https://statproofbook.github.io/P/dir-kl.html> ; DOI:10.5281/zenodo.4305949.

- [128] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [129] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.
- [130] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021.
- [131] Amin Ranem, Mohamed Afham Mohamed Aflal, Moritz Fuchs, and Anirban Mukhopadhyay. UnCLe SAM: Unleashing SAM’s potential for continual prostate MRI segmentation. In *MIDL*, 2024.
- [132] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- [133] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.
- [134] Stephanie Romutis and Randall Brand. Burden of new pancreatic cyst diagnosis. *Gastrointestinal Endoscopy Clinics of North America*, 33(3):487–495, 2023.
- [135] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

- [136] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.
- [137] Nisha I Sainani, Anuradha Saokar, Vikram Deshpande, Carlos Fernandez-del Castillo, Peter Hahn, and Dushyant V Sahani. Comparative performance of mdct and mri with mr cholangiopancreatography in characterizing small pancreatic cysts. *American Journal of Roentgenology*, 193(3):722–731, 2009.
- [138] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014.
- [139] F Proietto Salanitri. Deep recurrent-convolutional model for automated segmentation of craniomaxillofacial ct scans. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9062–9067. IEEE, 2021.
- [140] F Proietto Salanitri, Giovanni Bellitto, Simone Palazzo, Ismail Irmakci, M Wallace, C Bolan, Megan Engels, Sanne Hoogenboom, Marco Aldinucci, Ulas Bagci, et al. Neural transformers for intraductal papillary mucosal neoplasms (ipmn) classification in mri images. In *2022 44th annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 475–479. IEEE, 2022.
- [141] Federica Proietto Salanitri, Giovanni Bellitto, Salvatore Calcagno, Ulas Bagci, Concetto Spampinato, and Manuela Pennisi. Sam-guided prompt learning for multiple sclerosis lesion segmentation. *Pattern Recognition Letters*, 2025.
- [142] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [143] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.
- [144] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

- [145] Hyowoon Seo, Jihong Park, Seungeun Oh, Mehdi Bennis, and Seong-Lyun Kim. Federated knowledge distillation. *Machine Learning and Wireless Communications*, page 457, 2022.
- [146] Tal Shaharabany, Aviad Dahan, Raja Giryes, and Lior Wolf. Autosam: Adapting sam to medical images by overloading the prompt encoder. In *34th British Machine Vision Conference (BMVC)*, Aberdeen, UK, 2023.
- [147] Jeremy Sharib, Laura Esserman, Eugene J Koay, Anirban Maitra, Yu Shen, Kimberly S Kirkwood, and Elissa M Ozanne. Cost-effectiveness of consensus guideline based management of pancreatic cysts: The sensitivity and specificity required for guidelines to be cost-effective. *Surgery*, 168(4):601–609, 2020.
- [148] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.
- [149] Niclas Ståhl, Göran Falkman, Alexander Karlsson, and Gunnar Mathiason. Evaluation of uncertainty quantification in deep learning. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 556–568. Springer, 2020.
- [150] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.
- [151] Matteo Tacelli, Ciro Celsa, Bianca Magro, Marco Barchiesi, Luca Barresi, Gabriele Capurso, Paolo Giorgio Arcidiacono, Calogero Camma, and Stefano Francesco Crino. Diagnostic performance of endoscopic ultrasound through-the-needle microforceps biopsy of pancreatic cystic lesions: systematic review with meta-analysis. *Digestive Endoscopy*, 32(7):1018–1030, 2020.
- [152] Sahar Takshi. Unexpected inequality: disparate-impact from artificial intelligence in healthcare decisions. *JL & Health*, 34:215, 2020.
- [153] Domenico Tamburrino, Nicolò De Pretis, Enrique Pérez-Cuadrado-Robles, Laura Uribarri-Gonzalez, Zeeshan Ateeab, Giulio Belfiori, Patrick Maisonneuve, Gabriele Capurso, Giuseppe Vanella, Maria Chiara Petrone, et al. Identification of patients with branch-duct intraductal papillary mucinous neoplasm and very low risk of cancer: multicentre study. *British Journal of Surgery*, 109(7):617–622, 2022.

-
- [154] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8432–8440, 2022.
- [155] Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE transactions on neural networks and learning systems*, 35(9):13046–13055, 2023.
- [156] Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319. IEEE, 2022.
- [157] David Tobaly, Joao Santinha, Riccardo Sartoris, Marco Dioguardi Burgio, Celso Matos, Jérôme Cros, Anne Couvelard, Vinciane Rebours, Alain Sauvanet, Maxime Ronot, et al. Ct-based radiomics analysis to predict malignancy in patients with intraductal papillary mucinous neoplasm (ipmn) of the pancreas. *Cancers*, 12(11):3089, 2020.
- [158] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6021–6029, 2020.
- [159] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.
- [160] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, page 180161, 2018.
- [161] U.S. Food and Drug Administration. Artificial intelligence-enabled medical devices. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices>, 2025. Accessed: 2026-05-11.
- [162] Leonard Elia Van Dyck, Sebastian Jochen Denzler, and Walter Roland Gruber. Guiding visual attention in deep convolutional neural networks based on human eye movements. *Frontiers in Neuroscience*, 16:975639, 2022.
- [163] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

- [164] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017(1):4037190, 2017.
- [165] Paul Voigt and Axel Von dem Bussche. *The EU General Data Protection Regulation (GDPR)*. Springer, 2017.
- [166] Bin Wang, Armstrong Aboah, Zheyuan Zhang, Hongyi Pan, and Ulas Bagci. Gazesam: Interactive image segmentation with eye gaze and segment anything model. In *Gaze Meets Machine Learning Workshop*, pages 254–265. PMLR, 2024.
- [167] Jinfeng Wang, Sifan Song, Xinkun Wang, Yiyi Wang, Yiyi Miao, Jionglong Su, and S Kevin Zhou. Promise: Promptable medical image segmentation using sam. *arXiv preprint arXiv:2403.04164*, 2024.
- [168] Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review*, 58(8):227, 2025.
- [169] Meng Wang, Lianyu Wang, Xinxing Xu, Ke Zou, Yiming Qian, Rick Siow Mong Goh, Yong Liu, and Huazhu Fu. Federated uncertainty-aware aggregation for fundus diabetic retinopathy staging. *arXiv preprint arXiv:2303.13033*, 2023.
- [170] Haider J Warraich, Troy Tazbaz, and Robert M Califf. Fda perspective on the regulation of artificial intelligence in health care and biomedicine. *Jama*, 333(3):241–247, 2025.
- [171] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.
- [172] Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Lmeraser: Large model unlearning through adaptive prompt tuning. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Mai Khao, Thailand, May 3–5 2025. To appear.
- [173] Ming-ming Xu, Shi Yin, Ali A Siddiqui, Ronald R Salem, Beth Schrope, Amrita Sethi, John M Poneros, Frank G Gress, Jeanine M Genkinger, Catherine Do, et al. Com-

- parison of the diagnostic accuracy of three current guidelines for the evaluation of asymptomatic pancreatic cystic neoplasms. *Medicine*, 96(35):e7900, 2017.
- [174] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems*, 36:7093–7115, 2023.
 - [175] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023.
 - [176] Lanhong Yao, Zheyuan Zhang, Ugur Demir, Elif Keles, Camila Vendrami, Emil Agarunov, Candice Bolan, Ivo Schoots, Marc Bruno, Rajesh Keswani, et al. Radiomics boosts deep learning model for ipmn classification. In *International Workshop on Machine Learning in Medical Imaging*, pages 134–143. Springer, 2023.
 - [177] Lanhong Yao, Zheyuan Zhang, Elif Keles, Cemal Yazici, Temel Tirkes, and Ulas Bagci. A review of deep learning and radiomics approaches for pancreatic cancer diagnosis from medical imaging. *Current opinion in gastroenterology*, 39(5):436–447, 2023.
 - [178] Meng Ye, Bingyu Xin, Leon Axel, and Dimitris Metaxas. Continuous spatio-temporal memory networks for 4d cardiac cine mri segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9532–9542. IEEE, 2025.
 - [179] Songfeng Yu, Naoki Takasu, Toshihiro Watanabe, Tsuyoshi Fukumoto, Shinji Okazaki, Koji Tezuka, Shuichiro Sugawara, Ichiro Hirai, and Wataru Kimura. Validation of the 2012 fukuoka consensus guideline for intraductal papillary mucinous neoplasm of the pancreas from a single institution experience. *Pancreas*, 46(7):936–942, 2017.
 - [180] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020.
 - [181] Wenxi Yue, Jing Zhang, Kun Hu, Yong Xia, Jiebo Luo, and Zhiyong Wang. Surgical-sam: Efficient class promptable surgical instrument segmentation. *AAAI*, 38(7):6890–6898, Mar. 2024.
 - [182] Paul A Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C Gee, and Guido Gerig. User-guided 3d active contour segmentation of

- anatomical structures: significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006.
- [183] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, 2022.
- [184] Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. FedALA: Adaptive local aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.
- [185] Jingwei Zhang, Ke Ma, Saarthak Kapse, Joel Saltz, Maria Vakalopoulou, Prateek Prasanna, and Dimitris Samaras. Sam-path: A segment anything model for semantic segmentation in digital pathology. In *MICCAI*, pages 161–170. Springer, 2023.
- [186] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.
- [187] Xinkai Zhao, Zhenhua Wu, Shuangyi Tan, De-Jun Fan, Zhen Li, Xiang Wan, and Guanbin Li. Semi-supervised spatial temporal attention network for video polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 456–466. Springer, 2022.
- [188] Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, and Brian Lim. Exploiting explanations for model inversion attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 682–692, 2021.
- [189] Yuan Zhong, Chenhui Tang, Yumeng Yang, Ruoxi Qi, Kang Zhou, Yuqi Gong, Pheng Ann Heng, Janet H Hsiao, and Qi Dou. Weakly-supervised medical image segmentation with gaze annotations. In *MICCAI*, pages 530–540. Springer, 2024.
- [190] K. Zhou, J. Yang, C. C. Loy, et al. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337–2348, 2022.
- [191] Yu Zhou, Dian Zheng, Qijie Mo, Renjie Lu, Kun-Yu Lin, and Wei-Shi Zheng. Decoupled distillation to erase: A general unlearning method for any class-centric tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20350–20359, 2025.

- [192] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *DLMIA 2018 and ML-CDS 2018, MICCAI Workshop*, pages 3–11. Springer, 2018.
- [193] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019.
- [194] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

Appendices

Appendix A

PreForgettable Models

A.1 Inference-Time Uncertainty Estimation: Additional Results

This appendix section reports the per-dataset figures omitted from Section 5.1. Figures A.3–A.4 reproduce the DermaMNIST inference-time uncertainty analysis (entropy distributions and accuracy–coverage curves) on BloodMNIST and OrganSMNIST. The pattern matches: prompt-subset places incorrect predictions at higher H than MC dropout and yields a higher selective accuracy at every coverage level, with consistently higher AURC

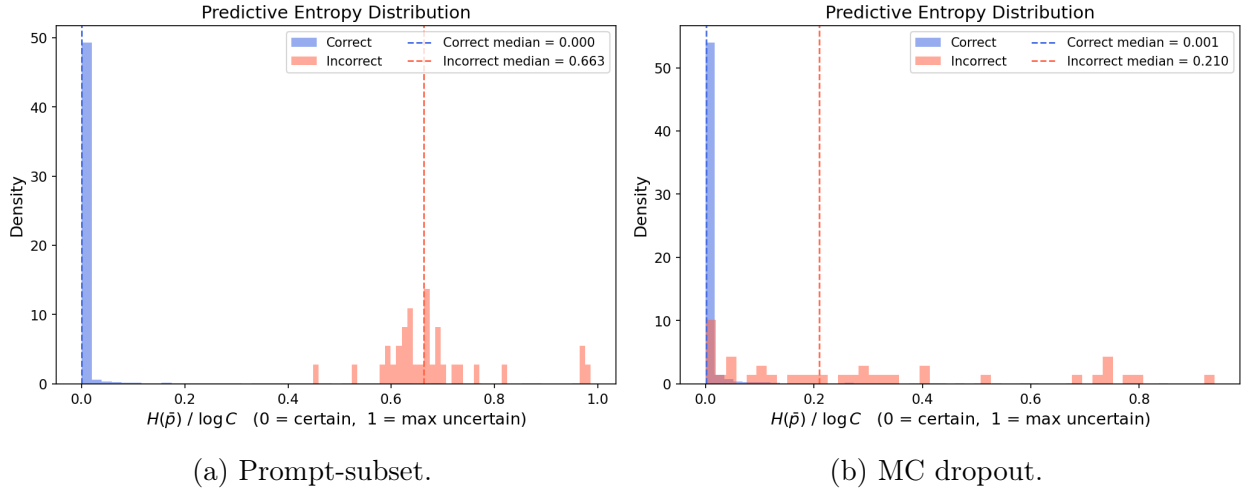


Figure A.1: Distribution of H on BloodMNIST test set, separated by prediction correctness.

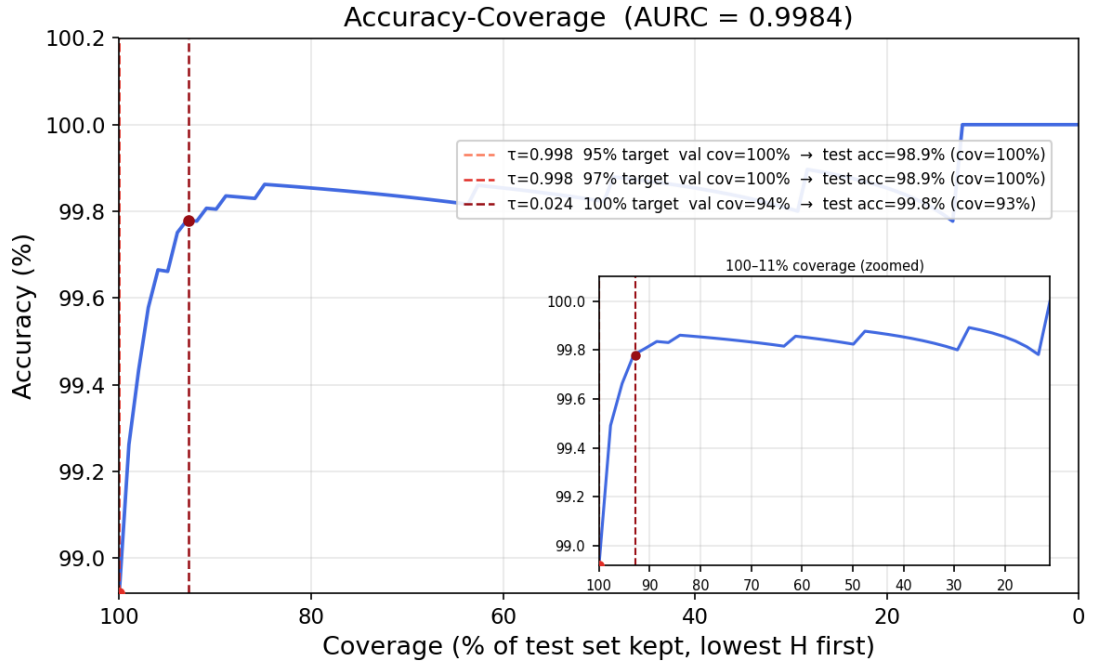
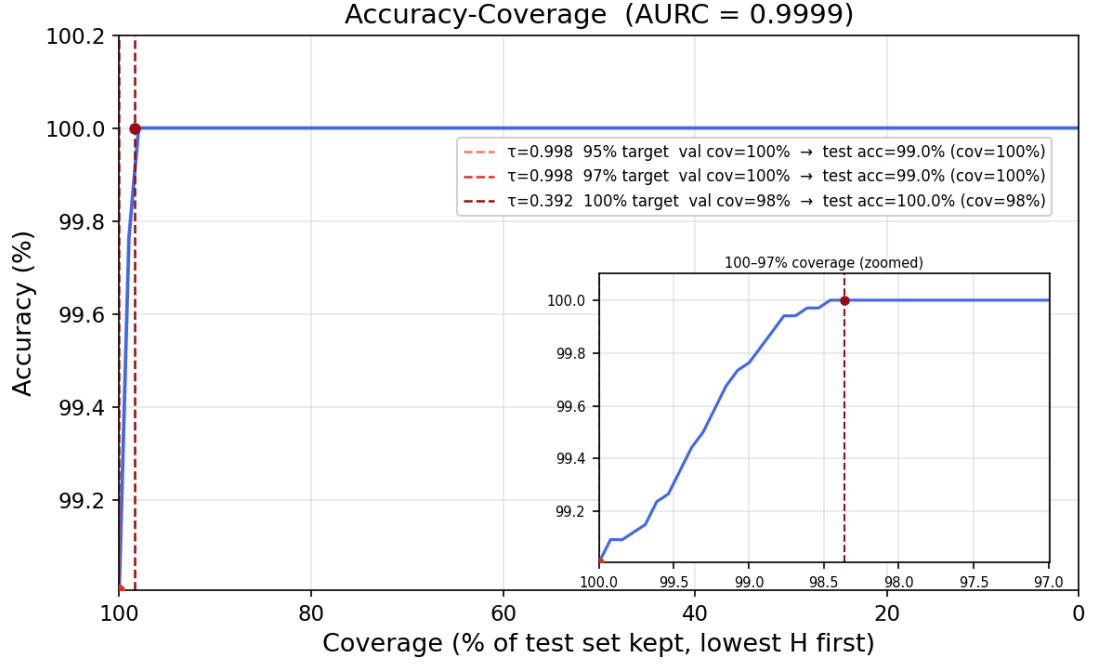


Figure A.2: Accuracy versus coverage on BloodMNIST, with val-calibrated thresholds applied to test.

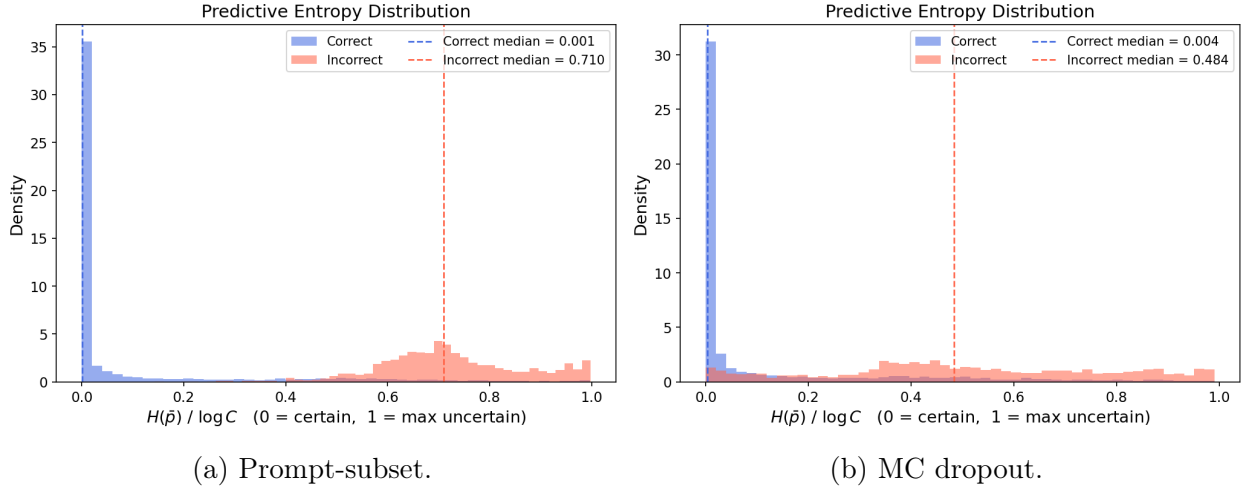
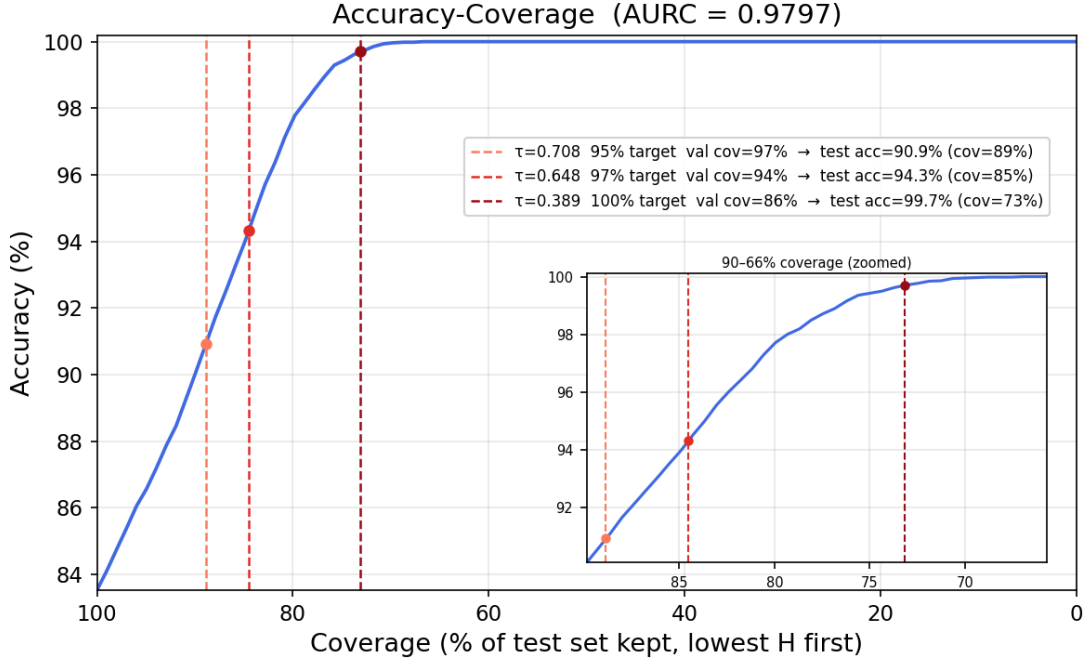
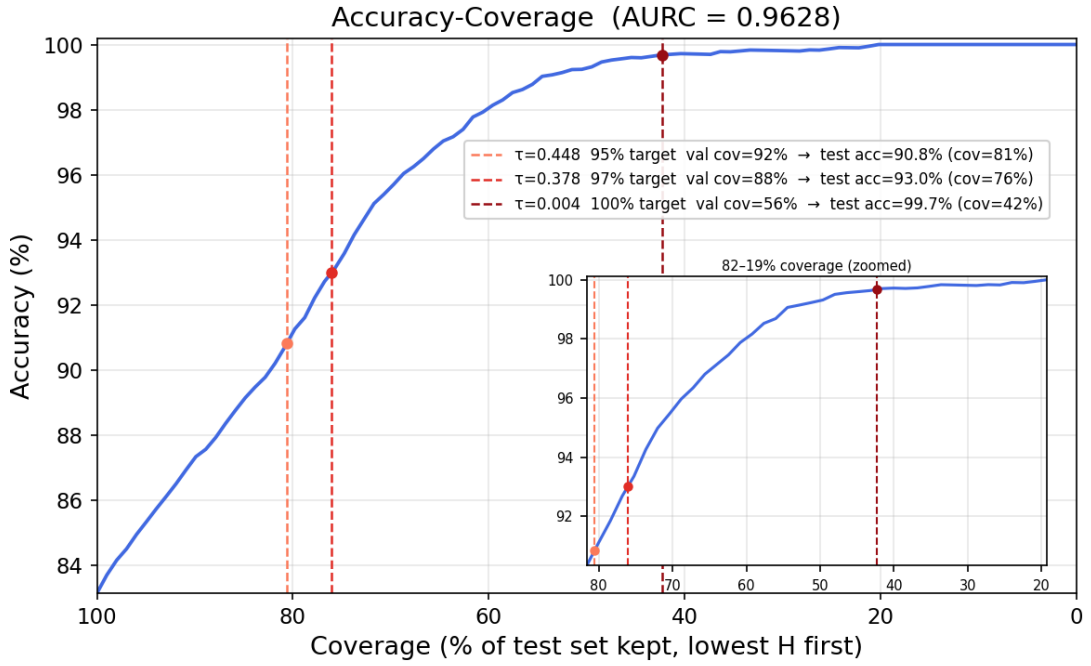


Figure A.3: Distribution of H on OrganSMNIST test set, separated by prediction correctness.



(a) Prompt-subset.



(b) MC dropout.

Figure A.4: Accuracy versus coverage on OrganSMNIST, with val-calibrated thresholds applied to test.

A.2 Sensitivity to the LoRA dropout rate

The main-text comparison uses LoRA dropout $p = 0.1$ to match the design choice for LoRA in the unlearning components of Section 5.1. However, MC dropout’s uncertainty signal comes from variation induced by the dropout layers at inference, so a low rate produces less variation and could in principle understate it. To evaluate this, we re-ran both methods on all three datasets with dropout raised to $p = 0.3$, keeping every other hyperparameter unchanged. Table A.1 reports the resulting AURC.

Table A.1: AURC of prompt-subset and MC dropout at two LoRA dropout rates. Higher is better.

Dataset	Method	$p = 0.1$	$p = 0.3$
BloodMNIST	Prompt-subset	0.9999	0.9999
BloodMNIST	MC dropout	0.9984	0.9984
DermaMNIST	Prompt-subset	0.9852	0.9835
DermaMNIST	MC dropout	0.9689	0.9729
OrganSMNIST	Prompt-subset	0.9797	0.9792
OrganSMNIST	MC dropout	0.9628	0.9653

Both methods move only slightly under the change of rate. MC dropout improves by at most 0.004 AURC, consistent with its signal benefiting from a stronger stochastic perturbation. Prompt-subset drops by at most 0.002 AURC. This can be explained by the higher training-time variation, which regularizes the routings and reduces how much they diverge across prompt subsets. Prompt-subset still outperforms MC dropout at $p = 0.3$. The comparison in Section 5.1 is therefore not an artifact of the dropout rate.

Appendix B

TrustErase

B.1 Experimental Setup

B.1.1 Training details and configurations

Architectures. Unless otherwise specified, experiments are conducted using Vision Transformers (ViT), specifically the ViT-Tiny, ViT-Small, and ViT-Base variants. All models share a depth of 12 layers and a patch size of 16×16 . The architectures differ in their embedding dimensions (d) and number of attention heads (h): ViT-Tiny ($d = 192, h = 3$), ViT-Small ($d = 384, h = 6$), and ViT-Base ($d = 768, h = 12$). While the proposed framework is architecture-agnostic and can be applied to other backbones (e.g., CNNs), we report ViT results in the main paper for consistency with prior work; additional experiments on alternative architectures are provided in the Appendix B.4.

Training During training, the low-rank matrices $\mathbf{A}$ and $\mathbf{B}$ are shared across all tasks, while task-specific passports $\mathbf{C}_k$ remain fixed. A hypernetwork variant of *TrustErase* is trained once after all atomic passports are obtained, enabling the synthesis of composite passports for arbitrary class combinations without retraining. Both stages use AdamW with weight decay 1×10^{-4} and cosine annealing learning rate scheduler, with a learning rate of 1×10^{-3} for base training and 1×10^{-4} for the hypernetwork, each for 10 epochs. Training uses an effective batch size of 96 via gradient accumulation. LoRA modules of rank 8, 32, and 64 are applied to the query and value projection matrices. For competing methods requiring both learning and unlearning phases, 10 unlearning epochs are applied to ensure fair comparison.

B.1.2 Hyperparameter Search for Competitors

To ensure a fair evaluation, all competing unlearning methods underwent a hyperparameter search. For all methods, we searched learning rates in the range $[10^{-6}, 10^{-2}]$, considering only multiples of 10, and used the configuration that achieved the best performance on CIFAR-10. For SSD and UNSIR, we searched α in the range $[10^{-1}, 10^2]$ and λ in the range $[10^{-1}, 5]$. All competitor methods were trained using AdamW with weight decay 1×10^{-4} and cosine annealing for 10 epochs, matching our setup.

B.1.3 Hypernetwork Architecture

The hypernetwork H_ϕ is implemented as a task-conditioned Transformer encoder that aggregates multiple atomic passports into a single, unified representation. Given a set of N atomic passports, the module synthesizes a multi-task passport $\mathcal{P}_{task}$ by integrating components selected via a task-specific mask.

Each vectorized atomic passport $P_i \in \mathbb{R}^{r^2}$ is projected into a shared latent space using a linear layer $W_e \in \mathbb{R}^{r^2 \times 768}$. A learnable $[CLS]$ token is prepended to serve as a global aggregator, and positional embeddings are added to preserve the identity of the component. The resulting input sequence Z_0 is

$$Z_0 = [z_{CLS}; (W_e P_1 + P E_1); (W_e P_2 + P E_2); \dots; (W_e P_N + P E_N)].$$

The fusion logic is controlled by a binary task descriptor $m \in \{0, 1\}^N$. We utilize this descriptor to generate a padding mask M applied within the Multi-Head Self-Attention (MSA) layers. For an attention head with query Q , key K , and value V , the attention weights are computed as

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

Here $M_{ij} = -\infty$ for all indices j where $m_j = 0$. This ensures that the $[CLS]$ token only integrates information from the passports designated by the mask, preventing interference from inactive components.

The backbone consists of 4 encoder layers with 8 attention heads and a feed-forward expansion ratio of 4. We utilize the final hidden state of the $[CLS]$ token, $h_{CLS} \in \mathbb{R}^{768}$, as a task-conditioned bottleneck. This state is mapped back to the parameter space through a

linear projection followed by a reshape operation

$$\mathcal{P}_{task} = \text{Reshape}(W_p h_{CLS} + b_p) \in \mathbb{R}^{r \times r},$$

recovering the passport matrix and distilling inter-task dependencies into a compact latent vector, enabling the synthesis of a unified parameterization that scales efficiently with the number of atomic components.

B.2 Parameter Dimensionality and Scaling Behavior

B.2.1 Lemma

Class-independent parameter dimensionality. Let $\mathbf{W}_0$ be a pretrained backbone and let $\mathbf{A} \in \mathbb{R}^{r \times k}$ and $\mathbf{B} \in \mathbb{R}^{d \times r}$ denote the LoRA parameters with fixed rank r . Consider a classification problem with C classes, where each class (or unlearning task) is associated with a fixed passport matrix $\mathbf{C}_k \in \mathbb{R}^{r \times r}$. The model prediction and training objective are

$$f(x; \mathbf{A}, \mathbf{B}) = (\mathbf{W}_0 + \mathbf{B}\mathbf{C}_y\mathbf{A})x, \quad (\text{B.1})$$

$$L(\mathbf{A}, \mathbf{B}) = \mathbb{E}_{(x,y)} [\ell(f(x; \mathbf{A}, \mathbf{B}), y)], \quad (\text{B.2})$$

where ℓ is a smooth and lower-bounded loss function. Then stochastic gradient descent (SGD) converges to an ε -stationary point at rate $\mathcal{O}(1/\sqrt{T})$, with constants independent of the number of classes C .

Proof. The only trainable parameters are $(\mathbf{A}, \mathbf{B})$, whose dimension is $\mathcal{O}(r(d+k))$ and therefore does not depend on C . Since the passport matrices $\mathbf{C}_k$ are fixed and the loss ℓ is smooth, the objective L has bounded-variance stochastic gradients with constants independent of C . Standard results for non-convex stochastic optimization [44] therefore yield

$$\mathbb{E} [\|\nabla L(\mathbf{A}_T, \mathbf{B}_T)\|^2] \leq \mathcal{O}(1/\sqrt{T}), \quad (\text{B.3})$$

which establishes convergence independently of the number of classes. $\square$

B.2.2 Implications and Scaling Limits

The lemma establishes that the parameter dimensionality of the optimization problem does not grow with the number of unlearning tasks K . Per-step compute, however, scales linearly

as $\mathcal{O}(K)$ via the per-task loss aggregation (Equation 5.20). The lemma’s smoothness assumption, combined with the gauge freedom of the passport-LoRA parameterization, means that for moderate K the shared low-rank subspace has sufficient capacity to express K near-orthogonal unlearning behaviors without optimization difficulty.

This regime is well-behaved within our scope. At $K \geq 1000$, the per-task gradients $\{\nabla_{(\mathbf{A}, \mathbf{B})} L_k\}$ increasingly compete for the same low-rank capacity, and the multi-task objective develops destructive interference. Concretely, when the rank- r subspace cannot simultaneously accommodate K task gradients, the pairwise gradients $\nabla_{\mathbf{A}} L_k$ and $\nabla_{\mathbf{A}} L_j$ for $k \neq j$ become non-orthogonal, and a fraction of pairs acquire negative cosine similarity. Under the aggregated update

$$\nabla_{\mathbf{A}, \mathbf{B}} L_{\text{total}} = \frac{1}{K} \sum_{k=1}^K \nabla_{\mathbf{A}, \mathbf{B}} L_k, \quad (\text{B.4})$$

the conflicting components cancel while the redundant components dilute the per-task signal, so SGD progresses suboptimally on individual L_k even when L_{total} continues to decrease. This is the gradient-conflict phenomenon studied in the multi-task optimization literature [143, 180, 101], and a related form of interference has been documented post-hoc when combining independently trained task-specific parameter updates [174].

Two mitigation strategies are natural extensions in this regime: gradient surgery, which projects per-task gradients onto mutually orthogonal subspaces before aggregation [180], and task grouping, which clusters passports by gradient similarity and trains each cluster against an independent $(\mathbf{A}, \mathbf{B})$ pair. Both restore the bounded-gradient-variance condition that underlies the lemma’s convergence guarantee: gradient surgery by removing destructive cross-task components, task grouping by reducing the effective K per parameter pair so that the rank- r subspace regains sufficient capacity. We leave a full empirical study to future work.

B.3 Extensions to Already-Trained Models

TrustErase as presented requires the passport library and the shared LoRA factors $(\mathbf{A}, \mathbf{B})$ to be jointly trained. For deployments where the target model already exists, two extension routes are possible depending on where the target knowledge resides.

Backbone knowledge: passport-augmented LoRA finetuning. If the unlearning target is encoded in the backbone itself, a passport-augmented LoRA module can be naturally attached on top of the frozen backbone and finetuned with the multi-task objective (Equation 5.20), with the target classes included in the per-task losses $\{L_k\}$.

LoRA-encoded knowledge: passport distillation. If the target knowledge already lives in an existing LoRA adapter ($\mathbf{A}_{\text{src}}, \mathbf{B}_{\text{src}}$), that adapter can be distilled into a passport-augmented LoRA via supervised passport distillation. Concretely, a fresh passport library $\{\mathbf{C}_k\}$ and shared factors ($\mathbf{A}, \mathbf{B}$) are trained to match the source adapter’s outputs on the retain set, while applying the masked forgetting loss (Equation 5.17) on the forget set.

Both routes inherit the pre-commitment requirement: the atomic forget targets must be specified at finetune or distillation time. Empirical evaluation of these extensions is left to future work.

B.4 Class Unlearning

Cifar20 evaluation. Here we extend our evaluation to CIFAR-20. Under the setup of competing methods, Table B.1 demonstrates that our approach achieves perfect forgetting in single-class scenarios while preserving strong retain performance and privacy guarantees close to retraining baselines. Table B.2 further shows that, under a uniform-distribution forgetting objective, we achieve on-par forget accuracy compared to Pre-Forgettable methods while exhibiting substantially smaller worst-case forgetting gaps, indicating more consistent and reliable forgetting behavior.

Class	metric	Original Model	Retrain	Finetune	Bad Teacher	UNSIR	Amnesiac	SSD	Ours
Veh2	$\mathcal{D}_r$	95.73 $\pm$ 0.00	94.85 $\pm$ 0.13	87.75 $\pm$ 1.64	93.59 $\pm$ 0.3	93.56 $\pm$ 0.32	93.88 $\pm$ 0.15	93.12 $\pm$ 0.00	94.99 $\pm$ 0.08
	$\mathcal{D}_f$	95.22 $\pm$ 0.00	0.00 $\pm$ 0.00	0.04 $\pm$ 0.12	4.88 $\pm$ 4.12	70.31 $\pm$ 5.03	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
	MIA	84.04 $\pm$ 0.00	22.96 $\pm$ 0.03	38.15 $\pm$ 0.08	0.02 $\pm$ 0.00	48.98 $\pm$ 0.07	1.20 $\pm$ 0.00	7.04 $\pm$ 0.00	23.74 $\pm$ 11.36
veg	$\mathcal{D}_r$	95.59 $\pm$ 0.00	94.54 $\pm$ 0.21	87.09 $\pm$ 1.24	92.92 $\pm$ 0.51	93.25 $\pm$ 0.35	93.29 $\pm$ 0.41	95.71 $\pm$ 0.00	94.78 $\pm$ 0.14
	$\mathcal{D}_f$	97.57 $\pm$ 0.00	0.00 $\pm$ 0.00	0.30 $\pm$ 0.29	8.28 $\pm$ 6.79	89.02 $\pm$ 2.41	0.02 $\pm$ 0.07	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
	MIA	91.32 $\pm$ 0.00	4.41 $\pm$ 0.01	14.72 $\pm$ 0.05	0.02 $\pm$ 0.00	58.67 $\pm$ 0.04	1.02 $\pm$ 0.00	1.88 $\pm$ 0.00	3.08 $\pm$ 1.37

Table B.1: **Class unlearning results on CIFAR-20 using ViT-Base Rank 32**, evaluating simultaneous forgetting of two classes from a single trained model. Results for baseline methods are reported from SSD [41].

Convolutional Neural Network evaluation. Although our method builds on LoRA, originally introduced for Transformers, it is not inherently Transformer-specific. To demonstrate backbone agnosticism, we evaluate TrustErase on an ImageNet-pretrained VGG-16 by inserting rank-128 passport C-LoRA into the last two classifier layers. As shown in Table B.3, TrustErase reaches complete forgetting ($\text{Acc}_{ft} = 0.0 \pm 0.0$) with $86.2 \pm 0.2\%$ retained accuracy, within 0.7 points of its retrain reference, and an MIA rate of 0.253 against the reference’s 0.246, yielding the smallest distance to the oracle of all methods (dist = 0.09, against 0.11 for Delete and 0.15 for Finetune). The over-forgetting pattern of Table 5.9

Table B.2: **Comparison with Pre-Forgettable Models on CIFAR-20 under the uniform-distribution forgetting objective.** All methods employ LoRA (ranks 32/64) with ViT-Small and ViT-Base backbones. Atomic passports are trained once, with one passport per class, and the reported results are averaged over all classes.

Method	$Acc_{rt} \uparrow$	$Acc_{ft} \downarrow$	$GAP_{ft} \downarrow$	$GAP_f^{\max} \downarrow$
ViT-Small				
Pre-Fgt-R32	94.46 ± 0.19	6.85 ± 4.14	1.85	12.00
Pre-Fgt-R64	94.56 ± 0.17	5.70 ± 2.74	0.70	5.00
Ours-R32	75.59 ± 2.93	1.90 ± 1.57	3.10	5.00
Ours-R64	83.93 ± 0.86	2.10 ± 1.37	2.90	4.60
ViT-Base				
Pre-Fgt-R32	95.41 ± 0.16	7.03 ± 4.67	2.03	14.20
Pre-Fgt-R64	95.65 ± 0.12	5.34 ± 5.46	0.34	21.00
Ours-R32	87.68 ± 1.33	5.29 ± 3.01	0.29	11.20
Ours-R64	92.25 ± 0.44	6.09 ± 1.99	1.09	5.20

Table B.3: **Single-class unlearning** performance on CIFAR-10 using an ImageNet-pretrained VGG-16 with passport C-LoRA rank 128 (3 seeds, mean $\pm$ std). dist is the aggregate distance to the retrain reference (Eq. 5.23); lower is better.

Method	Acc_{ft}	Acc_{rt}	MIA	dist $\downarrow$
<i>Retrain (full)</i>	0.0	95.2 ± 0.0	0.322 ± 0.01	0.00
<i>Retrain (LoRA)</i>	0.0	86.9 ± 0.4	0.246 ± 0.02	0.00
Finetune	1.3 ± 0.7	95.1 ± 0.1	0.190 ± 0.02	0.15 ± 0.02
Random Labels	8.1 ± 3.2	76.2 ± 3.5	0.000 ± 0.00	0.59 ± 0.05
Delete	0.0 ± 0.0	92.8 ± 0.7	0.235 ± 0.04	0.11 ± 0.04
Bad Teacher	0.0 ± 0.0	11.6 ± 0.6	0.000 ± 0.00	1.16 ± 0.01
Boundary Expand	0.0 ± 0.1	80.7 ± 6.6	0.171 ± 0.08	0.30 ± 0.01
Boundary Shrink	0.2 ± 0.1	68.9 ± 4.1	0.603 ± 0.52	0.76 ± 0.12
SalUn	9.4 ± 6.4	76.3 ± 2.3	0.000 ± 0.00	0.61 ± 0.08
UNSIR	0.3 ± 0.3	63.7 ± 2.3	0.440 ± 0.01	0.44 ± 0.02
Amnesiac	0.0 ± 0.0	93.3 ± 0.6	0.035 ± 0.01	0.31 ± 0.01
Ours	0.0 ± 0.0	86.2 ± 0.2	0.253 ± 0.12	0.09 ± 0.07

recurs: Amnesiac drives MIA to 0.035 where its retrain reference sits at 0.322. These results confirm that the mechanism transfers effectively beyond Transformers.

B.5 Instance-Level Unlearning

We additionally demonstrate our framework in the *instance-level* setting, where the goal is to forget a subset of training samples whose classes remain in the data. The forgotten samples are assigned to a dedicated forget passport, trained jointly with the retain-all passport following the main training setup and optimised with the loss-matching objective introduced below. We evaluate two settings: a *canary* setting on CIFAR-10, in which the forget set is constructed so that forgetting is measurable, and *unmodified* MNIST and CIFAR-10, which

probe behaviour on natural data and larger forget ratios.

Why the class-level objective does not transfer. The class-level objective masks the ground-truth label and drives $p(y_{\text{true}} | x) \rightarrow 0$. At class level this is exact, yet at instance level it targets forgotten samples belonging to classes that remain in training, so a retrained oracle still classifies them correctly (0.984 on held-out CIFAR-10 forget samples), and driving their accuracy to 0 is over-forgetting: a confidently incorrect prediction on exactly the forgotten set is a stronger signal to an adversary than the memorisation it removes.

The target is therefore the behaviour of a model that never observed the data, which differs in confidence rather than correctness (on CIFAR-10, 1.6 points of forget-set accuracy but a factor of 48 in predictive entropy). After each epoch we record the per-class distribution $\mathcal{R}_c$ of cross-entropy on the validation split, assign each forgotten sample a fixed quantile according to its difficulty rank within its class,

$$q_j = \frac{r_j + \frac{1}{2}}{n_{y_j}}, \quad (\text{B.5})$$

where r_j is the loss rank of x_j among the n_{y_j} forgotten samples of its class, measured at the end of the warm-up epoch, before memorisation flattens the per-sample losses. Its loss is then matched to the corresponding quantile of the current reference,

$$\ell^*(x_j) = Q_{q_j}(\mathcal{R}_{y_j}), \quad \mathcal{L}_{\text{match}} = \frac{1}{|\mathcal{D}_{\text{forget}}|} \sum_{x_j \in \mathcal{D}_{\text{forget}}} |\ell(x_j) - \ell^*(x_j)|, \quad (\text{B.6})$$

with $\ell(x) = -\log p(y | x)$. The rank-based assignment gives easy samples low and hard samples high targets, so the forget set reproduces both the marginal and the input-conditional loss distribution of unseen data.

Canary setting. On unmodified data the quantity to be erased is small: each forgotten image is redundant given the remaining samples of its class, so the memorised model and the oracle differ by only 1.6 points on the forget set, and distinguishing methods inside that margin is unreliable. In the canary setting we therefore construct the forget set from 10% of the CIFAR-10 training split (4,500 images): each image keeps its pixels, but its label is replaced by an incorrect one drawn uniformly at random, $y'_j \sim \mathcal{U}(\{0, \dots, C-1\} \setminus \{y_j\})$. These pairs can only be predicted through memorisation, meaning *assigned*-label accuracy measures memorisation that survived (oracle 0.001), *true*-label accuracy measures knowledge that survived (oracle 0.988).

Table B.4 reports both axes together with retained accuracy and MIA against thirteen

Table B.4: Instance-level unlearning with canaries, CIFAR-10, 4,500 forgotten images (10%), C-LoRA rank 64 (3 seeds; dist mean $\pm$ std).

Method	Assigned	True	Retain	MIA	dist
<i>Memorised model</i>	1.000	0.000	–	–	–
<i>Retrain oracle</i>	0.001	0.988	0.982	0.946	–
Ours	0.003	0.950	0.976	0.874	0.118$\pm$0.057
Finetune	0.010	0.907	0.910	0.799	0.308 $\pm$ 0.120
Gradient Ascent	0.157	0.703	0.897	0.652	0.819 $\pm$ 0.082
Wood Fisher	0.080	0.701	0.801	0.496	0.996 $\pm$ 0.080
Amnesiac	0.129	0.749	0.972	0.008	1.314 $\pm$ 0.101
Boundary Expand	0.163	0.321	0.459	0.347	1.950 $\pm$ 0.235
Delete	0.293	0.364	0.734	0.093	2.016 $\pm$ 0.026
Boundary Shrink	0.145	0.165	0.355	0.428	2.111 $\pm$ 0.091
L2UL (adv)	0.998	0.001	0.909	0.646	2.356 $\pm$ 0.106
L2UL (imp)	0.998	0.001	0.909	0.647	2.356 $\pm$ 0.106
Fisher	0.101	0.103	0.102	0.428	2.383 $\pm$ 0.163
SalUn	0.481	0.113	0.524	0.230	2.528 $\pm$ 0.034
Random Labels	0.362	0.120	0.414	0.197	2.545 $\pm$ 0.142
Bad Teacher	0.932	0.040	0.933	0.221	2.653 $\pm$ 0.078

competitors; here dist sums the absolute gaps to the oracle over all four columns. The second axis separates failures a single metric cannot: both L2UL variants score 0.998 on the assigned labels (they remain essentially the memorised model) while retaining 0.91 test accuracy, so an assigned-label metric alone would report them as healthy; Bad Teacher retains the memorisation and destroys the true-label knowledge; Fisher and the boundary methods collapse the model, depressing both axes for reasons unrelated to forgetting. Ours attains the smallest distance to the oracle; among the competitors only Finetune approaches the oracle on both axes. Competitor settings were transferred from the class-level protocol, so their figures are a lower bound. UNSIR is omitted as inapplicable, since it synthesises noise per forgotten class and an instance-level forget set provides no class to synthesise for.

Unmodified data and scale. In the second setting, without canaries, the forget set consists of 1% of the training data of MNIST and CIFAR-10. Averaged over both datasets (three seeds each), the forget passport reaches a distance to the oracle of 0.035, with test accuracy of 0.989 and forget-set accuracy of 0.997, matching the oracle’s generalization instead of suppressing it. The result is stable when the forget ratio grows to 50% on CIFAR-10 (dist = 0.026), a regime in which objectives that suppress the ground-truth label destabilize training, as they oppose the retain pass on half the data.

B.6 Backbone Statistics Unlearning

Beyond masked and uniform output objectives, we also demonstrate that the proposed method can be adapted to other unlearning objectives, such as those derived from backbone statistics. Specifically, we minimize the KL divergence between the model’s outputs and those produced by the frozen backbone combined with a randomly initialized classifier, effectively approximating the pre-learning state. Table B.5 summarizes the results obtained using these backbone-based statistics. Here, the unlearning gap is defined as the distance to the pre-learning classifier accuracy, rather than the chance level, as used previously for the uniform objective. The consistently small accuracy gaps across varying forgetting scenarios indicate that the backbone statistics objective provides a reliable and predictable unlearning target. By anchoring to the original backbone’s feature space, this approach preserves useful representations from the pretrained model. Overall, this method provides a practical and stable alternative to uniform objectives and demonstrates our framework’s capability to accommodate diverse unlearning objectives.

Table B.5: **Single-class unlearning performance using backbone statistics** with LoRA ranks R8 and R32 on MNIST and CIFAR-10 across various ViT backbones. Atomic passports are trained once, with one passport per class, and the reported results are averaged over all classes.

Method	MNIST				CIFAR10			
	$Acc_{rt} \uparrow$	$Acc_{ft} \downarrow$	$GAP_{ft} \downarrow$	$GAP_f^{\max} \downarrow$	$Acc_{rt} \uparrow$	$Acc_{ft} \downarrow$	$GAP_{ft} \downarrow$	$GAP_f^{\max} \downarrow$
ViT-Tiny								
Ours-R8	98.52±0.20	7.74±12.46	1.46	5.04	88.29±2.02	8.36±12.65	1.14	3.80
Ours-R32	99.24±0.08	8.13±13.31	1.08	3.40	92.87±0.34	8.63±13.05	1.53	3.20
ViT-Small								
Ours-R8	99.07±0.13	1.08±25.03	1.64	6.00	94.46±1.15	7.95±20.95	1.93	9.90
Ours-R32	99.38±0.09	10.57±25.04	1.89	6.14	95.58±0.33	7.93±21.33	2.35	12.10

B.7 Inference and Training Time Analysis

All runtime measurements were conducted on a NVIDIA H100 GPU for CIFAR-10 single-class unlearning. Across all competing unlearning methods, the average total unlearning time is 18.9 minutes, with a standard deviation of 8.7 minutes. The two slowest competitors are L2UL (32.1 minutes) and Finetune (26.7 minutes).

For a realistic comparison, we consider *TrustErase* trained for 11 initial tasks (one for

each class, plus an all-learning task). The total training time for our method is 62.53 minutes, which is about 32 minutes slower on average than the total training and unlearning time of competing methods. However, in return, our approach provides instant unlearning for predefined forget sets without requiring access to the training data or additional training time. If a forget set is not included in the predefined collection, the second hypernetwork training phase, which in the worst case learns all possible atomic-passport combinations, takes at most 1 day. This phase can be carried out after initial model deployment.

B.8 Reconstruction Attack Analysis

We examine whether the shared LoRA bottleneck C can be reconstructed from publicly released decomposed adapters. The setting assumes a frozen ViT backbone instrumented with LoRA adapters on all Q/V projections. For each layer ℓ , the victim applies a low-rank update

$$\Delta W_\ell = \frac{\alpha}{r} B_\ell C A_\ell, \quad (\text{B.7})$$

where the per-layer matrices $\{A_\ell, B_\ell\}$ and the shared $C \in \mathbb{R}^{r \times r}$ are private. Before release, the victim deterministically decomposes $C = C_1 C_2$ and publishes

$$A'_\ell = C_2 A_\ell, \quad B'_\ell = B_\ell C_1, \quad (\text{B.8})$$

so that the public adapter appears as a standard LoRA update $\Delta W_\ell = \frac{\alpha}{r} B'_\ell A'_\ell$.

The attacker instantiates its own factors $\{\hat{A}_\ell, \hat{B}_\ell\}$ and a shared $\hat{C}$, inducing

$$\Delta \hat{W}_\ell = \frac{\alpha}{r} \hat{B}_\ell \hat{C} \hat{A}_\ell. \quad (\text{B.9})$$

To remain consistent with the public decomposition, $\hat{C}$ is factored as $\hat{C} = \hat{C}_1 \hat{C}_2$, defining $\hat{A}'_\ell = \hat{C}_2 \hat{A}_\ell$ and $\hat{B}'_\ell = \hat{B}_\ell \hat{C}_1$. The attacker optimizes a combined objective

$$L_{\text{tot}} = L_{\text{cls}} + \lambda L_{\text{rec}}, \quad (\text{B.10})$$

where the supervised loss is

$$L_{\text{cls}} = \frac{1}{|D|} \sum_{(x,y) \in D} \text{CE}(f(x; W_\ell + \Delta \hat{W}_\ell), y), \quad (\text{B.11})$$

and the reconstruction loss enforces agreement with the released public factors:

$$L_{\text{rec}} = \sum_{\ell} \left(\|\hat{A}'_{\ell} - A'_{\ell}\|_F^2 + \|\hat{B}'_{\ell} - B'_{\ell}\|_F^2 \right). \quad (\text{B.12})$$

Because the LoRA parameterization admits an $r \times r$ gauge freedom (basis changes in $\hat{C}$ that can be absorbed by $\hat{A}_{\ell}$ and $\hat{B}_{\ell}$ without altering $\Delta\hat{W}_{\ell}$) the inverse problem is highly underdetermined. As a consequence, many attacker optima match the public factors only loosely while remaining incompatible with the victim’s private $\{A_{\ell}, B_{\ell}\}$.

Evaluation uses both parameter-level and function-level metrics. Factor reconstruction errors (FR-A, FR-B) and update-agreement error $E_{\Delta W}$ quantify alignment with public factors, while functional fidelity is measured by substituting $\hat{C}$ into the victim model and comparing predictions, accuracy, and logit similarity with the true- C model.

We optimized the attack objective for 20 epochs using AdamW with a learning rate of 10^{-3} , weight decay of 10^{-4} , and a cosine annealing scheduler. Despite this configuration, empirically, reconstruction fails. Errors remain large (FR-A ≈ 0.976 , FR-B ≈ 0.952 , $E_{\Delta W} \approx 0.947$), indicating poor alignment with the released factors. More critically, transplanting $\hat{C}$ into the victim collapses performance: $\text{Acc}(\hat{C}) = 0.0955$ versus $\text{Acc}(C) = 0.8712$, with negative logit cosine similarity and $\text{KL} \approx 3.70$. Although the attacker achieves low loss on its own model, optimization exploits gauge degrees of freedom rather than recovering the true bottleneck.

Under this reconstruction threat model, the shared LoRA bottleneck C is not identifiable. The decomposition-based release reveals insufficient structure, and the attacker’s optimization converges to degenerate solutions that fail both in parameter space and in function space. The passport matrix C therefore remains unrecoverable.

B.9 Audit Protocol and Use Cases

To guarantee the integrity of the unlearning process and foster trust in deployed models, we design an audit protocol. The objective of this protocol is to provide users with a formal assurance that the model they receive from a service provider correctly incorporates the specific unlearning task they requested. The protocol operates under a ‘trust-but-verify’ paradigm, as illustrated in Figure B.1, involving three main entities:

- **Verification Authority (VA):** The central entity that performs the initial training and secret distribution. It acts as the root of trust for certifying compliance.

- **Trusted Providers:** Entities initially entrusted with all secrets (base matrices and passports) to handle model serving. They are assumed to be honest-but-verifiable.
- **Users:** Clients who request specific unlearning tasks (e.g., forgetting a specific class) and verify the outcomes through the VA.

Beyond verifying the correctness of the unlearning operation itself, the protocol ensures that the resulting model exhibits functional behavior consistent with the intended task. Thus, verification encompasses both *structural validation*, checking that the correct passport was used, and *functional validation*, which assesses whether the produced model behaves similarly to one built with the user’s original task specification.

Secrets Distribution. After central training, the VA provisions the base LoRA matrices A, B together with the full set of original passports $\{C_k\}$ to a set of Trusted Providers. Each user U_i is securely issued their ground-truth passport C_i , which serves as a non-transferable receipt for task i . Providers are assumed to be honest-but-verifiable: they store the secrets and serve models, but the VA remains responsible for certifying that the correct unlearning configuration has been applied.

Verification Workflow. The verification process is initiated by the User. After receiving the unlearned model for task i , the user contacts the VA and submits both the received model parameters (specifically the decomposed adapters A'_i, B'_i) and their ground-truth passport C_i . Upon receipt, the VA independently audits the submitted model. Leveraging the private base matrices A and B (which are retained securely by the VA), it extracts the effective passport $\hat{C}_i$ embedded within the model weights via pseudo-inverse reconstruction:

$$\hat{C}_i = B^\dagger B'_i A'_i A^\dagger. \quad (5.12)$$

Simultaneously, the VA instantiates the reference ‘honest’ model by combining the secure base matrices with the user’s verified ground-truth passport C_i . The VA now holds two distinct objects: the ideal passport C_i and the passport $\hat{C}_i$ extracted from the deployed model. Verification consists of checking whether $\hat{C}_i$ is a faithful approximation

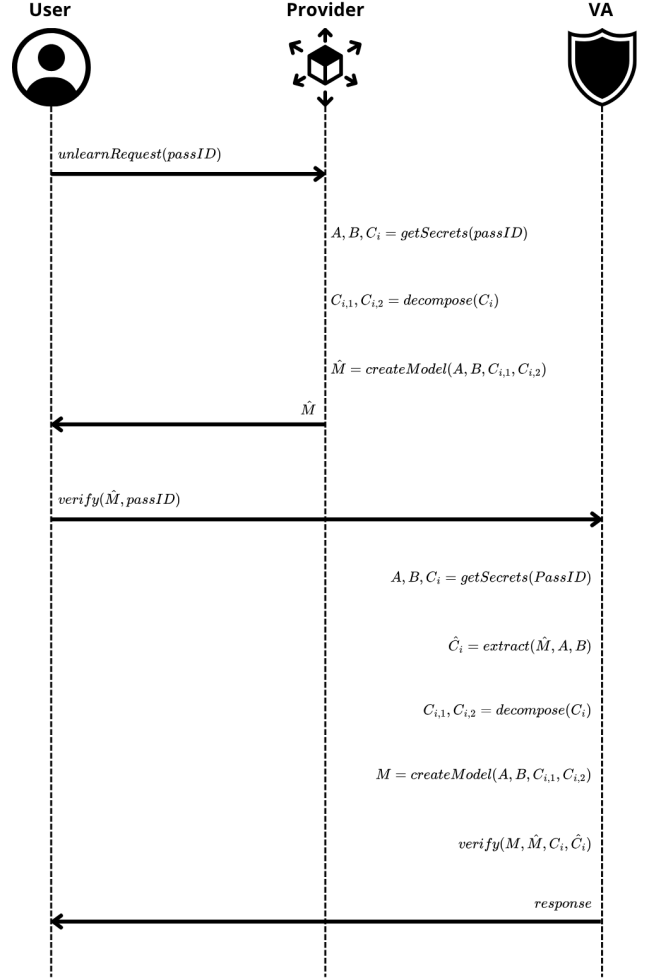


Figure B.1: **Overview of the proposed verification protocol for auditable unlearning.** The protocol follows a trust-but-verify paradigm involving three entities: a Verification Authority (VA) that conducts initial training and distributes secrets, trusted providers that perform model serving using designated passports, and users who request specific unlearning tasks. The VA validates both the structural correctness of the passport used and the functional behavior of the resulting model, ensuring that the delivered model faithfully reflects the user’s intended unlearning request.

of C_i within a predefined tolerance:

$$\|\psi(\hat{\mathbf{C}}_i) - \psi(\mathbf{C}_i)\| < \epsilon_T, \quad (5.22)$$

where $\psi(\cdot)$ denotes an arbitrary feature mapping and ϵ_T is a predefined tolerance accounting for numerical imprecision. If the condition holds, the VA certifies that the provider has correctly applied the requested unlearning task.

As discussed in Section B.8, structural similarity alone may not suffice; adversarial reconstruction attacks can exploit small deviations. Therefore, the VA complements the structural check with a *functional verification* step, comparing the behavior of the served model $(W_0 + B\hat{C}_iA)$ against a reference built with C_i . Agreement in logits or activations within ϵ_T ensures that the unlearning task has been faithfully executed.

Use Cases. This protocol supports multiple deployment scenarios: (i) regulatory audits, where third parties require verifiable evidence of compliance with ‘right-to-be-forgotten’ requests; (ii) multi-tenant services, where different users demand distinct unlearning configurations; and (iii) adversarial settings, where providers must prove that no unauthorized passport substitution has occurred. In all cases, the combination of structural and functional verification provides lightweight yet auditable guarantees of trustworthy unlearning.