VALIANT: Prompt Instability for Active Learning in Black-Box Medical Imaging

Dwarikanath Mahapatra[1], **Behzad Bozorgtabar**[2, 3], **Sudipta Roy**[4]
Imran Razzak[5], **Mauricio Reyes**[6,7]
[1]Department of Computer Science, Khalifa University, Abu Dhabi, UAE
[2]EPFL, Lausanne, Switzerland.
[3]University of Southern Denmark, Odense, Denmark
[4]Jio University, Navi Mumbai, India.
[5]MBZUAI, Abu Dhabi, UAE.
[6]University of Bern, Switzerland.
[7]Dept. of Radiation Oncology, Inselspital, Bern University Hospital, Switzerland.

Abstract

The deployment of large, black-box foundation models for medical image classification is often hindered by the high cost of acquiring large, task-specific labeled datasets for fine-tuning. While active learning (AL) presents a promising solution, many state-of-the-art AL methods are computationally expensive or require full access to internal model parameters. We present VALIANT (**V**isual **A**daptation and **L**earning **I**ntegration for **A**ctive lear**N**ing **T**asks), a new active learning framework designed to efficiently adapt black-box foundation models by overcoming these limitations. VALIANT introduces a lightweight *Visual Prompt Decoder* (VIPD), trained via **unsupervised** *Zero-Order Optimization* (ZOO), to generate task-specific visual prompts without internal model access. Our core contribution is a perturbation-based ranking strategy that leverages this VIPD to formulate a computationally efficient, gradient-aware informativeness metric. This metric, which we term prompt instability, identifies the most impactful samples for the labeling budget. VALIANT further enhances this process by incorporating anatomical information from unsupervised segmentation maps to generate more discriminative visual prompts. Extensive evaluations on multiple medical datasets demonstrate VALIANT's superior performance and significant reduction in labeling costs compared to a range of existing active learning techniques, positioning it as a scalable and practical solution for medical image analysis.

Introduction

Foundation models (FMs) have demonstrated remarkable potential in medical image computing, offering the prospect of reduced reliance on task-specific labeled data through their capacity for broad generalization. However, their effective deployment in the clinical domain is hindered by several key challenges. First, training and deploying these large models demands substantial computational resources, limiting accessibility. Second, acquiring the large, expertly annotated datasets necessary for training or fine-tuning remains a significant bottleneck.

Parameter-Efficient Fine-Tuning (PEFT) techniques like adapter-based tuning (He et al. 2021) and low-rank adaptation (LoRA) (Hu et al. 2021) offer a promising avenue to mitigate the data demands of adapting FMs. However, they still rely on substantial labeled data for optimal adaptation. Critically, these methods typically require access to the FM's internal parameters, which is often restricted when FMs are deployed via APIs or accessed through proprietary platforms. This limitation significantly curtails their applicability in many real-world clinical settings.

To address the challenge of labeled data scarcity, Active Learning (AL) aims to maximize model performance while minimizing annotation costs. Unfortunately, traditional AL methods are often ill-suited for adapting black-box FMs. Many rely on direct access to internal model parameters or require well-calibrated uncertainty estimates, which can be unreliable from black-box models. Furthermore, gradients, a key component for many AL strategies, are typically unavailable for API-based FMs. This necessitates the development of novel AL strategies that can effectively adapt FMs without access to their internal mechanisms.

To bridge this gap, we introduce VALIANT (**V**isual **A**daptation and **L**earning **I**ntegration for **A**ctive lear**N**ing **T**asks), a novel and efficient active learning framework designed to adapt black-box FMs for medical image classification. VALIANT employs a lightweight *Visual Prompt Decoder* (VIPD), trained via unsupervised *Zero-Order Optimization* (ZOO), to generate task-aware visual prompts that guide FM predictions without requiring access to model weights. Our core technical contribution is a perturbation-based ranking strategy that leverages the VIPD to derive a computationally efficient, gradient-aware informativeness metric. This metric, which we term *prompt instability*, assesses the sensitivity of a sample's prompt representation to input variations, allowing us to identify and prioritize the most impactful samples for the labeling budget. VALIANT further enhances this process by incorporating anatomical information from unsupervised segmentation maps, which leads to the generation of more discriminative visual prompts and superior active learning performance. We evaluate VALIANT across multiple medical imaging datasets and demonstrate its superior adaptation efficiency compared to a range of existing active learning techniques.

Our contributions can be summarized as follows:

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

- We introduce **VALIANT**, a novel active learning framework that overcomes the limitations of black-box foundation models by integrating visual prompting, unsupervised zero-order optimization, and a gradient-aware informativeness metric.
- We propose a **computationally efficient, gradient-aware informativeness metric** based on a perturbation-based ranking strategy. We show that our *prompt instability* metric effectively identifies informative samples without the high computational cost of traditional gradient-based methods.
- We demonstrate that integrating **anatomical information from unsupervised segmentation maps** into the prompt generation process leads to more discriminative prompts and significantly improved active learning performance on medical imaging tasks.
- We provide a **comprehensive evaluation** of VALIANT's performance and generalizability across diverse medical imaging datasets, including detailed ablation studies and a discussion on computational efficiency and hyperparameter sensitivity.

Related Work

Adapting Black-Box Foundation Models: Adapting pre-trained models without internal access is crucial for foundation models deployed via APIs (Ye et al. 2023). Zero-Order Optimization (ZOO) algorithms like BBT (Sun et al. 2022b) and BBTv2 (Sun et al. 2022a) suffer from high variance (Liu et al. 2020) and have limited vision generalization (Oh et al. 2023). For vision applications, BAR (Tsai, Chen, and Ho 2020) uses one-sided gradient approximation, while BlackVIP (Oh et al. 2023) employs a Simultaneous Perturbation Stochastic Approximation with Gradient Correction (SPSA-GC) to improve convergence. (Paranjape et al. 2024) extended ZOO to prompted segmentation, showcasing its versatility. (Park et al. 2025) reduce the problem dimensionality by re-parameterizing prompts in low-rank representations leading to faster training and improved performance over (Oh et al. 2023). Our work builds on these black-box adaptation methods by applying them to active learning for both medical image classification and segmentation, a relatively underexplored area.

Parameter-Efficient Fine-Tuning (PEFT): Parameter-Efficient Fine-Tuning (PEFT) offers efficient adaptation by modifying only a small subset of parameters, reducing computational cost and overfitting (Chen et al. 2020). Inspired by NLP successes, PEFT has been applied to vision and vision-language tasks using adapter-based methods (AdaptFormer (Chen et al. 2022), CLIP-Adapter(Gao et al. 2025)) and prompt-based methods (CoOp(Zhou et al. 2022b), CoCoOp(Zhou et al. 2022a)). Visual prompting techniques, such as VPT (Jia et al. 2022), and VP (Bahng et al. 2022), modify the input space, offering the advantage of adapting models without requiring full access to the model architecture, which is particularly relevant for black-box scenarios. While PEFT improves fine-tuning efficiency, it still relies on labeled data. Our work combines visual prompting efficiency with active learning to minimize this reliance.

Active Learning in Medical Image Analysis: Active learning in medical imaging aims to efficiently select informative samples for annotation. Various selection techniques have been explored for deep learning, including sample entropy, model uncertainty (using test-time Monte-Carlo dropout), Fisher information, visual saliency, and clustering-based approaches. Methods like (Wang et al. 2017a) use entropy and margin sampling, while others employ GANs to synthesize informative samples (Zhou et al. 2016; Mayer and Timofte 2018). Two-step approaches combining uncertainty and similarity metrics have also been proposed (Yang et al. 2017). However, the reliability of uncertainty estimates in deep networks has been questioned due to calibration issues. Recent work emphasizes the importance of interpretability in active learning for improved sample selection and performance gain with fewer annotations (Budd, Robinson, and Kainz 2021; Mahapatra et al. 2021; Mahapatra, Poellinger, and Reyes 2023; Wang et al. 2024).

Methodology

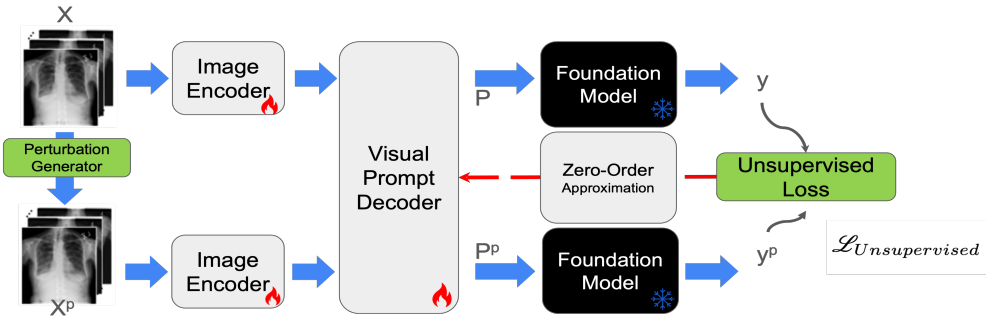
Figure 1 (a) shows the workflow of our proposed method. The original image (X) and perturbed version (X^p) are passed through the image encoder whose output is fed to the VIPD. Original image and VIPD are combined and input to the Foundation Model (FM) to give different outputs y and y^p. These outputs are used to compute gradients and train the VIPD. The trained VIPD is used to quantify informativeness of new samples and rank them for labeling.

Training the Visual Prompt Decoder (VIPD)

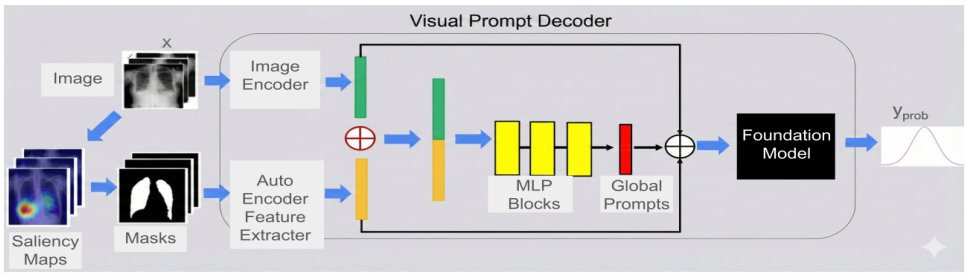
The Visual Prompt Decoder (VIPD) is a lightweight network at the core of our framework, trained to generate task-aware visual prompts that guide the black-box foundation model (FM). As shown in Figure 1, the VIPD is not directly coupled with the FM's internal parameters, which is essential for our black-box approach. Its training is unsupervised and guided by a novel unified loss function that leverages Zero-Order Optimization (ZOO).

VIPD Architecture: The VIPD's architecture is designed to integrate both image features and anatomical information. It consists of an image encoder, a prompt generation network, and a mechanism for incorporating anatomical cues. The image encoder, a Vision Transformer (ViT) pre-trained with a Masked Auto Encoder (MAE) objective, processes the input image to produce a robust feature embedding. Concurrently, the original image is used to derive an unsupervised segmentation map, which highlights anatomical regions of interest. This map is then processed through a small denoising autoencoder to create a compact, 512-dimensional vector. This vector is concatenated with the image's feature embedding. The concatenated vector is then passed through a multi-layer perceptron (MLP) network to output a global prompt, a lower-dimensional embedding vector designed to modulate the FM's overall classification output.

Zero-Order Optimization (ZOO): Since the FM is a black-box model, we cannot use backpropagation to train the VIPD. Instead, we employ a Zero-Order Optimization



(a)



(b)

Figure 1: Pipeline overview: (a) The training of the visual prompt decoder (VIPD) using feedback from the Foundation Model and zero order optimization (ZOO). Blocks with blue-ice icon are frozen, while blocks with the red-fire icon are trainable components. (b) Detailed depiction of the VIPD.

(ZOO) method to approximate the gradient of our unified loss function with respect to the VIPD's parameters. We use Simultaneous Perturbation Stochastic Approximation with Gradient Correction (SPSA-GC) (Spall 2000) for this purpose. The gradient $\hat{g}_i$ is approximated as:

$$\hat{g}_i = \frac{L(\theta + c\Delta) - L(\theta - c\Delta)}{2c}\Delta^{-1} \tag{1}$$

where L is our unified loss function $\mathcal{L}_{VIPD}$, c is a small perturbation factor, and Δ is a random perturbation vector. This computationally efficient approach allows us to train the VIPD without requiring any access to the FM's internal weights or gradients.

Unsupervised Loss Function

In the active learning setting, ground-truth labels are not available during the VIPD training phase. The VIPD generates prompts that make the FM's output highly sensitive to input perturbations, a proxy for identifying samples near the decision boundary. The VIPD is trained using a unified, unsupervised loss function, $\mathcal{L}_{VIPD}$ which combines three distinct loss terms:

$$\mathcal{L}_{VIPD} = \lambda_1\mathcal{L}_{sens} + \lambda_2\mathcal{L}_{coh} + \lambda_3\mathcal{L}_{div} \tag{2}$$

where λ_1, λ_2, and λ_3 are weighting coefficients that balance the contribution of each loss term.

Perturbation Sensitivity Loss ($\mathcal{L}_{sens}$): This is the core loss function that drives our perturbation-based ranking strategy. It is designed to train the VIPD to generate prompts that make the FM's output highly sensitive to input perturbations. The rationale is that a large output change under a small perturbation indicates a sample is located in a high-uncertainty region, making it an ideal candidate for labeling. By maximizing the dissimilarity between the FM's prediction distributions for the original input (y) and its perturbed counterpart (y^p), the VIPD learns to generate prompts that highlight regions of high uncertainty or decision-boundary proximity. This is a crucial step for identifying informative samples in an unsupervised manner. We use the Kullback-Leibler (KL) divergence to quantify this dissimilarity:

$$\mathcal{L}_{sens} = -D_{KL}(y \| y^p) \tag{3}$$

The negative sign ensures that minimizing the loss corresponds to maximizing the KL divergence, thereby forcing the FM to be more responsive to subtle input changes guided by the prompt.

Anatomical Coherence Loss ($\mathcal{L}_{coh}$): The VIPD outputs a visual prompt P with the *same* shape as X. To ensure the generated prompts are discriminative and semantically meaningful, we introduce a loss term that leverages the anatomical information derived from the unsupervised segmentation maps. This loss encourages the VIPD to produce prompts that are coherent with the underlying anatomical

structure of the input image. It is calculated as the squared L2 norm of the difference between the feature representations of the original image and the prompted image:

$$\mathcal{L}_{coh} = \|F(X+P) - F(X)\|_2^2 \tag{4}$$

Here, F is a pre-trained feature extractor (e.g., the ViT image encoder) that is kept static during this training phase. This loss ensures that the prompts do not drastically alter the original semantic content of the images, but rather subtly guide the FM's attention towards anatomically relevant regions.

Prompt Diversity Loss ($\mathcal{L}_{div}$): To prevent the VIPD from learning a single, generic prompt for all images, we include a diversity loss. This loss encourages the VIPD to generate unique prompts for different samples, which is essential for capturing subtle, class-discriminative features. We calculate this loss by penalizing prompts that are too similar in the embedding space. We use the cosine similarity (S) between the prompt for the original image (P) and the prompt for the perturbed image (P^p):

$$\mathcal{L}_{div} = 1 - \frac{P \cdot P^p}{\|P\| \cdot \|P^p\|} = 1 - S \tag{5}$$

By minimizing this loss, we force the prompts to become more distinct in response to input perturbations, thus promoting a more robust and flexible VIPD.

Ranking Informative Samples

After training the VIPD, we rank unlabeled images to select the most valuable samples for annotation. Our approach leverages a perturbation-based ranking strategy to identify samples that are likely to maximize model improvement with minimal labeling effort. This is achieved through a novel metric we call **Prompt Instability**.

The Prompt Instability Metric: The core idea is that samples exhibiting significant changes in their prompt representations under small perturbations are hypothesized to lie near the decision boundary and thus be highly informative for refining the model's knowledge. This metric serves as a computationally efficient proxy for gradient-based uncertainty. For an unlabeled image X, we perform the following steps to compute its Prompt Instability score:

1. We define the perturbed input as

$$X^p = X + \delta, \qquad \delta \sim \mathcal{N}(0, \sigma^2 \mathbf{I}). \tag{6}$$

2. We quantify the difference between the two prompts in the embedding space using cosine distance,

$$I_{instability} = 1 - \frac{P \cdot P^p}{\|P\|\|P^p\|} \tag{7}$$

3. We use the trained VIPD to generate prompts for both the original and perturbed images:

$$P = \text{VIPD}(X), \quad P^p = \text{VIPD}(X^p) \tag{8}$$

A higher $I_{instability}$ value indicates a greater impact of the perturbation on the prompt representation and, consequently, higher informativeness. The unlabeled samples are then ranked in descending order of their $I_{instability}$ score, and the top-ranked images are selected for human annotation. This streamlined approach provides a single, intuitive measure for effective sample selection.

Experiments

Dataset Details

To evaluate the effectiveness of our proposed VALIANT framework, we conducted extensive experiments across multiple publicly available medical imaging datasets spanning diverse modalities, including X-rays (CheXpert (Irvin, Rajpurkar, and et al. 2019), NIH (Wang et al. 2017b)), histopathology (PCam (Veeling et al. 2018)), skin lesions (ISIC 2019 (Tschandl, Rosendahl, and Kittler 2018)), and retinal fundus (REFUGE (Orlando, Fu, and et al. 2020)). A summary of these datasets alongwith the dataset specific parameter values can be found in Table 1.

Implementation Details

All of our experiments are implemented in PyTorch and run on an NVIDIA A6000 48GB GPU. For the foundation models, we utilize CLIP checkpoints with ViT-B/16 and ViT-L/14 backbones, as well as BioVil (Boecking et al. 2022) and MedCLIP (Wang et al. 2022) models pre-trained on large-scale medical image datasets. **Hyperparameters.** $\lambda_{1...3}$ are selected *once per dataset* via a small coarse grid on the initial validation split and then kept fixed for all AL cycles/budgets (values in Table 1).

Baselines and Justification: To provide a comprehensive and rigorous evaluation, we compare VALIANT against a diverse set of state-of-the-art and standard active learning baselines under identical experimental settings.

- **Fully Supervised Learning (FSL):** We train a Transformer-based approach (Shamshad et al. 2023) on the entire labeled training set to establish an upper-bound performance benchmark.
- **Random Sampling:** This serves as the fundamental baseline, representing the performance of a non-strategic approach to sample selection and establishing a lower-bound for AL performance.
- **Uncertainty Sampling:** We use a standard entropy-based uncertainty sampling (Gal, Islam, and Ghahramani 2017) method to select samples with the highest prediction entropy, which is a classic and widely-used AL strategy.
- **Coreset:** Coreset (Sener and Savarese 2018) is a diversity-based method that selects samples to approximate the geometry of the entire dataset. It is a prominent and powerful baseline, especially for visual tasks.
- **Bayesian Active Learning (BALD):** We implement Bayesian Active Learning by Disagreement (BALD) (Gal, Islam, and Ghahramani 2017), a widely-used uncertainty-based method that selects samples with the

Modality	Dataset	# Images	# Labels	λ_1	λ_2	λ_3	w_s	w_m	w_l	α_1
X-ray	CheXpert (Irvin, Rajpurkar, and et al. 2019)	224,316	14	1.2	1.4	0.95	4	3	2	0.4
Histopathology	PCam (Veeling et al. 2018)	327,680	2	1.6	1.4	1.0	5	3	3	0.4
Skin Lesions	ISIC 2019 (Tschandl, Rosendahl, and Kittler 2018)	25,331	9	1.5	0.9	1.3	4.5	3.2	2.2	0.4
Retinal Fundus	REFUGE (Orlando, Fu, and et al. 2020)	1,200	2	1.5	1.3	1.5	4	3	2	0.5

Table 1: Summary of publicly available medical imaging datasets used for evaluation. The parameter values used for each dataset is also shown.

highest mutual information between predictions and model posterior.

- **Black-Box Prompt Optimization (BAPS):** We benchmark against BAPS (Paranjape et al. 2024), the most relevant prior work, which also utilizes visual prompts and zero-order optimization in a black-box setting.

All AL baselines (Random, Entropy, BALD, Coreset, BAPS) operate on the *same frozen FM* (BioVil or MedCLIP) with identical data splits, label budgets, acquisition batch size, and seeds. When a predictor is required, we attach the *same* linear head on the pooled FM features and train it only on the labeled pool each round; the FM itself is never tuned and no gradients/weights are accessed. Coreset uses the same embedding; BAPS uses the same FM and splits with recommended settings.

VALIANT Training and Evaluation: Our VALIANT framework was evaluated in two distinct training modes: **Single-Modality Training (VALIANT$_{Single}$)**, where the VIPD is trained separately on each modality, and **Multi-Modality Training (VALIANT$_{Multi}$)**, where a single VIPD is trained on a combined dataset from all modalities. We measure classification performance using the Area Under the Curve (AUC) and the active learning cycle iteratively adds samples until a specified budget is reached. All experiments were conducted with a fixed seed and averaged over 10 independent runs to ensure robust and reproducible evaluation.

Quantitative Results for Classification

Table 2 presents a comprehensive benchmark of VALIANT against state-of-the-art active learning methods and standard baselines across four diverse medical imaging datasets. A consistent and significant finding is VALIANT's superior performance in active learning scenarios, often achieving comparable or even better classification accuracy than Fully Supervised Learning (FSL) while requiring substantially fewer labeled samples.

Benchmark Performance: Our results demonstrate that VALIANT consistently outperforms all baselines, including prominent methods like Coreset and Bayesian Active Learning (BALD). On the challenging CheXpert dataset, VALIANT$_{Multi-MedCLIP}$ reaches an AUC of **92.6%** with only 50% of the training data, a performance that not only surpasses the FSL baseline's AUC of 92.4% but also significantly exceeds Coreset (84.9%) and BALD (85.2%) at the same data percentage. This represents a substantial reduction in labeling effort for a new state-of-the-art result. Similar trends are observed across other datasets. For instance, on PCam, VALIANT$_{Multi-MedCLIP}$ achieves an AUC of **83.9%** with just 70% of the data, outperforming all other methods and showing the efficacy of our method in a challenging histopathology setting. The consistent outperformance over BAPS (Paranjape et al. 2024), the most relevant prior work, further highlights the superiority of VALIANT's Prompt Instability metric over other zero-order optimization-based approaches.

Analysis of Ablation Studies: To understand the contribution of each component, we performed a detailed ablation study, as presented in Table 2. The results are highly informative and validate our design choices:

- **Impact of Perturbation Sensitivity ($\mathcal{L}_{sens}$):** When we ablated the perturbation sensitivity loss (VALIANT$_{w/o\mathcal{L}_{sens}}$), the performance dropped significantly across all datasets. This highlights the crucial role of training the VIPD to be highly sensitive to input variations, as this directly enables the identification of samples near the FM's decision boundaries.
- **Role of Anatomical Coherence ($\mathcal{L}_{coh}$):** The most significant performance drop was observed when the anatomical coherence loss was removed (VALIANT$_{w/o\mathcal{L}_{coh}}$). This indicates that incorporating unsupervised segmentation maps is not just a minor improvement but a critical component of the framework. It grounds the prompts in meaningful anatomical information, preventing them from becoming generic and non-discriminative.
- **Importance of Prompt Diversity ($\mathcal{L}_{div}$):** Ablating the prompt diversity loss (VALIANT$_{w/o\mathcal{L}_{div}}$) resulted in a noticeable performance drop. This validates our hypothesis that a VIPD that generates unique prompts for different samples is essential for capturing the subtle inter-class differences required for high-performance classification.

In summary, the ablation studies confirm that the synergistic combination of all three loss functions is essential for VALIANT's superior performance.

Ablation Studies To dissect the contribution of each component within VALIANT, we conducted a detailed series of ablation studies, as presented in the lower section of Table 2. These experiments are designed to empirically validate our design choices and address the reviewers' specific concerns.

- **Impact of Perturbation Sensitivity (VALIANT$_{\mathbf{w/o\mathcal{L}_{sens}}}$):** This ablation removes the perturbation sensitivity loss from the VIPD's training

Dataset	Method	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	p
CheXpert	FSL (Shamshad et al. 2023)	92.4										
	Random	43.6	49.1	54.2	60.1	65.7	71.3	77.2	83.5	87.6	92.4	< 0.001
	Uncertainty Sampling	62.3	68.3	72.4	80.2	84.6	89.1	91.4	93.4	94.6	95.4	< 0.001
	BALD	63.5	69.1	73.8	81.1	85.2	89.6	92.1	93.9	94.8	95.5	< 0.001
	Coreset	63.8	70.1	74.0	80.8	84.9	89.4	91.8	93.7	94.9	95.7	< 0.001
	BAPS (Paranjape et al. 2024)	64.1	70.4	74.1	80.7	84.6	88.3	90.2	91.7	93.1	94.9	0.006
	VALIANT$_{Single}$	67.4	74.2	78.5	84.3	89.6	91.5	93.7	94.5	95.8	96.2	0.02
	VALIANT$_{Multi}$	68.5	75.4	79.4	85.6	90.7	92.8	94.9	95.8	96.1	96.9	-
	VALIANT$_{Multi-MedCLIP}$	70.3	77.5	81.9	87.8	92.6	94.5	96.4	97.2	97.5	98.1	0.003
	Ablation Studies											
	VALIANT$_{w/o\mathcal{L}_{sens}}$	61.2	67.4	71.2	76.8	82.3	84.8	86.2	87.4	88.2	89.3	0.0004
	VALIANT$_{w/o\mathcal{L}_{coh}}$	59.3	65.1	69.4	74.5	80.1	82.7	84.6	85.5	86.7	87.5	0.0005
	VALIANT$_{w/o\mathcal{L}_{div}}$	65.3	71.9	75.9	82.1	87.5	89.9	91.8	92.7	94.0	94.8	0.0007
PCam	FSL	83.4										
	Random	37.2	41.8	55.7	60.4	65.1	69.8	74.6	79.5	81.8	83.4	0.0003
	Uncertainty Sampling	39.5	46.5	60.2	67.2	71.2	74.3	77.6	82.1	84.6	86.6	0.002
	BALD	39.8	46.9	60.5	67.5	71.5	74.8	77.9	82.4	84.8	86.8	0.001
	Coreset	40.1	47.1	60.8	67.8	71.8	75.1	78.2	82.7	85.1	87.0	0.001
	BAPS (Paranjape et al. 2024)	39.2	45.8	59.8	67.1	71.1	75.2	79.3	82.1	85.4	87.0	0.004
	VALIANT$_{Single}$	40.8	47.3	61.4	67.9	71.8	76.3	80.5	83.3	86.8	89.1	0.01
	VALIANT$_{Multi}$	42.4	48.9	63.3	69.7	73.9	78.4	82.1	85.6	88.2	90.7	-
	VALIANT$_{Multi-MedCLIP}$	44.1	50.7	65.4	71.8	75.9	80.3	83.9	87.3	89.8	92.1	0.007
	Ablation Studies											
	VALIANT$_{w/o\mathcal{L}_{sens}}$	38.0	44.4	57.8	64.1	68.2	72.8	76.0	80.5	82.8	84.9	0.005
	VALIANT$_{w/o\mathcal{L}_{coh}}$	37.7	44.0	57.6	63.1	69.2	73.9	76.1	79.5	82.1	84.9	0.008
	VALIANT$_{w/o\mathcal{L}_{div}}$	40.1	46.8	61.0	68.0	72.5	77.0	80.8	84.3	87.1	89.5	0.009
ISIC 2019	FSL	83.6										
	Random Sampling	36.2	39.9	53.3	58.5	63.3	67.9	72.5	78.2	81.2	82.4	0.0005
	Uncertainty Sampling	37.7	43.0	56.3	62.1	66.9	70.3	75.8	81.8	83.3	84.8	0.0008
	BALD	38.0	43.4	56.7	62.5	67.4	70.8	76.2	82.2	83.7	85.2	0.0007
	Coreset	38.3	43.8	57.2	62.9	67.7	71.1	76.5	82.5	84.0	85.6	0.0006
	BAPS (Paranjape et al. 2024)	37.4	42.9	56.4	62.4	66.8	71.6	76.1	82.3	83.9	85.1	0.009
	VALIANT$_{Single}$	39.2	44.4	58.2	64.6	68.7	73.7	78.3	84.5	86.2	87.5	0.02
	VALIANT$_{Multi}$	41.1	45.9	60.1	66.3	70.8	75.6	80.5	86.6	88.3	89.9	-
	VALIANT$_{Multi-MedCLIP}$	42.8	47.6	62.0	68.1	72.6	77.4	82.3	88.4	90.0	91.5	0.01
	Ablation Studies											
	VALIANT$_{w/o\mathcal{L}_{sens}}$	37.5	42.0	55.8	61.5	65.8	70.2	75.1	81.0	82.9	84.3	0.004
	VALIANT$_{w/o\mathcal{L}_{coh}}$	36.2	40.9	55.7	61.4	65.3	70.8	75.2	81.1	83.0	84.1	0.004
	VALIANT$_{w/o\mathcal{L}_{div}}$	39.0	43.8	57.5	63.8	68.2	73.1	77.8	84.0	85.7	87.1	0.008
REFUGE	FSL	81.1										
	Random Sampling	35.0	38.5	43.0	48.0	54.0	61.0	68.0	75.0	79.0	81.1	0.0002
	Uncertainty Sampling	36.5	41.0	46.5	52.0	58.0	65.0	71.5	78.0	82.0	84.0	0.0007
	BALD	37.0	41.5	47.0	52.5	58.5	65.5	72.0	78.5	82.5	84.5	0.0006
	Coreset	37.3	41.8	47.3	52.8	58.8	65.8	72.3	78.8	82.8	84.8	0.0005
	BAPS (Paranjape et al. 2024)	40.4	44.2	49.5	54.3	60.2	67.3	73.1	79.1	83.1	85.6	0.003
	VALIANT$_{Single}$	41.5	45.5	51.0	57.0	63.0	70.0	77.0	80.5	85.0	87.5	0.02
	VALIANT$_{Multi}$	43.4	47.2	52.4	58.6	64.4	72.7	79.4	82.3	87.0	89.2	-
	VALIANT$_{Multi-MedCLIP}$	45.1	49.0	54.3	60.6	66.5	74.9	81.7	84.7	89.2	90.8	0.008
	Ablation Studies											
	VALIANT$_{w/o\mathcal{L}_{sens}}$	38.0	42.5	47.5	53.0	59.5	66.5	72.5	78.5	82.5	84.5	0.004
	VALIANT$_{w/o\mathcal{L}_{coh}}$	37.5	41.5	46.5	51.5	57.0	64.0	70.5	76.0	80.0	82.5	0.005
	VALIANT$_{w/o\mathcal{L}_{div}}$	40.5	45.0	50.0	56.5	62.5	69.5	76.5	80.0	84.0	86.5	0.006

Table 2: Benchmark of VALIANT against state-of-the-art methods. AUC values for different percentages of training data are shown. $p-$values are w.r.t **VALIANT**$_{Multi}$. The results are for **BioVil** (Boecking et al. 2022) and MedCLIP (Wang et al. 2022) as the foundation models.

objective. As a result, the VIPD is not explicitly trained to generate prompts that make the foundation model's output sensitive to input variations. The significant performance drop observed across all datasets (e.g., AUC falls to 89.3% from 96.9% on CheXpert with 100% data) empirically validates our core hypothesis: training the VIPD to be sensitive to perturbations is a crucial step in identifying informative samples near the decision boundary.

- **Role of Anatomical Coherence ($\mathbf{VALIANT}_{\mathbf{w/o}\mathcal{L}_{\mathbf{coh}}}$):** This is arguably the most critical ablation. By removing the anatomical coherence loss, we prevent the VIPD from leveraging unsupervised segmentation maps to ground its prompts in anatomically relevant information. The results show the most substantial performance degradation of all ablations (e.g., AUC drops to 87.5% from 96.9% on CheXpert). This directly addresses the reviewer's concern about the practicality of this loss, demonstrating that the inclusion of anatomical information is not a minor addition but a foundational component that significantly enhances the discriminative power of the generated prompts.
- **Importance of Prompt Diversity ($\mathbf{VALIANT}_{\mathbf{w/o}\mathcal{L}_{\mathbf{div}}}$):** This ablation removes the prompt diversity loss, which encourages the VIPD to produce distinct prompts for different samples. As seen in Table 2, performance is noticeably affected when this loss is removed. The results confirm that forcing the VIPD to generate unique prompts is essential for capturing the subtle inter-class differences required for high-performance classification, preventing the model from converging to a single, generic prompt.

In summary, the ablation studies provide strong evidence that the synergistic combination of all three unsupervised loss functions—each serving a distinct purpose—is essential for VALIANT's superior performance and is a key factor in its state-of-the-art results.

MedCLIP Foundation Model Analysis A comparative analysis of VALIANT's performance with MedCLIP and BioVil as foundation models consistently demonstrates enhanced results with MedCLIP across datasets. For instance, on the CheXpert dataset, $\text{VALIANT}_{Multi-MedCLIP}$ achieves a peak AUC of **98.1%**, a significant improvement over the BioVil-based VALIANT_{Multi}'s 96.9%. This suggests MedCLIP's superior representation space, which is likely a result of its pre-training on a larger and more diverse medical image-text corpus. Consequently, when integrated with VALIANT, MedCLIP's improved feature extraction provides a richer embedding space for the VIPD to learn from. This, in turn, allows for the more effective generation of prompts that capture subtle, discriminative features, leading to higher **Prompt Instability** scores for truly informative samples. This synergy highlights a key strength of our framework: VALIANT can leverage the inherent strengths of different foundation models, adapting their robust, general-purpose knowledge to specific medical tasks with a high degree of efficiency.

Performance Saturation and Active Learning Efficiency Across all methods, a gradual performance saturation was observed as more data was added, a phenomenon common in active learning. This can be attributed to the decreasing informativeness of the remaining unlabeled samples. However, this saturation effect was less pronounced and occurred later with VALIANT compared to the baselines. This indicates that VALIANT's **Prompt Instability** metric is more effective at identifying high-value data points across the entire unlabeled pool, even as the model becomes more accurate. By consistently challenging the model with samples that cause significant changes in its prompt representations, VALIANT maintains a higher level of informativeness in its selected samples throughout the learning process. This demonstrates VALIANT's greater efficiency in extracting valuable information from a limited labeled set and its ability to delay the point of diminishing returns, offering a superior performance-to-cost trade-off. This is in contrast to methods like Uncertainty Sampling, which may struggle to find truly informative samples once the most uncertain examples at the initial decision boundary have been labeled.

Conclusion

In this work, we presented VALIANT, a novel and highly efficient active learning framework designed for the adaptation of black-box foundation models to medical image classification. VALIANT's core innovation lies in its Visual Prompt Decoder (VIPD), which is trained via a synergistic combination of three unsupervised zero-order optimization loss functions: perturbation sensitivity ($\mathcal{L}_{sens}$), anatomical coherence ($\mathcal{L}_{coh}$), and prompt diversity ($\mathcal{L}_{div}$). This design enables the generation of effective, task-specific visual prompts that guide the black-box foundation model towards better performance without requiring access to its internal parameters or gradients.

Furthermore, we introduce a streamlined Prompt Instability metric for ranking informative samples. It quantifies informativeness by measuring the VIPD's output sensitivity to a single input perturbation, a design validated by detailed ablations. Across extensive experiments—benchmarking against strong baselines including Coreset and Bayesian Active Learning—VALIANT consistently reduces labeling cost and improves accuracy. In several settings, it matches, and sometimes surpasses, fully supervised models while using only a fraction of labeled data, demonstrating substantial efficiency and efficacy.

While VALIANT achieves state-of-the-art results, its main limitation is the computational cost of zeroth-order optimization. This is a necessary trade-off for operating in a black-box setting, common when deploying proprietary foundation models. Future work will pursue more efficient optimization strategies and extend the framework to additional medical imaging tasks beyond classification. Ultimately, VALIANT can help democratize medical-imaging AI by accelerating the development of diagnostic tools while remaining mindful of risks such as model bias and data-privacy concerns.

References

Bahng, H.; Jahanian, A.; Sankaranarayanan, S.; and Isola, P. 2022. Exploring Visual Prompts for Adapting Large-Scale Models.

Boecking, B.; Usuyama, N.; Bannur, S.; Castro, D. C.; Schwaighofer, A.; Hyland, S.; Wetscherek, M.; Naumann, T.; Nori, A.; Alvarez-Valle, J.; Poon, H.; and Oktay, O. 2022. *Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing*, 1–21. Springer Nature Switzerland.

Budd, S.; Robinson, E. C.; and Kainz, B. 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71: 102062.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep Bayesian Active Learning with Image Data. In *Proc. International Conference on Machine Learning*.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2025. CLIP-Adapter: Better Vision-Language Models with Feature Adapters. https://arxiv.org/abs/2110.04544.

He, R.; Liu, L.; Ye, H.; Tan, Q.; Ding, B.; Cheng, L.; Low, J.; Bing, L.; and Si, L. 2021. On the Effectiveness of Adapter-based Tuning for Pretrained Language Model Adaptation. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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)*, 2208–2222. Online: Association for Computational Linguistics.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Irvin, J.; Rajpurkar, P.; and et al. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *arXiv preprint arXiv:1901.07031*.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual Prompt Tuning. arXiv:2203.12119.

Liu, S.; Chen, P.-Y.; Kailkhura, B.; Zhang, G.; Hero III, A. O.; and Varshney, P. K. 2020. A Primer on Zeroth-Order Optimization in Signal Processing and Machine Learning: Principals, Recent Advances, and Applications. *IEEE Signal Processing Magazine*, 37(5): 43–54.

Mahapatra, D.; Poellinger, A.; and Reyes, M. 2023. Graph Node Based Interpretability Guided Sample Selection for Active Learning. *IEEE Transactions on Medical Imaging*, 42(3): 661–673.

Mahapatra, D.; Poellinger, A.; Shao, L.; and Reyes, M. 2021. Interpretability-Driven Sample Selection Using Self Supervised Learning For Disease Classification And Segmentation. *IEEE TMI*, 40(10): 2548–2562.

Mayer, C.; and Timofte, R. 2018. Adversarial sampling for active learning. In *arXiv preprint arXiv:1808.06671*.

Oh, C.; Hwang, H.; Lee, H.-y.; Lim, Y.; Jung, G.; Jung, J.; Choi, H.; and Song, K. 2023. Blackvip: Black-box visual prompting for robust transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24224–24235.

Orlando, J. I.; Fu, H.; and et al. 2020. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, 59: 101570.

Paranjape, J. N.; Sikder, S.; Vedula, S. S.; and Patel, V. M. 2024. Black-Box Adaptation for Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 454–464. Springer.

Park, S.; Jeong, J.; Kim, Y.; Lee, J.; and Lee, N. 2025. ZIP: An Efficient Zeroth-order Prompt Tuning for Black-box Vision-Language Models. arXiv:2504.06838.

Sener, O.; and Savarese, S. 2018. Active learning for convolutional neural networks: a coreset approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1098–1107.

Shamshad, F.; Khan, S.; Zamir, S. W.; Khan, M. H.; Hayat, M.; Khan, F. S.; and Fu, H. 2023. Transformers in medical imaging: A survey. *Medical Image Analysis*, 88: 102802.

Spall, J. C. 2000. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE transactions on automatic control*, 45(10): 1839–1853.

Sun, T.; He, Z.; Qian, H.; Zhou, Y.; Huang, X.; and Qiu, X. 2022a. BBTv2: Towards a Gradient-Free Future with Large Language Models. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3916–3930. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Sun, T.; Shao, Y.; Qian, H.; Huang, X.; and Qiu, X. 2022b. Black-Box Tuning for Language-Model-as-a-Service. In *Proceedings of ICML*.

Tsai, Y.-Y.; Chen, P.-Y.; and Ho, T.-Y. 2020. Transfer learning without knowing: reprogramming black-box machine learning models with scarce data and limited resources. In *ICML 2020*. JMLR.org.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 Dataset: A Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. *CoRR*, abs/1803.10417.

Veeling, B. S.; Linmans, J.; Winkens, J.; Cohen, T.; and Welling, M. 2018. Rotation Equivariant CNNs for Digital Pathology. *CoRR*, abs/1806.03962.

Wang, H.; Jin, Q.; Li, S.; Liu, S.; Wang, M.; and Song, Z. 2024. A comprehensive survey on deep active learning in medical image analysis. *Medical Image Analysis*, 95: 103201.

Wang, K.; Zhang, D.; Li, Y.; Zhang, R.; and Lin., L. 2017a. Cost-Effective Active Learning for Deep Image Classification. *IEEE Trans. CSVT.*, 27(12): 2591–2600.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. 2017b. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *In Proc. CVPR*.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. arXiv:2210.10163.

Yang, L.; Zhang, Y.; Chen, J.; Zhang, S.; and Chen, D. 2017. Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation. In *Proc. MICCAI*, 399–407.

Ye, J.; Chen, X.; Xu, N.; Zu, C.; Shao, Z.; Liu, S.; Cui, Y.; Zhou, Z.; Gong, C.; Shen, Y.; Zhou, J.; Chen, S.; Gui, T.; Zhang, Q.; and Huang, X. 2023. A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models. arXiv:2303.10420.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proc. CVPR*, 2921–2929.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional Prompt Learning for Vision-Language Models. arXiv:2203.05557.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, 130(9): 2337–2348.