Prompt and Probe: Data-Efficient Adaptation of Black-Box Foundation Models via Unified Active Visual Prompting

Dwarikanath Mahapatra[1,*], Abhijit Das[2,*], Sudipta Roy[3,11], Manish Pandey[4], Joy Dhar[5], Nayyar Zaidi[6], Zongyuan Ge[7], Behzad Bozorgtabar[8], Mauricio Reyes[9], Imran Razzak[2,10]

[1] Khalifa University, Abu Dhabi, UAE. [2] MBZUAI, Abu Dhabi, UAE. [3] Jio University, India, [4] Roentgen Health, India. [5] IIT Ropar, India. [6] Deakin University, Australia. [7] Monash University, Australia. [8] Aarhus University, Denmark. [9] University of Bern, Switzerland. [10] MedOS, Abu Dhabi, UAE. [11] UNSW Bengaluru, India.

Abstract. Deploying foundation models (FMs) in clinical imaging is limited by expert annotation cost and the opacity of API-served models that preclude gradient-based adaptation. Existing methods address these constraints separately: black-box prompt tuning assumes fixed labels, while active learning often requires gradients, internal features, logits, or calibrated probabilities. We present Prompt and Probe (P^2), a query-only framework for adapting frozen FMs to classification and segmentation while selecting informative unlabeled samples. A lightweight Visual Prompt Decoder (VIPD) maps shared semantic and anatomical embeddings to input-space prompts: global additive prompts for classification and spatial channel-concatenated prompts for segmentation, trained via zeroth-order optimization. For acquisition, the trained decoder probes unlabeled images with multi-scale perturbations; prompt and prediction shifts are summarized as image- and pixel-level instability scores and fused after per-batch normalization. Across seven public datasets, eight benchmarks, and six clinical modalities, P^2 reduces labeling by 40–60% while outperforming black-box active learning baselines, achieving 91.8% AUROC on CheXpert with 50% labels and 58.5% Dice on Kvasir-Seg with 10% labels. Boundary instability from model-predicted masks, without segmentation ground truth, improves classification acquisition by up to 24.5 AUROC points, suggesting that joint spatial–semantic probing identifies anatomically ambiguous samples missed by task-specific selection.

Keywords: Active Learning · Black-Box Adaptation · Visual Prompting · Foundation Models · Data-Efficient Learning

[*]Equal contribution. Contact: *dwarikanath.mahapatra@ku.ac.ae*

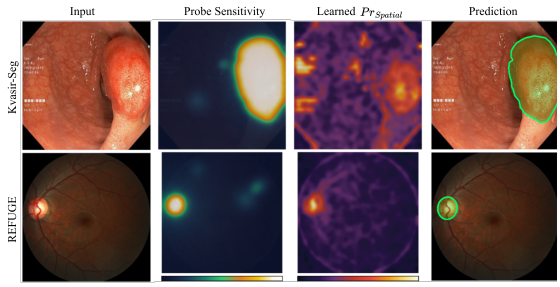

Fig. 1: **VIPD interpretability.** Left to right: input, Probe sensitivity (boundary-peaked instability), learned Pr_{Spatial} (broad target activation), predicted mask. Prompt and Probe capture complementary signals: target attention vs. boundary uncertainty.

1 Introduction

Foundation models (FMs) are reshaping clinical imaging [17], yet a widening gap separates research-grade performance from real-world deployment. On one side, pixel-level annotation remains a bottleneck—a single cardiac MRI study can demand over thirty minutes of specialist time—making large labeled pools impractical for most institutions. On the other, an increasing share of high-performing FMs are served as proprietary APIs whose weights, gradients, and internal representations are inaccessible by design [12], invalidating fine-tuning, parameter-efficient adaptation [6,5], and most active learning (AL) strategies that presuppose gradient access or calibrated predictive distributions [1,9,21]. These two constraints compound: even when an institution can afford annotation, the model it targets may refuse to reveal the signals needed to decide *what* to annotate next.

Existing work addresses fragments of this compound problem but never the whole. VALIANT [11] introduces black-box AL via zeroth-order prompt learning, yet supports only classification. BAPS [15] adapts black-box segmentation models but provides no active acquisition mechanism; annotation is assumed complete. ZIP [16] achieves query-efficient zeroth-order prompt tuning for vision–language models but operates on a fixed labeled set, leaving the data-selection problem untouched. BlackVIP [13] pioneers input-dependent visual prompting with SPSA-based optimization for robust transfer learning, yet neither incorporates active sample selection nor targets medical imaging. Entropy-based AL [21] requires the calibrated softmax distributions that API-served models do not expose. Each method satisfies at most one of three desiderata—black-box compatibility, annotation efficiency, and cross-task unification—and none jointly addresses classification and segmentation within a single acquisition loop.

We propose **Prompt and Probe ($\mathbf{P^2}$)**, a framework that closes this gap by adapting frozen, API-served FMs to both classification and segmentation using only input–output queries, while simultaneously learning *which* unlabeled sam-

ples to annotate. The framework operates in two stages that mirror its name. In the *Prompt* stage, a lightweight Visual Prompt Decoder (VIPD) generates task-aware prompts—global additive perturbations for classification, spatial channel-concatenated maps for segmentation—from shared image and anatomical embeddings, optimized via SPSA using only FM predictions. In the *Probe* stage, the trained VIPD is repurposed as an uncertainty instrument: multi-scale perturbations applied to unlabeled samples induce prediction shifts whose magnitude reveals decision-boundary proximity, and a rank-based fusion criterion unifies sample acquisition across tasks. Neither stage accesses model weights, gradients, or logits. P^2 contributes along three axes:

- **Unified global–spatial prompting.** A single VIPD produces both prompt types from a shared backbone, enforcing cross-task feature reuse while keeping the parameter count low enough for zeroth-order training where each parameter incurs query cost.
- **Multi-scale instability-driven acquisition.** Prediction instability is quantified at both image level (prompt-shift cosine dissimilarity and finite-difference gradient proxies) and pixel level (boundary disagreement, volumetric change, pixel entropy). A task-emphasis parameter γ fuses both scores into a single ranking, enabling one acquisition policy to serve classification, segmentation, or joint settings.
- **Dynamically weighted cross-task objective.** A unified loss jointly regularizes entropy, prompt sensitivity, and feature consistency with automatically balanced weights, enabling stable optimization across heterogeneous tasks without task-specific modules or manual scheduling.

Key finding. Ablations reveal that spatial instability signals designed for segmentation substantially improve *classification* sample selection—removing boundary disagreement on CheXpert drops AUROC by 24.5 points, even absent any segmentation labels. This demonstrates that unified acquisition surfaces anatomically ambiguous samples beneficial to both tasks, making task-specific active learning strictly sub-optimal in the medical imaging setting P^2 targets.

2 Methodology

P^2 operates in two stages (Fig. 2). The *Prompt* stage (§2.1) trains a lightweight decoder to generate input-space prompts that steer a frozen FM toward classification and segmentation simultaneously. The *Probe* stage (§2.2) then repurposes this decoder to rank unlabeled samples by prediction instability under controlled perturbations, selecting the most informative ones for annotation.

2.1 Prompt: Visual Prompt Decoder and Optimization

The VIPD takes an image X and produces two types of prompts—global and spatial—through a shared backbone that splits into task-specific heads. This design requires two ingredients: complementary embeddings that capture both semantics and structure, and a prompt-generation pathway light enough for gradient-free optimization.

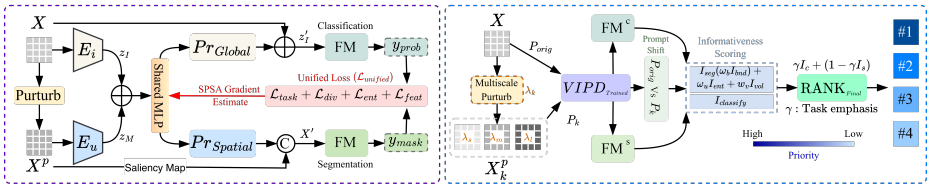

Fig. 2: **P^2 framework. (a)** *Prompt*: the VIPD fuses image (z_I) and anatomical (z_M) embeddings via shared MLPs into a global prompt (additive, for classification) and spatial prompt (channel-concat, for segmentation), trained by SPSA with the frozen FM. **(b)** *Probe*: multi-scale perturbations yield prompt pairs from the trained VIPD; prompt shift and output instability are scored into I_C and I_S, fused as $\gamma I_C + (1-\gamma) I_S$ to rank samples for annotation.

Embedding extraction. We extract two complementary representations (Fig. 2a). A frozen MAE-pretrained ViT-B/16 produces an *image embedding* $z_I \in \mathbb{R}^{768}$ encoding high-level semantics. In parallel, a saliency-based anatomical mask—obtained via Otsu thresholding and connected-component filtering—is processed by a frozen Uformer encoder to yield an *anatomical embedding* $z_M \in \mathbb{R}^{512}$, capturing structural priors such as organ boundaries and tissue layout. The Uformer's hierarchical window attention preserves multi-scale boundary cues that flat ViT features discard, making the two embeddings complementary by construction.

Prompt generation and integration. The concatenation $[z_I, z_M]$ passes through two shared MLP layers (dim 1024, GELU) before splitting into: a *global prompt* $Pr_{\text{Global}} \in \mathbb{R}^{512}$, and a *spatial prompt* $Pr_{\text{Spatial}} \in \mathbb{R}^{H \times W}$ from a transposed-convolution head—totaling ∼2.1M parameters. For classification, the global prompt additively modulates the image embedding ($z'_I = z_I + Pr_{\text{Global}}$) and the FM returns class probabilities $y_{\text{prob}} = \text{FM}(z'_I)$. For segmentation, the spatial prompt is concatenated channel-wise ($X' = [X \,\|\, Pr_{\text{Spatial}}]$) and the FM returns a mask $y_{\text{mask}} = \text{FM}(X')$. The FM remains completely frozen throughout—no weights, gradients, or activations are accessed.

Zeroth-order optimization. Since gradients are inaccessible, we optimize VIPD parameters θ via SPSA [18]. A Rademacher perturbation Δ and scale c yield the gradient estimate

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

updated with momentum: $m_{t+1}{=}\beta m_t {-} \alpha \hat{g}, \quad \theta_{t+1}{=}\theta_t {+} m_{t+1}$. Each step requires exactly two FM queries, keeping optimization practical under API rate limits.

Unified training objective. A single loss spans both tasks and three unsupervised regularizers:

$$\mathcal{L}_{\text{unified}} = \underbrace{\mathcal{L}_{\text{task}}}_{\text{supervised}} + \underbrace{\mathcal{L}_{\text{div}} + \mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{self}}}_{\text{unsupervised}}. \tag{2}$$

The *task loss* $\mathcal{L}_{\text{task}}=\alpha_1\mathcal{L}_{\text{cls}}+(1-\alpha_1)\mathcal{L}_{\text{seg}}$ balances classification consistency $\mathcal{L}_{\text{cls}}$ which is $\text{CE}(y_{\text{prob}}, y^p_{\text{prob}})$ under prompt perturbation with segmentation quality $\mathcal{L}_{\text{seg}}=\mathcal{L}_{\text{Dice}}+\mathcal{L}_{\text{CE-seg}}$. The three regularizers each address a distinct failure mode. *Diversity*: $\mathcal{L}_{\text{div}}=-D_{\text{KL}}(y_{\text{prob}}\|y^p_{\text{prob}})$ prevents prompt collapse by encouraging output divergence under perturbation. *Entropy*: $\mathcal{L}_{\text{ent}}=(1-S)\mathcal{H}(y_{\text{prob}}) + S\mathcal{H}(y^p_{\text{prob}})$, where $S=\cos(Pr_{\text{orig}}, Pr_{\text{pert}})$; when prompts are similar ($S\to1$) the perturbed output should be confident and its entropy is penalized, while divergent prompts ($S\to0$) shift the penalty to the original output—maintaining calibrated confidence proportional to prompt stability. *Decoder self-consistency.* To preserve the black-box setting, no FM activations are used. Let $h_\theta(\cdot)$ denote the shared 1024-d activation of the VIPD MLP trunk before the prompt heads. We define $L_{\text{self}} = \|h_\theta(X+P)-h_\theta(X)\|_2^2$, which regularizes the decoder response to small prompt perturbations without accessing FM weights, gradients, logits, or intermediate features.

2.2 Probe: Perturbation-Based Informativeness Ranking

The Prompt stage produces a decoder that is sensitive to input characteristics—precisely the property needed for sample selection. If a small perturbation to an unlabeled image causes the VIPD's prompt to shift and the FM's prediction to change substantially, the sample likely lies near a decision boundary or contains ambiguous anatomy. The Probe stage operationalizes this insight at multiple scales and across both tasks.

Multi-scale perturbation. For each unlabeled image X, we generate perturbed variants $X^p_k=X+\lambda_k\delta$ with $\delta\sim\mathcal{N}(0,\sigma^2 I)$ and base scales $\lambda_k\in\{0.01, 0.05, 0.1\}$ (fine, medium, coarse). Dataset-specific multipliers $\sigma_s, \sigma_m, \sigma_l$ (Table 1) modulate σ per scale. The VIPD produces prompt pairs (P_{orig}, P_k); the FM generates corresponding predictions. Small λ probes local decision-boundary proximity, while large λ reveals broader distributional fragility.

Classification informativeness (I_{Cls}). We quantify semantic instability via prompt-space and output-space signals, aggregated with scale-dependent weights $w_s>w_m>w_l$:

$$I_{\text{Cls}} = \frac{\sum_k w_k\,(1-\cos(P_{\text{orig}}, P_k))}{\sum_k w_k} + \underbrace{\sum_k w_k \left\| \frac{F(P_k)-F(P_{\text{orig}})}{\lambda_k} \right\|}_{I_{\text{grad}}\text{: finite-difference sensitivity}} . \tag{3}$$

Segmentation informativeness (I_{Seg}). Pixel-level instability complements the image-level score above through three measures between original and perturbed outputs:

$$I_{\text{Seg}} = \omega_u\, I_{\text{unc}} + \omega_b\, I_{\text{bnd}} + \omega_v\, I_{\text{vol}}, \tag{4}$$

where $I_{\text{unc}}=\frac{1}{HW}\sum_{i,j}\mathcal{H}(\bar{y}_{i,j})$ is mean pixel-wise entropy over scale-averaged predictions; $I_{\text{bnd}}=1-\text{Dice}(\partial M_{\text{orig}}, \partial M_k)$ is the Dice disagreement between boundary maps extracted via morphological erosion (3×3 kernel)—particularly informative for irregular structures like polyps or optic cups; $I_{\text{vol}}=|\sum M_{\text{orig}}-\sum M_k|/\sum M_{\text{orig}}$ captures relative volumetric change. We fix $\omega_u=0.3$, $\omega_b=0.5$, $\omega_v=0.2$ across all

datasets, weighting boundary disagreement highest given its dominant ablation impact (Section 4.1). For classification-only datasets such as CheXpert, I_{bnd} uses no segmentation ground truth: MedSAM predicts masks for the original and perturbed image, and boundary instability is computed only between these predicted masks.

Unified acquisition. Before fusion, each score is min–max normalized within the current unlabeled batch B: $\tilde{I} = (I - \min_B I)/(\max_B I - \min_B I + \epsilon)$. The two informativeness scores are fused into a single sample ranking:

$$\text{Rank}_{\text{Final}} = \gamma \tilde{I}_{\text{Cls}} + (1 - \gamma)\tilde{I}_{\text{Seg}}, \tag{5}$$

where $\gamma \in [0, 1]$ controls task emphasis (Table 1), and samples are ranked by sorting R. Top-ranked samples are selected for expert annotation, the VIPD is retrained on the expanded labeled set, and the cycle repeats until the budget is exhausted.

3 Experiments

Datasets. We evaluate P^2 on seven public datasets spanning six modalities and three task settings (C, S, joint C+S), summarized in Table 1. ISIC provides both 2019 classification labels (9-class) and 2018 segmentation annotations, yielding eight benchmark evaluations in total.

Protocol. All images are resized to 512×512, intensity-normalized, and augmented with standard geometric and photometric transforms. To simulate API-constrained deployment, we use open-source FMs with all internal access disabled: BioViL [4] and MedCLIP [22] for classification, MedSAM [10] for segmentation (v1 for reproducibility; the framework is architecture-agnostic and compatible with SAM2 backends). Baselines are restricted to methods operable under strict black-box access—no gradients, intermediate features, or logits: Random Sampling, Entropy Sampling [21], BAPS [15], and ZIP [16], all re-implemented under identical protocols. Gradient-dependent AL (CoreSet, BADGE, Learning Loss) and white-box adaptation (VPT, LoRA) are excluded as they require internal model access unavailable through APIs; BlackVIP [13], while black-box compatible, lacks acquisition—pairing it with random sampling evaluates only prompting quality, already subsumed by our w/o Probe ablation (Fig. **??**). FSL on 100% labels with DenseNet-121 serves as the performance ceiling. We report three variants: $\mathbf{P^2_{Single}}$ (per-modality VIPD), $\mathbf{P^2_{Multi}}$ (shared VIPD across modalities), and $\mathbf{P^2_{Multi\text{-}ZIP}}$ (ZIP prompt initialization).

Training. We use Adam (η=5×10^{-4}, weight decay 10^{-4}, cosine schedule) on an NVIDIA V100 for 100 epochs (single-modality) or 150 (multi-modality). The baseline classifier is ViT-B/16 initialized with a 3% stratified seed. All results report mean ± std over 10 runs with fixed seeds; classification is measured by AUROC, segmentation by Dice.

Table 1: **Dataset summary and per-dataset hyperparameters.** #L = number of classes/labels. Task: C = Classification, S = Segmentation.

					Loss weights		Perturbation scales			Ranking weights		
Modality	**Dataset**	**#Images**	**#L**	**Task**	α_1	γ	λ_s	λ_m	λ_l	w_s	w_m	w_l
Chest X-ray	CheXpert [7]	224,316	14	C	0.4	0.7	1.2	1.4	0.95	4.0	3.0	2.0
Histopathology	PCam [20]	327,680	2	C	0.4	0.7	1.6	1.4	1.0	5.0	3.0	3.0
Dermoscopy	ISIC 2019 [19]	25,331	9	C+S	0.5	0.5	1.5	0.9	1.3	4.5	3.2	2.2
Retinal Fundus	REFUGE [14]	1,200	2	C+S	0.5	0.5	1.5	1.3	1.5	4.0	3.0	2.0
GI Endoscopy	Kvasir-Seg [8]	1,000	2	S	0.5	0.3	1.4	1.5	1.4	4.6	3.1	2.2
Surgical	EndoVis 17 [2]	6,475	2	S	0.5	0.3	1.4	1.4	1.2	4.2	3.3	2.0
Cardiac MRI	ACDC [3]	1,000	5	C+S	0.5	0.5	1.5	1.4	1.4	4.1	3.2	2.1

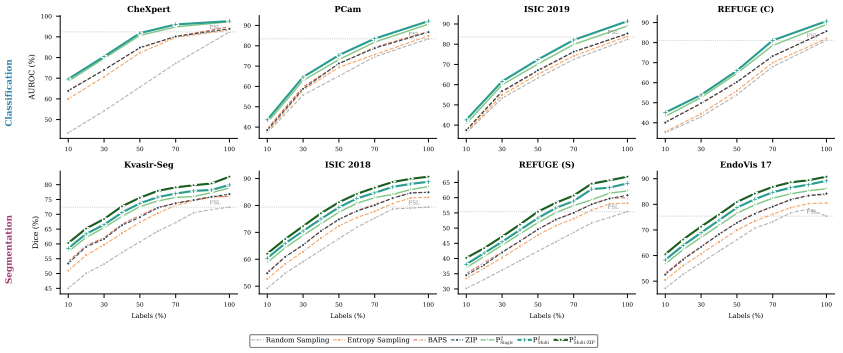

Fig. 3: **Label efficiency across tasks.** AUROC (top: classification) and Dice (bottom: segmentation) vs. labeling budget. P^2 variants (solid) consistently outperform baselines (dashed) across all eight benchmarks, with the largest margins at low budgets. Dotted line: FSL ceiling. All $p < 0.01$.

4 Results

Classification. Figure 3 reports AUROC at labeling budgets from 10% to 100%. P^2_{Multi} consistently leads all baselines, with the largest margins at low budgets. On CheXpert it reaches 91.8% at 50% labels—within 0.6 pts of FSL (92.4%)—and 97.6% at 100%, exceeding FSL by 5.2 pts owing to the regularizing effect of the unified prompt-perturbation objective. On PCam, P^2_{Multi} surpasses BAPS by 4.4 and 5.1 pts at 10% and 100% respectively. The multi-modality variant exceeds P^2_{Single} by 1–2 pts throughout, confirming cross-modality prompt learning provides transferable supervision ($p < 0.01$; paired t-test, 10 runs).

Segmentation. Figure 3 (bottom) reports Dice on four benchmarks using MedSAM as the frozen backbone. P^2_{Multi} achieves 58.5% on Kvasir-Seg with only 10% of labels—a 13.5-point gain over Random and 4.4 above BAPS. On ISIC 2018 it reaches 79.3% at 50% labels, matching FSL trained on the full dataset. $P^2_{\text{Multi-ZIP}}$ adds a further 2–3 pt boost, indicating that better initialization and perturbation-based acquisition are complementary ($p < 0.01$).

Joint task evaluation. On ACDC [3]—which pairs segmentation masks (LV, RV, myocardium) with pathology labels within the same study—P^2 achieves .887 AUROC and .865 Dice under our selected weighting (Table 2), outperforming all

Table 2: **Loss weight ablation (α_1, γ) on P^2.**

Dataset	Task	Metric	*Segmentation* α_1=.1, γ=.9	*Classification* α_1=.7, γ=.4	*Equal* α_1=.5, γ=.5	Ours
ISIC	C+S	AUC	.812	.902	.851	**.865**
		Dice	.891	.785	.811	**.844**
CheX	C	AUC	.887	.911	.901	**.905**
REF	S	Dice	.912	.813	.865	**.880**
ACDC	C+S	AUC	.854	.893	.882	**.887**
		Dice	.890	.801	.875	**.865**

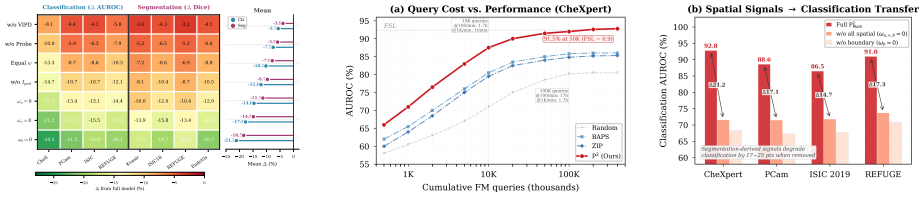

Fig. 4: **Ablation and deployment analysis. (Left)** Component ablation: Δ from full P^2_{Multi} for AUROC and Dice; lollipops show cross-dataset mean (p<0.001). **(Right-a)** AUROC on CheXpert (50% budget) vs. cumulative FM queries—P^2 reaches within 0.9 pts of FSL at ∼50K queries (<1 h at 1K queries/min). **(Right-b)** Ablating spatial signals ($\omega_{u,v,b}$=0) degrades *classification* AUROC by 14–21 pts on classification-only benchmarks.

alternative (α_1, γ) configurations and confirming that a single VIPD can drive both tasks without task-specific modules.

4.1 Ablations

Component ablation. Figure 4 (left) isolates each P^2 component by removal from P^2_{Multi}. Removing Probe (Δ_{cls}=−7.5, Δ_{seg}=−5.9) hurts more than removing the VIPD (−5.7, −3.9), confirming that *what* you label outweighs *how* you prompt. Among informativeness components, boundary disagreement is dominant: ω_b=0 alone causes the largest drop (−21.2 AUROC, −18.5 Dice), and ablating all spatial signals ($\omega_{u,v,b}$=0) degrades classification AUROC by 14–21 pts even on benchmarks without segmentation labels (Fig. 4, right-b). Equalizing scale weights (Δ_{cls}=−10.2) confirms fine-scale perturbations carry stronger boundary-proximity signal. All p<0.001.

Loss weight sensitivity. Table 2 varies α_1 (task balance) and γ (acquisition emphasis). Extreme configurations yield expected trade-offs: seg-priority (α_1=0.1, γ=0.9) maximizes Dice at the expense of AUROC, and cls-priority inverts this. Our dataset-specific choices consistently achieve the best joint performance, confirming that task weighting is a necessary design choice. Perturbation magnitudes and unsupervised loss weights were robust across datasets.

5 Discussion and Conclusion

The strongest ablation occurs on CheXpert: removing boundary disagreement reduces AUROC by 24.5 points. This cue uses no segmentation ground truth; for each unlabeled CXR, MedSAM predicts masks for the original and perturbed images, and I_{bnd} measures instability between these predicted boundaries. The effect is clinically plausible because cardiomegaly, pleural effusion, and atelectasis can alter or obscure anatomical contours. In our sample-level analysis, 62% of images uniquely selected by I_{bnd} contained one of these contour-related findings, compared with a 31% base rate. Thus, spatial probing helps identify anatomically ambiguous cases that image-level uncertainty may miss, without implying universal superiority of task-unified acquisition.

Overall, P^2 provides an annotation-efficient route for adapting black-box medical FMs: VIPD converges within 0.9 pts of FSL at $\sim$50K queries ($<$1 h at 1K queries/min), while $\text{P}^2_{\text{Multi}}$ amortizes prompt learning across modalities. Limitations include perturbation sensitivity, restriction to 2D data, and lack of validation against commercial APIs under real latency constraints.

Acknowledgments. This research was funded by Khalifa University of Science and Technology through the Faculty Startup grant under Project ID: KU-INT-FSU-2026-8471000024.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion **76**, 243–297 (dec 2021). https://doi.org/10.1016/j.inffus.2021.05.008
2. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., Herrera, L., Li, W., Iglovikov, V., Luo, H., Yang, J., Stoyanov, D., Maier-Hein, L., Speidel, S., Azizian, M.: 2017 robotic instrument segmentation challenge (2019)
3. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Transactions on Medical Imaging **37**(11), 2514–2525 (2018)
4. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., Poon, H., Oktay, O.: Making the most of text semantics to improve biomedical vision–language processing. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI. p. 1–21 (2022). https://doi.org/10.1007/978-3-031-20059-5_1

5. He, R., Liu, L., Ye, H., Tan, Q., Ding, B., Cheng, L., Low, J., Bing, L., Si, L.: On the effectiveness of adapter-based tuning for pretrained language model adaptation. In: Zong, C., Xia, F., Li, W., 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). pp. 2208–2222. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.172, https://aclanthology.org/2021.acl-long.172/
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685
7. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., others.: Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. pp. 590–597 (2019)
8. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset (2019), https://arxiv.org/abs/1911.07069
9. Jungo, A., Balsiger, F., Reyes, M.: Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Frontiers in neuroscience **14**, 282 (2020)
10. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1) (Jan 2024). https://doi.org/10.1038/s41467-024-44824-z, http://dx.doi.org/10.1038/s41467-024-44824-z
11. Mahapatra, D., Bozorgtabar, B., Roy, S., Razzak, I., Reyes, M.: VALIANT: Prompt instability for active learning in black-box medical imaging. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7901–7909 (2026)
12. Mahapatra, D., Poellinger, A., Shao, L., Reyes, M.: Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. IEEE TMI **40**(10), 2548–2562 (2021)
13. Oh, C., Hwang, H., Lee, H.y., Lim, Y., Jung, G., Jung, J., Choi, H., Song, K.: Blackvip: Black-box visual prompting for robust transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24224–24235 (2023)
14. Orlando, J.I., Fu, H., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical Image Analysis **59**, 101570 (Jan 2020). https://doi.org/10.1016/j.media.2019.101570, http://dx.doi.org/10.1016/j.media.2019.101570
15. Paranjape, J.N., Sikder, S., Vedula, S.S., Patel, V.M.: Black-box adaptation for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 454–464. Springer (2024)
16. Park, S., Jeong, J., Kim, Y., Lee, J., Lee, N.: ZIP: An efficient zeroth-order prompt tuning for black-box vision-language models. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=2OegVbwvY2
17. Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H.: Transformers in medical imaging: A survey. Medical Image Analysis **88**, 102802 (2023). https://doi.org/https://doi.org/10.1016/j.media.2023.102802, https://www.sciencedirect.com/science/article/pii/S1361841523000634

18. Spall, J.C.: Introduction to stochastic search and optimization: estimation, simulation, and control. John Wiley & Sons (2005)
19. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. CoRR **abs/1803.10417** (2018), http://arxiv.org/abs/1803.10417
20. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. CoRR **abs/1806.03962** (2018), http://arxiv.org/abs/1806.03962
21. Wang, H., Jin, Q., Li, S., Liu, S., Wang, M., Song, Z.: A comprehensive survey on deep active learning in medical image analysis. Medical Image Analysis **95**, 103201 (2024). https://doi.org/https://doi.org/10.1016/j.media.2024.103201, https://www.sciencedirect.com/science/article/pii/S1361841524001269
22. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text (2022), https://arxiv.org/abs/2210.10163