Biomedical Acquisition-induced Style Shifts as Mixture Shifts: Style-aware Mixture-of-Experts Multimodal Prompt Learning

Pingyi Miao[1] **Xianlai Chen**[1] **Ying An**[1]
Yunbo Wang[1*] **Linan Ren**[1] **Yuxin Peng**[2]
[1]Central South University [2]Peking University
{miaopingyi, chenxianlai, anying, wangyunbo, renlinan}@csu.edu.cn
pengyuxin@pku.edu.cn

Abstract

Vision-language models (VLMs), such as CLIP, have shown remarkable transferability in the biomedical domain, and prompt tuning enables efficient adaptation with limited supervision. In practice, biomedical images exhibit pronounced acquisition-induced style shifts, especially across imaging sites and protocols, while cross-modality transfer further introduces compound shifts from different physical acquisition processes. However, existing prompt-tuning methods ignore acquisition-induced style shifts, making the prompts sensitive to source-specific style and limiting the transferability to unseen medical domains. To address this, we formulate acquisition-induced style shifts as mixture shifts, where each domain is viewed as a mixture of latent style components. Under this formulation, the target risk depends on both mixture weights and latent component risks, motivating prompt adaptation that accounts for style-specific variation rather than relying on a single source-specific prompt. In this paper, we propose SaMoE, a **S**tyle-**a**ware **M**ixture-**o**f-**E**xperts Multimodal Prompt Learning framework for adapting VLMs to biomedical domains under acquisition-induced style shifts. To account for latent component risks in prompt adaptation, SaMoE represents multimodal prompts with latent style-conditioned prompt transformations and composes them through image-conditioned routing, enabling each image to induce style-adapted visual prompts. By extracting style cues from intermediate visual representations, the sparse expert composition adapts style-related prompt weights without domain labels and is injected across multiple VLM layers to align with hierarchical visual-language representations. Extensive experiments on 15 medical datasets across 12 modalities and 10 organs demonstrate significant improvements in both accuracy and generalizability over state-of-the-art methods. Our code is available at Code.

1 Introduction

Vision-language models like CLIP [31] demonstrate powerful open-domain representation and zero-shot learning abilities via pre-training on large-scale image-text pairs. Nevertheless, their substantial scale poses challenges for fine-tuning on downstream tasks while preserving strong generalization capabilities. Therefore, a series of methods have been proposed to fine-tune these foundational models by adding and adjusting a set of parameters [50, 49] or adapters [8, 47] while keeping pre-trained weights frozen. Among them, prompt-based approaches try to adapt VLMs to downstream tasks with fewer parameters and supervision by introducing a set of learnable weights in the image branch [12], text branch [50, 43, 42], or both

*Corresponding author: wangyunbo@csu.edu.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

of them [17, 3, 18]. Prompt-learning based approaches show strong promise in the biomedical field, where data scarcity, annotation difficulties, and subtle visual features are prevalent. However, biomedical VLM adaptation must generalize beyond the limited source supervision, as shown in Fig. 1. In base-to-novel transfer, prompts learned from base classes should remain transferable to unseen disease classes rather than overfitting source-class appearances. Beyond class transfer, variations in imaging sites and protocols can substantially alter image style through changes in contrast, noise, resolution, and scanner-dependent artifacts [21, 48]. Cross-modality transfer is even more challenging because different physical acquisition processes may additionally change anatomical visibility, structural patterns, and label spaces. Although prompt learning techniques have achieved remarkable success on natural images and benchmarks, most existing methods learn prompts from the source data without explicitly accounting for acquisition-induced style shifts. As a result, the learned prompts can become sensitive to source-specific styles, which limits their transferability to unseen biomedical domains.

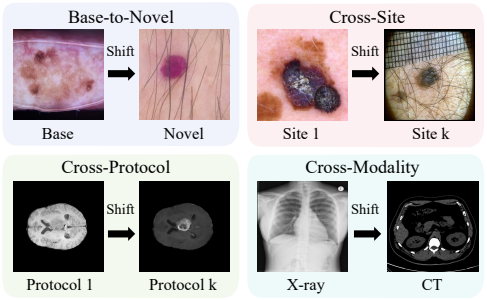


Figure 1: Biomedical VLM generalization settings considered in this work. Base-to-novel transfer evaluates generalization from base classes to unseen classes, while site and protocol changes mainly induce style shifts and modality changes may introduce stronger physical and structural differences.

Motivated by the observation that acquisition is a major source of distribution shifts in biomedical images [45], we formulate biomedical acquisition-induced style shifts as mixture shifts. Specifically, we focus on the style component of acquisition-induced variation: each domain is viewed as a mixture of latent style components, and cross-domain style variation is characterized by changes in the corresponding mixture weights. Under this formulation, the target risk depends on both mixture weights and latent component risks, which explains why a prompt optimized for the average source distribution may transfer poorly when unseen domains reweight latent acquisition styles. This perspective motivates prompt adaptation that accounts for style-specific variation rather than relying on a single source-specific prompt. To account for latent component risks in prompt adaptation, we propose SaMoE, a Style-aware Mixture-of-Experts Multimodal Prompt Learning framework for adapting VLMs to biomedical domains. SaMoE represents multimodal prompts through a style-aware prompt expert bank, where each expert acts as a latent style-conditioned prompt transformation to enable style-specific adaptation under acquisition shifts. The expert bank is coordinated by an image-conditioned router, forming a sparse prompt composition mechanism that estimates style-related prompt weights from style cues in intermediate visual representations without using domain labels or external domain-specific models. The composed visual prompts adapt the VLM to acquisition-induced style shifts, while the textual prompts provide class-level semantic context. We further inject the composed prompts across multiple layers of the vision and language branches so that style-aware prompt adaptation can interact with hierarchical visual-language representations. Our contributions are summarized as follows:

- We formulate the style component of biomedical acquisition-induced style shifts as mixture shifts, providing a principled perspective for understanding why source-specific prompts can transfer poorly when unseen domains change the mixture proportions of latent styles.
- We propose SaMoE, a Style-aware Mixture-of-Experts Multimodal Prompt Learning framework that coordinates a style-aware prompt expert bank with image-conditioned routing to compose visual prompts for adapting VLMs to the biomedical domain under acquisition-induced style shifts.
- Extensive experiments on 15 medical datasets spanning 12 modalities and 10 organs demonstrate consistent improvements over strong prompt-learning baselines under base-to-novel, cross-site, cross-protocol, and compound cross-modality settings, with analyses supporting the effectiveness of style cues, sparse routing, expert utilization, and prompt composition.

2 Related Work

Prompt Learning in Vision-Language Models. Vision-language models (VLMs) such as CLIP [31] align image and text representations through large-scale pretraining, enabling strong zero-shot

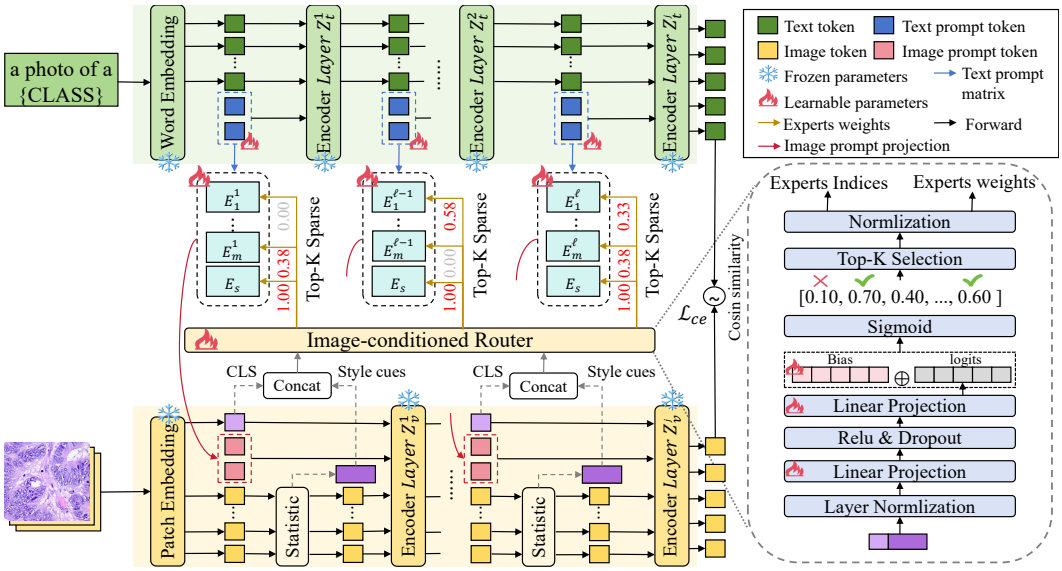


Figure 2: Overview of our proposed SaMoE prompting framework. SaMoE uses a style-aware expert bank to represent latent style-conditioned prompt transformations and performs image-conditioned sparse composition from intermediate visual style cues. The composed visual prompts are injected into corresponding transformer layers to adapt the frozen VLM under acquisition-induced style shifts.

transfer. Related cross-modal representation learning explores modality-agnostic retrieval [38], knowledge-guided hashing [44], and noisy correspondence rectification [39]. Prompt learning adapts VLMs efficiently by replacing hand-crafted templates with learnable contexts [50], conditioning prompts on image instances [49], regularizing updates to preserve pretrained knowledge [51, 42], or jointly optimizing visual and textual prompts across layers [17, 18, 3]. Biomedical prompt learning further incorporates medical ontologies and text priors, as in XCoOp [1] and BiomedCoOp [20]. vMFCoOp [33] aligns biomedical prompts through unified hyperspherical anchors, while GLAli [41] combines visually guided text refinement with global/local alignment for few-shot classification and OOD detection. BioDPP [23] learns a dynamic prompt policy with classification and context-consistency rewards for few-shot biomedical image classification. However, existing methods are not explicitly designed for acquisition-induced style shifts across sites and protocols, nor for broader modality-dependent gaps. Our approach addresses this limitation through style-adaptive multimodal prompts for biomedical VLM adaptation.

Mixture-of-Experts Prompting. Mixture-of-Experts (MoE) has been used in prompt learning and parameter-efficient adaptation to increase conditional capacity with sparse activation. MoE-style prompt methods learn multiple prompt experts and route them by tasks or instances [24, 7]; Prompt-MoE [34] composes textual state prompts for zero-shot anomaly detection, and MoECLIP [26] routes image patches to LoRA experts for patch-level anomaly adaptation. These methods demonstrate the utility of conditional routing, but focus on textual semantic composition, multi-backbone adaptation, or anomaly detection rather than biomedical classification under acquisition-induced style shifts. In contrast, we start from a mixture-shift formulation of biomedical acquisition styles and use compact visual style cues to route cross-modal prompt experts without domain labels.

3 Method

In this section, we formulate acquisition-induced style shifts as mixture shifts over latent style components and use the resulting risk perspective to motivate style-aware prompt composition. Fig. 2 illustrates the proposed SaMoE framework, which learns style-aware prompt experts and a shared image-conditioned router to sparsely compose prompts from input style cues without domain labels or external domain-specific models. The routed prompts are injected into multiple layers of both the vision and language branches to strengthen multi-level representations and cross-modal alignment.

3.1 Mixture-Shift Formulation of Acquisition-Induced Style Shifts

We study distribution shifts for prompt-tuned frozen VLMs in biomedical imaging, with a focus on acquisition-induced style shifts. A pretrained VLM (e.g., CLIP) consists of visual encoder $\phi(\cdot)$ and text encoder $\theta(\cdot)$ that embed an image x and a text t into a shared embedding space, respectively. For label space $\mathcal{Y}$, prompt tuning replaces hand-crafted templates with learnable context tokens. Specifically, we parameterize the class prompt as $t_y^\Theta = [v_1, \ldots, v_T, c_y]$, where $\{v_i\}_{i=1}^T$ are learnable context embeddings shared across classes and c_y is the class name token; Θ denotes the collection of all learnable prompting parameters. Prompt-based classification prediction is represented via temperature-scaled cosine similarity: the class score is $s_\Theta(x, y) = \cos\big(f_{\text{img}}^\Theta(x),\, w_y^\Theta(x)\big)/\tau$, where $f_{\text{img}}^\Theta(x)$ denotes the image embedding under prompts, and $w_y^\Theta(x) = \theta(t_y^\Theta)$ is the prompted text embedding. Learning is performed on a labeled source dataset $\mathcal{D}_s = \{(x_n, y_n)\}_{n=1}^N$ by minimizing the cross-entropy loss:

$$\mathcal{L}_{\text{ce}}(\Theta) = -\frac{1}{N}\sum_{n=1}^{N} \log\left(\frac{\exp(s_\Theta(x_n, y_n))}{\sum_{y'\in\mathcal{Y}}\exp(s_\Theta(x_n, y'))}\right). \tag{1}$$

Domain generalization (DG) setting. Training samples are drawn from a labeled source domain $\mathcal{D}_s = \{(x_n, y_n)\}_{n=1}^N$ with label space $\mathcal{Y}_s$. Evaluation is performed on unseen target domains $\mathcal{D}_t$ with label space $\mathcal{Y}_t$. In the open-set DG setting, $\mathcal{D}_t$ follows different image distributions and contains novel semantic classes unseen during training. The parameter Θ is optimized via minimizing the cross-entropy in Eq. (1) using only $\mathcal{D}_s$, while aiming to achieve low target risk defined as the expected loss on the $\mathcal{D}_t$:

$$\mathcal{R}_{\mathcal{D}_t}(\Theta) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}_t}[\ell(s_\Theta(x, y), y)], \tag{2}$$

where $\ell(\cdot, \cdot)$ denotes a bounded loss function. In biomedical imaging, the discrepancy between $\mathcal{D}_s$ and $\mathcal{D}_t$ is often influenced by acquisition-induced style changes; site and protocol shifts primarily affect style, whereas modality shifts may also introduce semantic and structural differences. We next formalize the style component of this structured shift to derive a criterion that guides the design of our prompting.

Latent style mixtures and target risk. To capture the acquisition-induced style component within the DG setting above, we formulate each domain as a mixture over latent style components:

$$\mathcal{D}_s = \sum_{m=1}^{M} \pi_s(m)\, \mathcal{D}_m, \qquad \mathcal{D}_t = \sum_{m=1}^{M} \pi_t(m)\, \mathcal{D}_m, \tag{3}$$

where $\{\mathcal{D}_m\}_{m=1}^M$ are joint distributions over (x, y) induced by latent acquisition styles and shared across domains, and the shift from $\mathcal{D}_s$ to $\mathcal{D}_t$ is represented by the change of mixture weights π_s vs. π_t. This shared-component view is a modeling abstraction for the acquisition-style component of the domain gap: it is most appropriate when target styles are contained in, near, or statistically related to style factors observable from the source data. It does not by itself guarantee performance on completely disjoint target styles whose acquisition statistics leave no learnable cues in the source domain. For predictor h, if $\mathcal{D}_t$ is a mixture distribution in the condition that a sample is generated by first drawing a component index $m \sim \pi_t$ and then drawing $(x, y) \sim \mathcal{D}_m$, the target risk satisfies

$$\mathcal{R}_{\mathcal{D}_t}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\big[\ell(h(x), y)\big] = \sum_{m=1}^{M} \pi_t(m)\mathbb{E}_{(x,y)\sim\mathcal{D}_m}\big[\ell(h(x), y)\big] = \sum_{m=1}^{M} \pi_t(m)\mathcal{R}_{\mathcal{D}_m}(h). \tag{4}$$

where $\pi_t(m) \geq 0$ and $\sum_{m=1}^M \pi_t(m) = 1$. Therefore, $\mathcal{R}_{\mathcal{D}_t}(h)$ can be expressed as a weighted average (convex combination) of the component risks $\{\mathcal{R}_{\mathcal{D}_m}(h)\}_{m=1}^M$ via Eq. (4). Since any weighted average of scalars is bounded above by their maximum, we obtain the following formulation:

$$\mathcal{R}_{\mathcal{D}_t}(h) \leq \max_{1\leq m\leq M} \mathcal{R}_{\mathcal{D}_m}(h). \tag{5}$$

Thus, achieving uniformly low component-wise risk (i.e., $\mathcal{R}_{\mathcal{D}_m}(h) \leq \varepsilon$ for all m) implies low target risk for any mixture weights π_t. More generally, when both source and target admit the mixture form in Eq. (3), the source-target risk gap satisfies:

$$\big|\mathcal{R}_{\mathcal{D}_t}(h) - \mathcal{R}_{\mathcal{D}_s}(h)\big| = \Big|\sum_{m=1}^{M} \big(\pi_t(m) - \pi_s(m)\big)\mathcal{R}_{\mathcal{D}_m}(h)\Big| \leq \|\pi_t - \pi_s\|_1 \cdot \max_{1\leq m\leq M} \mathcal{R}_{\mathcal{D}_m}(h). \tag{6}$$

Eq. (6) upper-bounds the target-source risk gap by the product of the mixture-weight shift $\|\pi_t - \pi_s\|_1$ and the worst-case component risk $\max_{1\le m\le M} \mathcal{R}_{\mathcal{D}_m}(h)$. The former is determined by the unseen target domain and cannot be controlled during training, whereas the latter depends on the learned hypothesis. Therefore, within the mixture-shift abstraction, improving adaptation to changes in mixture proportions amounts to reducing high component risks for acquisition styles that are represented or statistically inferable from the source data. More details in Appendix A. This motivates our mixture-of-experts prompting with image-conditioned routing, which composes prompts adaptively from a style-aware prompt expert bank based on input style cues to improve adaptation under acquisition-induced style shifts, while not claiming a formal guarantee for entirely unseen and unsupported style components.

3.2 Style-aware Prompt Composition under Mixture Shifts

3.2.1 Latent Style-aware Prompt Transformations

We keep the pretrained image encoder $\phi(\cdot)$ and text encoder $\theta(\cdot)$ frozen. Let $V \in \mathbb{R}^{T\times d_t}$ denote the learnable text prompt token matrix, where T is the prompt length and d_t is the text embedding dimension of $\theta(\cdot)$. For each class $y \in \mathcal{Y}$, we form the prompted text $t_y^{\Theta} = [V, c_y]$ and compute the class embedding $w_y(\Theta) = \theta(t_y^{\Theta})$. Given an image x, prediction is made by temperature-scaled cosine similarity between $f_{\text{img}}(x) = \phi(x)$ and $\{w_y(\Theta)\}_{y\in\mathcal{Y}}$ (Sec. 3.1).

To account for acquisition-induced style shifts, we equip prompts with a compact bank of M experts at each injection layer. Let $\mathcal{L}$ denote the set of prompt injection layers. Each expert maps text-side context tokens to visual-side prompt tokens with a width of d_v which are injected into the visual encoder (Sec. 3.3). Concretely, expert m at layer ℓ is formulated as a lightweight linear projection:

$$E_m^{(\ell)}(V^{(\ell)}) = V^{(\ell)} {W_m^{(\ell)}}^{\top}, \quad W_m^{(\ell)} \in \mathbb{R}^{d_v\times d_t}, \ \ m = 1, \dots, M. \tag{7}$$

Additionally, we consider a shared expert $E_0^{(\ell)}(V^{(\ell)}) = V^{(\ell)} {W_0^{(\ell)}}^{\top}$ that is always active to preserve a domain-invariant base behavior. Given an input image x and injection layer ℓ, the visual prompt is a shared component plus a normalized mixture over a sparse subset of experts:

$$\tilde{V}^{(\ell)}(x) = E_0^{(\ell)}(V^{(\ell)}) + \sum_{m\in\mathcal{S}^{(\ell)}(x)} \alpha_m^{(\ell)}(x)\, E_m^{(\ell)}(V^{(\ell)}), \tag{8}$$

where $\mathcal{S}^{(\ell)}(x)$ denotes the selected Top-K experts at layer ℓ and $\alpha_m^{(\ell)}(x) \ge 0$ with $\sum_{m\in\mathcal{S}^{(\ell)}(x)} \alpha_m^{(\ell)}(x) = 1$. Since $\tilde{V}^{(\ell)}(x)$ lies in the span of a small expert bank rather than being freely optimized per sample, this yields input-adaptive prompts while controlling capacity. We instantiate dedicated expert banks for each injection layer, while using a shared router to process layer-specific style cues from the visual encoder's multi-level representation hierarchy.

3.2.2 Image-Conditioned Sparse Routing from Style Statistics

The expert bank in Eq. (8) introduces multiple candidate prompt transformations $\{E_m^{(\ell)}\}_{m=1}^{M}$ at each injection layer ℓ, but the mixture-shift formulation requires selecting experts according to the input style rather than using a fixed mixture for all images. Thus, we introduce a shared image-conditioned router $g(\cdot)$ that is applied to layer-specific style cues from the corresponding intermediate visual tokens, so that the resulting mixture weights $\{\alpha_m^{(\ell)}(x)\}$ can adapt across both inputs and layers without introducing separate router parameters for every layer.

For token features $H^{(\ell)}(x) = [h_{\text{cls}}^{(\ell)}(x); h_1^{(\ell)}(x); \dots; h_P^{(\ell)}(x)] \in \mathbb{R}^{(P+1)\times d_v}$ entering layer ℓ, we summarize patch tokens by first- and second-order statistics:

$$\begin{aligned} \mu^{(\ell)}(x) &= \frac{1}{P}\sum_{p=1}^{P} h_p^{(\ell)}(x), \quad \sigma^{2,(\ell)}(x) = \frac{1}{P}\sum_{p=1}^{P} \big(h_p^{(\ell)}(x) - \mu^{(\ell)}(x)\big)^2, \\ s^{(\ell)}(x) &= \big[\, \mu^{(\ell)}(x) \,;\, \sigma^{2,(\ell)}(x) \,;\, h_{\text{cls}}^{(\ell)}(x) \,\big] \in \mathbb{R}^{3d_v}. \end{aligned} \tag{9}$$

Eq. (9) forms a compact style descriptor that captures acquisition-induced style statistics while the CLS token supplies complementary global context for stable routing.

Sparse prompt routing. A shared image-conditioned router $g(\cdot)$ maps the layer-specific style cues to expert affinity scores:

$$u^{(\ell)}(x) = \mathrm{sigm}\big(g(s^{(\ell)}(x)) + b\big) \in (0,1)^M, \tag{10}$$

where $\mathrm{sigm}(\cdot)$ is element-wise sigmoid, and $g(\cdot)$ and b are learned jointly with Θ. Although shared, the router produces layer-specific scores because $u^{(\ell)}(x)$ is computed from $s^{(\ell)}(x)$. We activate $\mathcal{S}^{(\ell)}(x) = \mathrm{TopK}(u^{(\ell)}(x))$ and normalize the selected scores:

$$\alpha_m^{(\ell)}(x) = \frac{u_m^{(\ell)}(x)}{\sum_{j\in\mathcal{S}^{(\ell)}(x)} u_j^{(\ell)}(x) + \epsilon}, \quad m \in \mathcal{S}^{(\ell)}(x), \tag{11}$$

Eq. (11) ensures $\sum_{m\in\mathcal{S}^{(\ell)}(x)} \alpha_m^{(\ell)}(x) = 1$ and $\alpha_m^{(\ell)}(x) \geq 0$, so the routed part of $\tilde{V}^{(\ell)}(x)$ in Eq. (8) has a consistent scale across samples. This avoids uncontrolled prompt-magnitude variations that could confound cross-domain generalization. Moreover, Top-K routing encourages sparsity, reducing interference among experts and keeping computation bounded by K active experts per image. Sparse routing may collapse to repeatedly selecting the same experts, which would reduce the effective coverage of style components. To encourage diverse selection, we regularize the average normalized routing mass within a minibatch $\{x_b\}_{b=1}^B$:

$$\mathcal{L}_{\mathrm{bal}} = \frac{1}{|\mathcal{L}|} \sum_{\ell\in\mathcal{L}} M \sum_{m=1}^{M} \left(\frac{1}{B} \sum_{b=1}^{B} \alpha_m^{(\ell)}(x_b) - \frac{1}{M} \right)^2, \tag{12}$$

where $\frac{1}{B}\sum_{b=1}^B \alpha_m^{(\ell)}(x_b)$ is the average normalized routing mass assigned to expert m at layer ℓ within a minibatch. Balancing $\alpha^{(\ell)}$ rather than raw scores directly regularizes the weights used in prompt composition and discourages degenerate single-expert solutions.

3.3 Multi-layer Multimodal Prompt Injection

Suppose that the visual encoder and text encoder contain L_v and L_t transformer blocks, respectively. We inject prompts into a set of early layers $\mathcal{L} \subseteq \{1, \ldots, \min(L_v, L_t)\}$. The overall prompt parameters are $\Theta = \{g, b\} \cup \{V^{(j)}, \{W_m^{(j)}\}_{m=0}^M\}_{j\in\mathcal{L}}$, where g and b define the shared router, and $V^{(j)} \in \mathbb{R}^{T\times d_t}$ is the learnable context matrix for layer j.

Deep text prompt injection. For class $y \in \mathcal{Y}$, let $Z_t^{(j-1)}(y) \in \mathbb{R}^{N_t\times d_t}$ be the input tokens to the j-th text block. We prepend layer-specific context tokens:

$$\hat{Z}_t^{(j-1)}(y) = \big[V^{(j)};\ Z_t^{(j-1)}(y)\big] \in \mathbb{R}^{(T+N_t)\times d_t}. \tag{13}$$

This sequence is processed by the frozen block, and the first T tokens are discarded to restore the original length N_t for subsequent layers. The final class embedding $w_y(\Theta)$ is derived from the output of the last text layer.

Deep visual prompt injection. Let $Z_v^{(j-1)}(x) \in \mathbb{R}^{(P+1)\times d_v}$ denote the input image tokens for the j-th visual block. With the defined style-aware experts in Eq. (7) and routing mechanism in Eq. (10), we compute the layer-specific image-conditioned visual prompts $\tilde{V}^{(j)}(x) \in \mathbb{R}^{T\times d_v}$. We prepend these prompts to the visual tokens:

$$\hat{Z}_v^{(j-1)}(x) = \big[\tilde{V}^{(j)}(x);\ Z_v^{(j-1)}(x)\big] \in \mathbb{R}^{(T+P+1)\times d_v}. \tag{14}$$

Similar to the text branch, the prompt tokens are discarded after processing to maintain the original token structure. The final image representation $f(x) = \phi(x)$ is extracted from the last visual block's output.

4 Experiment

We compare SaMoE with CLIP and representative prompt-learning baselines, including CoOp [50], ProGrad [51], KgCoOp [42], MaPLe [17], PromptSRC [18], DCPL [3], TCP [43], and Biomed-CoOp [20].

Table 1: **Comparison on base-to-novel generalization.** We report accuracy (%) on base classes (Base) and novel classes (Novel), and their harmonic mean (HM) across 11 biomedical datasets. Best and second-best results are highlighted in bold and underline, respectively.

Method	Average			BTMRI			BUSBRA			CHMNIST			COVID			CTKidney		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	43.00	38.63	40.70	48.25	56.05	51.86	55.17	25.81	35.17	32.85	26.60	29.39	47.46	38.30	42.39	52.75	41.42	46.40
CoOp	62.14	40.71	49.19	60.92	50.74	55.37	60.04	25.81	36.10	87.86	24.96	38.88	71.49	11.95	20.48	62.26	33.41	43.48
ProGrad	61.96	37.24	46.52	52.33	56.12	54.16	58.80	25.81	35.87	77.75	25.44	38.34	67.60	13.61	22.65	71.03	31.04	43.20
KgCoOp	56.43	41.19	47.62	55.58	56.71	56.14	55.26	26.52	35.84	69.33	29.25	41.14	66.35	37.04	47.54	53.40	44.31	48.44
MaPLe	76.37	39.52	52.09	77.67	57.07	65.80	61.68	41.94	49.93	90.51	17.02	28.65	**83.77**	40.32	54.44	76.31	35.90	48.83
PromptSRC	63.52	45.50	53.02	60.50	73.28	66.28	55.97	27.24	36.64	81.16	**28.99**	42.72	65.49	37.16	47.41	71.10	47.56	57.00
DCPL	75.80	39.54	51.97	72.33	45.84	56.11	61.63	41.76	49.78	92.11	22.07	35.61	82.95	56.03	66.88	**77.84**	31.65	45.00
TCP	58.87	40.62	48.07	54.25	41.62	47.10	60.57	45.16	51.74	66.18	25.05	36.34	59.70	11.70	19.57	66.38	44.52	53.30
BiomedCoOp	70.84	36.07	47.80	69.79	54.12	60.97	57.29	25.99	35.76	88.83	21.99	35.25	74.64	19.41	30.81	66.34	31.30	42.53
SaMoE (Ours)	**76.79**	**56.37**	**65.01**	**79.00**	**77.43**	**78.21**	**62.34**	**46.77**	**53.44**	**92.99**	28.86	**44.05**	81.14	**70.19**	**75.27**	69.37	**50.95**	**58.75**

Method	DermaMNIST			KneeXray			Kvasir			LUNGCOLON			OCTMNIST			RETINA		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	25.24	27.43	26.29	33.29	32.36	32.80	32.17	25.17	28.24	47.27	48.87	48.06	46.40	50.00	48.13	52.13	52.91	52.52
CoOp	23.30	52.59	32.29	43.29	60.22	50.37	65.95	21.61	32.55	69.76	55.56	61.85	66.27	50.00	57.00	72.41	61.00	66.22
ProGrad	32.69	26.85	29.48	35.36	39.66	37.39	70.05	24.89	36.73	72.51	56.25	63.35	72.40	49.73	58.96	71.04	60.26	65.21
KgCoOp	26.62	15.84	19.86	46.55	43.92	45.20	57.95	25.33	35.25	56.91	56.09	56.50	64.13	50.53	56.53	68.61	67.51	68.06
MaPLe	54.37	32.90	40.99	41.63	43.91	42.74	85.17	25.00	38.65	**98.20**	53.50	69.26	83.53	35.53	49.86	87.28	51.58	64.84
PromptSRC	42.64	20.78	27.94	39.10	67.03	49.39	71.41	23.67	35.55	78.39	51.47	62.14	74.27	51.33	60.71	58.72	71.97	64.67
DCPL	49.51	24.40	32.69	39.70	51.95	45.01	86.22	23.11	36.45	97.38	49.16	65.33	**86.60**	39.53	54.29	**87.47**	49.40	63.14
TCP	32.60	46.85	38.45	34.39	40.27	37.10	73.72	19.83	31.26	56.97	53.03	54.93	65.27	50.93	57.22	77.51	53.49	63.30
BiomedCoOp	44.42	18.66	26.28	48.26	42.46	45.17	79.72	23.84	36.70	91.60	55.06	68.78	72.87	50.00	59.31	85.52	53.91	66.13
SaMoE (Ours)	**57.20**	**61.58**	**59.31**	**50.22**	**67.15**	**57.45**	**87.28**	**25.83**	**39.86**	97.62	**56.33**	**71.44**	81.87	**51.47**	**63.20**	85.68	**83.46**	**84.56**

4.1 Evaluation settings

We report classification accuracy (%) for all settings. For base-to-novel generalization, we evaluate accuracy on base and novel classes separately and summarize their balance using the harmonic mean, $\mathrm{HM} = 2A_{\mathrm{base}}A_{\mathrm{novel}}/(A_{\mathrm{base}} + A_{\mathrm{novel}})$. For cross-domain generalization, we report source and target accuracies separately.

Base-to-Novel Generalization. In this setting, each dataset is split into disjoint base and novel class subsets. The model is adapted using a 16-shot labeled support set from the base classes only, and the novel classes are unseen in training. This protocol assesses within-dataset generalization to unseen classes, reflecting the ability to recognize novel classes of disease without fine-tuning.

Cross-site Domain Generalization. We evaluate generalization from a single source site to unseen target sites. Here, sites correspond to independent datasets from different institutions characterized by systematic acquisition shifts. The model is adapted with a 16-shot labeled support set from the source site and evaluated on unseen sites (zero-shot).

Cross-protocol Domain Generalization. We evaluate the effectiveness of models trained on one imaging protocol that are transferred to other protocols. A protocol corresponds to an MRI acquisition setting. The model is trained using a 16-shot labeled support set and evaluated on four target protocols, i.e., FLAIR, T1W, T1W-CE and T2W (zero-shot).

Cross-modality Domain Generalization. We evaluate the effectiveness of models under modality-induced compound domain shifts. We adapt the model on OCTMNIST as the source domain (16-shot) and test it on nine unseen target domains spanning different imaging modalities (zero-shot). This setting is used as a severe stress test because modality transfer involves not only style changes but also different physical acquisition mechanisms, anatomical structures, and label spaces.

Datasets. Experiments are conducted on 15 medical imaging datasets spanning 12 imaging modalities and 10 organs, including BTMRI [25], CHMNIST [14], LC25000 [2], COVID-QU-Ex [35], KneeXray [4], CTKidney [11], DermaMNIST [36, 6], Kvasir [29], OCTMNIST [16], RETINA [19, 30], BUSBRA [9], Derm7pt_C(Clinical), Derm7pt_D(Dermatoscopy) [15], PH2 [22], MILK10k [28] and Gazi [32]. Detailed dataset statistics and benchmark protocols are provided in Appendix F.

Implementation Details. We use CLIP [31] with a ViT-B/16 backbone for all experiments and report the average of three independent runs with different random seeds. We train using SGD with a learning rate of 0.0025 and a batch size of 4. We train for 50 epochs in the base-to-novel and 100 epochs in the cross-site, cross-protocol, and cross-modality settings. We initialize the first-layer text prompt with a photo of a {class}. For SaMoE, we set the prompt depth to 9 and use a prompt length of 8 tokens for the text branches, with 4 experts per layer and Top-K routing with $K = 2$. We set the

Table 2: **Cross-domain generalization under acquisition-induced style shifts.** Accuracy (%) after 16-shot source adaptation. Standard deviations are shown as lower-right subscripts ($\text{mean}_{\pm\text{std}}$). Frozen CLIP is shown without uncertainty. Avg. denotes mean target accuracy. Bold/underline indicate the best/second-best means.

Dataset	CLIP	CoOp	ProGrad	KgCoOp	MaPLe	Prompt SRC	DCPL	TCP	Biomed CoOp	SaMoE
Cross-site DG (Source: DermaMNIST)										
DermaMNIST	12.77	$31.55_{\pm 13.58}$	$30.01_{\pm 10.58}$	$15.83_{\pm 4.53}$	$53.82_{\pm 4.23}$	$49.92_{\pm 3.36}$	$\underline{55.70}_{\pm 2.02}$	$24.99_{\pm 3.53}$	$45.70_{\pm 2.62}$	$\mathbf{57.91}_{\pm 1.72}$
Derm7pt_C	22.65	$28.48_{\pm 18.83}$	$25.31_{\pm 3.60}$	$22.43_{\pm 0.57}$	$27.79_{\pm 10.09}$	$\underline{32.17}_{\pm 6.49}$	$24.98_{\pm 7.15}$	$17.83_{\pm 7.16}$	$30.63_{\pm 10.37}$	$\mathbf{34.91}_{\pm 7.51}$
Derm7pt_D	22.34	$27.57_{\pm 11.36}$	$23.95_{\pm 5.10}$	$24.39_{\pm 4.22}$	$32.71_{\pm 5.63}$	$32.49_{\pm 2.65}$	$30.92_{\pm 1.75}$	$18.91_{\pm 5.89}$	$\underline{32.75}_{\pm 8.44}$	$\mathbf{37.90}_{\pm 12.94}$
MILK10k	24.42	$20.71_{\pm 6.20}$	$\underline{33.49}_{\pm 7.79}$	$14.87_{\pm 1.81}$	$28.31_{\pm 7.08}$	$27.34_{\pm 3.93}$	$23.82_{\pm 13.33}$	$12.46_{\pm 3.13}$	$20.44_{\pm 6.29}$	$\mathbf{33.84}_{\pm 3.32}$
PH2	24.00	$34.50_{\pm 20.25}$	$34.83_{\pm 6.35}$	$30.00_{\pm 9.99}$	$34.33_{\pm 17.06}$	$25.83_{\pm 3.25}$	$\underline{39.83}_{\pm 12.07}$	$27.93_{\pm 16.54}$	$32.50_{\pm 28.46}$	$\mathbf{46.50}_{\pm 7.59}$
Avg.	23.35	$27.82_{\pm 11.34}$	$29.39_{\pm 3.14}$	$22.92_{\pm 3.27}$	$\underline{30.79}_{\pm 9.89}$	$29.46_{\pm 3.66}$	$29.89_{\pm 6.74}$	$19.28_{\pm 5.02}$	$29.08_{\pm 12.20}$	$\mathbf{38.29}_{\pm 6.27}$
Cross-protocol DG (Source: BTMRI)										
BTMRI	21.78	$67.73_{\pm 5.60}$	$59.74_{\pm 3.03}$	$55.72_{\pm 5.45}$	$80.70_{\pm 3.24}$	$58.82_{\pm 6.70}$	$\underline{80.88}_{\pm 5.67}$	$57.37_{\pm 2.99}$	$68.53_{\pm 2.99}$	$\mathbf{82.86}_{\pm 1.07}$
FLAIR	51.44	$56.00_{\pm 4.95}$	$\underline{57.00}_{\pm 3.46}$	$53.11_{\pm 1.26}$	$55.11_{\pm 3.29}$	$44.00_{\pm 2.00}$	$51.78_{\pm 1.02}$	$50.00_{\pm 0.00}$	$51.03_{\pm 4.16}$	$\mathbf{58.67}_{\pm 0.51}$
T1W	54.11	$57.67_{\pm 0.67}$	$\underline{60.84}_{\pm 3.84}$	$51.67_{\pm 2.65}$	$55.22_{\pm 5.50}$	$54.78_{\pm 0.51}$	$54.00_{\pm 2.33}$	$50.00_{\pm 0.00}$	$58.45_{\pm 4.67}$	$\mathbf{62.33}_{\pm 1.54}$
T1W-CE	61.47	$71.68_{\pm 8.37}$	$\underline{77.60}_{\pm 3.89}$	$58.78_{\pm 18.52}$	$70.25_{\pm 3.96}$	$49.10_{\pm 7.02}$	$68.81_{\pm 1.52}$	$44.64_{\pm 0.00}$	$72.76_{\pm 12.73}$	$\mathbf{80.11}_{\pm 23.34}$
T2W	51.44	$62.33_{\pm 1.34}$	$62.67_{\pm 5.43}$	$53.78_{\pm 0.77}$	$57.67_{\pm 4.49}$	$57.56_{\pm 4.16}$	$\underline{63.00}_{\pm 7.69}$	$50.00_{\pm 0.00}$	$61.78_{\pm 5.50}$	$\mathbf{68.89}_{\pm 3.56}$
Avg.	54.61	$61.92_{\pm 2.96}$	$\underline{64.53}_{\pm 2.72}$	$54.34_{\pm 3.86}$	$59.56_{\pm 2.10}$	$51.36_{\pm 3.38}$	$59.40_{\pm 4.43}$	$48.66_{\pm 0.00}$	$61.01_{\pm 5.06}$	$\mathbf{67.50}_{\pm 6.43}$
Compound cross-modality DG (Source: OCTMNIST)										
OCTMNIST	24.57	$45.17_{\pm 1.80}$	$56.20_{\pm 0.00}$	$38.93_{\pm 6.86}$	$76.10_{\pm 7.58}$	$35.13_{\pm 9.56}$	$\underline{78.90}_{\pm 1.86}$	$38.30_{\pm 4.24}$	$69.33_{\pm 5.35}$	$\mathbf{79.40}_{\pm 4.38}$
BTMRI	21.76	$25.76_{\pm 2.83}$	$24.14_{\pm 1.12}$	$\underline{26.30}_{\pm 4.37}$	$24.17_{\pm 3.15}$	$25.72_{\pm 0.49}$	$24.11_{\pm 1.01}$	$25.63_{\pm 7.68}$	$24.07_{\pm 0.99}$	$\mathbf{28.36}_{\pm 2.65}$
BUSBRA	8.59	$21.32_{\pm 6.11}$	$19.89_{\pm 14.32}$	$18.36_{\pm 8.77}$	$24.22_{\pm 10.54}$	$\underline{32.68}_{\pm 4.77}$	$17.11_{\pm 8.26}$	$24.21_{\pm 4.77}$	$12.91_{\pm 5.57}$	$\mathbf{36.23}_{\pm 13.10}$
CHMNIST	14.30	$13.52_{\pm 1.49}$	$16.69_{\pm 5.54}$	$16.05_{\pm 5.22}$	$\underline{18.15}_{\pm 6.45}$	$13.05_{\pm 2.67}$	$12.65_{\pm 0.89}$	$15.11_{\pm 7.16}$	$17.73_{\pm 7.28}$	$\mathbf{19.61}_{\pm 3.00}$
COVID	28.13	$27.96_{\pm 18.43}$	$17.41_{\pm 15.55}$	$14.96_{\pm 8.22}$	$21.33_{\pm 13.02}$	$\underline{28.31}_{\pm 0.16}$	$20.86_{\pm 12.22}$	$25.04_{\pm 6.55}$	$21.09_{\pm 12.70}$	$\mathbf{28.40}_{\pm 4.21}$
CTKidney	24.29	$21.65_{\pm 4.96}$	$19.13_{\pm 11.38}$	$20.97_{\pm 10.57}$	$24.64_{\pm 9.01}$	$26.81_{\pm 5.22}$	$25.93_{\pm 7.31}$	$\underline{29.80}_{\pm 8.19}$	$18.50_{\pm 0.28}$	$\mathbf{30.04}_{\pm 6.06}$
KneeXray	16.73	$17.93_{\pm 11.89}$	$\underline{28.23}_{\pm 14.65}$	$19.20_{\pm 10.93}$	$19.77_{\pm 17.63}$	$18.03_{\pm 0.28}$	$23.31_{\pm 6.00}$	$13.99_{\pm 2.87}$	$16.20_{\pm 18.46}$	$\mathbf{38.59}_{\pm 2.59}$
Kvasir	13.75	$14.53_{\pm 5.33}$	$\underline{17.00}_{\pm 6.48}$	$16.03_{\pm 5.21}$	$14.25_{\pm 1.64}$	$14.64_{\pm 3.13}$	$12.33_{\pm 2.76}$	$15.42_{\pm 2.11}$	$13.03_{\pm 2.98}$	$\mathbf{21.34}_{\pm 1.90}$
LUNGCOLON	20.02	$20.19_{\pm 0.33}$	$17.55_{\pm 3.47}$	$21.24_{\pm 5.67}$	$20.21_{\pm 0.49}$	$\underline{22.66}_{\pm 4.65}$	$19.66_{\pm 0.41}$	$20.95_{\pm 1.64}$	$20.01_{\pm 0.02}$	$\mathbf{28.03}_{\pm 0.83}$
RETINA	26.84	$26.24_{\pm 0.36}$	$28.79_{\pm 9.48}$	$28.26_{\pm 2.07}$	$\underline{31.94}_{\pm 7.26}$	$26.00_{\pm 0.05}$	$25.87_{\pm 2.69}$	$26.58_{\pm 5.85}$	$26.29_{\pm 0.74}$	$\mathbf{32.26}_{\pm 4.01}$
Avg.	19.38	$21.01_{\pm 0.71}$	$20.98_{\pm 0.83}$	$20.15_{\pm 4.44}$	$22.08_{\pm 1.01}$	$\underline{23.10}_{\pm 0.79}$	$20.20_{\pm 3.43}$	$21.86_{\pm 2.97}$	$18.87_{\pm 2.05}$	$\mathbf{29.21}_{\pm 0.44}$

hidden layer dimension for routing to 256 and $\lambda_{\text{bal}} = 0.1$, more details in Appendix B. Text prompt tokens at each layer are initialized randomly and experiments are conducted on RTX4090 GPU.

4.2 Base-to-Novel Generalization

As shown in Table 1, our SaMoE achieves the best Harmonic Mean (HM) of 65.01%, surpassing the second-best method PromptSRC by 11.99%. Different from CoOp sacrificing the accuracy of Novel classes for the performance of Base classes, SaMoE ranks first in both scenes (Base: 76.79%, Novel: 56.37%). This superior balance stems from our mixture-of-experts design, which captures diverse style components to prevent overfitting source styles. Notably, on the dataset BTMRI with strong acquisition shifts, SaMoE improves Novel accuracy by 20.36% compared to the MaPLe. This suggests that image-conditioned routing helps adapt prompt composition to acquisition-induced style variations, while the deep injection mechanism strengthens cross-modal alignment. Additional seed-level statistics are reported in Appendix B.5.

4.3 Cross-domain Generalization

Cross-site DG. As shown in Table 2, SaMoE transfers consistently from DermaMNIST to four unseen dermoscopy sites. It achieves the highest source accuracy (57.91%) and the best target accuracies on Derm7pt_C (34.91%), Derm7pt_D (37.90%), MILK10k (33.84%), and PH2 (46.50%). The gains are especially clear on PH2, where SaMoE exceeds the strongest baseline DCPL by 6.67%, and on Derm7pt_D, where it improves over BiomedCoOp by 5.15%. These results indicate that single prompt configurations remain sensitive to site-specific styles, while image-conditioned expert composition can better match input-dependent site styles.

Cross-protocol DG. For MRI protocol transfer, SaMoE also attains the best source accuracy on BTMRI (82.86%) and outperforms all baselines on the four unseen Gazi protocols. Compared with the strongest competing results, SaMoE improves FLAIR by 1.67%, T1W by 1.49%, T1W-CE by 2.51%, and T2W by 5.89%. Protocol changes directly alter intensity distributions and tissue contrast; the consistent gains therefore support our design choice of routing prompts according to compact visual style statistics rather than relying on a fixed source-protocol prompt.

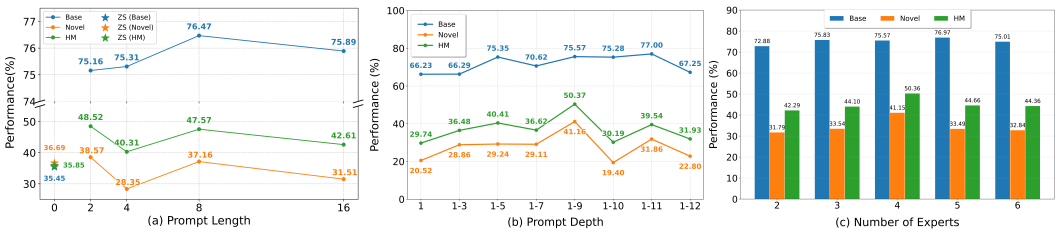


Figure 3: Analysis of Prompt Length (a), Prompt Depth (b), and the Number of Experts (c) on Base, Novel, and their HM. ZS denotes Zero-shot CLIP baseline, respectively.

Compound cross-modality DG. We further evaluate a severe compound shift by adapting on OCTMNIST and testing on nine target domains spanning MRI, ultrasound, etc. SaMoE achieves the best source accuracy (79.40%) and the highest average target accuracy (29.21%), improving over the second-best PromptSRC average by 6.11%. This setting involves more than acquisition-style shifts because cross-modality transfer also changes physical imaging mechanisms, anatomical visibility, and label spaces. We therefore treat it as a stress test: the improvements suggest that image-conditioned routing and prompt composition remain useful under broader domain gaps, likely because modality-dependent style statistics still provide informative cues for adapting prompt representations.

4.4 Ablation Study

Different Components of SaMoE. Table 3 presents a component ablation of SaMoE. Starting from a shallow prompting baseline ($j = 1$), deep prompt injection improves generalization and raises HM from 49.19% to 56.98% with only a small increase in GFLOPS. The "+ Expert bank" removes the image-conditioned router and combines experts with fixed uniform weights. Since our experts are linear projections, this static combination is algebraically equivalent to a single larger projection, so this row mainly controls for additional prompt transformation capacity rather than demonstrating input-adaptive expert selection. Introducing the image-conditioned router further improves HM to 63.43% by enabling style-dependent expert composition, and the load-balancing loss $\mathcal{L}_{\text{bal}}$ brings a final gain to 65.01% by preventing expert collapse and encouraging reliable specialization. Overall, the full SaMoE achieves the best balance between base fitting and robust generalization (Base: 76.79%, Novel: 56.37%,HM: 65.01%) with modest computational overhead. We analyze the learned prompts, computational efficiency in Appendix C and visual interpretability in Appendix E.

Table 3: Effect of SaMoE. Reporting Base/Novel accuracy, HM, and GFLOPs average on 11 datasets.

Method	Base	Novel	HM	GFLOPs
Shallow ($j = 1$)	62.14	40.71	49.19	19.99
+ Deep prompt	68.66	48.69	56.98	20.47
+ Expert bank	73.18	50.17	59.53	20.96
+ Router	76.20	54.32	63.43	20.96
+ $\mathcal{L}_{\text{bal}}$	**76.79**	**56.37**	**65.01**	20.96

Prompt Length. Fig. 3 (a) shows that increasing the number of prompt tokens improves performance over CLIP, but the trend is non-monotonic. A moderate length works best, with 8 tokens achieving the strongest overall balance (Base: 76.47%, Novel: 37.16%, HM: 47.57%). In contrast, longer prompts reduce Novel/HM, suggesting over-specialization.

Prompt Depth. Fig. 3 (b) indicates that deeper prompt injection generally improves generalization. The best result is obtained at depth 1-9 (Base: 75.57%, Novel: 41.16%, HM: 50.37%), whereas a larger depth (e.g., 1-10 or 1-12) tends to degrade the HM score, reflecting potential instability.

Number of Experts. Fig. 3 (c) shows that SaMoE is relatively robust to the expert count but performs best with a compact ExpertBank. Using 4 experts yields the highest HM (50.36%) and Novel (41.15%), while adding more experts provides no further gains and can even hurt performance. We provide further analysis of expert utilization and routing behavior across domains in Appendix D.

4.5 Additional Analysis

Capacity matched controls. We enlarge MaPLe's cross-modal coupling and activate/enlarge the dormant Meta-Net in the CoCoOp-derived BiomedCoOp implementation, keeping the remaining objectives and training settings unchanged. Table 4 shows that independent SaMoE exceeds parameter-matched Heavy MaPLe by 13.43%. Shared SaMoE also exceeds MaPLe by 5.82% with fewer

Table 4: Capacity-controlled base-to-novel results over 11 datasets (three seeds). Values are percentages; subscripts denote standard deviations. Params denotes trainable parameters.

Method	Params	Accuracy (%)			Macro-F1 (%)		
		Base	Novel	HM	Base	Novel	HM
BiomedCoOp	0.004M	$70.84_{\pm 2.62}$	$36.07_{\pm 8.83}$	$47.80_{\pm 7.78}$	$67.76_{\pm 1.95}$	$27.47_{\pm 4.45}$	$39.09_{\pm 4.48}$
Heavy BiomedCoOp	15.91M	$68.69_{\pm 0.21}$	$37.93_{\pm 3.73}$	$48.80_{\pm 3.14}$	$66.62_{\pm 0.86}$	$28.18_{\pm 3.76}$	$39.61_{\pm 3.72}$
MaPLe	3.63M	$76.37_{\pm 3.13}$	$39.52_{\pm 8.76}$	$52.09_{\pm 7.64}$	$74.06_{\pm 1.93}$	$32.67_{\pm 4.39}$	$45.34_{\pm 4.24}$
Heavy MaPLe	15.92M	$76.66_{\pm 1.59}$	$38.88_{\pm 1.79}$	$51.58_{\pm 1.81}$	$75.56_{\pm 1.66}$	$32.14_{\pm 0.54}$	$45.10_{\pm 0.61}$
SaMoE (Shared)	2.59M	$75.71_{\pm 1.63}$	$46.89_{\pm 7.91}$	$57.91_{\pm 6.05}$	$74.28_{\pm 1.32}$	$39.06_{\pm 3.87}$	$51.20_{\pm 3.34}$
SaMoE (Independent)	15.91M	$\mathbf{76.79}_{\pm 2.44}$	$\mathbf{56.37}_{\pm 5.91}$	$\mathbf{65.01}_{\pm 4.03}$	$\mathbf{76.05}_{\pm 1.07}$	$\mathbf{45.13}_{\pm 4.00}$	$\mathbf{56.65}_{\pm 3.16}$

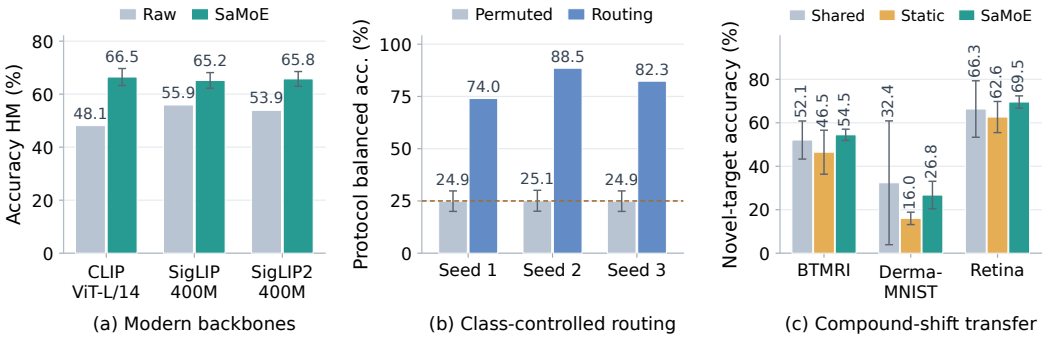


Figure 4: **Additional generalization experiments.** (a) Base-to-novel accuracy HM on modern backbones. (b) Leave-one-class-out MRI protocol prediction from routing versus class- and subject-preserving permutations (25% chance). (c) Shared prompt, Static experts, and SaMoE on novel classes under target style mixtures (64-shot). Bars report means; error bars show training-seed standard deviations, except the permutation bars in (b), which show null-distribution standard deviations.

parameters. This supports the contribution of input-adaptive transformations beyond increased capacity. Additional results are provided in Appendix G.3.

Generalization and Routing Analysis. Fig. 4 evaluates backbone transferability, acquisition-sensitive routing, and generalization under controlled compound shifts. Fig. 4 (a) SaMoE improves base-to-novel accuracy HM over each corresponding frozen modern backbone. Fig. 4 (b) In the class-controlled MRI analysis, routing representations predict acquisition protocol on a held-out class substantially above the permutation baseline, indicating that they encode acquisition information beyond class identity. Fig. 4 (c) Under the separate 64-shot controlled compound-shift setting, SaMoE outperforms static experts in novel-target accuracy on all three datasets, although the shared-prompt baseline remains stronger on DermaMNIST. Full protocols and statistics are provided in Appendix G.

5 Conclusion

This paper addresses biomedical VLM adaptation under acquisition-induced style shifts. We formulate such shifts from a mixture-shift perspective and propose SaMoE, a style-aware mixture-of-experts prompt learning framework that uses an image-conditioned router to compose multimodal prompts across the VLM's dual-branch, multi-layer architecture. Experiments across 15 datasets spanning 12 modalities and 10 organs demonstrate strong generalization over state-of-the-art baselines in cross-site, cross-protocol, and compound cross-modality settings. SaMoE also has several limitations. First, SaMoE is most suitable when acquisition-induced style variation is a major source of domain shift, and its gains may be smaller when the style gap is mild or when semantic and task-level differences dominate, especially in compound cross-modality transfer. Second, SaMoE is more complex than standard single-prompt tuning due to the expert bank and routing mechanism. Future work will extend SaMoE to dense prediction, develop more interpretable routing signals, and study continual or multi-source adaptation for evolving biomedical acquisition pipelines.

Acknowledgements

This work is supported in part by the Key Research and Development Program of Hunan Province (No. 2026JK2084), in part by the National Natural Science Foundation of China (No. 62402532), in part by the Hunan Provincial Natural Science Foundation of China (No. 2024JJ6526), and in part by the Hunan Provincial Degree and Postgraduate Teaching Reform Project (No. 2025JGYB050). This work is also supported in part by the Center for Computational Biology and Bioinformatics, Furong Laboratory and Bioinformatics Center, Xiangya Hospital, Central South University.

References

[1] Yequan Bie, Luyang Luo, Zhixuan Chen, and Hao Chen. Xcoop: Explainable prompt learning for computer-aided diagnosis via concept-guided context optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 773–783. Springer, 2024.

[2] Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019.

[3] Qinglong Cao, Zhengqin Xu, Yuntian Chen, Chao Ma, and Xiaokang Yang. Domain-controlled prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 936–944, 2024.

[4] Pingjun Chen. Knee osteoarthritis severity grading dataset. 2018. doi: 10.17632/56rmx5bjcr.1.

[5] Joon-Young Choi, Junho Kim, Jun-Hyung Park, Wing-Lam Mok, and SangKeun Lee. Smop: Towards efficient and effective prompt tuning with sparse mixture-of-prompts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14306–14316, 2023.

[6] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). 2019.

[7] Yu Du, Tong Niu, and Rong Zhao. Mixture of prompts learning for vision-language models. *Frontiers in Artificial Intelligence*, 8:1580973, 2025.

[8] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132:581–595, 2024.

[9] Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics*, 51(4):3110–3123, 2024.

[10] Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. Lp++: A surprisingly strong linear probe for few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23773–23782, 2024.

[11] Md Nazmul Islam, Mehedi Hasan, Md Kabir Hossain, Md Golam Rabiul Alam, Md Zia Uddin, and Ahmet Soylu. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ct-radiography. *Scientific Reports*, 12:11440, 2022.

[12] 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.

[13] Ruixiang Jiang, Lingbo Liu, and Changwen Chen. Mope: Mixture of prompt experts for parameter-efficient and scalable multimodal fusion. *arXiv preprint arXiv:2403.10568*, 2024.

[14] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6:1–11, 2016.

[15] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, mar 2019. ISSN 2168-2194. doi: 10.1109/JBHI.2018.2824327.

[16] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

[17] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023.

[18] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15190–15200, 2023.

[19] Thomas Köhler, Attila Budai, Martin F Kraus, Jan Odstrčilik, Georg Michelson, and Joachim Hornegger. Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In *Proceedings of the 26th IEEE international symposium on computer-based medical systems*, pages 95–100. IEEE, 2013.

[20] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Biomedcoop: Learning to prompt for biomedical vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14766–14776, 2025.

[21] Sarah Matta, Mathieu Lamard, Philippe Zhang, Alexandre Le Guilcher, Laurent Borderie, Béatrice Cochener, and Gwenolé Quellec. A systematic review of generalization research in medical image classification. *Computers in Biology and Medicine*, 183:109256, 2024. doi: 10.1016/j.compbiomed.2024.109256.

[22] Teresa Mendonça, M Celebi, T Mendonca, and J Marques. Ph2: A public database for the analysis of dermoscopic images. *Dermoscopy image analysis*, 2, 2015.

[23] Pingyi Miao, Xianlai Chen, Kai Sun, Yunbo Wang, Shuang Zhao, and Ying An. Biodpp: Dynamic prompt policy learning for biomedical vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 8052–8060, 2026.

[24] Shentong Mo, Xufang Luo, and Dongsheng Li. pmoe: Prompting diverse experts together wins more in visual adaptation. *arXiv preprint arXiv:2602.22938*, 2026.

[25] Msoud Nickparvar. Brain tumor mri dataset, 2021.

[26] Jun Yeong Park, JunYoung Seo, Minji Kang, and Yu Rang Park. Moeclip: Patch-specialized experts for zero-shot anomaly detection. *arXiv preprint arXiv:2603.03101*, 2026.

[27] Yuxin Peng, Zishuo Wang, Geng Li, Xiangtian Zheng, Sibo Yin, and Hulingxiao He. A survey on fine-grained multimodal large language models. *Chinese Journal of Electronics*, 35(2): 771–803, 2026.

[28] Tschandl Philipp, Akay Nisa Bengü, Rosendahl Cliff, Rotemberg Veronica, Todorovska Verche, Weber Jochen, Wolber Anna Katharina, Müller Christoph, Kurtansky Nicholas, Halpern Allan, et al. Milk10k: A hierarchical multimodal imaging learning toolkit for diagnosing pigmented and non-pigmented skin cancer and its simulators. *Journal of Investigative Dermatology*, 2025.

[29] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169, 2017.

[30] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid). 2018. doi: 10.21227/H25W98.

[31] 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.

[32] S Sagiroglu, R Terzi, E Celtikci, AÖ Börcek, Y Atay, B Arslan, MC Sahin, K Nernekli, U Demirezen, OB Ozdemir, K Özdem Karaca, and N Azgınoğlu. A novel brain tumor magnetic resonance imaging dataset (gazi brains 2020): initial benchmark results and comprehensive analysis. *PeerJ Computer Science*, 11:e2920, jun 2025. doi: 10.7717/peerj-cs.2920.

[33] Minye Shao, Sihan Guo, Xinrun Li, Xingyu Miao, Haoran Duan, and Yang Long. vMFCoOp: Towards equilibrium on a unified hyperspherical manifold for prompting biomedical VLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(11):8851–8859, 2026. doi: 10.1609/aaai.v40i11.37839.

[34] Yuheng Shao, Lizhang Wang, Changhao Li, Peixian Chen, and Qinyuan Liu. Promptmoe: Generalizable zero-shot anomaly detection via visually-guided prompt mixtures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 8878–8886, 2026.

[35] Anas M Tahir, Muhammad EH Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M Sohel Rahman, Somaya Al-Maadeed, et al. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in biology and medicine*, 139:105002, 2021.

[36] 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*, 5: 1–9, 2018.

[37] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL `https://arxiv.org/abs/2502.14786`.

[38] Yunbo Wang and Yuxin Peng. Mars: Learning modality-agnostic representation for scalable cross-media retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4765–4777, 2022.

[39] Yunbo Wang, YuJie Wu, Zhien Dai, Can Tian, Jun Long, and Jianhai Chen. Noisy correspondence rectification via asymmetric similarity learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21384–21392, 2025.

[40] Linhui Xiao, Guiping Cao, Mingyue Guo, Xianchao Guan, Fan Yang, Ming Tao, Xin Li, Yuxin Peng, and Yaowei Wang. A survey on the green development of large models: From resource-efficient architectures to hardware-software co-design. *arXiv preprint arXiv:2607.09084*, 2026.

[41] Jie Yan, Xiaoyuan Guan, Wei-Shi Zheng, Hao Chen, and Ruixuan Wang. Global and local vision-language alignment for few-shot learning and few-shot OOD detection. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15964, pages 208–218. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-04971-1_20.

[42] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023.

[43] Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23438–23448, 2024.

[44] Zhaoda Ye, Xiangteng He, and Yuxin Peng. Unsupervised cross-media hashing learning via knowledge graph. *Chinese Journal of Electronics*, 31(6):1081–1091, 2022.

[45] Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A. Mazurowski, and Heung-Il Suk. Domain generalization for medical image analysis: A review. *Proceedings of the IEEE*, 112 (10):1583–1609, 2024. doi: 10.1109/JPROC.2024.3507831.

[46] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. URL `https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html`.

[47] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022.

[48] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

[49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, pages 2337–2348, 2022.

[51] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15659–15669, 2023.

APPENDIX

Table of Contents

A Additional Discussion on Domain Discrepancy Under Acquisition Shifts 16

B Additional Implementation Details 19

- B.1 Top-K Routing Implementation and Soft Routing Comparison
- B.2 Additional Adapted Baseline Comparisons
- B.3 When is SaMoE Less Effective?
- B.4 Hyperparameter Sensitivity Analysis
- B.5 Seed Stability Analysis

C Prompt Analysis 23

- C.1 Prompting Complexity and Inference Throughput
- C.2 Text Prompt Embedding Analysis
- C.3 Supplementary Cross-Dataset Prompt Diversity
- C.4 Controlled Prompt Composition under Acquisition Shifts

D Qualitative Analysis of Expert Routing 24

- D.1 Expert Utilization and Routing Behavior
- D.2 Analysis on Router Input Features
- D.3 Analysis on Expert Load Balancing Mechanisms

E Visual Interpretability 27

F Datasets and Benchmarks 27

H Discussion on Limitations and Broader Impact 35

- H.1 Discussion on Limitations
- H.2 Impact Statement

A Additional Discussion on Domain Discrepancy Under Acquisition Shifts

This section complements Sec. 3.1 by detailing how acquisition-induced style shifts can enlarge the effective source-target discrepancy and thereby hinder generalization of prompted models under limited supervision.

Derivation of the domain adaptation bound. Let $\mathcal{H}$ denote the hypothesis class induced by prompted models, and let ℓ be the 0-1 loss for classification. We derive the target risk bound starting from the triangle inequality for classification error. For any hypotheses h, h', h'', the risk satisfies $\mathcal{R}(h, h'') \leq \mathcal{R}(h, h') + \mathcal{R}(h', h'')$, where $\mathcal{R}(h, h') = \mathbb{E}_x[\mathbb{I}(h(x) \neq h'(x))]$ represents the probability of disagreement. Let $h^* = \arg\min_{h \in \mathcal{H}} (\mathcal{R}_{\mathcal{D}_s}(h) + \mathcal{R}_{\mathcal{D}_t}(h))$ be the optimal joint hypothesis, with combined error $\lambda^* = \mathcal{R}_{\mathcal{D}_s}(h^*) + \mathcal{R}_{\mathcal{D}_t}(h^*)$.

First, we apply the triangle inequality to the target risk $\mathcal{R}_{\mathcal{D}_t}(h)$ with respect to the optimal joint hypothesis h^*:

$$\mathcal{R}_{\mathcal{D}_t}(h) \leq \mathcal{R}_{\mathcal{D}_t}(h^*) + \mathcal{R}_{\mathcal{D}_t}(h, h^*). \tag{15}$$

The term $\mathcal{R}_{\mathcal{D}_t}(h, h^*)$ represents the disagreement between h and h^* on the target domain. To connect this to the source domain, we add and subtract the source domain disagreement $\mathcal{R}_{\mathcal{D}_s}(h, h^*)$:

$$\begin{aligned} \mathcal{R}_{\mathcal{D}_t}(h, h^*) &= \mathcal{R}_{\mathcal{D}_s}(h, h^*) + (\mathcal{R}_{\mathcal{D}_t}(h, h^*) - \mathcal{R}_{\mathcal{D}_s}(h, h^*)) \\ &\leq \mathcal{R}_{\mathcal{D}_s}(h, h^*) + |\mathcal{R}_{\mathcal{D}_t}(h, h^*) - \mathcal{R}_{\mathcal{D}_s}(h, h^*)| \,. \end{aligned} \tag{16}$$

The absolute difference term in Eq. (16) quantifies how the disagreement probability changes across domains. This matches the definition of the $\mathcal{H}\Delta\mathcal{H}$-distance:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) = 2 \sup_{h_1, h_2 \in \mathcal{H}} \Big| \underbrace{\Pr_{x \sim \mathcal{D}_s}[h_1(x) \neq h_2(x)]}_{\mathcal{R}_{\mathcal{D}_s}(h_1, h_2)} - \underbrace{\Pr_{x \sim \mathcal{D}_t}[h_1(x) \neq h_2(x)]}_{\mathcal{R}_{\mathcal{D}_t}(h_1, h_2)} \Big|. \tag{17}$$

From Eq. (17), it follows that $|\mathcal{R}_{\mathcal{D}_t}(h, h^*) - \mathcal{R}_{\mathcal{D}_s}(h, h^*)| \leq \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t)$. Finally, we apply the triangle inequality again to the source risk term: $\mathcal{R}_{\mathcal{D}_s}(h, h^*) \leq \mathcal{R}_{\mathcal{D}_s}(h) + \mathcal{R}_{\mathcal{D}_s}(h^*)$. Substituting these results back into Eq. (15), we obtain the final bound:

$$\begin{aligned} \mathcal{R}_{\mathcal{D}_t}(h) &\leq \mathcal{R}_{\mathcal{D}_t}(h^*) + [\mathcal{R}_{\mathcal{D}_s}(h) + \mathcal{R}_{\mathcal{D}_s}(h^*)] + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) \\ &= \mathcal{R}_{\mathcal{D}_s}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) + \lambda^*. \end{aligned} \tag{18}$$

Eq. (17) measures how differently hypotheses in $\mathcal{H}$ can behave across domains. A large $d_{\mathcal{H}\Delta\mathcal{H}}$ indicates significant domain discrepancy, which can prevent target generalization even when the source risk is small. In summary, Eq. (18) establishes that minimizing source risk alone is insufficient; the domain discrepancy $d_{\mathcal{H}\Delta\mathcal{H}}$ must also be controlled. This result serves as the theoretical bridge to the next section, where we identify acquisition-induced style shifts (modeled as mixture shifts) as one important physical source of domain discrepancy in biomedical imaging.

Mixture shift induced by acquisition styles. In the main text, we consider that cross-domain shifts often arise from acquisition pipelines. Site and protocol changes predominantly affect image style, while modality changes can additionally alter physical imaging principles, anatomical visibility, and label semantics. Thus, the mixture-shift formulation is intended to model the acquisition-style component of the domain gap rather than reduce all cross-modality transfer to a pure style shift. To formalize this, consider the mixture form used in Sec. 3.1:

$$\mathcal{D}_s = \sum_{m=1}^{M} \pi_s(m) \mathcal{D}_m, \qquad \mathcal{D}_t = \sum_{m=1}^{M} \pi_t(m) \mathcal{D}_m, \tag{19}$$

where each $\mathcal{D}_m$ denotes a joint distribution over (x, y) induced by a latent acquisition style component, and cross-domain style shifts are captured by changes in mixture weights. When label spaces differ, as in open-set or cross-modality evaluation, an additional semantic shift is present and is not fully explained by this style-mixture model. We therefore use the mixture formulation to characterize the acquisition-style component of the domain gap, while treating label-space or anatomical changes as broader distribution shifts in the empirical evaluation. The formulation also assumes that the

target style components are shared with, close to, or statistically inferable from the style factors observed in the source domain. If a target domain contains an entirely disjoint acquisition style with no corresponding source evidence, the mixture bound remains a decomposition of the target risk but does not imply that source-only training can estimate or reduce that unseen component risk. For any measurable event A on the image space, the probability of A under a mixture distribution is the weighted sum of component probabilities, i.e., $\Pr_{\mathcal{D}}[A] = \sum_{m=1}^{M} \pi(m) \Pr_{\mathcal{D}_m}[A]$. Therefore, the difference in event probabilities between domains can be derived as follows:

$$\begin{aligned}
\big|\Pr_{\mathcal{D}_s}[A] - \Pr_{\mathcal{D}_t}[A]\big| &= \Big|\sum_{m=1}^{M} \pi_s(m) \Pr_{\mathcal{D}_m}[A] - \sum_{m=1}^{M} \pi_t(m) \Pr_{\mathcal{D}_m}[A]\Big| \\
&= \Big|\sum_{m=1}^{M} (\pi_s(m) - \pi_t(m)) \Pr_{\mathcal{D}_m}[A]\Big| \\
&\leq \sum_{m=1}^{M} \big|\pi_s(m) - \pi_t(m)\big| \cdot \underbrace{\Pr_{\mathcal{D}_m}[A]}_{\leq 1} \\
&\leq \sum_{m=1}^{M} |\pi_s(m) - \pi_t(m)| \\
&= \|\pi_s - \pi_t\|_1. \qquad (20)
\end{aligned}$$

We can now link this result back to the domain discrepancy defined in Eq. (17). Let $A_{h,h'} = \{x \mid h(x) \neq h'(x)\}$ be the disagreement event for any pair of hypotheses. Since Eq. (20) holds for *any* measurable event, it strictly bounds the probability divergence of this specific disagreement event:

$$\Big|\Pr_{\mathcal{D}_s}[A_{h,h'}] - \Pr_{\mathcal{D}_t}[A_{h,h'}]\Big| \leq \|\pi_s - \pi_t\|_1. \qquad (21)$$

Within the shared-component abstraction, the upper bound $\|\pi_s - \pi_t\|_1$ is determined by the acquisition-style mixture shift and is independent of the hypothesis space $\mathcal{H}$, so taking the supremum of the left-hand side over $h, h' \in \mathcal{H}$ (as required by Eq. (17)) does not exceed the bound. Multiplying by 2 yields the final bound on the discrepancy:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) \leq 2\|\pi_s - \pi_t\|_1. \qquad (22)$$

Eq. (22) upper-bounds the part of the $\mathcal{H}\Delta\mathcal{H}$ discrepancy induced by reweighting shared acquisition-style components. A larger mixture-weight shift permits a larger domain discrepancy, although the realized discrepancy also depends on how the component distributions affect hypothesis disagreements. This provides a concrete mechanism through which acquisition shifts can hinder standard prompt tuning: a prompt optimized for the source mixture π_s may fit the dominant source styles while transferring poorly to a target mixture π_t with different acquisition-style proportions.

Worst-case target risk under unknown mixture proportions. Under the mixture assumption in Eq. (19), the target risk $\mathcal{R}_{\mathcal{D}_t}(h)$ naturally decomposes into a weighted sum of component risks. Let $p_t(x, y)$ and $p_m(x, y)$ denote the probability density functions of $\mathcal{D}_t$ and $\mathcal{D}_m$, respectively. By definition, $p_t(x, y) = \sum_{m=1}^{M} \pi_t(m) p_m(x, y)$. Utilizing the linearity of expectation, we derive:

$$\begin{aligned}
\mathcal{R}_{\mathcal{D}_t}(h) &= \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\big[\ell(h(x), y)\big] \\
&= \int \ell(h(x), y)\, p_t(x, y)\, d(x, y) \\
&= \int \ell(h(x), y) \left(\sum_{m=1}^{M} \pi_t(m) p_m(x, y)\right) d(x, y) \\
&= \sum_{m=1}^{M} \pi_t(m) \underbrace{\int \ell(h(x), y) p_m(x, y)\, d(x, y)}_{\mathcal{R}_{\mathcal{D}_m}(h)} \\
&= \sum_{m=1}^{M} \pi_t(m) \mathcal{R}_{\mathcal{D}_m}(h). \qquad (23)
\end{aligned}$$

When the test-time mixture weights $\pi \in \Delta_M$ (the probability simplex) are unknown, we consider the worst-case risk. The derivation relies on the property that a weighted average is bounded by the maximum element. For any $\pi \in \Delta_M$:

$$\begin{aligned}\sum_{m=1}^{M} \pi(m)\mathcal{R}_{\mathcal{D}_m}(h) &\leq \sum_{m=1}^{M} \pi(m) \left(\max_{k\in[M]} \mathcal{R}_{\mathcal{D}_k}(h) \right) \\ &= \left(\max_{k\in[M]} \mathcal{R}_{\mathcal{D}_k}(h) \right) \underbrace{\sum_{m=1}^{M} \pi(m)}_{=1} \\ &= \max_{1\leq m\leq M} \mathcal{R}_{\mathcal{D}_m}(h). \end{aligned} \tag{24}$$

This upper bound is achievable (tight) if the target domain concentrates entirely on the hardest style component (i.e., π is a one-hot vector). Thus, taking the supremum yields:

$$\sup_{\pi\in\Delta_M} \sum_{m=1}^{M} \pi(m)\mathcal{R}_{\mathcal{D}_m}(h) = \max_{1\leq m\leq M} \mathcal{R}_{\mathcal{D}_m}(h). \tag{25}$$

Eq. (25) shows that stable performance under arbitrary mixture shifts requires controlling the component-wise risks, rather than only minimizing the average source risk $\mathcal{R}_{\mathcal{D}_s}(h) = \sum_m \pi_s(m)\mathcal{R}_{\mathcal{D}_m}(h)$. In source-only few-shot DG, this requirement should be read as a design criterion rather than a guarantee: component-wise risks can only be influenced for styles that are present, partially present, or statistically recognizable from source observations. This also yields a refined risk-gap bound:

$$\begin{aligned}\left|\mathcal{R}_{\mathcal{D}_t}(h) - \mathcal{R}_{\mathcal{D}_s}(h)\right| &= \Big| \sum_{m=1}^{M} \pi_t(m)\mathcal{R}_{\mathcal{D}_m}(h) - \sum_{m=1}^{M} \pi_s(m)\mathcal{R}_{\mathcal{D}_m}(h) \Big| \\ &= \Big| \sum_{m=1}^{M} (\pi_t(m) - \pi_s(m))\mathcal{R}_{\mathcal{D}_m}(h) \Big| \\ &\leq \sum_{m=1}^{M} |\pi_t(m) - \pi_s(m)| \cdot \mathcal{R}_{\mathcal{D}_m}(h) \\ &\leq \sum_{m=1}^{M} |\pi_t(m) - \pi_s(m)| \cdot \max_{k\in[M]} \mathcal{R}_{\mathcal{D}_k}(h) \\ &= \left(\sum_{m=1}^{M} |\pi_t(m) - \pi_s(m)| \right) \cdot \max_{1\leq m\leq M} \mathcal{R}_{\mathcal{D}_m}(h) \\ &= \|\pi_t - \pi_s\|_1 \cdot \max_{1\leq m\leq M} \mathcal{R}_{\mathcal{D}_m}(h). \end{aligned} \tag{26}$$

The bound in Eq. (26) decomposes the performance degradation into two factors: (1) the magnitude of the mixture shift $\|\pi_t - \pi_s\|_1$, and (2) the worst-case component risk $\max_m \mathcal{R}_{\mathcal{D}_m}(h)$.

Prompt-space interpretation of SaMoE. The domain adaptation bound above is stated in terms of 0-1 classification risk. The following optimization-level argument is separate and uses a differentiable surrogate objective, such as cross-entropy, over prompt parameters. It is intended to explain why prompt composition can reduce style-specific errors; it does not assume that the 0-1 loss itself is smooth or strongly convex.

Let $\phi \in \Phi$ denote the complete prompt parameters induced by the learnable text prompts and visual prompt tokens. For each latent acquisition-style component, define the surrogate component loss

$$\mathcal{L}_m^{\mathrm{ce}}(\phi) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}_m}\big[\ell_{\mathrm{ce}}(h_\phi(x), y)\big], \tag{27}$$

where h_ϕ is the predictor obtained by injecting prompt parameters ϕ into the frozen VLM. Let $\phi_m^* = \arg\min_\phi \mathcal{L}_m^{\mathrm{ce}}(\phi)$ be the optimal prompt parameters for style m under this surrogate loss.

Assumption A.1 (Local surrogate geometry)**.** For each component m, assume the cross-entropy surrogate is locally μ-strongly convex and L-smooth with respect to prompt parameters in the region of interest around ϕ_m^*:

$$\frac{\mu}{2}\|\phi - \phi_m^*\|^2 \le \mathcal{L}_m^{\mathrm{ce}}(\phi) - \mathcal{L}_m^{\mathrm{ce}}(\phi_m^*) \le \frac{L}{2}\|\phi - \phi_m^*\|^2. \tag{28}$$

This assumption is imposed only on the differentiable surrogate objective and is separate from the 0-1 risk bound above.

1. Standard Prompt Tuning (SPT): SPT minimizes the average source surrogate loss: $\phi_{std} = \arg\min_\phi \sum_{m=1}^M \pi_s(m)\mathcal{L}_m^{\mathrm{ce}}(\phi)$. When the source mixture is dominated by an acquisition style a, the shared prompt can be biased toward the corresponding style optimum. To make this explicit, suppose $\|\phi_{std} - \phi_a^*\| \le \rho$. For a target style k whose optimum is far from ϕ_a^*, the reverse triangle inequality gives $\|\phi_{std} - \phi_k^*\| \ge [\|\phi_a^* - \phi_k^*\| - \rho]_+$, where $[z]_+ = \max(z, 0)$. Using the lower side of Eq. (28), the worst-case surrogate component loss is bounded as

$$\begin{aligned}\max_m \mathcal{L}_m^{\mathrm{ce}}(\phi_{std}) &\ge \mathcal{L}_k^{\mathrm{ce}}(\phi_{std}) \\ &\ge \mathcal{L}_k^{\mathrm{ce}}(\phi_k^*) + \frac{\mu}{2}\|\phi_{std} - \phi_k^*\|^2 \\ &\ge \mathcal{L}_k^{\mathrm{ce}}(\phi_k^*) + \frac{\mu}{2}\left([\|\phi_a^* - \phi_k^*\| - \rho]_+\right)^2. \end{aligned} \tag{29}$$

The term $\|\phi_a^* - \phi_k^*\|$ represents the structural mismatch between style-specific prompt optima. In biomedical imaging, acquisition shifts such as protocol or site changes can make this mismatch non-negligible, increasing the loss of a single shared prompt on underrepresented target styles.

2. SaMoE: SaMoE does not linearly mix model outputs. Instead, it composes prompt parameters before they are passed through the frozen nonlinear VLM. Abstracting the layer-wise prompt bank into prompt vectors, the effective prompt at injection layer ℓ is

$$\phi_{\mathrm{eff}}^{(\ell)}(x) = \phi_0^{(\ell)} + \sum_{r \in \mathcal{S}^{(\ell)}(x)} \alpha_r^{(\ell)}(x)\phi_r^{(\ell)}, \tag{30}$$

where r indexes prompt experts within the layer. Let $\phi_{\mathrm{eff}}(x) = \{\phi_{\mathrm{eff}}^{(\ell)}(x)\}_{\ell \in \mathcal{L}}$ denote the complete input-conditioned prompt parameters, and let $h_{\phi_{\mathrm{eff}}(x)}$ be the nonlinear predictor induced by injecting these prompts into the frozen VLM. For inputs from style component $\mathcal{D}_m$, suppose the routed prompt composition remains in the local region of Eq. (28) and approximates the style-specific optimum with error

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_m}\left[\|\phi_{\mathrm{eff}}(x) - \phi_m^*\|^2\right] \le \delta_m. \tag{31}$$

Applying the smoothness side of Eq. (28) to the input-conditioned prompt parameters yields

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_m}\left[\ell_{\mathrm{ce}}(h_{\phi_{\mathrm{eff}}(x)}(x), y)\right] \le \mathcal{L}_m^{\mathrm{ce}}(\phi_m^*) + \frac{L}{2}\delta_m. \tag{32}$$

Thus, the worst-case surrogate component loss for SaMoE is controlled by the quality of style-aware prompt approximation:

$$\max_m \mathbb{E}_{(x,y)\sim\mathcal{D}_m}\left[\ell_{\mathrm{ce}}(h_{\phi_{\mathrm{eff}}(x)}(x), y)\right] \le \max_m \left(\mathcal{L}_m^{\mathrm{ce}}(\phi_m^*) + \frac{L}{2}\delta_m\right). \tag{33}$$

Conclusion: Comparing Eq. (29) and Eq. (33), a single shared prompt can incur a style-mismatch penalty when the target emphasizes components far from the source-dominant style optimum, whereas SaMoE can reduce this penalty when the router and expert bank make δ_m small for the relevant acquisition styles. This is a prompt-parameter-space abstraction, not an output-level mixture: experts are mixed as prompt tokens as in Eq. (8), and the frozen transformer then induces the final nonlinear prediction.

B Additional Implementation Details

B.1 Top-K Routing Implementation and Soft Routing Comparison

Hard Top-K routing. Given the router scores $u(x) \in (0,1)^M$ for an input image x, we perform hard expert selection by taking the indices of the K largest scores using `topk`. Only the selected

Table 5: We report mean Base, Novel, and harmonic mean (HM) over three seeds on three datasets.Hard Top-K activates only K experts per image with re-normalized weights, while soft routing aggregates all experts with dense weights.

Routing	BTMRI			DermaMNIST			RETINA		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
Hard Top-K (M=4, K=2) (Ours)	**79.00**	**77.43**	**78.21**	**57.20**	**61.58**	**59.31**	**85.68**	**83.46**	**84.56**
Hard Top-K (M=4, K=3)	79.20	74.50	76.78	56.80	58.40	57.59	85.20	80.10	82.57
Hard Top-K (M=4, K=4)	79.10	72.80	75.82	56.50	56.20	56.35	84.80	78.50	81.53
Soft routing (M=4, dense)	78.90	71.50	75.02	56.10	55.10	55.60	84.50	77.20	80.68

experts are activated, and their scores are re-normalized to form mixture weights so that the routed prompt has a stable scale across samples. Although the selected indices are discrete, gradients still back-propagate through the selected scores and the weighted expert outputs, while the index selection itself is treated as non-differentiable in practice. This yields sparse computation with K active experts per image and encourages experts to specialize to different acquisition-induced style modes.

Soft routing. As a dense alternative, soft routing removes the hard selection and aggregates all experts with continuous weights, e.g., using a softmax gate over M experts. Dense aggregation is fully differentiable but activates all experts for every image, increasing computation and making experts more prone to co-adaptation, which can blur expert specialization under strong style shifts.

Comparison. Table 5 presents a comprehensive comparison between hard Top-K routing (with varying K) and soft routing. The results indicate that hard routing with moderate sparsity ($K = 2$) consistently outperforms dense soft routing. Soft routing achieves suboptimal performance (e.g., HM 75.02% vs. 78.21% on BTMRI), likely because activating all experts for every sample encourages them to learn redundant, average features rather than specializing in distinct styles. Furthermore, as we increase K in the hard routing setting (from 2 to 4), the performance gradually degrades towards that of soft routing, confirming that sparsity is a key factor for the success of our style-aware expert bank. Therefore, we adopt $M = 4, K = 2$ as the default configuration, striking the best balance between expert specialization and representation capacity.

B.2 Additional Adapted Baseline Comparisons

We further compare SaMoE with adapter/probe baselines and prompt-mixture baselines adapted to our biomedical VLM setting. To make the comparison complete, Table 6 includes all methods from the main tables and reports the full results on representative base-to-novel, cross-protocol, and cross-site settings under the same 16-shot source protocol. SaMoE consistently outperforms these alternatives, suggesting that the gains are not merely due to adding adaptation capacity or using generic prompt mixtures, but stem from acquisition-style routing and visual prompt composition.

B.3 When is SaMoE Less Effective?

We further analyze when SaMoE brings smaller gains. Since SaMoE is designed to handle acquisition-induced style variation, it should be most beneficial when the source-target shift is dominated by acquisition style discrepancy. To quantify this effect, we compute simple domain-level image statistics as style proxies. For each image x, we define

$$v_2(x) = [\mu_{\text{gray}}(x), \sigma_{\text{gray}}(x)],$$
$$v_8(x) = [\mu_R(x), \sigma_R(x), \mu_G(x), \sigma_G(x), \mu_B(x), \sigma_B(x), \mu_{\text{gray}}(x), \sigma_{\text{gray}}(x)].$$

For a domain D, we average image-level statistics as $v(D) = \frac{1}{|D|} \sum_{x \in D} v(x)$, and define the source-target style gap as

$$\text{StyleGap}(S, T) = \|v(S) - v(T)\|_2.$$

Here, v_2 captures coarse grayscale intensity and contrast changes, while v_8 additionally includes RGB-channel statistics as a richer proxy for acquisition style.

Analysis. Table 7 shows that SaMoE can be less effective when the source-target style gap is mild. For example, FLAIR has the smallest style discrepancy under both v_2 and v_8, and SaMoE only

Table 6: **Additional comparison with adapted baselines on representative settings.** We include all methods from the main tables and additional adapter/probe and prompt-mixture baselines. For base-to-novel generalization, we report Base, Novel, and HM on BTMRI and DermaMNIST. For cross-domain DG, we report the source-domain accuracy and each unseen target-domain accuracy. All adapted baselines follow the same 16-shot source setting.

Method	Base-to-Novel Generalization						Cross-Protocol DG					Cross-Site DG				
	BTMRI			DermaMNIST			Src.	FLAIR	T1W	T1W-CE	T2W	Src.	D7pt_C	D7pt_D	MILK10k	PH2
	Base	Novel	HM	Base	Novel	HM	BTMRI	Gazi Protocols				Derma	Unseen Sites			
CLIP	48.25	56.05	51.86	25.24	27.43	26.29	21.78	51.44	54.11	61.47	51.44	12.77	22.65	22.34	24.42	24.00
CoOp	60.92	50.74	55.37	23.30	52.59	32.29	67.73	56.00	57.67	71.68	62.33	31.55	28.48	27.57	20.71	34.50
ProGrad	52.33	56.12	54.16	32.69	26.85	29.48	59.74	57.00	60.84	77.60	62.67	30.01	25.31	23.95	33.49	34.83
KgCoOp	55.58	56.71	56.14	26.62	15.84	19.86	55.72	53.11	51.67	58.78	53.78	15.83	22.43	24.39	14.87	30.00
MaPLe	77.67	57.07	65.80	54.37	32.90	40.99	80.70	55.11	55.22	70.25	57.67	53.82	27.79	32.71	28.31	34.33
PromptSRC	60.50	73.28	66.28	42.64	20.78	27.94	58.82	44.00	54.78	49.10	57.56	49.92	32.17	32.49	27.34	25.83
DCPL	72.33	45.84	56.11	49.51	24.40	32.69	80.88	51.78	54.00	68.81	63.00	55.70	24.98	30.92	23.82	39.83
TCP	54.25	41.62	47.10	32.60	46.85	38.45	57.37	50.00	50.00	44.64	50.00	24.99	17.83	18.91	12.46	27.93
BiomedCoOp	69.79	54.12	60.97	44.42	18.66	26.28	68.53	51.03	58.45	72.76	61.78	45.70	30.63	32.75	20.44	32.50
CLIP-Adapter [8]	51.00	56.16	53.46	28.40	31.32	29.79	54.28	58.33	53.33	59.68	55.00	21.60	24.62	26.62	16.57	24.32
TIP-Adapter [47]	63.25	56.16	59.50	45.87	26.87	33.89	71.64	54.67	56.67	59.57	64.33	28.28	22.54	20.81	21.21	26.74
LP++ [10]	58.00	56.05	57.01	28.16	26.74	27.43	73.97	55.38	57.00	69.43	65.00	32.47	22.65	20.70	20.03	25.51
SMoP [5]	49.63	47.76	48.68	16.02	14.00	14.94	63.12	50.00	50.00	19.35	50.00	30.29	24.60	24.60	26.50	28.00
MoPE [13]	61.25	67.50	64.22	27.43	19.96	23.11	68.84	55.00	55.67	64.52	51.67	30.17	26.26	28.81	21.10	20.00
SaMoE	**79.00**	**77.43**	**78.21**	**57.20**	**61.58**	**59.31**	**82.86**	**58.67**	**62.33**	**80.11**	**68.89**	**57.91**	**34.91**	**37.90**	**33.84**	**46.50**

Table 7: **Style-gap failure study on cross-protocol MRI.** We report source-target style gaps from BTMRI to each Gazi target protocol and the gain of SaMoE over the strongest non-SaMoE baseline among all compared methods in Table 6. Smaller gains indicate cases where SaMoE is less effective.

Target	StyleGap (v_2)	StyleGap (v_8)	Best Baseline	SaMoE	Gain
FLAIR	0.0616	0.1232	58.33	58.67	+0.34
T2W	0.0653	0.1305	65.00	68.89	+3.89
T1W-CE	0.0838	0.1676	77.60	80.11	+2.51
T1W	0.0902	0.1805	60.84	62.33	+1.49

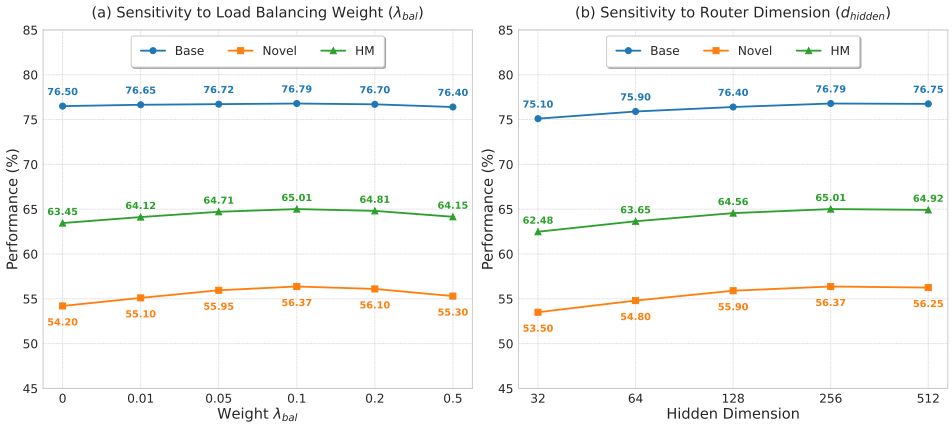


Figure 5: Post-hoc hyperparameter sensitivity analysis on Base-to-Novel generalization. We report the Harmonic Mean (HM) of Base and Novel accuracies to assess robustness after fixing the main experimental configuration. **(a)** Impact of the load balancing weight λ_{bal}. **(b)** Impact of the router's hidden dimension d_{hidden}.

improves over the strongest baseline by 0.34%. This suggests that when the target protocol is already close to the source in low-level acquisition statistics, a style-adaptive expert composition may provide limited additional benefit over strong prompt or adapter baselines. The gains are also not strictly monotonic with these simple pixel-level proxies: T1W has the largest measured v_8 gap but only a moderate gain. This indicates that v_2 and v_8 are useful but incomplete approximations of biomedical acquisition style, and cannot capture all clinically relevant factors such as protocol-specific contrast mechanisms, anatomical visibility, or residual dataset bias. Overall, SaMoE is expected to be less effective when acquisition style is not the dominant source of transfer difficulty, when the style gap is mild, or when semantic/task-level differences dominate the domain shift.

Table 8: **Seed stability on Base-to-Novel generalization.** We report means with standard deviations as subscripts over three independent random seeds, averaged across 11 biomedical datasets.

Method	Base Acc.	Novel Acc.
BiomedCoOp	$70.84_{\pm 2.62}$	$36.07_{\pm 8.83}$
CoOp	$62.14_{\pm 6.02}$	$40.71_{\pm 8.00}$
DCPL	$75.80_{\pm 3.30}$	$39.54_{\pm 10.54}$
KgCoOp	$56.43_{\pm 4.83}$	$41.19_{\pm 8.46}$
MaPLe	$76.37_{\pm 3.13}$	$39.52_{\pm 8.76}$
ProGrad	$61.96_{\pm 5.26}$	$37.24_{\pm 7.01}$
PromptSRC	$63.52_{\pm 5.22}$	$45.50_{\pm 10.28}$
TCP	$58.87_{\pm 5.71}$	$40.62_{\pm 10.22}$
SaMoE	$\mathbf{76.79}_{\pm 2.44}$	$\mathbf{56.37}_{\pm 5.91}$

B.4 Hyperparameter Sensitivity Analysis

To verify the robustness of SaMoE to hyperparameter variations, we conduct a post-hoc sensitivity analysis on the Base-to-Novel generalization benchmark. Specifically, we investigate the impact of the load balancing weight λ_{bal} and the router's hidden dimension d_{hidden} on the harmonic mean (HM) of Base and Novel accuracies. This analysis is used only to characterize sensitivity after the main configuration is fixed, rather than as a model-selection protocol on novel classes. The results are illustrated in Figure 5.

Sensitivity to Load Balancing Weight (λ_{bal}). The load balancing loss $\mathcal{L}_{bal}$ plays a critical role in preventing expert collapse by encouraging a balanced utilization of the expert bank. As shown in Figure 5(a), the performance follows an inverted-U trend. When $\lambda_{bal} = 0$ (i.e., no load balancing), the model achieves an HM of 63.45%, suggesting that the router may converge to a trivial solution where only a few experts are active, limiting the diversity of style adaptation. Increasing λ_{bal} to 0.1 yields the highest observed performance in this sweep (HM 65.01%), confirming that moderate regularization effectively encourages diverse expert specialization. However, further increasing λ_{bal} beyond 0.1 leads to performance degradation (e.g., 64.15% at $\lambda_{bal} = 0.5$). This is because an excessively large balancing weight forces the router towards a uniform distribution, overriding the style-specific selection logic and reducing the experts' ability to specialize in distinct acquisition shifts.

Sensitivity to Router Hidden Dimension (d_{hidden}). The hidden dimension determines the capacity of the router to process the style statistics $s(x) = [\mu(x); \sigma^2(x); h_{cls}(x)]$. As shown in Figure 5(b), increasing d_{hidden} from 32 to 256 results in a consistent performance improvement, with HM rising from 62.48% to a peak of 65.01%. This indicates that a sufficient feature capacity is necessary to encode complex acquisition-induced style variations from the low-level statistics. When d_{hidden} is further increased to 512, the performance saturates and slightly drops to 64.92%. This suggests that $d_{hidden} = 256$ provides a sufficient information bottleneck for our compact expert bank, while larger dimensions do not yield further gains and may introduce minor redundancy.

B.5 Seed Stability Analysis

To assess whether the base-to-novel improvements are stable across random seeds, we report the mean and standard deviation over three independent runs in Table 8. The results are averaged across the 11 datasets used in the base-to-novel benchmark. SaMoE achieves the best average Base and Novel accuracy while maintaining a moderate standard deviation compared with strong prompt-learning baselines. This indicates that the observed gain is not caused by a single favorable seed. Together with the domain-dependent routing analyses in Appendix D, these results support that SaMoE's performance and routing behavior are stable across random initializations.

Table 9: Prompting complexity and inference throughput of prompt-learning methods. "Params" reports trainable parameters and "%CLIP" is relative to the frozen CLIP backbone. FPS is measured under batch sizes 1/4/16. "HM" is the three-seed average harmonic mean of Base and Novel accuracies from Table 1. SaMoE (Independent) uses layer-wise ExpertBanks with a shared router, while SaMoE (Shared) further shares the ExpertBank across layers.

Method	**Params**	**% CLIP**	**FPS (BS=1)**	**FPS (BS=4)**	**FPS (BS=16)**	**HM**
CoOp	4.1K	0.003	11.93	44.23	156.18	49.19
ProGrad	4.1K	0.003	13.96	51.84	178.61	46.52
KgCoOp	267.3K	0.178	13.48	48.03	172.68	47.62
MaPLe	3.63M	2.421	11.72	36.12	122.25	52.09
PromptSRC	5.1K	0.003	10.68	40.05	125.09	53.02
DCPL	464.7K	0.310	9.62	14.02	14.86	51.97
TCP	334.0K	0.223	9.48	38.05	154.53	48.07
BiomedCoOp	4.1K	0.003	10.17	44.61	125.59	47.80
SaMoE (Independent)	15.91M	10.607	8.22	32.67	96.94	**65.01**
SaMoE (Shared)	2.59M	1.727	12.37	46.52	137.83	<u>57.91</u>

Table 10: Comparison of Methods: ℓ_2 Distance and Cosine Similarity

Method	ℓ_2 Distance ↑			Cosine Similarity ↓		
	BTMRI	DermaMNIST	RETINA	BTMRI	DermaMNIST	RETINA
MaPLe	0.8331	0.8899	0.5636	0.6530	0.6011	0.8412
PromptSRC	0.2621	0.2218	0.2863	0.9656	0.9727	0.9590
DCPL	0.4170	0.5885	0.4932	0.9131	0.8091	0.8784
BiomedCoOp	0.5884	0.6922	0.4084	0.8269	0.7417	0.9166
SaMoE	0.8678	0.9243	0.9503	0.6235	0.5708	0.5485

C Prompt Analysis

C.1 Prompting Complexity and Inference Throughput

Table 9 compares representative prompt-learning methods in terms of trainable parameters and inference throughput. Since all methods adapt a frozen CLIP backbone, the inference cost differences mainly stem from how prompts are injected (shallow vs. deep) and whether additional modules are introduced. Shallow prompt tuning (e.g., CoOp/ProGrad) adds only a few thousand trainable parameters and incurs limited overhead. Methods with deeper prompt injection or additional prompt transformations (e.g., MaPLe, DCPL) can increase both parameter count and runtime due to repeated prompt application across transformer layers. For SaMoE, the independent variant uses layer-wise ExpertBanks with a shared router, increasing trainable parameters and introducing routing/expert aggregation at each layer, which leads to the largest overhead. In contrast, the shared variant further amortizes the ExpertBank across layers, substantially reducing trainable parameters while maintaining competitive throughput. Overall, these results highlight a practical trade-off between adaptation capacity and efficiency, and show that SaMoE can be implemented in an efficient manner via sharing while still retaining strong generalization performance.

C.2 Text Prompt Embedding Analysis

Motivation. To better understand how different prompt-learning methods shape the textual prototype space, we analyze the separability of the *class text embeddings*. In CLIP-style classification, each class y is represented by a single prompted text prototype w_y, which is shared by all image samples of that class. Therefore, the geometry of $\{w_y\}_{y\in\mathcal{Y}}$ provides a direct and quantitative view of how prompts encode class discrimination.

Metrics. For each dataset, we compute all pairwise similarities between class prototypes and report the average statistics. Specifically, we measure (i) the average pairwise cosine similarity (lower is better, indicating less collapse and better angular separation), and (ii) the average pairwise normalized ℓ_2 distance (higher is better, indicating larger margins in the embedding space). The normalization

Table 11: **Supplementary cross-dataset prompt diversity of SaMoE.** We report ℓ_2 distance and cosine similarity between learned text prompts and composed vision prompts across dataset pairs. Since the compared datasets differ in semantics and label spaces, this table is used only to diagnose prompt diversity rather than to claim acquisition-style adaptation.

Dataset Pair	Text Prompts		Vision Prompts	
	ℓ_2 Distance ↑	Cosine Sim. ↓	ℓ_2 Distance ↑	Cosine Sim. ↓
BTMRI → DermaMNIST	1.1549	0.3322	1.1361	0.3546
BTMRI → RETINA	1.1874	0.2938	1.0393	0.4600
DermaMNIST → RETINA	1.1855	0.2965	1.0837	0.4128

follows the standard practice of comparing embeddings after ℓ_2-normalization, so that distances primarily reflect relative geometry rather than scale.

Results and discussion. Table 10 shows that SaMoE consistently yields *larger inter-class distances* and *lower cosine similarities* on BTMRI, DermaMNIST, and RETINA compared to strong prompt-learning baselines. This indicates that SaMoE produces a more discriminative text-prototype space with reduced prototype collapse, which is consistent with improved generalization when images undergo acquisition-induced style shifts. Intuitively, style shifts can perturb image embeddings and make classification more sensitive to the margin and angular separation among text prototypes. By composing prompts with style-aware experts and routing, SaMoE encourages the prompt representation to remain discriminative across diverse styles, leading to improved robustness in downstream transfer.

C.3 Supplementary Cross-Dataset Prompt Diversity

As a supplementary diagnostic, we also report cross-dataset prompt distances for SaMoE in Table 11. This analysis measures whether independently trained prompts collapse to a nearly identical configuration across datasets. However, because BTMRI, DermaMNIST, and RETINA differ in organs, label spaces, and visual semantics, these cross-dataset distances should be interpreted only as evidence of prompt diversity, not as isolated evidence of acquisition-style adaptation. The controlled analysis below therefore focuses on cross-protocol and cross-site settings where the task family is fixed and acquisition conditions are the main source of variation.

C.4 Controlled Prompt Composition under Acquisition Shifts

Cross-dataset prompt distances can be confounded by differences in organs, label spaces, and class semantics, so we do not use them as evidence for acquisition-style adaptation. Instead, we analyze prompt composition in controlled transfer settings where the task family is fixed and the main variation comes from acquisition conditions. Specifically, cross-protocol MRI shares the brain MRI task family but changes imaging protocols, while cross-site dermoscopy shares the skin-lesion task family but changes acquisition sites. For a model adapted on the source domain, we inspect the normalized expert weights that compose the visual prompts for each unseen target domain. The domain-dependent activation distributions in Fig. 6 and the layer-wise routing paths in Fig. 7 show that different protocols and sites induce different expert compositions. This controlled analysis supports style-aware prompt composition without relying on prompt comparisons across semantically different datasets.

D Qualitative Analysis of Expert Routing

D.1 Expert Utilization and Routing Behavior

We further visualize the routing behavior of SaMoE to examine whether the router adapts expert composition under acquisition-induced style shifts. We consider two evaluation settings. In cross-protocol MRI, the model is adapted on BTMRI and evaluated on four unseen Gazi protocols (T1W, T2W, FLAIR, and T1W-CE). In cross-site dermoscopy, the model is adapted on DermaMNIST and evaluated on four unseen target sites (Derm7pt_D, Derm7pt_C, PH2, and MILK10k). For each

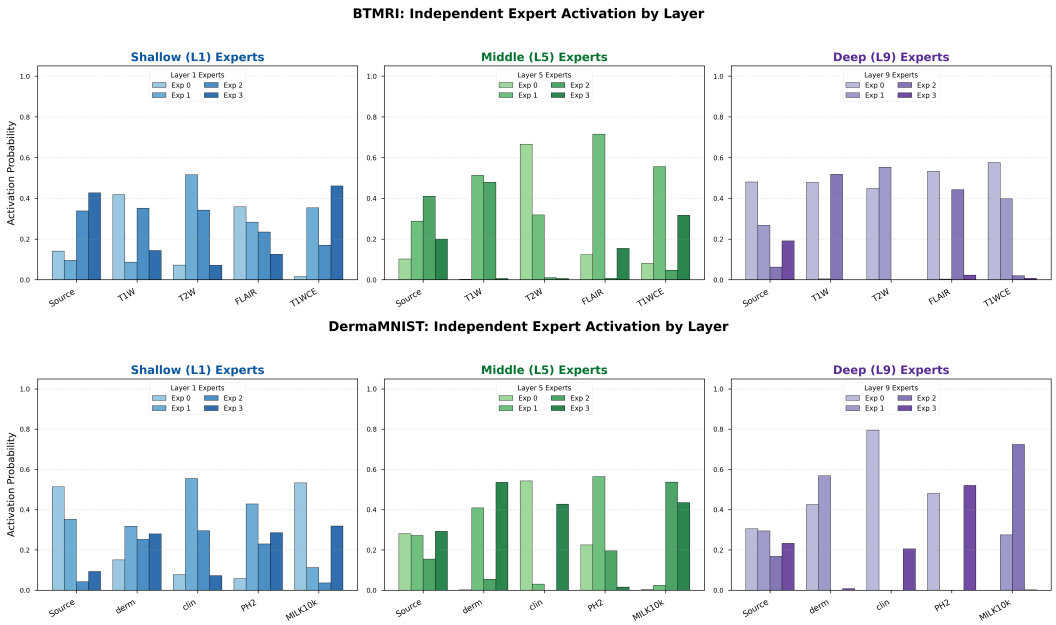


Figure 6: Expert activation distributions under domain shifts. We report the average routing weight assigned to each expert at three prompt injection depths (L1, L5, L9). Results are shown for cross-protocol MRI with BTMRI as the source and Gazi protocols as targets, and for cross-site dermoscopy with DermaMNIST as the source and four unseen target sites. Different target domains allocate routing mass to different experts, and the preference changes with depth, supporting style-adaptive expert composition.

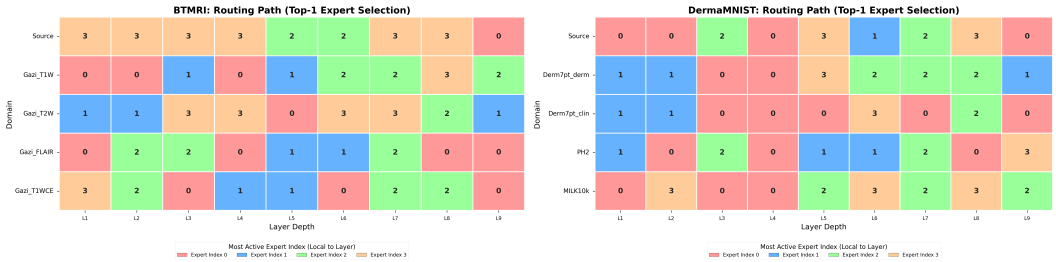


Figure 7: Layer-wise routing paths under unseen domains. Each cell shows the expert index with the largest average routing weight at a given layer and domain, computed on the corresponding test set. Expert indices are local to each layer since expert banks are independent across layers. Target protocols and sites follow different routing paths from the source, indicating that routing decisions respond to acquisition-induced style differences.

domain, we forward all test images through the trained model and record the routed expert weights at each prompt injection layer after Top-K selection and normalization.

For each injection layer ℓ and expert m, we compute the average routing mass

$$\bar{\alpha}_{d,m}^{(\ell)} = \frac{1}{|T_d|} \sum_{x \in T_d} \alpha_m^{(\ell)}(x),$$

where T_d is the test set of domain d and non-selected experts have weight zero. Fig. 6 visualizes $\bar{\alpha}_{d,m}^{(\ell)}$ at three representative depths (L1, L5, L9). We also summarize the layer-wise dominant expert by $\hat{m}_d^{(\ell)} = \arg\max_m \bar{\alpha}_{d,m}^{(\ell)}$ and visualize $\hat{m}_d^{(\ell)}$ across all layers in Fig. 7. Across both benchmarks, different protocols and sites exhibit distinct expert mixtures and routing paths, indicating that the router adjusts prompt composition according to domain-dependent style rather than relying on a fixed expert usage pattern.

Table 12: **Analysis on Router Input Representations.** We compare different style cues on the Base-to-Novel generalization benchmark (averaged across 11 datasets). Ours ($h_{\text{cls}} + [\mu, \sigma^2]$) achieves the best balance between performance and efficiency.

Router Input	Params (Router)	Base Acc.	Novel Acc.	HM
CLS Only (h_{cls})	$\sim$ 0.5M	75.80	52.68	62.15
Stats Only ($[\mu, \sigma^2]$)	$\sim$ 1.0M	76.10	54.89	63.80
Gram Matrix ($H^\top H$)	$\sim$ 200M	74.50	52.35	61.50
Fourier Amp. ($\mathcal{F}_{amp}$)	$\sim$ 0.5M	75.90	53.95	63.10
Ours ($[h_{\text{cls}}; \mu; \sigma^2]$)	$\sim$ 1.5**M**	**76.79**	**56.37**	**65.01**

Table 13: **Analysis on Load Balancing Strategies.** We report the Base-to-Novel performance (HM) and the Expert Utilization Rate (using $M = 4$ experts and Top-$K = 2$ routing). Expert Utilization is defined as the percentage of experts that are activated for at least 10% of the test samples, averaged across datasets and injection layers. Our approach achieves the best trade-off between specialization (high HM) and balanced utilization.

Balancing Strategy	Expert Utilization (%)	Base Acc.	Novel Acc.	HM
No Balancing (w/o $b, \mathcal{L}_{\text{bal}}$)	50.0% (Collapse)	**77.10**	35.40	48.52
Soft Gating (Dense)	100% (Always)	76.50	48.20	59.14
Noisy Top-K	85.5%	75.80	51.30	61.19
Loss Only (w/o b)	92.0%	76.20	53.50	62.86
Ours ($b + \mathcal{L}_{\text{bal}}$)	**98.5%**	76.79	**56.37**	**65.01**

D.2 Analysis on Router Input Features

To validate our design choice of using both the CLS token and feature statistics ($[\mu, \sigma^2]$) as style cues for the router, we compare SaMoE with variants using alternative routing inputs. As shown in Table 12, we consider: (1) **CLS Only:** Utilizing solely the global class token h_{cls}, which primarily encapsulates high-level semantic information. (2) **Statistics Only:** Employing the first- and second-order feature statistics $[\mu, \sigma^2]$ as a compact representation of the global style distribution, akin to AdaIN style codes. (3) **Gram Matrix:** Adopting the flattened Gram matrix of patch tokens, which captures feature correlations and serves as a standard style cues in neural style transfer, albeit with significantly higher computational complexity. (4) **Fourier Amplitude:** Leveraging the amplitude spectrum of the feature maps, a frequency-domain representation often used to disentangle style from content in domain generalization.

Analysis. The results validate our design:

- **Complementarity:** Using *Stats Only* outperforms *CLS Only* in novel class generalization (54.89% vs. 52.68%), confirming that acquisition shifts (e.g., contrast/noise) are better captured by distributional statistics than semantic tokens. However, combining them (*Ours*) yields the highest performance, suggesting that the CLS token provides necessary semantic context to disambiguate routing when styles are similar but content differs.
- **Efficiency vs. Expressiveness:** While the *Gram Matrix* is a rich style cues, its high dimensionality leads to overfitting (lower generalization) and excessive parameter costs in the router.
- **Stability:** Our statistic-based style cues ($[\mu, \sigma^2]$) offers a robust and compact summary of global style, outperforming frequency-based methods (*Fourier Amp.*) which can be sensitive to high-frequency artifacts common in medical imaging.

D.3 Analysis on Expert Load Balancing Mechanisms

To prevent expert collapse, where the router trivially selects the same experts for all inputs. We employ a dual mechanism consisting of a learnable bias term b and a load-balancing loss $\mathcal{L}_{\text{bal}}$. We compare this strategy against alternative routing and balancing mechanisms in Table 13. Specifically,

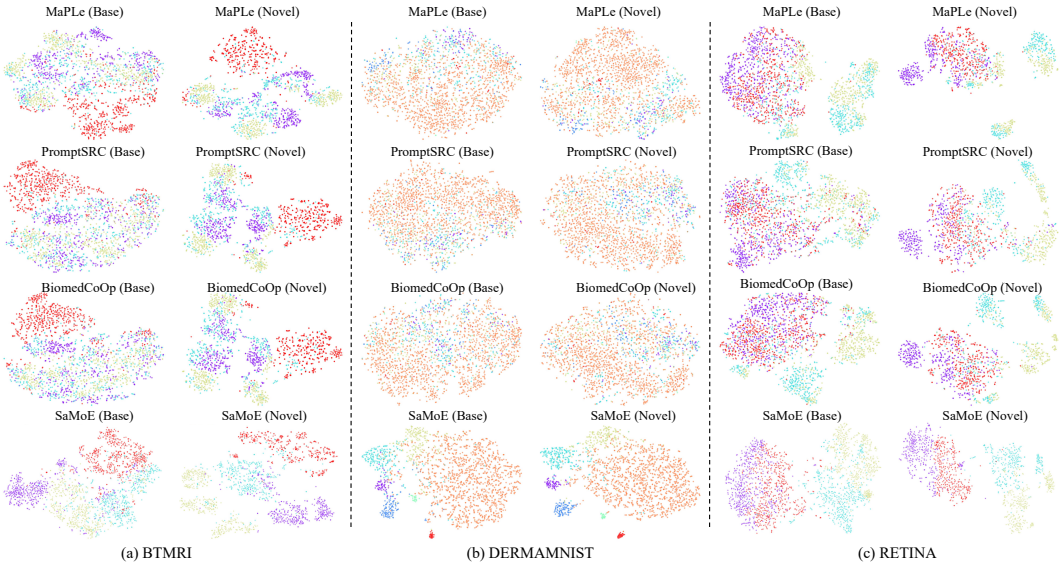


Figure 8: t-SNE visualization of prompted CLIP image embeddings for different prompt-learning methods. Each row corresponds to a method and each pair of panels shows the base and novel test samples for a dataset, with points colored by class. SaMoE yields clearer class-wise grouping and more consistent structure across base and novel splits, suggesting improved semantic separability under style shifts.

we evaluate: (1) **No Balancing:** Removing both b and $\mathcal{L}_{\text{bal}}$. (2) **Soft Gating:** Replacing Top-K selection with a dense Softmax weighted sum over all experts (dense computation). (3) **Noisy Top-K:** Adding Gaussian noise to the router logits to encourage exploration, a common technique in large-scale MoEs. (4) **Loss Only:** Using $\mathcal{L}_{\text{bal}}$ but removing the learnable bias b.

Analysis. As shown in Table 13, removing balancing mechanisms (*No Balancing*) leads to severe expert collapse (utilization $\approx 50\%$, i.e., the same two experts are repeatedly selected under Top-K=2 routing), which degrades generalization to Novel classes (HM 48.52%) as the model reverts to a nearly fixed prompt composition. While *Soft Gating* utilizes all experts, it lacks sparsity and performs worse (HM 59.14%), likely due to expert co-adaptation blurring the distinct style handling. *Noisy Top-K* improves utilization but introduces stochasticity that is suboptimal for stable few-shot adaptation. Crucially, comparing *Loss Only* (HM 62.86%) with *Ours* (HM 65.01%) highlights the importance of the learnable bias b. The bias term acts as a domain-agnostic "prior" that absorbs the base probability differences between experts, freeing the router's weights to focus solely on input-dependent style matching. Combined with $\mathcal{L}_{\text{bal}}$, this strategy ensures diverse and specialized expert coverage (98.5% utilization), yielding the highest generalization performance.

E Visual Interpretability

We provide a qualitative visualization of the learned representation using t-SNE embeddings of image features (see Fig. 8). For each method, we extract the CLIP image embedding of test samples after prompting and project the features into 2D with t-SNE, reporting results separately for base and novel splits on representative datasets. Compared with prior prompt-learning baselines, SaMoE tends to produce more compact and class-consistent clusters, with reduced overlap between categories, and the cluster structure remains more stable when moving from base to novel classes. This behavior is consistent with SaMoE's style-aware routing, which adapts prompts to acquisition-induced style shifts while preserving cross-modal alignment in the frozen backbone.

F Datasets and Benchmarks

We evaluate biomedical domain generalization on a diverse suite of datasets spanning multiple organs and imaging modalities. Table 14 summarizes all datasets used in our experiments, including

Table 14: Datasets used in our experiments across three settings: base-to-novel class generalization, dermatoscopy cross-site DG, and MRI cross-protocol DG. For each dataset, we report modality, anatomical region, label space, and train/val/test sizes; target-domain datasets in cross-site and cross-protocol benchmarks are evaluation-only.

Modality	Organ	Name	Label Space	Train	Val	Test
Base-to-Novel Experiments						
CT	Kidney	CTKidney	Cyst, stone, tumor, normal	6221	2487	3738
Dermatoscopy	Skin	DermaMNIST	Actinic keratosis, basal cell carcinoma, benign keratosis, dermatofibroma, nevus, melanoma, vascular lesion	7006	1003	2005
Fundus	Retina	RETINA	Cataract, diabetic retinopathy, glaucoma, normal	2108	841	1268
Histopath.	Colorectal	CHMNIST	Adipose, complex stroma, debris, empty background, immune cells, mucosal glands, simple stroma, tumor epithelium	2496	1000	1504
Histopath.	Lung/Col.	LUNGCOLON	Colon adenocarcinoma, colon benign tissue, lung adenocarcinoma, lung benign tissue, lung squamous carcinoma	12500	5000	7500
MRI	Brain	BTMRI	Glioma, meningioma, normal brain, pituitary tumor	2854	1141	1717
OCT	Retina	OCTMNIST	Choroidal neovascularization, drusen, diabetic macular edema, normal	97476	10832	1000
Ultrasound	Breast	BUSBRA	Benign, probably benign, suspicious, highly suggestive of malignancy	937	375	563
X-Ray	Chest/Lung	COVID	COVID-19, lung opacity, normal, viral pneumonia	10582	4232	6351
X-Ray	Knee	KneeXray	No, doubtful, minimal, moderate, severe osteoarthritis	5778	826	1656
Endoscopy	Colon	Kvasir	Dyed lifted polyps, normal cecum, esophagitis, dyed resection margins, normal pylorus, normal Z-line, polyps, ulcerative colitis	2000	800	1200
Dermatoscopy Cross-site DG Experiments						
Dermatoscopy	Skin	DermaMNIST	Actinic keratosis, basal cell carcinoma, benign keratosis, dermatofibroma, nevus, melanoma, vascular lesion	7006	1003	2005
Dermatoscopy	Skin	Derm7pt-D	Basal cell carcinoma, dermatofibroma, nevus, melanoma, vascular lesion	1011	0	1011
Clinical Photo	Skin	Derm7pt-C	Basal cell carcinoma, dermatofibroma, nevus, melanoma, vascular lesion	1011	0	1011
Dermatoscopy	Skin	PH2	Melanocytic Nevus, Melanoma	200	0	200
Dermatoscopy	Skin	MILK10k	Benign, Malignant	10000	0	10000
MRI Cross-Protocol DG Experiments						
MRI	Brain	BTMRI	Glioma, meningioma, normal brain, pituitary tumor	2854	1141	1717
MRI	Brain	Gazi (T1W)	Glioma Tumor, Normal Brain	300	0	300
MRI	Brain	Gazi (T2W)	Glioma Tumor, Normal Brain	300	0	300
MRI	Brain	Gazi (FLAIR)	Glioma Tumor, Normal Brain	300	0	300
MRI	Brain	Gazi (T1W-CE)	Glioma Tumor, Normal Brain	186	0	186

modality, anatomical region, label space, and the train/val/test splits. We consider three evaluation settings. (i) *Base-to-Novel*: each dataset is split into disjoint base and novel classes; we adapt the model with a 16-shot labeled support set from base classes and evaluate on both base and novel test samples. (ii) *Dermatoscopy Cross-site DG*: we adapt on DermaMNIST (source) and directly test on four unseen target sites (Derm7pt_derm, Derm7pt_clin, PH2, MILK10k) without target-specific

training. (iii) *MRI Cross-Protocol DG*: we adapt on BTMRI (source protocol) and evaluate on four unseen Gazi protocols (T1W/T2W/FLAIR/T1W-CE). Targets in the two DG benchmarks are used for evaluation only. *(The above protocol descriptions follow our main experimental setting.)*

G Additional Evaluation of Generalization and Routing

We report additional evidence on class-balanced performance, backbone generality, capacity-matched controls, acquisition-sensitive routing, and controlled compound shifts. Unless stated otherwise, the results use the 16-shot setting and three training seeds. Accuracy and Macro-F1 are percentages; uncertainties are standard deviations over three training seeds. Target averages are unweighted averages across the target domains in each setting.

G.1 Class-Balanced Evaluation

Macro-F1 averages the class-specific F1 scores and complements accuracy under class imbalance. Tables 15 and 16 report both metrics for the main baselines. Frozen CLIP is deterministic and has no training-seed uncertainty. SaMoE attains novel-class Macro-F1 of 45.13, versus 36.97 for PromptSRC, and target Macro-F1 of 36.01, 58.56, and 14.02 in cross-site, cross-protocol, and compound cross-modality transfer, respectively. These results support improved class-balanced performance in addition to overall accuracy.

Table 15: Base-to-novel generalization averaged across 11 biomedical datasets. Accuracy and Macro-F1 are shown separately for readability.

Accuracy (%)

Method	Base	Novel
CLIP	43.00	38.63
CoOp	$62.14_{\pm 6.02}$	$40.71_{\pm 8.00}$
ProGrad	$61.96_{\pm 5.26}$	$37.24_{\pm 7.01}$
KgCoOp	$56.43_{\pm 4.83}$	$41.19_{\pm 8.46}$
MaPLe	$76.37_{\pm 3.13}$	$39.52_{\pm 8.76}$
PromptSRC	$63.52_{\pm 5.22}$	$45.50_{\pm 10.28}$
DCPL	$75.80_{\pm 3.30}$	$39.54_{\pm 10.54}$
TCP	$58.87_{\pm 5.71}$	$40.62_{\pm 10.22}$
BiomedCoOp	$70.84_{\pm 2.62}$	$36.07_{\pm 8.83}$
SaMoE	$76.79_{\pm 2.44}$	$56.37_{\pm 5.91}$

Macro-F1 (%)

Method	Base	Novel
CLIP	33.29	34.42
CoOp	$59.05_{\pm 5.18}$	$28.14_{\pm 1.82}$
ProGrad	$55.99_{\pm 2.75}$	$27.18_{\pm 5.05}$
KgCoOp	$53.82_{\pm 3.80}$	$33.19_{\pm 4.51}$
MaPLe	$74.06_{\pm 1.93}$	$32.67_{\pm 4.39}$
PromptSRC	$55.21_{\pm 4.38}$	$36.97_{\pm 2.49}$
DCPL	$74.06_{\pm 1.87}$	$32.95_{\pm 5.49}$
TCP	$55.12_{\pm 3.58}$	$32.67_{\pm 3.22}$
BiomedCoOp	$67.76_{\pm 1.95}$	$27.47_{\pm 4.45}$
SaMoE	$76.05_{\pm 1.07}$	$45.13_{\pm 4.00}$

G.2 Generalization Across Vision-Language Backbones

We extend SaMoE to CLIP ViT-L/14, SigLIP so400m-patch14-224 [46], and SigLIP2 so400m-patch14-384 [37]. CLIP ViT-L/14 and SigLIP use 224×224 inputs; SigLIP2 uses its native 384×384 resolution. Data splits, 16-shot adaptation, SaMoE configurations, and optimization follow the main setting. Raw denotes each frozen backbone with deterministic dataset-specific text templates. CLIP ViT-L/14 uses the CLS token as the global descriptor. For SigLIP and SigLIP2, we use their native attention-pooled representation in place of CLS, retaining patch statistics and sparse routing. Experiments use FP32 on eight NVIDIA RTX 4090 GPUs (24 GB), with seeds 1, 2, and 3.

Table 16: Cross-domain accuracy and Macro-F1. Src. denotes source accuracy or Macro-F1, and Tgt. denotes the unweighted target-domain average of the corresponding metric.

Accuracy (%)

Method	Site Src.	Site Tgt.	Protocol Src.	Protocol Tgt.	Modality Src.	Modality Tgt.
CLIP	12.77	23.35	21.78	54.61	24.57	19.38
CoOp	$31.55_{\pm 13.58}$	$27.82_{\pm 11.34}$	$67.73_{\pm 5.60}$	$61.92_{\pm 2.96}$	$45.17_{\pm 1.80}$	$21.01_{\pm 0.71}$
ProGrad	$30.01_{\pm 10.58}$	$29.39_{\pm 3.14}$	$59.74_{\pm 3.03}$	$64.53_{\pm 2.72}$	$56.20_{\pm 0.00}$	$20.98_{\pm 0.83}$
KgCoOp	$15.83_{\pm 4.53}$	$22.92_{\pm 3.27}$	$55.72_{\pm 5.45}$	$54.34_{\pm 3.86}$	$38.93_{\pm 6.86}$	$20.15_{\pm 4.44}$
MaPLe	$53.82_{\pm 4.23}$	$30.79_{\pm 9.89}$	$80.70_{\pm 3.24}$	$59.56_{\pm 2.10}$	$76.10_{\pm 7.58}$	$22.08_{\pm 1.01}$
PromptSRC	$49.92_{\pm 3.36}$	$29.46_{\pm 3.66}$	$58.82_{\pm 6.70}$	$51.36_{\pm 3.38}$	$35.13_{\pm 9.56}$	$23.10_{\pm 0.79}$
DCPL	$55.70_{\pm 2.02}$	$29.89_{\pm 6.74}$	$80.88_{\pm 5.67}$	$59.40_{\pm 4.43}$	$78.90_{\pm 1.86}$	$20.20_{\pm 3.43}$
TCP	$24.99_{\pm 3.53}$	$19.28_{\pm 5.02}$	$57.37_{\pm 2.99}$	$48.66_{\pm 0.00}$	$38.30_{\pm 4.24}$	$21.86_{\pm 2.97}$
BiomedCoOp	$45.70_{\pm 2.62}$	$29.08_{\pm 12.20}$	$68.53_{\pm 2.99}$	$61.01_{\pm 5.06}$	$69.33_{\pm 5.35}$	$18.87_{\pm 2.05}$
SaMoE	$57.91_{\pm 1.72}$	$38.29_{\pm 6.27}$	$82.86_{\pm 1.07}$	$67.50_{\pm 6.43}$	$79.40_{\pm 4.38}$	$29.21_{\pm 0.44}$

Macro-F1 (%)

Method	Site Src.	Site Tgt.	Protocol Src.	Protocol Tgt.	Modality Src.	Modality Tgt.
CLIP	9.61	18.04	12.14	40.58	9.86	9.35
CoOp	$17.70_{\pm 7.69}$	$24.31_{\pm 7.61}$	$64.04_{\pm 6.13}$	$55.47_{\pm 3.35}$	$40.99_{\pm 1.49}$	$11.08_{\pm 0.94}$
ProGrad	$19.77_{\pm 4.09}$	$24.33_{\pm 1.77}$	$49.83_{\pm 3.75}$	$57.27_{\pm 2.28}$	$51.21_{\pm 0.00}$	$11.26_{\pm 1.42}$
KgCoOp	$13.95_{\pm 2.38}$	$20.29_{\pm 2.82}$	$40.55_{\pm 10.09}$	$42.85_{\pm 1.17}$	$36.43_{\pm 8.89}$	$13.12_{\pm 3.80}$
MaPLe	$42.35_{\pm 3.62}$	$27.21_{\pm 7.21}$	$79.99_{\pm 3.36}$	$47.87_{\pm 1.94}$	$75.51_{\pm 8.30}$	$11.57_{\pm 1.14}$
PromptSRC	$35.68_{\pm 1.74}$	$25.20_{\pm 2.48}$	$55.40_{\pm 9.19}$	$44.27_{\pm 4.28}$	$26.23_{\pm 13.04}$	$11.15_{\pm 2.93}$
DCPL	$40.52_{\pm 0.59}$	$27.04_{\pm 4.74}$	$80.06_{\pm 5.70}$	$46.17_{\pm 4.92}$	$81.45_{\pm 1.89}$	$12.15_{\pm 1.99}$
TCP	$12.46_{\pm 0.58}$	$18.38_{\pm 1.04}$	$52.60_{\pm 4.46}$	$36.16_{\pm 0.00}$	$35.02_{\pm 8.08}$	$13.69_{\pm 3.38}$
BiomedCoOp	$29.81_{\pm 0.98}$	$27.57_{\pm 3.36}$	$67.32_{\pm 2.83}$	$52.84_{\pm 5.55}$	$69.01_{\pm 5.68}$	$9.11_{\pm 1.63}$
SaMoE	$44.47_{\pm 0.40}$	$36.01_{\pm 4.22}$	$82.47_{\pm 1.31}$	$58.56_{\pm 5.18}$	$75.23_{\pm 4.63}$	$14.02_{\pm 1.00}$

Tables 17 and 18 compare each backbone with its own frozen Raw baseline. Base-to-novel accuracy HM reaches 66.46, 65.17, and 65.75 on CLIP ViT-L/14, SigLIP, and SigLIP2, respectively. Positive HM gains also occur in each cross-domain setting for both metrics. These are comparisons with the corresponding frozen backbone; they do not establish superiority over all prompt-learning methods on these architectures. HM and its standard deviation are aggregated over training seeds; reported differences use the unrounded results.

Table 17: Base-to-novel results on modern backbones, averaged across 11 datasets. ΔHM is SaMoE minus Raw.

Accuracy (%)

Backbone	Model	Base	Novel	HM	ΔHM
CLIP ViT-L/14	Raw	48.44	47.83	48.13	–
CLIP ViT-L/14	SaMoE	$73.74_{\pm 0.45}$	$60.48_{\pm 4.16}$	$66.46_{\pm 3.21}$	+18.32
SigLIP-400M	Raw	53.28	58.80	55.91	–
SigLIP-400M	SaMoE	$69.56_{\pm 0.51}$	$61.31_{\pm 3.85}$	$65.17_{\pm 2.96}$	+9.27
SigLIP2-400M	Raw	52.00	55.92	53.89	–
SigLIP2-400M	SaMoE	$68.68_{\pm 0.48}$	$63.06_{\pm 3.62}$	$65.75_{\pm 2.78}$	+11.86

Macro-F1 (%)

Backbone	Model	Base	Novel	HM	ΔHM
CLIP ViT-L/14	Raw	39.15	36.95	38.02	–
CLIP ViT-L/14	SaMoE	$72.02_{\pm 0.53}$	$45.82_{\pm 3.48}$	$56.01_{\pm 2.74}$	+17.99
SigLIP-400M	Raw	40.40	49.12	44.34	–
SigLIP-400M	SaMoE	$65.30_{\pm 0.62}$	$59.41_{\pm 3.15}$	$62.22_{\pm 2.42}$	+17.88
SigLIP2-400M	Raw	41.41	47.87	44.41	–
SigLIP2-400M	SaMoE	$69.19_{\pm 0.57}$	$55.39_{\pm 3.04}$	$61.53_{\pm 2.31}$	+17.12

G.3 Capacity-Matched Comparisons and Execution Cost

To distinguish the routing mechanism from additional capacity, we compare SaMoE (Independent) with heavier MaPLe and BiomedCoOp controls. Heavy MaPLe enlarges its native text-to-visual coupling function. Original BiomedCoOp does not use image conditioning, but its CoCoOp-derived

Table 18: Modern-backbone cross-domain results. Target metrics are averaged over unseen domains. ΔHM is SaMoE minus Raw.

Accuracy (%)

Setting	Backbone	Model	Source	Target avg.	HM	ΔHM
Site	CLIP ViT-L/14	Raw	47.58	53.64	50.43	–
Site	CLIP ViT-L/14	SaMoE	$52.82_{\pm 0.46}$	$58.34_{\pm 3.35}$	$55.44_{\pm 2.64}$	+5.01
Site	SigLIP-400M	Raw	15.66	27.65	20.00	–
Site	SigLIP-400M	SaMoE	$54.31_{\pm 0.50}$	$29.04_{\pm 3.72}$	$37.84_{\pm 2.88}$	+17.85
Site	SigLIP2-400M	Raw	55.96	59.39	57.62	–
Site	SigLIP2-400M	SaMoE	$59.55_{\pm 0.47}$	$61.29_{\pm 3.44}$	$60.41_{\pm 2.69}$	+2.78
Protocol	CLIP ViT-L/14	Raw	40.65	63.16	49.46	–
Protocol	CLIP ViT-L/14	SaMoE	$81.60_{\pm 0.44}$	$61.11_{\pm 5.26}$	$69.88_{\pm 2.56}$	+20.42
Protocol	SigLIP-400M	Raw	57.13	73.09	64.13	–
Protocol	SigLIP-400M	SaMoE	$88.70_{\pm 0.49}$	$82.89_{\pm 3.64}$	$85.70_{\pm 2.81}$	+21.56
Protocol	SigLIP2-400M	Raw	62.67	76.00	68.69	–
Protocol	SigLIP2-400M	SaMoE	$64.82_{\pm 0.45}$	$85.45_{\pm 7.32}$	$73.72_{\pm 2.60}$	+5.02
Modality	CLIP ViT-L/14	Raw	24.20	29.99	26.79	–
Modality	CLIP ViT-L/14	SaMoE	$71.10_{\pm 0.48}$	$31.72_{\pm 3.51}$	$43.87_{\pm 2.74}$	+17.08
Modality	SigLIP-400M	Raw	25.50	39.38	30.96	–
Modality	SigLIP-400M	SaMoE	$32.80_{\pm 0.51}$	$43.67_{\pm 3.78}$	$37.46_{\pm 2.93}$	+6.51
Modality	SigLIP2-400M	Raw	22.70	36.34	27.94	–
Modality	SigLIP2-400M	SaMoE	$25.20_{\pm 0.46}$	$38.43_{\pm 5.38}$	$30.44_{\pm 2.63}$	+2.50

Macro-F1 (%)

Setting	Backbone	Model	Source	Target avg.	HM	ΔHM
Site	CLIP ViT-L/14	Raw	19.97	31.35	24.40	–
Site	CLIP ViT-L/14	SaMoE	$37.00_{\pm 0.52}$	$33.65_{\pm 2.81}$	$35.25_{\pm 2.22}$	+10.85
Site	SigLIP-400M	Raw	14.00	26.68	18.36	–
Site	SigLIP-400M	SaMoE	$38.80_{\pm 0.58}$	$28.53_{\pm 3.06}$	$32.88_{\pm 2.36}$	+14.52
Site	SigLIP2-400M	Raw	30.03	42.96	35.35	–
Site	SigLIP2-400M	SaMoE	$31.09_{\pm 0.53}$	$49.92_{\pm 2.87}$	$38.32_{\pm 2.25}$	+2.97
Protocol	CLIP ViT-L/14	Raw	39.36	50.86	44.38	–
Protocol	CLIP ViT-L/14	SaMoE	$81.32_{\pm 0.51}$	$53.67_{\pm 4.74}$	$64.66_{\pm 2.16}$	+20.29
Protocol	SigLIP-400M	Raw	48.01	69.77	56.88	–
Protocol	SigLIP-400M	SaMoE	$88.49_{\pm 0.56}$	$76.45_{\pm 3.01}$	$82.03_{\pm 2.31}$	+25.15
Protocol	SigLIP2-400M	Raw	56.47	73.12	63.73	–
Protocol	SigLIP2-400M	SaMoE	$64.30_{\pm 0.52}$	$81.20_{\pm 6.78}$	$71.77_{\pm 2.18}$	+8.04
Modality	CLIP ViT-L/14	Raw	9.78	24.03	13.90	–
Modality	CLIP ViT-L/14	SaMoE	$70.91_{\pm 0.54}$	$25.05_{\pm 4.92}$	$37.02_{\pm 2.28}$	+23.12
Modality	SigLIP-400M	Raw	11.02	27.96	15.81	–
Modality	SigLIP-400M	SaMoE	$24.79_{\pm 0.59}$	$31.99_{\pm 3.11}$	$27.93_{\pm 2.40}$	+12.12
Modality	SigLIP2-400M	Raw	14.53	27.03	18.90	–
Modality	SigLIP2-400M	SaMoE	$18.58_{\pm 0.52}$	$28.05_{\pm 3.83}$	$22.35_{\pm 2.20}$	+3.45

implementation retains a dormant Meta-Net; Heavy BiomedCoOp activates and enlarges this Meta-Net. Other prompt objectives and training settings are unchanged. SaMoE (Shared) further shares the ExpertBank across layers, using fewer trainable parameters than MaPLe.

Tables 19 and 20 show that increasing baseline capacity alone does not reproduce SaMoE's generalization gains. SaMoE (Independent) exceeds parameter-matched Heavy MaPLe by 13.43 accuracy-HM points; the shared variant exceeds MaPLe by 5.82 points with fewer parameters. Independent SaMoE also achieves the highest target averages in all three cross-domain settings, although Heavy MaPLe has higher source accuracy in the MRI setting. Thus, source fitting alone does not explain target generalization.

Table 21 reports a separate execution-cost benchmark on BTMRI, using AMP on an NVIDIA GeForce RTX 2080 Ti (11 GiB), training batch size 4, and test batch size 100. Inference throughput measures steady-state model forward passes, excluding the first warm-up batch. Active/img counts trainable parameters executed per image. At Top-$K = 2$, two routed experts are activated per routed layer in addition to the always-active shared transformation. Consequently, stored parameter count is distinct from executed parameters and GFLOPs. Relative to MaPLe in this benchmark, independent SaMoE adds approximately 2.4% GFLOPs, reduces inference throughput by 0.63%, and increases training time from 90.97 to 97.54 seconds. The shared variant reduces trainable storage by 83.7% relative to independent SaMoE, at lower HM. These measurements quantify a storage/generalization

Table 19: Capacity-controlled base-to-novel comparison over 11 datasets. Params are trainable parameters; HM is the accuracy harmonic mean.

Accuracy (%)

Method	Params	Base	Novel	Acc. HM
BiomedCoOp	0.004M	$70.84_{\pm 2.62}$	$36.07_{\pm 8.83}$	$47.80_{\pm 7.78}$
Heavy BiomedCoOp	15.91M	$68.69_{\pm 0.21}$	$37.93_{\pm 3.73}$	$48.80_{\pm 3.14}$
MaPLe	3.63M	$76.37_{\pm 3.13}$	$39.52_{\pm 8.76}$	$52.09_{\pm 7.64}$
Heavy MaPLe	15.92M	$76.66_{\pm 1.59}$	$38.88_{\pm 1.79}$	$51.58_{\pm 1.81}$
SaMoE (Shared)	2.59M	$75.71_{\pm 1.63}$	$46.89_{\pm 7.91}$	$57.91_{\pm 6.05}$
SaMoE (Indep.)	15.91M	$76.79_{\pm 2.44}$	$56.37_{\pm 5.91}$	$65.01_{\pm 4.03}$

Macro-F1 (%)

Method	Params	Base	Novel	Acc. HM
BiomedCoOp	0.004M	$67.76_{\pm 1.95}$	$27.47_{\pm 4.45}$	$47.80_{\pm 7.78}$
Heavy BiomedCoOp	15.91M	$66.62_{\pm 0.86}$	$28.18_{\pm 3.76}$	$48.80_{\pm 3.14}$
MaPLe	3.63M	$74.06_{\pm 1.93}$	$32.67_{\pm 4.39}$	$52.09_{\pm 7.64}$
Heavy MaPLe	15.92M	$75.56_{\pm 1.66}$	$32.14_{\pm 0.54}$	$51.58_{\pm 1.81}$
SaMoE (Shared)	2.59M	$74.28_{\pm 1.32}$	$39.06_{\pm 3.87}$	$57.91_{\pm 6.05}$
SaMoE (Indep.)	15.91M	$76.05_{\pm 1.07}$	$45.13_{\pm 4.00}$	$65.01_{\pm 4.03}$

Table 20: Capacity-controlled cross-domain results. Src./Tgt. denote the source result and target average.

Accuracy (%)

Method	Params	Site Src.	Site Tgt.	Protocol Src.	Protocol Tgt.	Modality Src.	Modality Tgt.
BiomedCoOp	0.004M	$45.70_{\pm 2.62}$	$29.08_{\pm 12.20}$	$68.53_{\pm 2.99}$	$61.01_{\pm 5.06}$	$69.33_{\pm 5.35}$	$18.87_{\pm 2.05}$
Heavy BiomedCoOp	15.91M	$48.06_{\pm 4.04}$	$23.45_{\pm 1.89}$	$66.47_{\pm 2.59}$	$63.45_{\pm 4.97}$	$65.27_{\pm 5.40}$	$21.28_{\pm 2.56}$
MaPLe	3.63M	$53.82_{\pm 4.23}$	$30.79_{\pm 9.89}$	$80.70_{\pm 3.24}$	$59.56_{\pm 2.10}$	$76.10_{\pm 7.58}$	$22.08_{\pm 1.01}$
Heavy MaPLe	15.92M	$55.31_{\pm 1.49}$	$30.52_{\pm 7.08}$	$84.51_{\pm 1.54}$	$62.09_{\pm 6.24}$	$77.37_{\pm 4.53}$	$22.30_{\pm 1.16}$
SaMoE (Shared)	2.59M	$56.42_{\pm 2.11}$	$31.31_{\pm 6.29}$	$80.18_{\pm 0.79}$	$59.33_{\pm 2.99}$	$75.33_{\pm 4.41}$	$25.47_{\pm 1.03}$
SaMoE (Indep.)	15.91M	$57.91_{\pm 1.72}$	$38.29_{\pm 6.27}$	$82.86_{\pm 1.07}$	$67.50_{\pm 6.43}$	$79.40_{\pm 4.38}$	$29.21_{\pm 0.44}$

Macro-F1 (%)

Method	Params	Site Src.	Site Tgt.	Protocol Src.	Protocol Tgt.	Modality Src.	Modality Tgt.
BiomedCoOp	0.004M	$29.81_{\pm 0.98}$	$27.57_{\pm 3.36}$	$67.32_{\pm 2.83}$	$52.84_{\pm 5.55}$	$69.01_{\pm 5.68}$	$9.11_{\pm 1.63}$
Heavy BiomedCoOp	15.91M	$33.44_{\pm 1.11}$	$19.32_{\pm 2.91}$	$65.45_{\pm 2.03}$	$57.05_{\pm 5.17}$	$64.93_{\pm 5.24}$	$10.99_{\pm 1.75}$
MaPLe	3.63M	$42.35_{\pm 3.62}$	$27.21_{\pm 7.21}$	$79.99_{\pm 3.36}$	$47.87_{\pm 1.94}$	$75.51_{\pm 8.30}$	$11.57_{\pm 1.14}$
Heavy MaPLe	15.92M	$40.01_{\pm 2.37}$	$26.81_{\pm 5.97}$	$81.86_{\pm 1.66}$	$50.97_{\pm 7.35}$	$76.31_{\pm 4.41}$	$12.65_{\pm 1.45}$
SaMoE (Shared)	2.59M	$42.38_{\pm 1.62}$	$29.14_{\pm 3.97}$	$79.26_{\pm 1.51}$	$52.47_{\pm 2.31}$	$71.47_{\pm 4.29}$	$13.11_{\pm 1.37}$
SaMoE (Indep.)	15.91M	$44.47_{\pm 0.40}$	$36.01_{\pm 4.22}$	$82.47_{\pm 1.31}$	$58.56_{\pm 5.18}$	$75.23_{\pm 4.63}$	$14.02_{\pm 1.00}$

trade-off and do not imply equal performance on every device or implementation. They are a separate benchmark from the batch-size sweep in Table 9.

G.4 Class- and Subject-Controlled Protocol Information in Routing

We analyze the complete 32-dimensional routing vector, formed by the normalized routing masses from eight deep routed layers and four experts. Each seed uses 24 subjects from two shared classes, glioma tumor and normal brain (12 subjects per class), with four scans per subject under FLAIR, T1W, T1W-CE, and T2W. This yields 96 balanced vectors per seed. A leave-one-class-out (LOCO) probe trains on one class and predicts protocol only in the held-out class; four-way chance balanced accuracy is 25%. Partial R^2 separates Protocol and Class effects. Protocol-label permutations preserve class and repeated-subject structure. The analysis uses 1,000 within-class permutations per seed, 2,000 independent-cluster bootstraps per seed, and 5,000 hierarchical seed-and-cluster bootstraps (analysis seed 2026). Expert identities are not aligned across training seeds; only within-seed scalar statistics are aggregated.

Table 22 reports LOCO balanced accuracy of $81.6_{\pm 7.3}\%$, versus a 25% null baseline, with $p = 0.001$. The per-seed Protocol/Class partial-R^2 ratio averages 19.07. Table 23 complements this association test with layer-averaged Jensen–Shannon distances under protocol changes (Δ_{style}) and class changes

Table 21: Execution cost on BTMRI. Time is training time in seconds, FPS is images/s, latency is ms/image, and VRAM is MiB. Both panels use the same benchmark.

Parameters and training cost

Method	Params	Active/img	GFLOPs	Time (s)	Train FPS	Train VRAM
BiomedCoOp	0.004M	0.004M	19.99	87.68	18.25	956.9
MaPLe	3.63M	3.58M	20.47	90.97	17.59	1071.2
SaMoE (Shared)	2.59M	2.18M	20.96	95.19	16.30	1106.5
Heavy BiomedCoOp	15.91M	15.91M	20.01	88.57	17.70	1241.6
Heavy MaPLe	15.92M	15.92M	20.57	93.79	17.06	1243.9
SaMoE (Indep.)	15.91M	10.44M	20.96	97.54	16.40	1275.6

Inference cost

Method	Infer FPS	Latency	Infer VRAM
BiomedCoOp	713.89	1.401	1345.4
MaPLe	665.03	1.504	1087.8
SaMoE (Shared)	663.01	1.508	1080.3
Heavy BiomedCoOp	634.36	1.588	1499.6
Heavy MaPLe	654.85	1.527	1262.1
SaMoE (Indep.)	660.85	1.513	1248.9

(Δ_{class}). Their ratio, SDR, is 1.57 with hierarchical 95% interval [1.25, 2.11]. These measurements support acquisition-sensitive routing beyond category information; they do not require a one-to-one correspondence between expert identities and protocols. Seed 3's individual SDR interval includes one, so the aggregate result should not be interpreted as a separately significant effect in every seed.

Table 22: Class- and subject-controlled protocol information in routing. P/C is the Protocol/Class partial-R^2 ratio. SDR is the ratio of protocol-induced to class-induced routing distances.

Seed	LOCO BA	Null BA	Protocol R^2	Class R^2	P/C	p	SDR [95% CI]
1	74.0%	$24.9_{\pm 4.9}$%	0.558	0.026	21.80	0.001	1.51 [1.28,1.82]
2	88.5%	$25.1_{\pm 5.0}$%	0.642	0.022	29.19	0.001	1.92 [1.47,2.79]
3	82.3%	$24.9_{\pm 4.9}$%	0.303	0.049	6.20	0.001	1.27 [0.95,1.87]
Aggregate	$81.6_{\pm 7.3}$%	25.0%	$0.501_{\pm 0.177}$	$0.032_{\pm 0.015}$	$19.07_{\pm 11.74}$	0.001	1.57 [1.25,2.11]

Table 23: Layer-averaged routing distances. The aggregate SDR interval is obtained by hierarchical bootstrap.

Seed	Δ_{style}	Δ_{class}	SDR [95% CI]
1	0.000496	0.000328	1.51 [1.28,1.82]
2	0.047850	0.024896	1.92 [1.47,2.79]
3	0.014073	0.011107	1.27 [0.95,1.87]
Aggregate	$0.021_{\pm 0.024}$	$0.012_{\pm 0.012}$	1.57 [1.25,2.11]

G.5 Controlled Style-Mixture and Semantic Shifts

The style-mixture formulation describes acquisition variation when target styles are shared with, close to, or inferable from source-observed components. It does not completely characterize cross-modality changes in anatomy or label semantics. We therefore construct a supported-mixture experiment that separates changes in style proportions from changes in class semantics, rather than treating it as evidence for arbitrary unseen anatomy.

On BTMRI, DermaMNIST, and Retina, training uses 64 shots per base class and three seeds. This deliberately controlled experiment is separate from the 16-shot benchmark. Four label-preserving style components apply gamma response ($\gamma = 0.65$), sensor noise ($\sigma = 0.04$), Gaussian blur (radius 1.5), and $0.5\times$ resolution degradation. Source mixture weights are $(0.40, 0.20, 0.20, 0.20)$; each of three target directions assigns 0.70 to one non-primary component and 0.10 to each other component. Severity is fixed. Assignments are deterministic, class-balanced, and identical across methods, and

style labels are never exposed during learning. Each test image is rendered under all four components; metrics are reconstructed from component-wise confusion matrices and target scores average the three directions.

The same checkpoint is evaluated on seen/novel classes under source/target mixtures, producing IID, style-only, semantic-only, and compound cells. Table 24 compares a Shared prompt, Static experts, and SaMoE (Independent). Table 25 reports the routing advantage over Static experts, G, separately from an additional mixture-specific effect. Define $D_s = G_{s,\mathrm{target}} - G_{s,\mathrm{source}}$ for $s \in \{\mathrm{seen}, \mathrm{novel}\}$, interaction $D_{\mathrm{novel}} - D_{\mathrm{seen}}$, and retained gain $G_{\mathrm{novel,target}}/G_{\mathrm{novel,source}}$ (undefined if the denominator is non-positive). Confidence intervals use 2,000 hierarchical paired-bootstrap replicates, resampling training seeds and then paired images.

Novel-target routing gains over Static experts are 7.99, 10.73, and 6.91 accuracy points on BTMRI, DermaMNIST, and Retina, retaining 93.1%, 106.9%, and 96.5% of their respective novel-source advantages. The BTMRI gain interval crosses zero; DermaMNIST and Retina intervals exclude zero. SaMoE does not outperform the Shared prompt in every cell: DermaMNIST novel-target accuracy is 26.75 versus 32.39. All D_{novel} intervals include zero. Thus, these results support preservation of a routing advantage over Static experts under simultaneous semantic novelty and supported mixture reweighting, but not an additional mixture-specific benefit or uniform superiority over all controls.

Table 24: Controlled 2×2 factorial results at 64 shots per base class. Source/Target denotes mixture weights and Seen/Novel denotes class semantics.

Accuracy (%)

Dataset	Method	Seen+Source	Seen+Target	Novel+Source	Novel+Target
BTMRI	Shared prompt	$90.08_{\pm 1.37}$	$89.94_{\pm 1.27}$	$52.23_{\pm 9.19}$	$52.06_{\pm 8.76}$
BTMRI	Static experts	$89.52_{\pm 1.51}$	$89.25_{\pm 1.35}$	$47.19_{\pm 9.76}$	$46.47_{\pm 10.12}$
BTMRI	SaMoE (Indep.)	$90.72_{\pm 0.16}$	$90.66_{\pm 0.45}$	$55.78_{\pm 3.84}$	$54.46_{\pm 2.59}$
DermaMNIST	Shared prompt	$68.61_{\pm 2.02}$	$68.77_{\pm 1.80}$	$33.77_{\pm 27.59}$	$32.39_{\pm 28.48}$
DermaMNIST	Static experts	$68.88_{\pm 1.29}$	$68.94_{\pm 1.98}$	$16.13_{\pm 2.65}$	$16.02_{\pm 2.85}$
DermaMNIST	SaMoE (Indep.)	$69.16_{\pm 2.75}$	$69.35_{\pm 3.04}$	$26.17_{\pm 5.98}$	$26.75_{\pm 6.33}$
Retina	Shared prompt	$90.14_{\pm 0.52}$	$89.69_{\pm 0.69}$	$66.11_{\pm 12.54}$	$66.33_{\pm 12.99}$
Retina	Static experts	$88.96_{\pm 1.01}$	$88.63_{\pm 1.29}$	$62.32_{\pm 6.49}$	$62.64_{\pm 7.16}$
Retina	SaMoE (Indep.)	$89.24_{\pm 0.73}$	$89.17_{\pm 0.84}$	$69.48_{\pm 3.83}$	$69.55_{\pm 2.83}$

Macro-F1 (%)

Dataset	Method	Seen+Source	Seen+Target	Novel+Source	Novel+Target
BTMRI	Shared prompt	$90.08_{\pm 1.38}$	$89.93_{\pm 1.27}$	$47.73_{\pm 6.03}$	$46.87_{\pm 5.26}$
BTMRI	Static experts	$89.50_{\pm 1.52}$	$89.23_{\pm 1.37}$	$45.23_{\pm 10.34}$	$44.21_{\pm 10.42}$
BTMRI	SaMoE (Indep.)	$90.71_{\pm 0.16}$	$90.65_{\pm 0.45}$	$52.94_{\pm 2.12}$	$50.78_{\pm 1.39}$
DermaMNIST	Shared prompt	$64.22_{\pm 1.90}$	$64.52_{\pm 1.95}$	$20.90_{\pm 11.64}$	$20.10_{\pm 12.26}$
DermaMNIST	Static experts	$64.58_{\pm 1.76}$	$64.65_{\pm 1.97}$	$11.83_{\pm 2.32}$	$11.73_{\pm 2.42}$
DermaMNIST	SaMoE (Indep.)	$64.96_{\pm 3.10}$	$65.25_{\pm 3.18}$	$19.25_{\pm 4.84}$	$19.53_{\pm 4.29}$
Retina	Shared prompt	$90.14_{\pm 0.53}$	$89.68_{\pm 0.70}$	$65.72_{\pm 12.14}$	$65.96_{\pm 12.67}$
Retina	Static experts	$88.96_{\pm 1.01}$	$88.63_{\pm 1.29}$	$57.29_{\pm 9.04}$	$57.63_{\pm 10.30}$
Retina	SaMoE (Indep.)	$89.23_{\pm 0.74}$	$89.16_{\pm 0.85}$	$67.06_{\pm 4.71}$	$67.12_{\pm 3.25}$

Table 25: Routing advantage and mixture-specific effects, in accuracy percentage points. CI denotes a hierarchical paired-bootstrap 95% confidence interval.

Routing advantage

Dataset	$G_{\mathrm{seen,tgt}}$	$G_{\mathrm{novel,src}}$	$G_{\mathrm{novel,tgt}}$	Retained	CI $G_{\mathrm{novel,tgt}}$
BTMRI	$1.40_{\pm 1.15}$	$8.59_{\pm 8.17}$	$7.99_{\pm 10.90}$	93.1%	[-3.16,17.29]
DermaMNIST	$0.41_{\pm 1.06}$	$10.04_{\pm 4.76}$	$10.73_{\pm 4.53}$	106.9%	[5.71,14.20]
Retina	$0.54_{\pm 1.90}$	$7.16_{\pm 6.81}$	$6.91_{\pm 6.16}$	96.5%	[2.51,13.96]

Mixture-specific effects

Dataset	D_{seen}	D_{novel}	Interaction	CI D_{novel}
BTMRI	$0.20_{\pm 0.28}$	$-0.60_{\pm 3.27}$	$-0.80_{\pm 3.54}$	[-4.17,1.81]
DermaMNIST	$0.14_{\pm 0.44}$	$0.69_{\pm 1.15}$	$0.55_{\pm 1.39}$	[-0.39,1.81]
Retina	$0.27_{\pm 0.40}$	$-0.25_{\pm 0.65}$	$-0.52_{\pm 0.64}$	[-1.05,0.49]

H Discussion on Limitations and Broader Impact

H.1 Discussion on Limitations

SaMoE is designed for biomedical VLM adaptation under acquisition-induced style shifts, but several limitations should be considered when interpreting its scope. First, the mixture-shift view is a modeling abstraction rather than a literal physical decomposition of biomedical image formation. It treats each domain as a mixture of latent acquisition-style components and uses the change of mixture weights to characterize the style component of the domain gap. This abstraction is useful for motivating style-aware prompt composition, but it does not imply that the latent components are observed, uniquely identifiable, or aligned with discrete scanner/site/protocol labels. As discussed in Appendix A, the formulation is most appropriate when target styles are contained in, close to, or statistically inferable from style factors observed in the source data. For completely disjoint target styles with no corresponding source evidence, source-only few-shot training cannot guarantee low target component risk.

Second, SaMoE is most effective when acquisition-induced style variation is a major source of transfer difficulty. Its gains can be smaller when the source-target style gap is mild, because a single prompt or conventional prompt-learning baseline may already transfer sufficiently well. Appendix B.3 provides a failure study showing that the improvement over the strongest baseline is small on target protocols with mild measured style discrepancy. Conversely, when semantic, anatomical, or task-level differences dominate the domain gap, as in compound cross-modality transfer, style-aware routing can still provide useful adaptation cues but cannot fully explain or eliminate the broader distribution shift. Therefore, the cross-modality results should be interpreted as a stress test rather than as evidence that all modality gaps are reducible to style shifts.

Third, SaMoE introduces additional architectural complexity compared with standard single-prompt tuning. The expert bank and image-conditioned router increase the number of trainable parameters and add routing computation at each prompt injection layer. Although our efficiency analysis in Appendix C shows that the method remains practical with a frozen CLIP backbone, this cost may matter in highly resource-constrained clinical or edge deployment scenarios. The shared-expert variant reduces this overhead, but the trade-off between capacity, interpretability, and efficiency remains an important design consideration, consistent with broader concerns about resource-efficient model architectures [40].

Finally, this work focuses on classification-oriented biomedical VLM adaptation. The current evaluation does not cover dense prediction tasks such as segmentation, detection, or lesion localization, where acquisition shifts may interact with spatial precision and annotation protocols differently. The routing signals are also based on compact intermediate feature statistics and CLS context, which makes the method efficient but only partially interpretable. Future work should study more interpretable routing factors, multi-source and continual adaptation, and extensions to dense clinical tasks under evolving acquisition pipelines, with fine-grained multimodal understanding providing a broader context [27].

H.2 Impact Statement

The broader impact of this work lies in its potential to enhance the reliability and equity of AI-driven medical diagnostics by addressing real-world acquisition-induced style shifts. By improving transfer across sites, protocols, and broader biomedical domain gaps, SaMoE may help reduce performance degradation when models are applied beyond the acquisition conditions seen during adaptation. This can support more reliable deployment of biomedical VLMs in settings where data availability, annotation cost, and imaging protocols vary across institutions.

At the same time, this work should not be interpreted as validating autonomous clinical diagnosis. Incorrect predictions under unsupported target styles, severe modality gaps, or underrepresented patient populations may still lead to harmful decisions if a system is used without expert oversight. The method also relies on existing biomedical datasets and pretrained VLMs, whose biases, demographic coverage, and acquisition distributions may not reflect all clinical environments. Responsible deployment therefore requires prospective validation, monitoring for distribution shifts, human oversight, and careful assessment of fairness, privacy, and institutional constraints. Future societal benefits may

include reduced diagnostic errors and improved healthcare accessibility, but these benefits depend on rigorous evaluation and ethical governance aligned with the interests of patients and providers.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main contributions: a mixture-shift formulation of acquisition-induced style shifts, the SaMoE prompt-learning architecture, and empirical evaluation across base-to-novel, cross-site, cross-protocol, and compound cross-modality settings. The claims are scoped to biomedical VLM adaptation under acquisition-related shifts, and the paper discusses assumptions and limitations for mild style gaps, unsupported unseen styles, and compound modality shifts.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion discusses the main limitations of SaMoE, including the mixture-shift modeling assumption, smaller gains when acquisition style is not the dominant source of shift, and the additional complexity introduced by the expert bank and routing mechanism. Appendix B.3 further analyzes failure cases where SaMoE is less effective under mild style gaps or non-style-dominated transfer difficulty.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The main text states the mixture-shift abstraction and its scope in Section 3.1, including the caveat that unsupported unseen styles are not formally guaranteed. Appendix A provides the detailed domain-discrepancy derivation, mixture-risk analysis, worst-case mixture bound, and prompt-space interpretation used to motivate SaMoE.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the model architecture in Sections 3.2 and 3.3, the evaluation protocols and datasets in the Experiment section, and detailed dataset statistics and benchmark protocols in Appendix F. Implementation details, routing choices, hyperparameters, and additional adapted-baseline comparisons are provided in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper uses public biomedical datasets and cites their original sources, with dataset statistics and protocols summarized in Appendix F. We have provide source code in supplementary material and we will provide data, code, and instructions for the final paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The Experiment section describes the base-to-novel, cross-site, cross-protocol, and compound cross-modality settings, including source/target usage and 16-shot adaptation. The main text reports key hyperparameters such as prompt depth, prompt length, number of experts, Top-K, router hidden dimension, and load-balancing weight, while Appendix F and Appendix B provide dataset splits, benchmark protocols, and additional implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports averaged performance over three independent runs with different random seeds in the Experiment section. The cross-site setting reports mean and standard deviation in the main text, and Appendix B.5 provides additional means with standard deviations as subscripts seed-stability statistics for the Base-to-Novel benchmark averaged across 11 datasets.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The Experiment section states that experiments are conducted on an RTX4090 GPU and reports training time for the compared methods. The ablation and efficiency analyses further report GFLOPs and throughput, while Appendix C provides additional prompting-complexity and inference-throughput comparisons.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: We have carefully reviewed the Code of Ethics and affirm that our work fully complies with its principles.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The Impact Statement discusses potential positive impacts, including improved reliability, equity, and accessibility of AI-driven medical diagnostics in Appendix H.2. It also discusses negative or unintended impacts, including misuse, over-reliance on automated diagnosis, and the need for responsible deployment and ethical oversight.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a high-risk pretrained language model, image generator, scraped dataset, or other asset with substantial dual-use risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use publicly available datasets and pretrainedd models. All benchmarks and models used are properly cited in the main text, and their licenses and terms of use have been followed.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release a new dataset, benchmark, model checkpoint, or other standalone assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing, user studies, participant instructions, or compensation. Experiments are conducted on existing biomedical image datasets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not conduct new human-subject research or collect new patient data. It evaluates algorithms on existing biomedical image datasets, so IRB approval or equivalent review for new human-subject experiments is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method does not use LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.