BioMedVR: Confusion-Aware Mixture-of-Prompt Experts for Biomedical Visual Reprogramming

Jiaxiang Liu[1,2], Tianxiang Hu[2], Juwei Guan[3], Yujie Wu[4], Yusong Wang[5], Yao Mu[6], Zuozhu Liu[2], and Mingkun Xu[1]

[1] Guangdong Institute of Intelligence Science and Technology, Zhuhai, China
[2] Zhejiang University, Hangzhou, China
[3] Southeast University, Nanjing, China
[4] The Hong Kong Polytechnic University, Hong Kong SAR, China
[5] Institute of Science Tokyo, Tokyo, Japan
[6] Shanghai Jiao Tong University, Shanghai, China

Abstract. Recent advances in vision–language models (VLMs) such as CLIP have demonstrated strong generalization across natural-image domains. However, adapting these models to biomedical imaging is nontrivial: full-model fine-tuning is computationally expensive, while medical data are often scarce and exhibit subtle, fine-grained inter-class differences, making parameter-efficient adaptation particularly critical. Visual Reprogramming (VR) offers a parameter-efficient alternative by injecting learnable perturbations into the input space, but existing VR approaches for VLMs mainly focus on positive class prompts and overlook confusing negatives, leading to miscalibrated predictions in fine-grained medical scenarios. We present BioMedVR, the first VR-based framework for biomedical imaging, enabling few-shot adaptation of pretrained VLMs through compact learnable VR modules. To mitigate class confusion, we introduce a Confusion Minimization Mechanism that leverages LLM-generated confusion-aware attributes together with a Confusion-Suppression Loss to explicitly reduce false-positive alignment. Moreover, the designed Mixture-of-Prompt Experts combines a positive expert for main-class discrimination and a negative expert for confusion suppression, balanced via adaptive gating. Extensive experiments on 18 datasets—including 11 biomedical datasets and 7 natural image benchmarks—demonstrate that BioMedVR achieves superior accuracy and generalization, effectively bridging VR and VLMs in biomedical domains.

Keywords: Semantic Decoupling · Confusion-Aware Learning · Visual Reprogramming · Differential Diagnosis–Inspired Modeling

1 Introduction

Recent advances in vision-language models (VLMs), such as CLIP [8, 40, 48, 49], have opened new avenues for multimodal understanding. Unlike conventional supervised learning restricted to closed visual vocabularies, contrastive pretraining

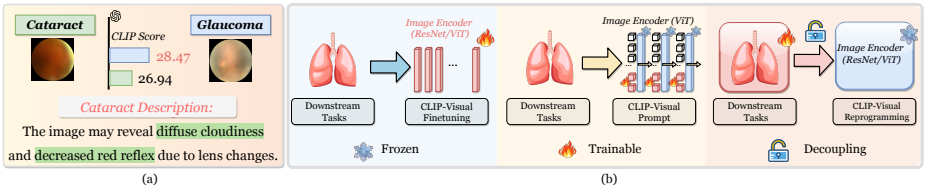

Fig. 1: (a) A cataract-specific description scores highly for glaucoma, exposing CLIP's semantic confusion and motivating the use of confusion-aware negative attributes to better separate similar diseases. (b) Comparison of CLIP adaptation strategies. (i) Finetuning updates all encoder parameters. (ii) Visual Prompting injects learnable tokens within ViT layers. (iii) VR learns lightweight input perturbations with frozen encoders, enabling efficient and architecture-agnostic adaptation. While prompt learning is parameter-efficient, it typically requires access to model internals such as token embeddings or encoder interfaces. In contrast, VR operates in the input space, making it architecture-agnostic and better suited for privacy-sensitive medical deployments.

aligns image and text representations via natural language supervision, enabling open-set recognition. However, adapting large VLMs to specialized domains remains challenging: full model training is computationally prohibitive, and performance is highly sensitive to the design of both visual inputs and textual prompts [11, 47].

To alleviate this, prompt learning [27, 30, 54, 55] offers a parameter-efficient alternative by inserting learnable prompts into text tokens, image patches, attention layers, or cross-modal mappings [7, 36, 44]. Representative methods include CoOp [55], CoCoOp [54], and MaPLe [27], with BiomedCoOp [28] extending this paradigm to biomedical imaging via LLM-guided prompt ensembles. However, existing prompt-learning methods still depend on model internals and struggle with fine-grained, modality-specific medical cues (Figure 1b).

An alternative and promising paradigm for adapting VLMs is Visual Reprogramming (VR) [5, 7, 9]. Unlike prompt learning, which requires additional parameters and access to model internals, VR learns a trainable visual pattern in the input space, achieving architecture-agnostic transfer without modifying pretrained weights. Originally explored for language reuse [17, 46], the reprogramming paradigm has since been extended to graphs [24], vision [6, 7, 43], and speech modeling [21, 51]. In the visual domain, VR overlays a learnable perturbation onto input images, which is optimized using task labels to repurpose frozen models. When applied to VLMs [2, 7], the perturbation functions as a *input-space visual prompt*, jointly learned with textual templates for multimodal alignment. Representative methods include VP [2], AR [7, 43], and AttrVR [6]. Among them, AttrVR [6] leverages descriptive and discriminative attributes to guide optimization, reducing intra-class variance and improving inter-class separability. However, existing VR methods typically rely on a single shared visual pattern, limiting their capacity to capture heterogeneous semantics. This limitation is especially pronounced in biomedical imaging, where class boundaries are

subtle and visual cues are highly fine-grained (Figure 1a). As a result, directly transferring a single VR pattern from natural images often leads to overfitting to local textures or boundary artifacts, degrading diagnostic discrimination. These challenges motivate a confusion-aware, multi-expert VR framework tailored to biomedical tasks.

To address the above challenges, we propose BioMedVR, the first framework that introduces VR into biomedical image analysis. BioMedVR enables parameter-efficient adaptation of VLMs such as CLIP for few-shot medical image diagnosis (Table 2). BioMedVR introduces a Confusion-aware Mixture-of-Prompt Experts (MoPE) structure that explicitly models the relationship between main and confusing categories. It comprises two complementary experts: a positive expert that enhances main-class recognition using positive attributes, and a negative expert that suppresses easily confused categories via large language model (LLM)-generated confusion-aware attributes, with their outputs adaptively fused through a learnable gating module. Furthermore, we design a confusion suppression loss that explicitly penalizes small similarity margins between true and confusing categories, effectively sharpening decision boundaries and improving training stability. Extensive experiments on 11 biomedical datasets spanning 9 imaging modalities, together with 7 natural benchmarks, show that BioMedVR consistently surpasses existing prompt-learning and VR methods in both few-shot and zero-shot settings, delivering superior accuracy, robustness, and interpretability while using over 500× fewer trainable parameters than full CLIP fine-tuning. These results validate the effectiveness of integrating VR into biomedical domains and establish BioMedVR as a lightweight yet powerful paradigm for explainable few-shot medical image diagnosis. Our main contributions are as follows:

- To our knowledge, we introduce VR into biomedical imaging for the first time, enabling a decoupled adaptation paradigm where downstream tasks are handled through VR rather than model modification, allowing parameter-efficient and privacy-preserving reuse of VLMs for few-shot medical image recognition.
- We propose a confusion-aware MoPE that decouples positive and negative experts: the positive expert exploits positive attributes for main-class recognition, while the negative expert uses confusion-aware attributes to suppress ambiguity. This confusion-aware design supports both zero-shot and few-shot medical adaptation.
- We design a confusion suppression loss that explicitly penalizes easily confused categories, reducing overconfidence and improving robustness. Extensive experiments on multiple medical datasets validate the superiority and interpretability of our approach.

2 Related Work

2.1 VLM Adaptation and Prompt Learning.

Large VLMs, such as CLIP [40] and ALIGN [23], have demonstrated strong zero-shot transfer by aligning visual and textual representations in a shared embedding space [31, 32]. To further adapt these pretrained models to downstream domains, recent works have explored various parameter-efficient strategies—such as prompt learning [27, 55], adapter tuning [19, 33], low-rank adaptation [20, 52], and VR [6, 7], which introduce learnable modules while keeping the backbone frozen. Prompt learning has gained notable attention due to its simplicity and strong compatibility with VLMs.

Prompt-based tuning methods, including CoOp [55], CoCoOp [54], MaPLe [27], and PromptKD [30], introduce learnable textual or visual tokens to adapt frozen backbones efficiently, while BiomedCoOp [28] extends this paradigm to medical imaging via LLM-generated prompts. However, these approaches remain constrained by their reliance on VPT (visual prompt tuning) and linguistic assumptions, making them inadequate for modeling fine-grained, pixel-level medical features such as microvascular and tissue textures. In contrast, VR modifies the input space through lightweight, learnable perturbations, enabling architecture-agnostic, privacy-preserving, and interpretable adaptation across heterogeneous biomedical modalities.

2.2 Visual Reprogramming.

VR [2, 6, 7, 15, 43] provides a parameter-efficient alternative for adapting pretrained models by learning input-space perturbations or task-specific visual prompts without modifying model parameters. Unlike prompt-based or adapter-based methods that require access to model internals, VR operates at the input level and is therefore architecture-agnostic, compatible with both transformer and convolutional backbones [50, 51]. Recent variants such as AR [7, 43] and AttrVR [6] enhance interpretability by introducing attribute-guided or semantic-aware reprogramming, effectively bridging image–text alignment within VLMs. Guided by LLMs, VR can optimize lightweight, input-level patterns to encode task-specific knowledge while keeping pretrained weights frozen, thereby reducing computational cost and avoiding catastrophic forgetting.

Given the heterogeneity of biomedical imaging modalities (e.g., ultrasound, MRI) and strict privacy constraints [14], VR offers an ideal mechanism for model reuse—it modifies only input patterns, ensuring data security and rapid adaptation across domains and organs [28, 31]. However, existing VR methods have mainly been validated on natural image benchmarks such as DTD, and OxfordPets, and typically rely on a single shared prompt for all categories. This design limits flexibility and tends to amplify overconfidence in visually ambiguous biomedical scenarios characterized by high inter-class similarity.

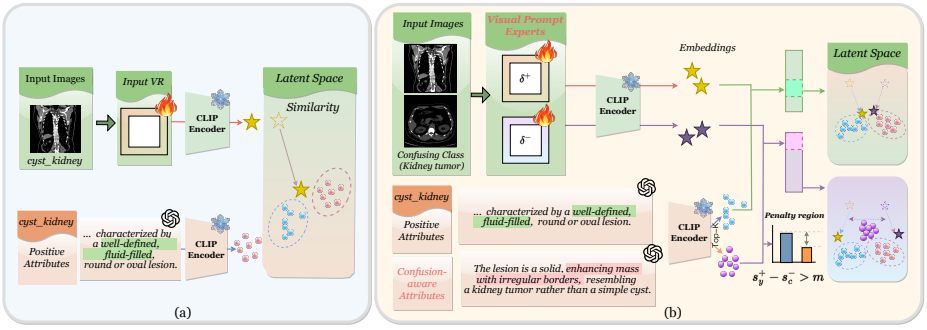

Fig. 2: Comparison between Conventional VR and BioMedVR. (a) Conventional VR methods (e.g., AttrVR) apply a single visual prompt to align input images with positive textual attributes (e.g., "well-defined, fluid-filled lesion"), but fail to handle visually similar yet semantically incorrect classes such as *kidney cyst* vs. *kidney tumor*, leading to overlapping latent embeddings. (b) In contrast, BioMedVR introduces a *Confusion-aware MoPE* that employs visual experts—positive (δ^+) and negative (δ^-)—guided by descriptive and confusion-aware attributes respectively. The penalty region suppresses inter-class confusion in the embedding space. This design explicitly models fine-grained semantic conflicts, yielding more discriminative and well-separated representations in medical image diagnosis.

3 Methodology

3.1 Preliminaries

Given a pretrained VLM, such as CLIP [40], we denote its frozen visual and textual encoders as $f_v(\cdot)$ and $f_t(\cdot)$, respectively. For an input image $x \in \mathbb{R}^{H\times W\times 3}$ and a category label $y \in \mathcal{Y}$, the corresponding textual prompt or attribute description t_y is first embedded as

$$z_v = \frac{f_v(x)}{\|f_v(x)\|}, \quad z_t = \frac{f_t(t_y)}{\|f_t(t_y)\|}, \tag{1}$$

where $z_v, z_t \in \mathbb{R}^d$ are the normalized feature embeddings in the shared multi-modal space. The similarity score between an image x and a text t_y is computed as

$$s(x, t_y) = \exp(\tau \cdot z_v^\top z_t), \tag{2}$$

where $\tau = \exp(\text{logit\_scale})$ is a learnable temperature parameter from CLIP.

Visual Reprogramming. In VR [6], instead of tuning model parameters, a learnable visual prompt δ is introduced at the input level to adapt the pretrained model to a new target domain. The reprogrammed image is formulated as

$$\tilde{x} = \mathcal{R}(x; \delta) = x \odot M + \delta \odot (1 - M), \tag{3}$$

where M denotes a binary mask controlling the region to be perturbed. The objective of VR is to learn δ such that the reprogrammed input $\tilde{x}$ can be correctly

aligned with textual representations of the target label y:

$$\mathcal{L}_{vr} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\tau\cdot f_v(\tilde{x}_i)^\top f_t(t_{y_i}))}{\sum_{y'\in\mathcal{Y}}\exp(\tau\cdot f_v(\tilde{x}_i)^\top f_t(t_{y'}))}. \tag{4}$$

Attribute-Guided Reprogramming. In practice, each class y can be described by a set of textual attributes $\mathcal{T}_y = \{\mathbf{t}_{y,1}, \mathbf{t}_{y,2}, \ldots, \mathbf{t}_{y,K}\}$ representing *descriptive* and *discriminative* semantics [6]. Accordingly, the VR objective can be reformulated as an attribute-level similarity aggregation:

$$\mathcal{L}_{attr} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\sum_{t\in\mathcal{T}_{y_i}}\exp(\tau\cdot f_v(\tilde{x}_i)^\top f_t(t))}{\sum_{y'\in\mathcal{Y}}\sum_{t'\in\mathcal{T}_{y'}}\exp(\tau\cdot f_v(\tilde{x}_i)^\top f_t(t'))}. \tag{5}$$

This formulation bridges image-text alignment at the attribute level, serving as the foundation for the proposed confusion-aware MoPE introduced in Section 3.3.

3.2 Analysis of the VR in Biomedical Imaging

In the VR paradigm, the optimization aims to learn an input-level perturbation δ that maximizes the similarity between the reprogrammed image $\tilde{x} = x + \delta$ and its corresponding text t_y while minimizing similarity to other classes:

$$\min_{\delta}\mathcal{L}_{vr} = -\log\frac{\exp(\tau\cdot f_v(\tilde{x})^\top f_t(t_y))}{\sum_{y'\in\mathcal{Y}}\exp(\tau\cdot f_v(\tilde{x})^\top f_t(t_{y'}))}. \tag{6}$$

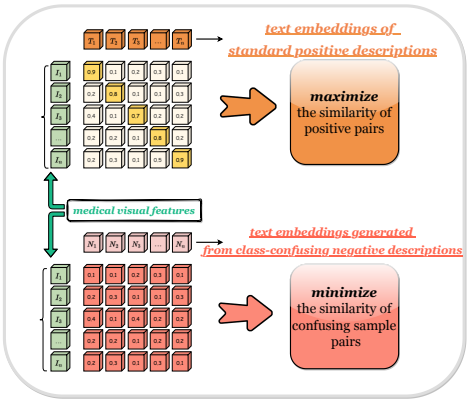

Fig. 3: Confusion-aware optimization objective. For an input image, BioMedVR learns to *maximize* similarity with descriptive positive text embeddings $\mathcal{T}_y^+$ (top) while *minimizing* similarity with confusing negative text embeddings $\mathcal{T}_y^-$ (bottom), generated from visually similar but semantically incorrect categories. This enlarges the semantic margin $s_y^+ - s_c^-$ and suppresses inter-class confusion in the embedding space.

However, directly applying this approach to biomedical image tasks raises two notable challenges. First, in biomedical imaging, categories often exhibit high inter-class similarity and low intra-class variance. Formally, for two classes $y_i, y_j \in \mathcal{Y}$ with embeddings $f_t(t_{y_i})$ and $f_t(t_{y_j})$, their cosine similarity $f_t(t_{y_i})^\top f_t(t_{y_j}) \approx 1$, especially when y_i and y_j correspond to subtle disease subtypes (e.g., *glaucoma* vs. *cataract*). As a result, the softmax distribution in Eq. (6) becomes flat and the gradients of the loss diminish, providing little discriminative signal for optimizing δ.

Second, the single shared perturbation δ is globally applied across all classes, implicitly assuming a

one-to-one alignment between visual prompts and class semantics. In practice, this can be decomposed as

$$f_v(\tilde{x}) = f_v(x+\delta) \approx f_v(x)+J_{f_v}(x)\cdot\delta, \tag{7}$$

where $J_{f_v}(x)$ is the Jacobian of the visual encoder. For heterogeneous biomedical modalities (e.g., MRI, CT), $J_{f_v}(x)$ can vary drastically across domains, causing a single δ to misalign with semantic boundaries and amplify confusion between visually similar categories.

These observations motivate the development of BioMedVR, summarized in Algorithm 1, to overcome the challenges of biomedical visual reprogramming.

3.3 Confusion-aware MoPE

We extend the visual reprogramming paradigm by introducing a *Confusion-aware MoPE*, which explicitly disentangles discriminative and confusable semantics via dual experts and margin-based optimization, as shown in Figure 2.

Given an image x with label y, two prompt experts—*positive* (δ^+) and *negative* (δ^-)—generate reprogrammed inputs:

$$\tilde{x}^+ = x + \delta^+, \qquad \tilde{x}^- = x + \delta^-, \tag{8}$$

whose normalized embeddings are

$$z^+ = \frac{f_v(\tilde{x}^+)}{\|f_v(\tilde{x}^+)\|}, \quad z^- = \frac{f_v(\tilde{x}^-)}{\|f_v(\tilde{x}^-)\|}. \tag{9}$$

The positive expert leverages descriptive and discriminative textual embeddings $\mathcal{T}_y^+$ [6], while the negative expert learns from confusion-aware textual attributes $\mathcal{T}_y^-$ automatically generated by an LLM. For each class y, we query the LLM with *"Generate the most visually confusing negative descriptions for class [Class Name]"*, obtaining

$$\mathcal{T}_y^- = \{\, t_i^- \,\}_{i=1}^{N_c}, \quad t_i^- \sim \mathrm{LLM}(y), \tag{10}$$

where N_c denotes the number of confusion-aware attributes. During optimization, the experts respectively compute their alignment scores:

$$s_y^+ = \frac{\tau}{k} \sum_{t \in \mathrm{Top}_k(\mathcal{T}_y^+)} z^{+\top} f_t(t), \quad s_y^- = \frac{\tau}{k} \sum_{t \in \mathrm{Top}_k(\mathcal{T}_y^-)} z^{-\top} f_t(t), \tag{11}$$

where τ denotes the temperature and k controls top-k attribute selection. The fused logit combines both experts through an adaptive gating mechanism:

$$s_y = g^+ s_y^+ + g^- s_y^-, \qquad [g^+, g^-] = \mathrm{softmax}(w), \tag{12}$$

where $w \in \mathbb{R}^2$ is a learnable gating vector shared across all samples. Each element of w corresponds to the confidence weight of the positive or negative expert, and is automatically optimized during training to balance discriminative enhancement and confusion suppression.

Instead of relying on any single logit, our MoPE formulation naturally induces a two-stream separation between a discriminative pathway s_y^+ and a confusion pathway $\max_{c \neq y} s_c^-$. This separation allows downstream objectives (Sec. 3.4) to directly manage the semantic margin:

$$\Gamma(x, y) = s_y^+ - \max_{c \neq y} s_c^- \quad \uparrow, \tag{13}$$

which quantifies how well the positive expert distinguishes the true class from the most confusable negatives. Our confusion-suppression loss explicitly enlarges this margin during training. Intuitively, $\max_{c \neq y} s_c^-$ can be viewed as the score of the most plausible alternative diagnosis candidate, so enlarging $\Gamma(x, y)$ directly promotes differential-diagnosis-like separation.

3.4 Confusion-Suppression Objective

The Confusion-Suppression (CS) loss is defined as:

$$\mathcal{L}_{cs} = \mathbb{E}_{(x,y)} \left[\max\big(0, \max_{c \neq y} s_c^- - s_y^+ + m\big) \right], \tag{14}$$

where m is a margin hyperparameter that defines the penalty region for confusion. Minimizing $\mathcal{L}$ enforces the constraint:

$$s_y^+ - \max_{c \neq y} s_c^- \geq m, \tag{15}$$

which explicitly enlarges the semantic margin between discriminative and confusable categories. This loss activates only when confusion occurs, preventing unnecessary suppression of legitimately ambiguous samples (Figure 3). By focusing on the strongest confusable alternative, the objective targets the most harmful errors in fine-grained biomedical recognition.

The final training objective combines standard cross-entropy with confusion suppression:

$$\mathcal{L} = \mathcal{L}_{CE} + \beta \, \mathcal{L}_{cs}, \tag{16}$$

where β balances classification accuracy and confusion robustness. This constraint reshapes the decision boundary by suppressing high-similarity negatives, yielding more stable logits and calibrated predictions in biomedical domains.

4 Experiments

4.1 Experimental Setup

Datasets. We comprehensively evaluate BioMedVR across 11 biomedical imaging datasets encompassing 10 organs and 9 imaging modalities, including Computerized Tomography (CTKidney [22]), Dermatoscopy (DermaMNIST [13, 45]),

Algorithm 1 Training Procedure of BioMedVR

Require: Frozen VLM encoders (f_v, f_t); dataset $\mathcal{D} = \{(x_i, y_i)\}$;
Descriptive & discriminative (positive) text sets $\mathcal{T}_y^+$; Confusion-aware text sets $\mathcal{T}_y^-$;
Hyperparameters $\beta, m, k, lr, Epoch$.
Ensure: Optimized visual reprogramming parameters $\{\delta^+, \delta^-, w\}$.
Initialize positive/negative experts $\{\delta^+, \delta^-\}$ and gating logits $w = [1, 1]$.
for epoch $= 1$ to $Epoch$ **do**
  for each mini-batch (x, y) in $\mathcal{D}$ **do**
    Reprogramming: $\tilde{x}^+ = x + \delta^+, \quad \tilde{x}^- = x + \delta^-$.
    Embedding: $z^+ = \frac{f_v(\tilde{x}^+)}{\|f_v(\tilde{x}^+)\|}, \; z^- = \frac{f_v(\tilde{x}^-)}{\|f_v(\tilde{x}^-)\|}$.
    Gating: $[g^+, g^-] = \mathrm{softmax}(w)$.
    Expert Logits:
    $s_y^+ = \frac{\tau}{k} \sum_{t \in \mathrm{Top}_k(\mathcal{T}_y^+)} z^{+\top} f_t(t)$,
    $s_y^- = \frac{\tau}{k} \sum_{t \in \mathrm{Top}_k(\mathcal{T}_y^-)} z^{-\top} f_t(t)$
    Fusion: $s_y = g^+ s_y^+ + g^- s_y^-, \quad p(y|x) = \frac{e^{s_y}}{\sum_c e^{s_c}}$.
    Loss: $\mathcal{L} = -\log p(y|x) + \beta[\max_{c \neq y} s_c^- - s_y^+ + m]_+$.
    Update $\{\delta^+, \delta^-, w\}$ via SGD with cosine annealing.
  end for
end for
return $\{\delta^+, \delta^-, w\}$.

Endoscopy (Kvasir [38]), Fundus Photography (RETINA [29, 39]), Histopathology (LungColon [3], CHMNIST [25]), Magnetic Resonance Imaging (BTMRI [34]), Optical Coherence Tomography (OCTMNIST [26]), Ultrasound (BUSI [1]), and X-Ray (COVID-19 [42], Knee X-ray [10]). This collection covers a broad range of diagnostic challenges—from high-contrast radiology to fine-grained histopathology—providing a rigorous testbed for cross-modal generalization. In addition, we further validate BioMedVR on seven natural image benchmarks—Caltech101 [16], Food101 [4], DTD [12], OxfordPets [37], Flowers102 [35], EuroSAT [18], and UCF101 [41]—to assess its generalization beyond the biomedical domain. Detailed dataset splits and task configurations are provided in the supplementary material.

Implementation Details. Encoders (f_v, f_t) are kept frozen, while only the $\{\delta^+, \delta^-\}$ and the gating vector w are trainable. We follow a few-shot learning setting with 4, 8, and 16 labeled samples per class, consistent with prior works [6]. Unless otherwise noted, we adopt the 16-shot setting with the remaining samples for testing and use GPT-4.1 as the default LLM. We also evaluate GPT-4o-mini, GPT-5-mini, and GPT-5 to study the impact of different LLMs on attribute generation. For generating positive attributes, we follow the procedure described in [6]. N_c and the choice of top-k follow the settings in [6]. Zero-shot BioMedVR combines positive cues with confusion-aware attribute suppression, using fused top-k similarities to select the most compatible class without training. Training is performed with CLIP backbones (ViT-B/16, ViT-B/32, and RN50) using SGD (initial learning rate 40, momentum 0.9) and cosine annealing scheduling for

Table 1: Accuracy (%) comparison of zero-shot and few-shot methods on 16-shot biomedical classification tasks using ViT-B/16–based CLIP as the pretrained model (the highest values are in **bold**; ZS denotes the zero-shot setting). In the few-shot block, rows with the light-peach background indicate VR-based methods, while the remaining rows correspond to prompt-learning approaches.

Method	BUSI	Knee X-ray	Kvasir	LungColon	OCTMNIST	BTMRI	CHMNIST	COVID-19	CT-Kidney	DermaMNIST	Retina	Avg.
Zero-shot												
CLIP [40] *[ICML 2021]*	30.5%	34.7%	18.7%	20.4%	33.2%	27.3%	32.0%	36.4%	38.1%	13.0%	25.5%	28.1%
AttrVR (ZS) *[ICLR 2025]* [6]	26.3%	38.6%	13.8%	40.8%	26.6%	44.5%	20.2%	07.6%	30.1%	09.7%	43.1%	27.4%
BioMedVR (ZS)	46.6%	38.3%	14.1%	39.3%	24.4%	46.2%	21.3%	28.4%	24.5%	13.9%	49.6%	31.6%
BioMedCLIP (ZS) *[NEJM AI 2025]* [53]	54.2%	25.5%	51.5%	38.1%	47.0%	57.2%	23.2%	59.4%	41.9%	19.3%	32.1%	40.9%
BioMedVR (ZS, BioMedCLIP)	54.2%	36.4%	62.9%	67.0%	64.7%	58.1%	36.2%	61.5%	32.9%	04.8%	45.0%	47.6%
Few-shot												
CoOp [55] *[IJCV'22]*	62.3%	27.1%	74.8%	89.6%	67.9%	79.2%	76.7%	73.6%	**81.8%**	43.0%	27.1%	63.9%
CoCoOp [54] *[CVPR'22]*	64.4%	30.0%	77.5%	89.4%	68.5%	80.4%	71.5%	76.2%	78.0%	44.0%	52.5%	66.6%
BioMedCoOp [28] *[CVPR'25]*	69.4%	36.8%	77.7%	91.5%	72.5%	81.0%	76.9%	77.0%	80.9%	59.9%	59.9%	71.2%
VP [2]	70.8%	41.6%	75.2%	90.8%	65.1%	65.9%	70.8%	67.8%	67.8%	64.9%	70.4%	68.3%
AR [7] *[CVPR'23]*	75.4%	39.4%	79.3%	94.3%	77.2%	75.7%	83.9%	70.1%	72.5%	58.0%	73.3%	72.6%
AttrVR [6] *[ICLR'25]*	78.0%	33.5%	79.4%	93.8%	**80.4%**	76.5%	**85.3%**	71.0%	71.6%	61.6%	71.5%	73.0%
BioMedVR (Ours)	**82.6%**	**45.7%**	**80.2%**	**94.7%**	80.3%	**81.7%**	84.5%	**77.4%**	74.0%	**65.3%**	**74.1%**	**76.4%**

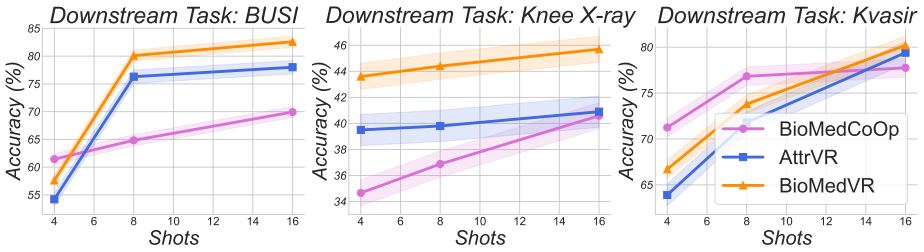

Fig. 4: Sample efficiency comparison across few-shot settings. Performance comparison of BioMedVR with AttrVR [6] and BioMedCoOp [28] on three representative biomedical datasets.

400 epochs with a batch size of 512. The confusion margin m and loss weight β are set to 0.5 and 0.3, respectively, based on hyperparameter search. We report training/inference efficiency, parameter overhead, and memory comparisons to PEFT baselines in the supplementary material. All experiments are conducted on NVIDIA H20 GPUs under identical data splits for fair comparison.

4.2 Comparison with State-of-the-Art (SOTA)

We evaluate BioMedVR under both *few-shot* and *zero-shot* settings for a comprehensive comparison. In the few-shot setting, we compare BioMedVR with SOTA VR-based VLM adaptation methods, including VP [2], which overlays learnable reprogramming patterns on resized images; AR [7, 43], which pads structured visual patterns around image borders; and AttrVR [6] performs attribute-based reprogramming using descriptive and discriminative textual cues to capture shared and class-specific semantics (GPT-5 is used since GPT-3.5 API is unavailable). To assess the adaptability of our method, we further compare with prompt-based

Table 2: Training cost of different VR methods, using the ViT-B16-based CLIP on BUSI.

	AR	AttrVR	BioMedVR	CLIP
Parameter Number	0.15M	0.15M	0.30M	150M
Training Time / Epoch (s)	7.12	6.55	9.10	–
Performance (%)	72.6	73.0	**76.4**	27.3

Table 3: Few-shot and zero-shot performance on seven natural-image benchmarks (%).

Dataset	AR	VP	AttrVR	BioMedVR	CLIP	BioMedVR (ZS)
	few-shot				*zero-shot*	
UCF101	77.7	73.2	78.6	**80.1**	59.4	**62.7**
Caltech101	95.5	93.8	**95.9**	95.7	85.4	**92.1**
Food101	85.4	82.3	86.0	**86.0**	72.8	**78.9**
DTD	59.8	61.0	65.9	**66.0**	37.4	**47.3**
EuroSAT	91.6	89.5	93.4	**93.9**	21.2	21.0
Oxford-Pets	92.6	91.3	93.1	**93.9**	77.6	**88.4**
Oxford-Flowers	86.2	83.3	**93.1**	93.0	59.6	**72.2**

baselines built on the BiomedCLIP backbone [53], including CoOp [55], CoCoOp [54], and BiomedCoOp [28]. For zero-shot evaluation, we include the original CLIP, the zero-shot variant of AttrVR [6], and zero-shot BioMedVR, which combines positive and confusion-aware textual cues without any fine-tuning.

Few-shot Setting: We compare BioMedVR with existing reprogramming methods using CLIP with the ViT-B/16. As shown in Table 1, BioMedVR consistently outperforms baseline VR approaches, including AttrVR, VP, and AR, achieving an average improvement of over 3% across 11 biomedical imaging datasets. The performance gain is particularly notable on confusion-prone datasets such as *Knee X-ray* and *DermaMNIST*, demonstrating that explicit confusion modeling effectively enhances robustness and generalization. In comparison with prompt-learning-based methods, our BioMedVR also surpasses CoOp, CoCoOp, and BiomedCoOp on most datasets. Although CoOp and BiomedCoOp perform slightly better on *CTKidney*, this likely stems from their use of the domain-specialized BiomedCLIP ViT-B/16. Nevertheless, despite leveraging a general-domain CLIP backbone, BioMedVR still achieves superior average accuracy, underscoring its strong adaptability and effectiveness in few-shot medical scenarios.

We further evaluate BioMedVR on seven natural image datasets (Table 3). It achieves the best or comparable results across all benchmarks, notably improving over AR, VP, and AttrVR on challenging datasets such as *UCF101*. These results confirm that the proposed confusion-aware reprogramming generalizes effectively beyond medical imaging, demonstrating strong cross-domain adaptability.

Zero-shot Setting. We further evaluate BioMedVR under the zero-shot setting using both CLIP and BiomedCLIP backbones. As shown in Table 1, BioMedVR (ZS) improves the average zero-shot accuracy over standard CLIP from 28.1% to 31.6%, and reaches 47.6% when built on BiomedCLIP, demonstrating strong overall performance across the 11 biomedical datasets. Notably, significant gains are observed on complex modalities such as Knee X-ray (+10.9%) and Lung-Colon (+28.9%), indicating that the confusion-aware attributes effectively enhance cross-modal generalization even without fine-tuning. Moreover, when extended to natural image benchmarks (Table 3), BioMedVR (ZS) achieves clear improvements over CLIP in zero-shot classification, particularly on *Caltech101* (+6.7%) and *Oxford-Pets* (+10.8%). These results demonstrate that BioMedVR

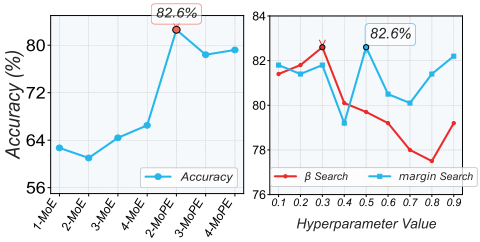

Fig. 5: Ablation on MoPE and hyperparameters. 2-MoPE yields the highest accuracy (82.6%), with the optimum at β=0.3, m=0.5.

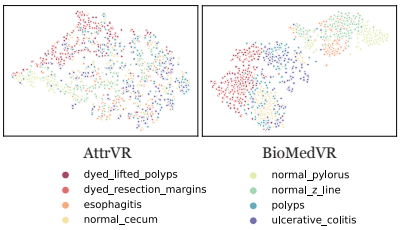

Fig. 6: t-SNE visualization on Kvasir. BioMedVR forms more compact and better-separated clusters than AttrVR.

leverages LLM-guided descriptive and confusion-aware attributes to perform reliable zero-shot reasoning, narrowing the performance gap between zero-shot and few-shot paradigms while maintaining strong transferability.

4.3 Analysis and Ablation Study

Results of Sample Efficiency. We evaluate BioMedVR under 4-, 8-, and 16-shot settings. As shown in Figure 4, BioMedVR outperforms CoOp and AttrVR across most configurations and remains competitive even with only 4 samples per class. Its performance saturates faster with increasing samples, demonstrating superior sample efficiency and effective utilization of limited supervision in data-scarce biomedical scenarios.

Results on Different Backbones. We evaluate the generality of BioMedVR across multiple CLIP backbones, including ViT-B/16, ViT-B/32, and RN50. As shown in Table 1 and Table 4, BioMedVR consistently outperforms its corresponding baselines on all backbones, demonstrating strong adaptability to both transformer- and convolution-based architectures. Compared with the baseline AttrVR, BioMedVR achieves an average improvement of +2.6% on RN50 and +1.5% on ViT-B/32, confirming its robustness even under weaker feature representations.

Visualization Examples. Figure 7 shows how BioMedVR distinguishes fine-grained semantics using both positive and confusion-aware attributes. For *Kidney Tumor*, the correct class exhibits "heterogeneous enhancement," while the confusing class *Cyst Kidney* is characterized by a "well-defined, fluid-filled lesion." By encoding cyst-like cues as confusion-aware attributes, BioMedVR learns to avoid misclassifying solid tumors as cysts. For *Cataract*, the positive attributes describe "diffuse cloudiness," whereas the confusing class *Glaucoma* presents cues such as "optic cup thinning." This contrast allows the model to suppress misleading similarities and improve differential diagnostic accuracy.

Visualization Results of Embedding Space. We visualize the t-SNE distributions of the visual embeddings on the Kvasir using the ViT-B/16. As shown in Figure 6, the embeddings produced by AttrVR (left) exhibit significant overlap

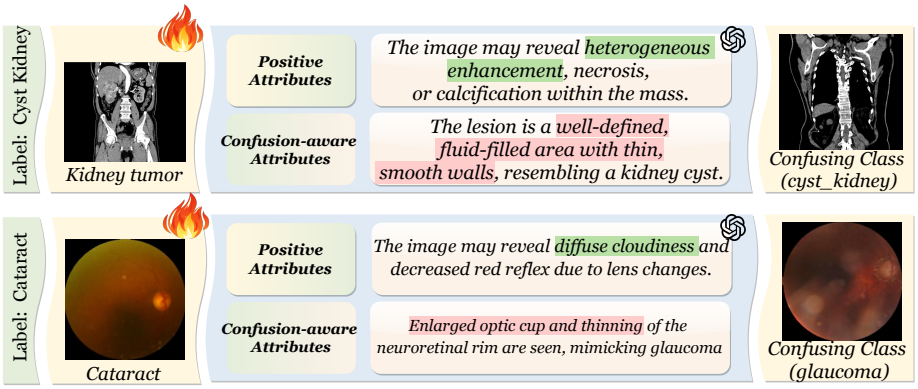

Fig. 7: Examples of confusion-aware attributes. BioMedVR constructs positive attributes for discriminative cues and LLM-generated confusion-aware attributes mimicking visually similar negatives, enabling the negative expert to suppress confusion.

Table 4: Performance (%) comparison of BioMedVR variants and baselines across 11 medical datasets. Green indicates accuracy drops relative to the corresponding BioMedVR backbone. Ablations (w/o CA (Confusion-aware Attributes), w/o MoPE, w/o CS Loss) show the contribution of each component, while different LLM-generated attribute sets (GPT-4o-mini, GPT-5-mini, GPT-5) demonstrate the robustness of BioMedVR to text–semantic sources. Results on RN50 and ViT-B/32 further confirm consistent gains across architectures.

Model	BUSI	Knee X-ray	Kvasir	Lung Colon	OCTMNIST	BTMRI	CHMNIST	COVID-19	CT-Kidney	DermaMNIST	Retina	Avg.
BioMedVR-ViT-B/16	82.6	45.7	80.2	94.7	80.3	81.7	84.5	77.4	74.0	65.3	74.1	76.4
w/o CA	81.4	39.4	78.2	94.2	80.8	80.3	85.6	76.7	27.6	67.8	73.5	71.4 (-5.0)
w/o MoPE	80.1	41.9	81.0	94.9	79.5	82.9	86.2	74.3	72.3	66.3	73.9	75.6 (-0.8)
w/o CS Loss	80.9	40.9	76.6	93.4	75.7	74.7	85.0	74.6	72.4	63.1	72.1	73.6 (-2.8)
w. gpt-4o-mini	80.5	44.1	82.7	94.9	78.1	79.1	86.1	78.3	76.9	65.8	73.5	76.4
w. gpt-5-mini	80.1	40.5	81.1	95.0	78.2	79.7	85.4	76.5	74.1	65.1	73.1	75.3
w. gpt-5	80.1	40.1	79.9	94.8	79.8	80.8	85.3	77.4	71.6	63.0	74.0	75.2
BioMedVR-**RN50**	58.5	38.5	47.4	77.7	50.3	58.9	67.0	54.8	40.7	36.6	36.9	51.6
Baseline (RN50)	50.0	36.9	44.7	74.3	33.8	49.4	68.4	48.8	40.7	32.0	59.9	49.0 (-2.6)
BioMedVR-**ViT-B/32**	65.7	36.0	62.1	89.2	29.9	70.2	74.7	55.7	61.6	54.5	66.0	60.5
Baseline (ViT-B/32)	61.0	38.2	57.8	85.6	44.4	56.3	75.0	55.9	59.3	54.0	61.6	59.0 (-1.5)

among classes such as *polyps*, *dyed_resection_margins*, and *dyed_lifted_polyps*, indicating severe inter-class confusion. In contrast, BioMedVR (right) yields more compact and clearly separated clusters, effectively enlarging the inter-class margins while preserving intra-class consistency. These results demonstrate that the BioMedVR facilitates more discriminative and semantically structured feature representations in the embedding space.

Ablation Studies. To verify the effectiveness of each component, we conduct a detailed ablation analysis across 11 medical datasets (Table 4). Removing the negative expert (*w/o MoPE*) leads to a sharp accuracy drop of 3.1% on COVID-19, confirming its role in confusion suppression. Without confusion-aware attributes (*w/o CA*), the negative descriptions fall back to NaN, causing ambiguous supervision and a 5% average decline over 11 datasets, which highlights the critical role of confusion-aware signals in stabilizing training and improving dis-

crimination. Excluding the CS Loss (*w/o CS Loss*) leads to less stable learning and higher cross-dataset variance, showing that margin calibration prevents overconfidence on ambiguous samples (Calibration metrics and reliability diagrams are provided in the supplementary material). Combining all components achieves the best overall performance, confirming the synergistic benefit of MoPE design. Beyond accuracy, we provide supplementary analyses on calibration (ECE and reliability diagrams) as well as human-vs-LLM attribute comparisons to further verify that BioMedVR is LLM-assisted rather than LLM-dependent.

Hyper-parameter Analyses. We investigate the sensitivity of BioMedVR to the CS loss weight β and margin m. As shown in Figure 5, the model achieves stable performance across a wide range of settings, with the optimal accuracy obtained at β=0.3 and m=0.5. This shows that BioMedVR is robust to moderate variations in hyperparameters and balances discrimination and confusion suppression.

MoE Framework Analyses. We compare standard *MoE* architectures with our tailored *Confusion-aware MoPE*. As shown in Figure 5, increasing the number of generic MoE experts yields limited gains, while the 2-MoPE—comprising experts specialized for positive, and confusion-aware attributes—achieves the best accuracy (82.6%). This demonstrates that semantic specialization, rather than expert quantity, is key to effective VR. Further details on MoE can be found in the supplementary material.

Effect of LLM-Generated Confusion-aware Attributes. We analyze the influence of different LLMs used for generating attribute descriptions. As shown in Table 4, BioMedVR achieves the highest average accuracy (76.4%) when using GPT-4.1-generated attributes. Alternative models (GPT-4o-mini, GPT-5-mini, GPT-5) yield slightly lower but comparable results (0% to -1.2%), indicating that BioMedVR is robust to the choice of LLM and benefits from the richer semantics captured by larger models. We further compare the performance of BioMedVR using human-generated vs. LLM-generated attributes in the supplementary material.

5 Conclusion

In this work, we present BioMedVR, a confusion-aware MoPE framework that brings VR to biomedical imaging for the first time. Unlike prompt learning, which depends on model internals and struggles with fine-grained medical ambiguity, BioMedVR provides an input-space, architecture-agnostic, and privacy-preserving adaptation mechanism for VLMs. By decoupling visual prompts into a positive expert for discriminative alignment and a negative expert guided by confusion-aware attributes, together with a confusion-suppression loss, BioMedVR effectively mitigates inter-class confusion among visually similar diseases. Experiments on 11 biomedical datasets and 7 natural-image benchmarks show that BioMedVR achieves SOTA few-shot and zero-shot performance, consistently surpassing prior VR and prompt-learning methods, with visualizations confirming clearer embedding separation. We believe this work provides a new paradigm for

efficient and interpretable medical adaptation of VLMs, paving the way toward reliable and data-efficient AI-assisted diagnosis in healthcare.

Limitations.

BioMedVR is designed for confusion-aware recognition and differential-diagnosis-inspired discrimination rather than full clinical decision support. The confusion-aware attributes are generated offline and may depend on the quality of the underlying language model and prompts. Future work will incorporate expert-curated candidate lists and clinically grounded evaluation protocols to further validate real-world diagnostic utility.

Bibliography

[1] Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020) 9

[2] Bahng, H., Jahanian, A., Sankaranarayanan, S., Isola, P.: Exploring visual prompts for adapting large-scale models. arXiv (2022) 2, 4, 10

[3] Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A., Mastorides, S.M.: Lung and colon cancer histopathological image dataset (lc25000) (2019), https://arxiv.org/abs/1912.12142 9

[4] Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: ECCV (2014) 9

[5] Cai, C., Ye, Z., Feng, L., Qi, J., Liu, F.: Sample-specific masks for visual reprogramming-based prompting. In: ICML (2024) 2

[6] Cai, C., Ye, Z., Feng, L., Qi, J., Liu, F.: Attribute-based visual reprogramming for vision-language models. In: The Thirteenth International Conference on Learning Representations (2025) 2, 4, 5, 6, 7, 9, 10, 11

[7] Chen, A., Yao, Y., Chen, P.Y., Zhang, Y., Liu, S.: Understanding and improving visual prompting: A label-mapping perspective. In: CVPR (2023) 2, 4, 10

[8] Chen, H., Wang, J., Shah, A., Tao, R., Wei, H., Xie, X., Sugiyama, M., Raj, B.: Understanding and mitigating the label noise in pre-training on downstream tasks. In: ICLR (2024) 1

[9] Chen, P.Y.: Model reprogramming: Resource-efficient cross-domain machine learning. In: AAAI (2024) 2

[10] Chen, P.: Knee osteoarthritis severity grading dataset (2018). https://doi.org/10.17632/56rmx5bjcr.1, https://www.kaggle.com/ds/3505991 9

[11] Chen, X., Lai, Z., Ruan, K., Chen, S., Liu, J., Liu, Z.: R-llava: Improving med-vqa understanding through visual region of interest. In: ICLR 2025 Workshop on Human-AI Coevolution 2

[12] Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: CVPR (2014) 9

[13] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019) 8

[14] Deng, Z., He, R., Liu, J., Wang, Y., Meng, Z., Jiang, S., Xie, Y., Liu, Z.: Med-glip: Advancing medical language-image pre-training with large-scale grounded dataset. arXiv preprint arXiv:2508.10528 (2025) 4

[15] Elsayed, G.F., Goodfellow, I., Sohl-Dickstein, J.: Adversarial reprogramming of neural networks. In: ICLR (2019) 4

[16] Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: CVPR workshop (2004) 9

[17] Hambardzumyan, K., Khachatrian, H., May, J.: Warp: Word-level adversarial reprogramming. In: ACL-IJCNLP (2021) 2

[18] Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2019) 9

[19] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International conference on machine learning. pp. 2790–2799. PMLR (2019) 4

[20] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022) 4

[21] Hung, Y.N., Yang, C.H.H., Chen, P.Y., Lerch, A.: Low-resource music genre classification with cross-modal neural model reprogramming. In: ICASSP (2023) 2

[22] Islam, M.N., Hasan, M., Hossain, M.K., Alam, M.G.R., Uddin, M.Z., Soylu, A.: Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ct-radiography. Scientific Reports **12**(1), 1–14 (2022) 8

[23] Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021) 4

[24] Jing, Y., Yuan, C., Ju, L., Yang, Y., Wang, X., Tao, D.: Deep graph reprogramming. In: CVPR (2023) 2

[25] Kather, J.N., Weis, C.A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A., Zöllner, F.G.: Multi-class texture analysis in colorectal cancer histology. Scientific reports **6**(1), 1–11 (2016) 9

[26] Kermany, D.S., Goldbaum, M., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5), 1122 – 1131.e9 (2018) 9

[27] Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: CVPR (2023) 2, 4

[28] Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: Biomedcoop: Learning to prompt for biomedical vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14766–14776 (2025) 2, 4, 10, 11

[29] Köhler, T., Budai, A., Kraus, M., Odstrcilik, J., Michelson, G., Hornegger, J.: Automatic no-reference quality assessment for retinal fundus images using vessel segmentation (06 2013). https://doi.org/10.1109/CBMS.2013.6627771 9

[30] Li, Z., Li, X., Fu, X., Zhang, X., Wang, W., Chen, S., Yang, J.: Promptkd: Unsupervised prompt distillation for vision-language models. In: CVPR (2024) 2, 4

[31] Liu, J., Hu, T., Du, J., Zhang, R., Zhou, J.T., Liu, Z.: Kpl: Training-free medical knowledge mining of vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 18852–18860 (2025) 4

[32] Liu, J., Hu, T., Xiong, H., Du, J., Feng, Y., Wu, J., Zhou, J.T., Liu, Z.: Vpl: Visual proxy learning framework for zero-shot medical image diagnosis. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 9978–9992 (2024) 4

[33] Liu, J., Hu, T., Zhang, Y., Feng, Y., Hao, J., Lv, J., Liu, Z.: Parameter-efficient transfer learning for medical visual question answering. IEEE Transactions on Emerging Topics in Computational Intelligence **8**(4), 2816–2826 (2023) 4

[34] Nickparvar, M.: Brain tumor mri dataset (2021). https://doi.org/10.34740/KAGGLE/DSV/2645886, https://www.kaggle.com/dsv/2645886 9

[35] Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian conference on computer vision, graphics & image processing (2008) 9

[36] Oh, C., Hwang, H., Lee, H.y., Lim, Y., Jung, G., Jung, J., Choi, H., Song, K.: Blackvip: Black-box visual prompting for robust transfer learning. In: CVPR (2023) 2

[37] Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: CVPR (2012) 9

[38] Pogorelov, K., Randel, K.R., Griwodz, C., Eskeland, S.L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.T., Lux, M., Schmidt, P.T., Riegler, M., Halvorsen, P.: Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on Multimedia Systems Conference. pp. 164–169. MMSys'17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3083187.3083212 9

[39] Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F.: Indian diabetic retinopathy image dataset (idrid) (2018). https://doi.org/10.21227/H25W98, https://dx.doi.org/10.21227/H25W98 9

[40] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 1, 4, 5, 10

[41] Soomro, K., Zamir, A.R., Shah, M.: A dataset of 101 human action classes from videos in the wild. Center for Research in Computer Vision (2012) 9

[42] Tahir, A.M., Chowdhury, M.E., Khandakar, A., Rahman, T., Qiblawey, Y., Khurshid, U., Kiranyaz, S., Ibtehaz, N., Rahman, M.S., Al-Maadeed, S., Mahmud, S., Ezeddin, M., Hameed, K., Hamid, T.: Covid-19 infection localization and severity grading from chest x-ray images. Computers in Biology and Medicine **139**, 105002 (2021). https://doi.org/https://doi.org/10.1016/j.compbiomed.2021.105002, https://www.sciencedirect.com/science/article/pii/S0010482521007964 9

[43] Tsai, Y.Y., Chen, P.Y., Ho, T.Y.: Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources. In: ICML (2020) 2, 4, 10
[44] Tsao, H.A., Hsiung, L., Chen, P.Y., Liu, S., Ho, T.Y.: Autovp: An automated visual prompting framework and benchmark. In: ICLR (2024) 2
[45] Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data p. 180161 (2018) 8
[46] Vinod, R., Chen, P.Y., Das, P.: Reprogramming language models for molecular representation learning. In: NeurIPS (2020) 2
[47] Wang, P., Tong, L., Wu, J., Liu, J., Liu, Z.: Fair-moe: Medical fairness-oriented mixture of experts in vision-language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 186–196. Springer (2025) 2
[48] Wang, Z., Liang, J., He, R., Wang, Z., Tan, T.: Connecting the dots: Collaborative fine-tuning for black-box vision-language models. In: ICML (2024) 1
[49] Xu, Z., Shi, Z., Wei, J., Mu, F., Li, Y., Liang, Y.: Towards few-shot adaptation of foundation models via multitask finetuning. In: ICLR (2024) 1
[50] Yang, C.H.H., Li, B., Zhang, Y., Chen, N., Prabhavalkar, R., Sainath, T.N., Strohman, T.: From english to more languages: Parameter-efficient model reprogramming for cross-lingual speech recognition. In: ICASSP (2023) 4
[51] Yang, C.H.H., Tsai, Y.Y., Chen, P.Y.: Voice2series: Reprogramming acoustic models for time series classification. In: ICML (2021) 2, 4
[52] Zanella, M., Ben Ayed, I.: Low-rank few-shot adaptation of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1593–1603 (2024) 4
[53] Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1), AIoa2400640 (2025) 10, 11
[54] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR (2022) 2, 4, 10, 11
[55] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV (2022) 2, 4, 10, 11