BioVLM: Routing Prompts, Not Parameters, for Cross-Modality Generalization in Biomedical VLMs

Mainak Singha[1], **Tanisha Gupta**[2], **Ankit Jha**[3], **Muhammad Haris Khan**[4], **Sayantani Ghosh**[5], **Biplab Banerjee**[6]
[1]University of Trento, Italy [2]Carnegie Mellon University, USA [3]LNMIIT Jaipur, India
[4]MBZUAI, UAE [5]Sunandan Divatia School of Science, Mumbai, India [6]IIT Bombay, India

Abstract

Pretrained biomedical vision-language models (VLMs) such as BioMedCLIP perform well on average but often degrade on challenging modalities where inter-class margins are small and acquisition-specific variations are pronounced, especially under few-shot supervision and when modality priors differ from pretraining corpora substantially. We propose BioVLM, a prompt-learning framework that improves cross-domain generalization without extensive backbone fine-tuning. BioVLM learns a diverse prompt bank and introduces dynamic prompt selection: for each input, it selects the most discriminative prompts via a low-entropy criterion on the predictive distribution, effectively coupling sparse few-shot evidence with rich LLM semantic priors. To strengthen this coupling, we distill high-confidence LLM-derived attributes and enforce robust knowledge transfer through strong/weak augmentation consistency. At test time, BioVLM adapts by choosing modality-appropriate prompts, enabling transfer to unseen categories and domains, while keeping training lightweight and inference efficient. On 11 MedMNIST+ 2D datasets, BioVLM achieves new state of the art across three distinct generalization settings. Codes are available at https://github.com/mainaksingha01/BioVLM.

1 Introduction

Foundation models pretrained on web-scale image-text pairs, such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), have reshaped vision-language learning by enabling strong zero-shot transfer via contrastive cross-modal alignment. Yet, their performance often collapses when deployed "as-is" in biomedicine. The root cause is a persistent domain mismatch: biomedical images (e.g., MRI, ultrasound, dermatoscopy) exhibit modality-dependent signal formation, characteristic textures, and acquisition artifacts that differ fundamentally from natural-image statistics. In addition, biomedical supervision is textually fragile reports and labels are sparse, jargon-heavy, and require domain expertise, making the language side noisier and less visually grounded than web captions, which challenges models trained on general-domain text.

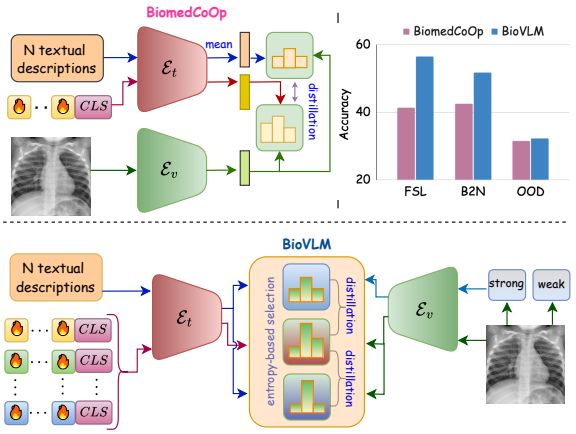

Figure 1: **Overview of our proposed BioVLM.** It selects high-confidence, optimal prompts using an entropy-based selection strategy and synergistically distills few-shot task semantics with the rich prior generic knowledge. BioVLM significantly outperforms the SOTA baseline, BioMedCoOp (Koleilat et al., 2025), across three distinct generalization settings.

To reduce this mismatch, biomedical VLMs such as BioMedCLIP (Zhang et al., 2025), PubMedCLIP (Eslami et al., 2021), and BioViL (Boecking et al., 2022) pretrain on in-domain image-text corpora, learning clinically relevant cross-modal representations. These models have enabled progress in downstream settings including pathology (Huang et al., 2023; Ikezogwo et al., 2023) and radiology (Wang et al., 2022; Wu et al., 2023), where fine-grained visual cues and structured medical semantics are critical. Consequently, they provide strong frozen backbones for medical AI, capturing complex morphology and modality-specific patterns that general VLMs typically miss.

Despite these advances, a key research gap remains: even specialized biomedical VLMs can be

Findings of the Association for Computational Linguistics: ACL 2026, pages 40986–41005
July 2-7, 2026

unreliable in zero-shot transfer to unseen conditions and modalities, especially when those distributions are underrepresented during pretraining and when clinical text contains ambiguity or semantic noise. Prompt learning is an attractive, parameter-efficient way to adapt frozen backbones, but existing methods remain limited. Approaches from CoOp (Zhou et al., 2022c) to BioMedCoOp (Koleilat et al., 2025) typically optimize a single prompt (or a small ensemble), which is insufficient to model the diversity of biomedical appearance across modalities and scanners. Moreover, BioMedCoOp integrates LLM knowledge (Achiam et al., 2023) by averaging and distilling multiple descriptions into one vector, which can propagate redundant or weakly grounded attributes and place disproportionate burden on a single prompt to generalize in unseen cases. In parallel, TPT (Shu et al., 2022) selects confident augmented views for test-time adaptation; however, for biomedical images, common augmentations (e.g., cropping/zooming) can perturb subtle diagnostic evidence and thereby bias confidence-based selection.

To address these limitations, we propose **BioVLM**, a prompt-learning framework that improves the generalization of frozen biomedical VLMs across heterogeneous medical imaging tasks (Fig. 1). Built on BioMedCLIP, BioVLM introduces *dynamic prompt engineering* with two explicit mechanisms. First, we maintain a *diverse bank of learnable prompts* and perform *low-entropy prompt selection* to choose the most informative prompts per input, rather than committing to a single prompt. Second, unlike TPT (Shu et al., 2022), we select *high-confidence textual features*: (i) we retrieve and distill the LLM descriptions that are most aligned with the input visual features (jointly under weak/strong augmentations), and (ii) we select the subset of prompts whose textual embeddings best match the input semantics in the shared vision-language space. This design filters noisy LLM supervision, reduces redundancy, and emphasizes the most discriminative cross-modal alignments. At inference, BioVLM adapts by selecting highly aligned prompts for the target modality, enabling robust transfer to unseen categories and modalities without requiring LLM plug-ins. Our primary contributions are summarized as follows:

- **A novel prompt learning framework, BioVLM,** specifically engineered to enhance the generalization of frozen biomedical VLMs across complex and unseen medical imaging modalities.
- **A low-entropy prompt selection strategy** that dynamically identifies the most discriminative prompts from a diverse learnable pool, effectively aligning few-shot task knowledge with LLM priors.
- **A robust cross-modal alignment technique** that distills high-confidence attributes from an LLM and leverages a dual strong/weak image augmentation strategy to improve knowledge transfer.
- **Demonstrated state-of-the-art performance** through extensive experiments on 11 datasets from the MedMNIST+ benchmark, where BioVLM consistently outperforms existing baselines in three distinct generalization settings (Fig. 1).

2 Related Works

(i) Biomedical Vision-Language Models: The success of generic VLMs such as CLIP (Radford et al., 2021) has motivated biomedical counterparts to reduce the mismatch between natural and clinical data. Models such as BioViL (Boecking et al., 2022), PubMedCLIP (Eslami et al., 2021), and BioMedCLIP (Zhang et al., 2025) pretrain on large-scale in-domain image-text corpora and adopt domain-aware choices (e.g., radiology-initialized language encoders, medical vocabularies, and clinical contrastive objectives) to improve cross-modal alignment. Beyond general biomedical VLMs, modality- or task-specific foundations target pathology (Huang et al., 2023; Ikezogwo et al., 2023), chest X-ray (Wang et al., 2022), radiology (Wu et al., 2023), and retinal imaging (Silva-Rodriguez et al., 2025), often using multi-modal co-attention and masking of domain-relevant terms (Gan et al., 2022) to boost zero-shot retrieval and recognition. Nevertheless, learning fine-grained, disease-specific semantics that transfer across modalities, scanners, and institutions remains challenging; as noted in (Zhou et al., 2022a), no single pretrained model covers all downstream needs.

Even specialized biomedical VLMs can be brittle under modality shift and long-tail conditions, motivating lightweight task/modality adaptation. BioVLM performs inference-time prompt routing to adapt a frozen biomedical VLM to the target modality without extensive fine-tuning.

(ii) Prompt Learning: Prompt learning adapts frozen VLMs by optimizing prompts rather than updating backbone parameters. In the general domain, CoOp (Zhou et al., 2022c) learns continuous textual prompts for few-shot classification, and CoCoOp (Zhou et al., 2022b) conditions

prompts on image features for better generalization. Later work injects structured priors (KgCoOp (Yao et al., 2023)) and semantic regularization (ProGrad (Zhu et al., 2022)). Prompting has expanded to multi-modal prompts (MaPLe (Khattak et al., 2023a), PromptSRC (Khattak et al., 2023b)), deeper prompt stacks (TCP (Yao et al., 2024)), and visual prompting (VPT (Jia et al., 2022)), and has been explored in domain adaptation (Singha et al., 2026; Monga et al., 2024), open-world segmentation (Zhao et al., 2025; Ghiasi et al., 2022), and federated learning (Lu et al.; Singha et al., 2025). In biomedicine, BioMedCoOp (Koleilat et al., 2025) builds on BioMedCLIP using LLM-generated descriptions with statistics-based selection, but prompting can also expose backdoor vulnerabilities in medical settings (Hanif et al., 2024; Khan et al., 2025).

Most prompt methods still optimize a single prompt (or small ensemble) and lack principled filtering of noisy supervision or dynamic selection under modality shift. BioVLM maintains a diverse prompt bank and uses low-entropy selection with LLM-attribute distillation to denoise and route the best prompts per input and modality.

3 Proposed Methodology

3.1 Problem Formulation

We formalize the task as follows. Let $\mathcal{D}_s = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the labeled source dataset, where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{C}_{\text{train}}$, with $|\mathcal{C}_{\text{train}}| = N_c$. We learn a classifier on $\mathcal{D}_s$ using supervision from $\mathcal{C}_{\text{train}}$. At inference, given a target dataset $\mathcal{D}_t$, for each query image $\mathbf{x} \in \mathcal{D}_t$ we compute class-wise similarity scores $\{s(\mathbf{x}, c)\}_{c \in \mathcal{C}_{\text{test}}}$ in a shared embedding space and predict $\hat{y}(\mathbf{x}) = \arg\max_{c \in \mathcal{C}_{\text{test}}} s(\mathbf{x}, c)$. Following standard biomedical evaluation protocols (Yang et al., 2023; Zhang et al., 2025), we report results under three generalization settings:

- **Few-Shot Learning:** Evaluates sample efficiency when only $\mathcal{K}$ labeled instances per class are available for training, with identical label spaces at train and test, i.e., $\mathcal{C}_{\text{train}} = \mathcal{C}_{\text{test}}$.
- **Base-to-New Generalization:** Assesses transfer to novel categories within the same modality by using disjoint label sets, $\mathcal{C}_{\text{train}} \cap \mathcal{C}_{\text{test}} = \emptyset$, where $\mathcal{C}_{\text{test}}$ contains unseen classes.
- **Out-of-Distribution (OOD) Generalization:** Tests robustness to domain shift where the target distribution differs from the source, i.e., $P_t(\mathbf{x}, y) \neq P_s(\mathbf{x}, y)$, while the label space may be shared or partially overlapping depending on the protocols.

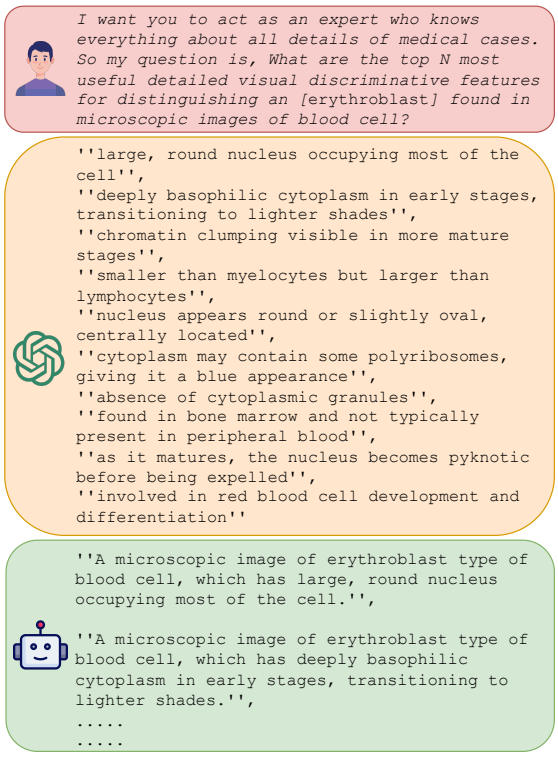

Figure 2: **LLM attribute generation.** To derive high-level clinical knowledge representations, we follow a three-stage approach. First (**top box**), an instructional query prompt is provided to a Large Language Model (LLM). In response, the LLM generates detailed visual and clinical descriptions (**middle box**). Finally (**bottom box**), we construct highly contextualized textual prompts by combining a modality-specific prefix template with the LLM-generated attributes.

3.2 The BioVLM Framework: From Static to Adaptive Prompting

Our BioVLM framework operates on a frozen CLIP backbone, which includes a vision encoder $\mathcal{E}_v$ and a text encoder $\mathcal{E}_t$. For any given image $\mathbf{x}$, a visual embedding is first extracted as $\mathbf{V} = \mathcal{E}_v(\mathbf{x})$. The core of our method lies in how we engineer the corresponding textual embeddings for robust cross-modal alignment.

We begin with a prompt structure inspired by CoOp (Zhou et al., 2022c), where a class prompt $\mathbf{t}_i$ is formed by learnable context vectors $\{\mathbf{p}_1, \ldots, \mathbf{p}_M\}$ prepended to a class token embedding $\mathbf{c}_i$. However, we posit that a single, generic prompt is insufficient for the diverse and complex nature of medical imaging. Our first key contribution is to create a **diverse prompt pool**. Using an LLM, we generate multiple rich, descriptive attributes for the symptomatic features of each medi-

cal condition. For each attribute, a unique learnable prompt is instantiated, forming a comprehensive set of textual descriptions per class.

With this diverse pool, our second contribution is a **dynamic prompt selection mechanism**. For a given image, we must identify the most suitable prompt. We achieve this with a confidence-based strategy that selects the prompt yielding the lowest entropy over the final classification probabilities. Low entropy signifies a high-confidence, unambiguous prediction, indicating a strong alignment between the image's visual evidence and the prompt's semantic meaning.

Finally, to ensure this selection process is robust and generalizes to unseen data, we introduce a **consistency-based distillation** step. We apply both strong and weak augmentations to the input image and require our model to make consistent, high-confidence predictions across these variations. By optimizing for a prompt that performs well across all views, we effectively distill robust visual features and guide the model to learn semantically meaningful representations that are invariant to superficial noise, shown in Table 4. The stages are detailed in the following.

(a) LLM attribute generation: We generate descriptive attributes for each class using GPT-4o (Hurst et al., 2024), queried with a structured instructional prompt adapted from (Menon and Vondrick, 2022) (Figure 2). To construct contextualized textual prompts, we combine a modality-specific prefix with the LLM-generated attributes. The prefix can be generic or tailored to the imaging modality (e.g., "`a microscopic image of a [class name] kidney cortex cell`"). This prefix is concatenated with an attribute using connectors such as "`which is`" or "`which has`", yielding modality-aware and semantically rich composite prompts. The third row of Figure 2 provides a complete example of such a prompt generated for the class "`erythroblast`".

(b) Diverse prompt tuning: While a single learnable prompt vector can effectively distill knowledge from multiple attributes and learn image-text alignment (Zhou et al., 2022c; Koleilat et al., 2025), this approach has inherent limitations. A single prompt tends to learn an averaged or blended representation of all attributes, which can compromise its specificity and reduce generalization, particularly when encountering unseen classes with unfamiliar feature distributions.

To overcome this, we propose a **multi-prompt framework**. Instead of a single vector, we initialize a **diverse prompt pool** for each class. For class i, we define N distinct prompts:

$$\mathbf{t}_i = \{\mathbf{t}_i^j\}_{j=1}^N, \quad \mathbf{t}_i^j = \{\mathbf{p}_1^j, \ldots, \mathbf{p}_M^j, \mathbf{c}_i\},$$

where each prompt combines learnable context vectors with the class token. For a given image, this pool produces N textual features, enabling multiple vision-language alignments, but raising the question: *how to select the most reliable prompt for accurate prediction?*

(c) Entropy-based prompt selection: To select the most discriminative prompt from the pool, we use an entropy-based mechanism: prompts that align well with an image's visual features yield low-entropy (high-confidence) class distributions. For a given image with visual features $\mathbf{V}$, the probability distribution from prompt variant j is computed as the softmax over cosine similarities:

$$p^j(y|\mathbf{x}) = \frac{\exp(\cos(\mathbf{V}, \mathbf{T}_y^j)/\beta)}{\sum_{i=1}^K \exp(\cos(\mathbf{V}, \mathbf{T}_i^j)/\beta)} \quad (1)$$

Here, $\mathbf{T}_i^j = \mathcal{E}_t(\mathbf{t}_i^j)$ denotes the feature vector of the j-th prompt for class i, K is the total number of classes, and β is a temperature hyperparameter. The self-entropy of this distribution, measuring its uncertainty, is defined as:

$$\mathrm{H}(p^j) = -\sum_{i=1}^K p^j(y_i|\mathbf{x}) \log p^j(y_i|\mathbf{x}). \quad (2)$$

Rather than using a fixed entropy threshold, which may not adapt well across different images or datasets, we employ a more robust percentile-based strategy. We calculate the entropy $\mathrm{H}(p^j)$ for all N prompts and identify the entropy value at a given cutoff percentile, ρ. Let this threshold be τ_ρ. We then select only the prompts whose entropy is below this threshold. The final class prediction probability is the normalized average of the probabilities from these high-confidence prompts:

$$p(y|\mathbf{x}) = \frac{\sum_{j=1}^N \mathbb{1}[\mathrm{H}(p^j) \leq \tau_\rho] \cdot p^j(y|\mathbf{x})}{\sum_{j=1}^N \mathbb{1}[\mathrm{H}(p^j) \leq \tau_\rho]}, \quad (3)$$

where $\mathbb{1}[\cdot]$ is the indicator function. This ensures that only reliable prompts influence the final decision, improving robustness and generalization.

3.3 Training & Inference

For training, we employ a composite loss to align learnable prompts with LLM-derived semantics.

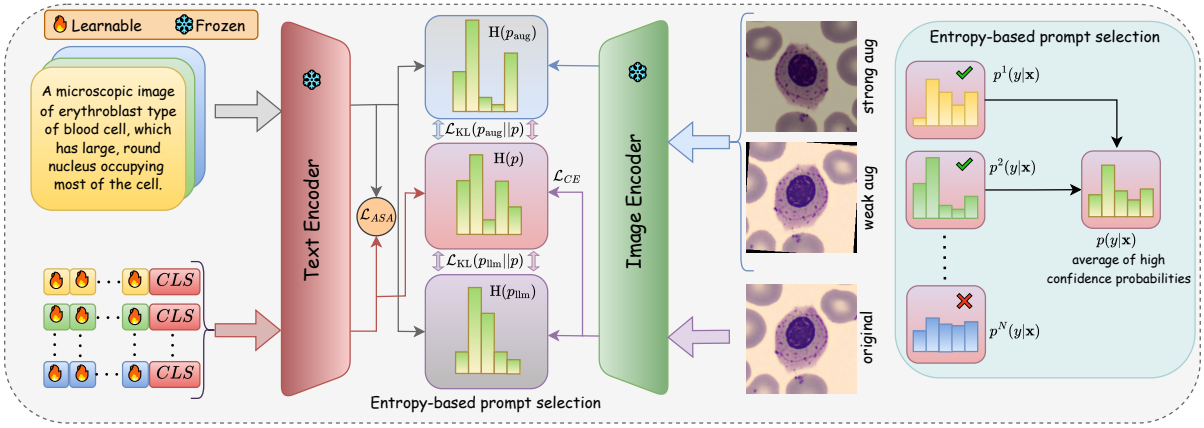

Figure 3: **Model architecture of BioVLM.** The proposed framework enhances the generalization capability of the pretrained BioMedCLIP by integrating learnable prompts with LLM-derived attributes processed through a frozen text encoder. An entropy-based selection strategy identifies the most discriminative prompts. Original, weakly augmented, and strongly augmented images are encoded using a frozen image encoder. The model is trained using cross-entropy loss ($\mathcal{L}_{CE}$), low-entropy regularization losses, KL-divergence distillation losses ($\mathcal{L}_{KL}$), and an alignment loss ($\mathcal{L}_{ASA}$). The final prediction is obtained by averaging high-confidence outputs.

We extract frozen text features for the N attributes of each class i using the text encoder $\mathcal{E}_t$, denoted as $\mathbf{T}_i^{\text{llm}} = \{\mathbf{T}_i^{\text{llm},(j)}\}_{j=1}^{N}$, which serve as semantic targets for prompt learning.

- Attribute-Semantic alignment ($\mathcal{L}_{ASA}$): A core component of our method is the one-to-one mapping between the N learnable prompts and the N LLM-generated attributes for each class. This correspondence ensures that each prompt specializes in distilling knowledge from a specific attribute, preventing the feature averaging that can dilute discriminative information. To enforce this, our Attribute-Semantic Alignment loss maximizes the cosine similarity between each learnable prompt's features ($\mathbf{T}_i^j$) and its corresponding LLM attribute's features ($\mathbf{T}_i^{\text{llm},(j)}$). This is equivalent to minimizing the negative similarity:

$$\mathcal{L}_{ASA} = -\sum_{i=1}^{K}\sum_{j=1}^{N} \frac{\mathbf{T}_i^j \cdot \mathbf{T}_i^{\text{llm},(j)}}{\|\mathbf{T}_i^j\| \cdot \|\mathbf{T}_i^{\text{llm},(j)}\|}. \quad (4)$$

- Low-Entropy regularization ($\mathcal{L}_{LER}$): To improve the model's prediction confidence and discriminative ability, we apply an entropy minimization loss to three key probability distributions.

(a) Student Confidence: We minimize the entropy of the final prediction $p(y|\mathbf{x})$ (from Eq. 3) to encourage confident and unambiguous classifications.

(b) Teacher Confidence: Since not all LLM-generated attributes are equally discriminative, we additionally enforce confidence on the teacher distributions. Specifically, $p_{\text{llm}}(y \mid \mathbf{x})$ is computed by matching the image embedding with frozen LLM-derived attribute embeddings, while the augmentation-based distribution is defined as $p_{\text{aug}}(y \mid \mathbf{x}) = \frac{1}{2}\big(p(y \mid \mathbf{x}^w) + p(y \mid \mathbf{x}^s)\big)$. We minimize the entropy of both p_{llm} and p_{aug} to encourage confident and reliable teacher supervision.

The combined low-entropy regularization is formulated in Eq. 5 with $\mathrm{H}(\cdot)$ is the entropy function defined in Eq. 2.

$$\mathcal{L}_{LER} = \mathrm{H}(p) + \mathrm{H}(p_{\text{llm}}) + \mathrm{H}(p_{\text{aug}}), \quad (5)$$

- Cross-Modal Knowledge Distillation ($\mathcal{L}_{CMD}$): We transfer semantic knowledge from LLM-derived distributions (teachers) to the learnable prompt-based model (student) by minimizing the Kullback-Leibler (KL) divergence between student logits p and both the LLM-based teacher p_{llm} and the augmented-view teacher p_{aug}. This encourages the student to match the teachers' confident and robust predictions:

$$\mathcal{L}_{CMD} = \mathcal{L}_{\text{KL}}(p_{\text{llm}}||p) + \mathcal{L}_{\text{KL}}(p_{\text{aug}}||p). \quad (6)$$

Overall Training Objective: The final training objective combines the standard cross-entropy classification loss ($\mathcal{L}_{CE}$) with our proposed regularization and distillation terms, weighted by hyperparameters $\lambda_1, \lambda_2, \lambda_3$:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1\mathcal{L}_{ASA} + \lambda_2\mathcal{L}_{LER} + \lambda_3\mathcal{L}_{CMD}. \quad (7)$$

From a ***generalization theory perspective***, our combined loss $\mathcal{L}$ is designed to minimize the true risk, $\mathcal{R}_{true}(f)$, by tackling both terms in the standard learning bound: $\mathcal{R}_{true}(f) \leq \mathcal{R}_{emp}(f) +$

$\Omega(\mathcal{H})$. While the cross-entropy loss ($\mathcal{L}_{CE}$) directly minimizes the empirical risk $\mathcal{R}_{emp}(f)$, the remaining objectives synergistically regularize the model to reduce the complexity term $\Omega(\mathcal{H})$. The attribute-semantic alignment loss ($\mathcal{L}_{ASA}$) prunes the hypothesis space $\mathcal{H}$ to a lower-complexity, semantically meaningful subspace defined by the LLM priors. Concurrently, the low-entropy loss ($\mathcal{L}_{LER}$) acts as a large-margin regularizer, pushing for more decisive predictions, which is known to improve generalization bounds. Finally, the knowledge distillation loss ($\mathcal{L}_{CMD}$) serves as a functional regularizer, encouraging the student model to find smoother functions that occupy flatter minima in the loss landscape by mimicking a stable teacher. This multifaceted regularization strategy ensures that the model learns a function that is not just accurate on training data but also simple, decisive, and smooth, all hallmarks of a model with a tight generalization bound and robust performance on unseen data.

Inference: During inference, our method is highly efficient, relying only on optimized learnable prompts and entropy-based selection. LLM-based attribute generation and image augmentations are used exclusively during training, substantially reducing test-time computational overhead.

4 Experimental Results

Dataset. We evaluate our method on 11 diverse 2D medical imaging datasets from the MedMNIST+ benchmark (Yang et al., 2023), covering nine imaging modalities and a wide range of anatomical structures and clinical protocols. The multi-label ChestMNIST dataset is excluded from the 2D setting. For all experiments, we use the standard 224×224 resolution images. A detailed description of each dataset is available in the `Appendix`.

Implementation Details. We conduct all experiments on a single NVIDIA A6000 GPU, using ViT-B/16 from pretrained BioMedCLIP (Zhang et al., 2025) as the backbone and GPT-4o (Achiam et al., 2023) for attribute generation. We report results averaged over three independent runs and train all models for 50 epochs across generalization settings. We set $N = 10$ prompts per category, each of length $M = 4$, with $\lambda_1 = \lambda_3 = 1$ and $\lambda_2 = 0.5$. We train using SGD with a learning rate of 2×10^{-3}, batch size 32, and $\mathcal{K} = 16$ shots, together with a cosine scheduler and a 1-epoch warm-up at 1×10^{-5}. We apply strong augmentations including random horizontal flip, color jitter, and Gaussian blur, while weak augmentations use horizontal flip and 10° rotation.

4.1 Comparison with the state-of-the-art methods

Few-shot learning: We evaluate BioVLM on the few-shot learning (FSL) task, considering only textual prompt learning methods as baselines for fair comparison (Koleilat et al., 2025). As shown in Table 1, BioVLM consistently outperforms prior methods, achieving at least an 8.36% gain over the second-best baseline. By leveraging entropy-based prompt selection and LLM-driven attribute alignment, our proposed BioVLM effectively addresses challenges such as class imbalance and low inter-class variance. Compared to deeper prompting approaches e.g., MaPLe, PromptSRC, TCP, it exhibits stronger generalization with lower complexity. Notably, BioVLM outperforms CoOp by 29.82% on BloodMNIST and TissueMNIST. While gains are smaller on RetinaMNIST and BreastMNIST due to dataset bias, BioVLM shows strong improvements in Base-to-New generalization.

Base-to-New (B2N) Generalization: We evaluate the model's ability to recognize novel categories within the same imaging modality, a stringent test of semantic generalization (Table 2). BioVLM achieves the highest harmonic mean (HM), outperforming prior baselines by an average of +8.04% across PathMNIST, DermaMNIST, and BloodMNIST. These gains demonstrate the effectiveness of entropy-guided prompt selection and attribute-aligned representations. PneumoniaMNIST and BreastMNIST are excluded, as each contains only one class in either the base or novel split.

The performance remains challenging on TissueMNIST and OrganMNIST due to pronounced intra-modality variation, where novel classes differ structurally and functionally from base classes (e.g., "bladder" vs. "lung"). This can limit prompt specialization and reduce performance relative to zero-shot BioMedCLIP. Nonetheless, BioVLM narrows this gap, confirming the utility of its LLM-grounded multi-prompt strategy. These results suggest future directions such as integrating domain-invariant regularization or hierarchical priors to handle anatomically diverse tasks.

Out-of-distribution Generalization: In the OOD generalization setting, we train models on a single source dataset and evaluate them on all remaining datasets (Bose et al., 2024). Table 3 reports the average accuracy across unseen target domains. BioVLM achieves the best overall per-

Table 1: **Comparison of methods on the few-shot learning task of MedMNIST+ benchmark.**

Dataset	BioMedCLIP NEJM AI'25	CoOp IJCV'22	CoCoOp CVPR'22	KgCoOp CVPR'23	MaPLe CVPR'23	ProGrad ICCV'23	PromptSRC ICCV'23	TCP CVPR'24	BiomedCoOp CVPR'25	BioVLM (Ours)	Δ (in %)
Train Params	0	3K	44K	3K	5.3M	3K	69K	0.5M	3K	30K	-
PathMNIST	43.30	76.93	75.92	76.61	75.50	60.26	76.42	76.21	72.59	**81.56**	+4.63
DermaMNIST	36.21	28.78	40.55	30.09	37.53	35.88	39.80	39.97	30.02	**45.27**	+4.72
OCTMNIST	36.80	59.20	60.47	57.67	60.76	47.03	59.35	59.92	54.77	**62.57**	+1.81
PneumoniaMNIST	59.46	72.70	76.01	73.24	74.81	57.64	75.06	72.48	70.51	**79.65**	+3.64
RetinaMNIST	35.25	31.58	**40.92**	30.83	38.45	29.17	38.95	39.68	34.50	40.08	-0.84
BreastMNIST	33.33	60.90	**67.31**	60.90	61.36	44.66	58.24	59.24	58.76	65.60	-1.71
BloodMNIST	12.66	40.33	33.75	39.84	42.78	21.78	38.17	27.56	22.98	**70.15**	+27.37
TissueMNIST	15.70	22.68	17.45	20.81	18.35	18.76	20.43	20.87	14.81	**27.34**	+4.66
OrganAMNIST	24.99	39.97	46.39	45.72	39.46	33.93	47.27	43.23	36.60	**54.75**	+7.48
OrganCMNIST	22.37	40.91	36.98	40.22	37.45	28.52	38.42	39.02	30.63	**49.22**	+8.31
OrganSMNIST	24.47	38.46	33.93	37.62	34.68	27.92	32.18	30.65	28.92	**45.41**	+6.95
Average	31.32	46.59	48.15	46.69	47.38	36.87	47.66	46.26	41.37	**56.51**	+8.36

Table 2: **Comparison of methods on the Base-to-New generalization task of MedMNIST+ benchmark.**

Dataset	Sets	BioMedCLIP NEJM AI'25	CoOp IJCV'22	CoCoOp CVPR'22	KgCoOp CVPR'23	MaPLe CVPR'23	ProGrad ICCV'23	PromptSRC ICCV'23	TCP CVPR'24	BioMedCoOp CVPR'25	BioVLM (Ours)	Δ (in %)
Average on 9 datasets	Base	36.45	42.32	41.62	42.65	43.42	35.90	43.49	43.03	40.01	**57.84**	+14.35
	New	43.49	44.97	42.20	44.86	44.67	44.54	45.35	44.39	45.40	**46.86**	+1.46
	H	39.66	43.60	41.91	43.73	44.04	39.76	44.40	43.70	42.54	**51.77**	+7.37
PathMNIST	Base	56.62	86.13	70.42	85.13	78.91	53.68	80.31	75.21	74.59	**92.73**	+6.60
	New	45.83	59.08	47.35	60.03	55.23	57.40	52.45	57.12	56.56	**63.39**	+3.36
	H	50.66	70.09	56.63	70.41	64.98	55.48	63.46	64.93	64.33	**75.30**	+4.89
DermaMNIST	Base	26.70	34.06	32.85	35.60	30.58	36.16	34.15	29.60	28.23	**52.83**	+16.67
	New	62.08	83.76	82.93	83.60	82.59	82.63	83.12	82.32	81.17	**83.87**	+0.11
	H	37.34	48.43	47.06	49.93	44.63	50.31	48.41	43.54	41.89	**64.83**	+14.52
OCTMNIST	Base	57.80	70.47	75.53	70.27	75.82	65.27	**76.07**	75.28	69.40	75.33	-0.74
	New	**52.80**	50.93	48.20	50.67	49.52	51.00	47.10	48.25	48.60	46.40	-6.40
	H	55.19	59.13	58.85	58.88	**59.91**	57.26	58.18	58.81	57.17	57.43	-2.48
RetinaMNIST	Base	45.83	36.00	44.87	38.46	44.29	31.52	42.50	41.98	40.81	**49.14**	+3.31
	New	39.77	55.15	59.47	53.26	60.74	**62.54**	62.06	62.87	57.74	61.74	-0.80
	H	42.59	43.57	51.15	44.67	51.23	41.91	50.45	50.34	47.82	**54.73**	+3.50
BloodMNIST	Base	36.63	33.16	30.96	34.15	34.83	35.11	35.51	36.02	32.35	**68.94**	+32.31
	New	23.81	41.21	27.22	42.20	42.48	37.00	46.23	40.72	50.07	**51.19**	+1.12
	H	28.86	36.75	28.97	37.75	38.28	36.03	40.17	38.23	39.31	**58.76**	+18.59
TissueMNIST	Base	6.53	16.52	12.27	15.62	20.48	7.58	17.43	21.70	9.79	**32.84**	+12.36
	New	**36.66**	26.22	20.09	25.86	19.32	22.12	23.98	18.75	22.90	19.03	-17.63
	H	11.09	20.27	15.23	19.47	19.88	11.29	20.19	20.12	13.71	**24.10**	+3.83
OrganAMNIST	Base	29.53	31.99	33.38	32.31	35.28	28.76	32.37	36.81	33.71	**48.00**	+11.19
	New	**48.40**	32.65	33.63	33.15	29.14	33.23	35.18	27.05	35.38	32.97	-15.43
	H	36.68	32.31	33.50	32.72	31.92	30.84	33.72	31.18	34.52	**39.09**	+2.41
OrganCMNIST	Base	32.67	36.81	37.67	37.00	35.46	32.59	36.02	34.19	36.01	**52.17**	+14.50
	New	**40.04**	31.65	32.07	29.92	34.67	27.84	32.54	35.40	27.93	33.17	-6.87
	H	35.98	34.04	34.65	33.08	35.06	30.03	34.19	34.78	31.46	**40.55**	+4.57
OrganSMNIST	Base	35.73	35.71	36.67	35.35	35.13	32.44	37.08	36.49	35.24	**48.54**	+11.46
	New	**42.01**	24.03	28.83	25.06	28.38	27.09	25.45	27.03	28.29	29.98	-12.03
	H	**38.62**	28.73	32.28	29.33	31.40	29.52	30.18	31.06	31.39	37.06	-1.56

formance, outperforming all prompt-learning baselines by +0.82% over the second-best method. Although margins are smaller than in other settings, BioVLM shows consistent generalization across diverse unseen medical imaging tasks, indicating that its entropy-based prompt selection and LLM-derived attributes yield domain-robust representations. Compared to CoOp and CoCoOp, which often overfit to domain-specific artifacts, BioVLM maintains strong performance on structurally diverse datasets such as RetinaMNIST, TissueMNIST, and OrganMNIST. While minor drops are observed on PathMNIST and BreastMNIST due to large domain shifts, the results confirm BioVLM's strong OOD transfer ability. Detailed per-source results are provided in the `Appendix`.

4.2 Ablation Experiments

Effect of different loss components: Table 4 presents an ablation study on Base-to-New generalization across all datasets, analyzing the impact of different loss components in Eq. 7. Using only cross-entropy loss ($\mathcal{L}_{\text{CE}}$) yields suboptimal performance (HM 45.01%), indicating that standard classification objectives are insufficient for robust biomedical generalization. Adding $\mathcal{L}_{\text{LER}}$ significantly improves performance, with gains of 5.10% and 2.38% on base and novel classes, respectively, highlighting its role in filtering uncertain prompts and promoting confident predictions. Incorporating $\mathcal{L}_{\text{CMD}}$ further boosts HM, demonstrating effective semantic transfer from LLM-derived embeddings to learned prompts. The full loss formulation

Table 3: **Comparison of methods on the out-of-distribution generalization task of MedMNIST+ benchmark.**

Source	CoOp IJCV'22	CoCoOp CVPR'22	KgCoOp CVPR'23	MaPLe CVPR'23	ProGrad ICCV'23	PromptSRC ICCV'23	TCP CVPR'24	BioMedCoOp CVPR'25	BioVLM (Ours)	Δ (in %)
PathMNIST	31.94	29.68	32.43	32.51	32.08	32.11	32.18	**32.72**	31.55	-1.17
DermaMNIST	22.73	26.67	23.29	22.79	22.86	24.42	23.17	29.38	**29.41**	+0.03
OCTMNIST	31.56	30.53	31.50	31.95	31.03	32.43	32.47	31.45	**32.89**	+0.42
PneumoniaMNIST	21.18	26.05	22.12	25.39	23.34	25.01	26.15	26.89	**28.19**	+1.30
RetinaMNIST	**34.68**	30.06	33.18	31.68	33.65	32.24	31.99	32.92	31.37	-3.31
BreastMNIST	29.27	25.54	29.78	30.16	30.14	**31.02**	30.75	30.35	28.93	-2.09
BloodMNIST	29.57	28.45	28.50	31.32	28.84	32.60	31.80	28.21	**30.38**	+0.81
TissueMNIST	31.06	30.73	34.88	32.96	33.60	34.14	32.58	34.91	**36.28**	+1.40
OrganAMNIST	33.83	**34.70**	33.02	32.44	34.68	33.30	33.80	33.51	34.34	-0.36
OrganCMNIST	34.20	35.47	33.55	34.09	33.83	34.23	34.55	32.62	**36.57**	+1.10
OrganSMNIST	34.78	32.38	35.04	33.33	34.32	33.24	32.68	33.74	**35.83**	+0.79
Average	30.44	30.02	30.66	30.78	30.76	31.34	31.10	31.52	**32.34**	+0.82

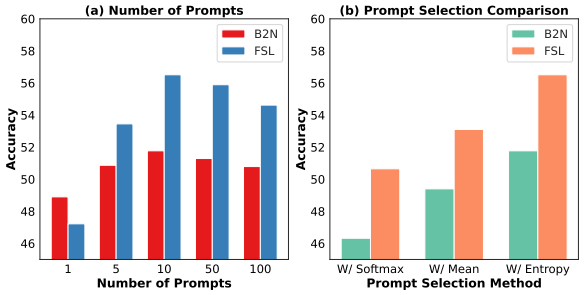

Figure 4: **Model ablation: Varying (a) number of prompts, (b) prompt selection methods (W/ = with).**

defined in Eq. 7 achieves the best performance, including an additional 0.74% improvement on novel classes. Similar trends are observed in few-shot learning, where combining all loss terms consistently improves accuracy, with performance increasing as the number of shots grows.

Table 4: **Effect of the loss functions on B2N generalization and FSL tasks. $\mathcal{K}$ is the number of shots.**

$\mathcal{L}_{CE}$	$\mathcal{L}_{LER}$	$\mathcal{L}_{CMD}$	$\mathcal{L}_{ASA}$	Base	Novel	HM	$\mathcal{K}$=1	$\mathcal{K}$=4	$\mathcal{K}$=8	$\mathcal{K}$=16
✓	×	×	×	46.82	43.34	45.01	35.07	41.27	47.82	53.26
✓	✓	×	×	51.92	45.72	48.62	37.21	43.11	48.72	54.67
✓	×	✓	×	52.74	44.65	48.36	36.24	42.89	48.14	53.98
✓	×	×	✓	50.56	44.10	47.11	34.78	41.66	46.22	53.78
✓	✓	✓	×	54.48	45.37	49.51	39.61	47.59	51.35	56.05
✓	×	✓	✓	55.20	45.82	50.07	38.56	45.72	51.19	55.34
✓	✓	×	✓	55.93	46.12	50.55	39.87	46.45	50.78	55.92
✓	✓	✓	✓	**57.84**	**46.86**	**51.77**	**40.23**	**47.94**	**51.70**	**56.51**

Ablation with number of learnable prompts: To study the effect of prompt diversity, we evaluate performance with varying numbers of learnable prompts per class (Figure 4(a)), reporting averages across all datasets. Increasing prompts from 1 to 10 steadily improves accuracy, from 48.9% to 51.77% on Base-to-New generalization and from 47.22% to 56.51% on FSL. This demonstrates that greater prompt diversity enhances generalization by providing multiple contextual views per class. However, gains saturate beyond 10 prompts, with performance slightly degrading at 50 and 100 prompts for example, B2N drops to 50.79% at 100, likely due to redundant or noisy prompt variants.

Performance against prompt selection methods: BioVLM enhances generalization by selecting the most informative prompts from a prompt pool. We compare entropy-based selection with softmax- and mean-based strategies (Figure 4(b)). The entropy-based approach consistently outperforms softmax by 5.45% (Base-to-New) and 5.86% (FSL), and mean-based selection by 2.33% and 3.36%, respectively, averaged across all datasets. Additional ablations on λ sensitivity and the impact of different LLMs are provided in the `Appendix`.

Computational complexity: Table 5 demonstrates that BioVLM attains superior performance in base-to-new (B2N) generalization and fewshot learning (FSL) settings, while utilizing substantially fewer trainable parameters compared to existing methods such as MaPLe, PromptSRC, and TCP.

Table 5: **Ablation with the trainable parameters.**

Method	Params. (in Millions)	B2N	FSL
CoOp	0.0031	43.60	46.59
CoCoOp	0.0448	41.91	48.15
KgCoOp	0.0031	43.73	46.69
MaPLe	5.3460	44.04	47.38
ProGrad	0.0031	39.76	36.87
PromptSRC	0.0690	44.40	47.66
TCP	0.4965	43.70	46.26
BioMedCoOp	0.0031	42.54	41.37
BioVLM (Ours)	0.0307	**51.77**	**56.51**

5 Conclusions

We introduce BioVLM, a prompt-based framework that significantly enhances the generalization of biomedical vision-language models. BioVLM, a novel framework, works by dynamically selecting the most informative prompts from a diverse pool, synergistically combining attribute alignment, entropy regularization, and knowledge distillation, guides the model to learn confident and semantically grounded representations. Building on the results, future work can develop methods to automatically refine or reduce the dependency on LLM-generated attributes, and extending the framework to 3D volumetric data.

6 Limitations

BioVLM relies on LLM-generated attributes to guide prompt learning, which introduces additional computational overhead during the attribute extraction stage and makes the approach dependent on the quality of the generated descriptions, although this process is performed offline. In addition, our experiments focus on 2D medical imaging datasets, and extending the framework to volumetric modalities like CT or MRI may require additional architectural and computational considerations. Finally, the use of multiple learnable prompts per class increases training complexity, which may affect efficiency in large-scale or highly diverse biomedical settings.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. 2019. Dataset of breast ultrasound images. data brief 28, 104863 (2020). *Crossref, Web of Science*.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, and 1 others. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer.

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

Shirsha Bose, Ankit Jha, Enrico Fini, Mainak Singha, Elisa Ricci, and Biplab Banerjee. 2024. Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5542–5552.

Pingjun Chen. 2018. Knee osteoarthritis severity grading dataset. *Mendeley Data*, 1(10.17632):30784984.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, and 1 others. 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*.

Sedigheh Eslami, Gerard De Melo, and Christoph Meinel. 2021. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*.

Zhe Gan, Yen-Chun Chen, Linjie Li, Yu Cheng, Jingjing Liu, and Lijuan Wang. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends in Computer Graphics and Vision*, 14(3):163–352.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Tsung-Yi Lin, Quoc V Le, Zhengyuan Deng, and Xiaohua Zhai. 2022. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Asif Hanif, Fahad Shamshad, Muhammad Awais, Muzammal Naseer, Fahad Shahbaz Khan, Karthik Nandakumar, Salman Khan, and Rao Muhammad Anwer. 2024. Baple: Backdoor attacks on medical foundational models using prompt learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 443–453. Springer.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. 2024. 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.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. 2023. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. 2023. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017.

Md Nazmul Islam, Mehedi Hasan, Md Kabir Hossain, Md Golam Rabiul Alam, Md Zia Uddin, and Ahmet Soylu. 2022. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ct-radiography. *Scientific Reports*, 12(1):11440.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer.

Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. 2016. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6(1):27988.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, and 1 others. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131.

Momin Ahmad Khan, Yasra Chandio, Eugene Bagdasarian, and Fatima Anwar. 2025. Compromising federated medical ai-backdoor risks in prompt learning. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pages 630–631.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023a. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2023b. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200.

Konwoo Kim, Michael Laskin, Igor Mordatch, and Deepak Pathak. 2021. How to adapt your large-scale vision-and-language model.

Thomas Köhler, Attila Budai, Martin F Kraus, Jan Odstrčilik, Georg Michelson, and Joachim Hornegger. 2013. 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.

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

Wang Lu, HU Xixu, Jindong Wang, and Xing Xie. Fedclip: Fast generalization and personalization for clip in federated learning. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

Sachit Menon and Carl Vondrick. 2022. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*.

Munish Monga, Sachin Kumar Giroh, Ankit Jha, Mainak Singha, Biplab Banerjee, and Jocelyn Chanussot. 2024. Cosmo: Clip talks on open-set multi-target domain adaptation. *arXiv preprint arXiv:2409.00397*.

Msoud Nickparvar and 1 others. 2021. Brain tumor mri dataset.

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, and 1 others. 2017. 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.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. 2018. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. *arXiv preprint arXiv:2209.07511*.

Julio Silva-Rodriguez, Hadi Chakor, Riadh Kobbi, Jose Dolz, and Ismail Ben Ayed. 2025. A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Medical Image Analysis*, 99:103357.

Mainak Singha, Sarthak Mehrotra, Paolo Casari, Subhasis Chaudhuri, Elisa Ricci, and Biplab Banerjee. 2026. Clipoint3d: Language-grounded few-shot unsupervised 3d point cloud domain adaptation. *arXiv preprint arXiv:2602.20409*.

Mainak Singha, Subhankar Roy, Sarthak Mehrotra, Ankit Jha, Moloud Abdar, Biplab Banerjee, and Elisa Ricci. 2025. Fedmvp: Federated multimodal visual prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17869–17878.

Anas M Tahir, Muhammad EH Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M Sohel Rahman, Somaya Al-Maadeed, and 1 others. 2021. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in biology and medicine*, 139:105002.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):180161.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Junhao Yang, Rui Shi, Dapeng Wei, Zai Wang, Ling Wang, and Jiayu Zhou. 2023. Medmnist v2: A large-scale lightweight benchmark for biomedical image classification. *IEEE Transactions on Medical Imaging*.

Hantao Yao, Rui Zhang, and Changsheng Xu. 2023. 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.

Hantao Yao, Rui Zhang, and Changsheng Xu. 2024. 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.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*.

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

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2025. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640.

Minfan Zhao, Ziqi Zhu, Jun Shi, Zhaohui Wang, Junshi Chen, Hong An, and Bing Yan. 2025. Promptseg: Learning to segment medical image via visual prompts. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Hong-Yu Zhou, Cheng Li, Zhihong Jiang, Felix X Yu, Xiang Bai, Sanjiv Kumar, and Thomas S Huang. 2022a. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. In *International Conference on Learning Representations (ICLR)*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022c. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. 2022. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*.

A Appendix

1. **Overview of the MedMNIST+ datasets**: In Table A3, we provide a detailed overview of the datasets included in MedMNIST+, highlighting key attributes such as imaging modality, number of classes, total samples, and the standardized training, validation, and testing splits. MedMNIST+ comprises 11 publicly available 2D medical image classification datasets, namely PathMNIST, ChestMNIST, DermaMNIST, OCTMNIST, PneumoniaMNIST, RetinaMNIST, BreastMNIST, OrganAMNIST, OrganCMNIST, OrganSMNIST, and BloodMNIST. These datasets cover a diverse set of imaging modalities, including pathology slides, chest and hand X-rays, abdominal CT scans, dermoscopy, retinal fundus, and OCT imaging and support both binary and multi-class classification tasks. Their heterogeneity in anatomical focus, class granularity, and domain distribution makes MedMNIST+ a comprehensive benchmark for evaluating the robustness and generalization ability of vision-language models across different tasks and domains in medical imaging.

2. **Overview of the 11 biomedical datasets**: We also conduct experiments on 11 biomedical datasets spanning diverse imaging modalities, following the same train-validation-test splits as in (Koleilat et al., 2025). These datasets include CTKidney (Islam et al., 2022) for computed tomography; DermaMNIST (Codella et al., 2019; Tschandl et al., 2018) for dermatoscopy; Kvasir (Pogorelov et al., 2017) for endoscopy; RETINA (Köhler et al., 2013; Porwal et al., 2018) for fundus photography; LC25000 (Borkowski et al., 2019) and CHMNIST (Kather et al., 2016) for histopathology; BTMRI (Nickparvar et al., 2021) for magnetic resonance imaging; OCTMNIST (Kermany et al., 2018) for optical coherence tomography; BUSI (Al-Dhabyani et al., 2019) for ultrasound; and COVID-QU-Ex (Tahir et al., 2021) and KneeXray (Chen, 2018) for X-ray imaging.

3. **Few-shot learning and base-to-new generalization on the biomedical benchmark datasets:** Tables A4 and A5 present the comparison of BioVLM with state-of-the-art baseline methods across the biomedical datasets used by BioMedCoOp (Koleilat et al., 2025) under few-shot learning and base-to-new generalization settings respectively. We consider CLIP-based adapter baselines as CLIP-Adapter (Gao et al., 2021), and Tip-Adapter (Zhang et al., 2022), linear probing baselines as Standard LP (Radford et al., 2021), LP++ (Huang et al., 2024), and prompt learning baselines same as Table 1. The results clearly demonstrate that BioVLM consistently outperforms competing approaches, highlighting its robustness and superior generalization capability in challenging biomedical scenarios.

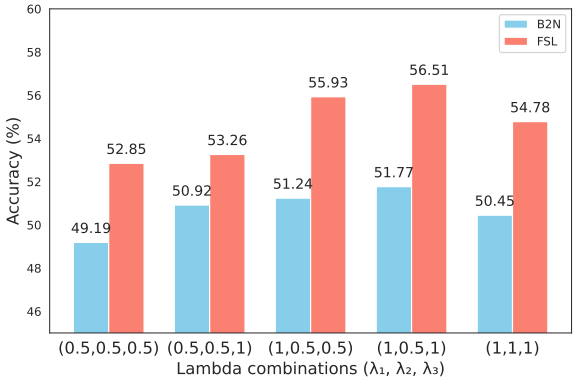

Figure A1: Ablation for the hyperparameter λ_1, λ_2 and λ_3 for the Base-to-New Generalization and Few-shot learning tasks.

4. **Model Calibration Performance:** Beyond classification accuracy, we analyze model calibration on the few-shot learning task across 11 biomedical datasets. using Expected Calibration Error (ECE) (Guo et al., 2017). As shown in Table A6, our BioVLM demonstrates consistently stronger calibration than the compared methods across diverse biomedical modalities. In particular, BioVLM substantially improves over the zero-shot BiomedCLIP baseline, indicating that the proposed adaptation strategy leads to more reliable confidence estimates. Compared with existing prompt-learning approaches, BioVLM also shows more stable calibration behavior across datasets, whereas several baselines exhibit modality-dependent fluctuations. Although MaPLe remains competitive on a few datasets, BioVLM achieves the best overall calibration trend, suggesting that biomedical-specific visual-language modeling better captures uncertainty under few-shot supervision.

5. **λ-Hyperparameter Sensitivity:** We ablate our proposed BioVLM for the best combination of hyperparameters (λ_1, λ_2 and λ_3) defined in Eq.7 and showcase the results in Figure A1 for the Base-to-New Generalization and Few-shot learning tasks.

6. **Effect of different LLMs:** In Table A2, we present the impact of different LLMs including Llama-3.2-3B (Grattafiori et al., 2024), Qwen2.5-14B (Yang et al., 2024), Phi-4 (Abdin et al., 2024) and GPT-4o (Hurst et al., 2024), on the Base-to-New Generalization and Few-shot learning tasks. The results show that variations in the choice of LLM have minimal effect on our proposed method, with BioVLM consistently outperforming all baselines across both tasks.

7. **Ablation with additional prompt-selection methods:** Table A1 compares different strategies for selecting or aggregating prompts from the prompt bank. In our method, each class is represented by multiple LLM-guided learnable prompts, and each prompt produces image-text similarity scores through the frozen encoders. *Softmax* directly uses probability scores but may emphasize noisy high-confidence prompts, while *Mean* averages all prompt predictions and can dilute discriminative cues. *Average Logits* combines raw similarity logits before normalization, but does not explicitly filter uncertain prompts. *Argmax* selects the single most confident prompt based on prediction probabilities, whereas *Top-2* and *Top-5* average predictions from the most confident prompts using the maximum class probability as the confidence score. However, these fixed selection strategies may not adapt well across images or modalities. In contrast, *entropy* selection chooses low-uncertainty prompts, which are more confident and discriminative for each input. This better aligns with BioVLM's prompt-routing design and achieves the strongest performance in both base-to-new generalization and few-shot learning.

8. **Comparison with PEFT methods.** We compare BioVLM with representative PEFT methods on the few-shot learning task of the MedMNIST+ benchmark. As shown in Table A7, **BioVLM** consistently outperforms all PEFT baselines, demonstrating stronger adaptation under limited supervision. Compared with LoRA (Hu et al., 2022), AdaLoRA (Zhang et al., 2023), LayerNorm (Kim et al., 2021), and BitFit (Zaken et al., 2021), BioVLM better leverages biomedical visual-language alignment rather than only adapting model parameters. It also surpasses adapter-based methods such as CLIP-Adapter (Gao et al., 2021) and Tip-Adapter (Zhang et al., 2022), which show less consistent performance across modalities, highlighting BioVLM as a more effective PEFT strategy for few-shot biomedical image recognition.

9. **Qualitative results:** Figure A2 presents the t-SNE visualization of the logits from BioMedCoOp and BioVLM on the PathMNIST dataset in the few-shot setting. The plot clearly shows that BioVLM achieves better class separation compared to BioMedCoOp.

10. **Details results on Out-of-Domain generalization task:** In Tables A8 - A18, we showcase the performance of Out-of-Domain (OOD) generalization task on 11 datasets, where our proposed BioVLM outperforms the state-of-the-art prompt learning methods by significant margin.

Table A1: **Ablation of prompt selection methods in BioVLM on Base-to-New Generalization and Few-shot learning settings.**

Method	B2N	FSL
Softmax	46.32	50.65
Mean	49.62	53.06
Avg. Logits	48.78	52.57
Argmax	50.24	54.19
Top-2	49.70	53.56
Top-5	49.52	53.40
Entropy (ours)	51.77	56.51

Table A2: **Effect of different LLMs on base-to-new generalization and few-shot learning tasks.** We also showcase a snippet of attributes of the class 'lymphocyte', belongs to the BloodMNIST dataset.

Methods	Attribute Snippet	B2N	FSL
Llama-3.2-3B	"small round cell with the nucleus making up most of the cell volume"	49.92	54.10
Qwen2.5-14B	"small, rounded cell characterized by a high nuclear-to-cytoplasmic proportion"	50.56	55.32
Phi-4	"small round cell"	49.34	54.88
GPT-4o	"small, round cell with a high nuclear-to-cytoplasmic ratio"	**51.77**	**56.51**

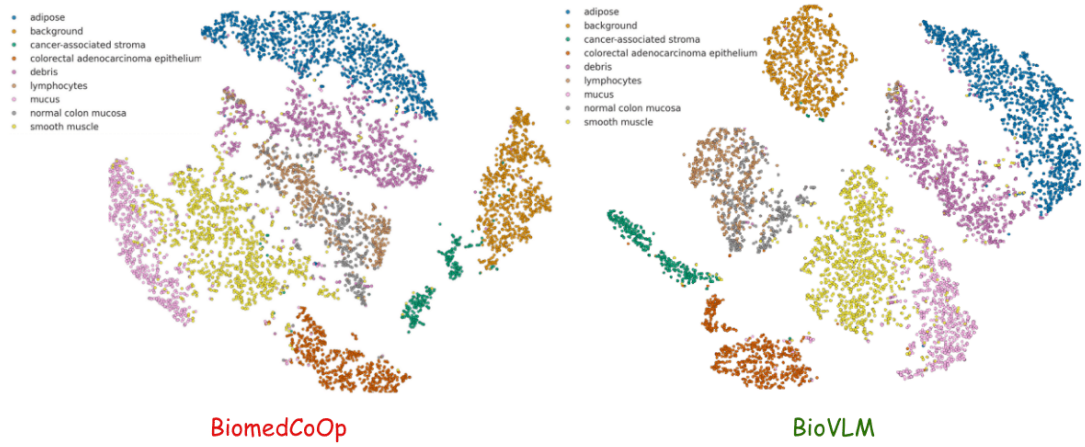

Figure A2: t-SNE visualization of BioMedCoOp and BioVLM in few-shot evaluation on the PathMNIST dataset.

Table A3: **Overview of the MedMNIST+ datasets, their modalities, classification tasks, number of samples, and dataset splits.**

Dataset	Data Modality	Classes	# Samples	# Training / Validation / Test
PathMNIST	Colon Pathology	adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, colorectal adenocarcinoma epithelium	107,180	89,996 / 10,004 / 7,180
DermaMNIST	Dermatoscope	actinic keratoses and intraepithelial carcinoma, basal cell carcinoma, benign keratosis, dermatofibroma, melanoma, melanocytic nevi, vascular lesions	10,015	7,007 / 1,003 / 2,005
OCTMNIST	Retinal OCT	choroidal neovascularization, diabetic macular edema, drusen, normal	109,309	97,477 / 10,832 / 1,000
PneumoniaMNIST	Chest X-Ray	normal, pneumonia	5,856	4,708 / 524 / 624
RetinaMNIST	Fundus Camera	normal retina, mild diabetic retinopathy, moderate diabetic retinopathy, severe diabetic retinopathy, proliferative diabetic retinopathy	1,600	1,080 / 120 / 400
BreastMNIST	Breast Ultrasound	malignant, benign	780	546 / 78 / 156
BloodMNIST	Blood Cell Microscope	basophil, eosinophil, erythroblast, immature granulocytes (myelocytes, metamyelocytes and promyelocytes), lymphocyte, monocyte, neutrophil, platelet	17,092	11,959 / 1,712 / 3,421
TissueMNIST	Kidney Cortex Microscope	Collecting Duct, Connecting Tubule, Distal Convoluted Tubule, Glomerular endothelial cells, Interstitial endothelial cells, Leukocytes, Podocytes, Proximal Tubule Segments, Thick Ascending Limb	236,386	165,466 / 23,640 / 47,280
OrganAMNIST	Abdominal CT	bladder, femur-left, femur-right, heart, kidney-left, kidney-right, liver, lung-left, lung-right, pancreas, spleen	58,850	34,561 / 6,491 / 17,778
OrganCMNIST	Abdominal CT	bladder, femur-left, femur-right, heart, kidney-left, kidney-right, liver, lung-left, lung-right, pancreas, spleen	23,583	12,975 / 2,392 / 8,216
OrganSMNIST	Abdominal CT	bladder, femur-left, femur-right, heart, kidney-left, kidney-right, liver, lung-left, lung-right, pancreas, spleen	25,211	13,932 / 2,452 / 8,827

Table A4: **Comparison of methods on the few-shot learning task of the average of 11 biomedical datasets.**

Method	$K=1$	$K=2$	$K=4$	$K=8$	$K=16$
Zero-shot Methods					
BioMedCLIP			42.05		
BioMedCLIP + Ensemble			52.27		
BioMedCLIP + Selective Ensemble			53.72		
CLIP-based Adapter Methods					
CLIP-Adapter	44.66	43.91	44.36	45.42	46.69
Tip-Adapter	49.19	52.36	57.33	61.98	67.15
Tip-Adapter-F	51.17	52.74	61.23	65.91	70.91
Linear Probing Methods					
Standard LP	47.25	54.21	61.00	65.85	69.40
LP++	47.24	53.18	59.02	63.69	68.35
Prompt Learning Methods					
CoOp	50.16	54.18	59.75	65.84	69.62
CoCoOp	48.49	51.28	54.69	61.08	65.09
KgCoOp	50.85	53.18	57.82	62.08	62.84
MaPLe	52.30	56.17	64.10	69.29	70.48
ProGrad	51.88	54.71	60.42	65.61	67.13
PromptSRC	57.54	59.73	63.27	68.36	73.25
TCP	55.68	60.06	64.35	70.57	71.80
BioMedCoOp	57.03	59.13	63.95	68.32	72.42
BioVLM (Ours)	**59.91**	**62.34**	**66.50**	**72.86**	**77.62**
Δ (in %)	+2.37	+2.28	+2.15	+2.29	+4.37

Table A5: **Comparison of methods on the Base-to-New generalization task of 10 biomedical datasets.**

Dataset	**Sets**	BioMedCLIP NEJM AI'25	CoOp IJCV'22	CoCoOp CVPR'22	KgCoOp CVPR'23	MaPLe CVPR'23	ProGrad ICCV'23	PromptSRC ICCV'23	TCP CVPR'24	BioMedCoOp CVPR'25	BioVLM (Ours)	Δ (in %)
Average on 10 Datasets	Base	47.84	73.85	72.26	68.36	73.05	71.67	75.28	73.82	76.26	**77.19**	+0.93
	New	65.42	64.75	67.03	64.08	71.46	66.93	71.33	70.74	73.92	**75.15**	+1.23
	H	53.81	67.23	67.22	64.61	72.25	67.43	73.25	72.24	75.07	**76.15**	+1.08
BTMRI	Base	40.88	82.25	77.88	78.03	80.28	82.13	**82.78**	81.96	82.42	82.62	-0.16
	New	96.18	94.51	94.84	95.05	95.89	94.98	94.56	94.96	**96.84**	97.29	+0.45
	H	57.37	87.95	85.53	85.69	87.39	88.09	88.28	87.98	89.05	**89.36**	+0.31
COVID-QU-Ex	Base	53.96	75.92	77.28	75.42	76.21	75.19	77.49	75.50	75.91	**78.19**	+0.70
	New	89.43	90.07	87.61	89.61	87.10	90.34	85.24	90.37	91.63	89.80	-1.83
	H	67.31	82.39	82.12	81.90	81.29	82.07	81.18	82.27	83.03	**83.59**	+0.56
CTKIDNEY	Base	38.55	82.24	81.96	81.67	85.24	83.86	**87.03**	84.69	86.93	86.13	-0.90
	New	52.99	67.92	56.56	58.45	76.27	63.01	74.82	75.20	**78.94**	76.02	-2.92
	H	44.63	74.40	66.93	68.14	80.51	71.96	80.46	79.66	**82.74**	80.76	-1.98
DermaMNIST	Base	34.95	48.06	42.88	36.41	40.52	35.52	45.76	43.11	54.86	**55.83**	+0.97
	New	49.59	59.41	60.66	47.31	68.13	63.28	72.30	69.38	74.10	**76.08**	+1.98
	H	41.00	53.14	50.24	41.15	50.82	45.50	56.05	53.18	63.04	**64.40**	+1.36
Kvasir	Base	75.00	86.22	85.94	81.56	83.28	82.89	84.59	85.00	**86.50**	86.17	-0.33
	New	60.50	58.06	53.95	59.00	62.78	60.45	60.14	**61.85**	61.83	61.67	-0.18
	H	66.97	69.39	66.29	68.47	71.59	69.91	70.30	71.60	**72.11**	71.89	-0.22
CHMNIST	Base	37.63	89.41	87.77	75.45	90.26	82.98	88.12	91.34	88.87	**92.82**	+1.48
	New	40.69	35.11	42.51	38.70	46.14	44.19	50.62	48.50	42.73	**62.10**	+11.48
	H	39.10	50.42	57.28	51.16	61.06	57.67	64.30	63.36	57.71	**74.41**	+10.11
LC25000	Base	59.73	90.12	88.33	88.13	90.31	90.29	92.78	88.42	**93.77**	92.00	-1.77
	New	87.60	87.55	95.02	86.44	90.16	85.47	94.16	85.21	97.00	**98.80**	+1.80
	H	71.03	88.82	91.55	87.28	90.23	87.81	93.46	86.79	**95.36**	95.28	-0.08
RETINA	Base	45.18	70.98	66.88	60.77	66.16	68.77	69.37	64.75	68.46	**72.99**	+2.01
	New	55.28	56.90	65.56	54.91	62.48	58.43	56.95	60.38	**67.72**	59.06	-8.66
	H	49.72	63.16	66.21	57.69	64.27	63.18	62.55	62.49	**68.09**	65.29	-2.80
KneeXray	Base	35.89	38.28	34.08	37.94	39.58	40.88	42.58	43.19	44.23	**44.93**	+0.70
	New	71.90	47.69	63.14	61.19	74.10	59.12	76.29	72.49	78.35	**80.66**	+2.31
	H	47.88	42.47	44.27	46.84	51.60	48.34	54.66	54.13	56.54	**57.71**	+1.17
OCTMNIST	Base	56.60	75.00	79.60	68.20	78.66	74.20	**82.33**	80.20	80.33	80.20	-2.13
	New	50.00	50.23	50.47	50.13	**51.59**	50.02	48.17	49.02	50.07	50.00	-1.59
	H	53.10	60.17	61.77	57.79	**62.31**	59.76	60.78	60.85	61.69	61.60	-0.71

Table A6: **Calibration performance using the ECE metric, on few-shot learning task across 11 biomedical datasets.**

Source	BioMedCLIP NEJM AI'25	CoOp IJCV'22	CoCoOp CVPR'22	KgCoOp CVPR'23	MaPLe CVPR'23	ProGrad ICCV'23	PromptSRC ICCV'23	TCP CVPR'24	BiomedCoOp CVPR'25	BioVLM (Ours)
BTMRI	0.17	0.11	0.13	0.11	0.08	0.12	0.09	0.08	0.09	**0.06**
BUSI	0.20	0.12	0.15	0.11	0.13	0.15	0.14	0.13	0.12	**0.10**
COVID-QU-Ex	0.08	0.05	0.10	0.04	0.03	0.06	0.07	0.04	0.03	**0.01**
CTKIDNEY	0.22	0.12	0.20	0.11	0.11	0.14	0.13	0.15	0.13	**0.09**
DermaMNIST	0.14	0.11	0.16	0.10	0.09	0.14	0.10	0.09	0.10	**0.07**
Kvasir	0.10	0.09	0.11	0.08	**0.05**	0.09	0.07	0.08	0.08	0.06
CHMNIST	0.08	0.06	0.06	0.05	0.05	0.08	0.09	0.05	0.04	**0.03**
LC25000	0.07	0.05	0.06	0.06	**0.03**	0.09	0.05	0.06	0.05	0.04
RETINA	0.12	0.10	0.11	0.10	0.07	0.10	0.08	0.07	0.07	**0.05**
KneeXray	0.09	0.05	0.06	0.05	0.04	0.08	0.06	0.06	0.06	**0.02**
OCTMNIST	0.16	0.11	0.11	0.10	0.11	0.13	0.10	0.11	0.14	**0.09**

Table A7: **Comparison of PEFT methods on the few-shot learning task of MedMNIST+ benchmark.**

Dataset	LoRA	AdaLoRA	LayerNorm	BitFit	CLIP-Adapter	Tip-Adapter	BioVLM (Ours)
PathMNIST	74.73	75.29	61.39	59.28	71.47	72.59	**81.56**
DermaMNIST	36.18	37.50	34.48	32.52	37.29	38.11	**45.27**
OCTMNIST	59.19	59.91	48.28	44.59	58.12	57.30	**62.57**
PneumoniaMNIST	70.36	72.72	58.42	52.96	68.40	69.24	**79.65**
RetinaMNIST	35.49	36.81	30.58	27.32	34.67	33.78	**40.08**
BreastMNIST	58.50	57.28	47.38	45.10	56.39	57.00	**65.60**
BloodMNIST	45.28	45.43	25.28	22.49	40.42	41.58	**70.15**
TissueMNIST	17.39	18.06	16.47	15.59	15.93	16.78	**27.34**
OrganAMNIST	36.61	37.06	34.08	32.81	36.51	36.90	**54.75**
OrganCMNIST	35.40	37.47	29.48	26.33	35.02	34.85	**49.22**
OrganSMNIST	33.95	35.19	28.70	25.60	32.76	33.04	**45.41**
Average	45.73	46.61	37.69	34.96	44.27	44.65	**56.51**

Table A8: **OOD generalization, with source dataset as PathMNIST.**

Method	Derma	OCT	Pneumonia	Retina	Breast	Blood	Tissue	OrganA	OrganC	OrganS	Average
CoOp	**66.90**	31.80	65.97	19.42	66.67	12.86	3.93	18.77	16.11	16.96	31.94
CoCoOp	66.25	**34.40**	62.50	16.50	50.85	16.92	3.89	16.88	14.05	14.53	29.68
KgCoOp	66.81	30.90	66.77	18.75	**67.09**	16.28	4.33	19.86	16.62	16.88	32.43
MaPLe	66.51	29.85	66.76	20.50	64.93	**19.04**	4.27	19.34	16.90	17.04	32.51
ProGrad	66.76	32.13	68.75	**23.17**	48.72	18.28	5.53	21.45	17.55	**18.47**	32.08
PromptSRC	66.73	31.45	65.23	18.47	62.49	17.38	**5.83**	18.76	17.25	17.46	32.11
TCP	66.34	30.78	69.45	17.24	61.82	17.65	4.05	19.78	17.35	17.32	32.18
BioMedCoOp	66.85	33.27	**71.69**	17.33	61.75	18.99	4.41	19.36	16.33	17.26	**32.72**
BioVLM (Ours)	66.52	26.33	62.23	17.75	61.32	16.83	4.55	**23.20**	**18.42**	18.32	31.55

Table A9: **OOD generalization, with source dataset as DermaMNIST.**

Method	Path	OCT	Pneumonia	Retina	Breast	Blood	Tissue	OrganA	OrganC	OrganS	Average
CoOp	16.72	25.33	54.54	23.50	32.27	16.91	**4.39**	17.92	17.38	18.36	22.73
CoCoOp	17.70	**38.50**	65.33	28.42	**47.87**	19.76	3.83	14.33	14.61	16.37	26.67
KgCoOp	15.46	25.27	55.29	33.67	29.28	16.91	4.35	18.24	16.91	17.56	23.29
MaPLe	20.56	26.43	55.21	30.53	30.26	12.78	2.65	15.79	16.72	16.98	22.79
ProGrad	16.94	25.03	56.04	25.67	29.91	11.32	3.80	19.24	17.83	19.36	22.51
PromptSRC	24.87	28.42	59.23	35.61	25.47	16.21	2.57	18.94	17.33	15.53	24.42
TCP	16.42	25.89	60.31	26.78	32.47	13.45	2.99	16.43	17.31	19.67	23.17
BioMedCoOp	29.08	31.93	**66.45**	43.50	29.91	**23.64**	3.78	21.73	**20.59**	**23.16**	29.38
BioVLM (Ours)	**34.25**	27.87	62.45	**43.67**	41.24	17.64	3.72	**23.08**	20.20	19.98	**29.41**

Table A10: **OOD generalization, with source dataset as OCTMNIST.**

Method	Path	Derma	Pneumonia	Retina	Breast	Blood	Tissue	OrganA	OrganC	OrganS	Average
CoOp	41.91	64.24	62.50	**37.25**	30.34	18.44	3.64	19.00	17.77	20.50	31.56
CoCoOp	21.06	55.68	62.82	35.67	**35.90**	16.93	**6.20**	20.50	18.56	18.38	29.17
KgCoOp	41.62	62.34	62.50	35.58	30.98	18.69	3.71	19.84	18.66	21.09	31.50
MaPLe	42.12	62.78	62.53	33.67	31.37	**20.24**	4.78	19.45	20.13	22.46	31.95
ProGrad	38.90	62.69	62.50	35.50	29.70	17.21	4.14	20.12	18.36	21.16	31.03
PromptSRC	40.56	61.37	**63.42**	35.91	32.55	18.43	5.84	23.65	21.18	21.39	32.43
TCP	40.96	61.02	62.15	36.24	34.98	19.54	6.32	21.52	21.05	20.87	32.47
BioMedCoOp	37.98	61.88	62.50	33.58	29.70	19.65	4.54	21.49	20.24	22.94	31.45
BioVLM (Ours)	**44.62**	**66.05**	62.50	30.25	30.34	16.87	4.46	**28.00**	**22.32**	**23.46**	**32.89**

Table A11: **OOD generalization, with source dataset as PneumoniaMNIST.**

Method	Path	Derma	OCT	Retina	Breast	Blood	Tissue	OrganA	OrganC	OrganS	Average
CoOp	17.13	40.00	25.73	24.33	28.85	16.91	**5.58**	19.62	16.86	16.78	21.18
CoCoOp	17.87	54.50	31.37	37.25	**43.78**	16.50	4.15	21.77	15.85	17.49	26.05
KgCoOp	17.88	43.76	26.27	25.83	27.78	16.91	5.55	21.09	17.87	18.26	22.12
MaPLe	22.32	56.93	31.45	39.67	27.89	16.37	4.87	20.64	16.34	17.38	25.39
ProGrad	17.89	52.43	27.60	32.83	27.56	16.91	5.48	19.02	16.67	17.04	23.34
PromptSRC	20.56	52.46	32.07	36.68	31.45	16.06	4.93	21.05	16.89	17.93	25.01
TCP	24.78	57.25	30.35	40.29	34.67	16.14	4.37	19.53	16.23	17.85	26.15
BioMedCoOp	26.38	55.31	**33.87**	42.00	35.90	**17.00**	4.91	17.09	16.69	**19.80**	26.89
BioVLM (Ours)	**30.14**	**63.66**	33.20	**44.25**	30.56	16.87	3.85	**22.09**	**18.48**	18.76	**28.19**

Table A12: **OOD generalization, with source dataset as RetinaMNIST.**

Method	Path	Derma	OCT	Pneumonia	Breast	Blood	Tissue	OrganA	OrganC	OrganS	Average
CoOp	**41.93**	64.35	45.73	64.10	**45.67**	19.71	6.85	20.18	17.94	20.37	**34.68**
CoCoOp	26.54	58.72	32.27	61.91	42.73	14.31	3.69	**23.24**	18.67	18.50	30.06
KgCoOp	40.43	63.70	45.93	63.94	35.90	18.48	6.84	19.76	16.98	19.81	33.18
MaPLe	38.53	62.47	42.66	62.40	33.56	14.68	4.09	19.47	16.90	**22.05**	31.68
ProGrad	38.78	63.89	43.80	**65.33**	39.31	17.71	**6.95**	20.15	19.11	21.48	33.65
PromptSRC	35.69	62.89	45.15	63.85	37.64	16.43	4.78	18.28	17.28	20.38	32.24
TCP	34.65	62.50	46.34	64.16	32.59	16.78	4.15	19.78	20.67	18.30	31.99
BioMedCoOp	34.69	**64.66**	**48.63**	62.50	31.84	18.56	5.64	20.89	**20.05**	21.76	32.92
BioVLM (Ours)	37.57	63.61	32.90	62.34	34.83	**21.88**	4.96	20.54	16.63	18.46	31.37

Table A13: **OOD generalization, with source dataset as BreastMNIST.**

Method	Path	Derma	OCT	Pneumonia	Retina	Blood	Tissue	OrganA	OrganC	OrganS	Average
CoOp	32.08	**65.68**	34.33	74.62	36.83	**19.08**	4.06	22.62	19.40	20.37	32.91
CoCoOp	26.15	51.89	32.97	63.03	28.83	16.56	3.59	20.72	18.62	19.40	28.18
KgCoOp	36.46	63.09	29.73	**75.38**	41.42	17.90	3.46	21.75	20.35	21.61	**33.12**
MaPLe	33.54	50.42	30.52	60.17	40.23	16.49	3.58	24.79	20.47	21.35	30.16
ProGrad	33.86	46.04	**37.63**	72.60	32.17	17.26	**5.76**	24.68	**23.69**	23.59	31.73
PromptSRC	34.58	55.18	34.11	63.35	36.14	18.32	3.16	25.68	19.13	20.55	31.02
TCP	33.16	52.90	31.38	67.49	39.87	17.03	3.96	25.31	17.43	18.94	30.75
BioMedCoOp	31.76	45.47	33.60	66.78	**44.08**	17.07	3.46	26.46	23.65	**26.29**	31.86
BioVLM (Ours)	**41.52**	65.20	32.60	62.39	37.08	16.92	4.19	**26.69**	19.37	19.62	32.56

Table A14: **OOD generalization, with source dataset as BloodMNIST.**

Method	Path	Derma	OCT	Pneumonia	Retina	Breast	Tissue	OrganA	OrganC	OrganS	Average
CoOp	**35.04**	65.10	26.60	67.63	23.98	52.14	5.18	19.63	17.18	18.74	33.12
CoCoOp	26.50	63.03	29.20	66.93	24.00	**57.48**	**5.29**	17.24	15.02	14.40	31.91
KgCoOp	30.78	63.49	26.80	66.99	22.58	48.51	4.94	20.27	17.17	18.51	32.00
MaPLe	30.67	65.38	28.04	63.59	20.14	45.09	4.65	18.43	18.32	18.84	31.32
ProGrad	29.08	66.83	25.43	68.64	29.42	42.74	5.08	20.47	**18.43**	20.27	32.64
PromptSRC	33.14	65.82	28.59	64.72	24.64	47.30	4.87	18.94	17.45	**20.56**	32.60
TCP	31.08	64.34	30.15	61.46	22.60	48.85	4.96	19.07	16.95	18.54	31.80
BioMedCoOp	32.67	**66.86**	29.00	**68.91**	21.67	42.73	**5.29**	18.69	16.41	18.48	32.07
BioVLM (Ours)	29.30	66.66	**34.77**	62.71	**30.75**	50.64	5.04	**23.27**	17.77	19.17	**34.01**

Table A15: **OOD generalization, with source dataset as TissueMNIST.**

Method	Path	Derma	OCT	Pneumonia	Retina	Breast	Blood	OrganA	OrganC	OrganS	Average
CoOp	18.67	53.05	28.40	52.46	29.92	**64.74**	**16.92**	18.00	14.44	14.03	31.06
CoCoOp	32.66	49.34	27.77	61.32	**43.50**	34.19	16.91	13.75	13.58	14.26	30.73
KgCoOp	27.84	65.65	33.57	62.66	33.58	62.39	16.90	17.38	14.00	14.82	34.88
MaPLe	37.58	65.31	29.31	61.39	27.48	39.47	16.90	18.05	16.58	17.48	32.96
ProGrad	38.26	65.83	25.90	62.61	25.33	36.96	16.90	**25.66**	**19.94**	18.63	33.60
PromptSRC	40.25	64.64	30.54	62.58	27.92	42.48	16.89	21.36	16.84	17.85	34.14
TCP	38.54	64.92	30.26	62.80	28.03	30.42	16.90	20.58	17.07	16.30	32.58
BioMedCoOp	40.18	65.98	**44.67**	**64.74**	27.25	37.82	16.90	19.66	15.33	16.59	34.91
BioVLM (Ours)	**48.29**	**67.00**	31.20	62.34	**36.83**	34.83	16.90	25.05	19.63	**20.73**	**36.28**

Table A16: **OOD generalization, with source dataset as OrganAMNIST.**

Method	Path	Derma	OCT	Pneumonia	Retina	Breast	Blood	Tissue	OrganC	OrganS	**Average**
CoOp	32.08	39.97	34.33	**66.40**	37.33	32.48	16.88	5.01	37.21	36.63	33.83
CoCoOp	26.15	54.68	32.97	57.69	35.25	**46.37**	16.85	5.10	36.66	35.23	**34.70**
KgCoOp	**36.46**	44.17	29.73	57.58	30.67	34.19	**16.94**	**5.73**	38.08	36.63	33.02
MaPLe	22.57	48.42	35.28	60.93	35.52	32.64	16.78	5.02	36.53	30.68	32.44
ProGrad	33.86	**59.27**	**37.63**	65.33	35.75	37.18	16.66	4.75	28.48	27.92	34.68
PromptSRC	25.49	48.96	34.50	62.05	36.04	33.49	16.92	4.85	35.89	34.84	33.30
TCP	26.31	52.06	31.06	63.87	33.18	37.98	16.88	4.98	36.58	35.05	33.80
BioMedCoOp	31.76	57.34	33.60	62.29	34.25	33.76	16.90	3.72	31.09	30.43	33.51
BioVLM (Ours)	21.15	51.89	30.97	62.18	**39.42**	28.85	16.86	5.50	**44.51**	**42.04**	34.34

Table A17: **OOD generalization, with source dataset as OrganCMNIST.**

Method	Path	Derma	OCT	Pneumonia	Retina	Breast	Blood	Tissue	OrganA	OrganS	**Average**
CoOp	29.05	48.08	31.53	62.55	35.33	33.33	17.04	6.54	42.36	36.20	34.20
CoCoOp	21.23	59.68	**37.17**	62.34	40.58	**41.67**	**17.32**	4.09	37.63	33.02	35.47
KgCoOp	29.22	45.83	26.47	61.27	31.17	**41.67**	16.90	6.76	41.84	34.41	33.55
MaPLe	18.43	58.73	35.45	62.34	36.58	35.78	16.45	5.83	34.75	36.59	34.09
ProGrad	**32.57**	**63.01**	30.37	**64.05**	34.08	34.40	15.19	**6.85**	31.11	26.67	33.83
PromptSRC	20.59	56.09	32.07	61.48	37.07	38.61	16.70	5.05	36.58	38.10	34.23
TCP	22.05	60.18	33.86	62.95	36.14	39.57	16.36	4.87	40.25	29.28	34.55
BioMedCoOp	24.86	60.28	34.13	62.34	34.25	30.99	16.72	5.21	30.74	26.71	32.62
BioVLM (Ours)	19.72	62.94	30.33	62.13	**43.33**	31.84	16.88	4.41	**50.40**	**43.68**	**36.57**

Table A18: **OOD generalization, with source dataset as OrganSMNIST.**

Method	Path	Derma	OCT	Pneumonia	Retina	Breast	Blood	Tissue	OrganA	OrganC	**Average**
CoOp	33.43	62.69	27.27	62.29	32.50	31.84	**17.01**	**5.42**	39.08	36.25	34.78
CoCoOp	19.63	43.44	29.17	61.75	34.00	**46.36**	16.52	4.56	34.99	33.39	32.38
KgCoOp	31.80	57.12	26.77	61.86	34.25	39.75	16.94	5.35	39.32	37.25	35.04
MaPLe	20.55	62.68	28.95	61.58	33.28	32.57	16.70	4.27	36.44	36.27	33.33
ProGrad	**34.58**	**65.72**	26.33	**68.27**	34.17	33.76	16.64	4.62	30.96	28.10	34.32
PromptSRC	22.97	64.00	27.40	63.20	30.57	30.18	16.54	4.65	35.98	36.90	33.24
TCP	25.48	61.49	23.56	62.45	35.19	33.79	16.49	4.50	33.29	30.51	32.68
BioMedCoOp	28.65	62.14	**35.83**	62.39	33.67	31.62	16.77	4.11	31.94	30.29	33.74
BioVLM (Ours)	22.38	63.19	26.50	62.23	**41.08**	37.18	16.90	4.40	**41.91**	**42.51**	**35.83**