Geometry-Aware Distillation for Prompt Tuning Biomedical Vision-Language Models

Anonymous authors
Paper under double-blind review

Abstract

Current prompt-based and adapter-based tuning of vision-language models (VLMs) is attractive for medical imaging, where clinical data sensitivity favors frozen backbones and annotations are limited. However, these methods typically optimize only the ground-truth class, treating all other classes as equally incorrect, ignoring clinically meaningful class relations and yielding unstable decision boundaries in limited-supervision settings. We propose Omni-Geometry Knowledge Distillation (OGKD), a new framework that injects class-relation structure into the teacher to produce directional targets that preserve the ground truth while respecting inter-class geometry. Using these targets, we develop two distillation losses: Global Geometry-Aware Distillation (GAD) operates on the global image token, and Label-Guided Geometry Distillation (LGD) applies the same geometry to attentive patch tokens to improve fine-grained alignment. Across comprehensive experiments and analyses on 11 widely-used medical datasets for base-to-novel and few-shot evaluations, our OGKD achieves substantially better performance, consistently improving accuracy by an average absolute gain of **1.7%–2.8%** over all prior state-of-the-art VLM adaptation counterparts. It also robustly generalizes to unseen classes and yields more reliable predictions than other approaches. Our code is anonymously available at `https://anonymous.4open.science/r/OGKD-02B2`.

1 Introduction

Annotated clinical imaging data are often scarce, particularly for rare conditions and novel acquisition protocols (Litjens et al., 2017). This motivates methods that generalize from minimal supervision. Few-shot learning addresses this setting by enabling generalization from limited labeled samples (Vinyals et al., 2016; Snell et al., 2017; Finn et al., 2017). When combined with vision–language pretraining, models can condition on textual descriptors, which is a powerful route to improve data efficiency (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2022). In medical imaging, domain-specific vision–language models (VLMs) show that this route can unlock broad transfer with limited labels (Wang et al., 2022; Zhang et al., 2023; Bannur et al., 2023). These developments make few-shot adaptation of biomedical VLMs a promising direction for building data-efficient diagnostic tools.

Adapting pretrained VLMs typically follows two directions. (1) *Prompt learning* optimizes lightweight textual prompts while freezing the backbone; representative methods include CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a). (2) *Adapter-based* tuning adds small trainable modules on top of a frozen backbone, including Tip-Adapter (Zhang et al., 2021), CLIP-Adapter (Gao et al., 2024); linear probing (Radford et al., 2021; Huang et al., 2024) is also widely used as a strong baseline. Recently, knowledge distillation (KD) methods (Hinton et al., 2015) have also been introduced for CLIP-style models (Yang et al., 2024; Li et al., 2024), but they typically require heavy student fine-tuning or two-stage training with additional layers.

Biomedical imaging poses additional challenges due to complex visual features, diverse modalities, and heterogeneous acquisition devices. BiomedCoOp (Koleilat et al., 2025) uses LLM-generated templates as teacher prefix prompts and trains student prompts by transferring knowledge from the teacher, achieving strong performance on biomedical datasets. However, existing prompt-, adapter-, and KD-based methods typically (i) treat non-target classes uniformly, ignoring semantic relations (e.g., *benign* is closer to *malignant*

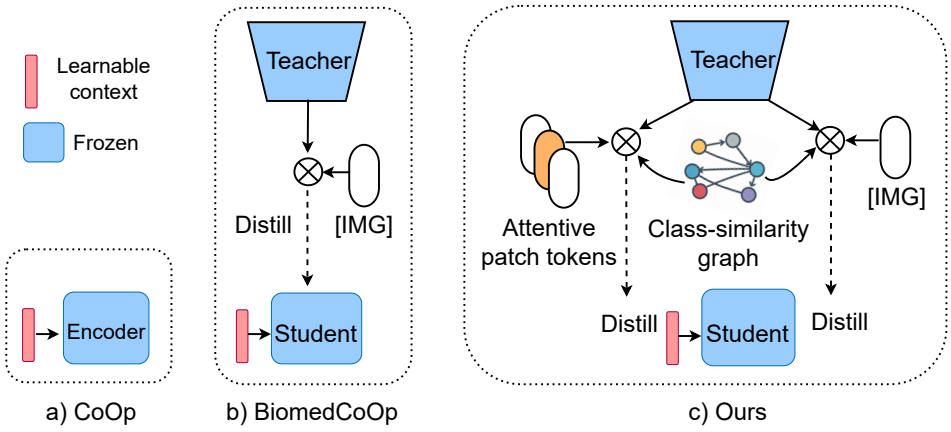

Figure 1: Comparison of prompt-learning pipelines: (a) CoOp (Zhou et al., 2022b), (b) BiomedCoOp (Koleilat et al., 2025), and (c) Ours. Prior methods typically treat classes independently and rely mainly on global image representations; our approach leverages class semantic structure by shaping the teacher distribution with a class graph W and distills geometry-aware supervision over both global `[IMG]` token and attentive patch tokens. Only prompts are learned; the encoders remain frozen.

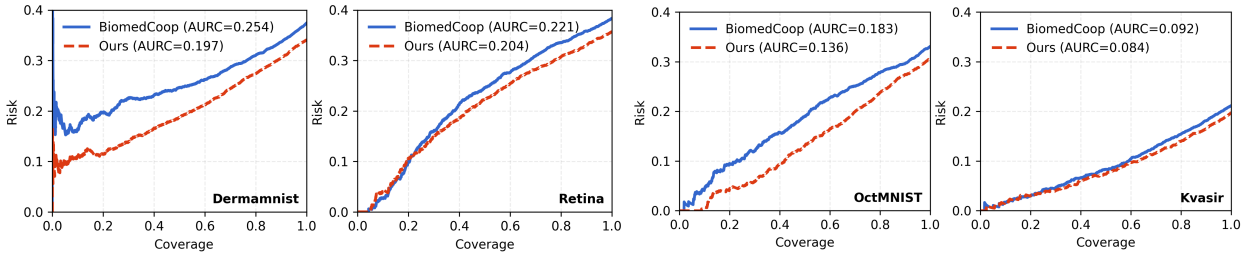

Figure 2: **Risk–coverage across four biomedical datasets.** Our method yields consistently lower curves (smaller AURC) than BiomedCoOp (Koleilat et al., 2025), indicating fewer high-confidence errors at the same coverage and improved selective reliability.

than *malignant* is to *normal*); and (ii) distill only the global `[IMG]` token, missing fine-grained information at the patch-token level.

In safety-critical domains such as medical diagnostics, models must also know when to abstain (Geifman & El-Yaniv, 2017). The risk-coverage curve in Fig. 2 measures selective reliability: as coverage decreases (keeping only the most confident predictions), risk (error among retained samples) should drop quickly, yielding a small area under the risk–coverage curve (AURC). Empirically, we find that strong methods such as BiomedCoOp (Koleilat et al., 2025), which emphasize the ground-truth logit while treating all non-targets as equally incorrect, can produce overconfident mistakes and increase AURC. Our method shifts probability mass toward semantically related classes and reduces spurious confidence in distant ones, lowering AURC and complementing accuracy gains. Recent work (Fan et al., 2024) explores geometry in biomedical settings for novel concept discovery rather than supervised few-shot VLM adaptation. More broadly, recent VLM adaptation methods have begun to model inter-class structure through graph-based modules or similarity-preserving prompt regularization (Li et al., 2023; Jung & Lee, 2025). Our setting is different: rather than learning additional graph/adaptation modules or regularizing prompt embeddings directly, we construct a fixed text-derived class graph and use it to shape teacher targets for few-shot biomedical prompt tuning.

Medical VLMs have also explored global–local alignment between image regions and report tokens (e.g., GLoRIA (Huang et al., 2021), LoVT (Müller et al., 2022)), supporting the need for local supervision beyond global embeddings. In this work, we focus on few-shot prompt tuning with geometry-aware teacher shaping at both the global `[IMG]` token and label-guided patch-token levels (Fig. 1).

Motivation. Expert clinicians rarely weigh a target diagnosis equally against all alternatives. A radiologist might form an initial impression of an anomaly and then focus on a narrow differential diagnosis rather than exhaustively evaluating every option (Krupinski, 2010). Existing models, however, often treat non-targets uniformly despite strong semantic relations among medical classes. Motivated by this gap, we introduce an analogous prior for VLMs: when supervision is limited, errors should fall on semantically related classes rather than unrelated ones.

Our contributions are summarized as follows:

- We encode class-relation structure into the VLM distillation framework by constructing a *text-driven* class graph W from frozen biomedical text prototypes, without using visual samples. This mitigates overfitting in few-shot regimes while injecting clinically meaningful inter-class geometry.
- We introduce an Omni-Geometry Knowledge Distillation method (OGKD) with two losses: (i) *Global Geometry-Aware Distillation (GAD)* shapes the teacher distribution at the global `[IMG]` token and distills the full class distribution; and (ii) *Label-Guided Geometry Distillation (LGD)* applies the same geometry at label-guided patch tokens and distills the label channel to emphasize fine-grained alignment. The approach is parameter-efficient: only student prompts are learned.
- Across 11 biomedical datasets, OGKD delivers consistent gains over state-of-the-art VLM adaptation methods in base→novel generalization, few-shot classification, and selective reliability (risk–coverage), indicating both higher accuracy and clinically safer decisions.

2 Related Work

Vision–language Models (VLMs). Large-scale VLMs such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) learn open-vocabulary recognition by contrasting images and texts at scale. Although highly effective on natural images, multiple works have documented a domain gap when transferring these models to medical images, motivating domain-specific pretraining (Boecking et al., 2022; Eslami et al., 2021; Zhang et al., 2023). BioViL (Boecking et al., 2022) aligns radiology reports with images. PubMedCLIP (Eslami et al., 2021) and BiomedCLIP (Zhang et al., 2023) scale up with curated biomedical pairs. These works improve zero/few-shot transfer to medical images, but they still struggle with fine-grained categories due to limited localized supervision and medical data imbalance (Boecking et al., 2022; Zhang et al., 2023).

Recent work has also developed specific medical vision–language foundation models, such as FLAIR (Silva-Rodriguez et al., 2025) for ophthalmology, as well as PLIP (Huang et al., 2023), QuiltNet (Oluchi Ikezogwo et al., 2023), and CONCH (Lu et al., 2024) for pathology. These efforts highlight the impact of foundation models for healthcare, but downstream use still requires reliable adaptation to new label spaces and modality shifts, especially across settings like our benchmark spanning nine modalities and ten anatomical sites. We therefore focus on a parameter-efficient distillation framework that transfers a frozen biomedical VLM across diverse domains without training a new foundation model per specialty.

Prompt learning and adapters. Prompt learning adapts VLMs by optimizing a small set of textual tokens instead of all model weights. CoOp (Zhou et al., 2022b) learns class-specific prompts and achieves strong accuracy, but subsequent studies (e.g., CoCoOp (Zhou et al., 2022a) and KgCoOp (Yao et al., 2023)) show that naive prompt tuning can *overfit* to base classes and generalize poorly under distribution shift. MaPLe (Khattak et al., 2023a) and PromptSRC (Khattak et al., 2023b) argue that using both visual and textual prompts can mitigate this instability. KgCoOp (Yao et al., 2023) constrains prompts with a knowledge-guided objective to preserve CLIP priors. ProGrad (Zhu et al., 2023) regularizes gradients to avoid negative transfer. Another direction is to use *adapter* modules: CLIP-Adapter (Gao et al., 2024) inserts lightweight residual adapters, while Tip-Adapter (Zhang et al., 2021) caches features to enable training-free or fast tuning. LP++ (Huang et al., 2024) improves linear probing with strong regularization. These methods reduce trainable parameters but operate on *global* image tokens and thus miss fine-grained signals.

Inter-class relations in VLM adaptation. Traditional prompt tuning typically treats categories as independent targets, which can overlook the semantic relationships originally captured by pretrained VLMs. Recent work has been exploring inter-class structure to regularize adaptation. For example, GraphAdapter (Li et al., 2023) is an adapter-style method that models class relations with dual knowledge graphs built from

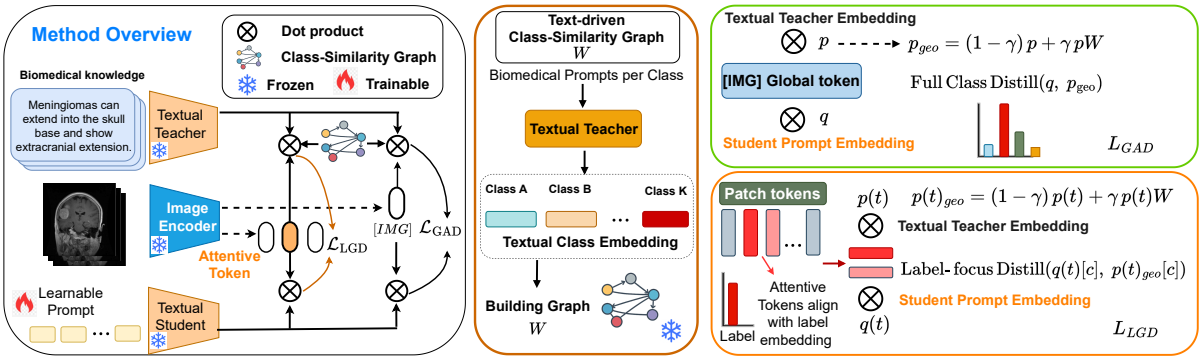

Figure 3: From biomedical text prototypes, we build a class graph $\mathbf{W}$ that captures semantic relations among classes. A geometry strength γ smooths the teacher distribution and supervises two distillation losses: (1) **GAD** operates at the global `[IMG]` token; (2) **LGD** operates at patch-token level, where c denotes the ground-truth class. Only the student prompts are updated; the encoders and $\mathbf{W}$ remain frozen.

textual and visual structure, while SAR (Jung & Lee, 2025) regularizes learnable prompts so that pairwise similarities among prompt-induced text embeddings remain aligned with those from hand-crafted prompts, using LLM-generated related novel classes as auxiliary anchors.

Our OGKD framework shares the high-level goal of preserving class semantics, but differs in mechanism and supervision granularity. Unlike GraphAdapter (Li et al., 2023), which introduces additional trainable adapter and graph-learning components, and unlike SAR (Jung & Lee, 2025), which directly regularizes prompt-induced text embeddings using auxiliary LLM-generated classes, OGKD (ours) constructs a fixed class graph $\mathbf{W}$ once from frozen biomedical text prototypes over the predefined label set. This design is intentional for few-shot biomedical settings, where estimating class relations from limited visual samples can be unstable and may overfit dataset-specific artifacts or annotation noise. In addition, while these methods operate exclusively at the global feature level, OGKD utilizes an *omni-geometry* approach that projects this class structure down to *label-guided attentive patch tokens* (Section 3.3). This local distillation is essential for biomedical imaging, where diagnostic evidence is highly localized.

Prompt learning in biomedical imaging. BiomedCoOp (Koleilat et al., 2025) applies knowledge distillation (Hinton et al., 2015) with a medical encoder backbone (Zhang et al., 2023) and shows competitive performance. In line with most prompt-tuning methods, it (i) treats classes independently and (ii) distills only at the *global* token level. As widely observed in medical recognition, relying only on global signals can miss discriminative local pathology cues. In contrast, we address these limitations by (i) exploiting semantic structure among classes, and (ii) introducing *omni-geometry distillation* that preserves geometry-aware structure at both the global token and label-guided local tokens.

3 Method

Preliminaries. We are given a few-shot training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with classes $\mathcal{C} = \{1, \ldots, C\}$. We build on a frozen VLM (Zhang et al., 2023) with an image encoder g^{img} and a text encoder g^{text}. Let d denote the shared embedding dimension. Following prompt learning, we optimize only a small learnable context $\mathbf{C} \in \mathbb{R}^{n_{\text{ctx}} \times d}$ inserted into the text prompt; all encoder parameters are frozen. Given an image $\mathbf{x}$, the global image token $\mathbf{v}(\mathbf{x}) \in \mathbb{R}^d$ is the projected `[IMG]` token produced by g^{img} and is L2-normalized. For a class $c \in \mathcal{C}$, the textual prototype $\mathbf{t}_c \in \mathbb{R}^d$ is the L2-normalized average of text features across a bank of prompt templates; we stack them into $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_C]^\top \in \mathbb{R}^{C \times d}$. Global logits (pre-softmax) are $\mathbf{z}(\mathbf{x}) = \tau\, \mathbf{T}\, \mathbf{v}(\mathbf{x}) \in \mathbb{R}^C$, where $\tau > 0$ is the fixed logit scale of the VLM. The overall architecture of OGKD framework is illustrated in Fig. 3.

3.1 Class-Similarity Graph

Our prior is that semantically related classes (e.g., *benign* vs. *malignant*) should be closer than unrelated ones. We encode this relation in a fixed class-similarity matrix $\mathbf{W} \in \mathbb{R}^{C\times C}$ computed once from the frozen text encoder and kept constant during training. Each row $\mathbf{W}_{c:}$ contains non-negative weights over $\{1,\dots,C\}$ that sum to one, quantifying the nearest neighbors of class c.

Text prototypes. For each class $c \in \{1,\dots,C\}$, we obtain a textual prototype $\mathbf{t}_c \in \mathbb{R}^d$ by averaging text features computed with a *frozen* text encoder over a bank of prompt templates (e.g., 50 per class), followed by L2 normalization. Stacking them row-wise gives $\mathbf{T} = [\mathbf{t}_1,\dots,\mathbf{t}_C]^\top \in \mathbb{R}^{C\times d}$, whose c-th row equals $\mathbf{t}_c^\top$.

Pairwise similarities. We quantify semantic relations between classes i and j using cosine similarity

$$\begin{aligned} s_{ij} &= \cos\big(\mathbf{t}_i, \mathbf{t}_j\big) = \mathbf{t}_i^\top \mathbf{t}_j \in [-1,1], \\ \mathbf{s}_{c:} &= [\, s_{c1},\dots,s_{cC} \,] \in \mathbb{R}^C. \end{aligned} \tag{1}$$

We then convert each similarity row $\mathbf{s}_{c:}$ into a probability-like distribution with a row-wise softmax:

$$\mathbf{W}_{c:} = \mathrm{softmax}\big(\alpha \mathbf{s}_{c:}\big) \in \mathbb{R}^C, \qquad \mathbf{W}_{cd} \ge 0, \quad \sum_{d=1}^{C} \mathbf{W}_{cd} = 1. \tag{2}$$

Stacking all rows gives $\mathbf{W} = \begin{bmatrix} \mathbf{W}_{1:} \\ \vdots \\ \mathbf{W}_{C:} \end{bmatrix} \in \mathbb{R}^{C\times C}$, where $\alpha > 0$ controls graph sharpness: larger α yields a sharper neighborhood; smaller α spreads weight more broadly. $\mathbf{W}$ depends only on frozen, text-driven prototypes and a fixed teacher encoder-not on any training images to avoid overfitting in few-shot settings.

3.2 Global Geometry-Aware Distillation

We shape the teacher signal so that the probability mass does not flow uniformly to all non-targets but is redistributed to *semantically related* classes according to the graph $\mathbf{W}$ from Sec. 3.1. This yields a softer yet *directional* target that reduces spurious confidence on distant classes.

Student and teacher logits. Let $\hat{\mathbf{T}} \in \mathbb{R}^{C\times d}$ denote the student textual features obtained from frozen text encoder with the learnable context $\mathbf{C}$, and let $\tilde{\mathbf{T}} \in \mathbb{R}^{C\times d}$ denote the *teacher* textual prototypes formed by averaging a subset of LLM prompts retained by a statistics-based selector per class from BiomedCoOp (Koleilat et al., 2025). Given an image $\mathbf{x}$ with global image token $\mathbf{v}(\mathbf{x}) \in \mathbb{R}^d$ and fixed logit scale $\tau > 0$, the student and teacher global logits are:

$$\mathbf{z}_{\mathrm{stu}}(\mathbf{x}) = \tau\, \hat{\mathbf{T}}\, \mathbf{v}(\mathbf{x}) \in \mathbb{R}^C, \;\; \tilde{\mathbf{z}}(\mathbf{x}) = \tau\, \tilde{\mathbf{T}}\, \mathbf{v}(\mathbf{x}) \in \mathbb{R}^C. \tag{3}$$

Then we define the student and teacher log-probabilities with temperature T

$$\begin{aligned} \boldsymbol{\ell}_{\mathrm{stu}}(\mathbf{x}) &= \log \mathrm{softmax}\big(T^{-1}\, \mathbf{z}_{\mathrm{stu}}(\mathbf{x})\big), \\ \boldsymbol{\ell}_{\mathrm{tea}}(\mathbf{x}) &= \log \mathrm{softmax}\big(T^{-1}\, \tilde{\mathbf{z}}(\mathbf{x})\big). \end{aligned} \tag{4}$$

To inject class geometry, we diffuse the teacher with graph $\mathbf{W}$ using a hyperparameter $\gamma \in [0,1]$ that controls the geometry-aware strength:

$$\boldsymbol{\ell}^\star_{\mathrm{tea}}(\mathbf{x}) = (1-\gamma)\, \boldsymbol{\ell}_{\mathrm{tea}}(\mathbf{x}) + \gamma\, \boldsymbol{\ell}_{\mathrm{tea}}(\mathbf{x})\, \mathbf{W}, \tag{5}$$

which partially reallocates teacher knowledge toward graph neighbors of ground-truth class. Given a mini-batch $\mathcal{B}$, the Global Geometry–Aware Distillation loss (GAD) is:

$$\mathcal{L}_{\mathrm{GAD}} = \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x}\in\mathcal{B}} \mathrm{KL}(\boldsymbol{\ell}^\star_{\mathrm{tea}}(\mathbf{x}) \,\|\, \boldsymbol{\ell}_{\mathrm{stu}}(\mathbf{x}))\,, \tag{6}$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the KL divergence with *log-targets*. $\mathcal{L}_{\mathrm{GAD}}$ encourages the student to remain confident on ground truth while considering *semantically nearby* classes, reducing spurious confidence in unrelated alternatives.

3.3 Label-Guided Geometry Distillation

Beyond global [IMG] token, patch tokens contain fine-grained evidence that is valuable in biomedical images. We therefore distill at the level of *patch tokens* while injecting the same class geometry used in Section 3.2. This produces a supervision signal that is both attentive (focused on the most informative tokens) and geometry-aware.

Patch tokens and logits. Let the frozen image encoder produce P patch embeddings for $\mathbf{x}$, projected to the shared space as $\mathbf{Z}(\mathbf{x}) = [\mathbf{z}_1(\mathbf{x}), \ldots, \mathbf{z}_P(\mathbf{x})]^\top \in \mathbb{R}^{P\times d}$. Given the student text embedding $\hat{\mathbf{T}} \in \mathbb{R}^{C\times d}$ and teacher text embedding $\tilde{\mathbf{T}} \in \mathbb{R}^{C\times d}$ (see Section 3.2), the student and teacher *patch–class* logits are

$$\begin{aligned} \mathbf{S}(\mathbf{x}) &= \tau\, \mathbf{Z}(\mathbf{x})\, \hat{\mathbf{T}}^\top \in \mathbb{R}^{P\times C}, \\ \tilde{\mathbf{S}}(\mathbf{x}) &= \tau\, \mathbf{Z}(\mathbf{x})\, \tilde{\mathbf{T}}^\top \in \mathbb{R}^{P\times C}. \end{aligned} \tag{7}$$

We distill the student using informative patch tokens, supervised on regions aligned with ground-truth signal. For each $(\mathbf{x}, y)$, we score patch token n by its *alignment* with teacher text embedding for *ground-truth* class:

$$s_n(\mathbf{x}, y) = \cos\big(\mathbf{z}_n(\mathbf{x}),\, \tilde{\mathbf{t}}_y\big) = \frac{\mathbf{z}_n(\mathbf{x})^\top \tilde{\mathbf{t}}_y}{\|\mathbf{z}_n(\mathbf{x})\|_2\, \|\tilde{\mathbf{t}}_y\|_2}. \tag{8}$$

We then retain the K most attentive patches (highest s_n):

$$\mathcal{I}_K(\mathbf{x}, y) = \mathrm{TopK}_K\big(\{\, s_n(\mathbf{x}, y)\,\}_{n=1}^{P}\big). \tag{9}$$

(For example, $K = \lfloor 0.10\, P \rfloor$ retains the top 10%.) Restricting $\mathbf{S}(\mathbf{x})$ and $\tilde{\mathbf{S}}(\mathbf{x})$ to the selected patch tokens reduces their shapes from $\mathbb{R}^{P\times C}$ to $\mathbb{R}^{K\times C}$. For the selected patches, we form log-probabilities over classes,

$$\begin{aligned} \mathbf{L}^{\mathrm{stu}}(\mathbf{x}) &= \log \mathrm{softmax}\big(T^{-1}\, \mathbf{S}(\mathbf{x})\big), \\ \mathbf{L}^{\mathrm{tea}}(\mathbf{x}) &= \log \mathrm{softmax}\big(T^{-1}\, \tilde{\mathbf{S}}(\mathbf{x})\big). \end{aligned} \tag{10}$$

and diffuse the teacher along the graph $\mathbf{W}$:

$$\mathbf{L}^{\mathrm{tea}\star}(\mathbf{x}) = (1-\gamma)\, \mathbf{L}^{\mathrm{tea}}(\mathbf{x}) + \gamma\, \mathbf{L}^{\mathrm{tea}}(\mathbf{x})\, \mathbf{W}. \tag{11}$$

Let $\mathbf{e}_y \in \{0,1\}^C$ be the one-hot vector for class y. Gathering the *ground-truth* coordinate and restricting to $\mathcal{I}_K(\mathbf{x}, y)$ gives

$$\begin{aligned} \mathbf{u}^{\mathrm{stu}}_{\mathcal{I}}(\mathbf{x}, y) &= \big(\mathbf{L}^{\mathrm{stu}}(\mathbf{x})\, \mathbf{e}_y\big)\Big|_{\mathcal{I}_K(\mathbf{x},y)} \in \mathbb{R}^K, \\ \mathbf{u}^{\mathrm{tea}\star}_{\mathcal{I}}(\mathbf{x}, y) &= \big(\mathbf{L}^{\mathrm{tea}\star}(\mathbf{x})\, \mathbf{e}_y\big)\Big|_{\mathcal{I}_K(\mathbf{x},y)} \in \mathbb{R}^K. \end{aligned} \tag{12}$$

The Label-Guided Geometry Distillation (LGD) loss is

$$\mathcal{L}_{\mathrm{LGD}} = \frac{1}{|\mathcal{B}|} \sum_{(\mathbf{x},y)\in\mathcal{B}} \mathrm{KL}\big(\mathbf{u}^{\mathrm{tea}\star}_{\mathcal{I}}(\mathbf{x}, y) \,\big\|\, \mathbf{u}^{\mathrm{stu}}_{\mathcal{I}}(\mathbf{x}, y)\big)\,. \tag{13}$$

Discussion. LGD introduces no trainable parameters and reuses frozen patch tokens and text prototypes. While GAD (Section 3.2) distills the *full* class distribution at the global [IMG] token, LGD is deliberately *label-focused*: after making the teacher geometry-aware, we gather only the ground-truth channel at the selected patches (see Equation (12)) to emphasize fine-grained, label-specific evidence and reduce compute by avoiding all patch tokens.

3.4 Overall Training Objective

Classification loss. Given student logits $\mathbf{z}_{\mathrm{stu}}(\mathbf{x}) = \tau\, \hat{\mathbf{T}}\, \mathbf{v}(\mathbf{x}) \in \mathbb{R}^C$ from Equation (3), the predicted probabilities are $\mathbf{p}_{\mathrm{stu}}(\mathbf{x}) = \mathrm{softmax}\big(\mathbf{z}_{\mathrm{stu}}(\mathbf{x})\big)$. For a mini-batch $\mathcal{B}$ with labels $y \in \{1, \ldots, C\}$, we use conventional cross-entropy loss for the classification task:

$$\mathcal{L}_{\mathrm{CE}} = \frac{1}{|\mathcal{B}|} \sum_{(\mathbf{x},y)\in\mathcal{B}} \Big[-\log\big(\mathbf{p}_{\mathrm{stu}}(\mathbf{x})\big)_y\Big]. \tag{14}$$

Semantic Consistency by Contextual Mapping (SCCM). This SCCM loss from BiomedCoOp (Koleilat et al., 2025) is defined as: Let $\hat{\mathbf{T}} = [\hat{\mathbf{t}}_1, \ldots, \hat{\mathbf{t}}_C]^\top \in \mathbb{R}^{C\times d}$ denote the *student* text features produced by the frozen text encoder with the learnable context $\mathbf{C}$, and let $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_C]^\top \in \mathbb{R}^{C\times d}$ be the *frozen* per-class textual prototypes obtained by averaging zero-shot prompt templates (Section 3.1). SCCM encourages the learned context to remain semantically anchored to the frozen prototypes via a per-class squared error:

$$\mathcal{L}_{\text{SCCM}} = \frac{1}{C}\sum_{c=1}^{C}\left\|\hat{\mathbf{t}}_c - \mathbf{t}_c\right\|_2^2 = \frac{1}{C}\left\|\hat{\mathbf{T}} - \mathbf{T}\right\|_F^2, \tag{15}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

The total training loss

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1\,\mathcal{L}_{\text{SCCM}} + \lambda_2\,\mathcal{L}_{\text{GAD}} + \lambda_3\,\mathcal{L}_{\text{LGD}}, \tag{16}$$

where $\lambda_1, \lambda_2, \lambda_3 > 0$ are weights of the respective loss.

4 Experiments and Analysis

4.1 Experimental setup and results

Protocols. We evaluate our approach under two few-shot recognition protocols. (i) *Few-shot learning:* for each dataset, we sample $K \in \{1, 2, 4, 8, 16\}$ labeled images per class to form a support set, and train only the prompt learner while keeping the vision–language backbone frozen. Performance is reported on the test split. (ii) *Base–to–novel generalization:* each dataset is divided into base and novel classes. The model is trained on the base classes with $K = 16$ shots and then evaluated on both base and novel test sets, reporting base accuracy, novel accuracy, and harmonic mean (HM) to quantify transfer to unseen categories. We also evaluate the robustness of our method and BiomedCoOp (Koleilat et al., 2025) under label-noise few-shot conditions (details are provided in Section A).

Datasets. We conduct experiments on 11 biomedical image datasets spanning nine imaging modalities and ten anatomical sites: CTKIDNEY (Islam et al., 2022) (CT), DermaMNIST (Codella et al., 2019; Tschandl et al., 2018) (dermatoscopy), Kvasir (Pogorelov et al., 2017) (endoscopy), RETINA (Köhler et al., 2013; Porwal et al., 2018) (fundus photography), LC25000 (Borkowski et al., 2019) and CHMNIST (Kather et al., 2016) (histopathology), BTMRI (Nickparvar, 2021) (brain MRI), OCTMNIST (Kermany et al., 2018) (OCT), BUSI (Al-Dhabyani et al., 2020) (ultrasound), COVID-QU-Ex (Tahir et al., 2021) and KneeXray (Chen, 2018) (X-ray). We follow the official train/validation/test splits for each dataset. For the base–to–novel benchmark, BUSI (Al-Dhabyani et al., 2020) is excluded due to limited class diversity; it is still included in few-shot evaluation.

Implementation details. We use BiomedCLIP (Zhang et al., 2023) (ViT-B/16) as the vision–language backbone. Only learnable context tokens are optimized; all encoder parameters remain frozen. For base–to–novel generalization, we train for 50 epochs; for few-shot evaluation, we train for 100 epochs. We use SGD with learning rate 2.5×10^{-3}, batch size 4, and random resized crop for data augmentation. We initialize the learnable context with "a photo of a", and set the LLM prompt bank to 50 templates per class when building teacher text prototypes. We keep the backbone and training recipe consistent with BiomedCoOp (Koleilat et al., 2025) to ensure a fair comparison. The geometry strength γ, top-token ratio $k\%$, λ_2, λ_3 for each dataset are detailed in the appendix document. For λ_1, we use the same weight from BiomedCoOp. We set graph sharpness α to 1 in base–to–novel generalization and 4 in few-shot evaluation. All main and ablation results in this work are averaged over 3 independent runs. All experiments were run on a single NVIDIA A100 GPU (40 GB). Additional hyperparameters and pseudocode are provided in Section B.

Base–to–novel generalization results. As shown in Table 1, our method outperforms BiomedCoOp (Koleilat et al., 2025) by (+2.77/1.68/2.21)% (Base/Novel/HM) on average across 10 datasets. It achieves the best HM on 9/10 datasets, indicating strong transfer to novel classes. In particular, compared to BiomedCoOp, we observe large gains on RETINA (+7.11/5.98/6.53)%, CHMNIST (+0.67/7.09/6.31)%,

Table 1: **Accuracy comparison on base-to-novel generalization with state-of-the-art prompt-learning methods (16-shot).** Average accuracy (%) over 3 runs. Best per column is in **bold**; second best is underlined. HM denotes the harmonic mean between base and novel. Citations are provided in the "AVERAGE over 10 DATASETS" table.

AVERAGE over 10 DATASETS

Method	Base	Novel	HM
BiomedCLIP (Zhang et al., 2023)	47.84	65.42	53.81
CoOp (Zhou et al., 2022b)	73.85	64.75	67.23
CoCoOp (Zhou et al., 2022a)	72.26	67.03	67.22
KgCoOp (Yao et al., 2023)	68.36	64.08	64.61
ProGrad (Zhu et al., 2023)	71.67	66.93	67.43
BiomedCoOp (Koleilat et al., 2025)	76.26	73.92	75.07
Ours	**79.03** $^{\uparrow 2.77}$	**75.60** $^{\uparrow 1.68}$	**77.28** $^{\uparrow 2.21}$

BTMRI

Method	Base	Novel	HM
BiomedCLIP	40.88	96.18	57.37
CoOp	82.25	94.51	87.95
CoCoOp	77.88	94.84	85.53
KgCoOp	78.03	95.05	85.69
ProGrad	82.13	94.98	88.09
BiomedCoOp	82.42	**96.84**	89.05
Ours	**84.88**	96.11	**90.15**

COVID-QU-Ex

Method	Base	Novel	HM
BiomedCLIP	53.96	89.43	67.31
CoOp	75.92	90.07	82.39
CoCoOp	77.28	87.61	82.12
KgCoOp	75.42	89.61	81.90
ProGrad	75.19	90.34	82.07
BiomedCoOp	75.91	91.63	83.03
Ours	**78.86**	**93.07**	**85.38**

CHMNIST

Method	Base	Novel	HM
BiomedCLIP	37.63	40.69	39.10
CoOp	89.41	35.11	50.42
CoCoOp	87.77	42.51	57.28
KgCoOp	75.45	38.70	51.16
ProGrad	82.98	44.19	57.67
BiomedCoOp	88.87	42.73	57.71
Ours	**89.54**	**49.82**	**64.02**

DermaMNIST

Method	Base	Novel	HM
BiomedCLIP	34.95	49.59	41.00
CoOp	48.06	59.41	53.14
CoCoOp	42.88	60.66	50.24
KgCoOp	36.41	47.31	41.15
ProGrad	35.52	63.28	45.50
BiomedCoOp	54.86	**74.10**	63.04
Ours	**65.45**	67.31	**66.37**

Kvasir

Method	Base	Novel	HM
BiomedCLIP	75.00	60.50	66.97
CoOp	86.22	58.06	69.39
CoCoOp	85.94	53.95	66.29
KgCoOp	81.56	59.00	68.47
ProGrad	82.89	60.45	69.91
BiomedCoOp	86.50	61.83	72.11
Ours	**86.83**	**66.61**	**75.39**

CTKIDNEY

Method	Base	Novel	HM
BiomedCLIP	38.55	52.99	44.63
CoOp	82.24	67.92	74.40
CoCoOp	81.96	56.56	66.93
KgCoOp	81.67	58.45	68.14
ProGrad	83.86	63.01	71.96
BiomedCoOp	**86.93**	78.94	**82.74**
Ours	83.88	**81.11**	82.47

LC25000

Method	Base	Novel	HM
BiomedCLIP	59.73	87.60	71.03
CoOp	90.12	87.55	88.82
CoCoOp	88.33	95.02	91.55
KgCoOp	88.13	86.44	87.28
ProGrad	90.29	85.47	87.81
BiomedCoOp	93.77	97.00	95.36
Ours	**94.34**	**97.98**	**96.13**

RETINA

Method	Base	Novel	HM
BiomedCLIP	45.18	55.28	49.72
CoOp	70.98	56.90	63.16
CoCoOp	66.88	65.56	66.21
KgCoOp	60.77	54.91	57.69
ProGrad	68.77	58.43	63.18
BiomedCoOp	68.46	67.72	68.09
Ours	**75.57**	**73.70**	**74.62**

KneeXray

Method	Base	Novel	HM
BiomedCLIP	35.89	71.90	47.88
CoOp	38.28	47.69	42.47
CoCoOp	34.08	63.14	44.27
KgCoOp	37.94	61.19	46.84
ProGrad	40.88	59.12	48.34
BiomedCoOp	44.23	**78.35**	56.54
Ours	**45.56**	77.49	**57.38**

OCTMNIST

Method	Base	Novel	HM
BiomedCLIP	56.60	50.00	53.10
CoOp	75.00	50.23	60.17
CoCoOp	79.60	50.47	61.77
KgCoOp	68.20	50.13	57.79
ProGrad	74.20	50.02	59.76
BiomedCoOp	80.33	50.07	61.69
Ours	**85.40**	**52.80**	**65.25**

OCTMNIST (+5.07/2.73/3.56)%, and COVID-QU-Ex (+2.95/1.44/2.35)%. These results highlight the benefit of geometry-aware distillation across diverse imaging modalities.

Few-shot evaluation results. As shown in Table 2, our method achieves strong accuracy and consistently improves over BiomedCoOp (Koleilat et al., 2025) by about 2% on average across $K \in \{1, 2, 4, 8, 16\}$. In particular, at $K = 1$ and $K = 8$, we outperform BiomedCoOp by about 2.4%, demonstrating the effectiveness of the method.

4.2 Method analysis

Why geometry-aware design helps. On samples that our method classifies correctly but BiomedCoOp (Koleilat et al., 2025) misclassifies, the margin distribution (as shown in Fig. 4, left panel) shifts right and exhibits a larger Σ, indicating a stronger separation between the true class and its closest neighbors. In

Table 2: **Few-shot evaluation against state-of-the-art methods.** Average classification accuracy (%) over 11 datasets, reported as mean±std over 3 runs. Best results for each K (shots) are in **bold**.

Method	$K = 1$	$K = 2$	$K = 4$	$K = 8$	$K = 16$
Zero-shot methods					
BiomedCLIP (Zhang et al., 2023)	—	—	42.05	—	—
BiomedCLIP (Zhang et al., 2023) + Ensemble	—	—	52.27	—	—
BiomedCLIP (Zhang et al., 2023) + Selective Ensemble	—	—	53.72	—	—
CLIP-based adapter methods					
CLIP-Adapter (Gao et al., 2024)	44.66 ± 2.97	43.91 ± 2.48	44.36 ± 1.94	45.42 ± 2.38	46.69 ± 1.71
Tip-Adapter (Zhang et al., 2021)	49.19 ± 4.84	52.36 ± 6.57	57.33 ± 5.07	61.98 ± 5.76	67.15 ± 4.25
Tip-Adapter-F (Zhang et al., 2021)	51.17 ± 8.33	52.74 ± 5.88	61.23 ± 6.22	65.91 ± 3.64	70.91 ± 2.65
Linear probing methods					
LP++ (Huang et al., 2024)	47.24 ± 7.68	53.18 ± 7.29	59.02 ± 6.93	63.69 ± 4.68	68.35 ± 3.59
Prompt learning methods					
CoOp (Zhou et al., 2022b)	50.16 ± 6.93	54.18 ± 4.31	59.75 ± 3.72	65.84 ± 3.66	69.62 ± 2.83
CoCoOp (Zhou et al., 2022a)	48.49 ± 4.39	51.28 ± 5.06	54.69 ± 4.79	61.08 ± 3.49	65.09 ± 2.87
KgCoOp (Yao et al., 2023)	50.85 ± 5.59	53.18 ± 4.33	57.82 ± 4.50	62.08 ± 2.59	62.84 ± 1.72
ProGrad (Zhu et al., 2023)	51.88 ± 6.39	54.71 ± 4.46	60.42 ± 4.78	65.61 ± 3.02	67.13 ± 3.00
BiomedCoOp (Koleilat et al., 2025)	57.03 ± 2.80	59.13 ± 3.64	63.95 ± 2.42	68.32 ± 2.65	72.42 ± 1.69
Ours	$\mathbf{59.40 \pm 2.86}$	$\mathbf{61.10 \pm 2.08}$	$\mathbf{65.32 \pm 2.26}$	$\mathbf{70.84 \pm 1.81}$	$\mathbf{73.80 \pm 1.38}$

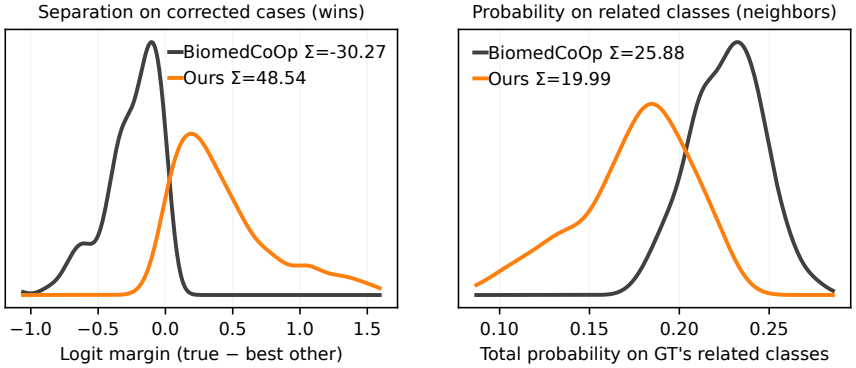

Figure 4: **Wins analysis via KDE curves** on the RETINA dataset. We compare BiomedCoOp and our method on challenging test samples that our method classifies correctly while BiomedCoOp is wrong ("wins"). *Left:* KDE of the logit margin $z(y) - \max_{j \neq y} z(j)$. *Right:* KDE of the probability assigned to the ground-truth class's top-k neighbors in W (here k=3). The legend in each panel also reports Σ, the sum over all win samples.

the right panel, BiomedCoOp (Koleilat et al., 2025) allocates more probability to *neighbors* of the ground-truth class (larger Σ on the neighbor-sum curve), revealing near-miss behavior. Our geometry-aware design leverages the neighborhood structure to discourage spreading mass onto related but non-target classes, which sharpens the target logit and increases the margin. These two diagnostics: fewer near-misses neighbor and larger decision margins explain the observed gains in few-shot accuracy.

Analysis on misclassified samples. We also analyze samples that *both methods misclassify* and show some interesting insights. In the left panel of Fig. 5, where the ground truth is a disease class, the density of our method is shifted to the right and shows a heavier right tail, and its Σ is larger than BiomedCoOp (Koleilat et al., 2025). Although both models predict the wrong *specific* disease, our geometry-aware learner allocates more probability mass to *disease* classes overall. If we change our perspective from multi to binary classification, through a screening lens ("disease vs. normal"), these are *clinically safer* errors: patients would still be identified as "likely diseased" and sent for further examination, rather than being incorrectly diagnosed as normal. This inclination is necessary in healthcare system.

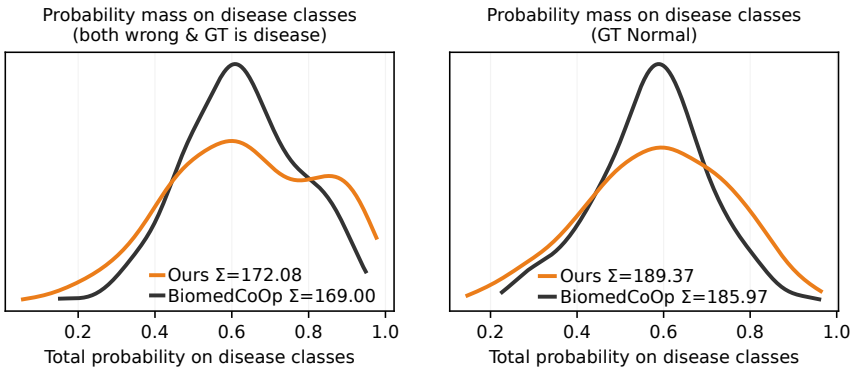

Figure 5: **Probability mass on disease classes (excluded GT)** on Retina dataset. Each curve is a kernel density estimate over per-image quantity $s(x) = \sum_{c \in \mathcal{D}} p_\theta(y = c \mid x)$, where $\mathcal{D}$ are the disease classes (Normal excluded). Legends report $\sum_x s(x)$ (Σ): total probability mass placed on disease classes across all images in the subset. **Left:** subset where both methods misclassify and the ground-truth class is a disease. **Right:** all images with normal GT.

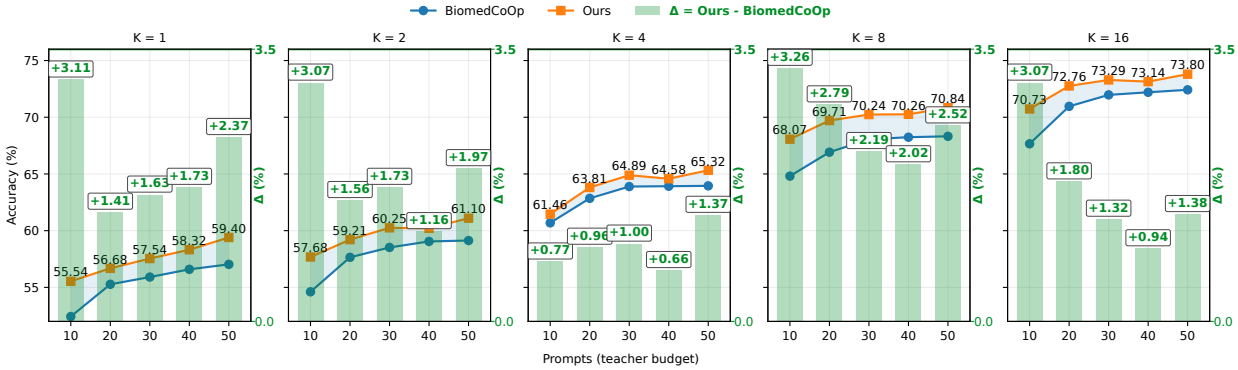

Figure 6: **Few-shot evaluation averaged over 11 datasets under different teacher budgets** (number of LLM prompt templates per class) with shots $K \in \{1, 2, 4, 8, 16\}$. The left y-axis shows absolute accuracies (%) for BiomedCoOp (Koleilat et al., 2025) and our method; the right y-axis and the green strip show the per-budget improvement Δ = Ours − BiomedCoOp (%), annotated at each budget.

Discussion. While beneficial, our method also exhibits a trade-off. In the right panel (Fig. 5), we investigate the case when the ground truth is normal. The curve of our method also shifts to the right and its Σ increases, revealing a higher tendency to place mass on disease classes—i.e., a higher false-positive rate. This is an expected trade-off of our geometry-aware design: avoiding *distant* mistakes on ambiguous disease cases can raise the probability on disease for truly normal images. We provide ideas that can help alleviate this phenomenon in the conclusion section.

5 Ablation study and analysis

Components effect. As shown in Table 3, starting from BiomedCoOp (Koleilat et al., 2025), replacing its vanilla KD loss with our GAD loss significantly improves both base-to-novel generalization and few-shot evaluation. Adding LGD on top yields the best performance. Overall, GAD and LGD provide consistent gains across settings.

Teacher prompt budget. We vary the teacher prompt budget across few-shot settings and report averages over 11 datasets. As shown in Figure 6, our method consistently outperforms BiomedCoOp (Koleilat et al., 2025) for every budget and shot. The gains are largest in the low-budget regime: at N=10, improvements

Table 3: Impact of our components. B denotes BiomedCoOp (Koleilat et al., 2025).

Method	Base-to-novel			Few-shot				
	Base	Novel	HM	K=1	K=2	K=4	K=8	K=16
B	76.26	73.92	75.07	57.03	59.13	63.95	68.32	72.42
B + GAD	78.61	74.60	76.56	57.95	60.39	64.51	69.99	73.41
B + GAD + LGD	**79.03**	**75.60**	**77.28**	**59.40**	**61.10**	**65.32**	**70.84**	**73.80**

Table 4: $\mathcal{L}_{\text{LGD}}$ **variants.** "Full (w/o geom)" disables geometry; "Full geom" distills all patch tokens with geometry; "Label-guided geom" applies geometry only at attentive tokens.

Few-shot	Full (w/o)	Full geom	Label-guided geom
1	58.00	58.87	**59.40**
2	60.17	60.70	**61.10**
4	64.86	64.79	**65.32**
8	69.88	70.29	**70.84**
16	72.48	73.27	**73.80**
Base-to-novel			
Base	78.44	78.49	**79.03**
Novel	73.79	74.38	**75.60**
HM	76.04	76.38	**77.28**

reach 3.11% (K=1), 3.07% (K=2), 3.26% (K=8), and 3.07% (K=16). These results suggest that our method yields robust gains when textual supervision is limited, while remaining beneficial at higher budgets.

Geometry strength. We vary γ, which controls how strongly geometry is injected into the teacher. As shown in Fig. 7, without geometry (γ=0) accuracy drops substantially; moderate geometry consistently helps, whereas too much geometry (γ=1) over-regularizes class boundaries.

Attentive patch tokens vs. full patch tokens. In Table 4, we compare variants of $\mathcal{L}_{\text{LGD}}$ to evaluate: (i) how much the geometry-aware component improves patch-token distillation, and (ii) how label-guided alignment helps student focus on informative tokens. We observe that distilling over all patch tokens with geometry (*Full geom*) outperforms the version without geometry (*Full (w/o)*), demonstrating the benefit of the class-relation graph. The label-guided variant (*Label-guided geom*), which distills only ground-truth channel on top-K attentive patches, further improves accuracy by concentrating supervision on regions that align strongly with ground-truth signals. Overall, focusing distillation on informative patches strengthens fine-grained evidence.

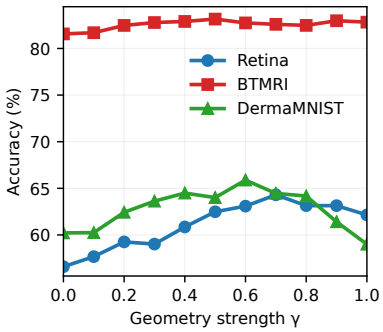

Figure 7: **Effect of geometry strength γ on accuracy.**

6 Conclusion

We introduced OGKD, a framework for prompt tuning biomedical VLMs that distills geometry-aware knowledge at both global and patch-token levels. OGKD improves classification accuracy over state-of-the-art VLM prompt-learning methods across 11 datasets and multiple evaluation settings. Using risk-coverage analysis, we further show that OGKD yields more reliable predictions, which is important for medical applications. One trade-off we observe (right panel of Fig. 5) is a slight increase in false positives on *Normal* class. A straightforward mitigation is to set geometry strength γ to zero (or a small value) for *Normal*, preventing probability mass from diffusing from *Normal* to its disease neighbors. This class-conditional design may reduce false positives on *Normal* while retaining geometry-aware benefits for disease categories; we leave assessing this mitigation to future work.

Limitations and Ethical Statement

Limitations. Our method is evaluated on public biomedical benchmarks, and its performance may not directly transfer to clinical deployment settings with different scanners, patient demographics, or acquisition protocols. Further validation on larger datasets collected from multiple institutions is needed before considering real-world clinical use.

Ethical statement. This work studies few-shot adaptation of biomedical vision-language models for medical image recognition. A potential positive impact is improved data-efficient adaptation for clinical imaging tasks where annotations are limited. However, the method is not intended for autonomous clinical decision-making. Final decisions should be made by qualified clinicians, and model performance may vary across institutions, devices, and patient demographics. All experiments are conducted on publicly available benchmark datasets, and we do not collect new patient data. Future deployment should include careful validation and human clinical supervision.

References

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15016–15027, 2023.

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

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

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

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, 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.

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

Jianan Fan, Dongnan Liu, Hang Chang, Heng Huang, Mei Chen, and Weidong Cai. Seeing unseen: Discover novel biomedical concepts via geometry-constrained probabilistic modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11524–11534, 2024.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3942–3951, 2021.

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

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

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

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

Sehun Jung and Hyang-won Lee. Learning generalizable prompt for clip with class similarity knowledge. *arXiv preprint arXiv:2502.11969*, 2025.

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

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

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

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

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

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

Elizabeth A Krupinski. Current perspectives in medical image perception. *Attention, Perception, & Psychophysics*, 72(5):1205–1217, 2010.

Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36:13448–13466, 2023.

Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26617–26626, 2024.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint learning of localized representations from medical images and reports. In *European conference on computer vision*, pp. 685–701. Springer, 2022.

Msoud Nickparvar. Brain tumor mri dataset. *Kaggle*, 2021.

Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *arXiv e-prints*, pp. arXiv–2306, 2023.

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

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

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

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 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, pp. 3876, 2022.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15952–15962, 2024.

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

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18123–18133, 2022.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

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

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

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

Appendix

A Method robustness under label noise

This ablation compares how our method and BiomedCoOp (Koleilat et al., 2025) perform under label-noise conditions in both few-shot evaluation and base-to-novel generalization settings.

Setup. In base→novel generalization, we corrupt the *16-shot* support set by randomly flipping the labels of k samples per class to other *base* classes. In few-shot evaluation, since training uses all classes, a flipped label can be reassigned to any other class. We vary $k \in \{2, 8, 12, 16\}$, where "Num flips = k" means that k of the 16 shots per class are relabeled at random. This protocol reflects realistic medical curation issues (clinician disagreement, ambiguous findings, data-entry errors), and tests whether our geometry-aware component can mitigate noise.

Result of base→novel generalization. Table 6 reports Base/Novel/HM (%) averaged across 10 datasets. Across all noise levels, our geometry-aware method is consistently more robust to label flips: for **2** flips, we outperform BiomedCoOp (Koleilat et al., 2025) by $(+1.65/+2.10/+1.89)\%$ (Base/Novel/HM); for **8** flips by $(+2.70/-0.14/+1.27)\%$ (a slight drop on Novel); for **12** flips by $(+1.61/+0.63/+1.16)\%$; and for the extreme **16** flips by $(+2.18/+0.60/+1.53)\%$.

Result of few-shot evaluation. Table 7 reports robustness comparisons under different label flip settings. We observe consistently improved robustness across flip settings. Under heavier noise (12 and especially 16 flips), our method shows markedly stronger tolerance than BiomedCoOp (Koleilat et al., 2025).

Why geometry helps. When some supports are mislabeled, our geometry-aware prior *provides stable, structured targets*, reduces spurious confidence on distant classes, and retains stronger robustness compared to BiomedCoOp (Koleilat et al., 2025) as noise increases. Our geometry-aware component does not rely on any training samples, as our class graph $\mathbf{W}$ is *text-driven*. Even with all 16 shots per class flipped (16/16), the prior still provides useful guidance, translating into consistent HM gains over BiomedCoOp (Koleilat et al., 2025). This behavior is desirable in medical settings, where imperfections in annotation are common, but safety performance is required.

B Additional hyperparameters and pseudocode

Table 5 includes the value of hyperparameters used to train the proposed method. Following BiomedCoOp (Koleilat et al., 2025), we select these values using validation set from official datasets.

Pseudocode of proposed method. We use algorithm 1 to construct the class graph $\mathbf{W}$ and algorithm 2 for the overall training process. The class graph $\mathbf{W}$ is computed once and kept frozen during training.

Table 5: Hyperparameters per dataset and benchmark for our method. λ_{GAD} and λ_{LGD} are loss weights; γ is the geometry strength; K/P is the top-token ratio.

Dataset	Benchmark	λ_{GAD}	λ_{LGD}	γ	K/P
BTMRI	Base-to-Novel	0.50	0.01	0.80	0.50
	Few-shot	1.75	0.05	0.50	0.40
BUSI	Base-to-Novel	—	—	—	—
	Few-shot	0.10	0.75	0.05	0.80
COVID-QU-Ex	Base-to-Novel	0.50	2.25	0.05	0.50
	Few-shot	3.50	0.75	0.10	0.20
CTKIDNEY	Base-to-Novel	2.00	0.25	0.90	0.30
	Few-shot	1.25	0.10	0.70	0.03
DermaMNIST	Base-to-Novel	3.50	0.10	0.10	0.30
	Few-shot	24.00	4.00	0.60	0.10
Kvasir	Base-to-Novel	3.50	22.00	0.30	0.03
	Few-shot	0.50	1.00	0.01	0.01
CHMNIST	Base-to-Novel	3.75	0.01	0.30	0.05
	Few-shot	2.25	0.05	0.70	0.80
LC25000	Base-to-Novel	0.25	2.00	0.40	0.60
	Few-shot	1.00	0.05	0.50	0.05
RETINA	Base-to-Novel	0.50	0.01	0.30	0.90
	Few-shot	0.50	1.25	0.70	0.10
KneeXray	Base-to-Novel	4.75	0.01	0.30	0.60
	Few-shot	20.00	1.75	0.40	0.80
OCTMNIST	Base-to-Novel	0.25	0.03	0.70	0.01
	Few-shot	4.00	3.25	0.70	0.03

Algorithm 1 Build class graph $\mathbf{W}$

Input: Frozen text encoder g^{text}; class set $\mathcal{C} = \{1, \ldots, C\}$.
Note that in few-shot evaluation, $\mathcal{C}$ means using all classes; whereas in base-to-novel generalization, $\mathcal{C}$ means using only base classes.
Prompt bank $\{\mathcal{P}_c\}_{c=1}^{C}$ (e.g., 50 prompts per class); graph sharpness $\alpha > 0$.

Output: class graph $\mathbf{W} \in \mathbb{R}^{C \times C}$ (kept frozen).

for $c \in \{1, \ldots, C\}$ **do**
  Text prototype for class c (average of frozen prompt features)
  $\mathbf{t}_c \leftarrow \frac{1}{|\mathcal{P}_c|} \sum_{p \in \mathcal{P}_c} \text{normalize}\big(g^{\text{text}}(p)\big)$ // L2-normalize
end for
$\mathbf{T} \leftarrow [\,\mathbf{t}_1, \ldots, \mathbf{t}_C\,]^\top \in \mathbb{R}^{C \times d}$
$\mathbf{S} \leftarrow \mathbf{T}\mathbf{T}^\top \in [-1, 1]^{C \times C}$
for $c \in \{1, \ldots, C\}$ **do**
  Row-wise softmax to get non-negative weights that sum to 1
  $\mathbf{W}_{c:} \leftarrow \text{softmax}\big(\alpha\, \mathbf{S}_{c:}\big)$ // $\alpha \uparrow$: sharper neighborhoods; $\alpha \downarrow$: more spread
end for
$\mathbf{W} \leftarrow$ `torch.stack`$(\,[\mathbf{W}_{c:}]_{c=1}^{C}$, `dim=0`$)$
return $\mathbf{W}$

Table 6: **Robustness comparison under label-noise (random label flips) in base→novel generalization.** Base/Novel/HM accuracies (%) across 10 datasets with 16 shots over 3 runs. Arrows in the **AVERAGE** block denote (Ours − BiomedCoOp (Koleilat et al., 2025)): green ↑ = improvement; red ↓ = drop.

Num flips = 2

	Ours			BiomedCoOp		
Dataset	Base	Novel	HM	Base	Novel	HM
BTMRI	81.79	96.40	88.50	79.29	95.35	86.58
CHMNIST	88.34	45.70	60.24	85.88	44.64	58.98
COVID-QU-Ex	75.72	91.25	82.76	74.77	90.57	81.92
CTKIDNEY	80.58	81.13	80.85	85.77	74.24	79.59
DermaMNIST	64.08	45.03	52.89	63.80	55.74	54.75
KneeXray	42.16	77.01	54.49	43.54	63.14	51.54
Kvasir	86.39	61.28	71.70	84.94	58.00	68.93
LC25000	89.80	95.00	92.33	93.92	93.29	91.35
OCTMNIST	81.80	52.67	64.08	77.00	50.00	60.63
RETINA	69.25	66.25	67.72	63.46	65.14	66.76
AVERAGE	75.99	71.17	73.50	74.34	69.07	71.61
	↑1.65	↑2.10	↑1.89			

Num flips = 8

	Ours			BiomedCoOp		
Dataset	Base	Novel	HM	Base	Novel	HM
BTMRI	72.67	96.33	82.84	69.00	84.88	76.12
CHMNIST	86.08	44.42	58.60	85.81	39.67	54.26
COVID-QU-Ex	63.52	91.56	75.01	65.06	91.37	76.00
CTKIDNEY	71.55	81.31	76.12	73.76	77.87	75.76
DermaMNIST	62.46	47.19	53.76	50.89	61.67	55.76
KneeXray	41.25	64.72	50.39	41.44	66.79	51.15
Kvasir	85.33	58.78	69.61	82.78	55.17	66.21
LC25000	86.97	95.33	90.96	85.44	95.29	90.10
OCTMNIST	74.93	49.60	59.69	70.00	50.07	58.59
RETINA	65.03	61.57	63.25	58.03	69.45	63.23
AVERAGE	70.98	69.08	70.02	68.28	69.22	68.75
	↑2.70	↓0.14	↑1.27			

Num flips = 12

	Ours			BiomedCoOp		
Dataset	Base	Novel	HM	Base	Novel	HM
BTMRI	64.00	95.45	76.65	59.20	84.04	69.45
CHMNIST	81.30	44.99	57.91	82.60	37.41	51.50
COVID-QU-Ex	49.50	91.19	64.18	58.10	90.76	70.85
CTKIDNEY	53.90	78.99	64.08	51.20	75.99	61.19
DermaMNIST	62.40	53.71	57.72	52.50	65.31	58.21
KneeXray	42.90	65.45	51.80	42.70	75.43	54.56
Kvasir	84.80	55.56	67.14	82.00	49.84	61.98
LC25000	79.50	89.02	83.99	76.90	94.26	84.71
OCTMNIST	68.70	50.00	57.87	71.00	48.20	57.42
RETINA	59.20	70.08	64.21	53.80	66.93	59.66
AVERAGE	64.60	69.44	66.94	63.00	68.82	65.78
	↑1.61	↑0.63	↑1.16			

Num flips = 16

	Ours			BiomedCoOp		
Dataset	Base	Novel	HM	Base	Novel	HM
BTMRI	51.42	96.73	67.15	45.63	89.82	60.52
CHMNIST	78.10	41.89	54.53	79.12	39.89	53.04
COVID-QU-Ex	43.56	91.34	58.99	49.31	89.77	63.65
CTKIDNEY	55.91	72.79	63.24	50.17	74.04	59.81
DermaMNIST	62.22	62.71	62.46	49.11	82.44	61.55
KneeXray	41.08	71.53	52.19	41.15	49.52	44.95
Kvasir	82.28	55.28	66.13	80.17	54.11	64.61
LC25000	80.24	93.15	86.21	75.24	92.24	82.88
OCTMNIST	62.67	49.20	55.12	64.07	51.67	57.21
RETINA	48.71	67.45	56.57	50.40	72.60	59.50
AVERAGE	60.62	70.21	65.06	58.44	69.61	63.54
	↑2.18	↑0.60	↑1.53			

Table 7: **Robustness comparison under label noise (random label flips) in few-shot evaluation.** Average accuracy (%) across 11 datasets under different numbers of random label flips (16-shot), averaged over 3 runs.

	Num flips = 2		**Num flips = 8**		**Num flips = 12**		**Num flips = 16**	
Dataset	**Ours**	**BiomedCoOp**	**Ours**	**BiomedCoOp**	**Ours**	**BiomedCoOp**	**Ours**	**BiomedCoOp**
BTMRI	82.43	81.11	79.15	78.22	74.41	76.86	73.11	73.07
BUSI	67.09	65.68	63.98	63.14	65.54	62.29	62.01	57.06
CHMNIST	78.64	78.50	77.26	77.52	74.64	73.85	75.22	75.31
COVID-QU-Ex	77.95	77.76	77.56	77.42	77.06	76.86	74.83	75.85
CTKIDNEY	80.78	82.81	76.61	79.31	78.47	75.51	72.89	69.53
DermaMNIST	65.65	62.76	66.22	62.14	64.42	61.61	64.29	61.39
KneeXray	41.49	39.72	40.82	39.41	40.52	39.57	40.60	39.90
Kvasir	80.14	78.67	77.94	77.75	77.44	76.06	76.00	74.81
LC25000	90.71	91.85	87.83	89.16	85.00	83.99	83.97	86.95
OCTMNIST	67.90	66.90	66.60	64.83	66.10	61.87	65.17	59.57
RETINA	60.81	61.25	54.13	55.65	49.97	53.58	47.08	49.34
AVERAGE	**72.14**	71.55	**69.83**	69.50	**68.51**	67.46	**66.83**	65.71
	↑0.60		↑0.32		↑1.05		↑1.13	

Algorithm 2 Training loop of proposed method in PyTorch-like style

Input: $\mathbf{T} \in \mathbb{R}^{C\times d}$, $\mathbf{W}$ (Alg. 1), $\tilde{\mathbf{T}} \in \mathbb{R}^{C\times d}$ (frozen); $\tau, T, \gamma, K, \lambda_1, \lambda_2, \lambda_3$; context $\mathbf{C}$; dataset $\mathcal{D}$; optimizer on $\mathbf{C}$ only.
Output: Updated context $\mathbf{C}$ (encoders and $\mathbf{W}$ remain frozen).
for each epoch **do**
  for mini-batch $\mathcal{B} = \{(\mathbf{x}, y)\}$ **do**
    $\hat{\mathbf{T}} \leftarrow g^{\text{text}}(\text{class prompts with context } \mathbf{C})$ // student text (L2-normalized rows)
    $\mathbf{v}(\mathbf{x}) \leftarrow g^{\text{img}}(\mathbf{x}).\texttt{global_token}()$ // global `[IMG]` token, L2-normalized
    $\mathbf{z}_{\text{stu}}(\mathbf{x}) \leftarrow \tau\, \hat{\mathbf{T}}\, \mathbf{v}(\mathbf{x})$
    $\tilde{\mathbf{z}}(\mathbf{x}) \leftarrow \tau\, \tilde{\mathbf{T}}\, \mathbf{v}(\mathbf{x})$
    $\boldsymbol{\ell}_{\text{stu}}(\mathbf{x}) \leftarrow \text{logsoftmax}\big(\mathbf{z}_{\text{stu}}(\mathbf{x})/T\big)$
    $\boldsymbol{\ell}_{\text{tea}}(\mathbf{x}) \leftarrow \text{logsoftmax}\big(\tilde{\mathbf{z}}(\mathbf{x})/T\big)$
    $\boldsymbol{\ell}^{\star}_{\text{tea}}(\mathbf{x}) \leftarrow (1-\gamma)\, \boldsymbol{\ell}_{\text{tea}}(\mathbf{x}) + \gamma\, \boldsymbol{\ell}_{\text{tea}}(\mathbf{x})\, \mathbf{W}$ // geometry-aware teacher (global)
    Global geometry-aware distillation (GAD)
    $\mathcal{L}_{\text{GAD}} \leftarrow \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x}\in\mathcal{B}} \text{KL}\big(\boldsymbol{\ell}^{\star}_{\text{tea}}(\mathbf{x}) \,\|\, \boldsymbol{\ell}_{\text{stu}}(\mathbf{x})\big)$
    $\mathbf{Z}(\mathbf{x}) \leftarrow g^{\text{img}}(\mathbf{x}).\texttt{patch_tokens}()$ // patch tokens, shape $[P, d]$
    $\mathbf{S}_{\text{stu}}(\mathbf{x}) \leftarrow \tau\, \mathbf{Z}(\mathbf{x})\, \hat{\mathbf{T}}^{\top}$ // patch logits
    $\mathbf{S}_{\text{tea}}(\mathbf{x}) \leftarrow \tau\, \mathbf{Z}(\mathbf{x})\, \tilde{\mathbf{T}}^{\top}$
    $\mathbf{L}_{\text{stu}}(\mathbf{x}) \leftarrow \text{logsoftmax}\big(\mathbf{S}_{\text{stu}}(\mathbf{x})/T\big)$
    $\mathbf{L}_{\text{tea}}(\mathbf{x}) \leftarrow \text{logsoftmax}\big(\mathbf{S}_{\text{tea}}(\mathbf{x})/T\big)$
    $\mathbf{L}^{\star}_{\text{tea}}(\mathbf{x}) \leftarrow (1-\gamma)\, \mathbf{L}_{\text{tea}}(\mathbf{x}) + \gamma\, \mathbf{L}_{\text{tea}}(\mathbf{x})\, \mathbf{W}$ // geometry-aware teacher (patch)
    (Shown as loop for clarity; implement in batch mode.)
    Initialize accumulator $A \leftarrow 0$
    for each $(\mathbf{x}, y) \in \mathcal{B}$ **do**
      $s_n \leftarrow \cos\big(\mathbf{z}_n(\mathbf{x}), \tilde{\mathbf{t}}_y\big)$ for $n = 1, \ldots, P$ // align w/ teacher GT text
      $\mathcal{I}_K(\mathbf{x}, y) \leftarrow \text{TopK}\left(\{s_n\}_{n=1}^{P}, K\right)$
      $u^{\text{stu}} \leftarrow \big(\mathbf{L}_{\text{stu}}(\mathbf{x})\, \mathbf{e}_y\big)\big|_{\mathcal{I}_K(\mathbf{x},y)}$
      $u^{\text{tea}\star} \leftarrow \big(\mathbf{L}^{\star}_{\text{tea}}(\mathbf{x})\, \mathbf{e}_y\big)\big|_{\mathcal{I}_K(\mathbf{x},y)}$
      $A \leftarrow A + \text{KL}\big(u^{\text{tea}\star} \,\|\, u^{\text{stu}}\big)$
    end for
    Label-guided Geometry Distillation (LGD)
    $\mathcal{L}_{\text{LGD}} \leftarrow \frac{1}{|\mathcal{B}|}\, A$
    $\mathbf{p}_{\text{stu}}(\mathbf{x}) \leftarrow \text{softmax}\big(\mathbf{z}_{\text{stu}}(\mathbf{x})\big)$
    $\mathcal{L}_{\text{CE}} \leftarrow \frac{1}{|\mathcal{B}|} \sum_{(\mathbf{x},y)\in\mathcal{B}} \Big[-\log \mathbf{p}_{\text{stu}}(\mathbf{x})_y\Big]$
    $\mathcal{L}_{\text{SCCM}} \leftarrow \frac{1}{C}\, \|\hat{\mathbf{T}} - \mathbf{T}\|_F^2$
    Overall training loss
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{SCCM}} + \lambda_2 \mathcal{L}_{\text{GAD}} + \lambda_3 \mathcal{L}_{\text{LGD}}$
    `optimizer.zero_grad()`
    backprop $\nabla_{\mathbf{C}} \mathcal{L}$; `optimizer.step()` // encoders & $\mathbf{W}$ frozen
  end for
end for