vMFCoOp: Towards Equilibrium on a Unified Hyperspherical Manifold for Prompting Biomedical VLMs

Minye Shao[1], Sihan Guo[1], Xinrun Li[1], Xingyu Miao[1], Haoran Duan[2], Yang Long [1*]

[1]Department of Computer Science, Durham University, Durham, UK
[2]Department of Automation, Tsinghua University, Beijing, China
{minye.shao, yang.long}@ieee.org

Abstract

Recent advances in context optimization (CoOp) guided by large language model (LLM)–distilled medical semantic priors offer a scalable alternative to manual prompt engineering and full fine-tuning for adapting biomedical CLIP-based vision-language models (VLMs). However, prompt learning in this context is challenged by semantic misalignment between LLMs and CLIP variants due to divergent training corpora and model architectures; it further lacks scalability across continuously evolving families of foundation models. More critically, pairwise multimodal alignment via conventional Euclidean-space optimization lacks the capacity to model unified representations or apply localized geometric constraints, which tends to amplify modality gaps in complex biomedical imaging and destabilize few-shot adaptation. To address these challenges, we propose vMFCoOp, a framework that inversely estimates von Mises–Fisher (vMF) distributions on a shared Hyperspherical Manifold, aligning semantic biases between arbitrary LLMs and CLIP backbones via Unified Semantic Anchors to achieve robust biomedical prompting and superior few-shot classification. Grounded in three complementary constraints, vMFCoOp achieves consistent gains across 14 medical datasets, 12 imaging modalities, and 13 anatomical regions, surpassing state-of-the-art methods in accuracy, generalization, and clinical applicability.

Extended version — https://arxiv.org/abs/2511.09540

Introduction

Recent advances in vision-language models (VLMs), such as CLIP (Radford et al. 2021), have significantly advanced general vision tasks by aligning visual and linguistic semantics through large-scale contrastive learning, enabling strong zero-shot and few-shot generalization. Biomedical images exhibit highly structured semantics and demand domain-specific knowledge, making it challenging for VLMs pretrained on natural images to generalize effectively. Previous adaptations of CLIP to the medical domain often require heavy supervision, limiting scalability and generalization under low-data settings (Li et al. 2024). In response, recent medical-specific CLIP variants (Wang et al. 2022; Eslami et al. 2023) have emerged, pretrained on large-scale medical image–text pairs, and exhibit improved performance (Zhang et al. 2022b). However, these VLMs largely hinge on the quality of the textual prompts used, and full fine-tuning of large-scale pretrained models is often impractical due to computational and data constraints (Schmalfuss et al. 2025). To address this, CLIP-based prompt learning has emerged as a lightweight adaptation strategy that optimizes input prompts to guide cross-modal predictions without modifying model parameters. Methods such as CoOp (Zhou et al. 2022b) learn tunable context vectors to construct task-specific semantic prompts, substantially enhancing CLIP's performance across diverse downstream tasks.

Unlike natural images, biomedical images present greater visual complexity and modality diversity (Shao et al. 2025b), with fine-grained structures, strong anatomical priors (Shao et al. 2025a), and significant cross-scale variations as illustrated by the medical imaging cases in fig. 1, imposing higher demands on model discrimination. Meanwhile, the acquisition of high-quality labeled data is severely limited by privacy concerns, high annotation costs, and the rarity or inaccessibility of certain patient populations, making few-shot learning a central paradigm in medical imaging (Huang et al. 2023). Multimodal large language models (MLLMs), trained on rich medical corpora, show promising capabilities in basic recognition tasks but remain unstable in structural understanding and high-level semantics, often exhibiting hallucinations, particularly in complex scenarios like multi-stage diagnosis or surgical planning. In contrast, CLIP-based medical VLMs offer lightweight, interpretable architectures with stronger visual–language alignment and better clinical deployability (Chen et al. 2025). To leverage the strengths of both LLMs and CLIP, recent work such as BiomedCoOp (Koleilat et al. 2025) introduces LLM-generated prompts to guide CLIP's contextual learning. It implicitly assumes a direct compatibility between the language priors of LLMs and the multimodal embedding space of CLIP. However, prompt learning across such heterogeneous modalities entails more than simple token-level substitution; it requires reconciling differing semantic abstractions, representational granularities, and alignment dynamics. As with many conventional approaches to cross-modal alignment (Qiu et al. 2025), it often relies on independent pairwise matching within flat Euclidean spaces, which is in-

*Corresponding author.

sufficient to capture the intrinsic relational geometry and directional semantics among modalities (Zheng et al. 2025).

To overcome the limitations of flat Euclidean alignment in cross-modal prompt learning, recent efforts have explored manifold optimization to better respect the geometric structure of high-dimensional representation spaces (Park, Ko, and Kim 2024). In particular, methods combining CoOp-style prompt tuning with geometry-aware learning have demonstrated improved robustness and adaptability by constraining features or prompts to lie on Riemannian manifolds such as hyperspheres (Cho, Kim, and Kim 2023). However, these approaches often treat vision and language representations independently, neglecting the heterogeneous semantic granularity and alignment dynamics between LLMs and CLIP. Moreover, existing designs typically impose fixed geometric priors without accounting for distributional mismatch or modality-specific biases, leading to unstable convergence or suboptimal generalization under few-shot regimes (Wang et al. 2023). To address all the above challenges, we introduce vMFCoOp, a novel biomedical prompting framework that leverages a unified hyperspherical manifold and inverse vMF mapping estimation to achieve stable, generalizable, and model-agnostic few-shot learning. In summary, our contributions are threefold:

- This is the first framework to introduce an inverse-estimation based on hyperspherical manifold probabilistic modeling, enabling explicit capture and reconciliation of semantic biases inherent in diverse foundation models, and ensuring scalable adaptability across rapidly evolving model families.
- We revisit few-shot prompt learning for biomedical VLMs on a Unified Hyperspherical Manifold, where semantic anchors bridge cross-modal gaps and improve generalization beyond Euclidean-space approaches.
- Extensive validation on 14 realistic clinical few-shot scenarios, including challenging datasets from UK Biobank, confirms that vMFCoOp substantially improves biomedical prompt-based classification performance, supporting its potential for clinical deployment.

Related Work

Few-shot Adaptation of VLMs

Prompt-based methods have become effective for adapting vision-language models (VLMs) like CLIP (Radford et al. 2021) to data-limited downstream tasks. CoOp (Zhou et al. 2022b) learns continuous prompts but overfits to base classes, while CoCoOp (Zhou et al. 2022a) introduces instance-conditioned prompts for better generalization. ProGrad (Zhu et al. 2023) and KgCoOp (Yao, Zhang, and Xu 2023) align gradients or knowledge to preserve pretrained semantics, though with reduced flexibility. PLOT (Chen et al. 2022a) improves fine-grained alignment via multi-prompt optimal transport, and MaPLe (Khattak et al. 2023a) together with PromptSRC (Khattak et al. 2023b) enhance prompt diversity and generalization. Adapter-based methods such as CLIP-Adapter (Gao et al. 2024) and Tip-Adapter (Zhang et al. 2022a) enable efficient black-box tuning but suffer from hyperparameter sensitivity and limited text feature utilization. In medical imaging, MedFocusCLIP (Arora and Namboodiri 2025) argues that CoOp is inadequate for fine-grained classification, introducing SAM-based ROIs for attention guidance, while CLIPath (Lai et al. 2023) employs residual adapters for pathology. Despite these advances, most methods still lack unified representations, explicit semantic alignment, and robust cross-family adaptation.

Prompt Learning with Manifold

To improve prompt tuning's adaptability and generalization, manifold-based extensions to CoOp have been explored. ProMetaR (Park, Ko, and Kim 2024) frames prompt tuning as bi-level meta-regularization with Manifold Mixup to enhance generalization, but depends on curated meta-tasks and lacks scalability to diverse models and evolving prompts. ProDA (Lu et al. 2022) models prompts as Gaussians to capture intra-class variance, yet assumes unimodal linear geometry. ProGrad (Zhu et al. 2023) constrains updates along zero-shot gradient directions on the CLIP manifold, improving stability but limiting semantic flexibility. (Chen, Su, and Bao 2024) project prompts into a Poincaré disk to model textual hierarchies, but ignore cross-modal alignment, vision-language gaps, and VLM scalability. MERU (Desai et al. 2023) uses hyperbolic spaces for semantic hierarchies and uncertainty, improving interpretability and alignment but requiring complex optimization and curvature tuning. DualCoOp (Sun, Hu, and Saenko 2022) embeds prompts in label-conditioned manifolds for multi-label tuning, yet overlooks inter-label dependency and efficiency. Collectively, these methods show that distributional or non-Euclidean manifolds enhance prompt alignment and generalization beyond Euclidean tuning, though scalability and adaptability remain key challenges in biomedical settings.

vMF-based Representation Learning

The von Mises–Fisher (vMF) distribution models data on the unit hypersphere via a mean direction and concentration parameter, serving as a fundamental tool for statistical analysis on Riemannian manifolds and modern representation learning (Mardia and Jupp 2009). ProCo (Du et al. 2024) employs class-wise vMF modeling to derive closed-form contrastive losses without large batches, achieving strong long-tailed recognition but lacking cross-modal alignment. T-vMF (Kobayashi 2021) introduces vMF-based angular similarity for compact visual feature clustering, yet does not extend to multimodal learning. In multimodal contexts, CLIP-Enhance (Sandhu et al. 2024) fits vMF mixtures to CLIP embeddings to mitigate modality gaps and long-tail bias, while SSP (Zhu et al. 2024) identifies vMF-based misalignment in CLIP and adopts a projection-based modality bridging strategy, though without class-level adaptability. Meanwhile, vMF Loss (Scott, Gallagher, and Mozer 2021) exploits spherical embedding geometry for calibrated classification but is not designed for prompt-based or multimodal tasks. Our task inherently resides on a hyperspherical manifold, as both CLIP and LLM embeddings are unit-normalized, making vMF modeling particularly suitable for geometrically consistent multimodal alignment.

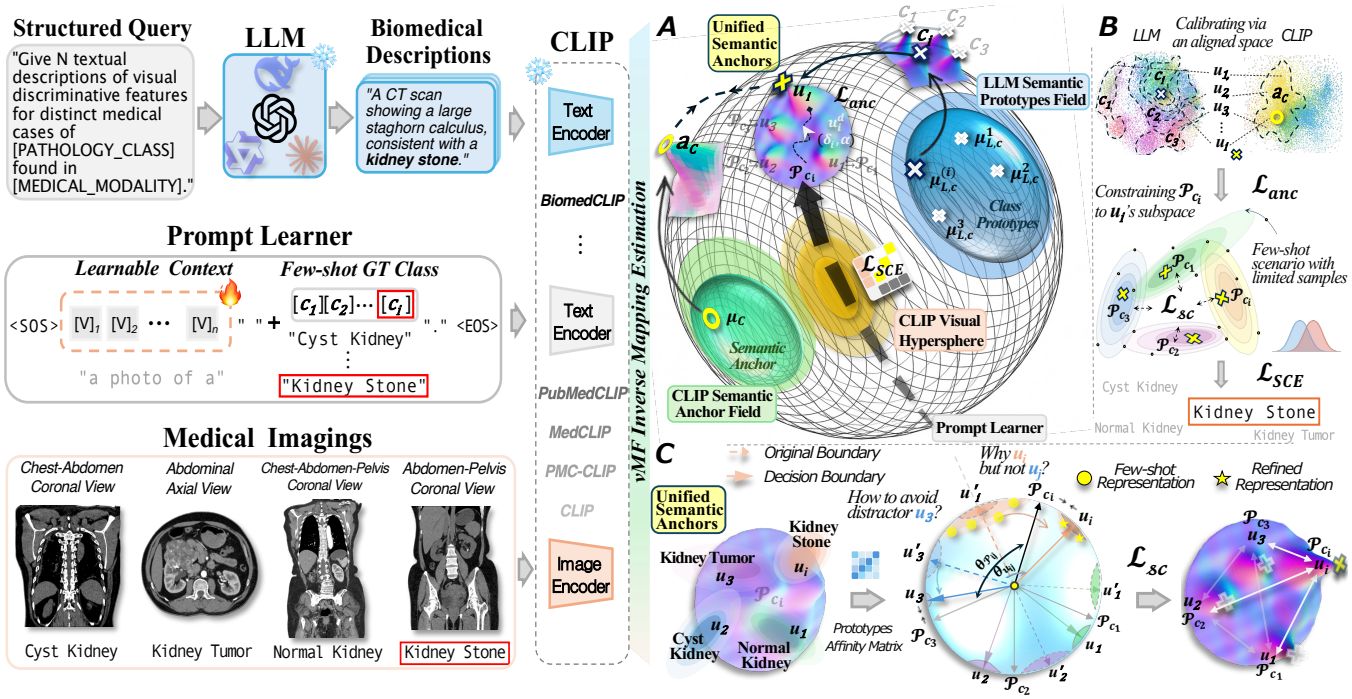

Figure 1: vMFCoOp framework combines structured prompts from an arbitrary LLM, learnable context tokens, and any biomedical CLIP variant to model complex medical images under few-shot settings, with vMF inverse mapping estimation enabling unified semantic alignment on the manifold, guided by 3 constraints. **A**.Unified semantic anchor construction and prompt optimization on the hyperspherical manifold. **B**.Low-dimensional overview of the aligned framework. **C**.Details of the $\mathcal{L}_{sc}$ loss.

Methodology

Our vMFCoOp framework flexibly plugs into any LLM and biomedical-CLIP backbone to jointly learn prompt contexts. It imposes complementary constraints on a shared hyperspherical manifold, guiding prompt and visual embeddings toward a unified semantic anchor field that reconciles multimodal textual biases via the *vMF Inverse Mapping Estimation*, thereby mitigating textual semantic transfer misalignment across diverse model priors and granularity levels in biomedical imaging. Fig. 1B offers an intuitive low-dimensional view, highlighting three hierarchical constraints: *Semantic Anchor Loss* ($\mathcal{L}_{anc}$) aligns the LLM and CLIP embedding spaces; *Spherical Contrastive Loss* ($\mathcal{L}_{sc}$) introduces angular margins to enhance inter-class separability; and *Symmetric Cross-Entropy Loss* ($\mathcal{L}_{SCE}$) for minimizing forward-reverse distributional divergences. This unified manifold ensures that few-shot samples converge toward semantic equilibrium for robust prediction.

Pre-trained Backbones and Prompt Learner Setup

We use BiomedCLIP and other biomedical CLIP variants take a tokenized prompt sequence $\boldsymbol{X_t} \in \mathbb{R}^{B\times L}$ and a preprocessed RGB medical image $\boldsymbol{X_v} \in \mathbb{R}^{B\times 3\times H\times W}$ as input, and encodes them into the embedding space via a text encoder $\boldsymbol{E_t}$ and a vision encoder $\boldsymbol{E_v}$, respectively. The text $\boldsymbol{T}$ and vision embeddings $\boldsymbol{V}$ are given by $\boldsymbol{E_t}(\boldsymbol{X_t}) \in \mathbb{R}^{C\times D}$ and $\boldsymbol{E_v}(\boldsymbol{X_v}) \in \mathbb{R}^{B\times D}$, where B is the batch size, L is the maximum token length, H and W are image dimensions, C is the number of classes, and D is the embedding dimension. Our prompt learner constructs class-specific prompts of the form *"a photo of a* `[CLASS]`*."*, where *"a photo of a"* serves as the initialized learnable context, and `[CLASS]` denotes the class name token. Classification is performed by computing cosine similarities $\cos(,)$ between the image features V and each class prompt embedding $\boldsymbol{T_p^{(i)}}$, scaling these scores by the temperature τ, applying $\mathrm{Softmax}\big(\cos(\boldsymbol{V}, \boldsymbol{T_p^{(i)}})/\tau\big)$ to obtain class probabilities p_i, and selecting the final label as $\arg\max_i\, p(y=i \mid \boldsymbol{V}, \boldsymbol{T_p^{(i)}})$.

Following the setup of BiomedCoOp, we generate prompts for each class c_i in medical modality by querying each of several recent LLMs separately with the standardized instruction:*"Give N textual descriptions of visual discriminative features for distinct medical cases of* `[CLASS]` *found in* `[MODALITY]`*."* We treat all LLM-provided template features as samples distributed on a shared manifold space and estimate a class-level semantic prototype field over this space. For each class c_i, we extract a semantic prototype vector $\boldsymbol{\mu}_{L,c}^{(i)}$, which guides the prompt learner in balancing multi-modal alignment and semantic expressiveness.

Inverse Estimation on Hyperspherical Manifold

In CLIP, both the text and vision encoders produce ℓ_2-normalized output vectors, placing all textual embeddings $\{\boldsymbol{w}_i\}$ naturally on the unit hypersphere S^{d-1}. Likewise,

for each class, n_p prompt templates are generated by each LLM and projected through the CLIP text encoder, yielding prompt embeddings that also lie on S^{d-1}. This allows us to formulate a unified embedding space in $\mathbb{R}^{512}$ by mapping all modalities into a shared hyperspherical manifold, defined as

$$S^{d-1} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}. \tag{1}$$

Definition 1 (von Mises-Fisher Distribution). The vMF distribution, which is particularly suited for modeling directional data, defines a probability density over the unit hypersphere S^{d-1} as: $p_{\text{vMF}}(\mathbf{x}; \boldsymbol{\mu}, \kappa) = C_d(\kappa) \exp\big(\kappa\, \boldsymbol{\mu}^\top \mathbf{x}\big)$, where the mean direction $\boldsymbol{\mu} \in S^{d-1}$ represents the orientation toward which samples concentrate, $\kappa \geq 0$ is the concentration parameter controlling the spread around $\boldsymbol{\mu}$, and $C_d(\kappa)$ is the normalization constant.

We then fit the vMF distribution to CLIP's text-token embeddings and LLM-generated prompt embeddings as directional vectors on the hypersphere to map their semantic bias on the hypersphere, which yields a CLIP Semantic Anchor Field $\mathcal{F}_C$ and an LLM Semantic Prototype Field $\mathcal{F}_{L,c}$. These two fields are subsequently fused to estimate Unified Semantic Anchors that serve as target attractors to guide prompt optimization on the shared hyperspherical manifold.

CLIP Semantic Anchor Field. Let $[\boldsymbol{w}_1; \ldots; \boldsymbol{w}_V] \in \mathbb{R}^{V \times d}$ denote the vocabulary embedding matrix $\boldsymbol{W}$ from the CLIP text encoder, where V is the vocabulary size and d is the embedding dimension, each word embedding $\boldsymbol{w}_i \in S^{d-1}$ lies on the unit hypersphere. We assume that these embeddings are independent and identically distributed samples drawn from a vMF distribution: $\mathcal{F}_C = p_{\text{vMF}}(\boldsymbol{w}; \boldsymbol{\mu}_C, \hat{\kappa}_C) = C_d(\hat{\kappa}_C) \exp(\hat{\kappa}_C \boldsymbol{\mu}_C^\top \boldsymbol{w})$. We estimate the parameters $\boldsymbol{\mu}_C$ and $\hat{\kappa}_C$ using maximum likelihood estimation (MLE), by computing the empirical mean of the vocabulary embeddings $\bar{\boldsymbol{w}} = \frac{1}{V}\sum_{i=1}^{V} \boldsymbol{w}_i$, the Euclidean norm $R = \|\bar{\boldsymbol{w}}\|_2$, and applying:

$$\boldsymbol{\mu}_C = \frac{\bar{\boldsymbol{w}}}{R}, \quad \kappa_C \approx \frac{R(d - R^2)}{1 - R^2 + \epsilon}. \tag{2}$$

This produces our CLIP Semantic Anchor Field $\{\boldsymbol{\mu}_C, \kappa_C\}$, where ϵ is a small constant to avoid division by zero. $\boldsymbol{\mu}_C$ represents the global semantic direction of CLIP's vocabulary, and κ_C quantifies its semantic concentration.

LLM Semantic Prototype Field. For each class c, the LLM generates n_p prompt templates. Let $\boldsymbol{T}_c = \{\boldsymbol{t}_{c,1}, \ldots, \boldsymbol{t}_{c,n_p}\} \subset S^{d-1}$ denote their CLIP-encoded embeddings. We model these class-specific prompt features using a class-conditional vMF distribution: $\mathcal{F}_{L,c} = p_{\text{vMF}}(\boldsymbol{t}; \boldsymbol{\mu}_{L,c}, \kappa_{L,c}) = C_d(\kappa_{L,c}) \exp(\kappa_{L,c} \boldsymbol{\mu}_{L,c}^\top \boldsymbol{t})$. Similarly, with the empirical mean of class-c prompt embeddings $\bar{\boldsymbol{t}}_c = \frac{1}{n_p}\sum_{j=1}^{n_p} \boldsymbol{t}_{c,j}$, and its norm $R_{L,c} = \|\bar{\boldsymbol{t}}_c\|_2$, we estimate:

$$\boldsymbol{\mu}_{L,c} = \frac{\bar{\boldsymbol{t}}_c}{R_{L,c}}, \quad \kappa_{L,c} \approx \frac{R_{L,c}(d - R_{L,c}^2)}{1 - R_{L,c}^2 + \epsilon}. \tag{3}$$

This defines the LLM Semantic Prototype Field $(\boldsymbol{\mu}_{L,c}, \kappa_{L,c})$ for each category c, reflecting the intrinsic semantic prototype bias from the LLM for that class.

Unified Semantic Anchors. On the two hyperspherical fields $\mathcal{F}_C$ and $\mathcal{F}_{L,c}$, to guide prompt learning, we fuse the global CLIP anchor and the class-specific LLM prototype into a unified semantic anchor $\boldsymbol{u}_i$ for class i:

$$\boldsymbol{a}_C = \kappa_C \boldsymbol{\mu}_C, \quad \boldsymbol{c}_i = \kappa_{L,c} \boldsymbol{\mu}_{L,c}, \quad \boldsymbol{u}_i = \frac{\boldsymbol{a}_C + \boldsymbol{c}_i}{\|\boldsymbol{a}_C + \boldsymbol{c}_i\|_2}. \tag{4}$$

Each fused vector $\boldsymbol{u}_i \in S^{d-1}$ serves as a unified semantic anchor, blending CLIP's global directional prior with class-specific prototype semantics derived from the LLM, and thereby steering the prompt learner along the shared hyperspherical manifold.

Prompt Optimization on Hyperspherical Manifold

With a unified hyperspherical manifold that integrates global CLIP semantics and class-specific LLM priors, our next step is to learn prompts that are semantically expressive and cross-modally aligned with visual features on this shared hyperspherical space. While conventional prompt learning operates in flat Euclidean space, our method optimizes over a unit hypersphere, where representations lie on a curved Riemannian manifold governed by angular geometry.

Grounded in our task formulation and hyperspherical geometry, we design prompt optimization around three objectives. First, each class prompt $\boldsymbol{\mathcal{P}}_{c_i}$ is guided toward its unified semantic anchor $\boldsymbol{u}_i$ for balanced alignment between CLIP and LLM spaces. Second, under few-shot supervision, we enhance class discriminability by maintaining angular separation from irrelevant categories. Third, we ensure consistency between visual and prompt-based predictions for cross-modal understanding. The overall loss is defined as $\mathcal{L} = \lambda_{\boldsymbol{anc}} \mathcal{L}_{\boldsymbol{anc}} + \lambda_{\boldsymbol{sc}} \mathcal{L}_{\boldsymbol{sc}} + \mathcal{L}_{\boldsymbol{SCE}}$, where $\lambda_{\boldsymbol{anc}}$ and $\lambda_{\boldsymbol{sc}}$ are balancing weights, with minor adjustments applicable for dataset-specific refinement.

Semantic Anchor Loss. To align each prompt embedding $\boldsymbol{\mathcal{P}}_{c_i}$ with its unified semantic anchor $\boldsymbol{u}_i$, *Semantic Anchor Loss* explicitly constrains the prompt embedding $\boldsymbol{\mathcal{P}}_{c_i}$ to approach its unified semantic target $\boldsymbol{u}_i$ along a directionally controllable path for each class c_i. Instead of altering the anchor itself, we introduce a learnable offset $\boldsymbol{\delta}_i \in \mathbb{R}^d$ and a global scaling factor $\alpha \in \mathbb{R}_{>0}$ to adapt the target direction during optimization. The offset $\boldsymbol{\delta}_i$ navigates the direction on S^{d-1}, while α controls the step size. Together, they enable continuous navigation on the hypersphere during training:

$$\boldsymbol{u}_i^d = \boldsymbol{u}_i + \alpha \cdot \boldsymbol{\delta}_i, \quad \tilde{\boldsymbol{u}}_i^d = \frac{\boldsymbol{u}_i^d}{\|\boldsymbol{u}_i^d\|_2}. \tag{5}$$

Here, dynamic target $\boldsymbol{u}_i^d$ guides the update of the prompt embedding in each training step. The loss is defined as the average squared distance between the normalized prompt embedding $\tilde{\boldsymbol{\mathcal{P}}}_{c_i}$ and its current directional anchor target $\tilde{\boldsymbol{u}}_i^d$ across C classes:

$$\mathcal{L}_{\boldsymbol{anc}} = \frac{1}{C}\sum_{i=1}^{C} \left\| \tilde{\boldsymbol{\mathcal{P}}}_{c_i} - \tilde{\boldsymbol{u}}_i^d \right\|_2^2, \quad \tilde{\boldsymbol{\mathcal{P}}}_{c_i} = \frac{\boldsymbol{\mathcal{P}}_{c_i}}{\|\boldsymbol{\mathcal{P}}_{c_i}\|_2}. \tag{6}$$

This formulation enables the prompt embedding to adaptively follow class-specific directional guidance over the hyperspherical manifold.

Spherical Contrastive Loss. As shown in fig. 1C, to better address few-shot classification within each class c_i, *Spherical Contrastive Loss* aims to increase the angular separation between the prompt embedding $\boldsymbol{\mathcal{P}}_{c_i}$ and distractor-class unified anchors $\boldsymbol{u}_j$, thereby refining the few-shot representations. It does so by formulating two key questions during optimization. *1. Why encourage alignment with $\boldsymbol{u}_i$ rather than $\boldsymbol{u}_j$?* On the hypersphere, decision boundaries are defined by equal angular proximity to class anchors. Thus, maximizing the similarity between prompt $\boldsymbol{\mathcal{P}}_{c_i}$ and its corresponding unified anchor $\boldsymbol{u}_i$ ensures angular closeness to the correct class. We construct a prototype affinity matrix $\boldsymbol{S}$:

$$\boldsymbol{S} = \tau\, \boldsymbol{P}\, \boldsymbol{U}^\top, \quad S_{ij} = \tau\, \boldsymbol{p}_{c_i}^\top \boldsymbol{u}_j = \tau \cos\theta_{ij}, \tag{7}$$

where $\boldsymbol{P} = [\boldsymbol{\mathcal{P}}_1; \ldots; \boldsymbol{\mathcal{P}}_{c_i}]$ and $\boldsymbol{U} = [\boldsymbol{u}_1; \ldots; \boldsymbol{u}_i]$ are $C \times d$ matrices of ℓ_2-normalized prompt features and semantic anchors, respectively, and θ_{ij} is the angle between $\boldsymbol{\mathcal{P}}_i$ and $\boldsymbol{u}_j$. *2. How to avoid distractors?* We apply a row-wise softmax cross-entropy across C classes:

$$\mathcal{L}_{\boldsymbol{sc}} = -\frac{1}{C} \sum_{i=1}^{C} \log \frac{\exp(S_{ii})}{\sum_{j=1}^{C} \exp(S_{ij})}, \tag{8}$$

which simultaneously pulls each prompt toward its correct anchor and pushes it away from all others. To progressively sharpen angular margins, we anneal the temperature τ from τ_0 to τ_{max} using a cosine schedule: $\tau(t) = \tau_{\text{max}} + \frac{1}{2}(\tau_0 - \tau_{\text{max}}) \left(1 + \cos\left(\frac{\pi t}{T}\right)\right)$, resulting in original decision boundaries early in training and increasingly discriminative separation as convergence progresses.

Symmetric Cross-Entropy Loss. To tightly couple CLIP's visual and prompt embeddings, the *Symmetric Cross Entropy Loss* jointly rewards correct class probability and penalizes any mass on incorrect classes. Given logits $L \in \mathbb{R}^{B \times C}$ produced by the dot product of normalized image features and prompt embeddings, with $L_{b,c}$ denoting the score of the b-th sample in the batch B for the c-th class among all C classes, we compute class probabilities $p_{b,c} = \mathrm{Softmax}(L)_{b,c}$. The loss is defined as:

$$\mathcal{L}_{\boldsymbol{SCE}} = -\frac{1}{B} \sum_{b=1}^{B} \Big[\log p_{b,y_b} + \sum_{c=1}^{C} p_{b,c} \log\big(q_{b,c} + \epsilon\big) \Big], \tag{9}$$

where y_b is the ground-truth label for sample b, and $q_{b,c} = \mathbb{1}_{\{c = y_b\}}$ is the corresponding one-hot target. The first term encourages confident predictions on the correct class, aligning visual and prompt representations, while the second term penalizes distributional ambiguity by suppressing incorrect class probabilities, with ϵ preventing instability in the log-domain. Together, $\mathcal{L}_{\boldsymbol{SCE}}$ enhances cross-modal alignment by jointly calibrating certainty and discrimination.

Experiments

We evaluate the vMFCoOp framework under multiple protocols on diverse biomedical benchmarks, including clinically complex datasets like the UK Biobank, to assess accuracy, generalization, and few-shot applicability in realistic medical imaging scenarios. **More experimental results and methodological details are provided in** *Extended version*.

Experimental Setup

Datasets. We conduct experiments on 14 biomedical imaging datasets spanning 13 anatomical regions and 12 medical imaging modalities, including Cardiac Cine MRI (CMRI), Liver MRI (LMRI) and Pancreas MRI (PMRI) from UK Biobank (Sudlow et al. 2015); and Endoscopy (Kvasir (Pogorelov et al. 2017)), Computerized Tomography (CTKidney (Islam et al. 2022)), Brain MRI (BTMRI (Nickparvar 2021)), Histopathology (LC25000 (Borkowski et al. 2019), CHMNIST (Kather et al. 2016)), Dermoscopy (DermaMNIST (Codella et al. 2019; Tschandl, Rosendahl, and Kittler 2018)), Fundus Photography (RETINA (K"ohler et al. 2013)), Optical Coherence Tomography (OCTMNIST (Kermany et al. 2018)), Ultrasound (BUSI (Al-Dhabyani et al. 2020)), and X-ray (COVID-QU-Ex (Tahir et al. 2021), KneeXray (Chen 2018; Porwal et al. 2018)). For the UK Biobank, we select patients active after July 1st, 2020, and derive fine-grained disease labels using ICD-10 codes (Stroganov et al. 2022) with descriptions. This diverse and clinically realistic setup enables comprehensive evaluation across varied biomedical imaging conditions. See the *Extended version* for details on patient lists, data splits, and task definitions.

Few-Shot Learning. We follow the few-shot setting of BiomedCoOp to evaluate model performance under limited supervision, using $K = 1, 2, 4, 8, 16$ labeled samples per class. Additionally, we include evaluations under 32- and 64-shot settings, given that all datasets (except BTMRI) contain over 2,000 training samples. These higher-shot scenarios better reflect realistic clinical conditions and help mitigate overfitting and variance commonly observed in extremely low-shot regimes.

Base-to-Novel Class Generalization. To evaluate generalization, we follow BiomedCoOp by splitting each dataset into base and novel classes. Models are trained on base classes with 16-shot supervision and tested on both base and novel classes, assessing their ability to recognize unseen disease types without fine-tuning. Results for 32- and 64-shot settings are included in the *Extended version*.

Implementation Details. For biomedical CLIPs, we adopt BiomedCLIP (Zhang et al. 2023) with a ViT-B/16 backbone as the default VLM, and additionally evaluate compatibility with PubMedCLIP (Eslami et al. 2023), MedCLIP (Wang et al. 2022), and PMC-CLIP (Lin et al. 2023). We access leading LLMs from major families, including Qwen2.5-72B-Instruct (QwenTeam 2025), Claude 3.5–1022 (Anthropic 2024), and DeepSeek R1 (DeepSeek 2025). Following BiomedCoOp, we use GPT-4 (OpenAI 2024) with 50 prompts per class as the default. To explore the effect of prompt quantity, additional results with 150 prompts per class are provided in the *Extended version*. We initialize the learnable context using the embedding of *"a photo of a"*. The learning rate is set to 0.003 with a batch size of 4. We use SGD with cosine learning rate scheduling. The constraint weights $\lambda_{\boldsymbol{anc}}$ and $\lambda_{\boldsymbol{sc}}$ for each datasets are detailed in the *Extended version*. All experiments are conducted on an NVIDIA A100 GPU (80GB).

Method	$K=1$	$K=2$	$K=4$	$K=8$	$K=16$	$K=32$	$K=64$
Zero-shot Methods							
BiomedCLIP (Zhang et al. 2023)				46.32			
BiomedCLIP (Zhang et al. 2023) + Ensemble				54.46			
BiomedCLIP (Zhang et al. 2023) + Selective Ensemble				56.98			
CLIP-based Adapter Methods							
CLIP-Adapter (Eslami, De Melo, and Meinel 2021)	43.52 ± 1.25	43.27 ± 2.36	45.66 ± 2.01	44.98 ± 3.13	47.25 ± 2.58	48.10 ± 1.55	50.85 ± 1.40
Tip-Adapter (Zhang et al. 2021)	47.58 ± 4.36	51.88 ± 5.69	59.04 ± 4.51	61.37 ± 4.62	66.29 ± 2.10	67.02 ± 2.86	66.45 ± 2.97
Tip-Adapter-F (Zhang et al. 2021)	49.82 ± 7.64	53.34 ± 4.36	59.21 ± 3.28	60.35 ± 2.84	64.27 ± 4.33	68.58 ± 3.59	68.27 ± 0.69
Linear Probing Methods							
Standard LP (Radford et al. 2021)	49.38 ± 5.77	52.85 ± 5.39	58.78 ± 3.87	60.88 ± 7.62	65.75 ± 4.27	70.80 ± 3.45	70.25 ± 2.40
LP++ (Huang et al. 2024)	49.39 ± 7.28	58.62 ± 5.43	62.39 ± 3.10	64.37 ± 5.64	69.40 ± 2.91	71.50 ± 2.30	71.89 ± 2.87
Prompt Learning Methods							
CoOp (Zhou et al. 2022b)	47.63 ± 4.52	49.88 ± 3.27	54.05 ± 2.19	60.32 ± 1.07	68.43 ± 1.57	72.46 ± 1.20	70.30 ± 3.25
CoCoOp (Zhou et al. 2022a)	49.43 ± 3.25	50.77 ± 3.29	54.69 ± 4.79	61.08 ± 3.49	65.09 ± 2.87	66.59 ± 6.50	69.75 ± 4.10
KgCoOp (Yao, Zhang, and Xu 2023)	52.46 ± 3.89	54.78 ± 3.29	58.35 ± 4.05	60.78 ± 6.34	60.95 ± 7.48	71.89 ± 3.54	70.36 ± 2.38
ProGrad (Zhu et al. 2023)	54.97 ± 5.39	56.27 ± 4.46	63.28 ± 4.23	68.87 ± 4.69	70.33 ± 4.98	74.39 ± 1.58	71.25 ± 4.38
BiomedCoOp (Koleilat et al. 2025)	55.08 ± 5.85	57.98 ± 4.20	63.65 ± 3.27	71.29 ± 2.19	73.63 ± 1.27	75.08 ± 3.28	73.65 ± 3.98
vMFCoOp (Ours)	$\mathbf{57.25 \pm 4.75}$	$\mathbf{58.88 \pm 3.95}$	$\mathbf{68.29 \pm 2.07}$	$\mathbf{72.07 \pm 1.98}$	$\mathbf{75.45 \pm 1.48}$	$\mathbf{77.08 \pm 1.36}$	$\mathbf{77.49 \pm 1.05}$

Table 1: Comparison with state-of-the-art methods: average classification accuracy (%) on 14 biomedical datasets, reported as mean ± standard deviation over 3 random support sets per dataset. Best results are highlighted in bold.

Few-shot Benchmarking

We benchmark our method against five representative text-only prompt learning baselines (CoOp, CoCoOp, KgCoOp, ProGrad, BiomedCoOp), three CLIP-based adapter methods (CLIP-Adapter, Tip-Adapter, Tip-Adapter-F), and two linear probing approaches (Standard LP, LP++), all methods that use BiomedCLIP as the backbone are tuned to their optimal settings. Following BiomedCoOp, we focus on shallow text-side adaptations, excluding joint text-image prompt tuning methods (e.g., MaPLe) to ensure fair comparison. As shown in table 1, vMFCoOp consistently outperforms all baselines, with a notable relative gain of 7.29% at 4-shot and 5.22% at 64-shot over the second-best method, BiomedCoOp. These two settings are particularly reflective of clinical conditions, where either only a few labeled cases are available or moderate supervision is feasible. Beyond raw accuracy, we observe that while linear probes benefit from more data, their performance quickly saturates. Prompt learning methods improve with more shots but exhibit diminishing variance reduction, and in some cases (e.g., CoOp), performance degrades due to overfitting. In contrast to competing approaches whose gains plateau or even reverse at higher shot counts, vMFCoOp sustains both accuracy improvements and low variance across all K-shot regimes. This robustness derives from our unified hyperspherical optimization, which preserves the intrinsic semantic topology between LLM and CLIP embeddings and curbs semantic drift.

Base-to-Novel Generalization

We evaluate the generalization capability of vMFCoOp against state-of-the-art prompt-learning methods from base to novel biomedical categories using harmonic mean (HM) accuracy to balance performance across both sets. Due to limited class diversity, the BUSI dataset is excluded from this experiment. Results presented in table 2 demonstrate that vMFCoOp substantially enhances generalization, achieving an average HM of 77.35%, surpassing BiomedCoOp by approximately 4.8%. Notably, vMFCoOp consistently outperforms baseline methods across datasets, particularly in clinically nuanced cases such as BTMRI, CTKIDNEY, and cardiac MRI, reflecting improvements in both base-class and novel-class recognition. This substantial gain is attributed to vMFCoOp's ability to mitigate semantic misalignment through the unified hyperspherical manifold, enabling more precise multimodal semantic alignment, thus providing robust generalization across diverse and complex clinical imaging conditions.

Dataset	Metric	BiomedCLIP	CoCoOp	BiomedCoOp	**vMFCoOp**
Avg on 13 datasets	Base	45.64	67.94	73.26	**78.02**
	Novel	64.66	65.57	71.91	**76.70**
	HM	53.51	66.73	72.58	**77.35**
BTMRI	Base	44.32	79.43	82.73	**83.72**
	Novel	92.57	92.64	95.26	**97.34**
COVID-QU-Ex	Base	51.38	73.47	75.19	**77.42**
	Novel	89.25	89.79	91.08	**93.09**
CTKIDNEY	Base	31.59	80.35	86.64	**87.53**
	Novel	49.21	58.42	75.09	**80.47**
DermaMNIST	Base	36.37	46.53	52.89	**55.67**
	Novel	48.55	64.85	75.28	**76.98**
Kvasir	Base	73.29	85.85	86.72	**87.35**
	Novel	58.47	56.39	60.98	**62.11**
CHMNIST	Base	35.49	86.53	**88.98**	88.35
	Novel	41.38	42.13	42.19	**53.49**
LC25000	Base	55.42	87.98	92.91	**94.07**
	Novel	86.59	94.62	96.85	**97.80**
RETINA	Base	43.52	**70.37**	68.77	70.32
	Novel	51.89	62.29	67.30	**69.22**
KneeXray	Base	31.07	29.85	42.48	**50.19**
	Novel	68.43	62.59	79.32	**79.05**
OCTMNIST	Base	53.25	79.54	79.32	**80.08**
	Novel	49.57	**51.37**	49.27	51.25
CardiacMRI	Base	48.25	53.46	52.59	**64.59**
	Novel	52.78	49.78	47.36	**70.85**
LiverMRI	Base	43.77	39.29	48.29	**66.39**
	Novel	64.38	48.98	66.59	**73.97**
PancreasMRI	Base	45.62	70.59	74.89	**83.58**
	Novel	87.55	78.56	88.23	**91.45**

Table 2: Base-to-Novel Generalization: Accuracy (%) Comparison of BiomedCoOp and SOTA Prompting Approaches.

Visual Interpretability

Interpretability is essential in clinical decision-making. Follow BiomedCoOp, we adopt gScoreCAM (Chen et al. 2022b) to generate saliency maps for five representative

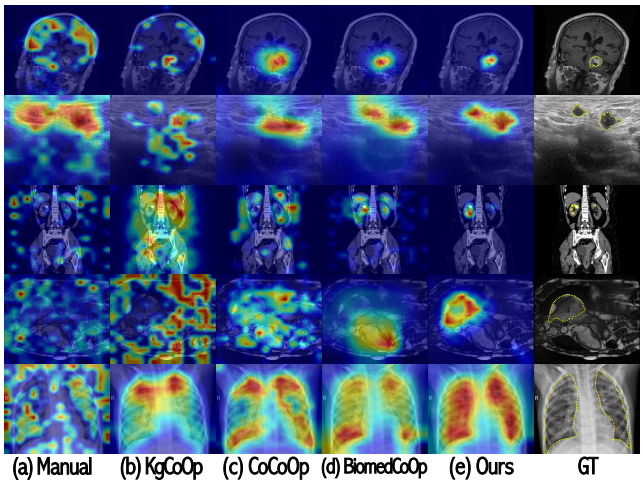

Figure 2: Effect of prompting variations on saliency maps, where (a)–(e) show different strategies, zoom in for details.

samples spanning diverse modalities. As shown in fig. 2, ground-truth lesion contours are overlaid in yellow on the raw images for reference. (a) Manual Prompt with only *"a photo of a* [CLASS]*."* yields scattered focus and often misses fine details. (b) KgCoOp adds manual priors, improving low-level localization (e.g., lung fields), but remains diffuse in complex cases (e.g., tiny kidney stones). (c) CoCoOp learns prompts without prior knowledge, leading to highly unstable and noisy saliency across all examples. (d) BiomedCoOp, with LLM-derived textual priors, better captures coarse organ-level semantics but exhibits modality-dependent bias, it focuses narrowly on the heart region in CMRI and misses the obvious posterior mediastinal tumor. In contrast, our (e) vMFCoOp consistently highlights lesion-centric regions aligned with ground truth, demonstrating superior interpretability and robust cross-modal grounding.

Ablation Study

Effect of Main Components. We summarize the three hierarchical constraints that define the core components of vMFCoOp. In this ablation, we systematically isolate each component under identical settings: $\mathcal{L}_{SCE}$ denotes replacing standard cross-entropy with Symmetric Cross-Entropy Loss atop BiomedCLIP; $\mathcal{L}_{anc}$ introduces a unified semantic anchor field via inverse vMF estimation on the hyperspherical manifold; and $\mathcal{L}_{sc}$ adds Spherical Contrastive Loss to enforce angular margins. As shown in table 3, each constraint provides incremental improvements, while their combination yields the strongest gains. Notably, incorporating $\mathcal{L}_{anc}$ consistently enhances both few-shot performance and base-to-novel generalization, underscoring the importance of aligning multimodal representations via a shared semantic anchor. Adding $\mathcal{L}_{sc}$ further improves novel class accuracy, especially in higher-shot regimes, by reducing overfitting through explicit angular separation. While $\mathcal{L}_{SCE}$ alone offers modest improvements, its integration with the other two constraints amplifies robustness and cross-modal alignment. These results collectively demonstrate that the three constraints are complementary and jointly contribute to optimal generalization and stability in biomedical few-shot learning.

Components			Base-to-Novel			Few-shot				
$\mathcal{L}_{SCE}$	$\mathcal{L}_{anc}$	$\mathcal{L}_{sc}$	Base	Novel	HM	1	4	8	16	32
✗	✗	✗	68.43	42.11	52.14	43.22	47.81	55.27	60.90	62.34
✗	✓	✗	72.65	69.38	70.98	50.29	53.68	60.29	63.87	70.51
✗	✓	✓	73.26	73.59	73.42	49.87	54.23	60.58	73.25	74.02
✓	✗	✗	69.20	47.58	56.39	46.35	48.98	57.34	65.72	67.77
✓	✓	✗	74.35	74.08	74.71	54.88	63.20	70.53	73.58	75.34
✓	✓	✓	**78.11**	**76.22**	**77.15**	**57.25**	**68.29**	**72.07**	**75.45**	**77.08**

Table 3: Ablation of vMFCoOp components with 3 constraints on few-shot and base-to-novel accuracy (%).

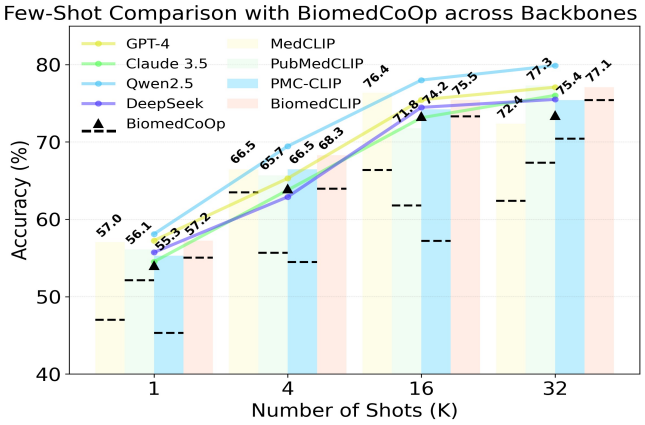

Figure 3: Few-shot across backbones with 50 LLM prompts. Black dashed line and triangle denote BiomedCoOp.

Effect of Backbone Configurations. To further validate vMFCoOp's plug-and-play flexibility across different foundation backbones, fig. 3 contrasts performance under two complementary views. The vertical bars show few-shot accuracy for various CLIP variants when paired with GPT-4, with the black dashed line indicating the corresponding BiomedCoOp baseline. Conversely, the overlaid curves trace accuracy across four different LLMs using BiomedCLIP as the fixed visual backbone, where black triangles mark the average performance over all LLMs. Notably, vMFCoOp consistently outperforms BiomedCoOp across all backbone combinations, yielding both higher accuracy and a markedly flatter response curve, demonstrating robust, stable gains regardless of the underlying CLIP or LLM choice. *Extended version* provides detailed results, including analysis of how the number of LLM prompts affects performance.

Conclusion

Our work bridges the critical semantic gap in prompting biomedical vision-language models by proposing a unified geometric framework that elegantly aligns heterogeneous textual biases from diverse foundational models. By moving beyond Euclidean methods and leveraging the structured semantic manifold defined by vMF distributions, our approach significantly advances few-shot adaptation in complex biomedical imaging scenarios. This geometric perspective resolves semantic misalignment, providing a scalable solution for cross-modal integration in evolving clinical AI.

Ethics Statement

The UK Biobank received ethical approval from the NHS National Research Ethics Service North West (11/NW/0382; 16/NW/0274), and no additional ethics review was required for this study.

Acknowledgments

This research was conducted using data from the UK Biobank Resource under Application Number 603483, as part of an existing approved project. The authors express their sincere gratitude to all participants and research personnel involved in the UK Biobank initiative. This work was further supported by the Scheme Towards Collaborative Cloud-Edge Deep Learning Deployment (Grant IEC/NSFC/223523), the National Edge AI Hub for Real Data: Edge Intelligence for Cyber-Disturbances and Data Quality (EP/Y028813/1), and the UK Medical Research Council (MRC) Innovation Fellowship (Grant MR/S003916/2).

References

Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; and Fahmy, A. 2020. Dataset of breast ultrasound images. *Data in Brief*.

Anthropic. 2024. Claude by Anthropic.

Arora, A.; and Namboodiri, V. 2025. MedFocusCLIP: Improving few shot classification in medical datasets using pixel wise attention. In *IEEE ICASSP*.

Borkowski, A. A.; Bui, M. M.; Thomas, L. B.; Wilson, C. P.; DeLand, L. A.; and Mastorides, S. M. 2019. Lung and colon cancer histopathological image dataset (LC25000). *arXiv*.

Chen, G.; Yao, W.; Song, X.; Li, X.; Rao, Y.; and Zhang, K. 2022a. Prompt learning with optimal transport for vision-language models. *arXiv*.

Chen, N.; Su, X.; and Bao, F. 2024. Hyperbolic representations for prompt learning. In *LREC-COLING*.

Chen, P. 2018. Knee Osteoarthritis Severity Grading Dataset.

Chen, P.; Li, Q.; Biaz, S.; Bui, T.; and Nguyen, A. 2022b. gScoreCAM: What objects is CLIP looking at? In *ACCV*.

Chen, Y.; Xu, D.; Huang, Y.; Zhan, S.; Wang, H.; Chen, D.; Wang, X.; Qiu, M.; and Li, H. 2025. MIMO: A Medical Vision Language Model with Visual Referring Multimodal Input and Pixel Grounding Multimodal Output. In *CVPR*.

Cho, E.; Kim, J.; and Kim, H. J. 2023. Distribution-aware prompt tuning for vision-language models. In *ICCV*.

Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M. E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv*.

DeepSeek. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv*.

Desai, K.; Nickel, M.; Rajpurohit, T.; Johnson, J.; and Vedantam, S. R. 2023. Hyperbolic image-text representations. In *ICML*.

Du, C.; Wang, Y.; Song, S.; and Huang, G. 2024. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE TPAMI*.

Eslami, S.; De Melo, G.; and Meinel, C. 2021. Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv*.

Eslami et al. 2023. PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain? In *EACL(ACL Findings)*.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. CLIP-Adapter: Better vision-language models with feature adapters. *IJCV*.

Huang, S.; Xu, T.; Shen, N.; Mu, F.; and Li, J. 2023. Rethinking few-shot medical segmentation: a vector quantization view. In *CVPR*.

Huang, Y.; Shakeri, F.; Dolz, J.; Boudiaf, M.; Bahig, H.; and Ben Ayed, I. 2024. LP++: A surprisingly strong linear probe for few-shot CLIP. In *CVPR*.

Islam, M. N.; Hasan, M.; Hossain, M. K.; Alam, M. G. R.; Uddin, M. Z.; and Soylu, A. 2022. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. *Scientific Reports*.

Kather, J. N.; Weis, C.-A.; Bianconi, F.; Melchers, S. M.; Schad, L. R.; Gaiser, T.; Marx, A.; and Z"ollner, F. G. 2016. Multi-class texture analysis in colorectal cancer histology. *Scientific Reports*.

Kermany, D. S.; Goldbaum, M.; Cai, W.; Valentim, C. C.; Liang, H.; Baxter, S. L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023a. MaPLe: Multi-modal prompt learning. In *CVPR*.

Khattak, M. U.; Wasim, S. T.; Naseer, M.; Khan, S.; Yang, M.-H.; and Khan, F. S. 2023b. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*.

Kobayashi, T. 2021. T-vMF similarity for regularizing intra-class feature distribution. In *CVPR*.

K"ohler, T.; Budai, A.; Kraus, M. F.; Odstrčilik, J.; Michelson, G.; and Hornegger, J. 2013. Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In *26th IEEE International Symposium on Computer-Based Medical Systems*.

Koleilat, T.; Asgariandehkordi, H.; Rivaz, H.; and Xiao, Y. 2025. BiomedCoOp: Learning to prompt for biomedical vision-language models. In *CVPR*.

Lai, Z.; Li, Z.; Oliveira, L. C.; Chauhan, J.; Dugger, B. N.; and Chuah, C.-N. 2023. CLIPaTH: Fine-tune CLIP with visual feature fusion for pathology image analysis towards minimizing data collection efforts. In *ICCV*.

Li, Z.; Li, X.; Fu, X.; Zhang, X.; Wang, W.; Chen, S.; and Yang, J. 2024. Promptkd: Unsupervised prompt distillation for vision-language models. In *CVPR*.

Lin, W.; Zhao, Z.; Zhang, X.; Wu, C.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. PMC-CLIP: Contrastive language-image pre-training using biomedical documents. In *MICCAI*.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt distribution learning. In *CVPR*.

Mardia, K. V.; and Jupp, P. E. 2009. *Directional Statistics*. John Wiley & Sons.

Nickparvar, M. 2021. Brain tumor MRI dataset. *Kaggle*.

OpenAI. 2024. GPT-4 Technical Report. *arXiv*.

Park, J.; Ko, J.; and Kim, H. J. 2024. Prompt learning via meta-regularization. In *CVPR*.

Pogorelov, K.; Randel, K. R.; Griwodz, C.; Eskeland, S. L.; de Lange, T.; Johansen, D.; Spampinato, C.; Dang-Nguyen, D.-T.; Lux, M.; Schmidt, P. T.; et al. 2017. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *8th ACM on Multimedia Systems Conference*.

Porwal, P.; Pachade, S.; Kamble, R.; Kokare, M.; Deshmukh, G.; Sahasrabuddhe, V.; and Meriaudeau, F. 2018. Indian Diabetic Retinopathy Image Dataset (IDRiD): a database for diabetic retinopathy screening research. *Data*.

Qiu, K.; Gao, Z.; Zhou, Z.; Sun, M.; and Guo, Y. 2025. Noise-Consistent Siamese-Diffusion for Medical Image Synthesis and Segmentation. In *CVPR*.

QwenTeam. 2025. Qwen2.5 Technical Report. *arXiv*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Sandhu, M.; Pequignot, Y. B.; Nashed, S. B.; Sahoo, S.; and Paull, L. 2024. CLIP-Enhance: Improving CLIP Zero-Shot Classification via von Mises-Fisher Clustering. *arXiv*.

Schmalfuss, J.; Chang, N.; VS, V.; Shen, M.; Bruhn, A.; and Alvarez, J. M. 2025. PARC: A Quantitative Framework Uncovering the Symmetries within Vision Language Models. In *CVPR*.

Scott, T. R.; Gallagher, A. C.; and Mozer, M. C. 2021. von Mises-Fisher loss: An exploration of embedding geometries for supervised learning. In *ICCV*.

Shao, M.; Miao, X.; Duan, H.; Wang, Z.; Chen, J.; Huang, Y.; Wu, X.; Deng, J.; Long, Y.; and Zheng, Y. 2025a. TRACE: Temporally Reliable Anatomically-Conditioned 3D CT Generation with Enhanced Efficiency. In *MICCAI*.

Shao, M.; Wang, Z.; Duan, H.; Huang, Y.; Zhai, B.; Wang, S.; Long, Y.; and Zheng, Y. 2025b. Rethinking Brain Tumor Segmentation from the Frequency Domain Perspective. *IEEE Transactions on Medical Imaging*.

Stroganov, O.; Fedarovich, A.; Wong, E.; Skovpen, Y.; Pakhomova, E.; Grishagin, I.; Fedarovich, D.; Khasanova, T.; Merberg, D.; Szalma, S.; et al. 2022. Mapping of UK Biobank clinical codes: Challenges and possible solutions. *PLOS ONE*.

Sudlow, C.; Gallacher, J.; Allen, N.; Beral, V.; Burton, P.; Danesh, J.; Downey, P.; Elliott, P.; Green, J.; Landray, M.; et al. 2015. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*.

Sun, X.; Hu, P.; and Saenko, K. 2022. DualCoOp: Fast adaptation to multi-label recognition with limited annotations. *NeurIPS*.

Tahir, A. M.; Chowdhury, M. E.; Khandakar, A.; Rahman, T.; Qiblawey, Y.; Khurshid, U.; Kiranyaz, S.; Ibtehaz, N.; Rahman, M. S.; Al-Maadeed, S.; et al. 2021. COVID-19 infection localization and severity grading from chest X-ray images. *Computers in Biology and Medicine*.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*.

Wang, R.; Zheng, H.; Duan, X.; Liu, J.; Lu, Y.; Wang, T.; Xu, S.; and Zhang, B. 2023. Few-shot learning with visual distribution calibration and cross-modal distribution alignment. In *CVPR*.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022. MedCLIP: Contrastive learning from unpaired medical images and text. In *EMNLP*.

Yao, H.; Zhang, R.; and Xu, C. 2023. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*.

Zhang, R.; Fang, R.; Zhang, W.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2021. Tip-Adapter: Training-free CLIP-Adapter for better vision-language modeling. *arXiv*.

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022a. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *ECCV*.

Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; et al. 2023. BioMedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv*.

Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C. D.; and Langlotz, C. P. 2022b. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*.

Zheng, H.; Yang, S.; He, Z.; Yang, J.; and Huang, Z. 2025. Hierarchical Cross-modal Prompt Learning for Vision-Language Models. *arXiv*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *CVPR*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *IJCV*.

Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; and Zhang, H. 2023. Prompt-aligned gradient for prompt tuning. In *ICCV*.

Zhu, X.; Zhu, B.; Tan, Y.; Wang, S.; Hao, Y.; and Zhang, H. 2024. Selective vision-language subspace projection for few-shot CLIP. In *32nd ACMMM*.