BioDPP: Dynamic Prompt Policy Learning for Biomedical Vision-Language Models

Pingyi Miao[1,2], Xianlai Chen[1,2], Kai Sun[3], Yunbo Wang[1,2*], Shuang Zhao[4], Ying An[1,2]

[1]Big Data Institute, Central South University, China
[2]National Engineering Research Center for Medical Big Data Application Technology, Central South University, China
[3] Furong Laboratory, Central South University, China
[4] Xiangya Hospital, Central South University, China
{241812022, chenxianlai, sunkai16, wangyunbo, shuangxy, anying}@csu.edu.cn

Abstract

Foundational vision-language models (VLMs), such as CLIP, are emerging as a promising paradigm in vision tasks due to their strong generalization ability. Nevertheless, adapting them to downstream tasks remains challenging, especially in biomedical imaging, where scarce annotations, low-contrast features and complex patterns hinder model adaptation. Thus, prompt tuning is employed to facilitate the adaptation of VLMs. However, current prompt tuning methods like Context Optimization (CoOp) mainly learn a single yet static prompt which is applied to all images, and such *one-size-fits-all* prompt cannot describe the case-specific diagnostic cues in biomedical data, compromising the adaptation of VLMs. To this end, we propose a Dynamic Prompt Policy learning method that enables efficient adaptation of Biomedical VLMs (BioDPP) for accurate and highly generalizable few-shot biomedical image classification. Specifically, we conceptualize the learnable context as an agent, and present a paradigm of learning a dynamic prompting policy, rather than obtaining a single yet static prompt. Wherein, a dual-reward mechanism is developed to guide policy learning via the feedback on both classification decision and the consistency between the prompt and the context, steering the agent to generate context-aware prompts. Moreover, we devise adaptive baseline stabilization to dynamically regulate reward advantage value throughout the training process, enabling policy refinement in a complex reward space tailored to biomedical VLMs. Extensive experiments are conducted on 10 biomedical datasets, and the results reveal that our BioDPP achieves superior performance, demonstrating more efficient prompt optimization in biomedical VLMs.

Code — https://github.com/Miaopingyi/BioDPP

Introduction

Recent breakthroughs in vision-language models (VLMs) have opened new avenues for utilizing multimodal data across various applications. Unlike traditional supervised models that focus on closed-set visual concepts, current VLMs like CLIP (Radford et al. 2021) use large-scale image-text pairs to align visual and textual information via contrastive pre-training. With the help of natural language supervision, they explore open-set visual concepts and show strong generalization capability. Although these models are adept at processing visual and textual information, training them requires large-scale, high-quality multimodal data. However, in real-world tasks, collecting sufficient task-relevant data is particularly challenging. Besides, VLMs heavily depend on the quality of textual prompts (Zhou et al. 2022b). For instance, CLIP employs manually-designed prompt like "a photo of [CLASS]". Yet such static prompts are often ineffective in downstream tasks. Studies show that adding attributes (An et al. 2023) or incorporating detailed category descriptions (Pratt et al. 2023; Jin et al. 2021) improves VLM performance. However, designing high-quality prompts that accurately describe text or image context remains challenging.

Therefore, prompt learning is proposed to automatically optimize the textual prompt in VLMs, improving the adaptability and generalization ability of VLMs in downstream tasks. The prompt learning-based method aims to obtain discriminative and task-specific textual prompt, rather than relying on hand-crafted prompt for textual description. Representative Context Optimization approaches like CoOp (Zhou et al. 2022b) and CoCoOp (Zhou et al. 2022a) replace manual prompt with learnable context vectors to better capture category semantics. The learned task-specific knowledge in these methods effectively improves the performance of VLMs on known categories. However, these methods often show poor generalization ability when facing new categories. In addition, some methods (Gao et al. 2024; Zhang et al. 2022) exploit lightweight or few-shot adaptation by using adapters (Houlsby et al. 2019), and Linear Probes (Huang et al. 2024) further presents parameter-efficient solution to adapt VLMs to downstream tasks. Meanwhile, BiomedCoOp (Koleilat et al. 2025) and ProText (Khattak et al. 2025) leverage Large Language Model (LLM) to automatically generate category descriptions, replacing handcrafted descriptions.

Different from natural images, biomedical images encompass a wide range of contrasts and modalities. Images of different modalities like CT, MRI and Pathology often exhibit subtle, highly-variable diagnostic features, while VLMs are also constrained by the scarcity of well-annotated data due to

*Corresponding author

privacy concerns and high-cost annotation. Although prompt tuning shows promise for adapting VLMs to this field, current methods typically focus on obtaining a single static prompt (Zhou et al. 2022b; Koleilat et al. 2025) applied to all data analysis. In the complex context of biomedical imaging(Ma et al. 2024), such one-size-fits-all prompt can be suboptimal, as it is optimized on the most common characteristics within a dataset, failing to capture case-specific diagnostic cues in each image. Furthermore, these methods lack a dynamic feedback mechanism to guide their prompt optimization when generalizing from familiar base classes to new ones. Therefore, it is significant to explore the transformation of prompt optimization from obtaining a single yet static prompt to active prompt policy learning, which enables generating specific context-aware prompts for adapting VLMs to the complex biomedical domain effectively.

To address the above issues, we propose Dynamic Prompt Policy learning for Biomedical vision-language models (BioDPP), shifting the paradigm from obtaining a single static prompt to learning a dynamic prompt policy, reframing prompt optimization as a reinforcement learning problem. In BioDPP, the learnable context is conceptualized as an agent, which learns a dynamic policy to generate context-aware prompts for each image via observing its distinct characteristics. A Dual-Reward Mechanism (DRM) is constructed to guide policy learning, where feedback from both the classification decision and the consistency between the prompt and context drives the agent to generate specific prompts. Meanwhile, Adaptive Baseline Stabilization (ABS) is devised to refine the policy by dynamically regulating the reward advantage throughout training.

The contributions of this work are summarized as follows:

- We propose a dynamic prompt policy learning method that reframes prompt optimization by conceptualizing the learnable context as an agent. This agent learns a prompting policy to generate context-aware prompts tailored to the content of each biomedical image.
- We develop a dual-reward mechanism that provides multi-dimensional feedback to guide the agent's policy learning by simultaneously evaluating the classification decision and the prompt semantic representation.
- We design an adaptive baseline stabilization that dynamically regulates reward advantage value throughout the training process, enabling policy refinement within the complex reward space of biomedical vision-language models.
- Extensive experiments are conducted on 10 biomedical datasets, spanning 8 modalities and 9 organs in few-shot and base-to-novel benchmarks. The results demonstrate BioDPP's superior generalization ability across various medical conditions and imaging modalities.

Related Works

Foundation Vision-Language Models

Vision-language models, such as CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), are pre-trained in a self-supervised manner using large-scale image-text data. With a contrastive learning objective, VLMs learn rich multimodal representations by drawing matched image-text pairs closer together while pushing unmatched pairs away. This process enables them to achieve exceptional performance in downstream tasks, including zero-shot classification and cross-modal retrieval. Recent extensions to the biomedical domain, BiomedCLIP (Zhang et al. 2023), BiomedGPT (Zhang et al. 2024) and CONCH (Lu et al. 2024) leverage extensive biomedical data to acquire generalizable biomedical knowledge to enhance representations. However, these models usually require additional task-specific adaptation to capture the disease-specific cues for an ideal performance in the biomedical domain.

Prompt Learning for VLMs Adaptation

Prompt learning is an effective solution for fine-tuning VLMs on downstream tasks. For example, CoOp (Zhou et al. 2022b) and CoCoOp (Zhou et al. 2022a) replace manual prompts with learnable soft prompts. Such approach is crucial for few-shot adaptation in data-scarce fields like biomedicine, as it avoids altering the weights in pre-trained VLMs. Recent advancements in the general domain offer diverse strategies. ProDA (Lu et al. 2022) incorporates prompts distributions, and MaPle (Khattak et al. 2023a) introduces multimodal prompting for both the vision and language branches. To improve generalization, KgCoOp (Yao, Zhang, and Xu 2023) utilizes knowledge graphs, ProGrad (Zhu et al. 2023) employs gradient-guided optimization, and PromptSRC (Khattak et al. 2023b) jointly considers features from multiple sources. Other approaches use adapters for VLMs adaptation: DenseCLIP (Rao et al. 2022) adapts the VLMs for dense prediction tasks, and CLIP-Adapter (Gao et al. 2024) and Tip-Adapter modify visual or textual embeddings. Lastly, ProText (Khattak et al. 2025) leverages LLMs to enhance prompts for better transferability. The above general methods are limited in specialized biomedical applications, such as explaining lesion morphology. Thus, BiomedCoOp (Koleilat et al. 2025) presents a tailored solution by integrating LLM knowledge with a biomedical VLMs. However, these customized methods generally require complex adaptations to their vision and language components.

Reinforcement Learning for VLMs Tuning

Reinforcement Learning (RL) has been explored to facilitate VLMs tuning. One line of works uses RL to optimize prompts for VLMs. These methods frame prompt generation as a sequential decision-making process beyond fixed optimization objectives. For example, RLPROMPT (Deng et al. 2022) utilizes policy gradients to find optimal discrete text prompts. Vision-R1 (Zhan et al. 2025) applies a visually-guided RL approach where visual feedback helps refine task understanding. Another stream of works leverages VLMs to assist RL agents as zero-shot reward models. This allows VLMs to provide dense and semantic feedback by evaluating visual states against language goals without task-specific fine-tuning (Rocamonde et al. 2024). Additionally, PR2L (Chen et al. 2023) shows how prompts can instruct VLMs to extract task-relevant features for the state representation

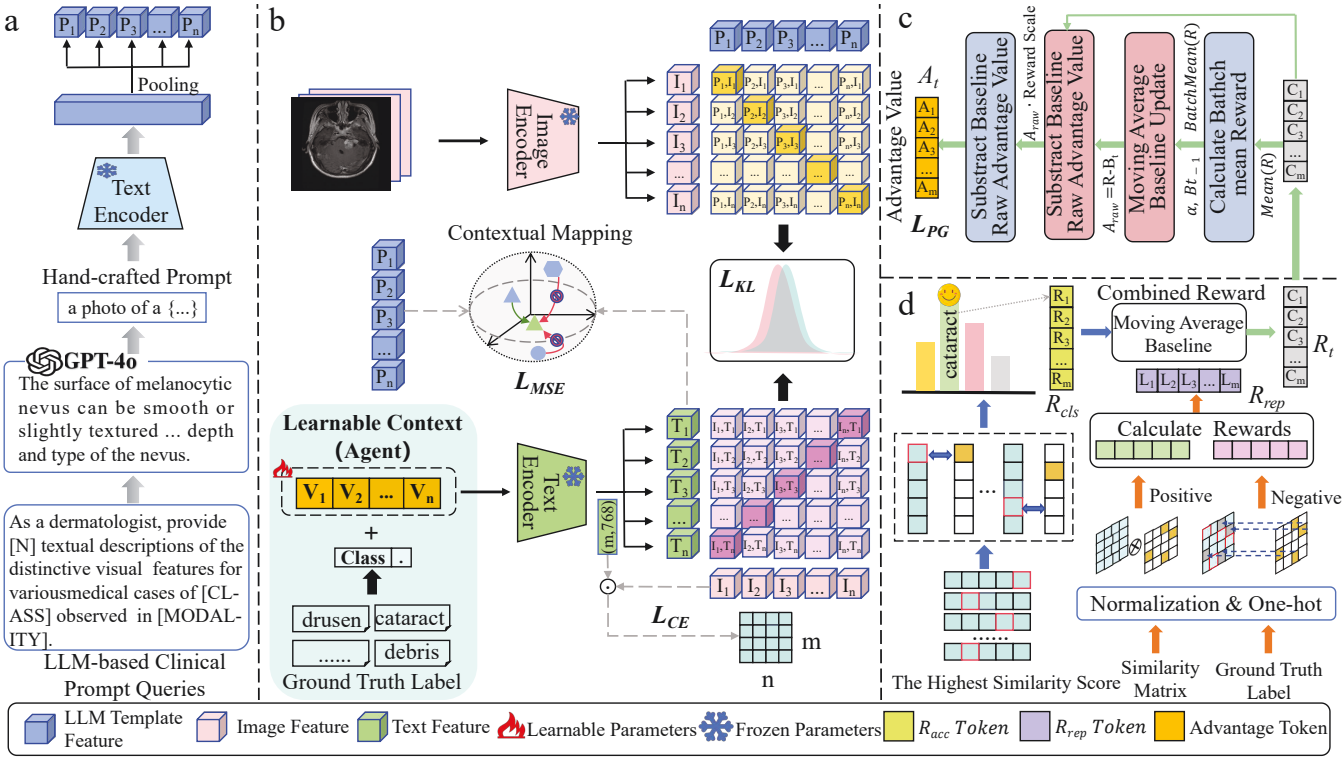

Figure 1: An overview of the BioDPP. **(a)** A Large Language Model (LLM) generates teacher prompts. **(b)** The student model is supervised via classification ($\mathcal{L}_{CE}$) and knowledge distillation ($\mathcal{L}_{MSE}$, $\mathcal{L}_{KL}$) losses. **(c)** The Adaptive Baseline Stabilization (ABS) module calculates a stable advantage signal for the policy gradient loss ($\mathcal{L}_{PG}$). **(d)** The Dual-Reward Mechanism (DRM) provides policy learning feedback by fusing rewards from both the classification decision and prompt semantic representation.

of the agent. However, these methods mainly rely on a single objective reward like classification accuracy, and fail to ensure the semantic understanding of critical diagnostic cues when applied in the biomedical domain.

Methodology

This section first outlines the foundational concepts of our work, including prompt learning in VLMs and prompt ensembles creation via LLMs. Upon these concepts, we introduce the proposed BioDPP, and the overall framework of the BioDPP is shown in Fig. 1.

Preliminaries

Contrastive Vision-Language Pre-training. CLIP is pre-trained on a massive dataset of image-text pairs from the Internet via using a contrastive learning objective. The crucial idea is to learn a shared embedding space, where the features of an image and its matched text are pulled closer together, while features from non-matched image and text are pushed away(Wang and Peng 2021; Wang et al. 2020b). This allows to acquire generalizable visual concepts that are aligned with natural language.

Zero-shot Transfer Inference. A key capability of pre-trained VLMs like CLIP is capable of performing zero-shot classification on downstream tasks without fine-tuning. This is achieved by reformulating the classification task as an image-text matching problem. For a given set of classes, a hand-crafted text prompt, such as "a photo of a [CLASS]", is created for each class. The image is encoded by the visual encoder to produce an image feature f, and different types of prompts are encoded by the text encoder to produce a set of text features $\{w_1, w_2, ..., w_K\}$. The probability of the image x belonging to the i-th class can be calculated as follows:

$$p(y=i|x) = \frac{\exp(s(w_i, f)/\tau)}{\sum_{j=1}^{K} \exp(s(w_j, f)/\tau)} \quad (1)$$

where τ is a temperature parameter, $s(\cdot,\cdot)$ denotes the cosine similarity.

Prompt Learning. Although VLMs show great potential for zero-shot inference, the constructed hand-crafted prompt is often suboptimal for specific downstream tasks (Zhou et al. 2022b). Therefore, prompt learning-based methods like CoOp (Zhou et al. 2022b) and KgCoOp (Yao, Zhang, and Xu 2023) are proposed. Instead of using a fixed text template, these methods try to learn a set of continuous vectors, known as context vectors $V = \{v_1, v_2, ..., v_M\}$, which are prepended to the class token embedding c_i. The entire prompt $t_i = \{v_1, v_2, ..., v_M, c_i\}$ is then fed into the text encoder. These context vectors are optimized directly on downstream data by minimizing the cross-entropy loss:

$$\mathcal{L}_{ce}(V) = -\sum_{(x,y)\in\mathcal{D}_{train}} \log p(y|x;V) \tag{2}$$

where (x, y) denotes the image-label pairs on training dataset $\mathcal{D}_{train}$. $p(y|x;V)$ denotes the probability from Eq. (1). With th Eq. (2), the text features w_i are generated conditioned on the learnable context vectors V.

Prompt Ensembling with LLM Descriptions. To enrich the semantic guidance in prompt tuning, recent works try to incorporate LLMs into prompt learning. The powerful generative capability of LLMs can be leveraged to create a diverse set of descriptive teacher prompts tailored to each class. These ensembles capture a much wider range of semantic variations and contextual nuances than a single handcrafted template. In our BioDPP, the LLM-generated descriptions serve as a source of general knowledge to guide the optimization of learnable student prompts via knowledge distillation. Specifically, we align the student model's prompt embeddings and feature distributions with those of teacher model, and the Mean Squared Error and KL-Divergence are used as the constraint.

Since the context to be learned is unified among all the classes. Given that the learnable context vectors are a set $V = \{v_1, v_2, ..., v_M\}$, the Mean Squared Error ($\mathcal{L}_{\text{MSE}}$) is adopted to minimize the difference between the student prompt embedding (V_s) and the teacher prompt embedding (V_t), aligning their semantic representations. The formulation is listed as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{M}\sum_{i=1}^{M} \|v_{s_i} - v_{t_i}\|_2^2 \tag{3}$$

where M is the number of context vectors.

To align the distribution of the probability from image embeddings with learnable context prompts (student P_s) and the distribution from image embeddings with selective LLM-generated text embeddings (teacher P_t), we minimize the KL divergence between these two distributions:

$$\mathcal{L}_{\text{KL}} = \sum_{c=1}^{K} P_t(y=c|x) \log \frac{P_t(y=c|x)}{P_s(y=c|x)} \tag{4}$$

where K is the number of classes. This alignment ensures that the learned embeddings maintain essential information about the biomedical images, preventing the model from diverging into unrelated semantic spaces.

Prompt Optimization via Policy Learning

Our approach formulates prompt optimization as a dynamic process within a reinforcement learning framework. In our BioDPP, we conceptualize the learnable context as an agent that learns a policy (π_θ) for prompt generation. This policy guides the agent to construct a context-aware prompt, dynamically tailored to the distinct visual information of each image. At each step, the agent observes an image's visual state (s_t) and performs an action (a_t), generating a set of learnable context-aware vectors P_t. This process is governed by the policy (π_θ), which is a parameterized function for mapping the states to a distribution over actions, $\pi_\theta(a_t|s_t)$. After each action, the agent receives a reward (R_t) that evaluates the quality of the generated prompts.

A reward based solely on final task performance (i.e., classification accuracy) is insufficient. The dynamic process of generating context-aware prompts requires a more nuanced feedback mechanism to guide the agent. Therefore, we design a multi-dimensional reward combining the classification decision reward and the prompt semantic representation reward.

Classification Decision Reward ($\mathcal{R}_{cls}$). To effectively guide the policy, we employ a direct and clear binary reward signal. Given an input image x and its ground truth label y_{truth}, the agent receives a positive reward of 1 if the model's predicted class matches the ground truth, and 0 otherwise. This reward directly encourages the policy to learn prompts that produce correct classifications. The formulation is as follows:

$$\mathcal{R}_{cls} = \begin{cases} 1 & \text{if } \arg\max_i p(y=i|x) == y_{\text{truth}} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

where i is an index over the set of all classes, and the $\arg\max$ function returns the class index that maximizes the probability $p(y = i|x)$.

Semantic Representation Reward ($\mathcal{R}_{rep}$). To guide the agent generating semantically rich context-aware prompts, we introduce the semantic representation reward $\mathcal{R}_{rep}$, based on noise contrastive estimation (van den Oord, Li, and Vinyals 2019; Wang et al. 2025). Instead of only penalizing the most competitive negative class, this reward contrasts the positive pair (image and its ground truth class) against all possible negative pairs (same image and other classes), encouraging more discriminative feature representations(Wang et al. 2020a; Yin et al. 2025). The reward can be obtained according to the negative log-likelihood of correctly classifying the positive sample:

$$\mathcal{R}_{rep} = \frac{1}{\tau} s(f, w_{y_{\text{true}}}) - \log\left(\sum_{j=1}^{K} \exp(s(f, w_j)/\tau)\right) \tag{6}$$

where $s(\cdot,\cdot)$ denotes the cosine similarity, K is the number of classes, and τ is a temperature hyperparameter that controls the sharpness of the distribution. A higher reward signifies that the true class is more distinct from all other classes. To enhance this policy learning, we combine these two rewards. The final reward R_t is defined as a weighted combination of the above two components:

$$R_t = \omega \cdot \mathcal{R}_{cls} + (1-\omega) \cdot \mathcal{R}_{rep} \tag{7}$$

where $\omega \in [0, 1]$ is a hyperparameter to balance the two reward signals for a more stable and effective policy learning.

Adaptive Baseline Stabilization

The constructed DRM can provide a rich reward signal on prompt policy learning, but such signal is prone to high variance in gradient estimation, resulting in unstable training and slow convergence(Su et al. 2025). Thus, a common

Method	K = 1	K = 2	K = 4	K = 8	K = 16
Zero-shot Methods					
CLIP			24.69		
BiomedCLIP			43.26		
BiomedCLIP+Ensemble			52.05		
CLIP-based Adapter Methods					
CLIP-Adapter	45.94±1.24	44.44±1.13	44.74±1.33	45.65±1.14	47.25±1.04
Tip-Adapter-F	55.65±1.03	53.33±2.25	61.79±2.05	66.67±2.88	69.89±1.36
Linear Probing Methods					
Standard LP	52.28±4.00	55.87±3.76	62.39±2.12	68.10±2.14	69.07±2.40
LP++	53.29±5.78	53.77±4.21	58.42±2.16	65.06±2.93	68.99±1.81
Prompt Learning Methods					
CoCoOp	48.45±3.85	51.32±3.62	55.31±3.83	61.66±2.94	65.54±2.17
CoOp	49.83±5.84	54.22±3.79	60.41±3.35	65.38±1.63	70.15±1.69
KgCoOp	52.01±4.68	53.53±4.18	59.29±3.84	63.78±1.91	65.11±1.44
ProGrad	51.58±5.73	54.54±3.71	60.72±3.99	65.59±2.07	67.50±2.05
BiomedCoOp	55.05±4.91	58.43±2.78	65.47±2.17	69.50±2.09	72.70±1.15
BioDPP (Ours)	**57.10±3.21**	**59.37±2.91**	**66.67±1.91**	**70.57±2.36**	**73.61±1.20**

Table 1: Comparison with State-of-the-Art Methods. The table reports the average classification accuracy (%) across 10 benchmark datasets. For each dataset, results are averaged over 3 randomly sampled support sets, and reported as mean ± standard deviation. The best performance in each setting is highlighted in bold.

technique is to subtract a baseline from the reward, reshaping the learning signal into an advantage value. This advantage value represents how much better an action is than the expected average, thereby reducing the variance and stabilizing the policy learning process(Wang et al. 2023).

Instead of training a complex and computationally expensive neural network critic to estimate the baseline B_t, we introduce a simple yet highly effective strategy: Adaptive Baseline Stabilization (ABS). ABS estimates the expected reward using an Exponential Moving Average (EMA) of the rewards observed in recent batches. At any given training step t, the baseline B_t is formally defined as a weighted sum of the mean rewards from all past batches $(\bar{R}_1, \ldots, \bar{R}_t)$. This strategy starts with an initial baseline B_0 and uses a momentum hyperparameter $\alpha \in [0, 1]$ to apply exponentially decaying weights to the reward history. The specific equation is formulated as follows:

$$B_t = \alpha^t B_0 + (1 - \alpha) \sum_{k=1}^{t} \alpha^{t-k} \bar{R}_k \tag{8}$$

where $\bar{R}_k$ represents the mean reward within the batch at step k. This equation is mathematically equivalent to the original EMA update, and explicitly shows the exponentially decaying weights applied to the history of all rewards. This EMA approach provides a smooth and stable estimate of the expected reward with minimal computational overhead. With the reward R_t from DRM and the baseline B_t from ABS, the final advantage value used for updating the policy is listed as follows:

$$A_t = R_t - B_t \tag{9}$$

This advantage value A_t is then used directly within the standard policy gradient loss function. The objective is to adjust the policy parameters θ to encourage actions that yield a positive advantage and discourage those that result in a negative advantage. The policy gradient loss is formulated as:

$$\mathcal{L}_{PG} = -\mathbb{E}_{t \sim \pi_\theta}[A_t \cdot \log \pi_\theta(a_t|s_t)] \tag{10}$$

where A_t is the advantage value and $\pi_\theta(a_t|s_t)$ is the policy's probability of taking action a_t given state s_t. By using the advantage A_t calculated via our DRM and ABS, we ensure a stable and efficient prompt optimization for the policy learning.

Overall Training Objective

Our BioDPP is trained in an end-to-end fashion via optimizing an objective function, which integrates the reinforcement learning signal with traditional supervised and knowledge distillation losses. The total loss $\mathcal{L}_{total}$ is a weighted sum of four components: standard cross-entropy loss $\mathcal{L}_{CE}$ for the primary classification task, two knowledge distillation losses, Mean Squared Error $\mathcal{L}_{MSE}$ and KL-Divergence $\mathcal{L}_{KL}$, aligning the student's prompt embeddings and feature distributions with those of the teacher model as well as the policy gradient loss $\mathcal{L}_{PG}$ derived from our DRM and ABS. The final objective is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{MSE} + \lambda_2 \mathcal{L}_{KL} + \lambda_3 \mathcal{L}_{PG} \tag{11}$$

where λ_1, λ_2, and λ_3 are hyperparameters that balance the contribution of each component.

Experiments

We evaluate the effectiveness of the proposed BioDPP on a comprehensive set of biomedical imaging benchmarks, where we use multiple evaluation protocols designed to test

	CoOp			CoCoOp			KgCoOp			ProGrad			BiomedCoOp			BioDPP (Ours)		
Dataset	**Base**	**Novel**	**HM**	**Base**	**Novel**	**HM**	**Base**	**Novel**	**HM**	**Base**	**Novel**	**HM**	**Base**	**Novel**	**HM**	**Base**	**Novel**	**HM**
BTMRI	82.25	94.51	87.95	77.88	94.86	85.54	78.04	95.05	85.71	81.33	94.18	87.28	82.62	95.82	88.73	**86.00**	**97.16**	**91.24**
CHMNIST	89.41	35.11	50.42	87.77	42.51	57.28	75.49	38.70	51.17	82.31	44.64	57.89	89.27	43.31	58.32	**90.03**	**45.61**	**60.55**
COVID-QU-Ex	75.92	90.06	82.39	77.27	87.61	82.12	75.42	89.61	81.90	75.02	90.33	81.97	76.91	90.13	82.99	**80.10**	**90.58**	**85.02**
CTKIDNEY	82.26	67.92	74.41	81.96	56.56	66.93	81.67	58.45	68.14	81.33	68.52	74.38	86.32	**78.18**	82.05	**88.68**	76.83	**82.33**
DermaMNIST	48.06	59.41	53.14	42.88	60.66	50.24	36.41	47.33	41.16	35.84	63.72	45.88	55.50	66.10	60.34	**62.86**	**78.53**	**69.83**
KneeXray	38.25	47.69	42.45	34.08	63.14	44.27	37.94	61.19	46.84	40.47	58.64	47.89	45.18	**80.78**	57.95	**49.64**	75.43	**59.88**
Kvasir	86.22	58.06	69.40	85.94	53.95	66.29	81.56	59.00	68.47	82.83	60.45	69.90	86.95	57.22	69.02	**87.83**	**67.00**	**76.01**
LC25000	90.12	87.57	88.83	88.33	95.02	91.55	88.13	86.43	87.27	90.81	84.03	87.29	93.99	96.41	95.18	**97.07**	**99.80**	**98.42**
RETINA	70.98	56.90	63.16	66.88	**65.56**	66.21	60.77	54.91	57.69	70.09	58.11	63.51	68.88	62.57	65.57	**71.14**	64.04	**67.41**
Average	73.72	66.36	68.02	71.44	68.87	67.82	68.38	65.63	65.37	71.11	69.18	68.44	76.18	74.50	73.35	**79.26**	**77.22**	**76.75**

Table 2: Comparison of our BioDPP and state-of-the-art prompt learning methods in terms of accuracy (%) on generalization from base classes to novel classes. HM denotes the harmonic mean of classification accuracy on base and novel classes.

accuracy and generalization capabilities within and across various few-shot image classification tasks.

Experimental Setup

To comprehensively evaluate the efficacy of our BioDPP, we conduct experiments under two rigorous evaluation protocols designed to assess accuracy and generalization capabilities in challenging biomedical visual tasks. The compared baselines include *adapter-based approaches* CLIP-Adapter (Gao et al. 2024), Tip-Adapter-F (Zhang et al. 2022)), *linear probing methods* Standard LP (Radford et al. 2021), LP++ (Huang et al. 2024), and *prompt learning-based methods* CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a), KgCoOp (Yao, Zhang, and Xu 2023), ProGrad (Zhu et al. 2023), as well as BiomedCoOp (Koleilat et al. 2025).

Few-shot Learning. As a common scenario in clinical applications, we conduct few-shot classification experiments under limited data supervision. Following standard practice, we train all models with varying numbers of labeled examples per class, specifically for $K = 1, 2, 4, 8$, and 16 shots. This setup is critical for evaluating the model's ability to learn effectively from sparse data.

Base-to-Novel Class Generalization. To evaluate the generalization ability of models from seen to unseen classes, we adopt the base-to-novel generalization benchmark. For each dataset, all classes are split into two disjoint sets base (seen) and novel (unseen). Models are trained exclusively on a 16-shot setup from the base classes and are subsequently evaluated on the test sets of both base and novel classes. The Harmonic Mean (HM) of the base and novel accuracies is reported as the metric to assess the generalization performance.

Datasets. Experiments are conducted on 10 medical imaging datasets spanning 9 organs and 8 imaging modalities, including BTMRI (Nickparvar 2021), CHMNIST (Kather et al. 2016), LC25000 (Borkowski et al. 2019), COVID-QU-Ex (Tahir et al. 2021), KneeXray (Chen 2018), CTKidney (Islam et al. 2022), DermaMNIST (Tschandl, Rosendahl, and Kittler 2018; Codella et al. 2019), Kvasir (Pogorelov et al. 2017), RETINA (Köhler et al. 2013; Porwal et al. 2018) and BUSI (Al-Dhabyani et al. 2020). This diverse benchmarks ensures a comprehensive evaluation of model performance across varied biomedical imaging settings.

Implementation Details. We use the BiomedCLIP model (Zhang et al. 2023) with a ViT-B/16 backbone for all experiments, and results are averaged over three independent runs with different random seeds. In training, we use the SGD optimizer with a learning rate of 0.0025 and a batch size of 4. The training is set to 100 epochs for few-shot and 50 epochs for base-to-novel benchmarking. The learnable context is initialized with the embedding of "a photo of a", and we use an ensemble of 50 prompts generated by ChatGPT-4o (Achiam et al. 2023). Optimal values for the hyperparameters ω, λ_1, λ_2, and λ_3 are set to 0.6, 0.75, 0.75 and 0.25. All experiments are conducted on RTX 4090 GPUs.

Few-shot Evaluation

As shown in Table 1, BioDPP consistently outperforms a wide range of state-of-the-art methods across all few-shot scenarios (K=1, 2, 4, 8 and 16). In the most challenging 1-shot setting, our method achieves a leading accuracy of 57.10%, surpassing strong methods like Tip-Adapter-F (55.65%) and BiomedCoOp (55.05%). In the 4-shot setting, where BioDPP achieves 66.67%, further extending its advantage over other methods. This competitive advantage is maintained as the number of examples increases, culminating in a top performance of 73.61% in the 16-shot scenario, showing a notable margin over the second-best performing method BiomedCoOp (72.70%). The reason is that our dynamic prompt policy enables the agent to generate context-aware prompts for each image, capturing case-specific cues from few examples. Moreover, our dual-reward mechanism provides a richer training signal by combining feedback on classification decision and prompt semantic representation, enabling efficient prompt optimization under limited data.

Base-to-Novel Generalization

In the base-to-novel generalization task, BioDPP demonstrates superior performance shown in Table 2. It achieves the highest average Harmonic Mean (HM) of 76.75%, significantly surpassing the baselines like CoOp (68.02%) and BiomedCoOp (73.35%), indicating a better balance between retaining base knowledge and adapting to novel classes. This strong generalization ability is attributes to two key designs. First, the semantic representation reward ($\mathcal{R}_{rep}$) explicitly pushes the embeddings of different classes apart in the feature space. It encourages generalization learning, rather than memorizing only the features specific to base classes. Second, our adaptive baseline stabilization module ensures stable policy refinement, and reduces gradient variance to prevent overfitting on base classes, thereby allowing the policy to generalize effectively.

Ablation Study

Effect of Different Components. We conduct an ablation study to validate the effectiveness of each component in BioDPP shown in Table 3. The analysis confirms the effectiveness and synergy of $\mathcal{R}_{cls}$, $\mathcal{R}_{rep}$, and ABS. For base-to-novel generalization, a baseline using only $\mathcal{R}_{cls}$ achieves a 64.41% HM. The most significant gain comes from incorporating $\mathcal{R}_{rep}$, which boosts the HM to 67.00%, confirming its effectiveness in learning generalizable features. The full model with all components reaches the peak performance of 67.41% HM. A similar trend is observed in the few-shot scenarios, where our full model consistently outperforms all other variants. This is especially evident in the challenging 1-shot setting (37.15%), and the advantage is maintained in different shots, where the best accuracy is up to 66.32% at 16-shots.

Components			Base-to-Novel			Few-shot				
$\mathcal{R}_{cls}$	$\mathcal{R}_{rep}$	ABS	Base	Novel	HM	1	2	4	8	16
✓			67.14	61.89	64.41	34.20	38.22	48.13	55.57	64.43
✓	✓		70.77	63.62	67.00	34.99	38.90	49.08	55.81	65.56
✓		✓	68.47	63.15	65.70	34.17	38.46	48.08	55.68	65.62
	✓	✓	67.53	63.57	65.49	35.12	39.01	49.13	57.78	64.53
✓	✓	✓	**71.14**	**64.04**	**67.41**	**37.15**	**41.56**	**51.03**	**58.84**	**66.32**

Table 3: Accuracy comparisons of different component combinations on the Retina dataset in the Base-to-Novel and Few-shot tasks.

Effect of Different Backbone Networks. We evaluate different CLIP-based backbones, and the results are shown in Fig. 2. The BiomedCLIP (ViT-B/16)(Zhang et al. 2023) backbone performs best across all configurations, achieving the highest accuracy from zero-shot (59.8%) to 16-shot scenarios (69.8%). Thus, we adopt BiomedCLIP (ViT-B/16) as the backbone in all experiments.

Effect of Context Length. We further analyze the effect of context length on the BTMRI dataset. As shown in Fig. 3(a), a shorter length is superior, with an optimal length of 4 achieving the highest HM of 89.21%.

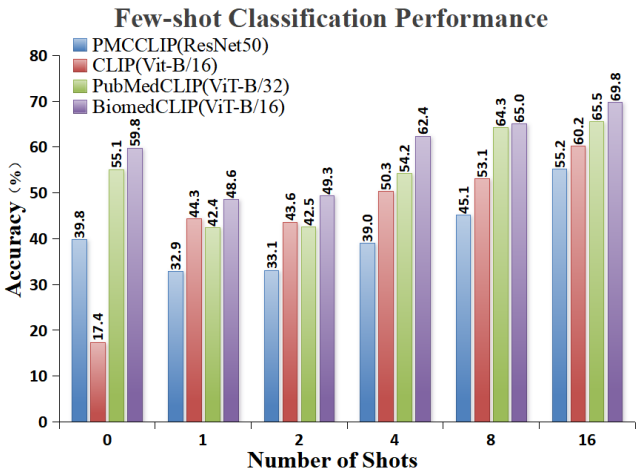

Figure 2: Classification accuracy (%) of different CLIP-based backbone models in BioDPP on the BUSI dataset.

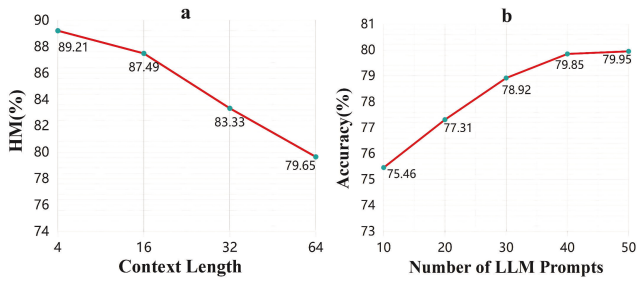

Figure 3: (a) The effect of context vector length on classification accuracy (%) in the Base-to-Novel generalization task. (b) The effect of different numbers of LLM prompts on classification accuracy (%) in 8-shot settings.

Effect of Number of LLM Prompts. Fig. 3(b) shows the accuracy regarding different numbers of LLM generated prompts in 8-shot settings. The accuracy is up to 79.95% with 50 prompts, and reachs a saturating status beyond this point. Therefore, the number of prompts is set to 50 to balance performance and efficiency.

Conclusion

In this study, we propose a dynamic prompt policy learning method that enables efficient adaptation of VLMs for accurate and generalizable few-shot classification across diverse biomedical datasets. Our approach shifts the paradigm from a single static prompt to learning a dynamic policy, reframing prompt optimization as a reinforcement learning problem. We introduce a dual-reward mechanism that evaluates classification and prompt semantic representation to provide multi-dimensional feedback for context-aware prompt optimization. Additionally, an adaptive baseline stabilization is devised to ensure stable policy refinement. Extensive experiments show BioDPP significantly outperforms SOTA methods in few-shot and base-to-novel tasks. Future work could extend this approach to domain-specific adaptation.

Acknowledgments

This paper is supported by the National Key Research and Development Program of China (2023YFC3604601), and the Key Research and Development Program of Hunan Province (2025JK2118). This work is also supported in part by the National Nature Science Foundation of China (No. 62402532), in part by the Hunan Provincial Natural Science Foundation of China (No. 2024JJ6526), in part by the Hunan Provincial degree and Postgraduate Teaching Reform (No. 2025JGYB050), and in part by the Center for Computational Biology and Bioinformatics, Furong Laboratory and Bioinformatics Center, Xiangya Hospital, Central South University.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; and Fahmy, A. 2020. Dataset of breast ultrasound images. *Data in Brief*, 28: 104863.

An, B.; Zhu, S.; Panaitescu-Liess, M.-A.; Mummadi, C. K.; and Huang, F. 2023. More context, less distraction: Improving zero-shot inference of clip by inferring and describing spurious features. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*.

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 preprint arXiv:1912.12142*.

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

Chen, W.; Mees, O.; Kumar, A.; and Levine, S. 2023. Vision-Language Models Provide Promptable Representations for Reinforcement Learning. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M. E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; Kittler, H.; and Halpern, A. 2019. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC).

Deng, M.; Wang, J.; Hsieh, C.-P.; Wang, Y.; Guo, H.; Shu, T.; Song, M.; Xing, E. P.; and Hu, Z. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*.

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. *International Journal of Computer Vision*, 132: 581–595.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23773–23782.

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*, 12: 11440.

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

Jin, W.; Cheng, Y.; Shen, Y.; Chen, W.; and Ren, X. 2021. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*.

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

Khattak, M. U.; Naeem, M. F.; Naseer, M.; Van Gool, L.; and Tombari, F. 2025. Learning to prompt with text only supervision for vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4230–4238.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023a. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19113–19122.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, 15190–15200.

Köhler, 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 *Proceedings of the 26th IEEE international symposium on computer-based medical systems*, 95–100. IEEE.

Koleilat, T.; Asgariandehkordi, H.; Rivaz, H.; and Xiao, Y. 2025. Biomedcoop: Learning to prompt for biomedical vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14766–14776.

Lu, M. Y.; Chen, B.; Williamson, D. F.; Chen, R. J.; Liang, I.; Ding, T.; Jaume, G.; Odintsov, I.; Le, L. P.; Gerber, G.; et al. 2024. A visual-language foundation model for computational pathology. *Nature Medicine*, 30: 863–874.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5206–5215.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15: 654.

Nickparvar, M. 2021. Brain Tumor MRI Dataset.

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 *Proceedings of the 8th ACM on Multimedia Systems Conference*, 164–169.

Porwal, P.; Pachade, S.; Kamble, R.; Kokare, M.; Deshmukh, G.; Sahasrabuddhe, V.; and Meriaudeau, F. 2018. Indian Diabetic Retinopathy Image Dataset (IDRiD).

Pratt, S.; Covert, I.; Liu, R.; and Farhadi, A. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15691–15701.

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 *International conference on machine learning*, 8748–8763. PmLR.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18082–18091.

Rocamonde, J.; Montesinos, V.; Nava, E.; Perez, E.; and Lindner, D. 2024. Vision-Language Models are Zero-Shot Reward Models for Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

Su, C.; Zheng, H.; Peng, D.; and Wang, X. 2025. DiCA: Disambiguated Contrastive Alignment for Cross-Modal Retrieval with Partial Labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 20610–20618.

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*, 139: 105002.

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*, 5: 1–9.

van den Oord, A.; Li, Y.; and Vinyals, O. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Wang, X.; Hu, P.; Liu, P.; and Peng, D. 2020a. Deep semisupervised class-and correlation-collapsed cross-view learning. *IEEE transactions on cybernetics*, 52(3): 1588–1601.

Wang, X.; Peng, D.; Yan, M.; and Hu, P. 2023. Correspondence-free domain alignment for unsupervised cross-domain image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10200–10208.

Wang, Y.; Ou, X.; Liang, J.; and Sun, Z. 2020b. Deep semantic reconstruction hashing for similarity retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(1): 387–400.

Wang, Y.; and Peng, Y. 2021. MARS: Learning modality-agnostic representation for scalable cross-media retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4765–4777.

Wang, Y.; Wu, Y.; Dai, Z.; Tian, C.; Long, J.; and Chen, J. 2025. Noisy Correspondence Rectification via Asymmetric Similarity Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21384–21392.

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

Yin, Z.; Feng, Y.; Yan, M.; Song, X.; Peng, D.; and Wang, X. 2025. RoDA: Robust Domain Alignment for Cross-Domain Retrieval Against Label Noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9535–9543.

Zhan, Y.; Zhu, Y.; Zheng, S.; Zhao, H.; Yang, F.; Tang, M.; and Wang, J. 2025. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *arXiv preprint arXiv:2503.18013*.

Zhang, K.; Zhou, R.; Adhikarla, E.; Yan, Z.; Liu, Y.; Yu, J.; Liu, Z.; Chen, X.; Davison, B. D.; Ren, H.; et al. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 1–13.

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, 493–510. Springer.

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 preprint arXiv:2303.00915*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2337–2348.

Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; and Zhang, H. 2023. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15659–15669.