BiomedCoOp: Learning to Prompt for Biomedical Vision-Language Models

Taha Koleilat* Hojat Asgariandehkordi Hassan Rivaz Yiming Xiao
Concordia University, Montreal, Canada
{taha.koleilat, hojat.asgariandehkordi, hassan.rivaz, yiming.xiao}@concordia.ca

Abstract

Recent advancements in vision-language models (VLMs), such as CLIP, have demonstrated substantial success in self-supervised representation learning for vision tasks. However, effectively adapting VLMs to downstream applications remains challenging, as their accuracy often depends on time-intensive and expertise-demanding prompt engineering, while full model fine-tuning is costly. This is particularly true for biomedical images, which, unlike natural images, typically suffer from limited annotated datasets, unintuitive image contrasts, and nuanced visual features. Recent prompt learning techniques, such as Context Optimization (CoOp) intend to tackle these issues, but still fall short in generalizability. Meanwhile, explorations in prompt learning for biomedical image analysis are still highly limited. In this work, we propose BiomedCoOp, a novel prompt learning framework that enables efficient adaptation of BiomedCLIP for accurate and highly generalizable few-shot biomedical image classification. Our approach achieves effective prompt context learning by leveraging semantic consistency with average prompt ensembles from Large Language Models (LLMs) and knowledge distillation with a statistics-based prompt selection strategy. We conducted comprehensive validation of our proposed framework on 11 medical datasets across 9 modalities and 10 organs against existing state-of-the-art methods, demonstrating significant improvements in both accuracy and generalizability. The code is publicly available at https://github.com/HealthX-Lab/BiomedCoOp.

1. Introduction

The latest breakthroughs in vision-language models (VLMs) have opened new possibilities for leveraging multi-modal data in diverse applications. Unlike traditional supervised learning that focuses on closed-set visual concepts, models like Contrastive Language-Image Pre-training (CLIP) [37], which align visual and textual information through contrastive pre-training, allow the exploration of open-set visual concepts, thanks to the adoption of natural language supervision. However, the success of these models often relies heavily on the quality of the textual prompts that guide their predictions while full-model fine-tuning for large-scale VLMs is impractical. To mitigate these, prompt learning that optimizes textual prompts in vision-language models [25, 50, 51] has emerged as one of the critical techniques to enhance performance without the need for extensive fine-tuning. Notably, the pioneering work of Context Optimization (CoOp) [51] introduced this approach for CLIP by treating text prompts as learnable context vectors and preserving the pre-trained model weights. Meanwhile, other approaches [16, 19, 47] focus on lightweight few-shot adaptation through Adapters [18] and Linear Probes [37] to offer parameter-efficient solutions for model adaptation in downstream tasks.

Different from natural images, biomedical images include a wide range of contrasts and modalities, depending on the image acquisition devices and parameters. These images, such as MRI and ultrasound, often have unique visual appearances that can be more difficult to interpret than typical photographs. In addition, image features (e.g., color, texture, shape, and anatomical context) that are related to physiological and pathological changes are more nuanced and complex to describe, and can differ between image modalities. Finally, due to privacy concerns and the high requirement for clinical expertise, large datasets of well-annotated biomedical images are scarce for developing clinical deep learning models. While VLMs and the associated prompt learning techniques have shown success across natural image datasets and benchmarks, their application in the biomedical imaging domain (e.g., diagnosis), which has distinct challenges, remains largely under-explored.

Due to the unique domain knowledge of biomedical images, the backbone vision-language model for prompt learning may require tailored pre-training for the best outcome. Biomed-specific VLMs, such as BiomedCLIP [48]—pretrained on 15 million biomedical image-text pairs from internet resources—are better suited for biomedical tasks

*Corresponding author

[49]. Although few recent works [3, 6, 15] have investigated prompt learning for biomedical image classification by using natural-image-trained CLIP backbones, their explorations remain in a narrow range of tasks with larger datasets (e.g., chest X-ray and dermatology), which may not be common for other clinical tasks. Notably, these approaches either employ the full training dataset [3, 15] or domain-controlled prompt learning via an additional foundation model, i.e., MedSAM [33], reducing their computational efficiency. In addition, further improvements in accuracy and generalizability to unseen classes are still needed over the existing prompt learning approaches. Therefore, it is highly instrumental and desirable to explore more efficient and robust novel prompt learning techniques based on biomed-specific VLMs (e.g., BiomedCLIP) and conduct large-scale validation on diverse benchmarking datasets of biomedical images.

In this work, we present an innovative prompt-learning framework based on CoOp [51], called BiomedCoOp, to facilitate efficient adaptation of CLIP-like VLMs, such as BiomedCLIP, in few-shot biomedical image classification for differential diagnosis. Instead of full model fine-tuning, our approach exclusively focuses on effective strategies for textual prompt optimization with the help of large language models (LLMs), which not only reduce computational overhead, but also preserve the pre-trained model's foundational knowledge. Specifically, with high disparity across imaging modalities (e.g., ultrasound vs. MRI), we hypothesize that a few-shot learning strategy will benefit the application while addressing data limitations in the biomedical domain. Furthermore, LLMs can help mitigate limitations in human-designed prompt templates in context learning. Our four key contributions include:

1. We introduce a novel prompt-learning approach that enhances context vector learning by enforcing semantic consistency through prompt ensembles derived from large language models (i.e., GPT-4) and an effective knowledge distillation strategy.
2. To address the challenge of outlier prompts from LLMs during context learning, which can lead to over-specialization and hinder generalization, we employ a statistics-based pruning strategy to mitigate the risk of "forgetting" essential biomedical patterns while maintaining sensitivity to diverse disease presentations.
3. Our study adopts BiomedCLIP [48] for prompt learning for the first time, and we demonstrate the benefits against general knowledge CLIP in downstream clinical tasks.
4. We conduct a comprehensive evaluation of our proposed method against existing CLIP prompt learning techniques using 11 diverse biomedical image classification datasets across 9 modalities and 10 organs in few-shot and base-to-novel generalization benchmarks. Our results highlight BiomedCoOp's superior generalizability and robustness across a wide range of medical conditions and imaging modalities.

2. Related Work

2.1. Vision-Language Models

Vision-language models like CLIP [37] and ALIGN [21] have transformed multi-modal learning with self-supervised visual and textual representations in a common feature space, enabling remarkable performance in various tasks, including zero-shot classification and cross-modal retrieval. Recent extensions in the biomedical domain include BioViL [4], PubMedCLIP [14], and BiomedCLIP [48], which adapt VLMs to biomedical data using millions of biomedical image-text pairs. However, while effective, these models still require additional task-specific adaptation to capture the fine-grained, disease-specific nuances crucial in clinical applications for better performance [42, 49]. Such limitations highlight the need for methods that conduct further domain-specific adaptations for target clinical tasks.

2.2. Prompt Learning

Prompt learning has emerged as a powerful alternative to traditional model fine-tuning for VLMs, with methods like CoOp [51] and CoCoOp [50] treating text prompts as the primary learnable component to steer VLMs toward specific tasks. In biomedical applications, prompt learning has shown preliminary potential for few-shot adaptation without altering pre-trained weights [13]. Recent methods in the natural vision domain like MaPLe [24] adapt both the vision and language components of CLIP while PromptSRC [26] applies self-regulation techniques to improve prompt generalization without sacrificing prior knowledge. Complementing these approaches, KgCoOp [43] and ProGrad [53] refine textual prompts using knowledge and gradient guidance, respectively, to enhance model generalizability. Additionally, ProText [27] stands out with deep prompt learning solely from text data derived from LLMs, enabling cross-dataset and cross-class transfer without requiring labeled images. However, these methods may not be well suited for biomedical applications due to the aforementioned challenges, especially cross-modal differences in anatomical and pathological feature descriptions. Recently, biomed-specific approaches, including ViP [15] and XCoOp [3], have enhanced the adaptability of VLMs in clinical tasks by integrating disease-specific terminology and contextual tokens. Domain-Controlled Prompt Learning (DCPL) [6] further extends this adaptability by integrating domain-specific biases for both vision and language branches, particularly useful in specialized fields such as remote sensing and medical imaging, where natural-domain prompts fall short. Despite more tailored solutions for biomedical tasks, these methods often require heavier adaptation of both vision and

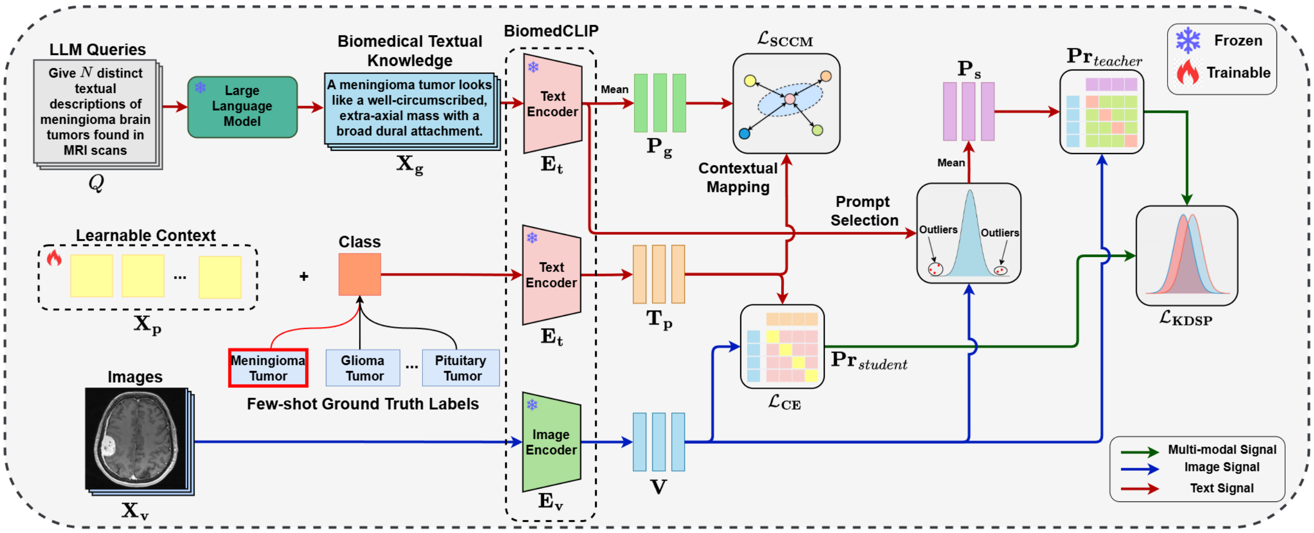

Figure 1. Overview of the BiomedCoOp framework, which combines LLM queries, learnable context tokens, and BiomedCLIP to generate multi-modal representations for biomedical tasks. The method integrates text and image features using prompt ensembling strategies, minimizes cross-entropy and semantic differences, and aligns teacher-student logits, enabling effective few-shot learning for novel biomedical categories.

language components and clinical expert intervention, thus potentially limiting their performance, ease of implementation, and generalizability to wider tasks. As suggested by Khattak et al. [27], robust mapping between learnable context and knowledge from LLMs can benefit the performance and adaptability of VLMs. With human-level performance and stability upgrades of recent LLMs like GPT-4 in diagnostic reports [17, 30], integration of the latest LLMs, in addition to biomed-specific VLMs (e.g., BiomedCLIP) may offer an alternative avenue to more conveniently incorporate biomedical domain expertise in VLMs for a data-efficient prompt learning solution, suitable for diverse clinical needs. Yet, this remains to be explored, and we intend to take advantage of these advances in our methods.

2.3. Few-shot Adaptation of VLMs

Few-shot adaptation techniques for vision-language models enable task specialization with minimal labeled data to balance generalization and domain specificity. Aside from few-shot-based prompt learning, adapter-based methods like CLIP-Adapter [16], and Tip-Adapter [47] incorporate lightweight modules that adjust visual features while preserving zero-shot capabilities. Specifically, CLIP-Adapter modifies visual embeddings via a compact MLP. Tip-Adapter uses a similarity-based mechanism that blends visual features from the support set directly into the model's predictor to enhance accuracy with minimal data, but requires careful tuning. Enhanced linear probing methods, such as LP++ [19], extend adaptation by mixing visual and text features and using data-informed implicit learning rates to achieve competitive few-shot performance without extensive hyperparameter optimization. Finally, CLAP [38] further constrains adaptation to stay close to original zero-shot prototypes by using adaptive penalties to prevent overfitting. These categories of methods refine visual embeddings typically at the final layers of VLMs, focusing on adjusting model features. In comparison, prompt learning approaches that target optimizing textual prompt inputs may be more advantageous in computational efficiency and adaptation to unseen classes, particularly in biomedical imaging domains.

3. Methodology

An overview of the proposed BiomedCoOp framework is illustrated in Fig. 1. By leveraging the BiomedCLIP backbone to encode the rich image and text features, our method proposes two major components to enable effective prompt context learning jointly. Specifically, while the *Semantic Consistency by Contextual Mapping* (SCCM) component aligns text embeddings with general biomedical knowledge via minimizing their distance to class-specific prompts generated by an LLM, we design the *Knowledge Distillation with Selective Prompting* (KDSP) component to refine context mapping through statistics-based prompt selection. The unified learning objective integrates cross-entropy loss, MSE loss of *SCCM*, and Kullback-Leibler (KL)-divergence loss of *KDSP* to ensure accurate and robust model representations.

3.1. Contrastive Language-Image Pre-training

CLIP utilizes a vision encoder $\boldsymbol{E_v}$ and a text encoder $\boldsymbol{E_t}$ to take batches of pre-processed images $\boldsymbol{X_v} \in \mathbb{R}^{B\times 3\times H\times W}$

and texts $\boldsymbol{X_t} \in \mathbb{R}^{B\times L}$ as inputs. Here, B is the number of image/text pairs in a single batch, H and W are the height and width of the images, respectively, and L is the maximum sequence length of the tokenized text input. Specifically, the vision and text encoder networks produce the corresponding feature vectors $\mathbf{V} = \boldsymbol{E_v}(\boldsymbol{X_v})$ and $\mathbf{T} = \boldsymbol{E_t}(\boldsymbol{X_t})$, where $\mathbf{V} \in \mathbb{R}^{B\times D}$,$\mathbf{T} \in \mathbb{R}^{B\times C\times D}$, C is the number of classes and D is the embedding dimension. Finally, the multi-modal feature vectors are aligned in the latent space based on their similarity scores through a contrastive learning objective. The pre-trained CLIP model with rich multi-modal representation has the capacity for zero-shot classification. Given C possible classes, a new image can be classified by computing its similarity to a set of C different text prompts as "`a photo of a [CLASS]`", where "`a photo of a`" represents the pre-defined context of the textual template while `[CLASS]` is the class name token. Notably, in CoOp, the context (e.g., "`a photo of a`") is learnable and optimized according to a categorical objective. Specifically, the zero-shot inference is made by comparing each text feature $\boldsymbol{T}^{(i)}$ of the different text prompts with the image embeddings $\boldsymbol{V}$ to get a prediction probability:

$$p(\boldsymbol{Y} = i|\boldsymbol{V}, \boldsymbol{T}^{(i)}) = \frac{\exp(\cos(\boldsymbol{T}^{(i)}, \boldsymbol{V})/\tau)}{\sum_{j=1}^{C}\exp(\cos(\boldsymbol{T}^{(j)}, \boldsymbol{V})/\tau)}, \quad (1)$$

where τ denotes the learnable temperature parameter and $\cos(\cdot,\cdot)$ denotes cosine similarity. The class of the image is determined by taking:

$$\arg\max_{\boldsymbol{i}} p(\boldsymbol{Y} = i|\boldsymbol{V}, \boldsymbol{T}^{(i)}) \quad (2)$$

3.2. LLM Prompt Ensembling

While prompt ensembling is shown to benefit prompt learning by bringing diverse text representations [46], human-designed prompt templates that were previously adopted in natural vision tasks [32] can face obstacles for biomedical domains in sourcing relevant clinical knowledge and sufficient description diversity. Notably, recent studies [17, 30] on GPT-4 [1] have validated its performance in tasks related to clinical case reports. Thus, we decided to leverage GPT-4 for LLM-generated prompt ensembling with diverse textual descriptions of class-specific lesions and abnormalities in diagnostic scans. This approach ensures that the contextual structure of the learned prompts in BiomedCoOp reflects the required domain knowledge to help effectively capture essential semantic features. Specifically, for C different classes in a dataset, we generate text prompts $\boldsymbol{X_g} \in \mathbb{R}^{N\times C\times L}$ comprising N different outputs for each class from text query Q: "`Give` N `textual descriptions of visual discriminative features for distinct medical cases of [CLASS] found in [MODALITY].`" Unique for biomedical images, we specifically mention the imaging modality in Q as certain classes might overlap in different modalities. $\boldsymbol{X_g}$ from the LLM are then encoded to $\boldsymbol{T_g} = \boldsymbol{E_t}(\boldsymbol{X_g}) \in \mathbb{R}^{N\times C\times D}$. For the SCCM component, all N text embeddings for each class are ensembled by taking the mean to get $\boldsymbol{P_g} \in \mathbb{R}^{C\times D}$:

$$\boldsymbol{P_g}^{(c)} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{T}_{\mathbf{g,i}}^{(c)} \quad (3)$$

3.3. Selective Prompting via Outlier Exclusion

In the KDSP component, we further refine the learned context by also considering the probability distribution of the LLM-generated prompts. As overly specific prompts can risk overfitting the model on particular disease states and low-scoring prompts that do not align well with relevant biomedical features can hurt accuracy, we propose selective prompt ensembling with outlier pruning. By removing these outliers, the distribution of prompts is refined, ensuring that the selected textual prompt distribution reflects broader biomedical insights, helping the model avoid "forgetting" key general knowledge while preserving the versatility of the BiomedCLIP model. This enables the model to remain sensitive to both typical and atypical features across diverse disease presentations in biomedical images.

Given a batch of B images, each representing a unique case for a disease class, we first encode these images using the vision encoder $\boldsymbol{E_v}$ to obtain image features $\boldsymbol{V}$. For each image embedding in $\boldsymbol{V}$, we compute its similarity with the corresponding prompt embedding in $\boldsymbol{T_g}$. The similarity measure, typically cosine similarity, helps identify the prompts most relevant to the specific characteristics of the image. The scores $\boldsymbol{S}$ are calculated for each prompt by taking the mean of the maximum similarity logits across all image embeddings:

$$\boldsymbol{S} = \frac{1}{B}\sum_{i=1}^{B}\max_{j}(\beta \cdot \boldsymbol{T_g}^{(j)} \cdot \boldsymbol{V}^{\intercal}), \quad (4)$$

where β is a scaling factor applied to the logits.

To detect and handle anomalous prompts that deviate from the general distribution, we apply an outlier detection approach using the *Median Absolute Deviation* test statistic. Specifically, we calculate the median M_s of the prompt scores $\boldsymbol{S}$ and the median absolute deviation D:

$$M_s = \text{median}(\boldsymbol{S}) \quad (5)$$

$$D = \text{median}(|\boldsymbol{S} - M_s|) \quad (6)$$

For a given prompt, we compute the modified z-score:

$$z = \frac{\boldsymbol{S} - M_s}{D}. \quad (7)$$

We select only the N_s prompts that are associated with the modified z-scores with absolute values that are lower than a selection threshold ζ_s. Following a similar approach to Eq. 3, we obtain an average prompt encoding $\boldsymbol{P_s}$ based on the selections.

3.4. Overall Learning Objective

The overall learning objective of our few-shot prompt learning framework entails a cross-entropy loss for classification accuracy and the context-mapping-related losses from the *Semantic Consistency by Contextual Mapping* and *Knowledge Distillation with Selective Prompting* components. Their details are described below:

First, we aim to learn prompts $\boldsymbol{X_p}$ by optimizing the overall learning objective given K few-shot ground truth examples for each class. The objective includes the cross-entropy loss of the image-text logits, which is defined as:

$$\mathcal{L}_{\texttt{CE}} = -\sum_{i=1}^{C} \boldsymbol{Y}^{(i)} \log p(\boldsymbol{Y} = i | \boldsymbol{V}, \boldsymbol{T_p}^{(i)}), \tag{8}$$

where $\boldsymbol{Y}^{(i)}$ is the ground truth label for class i, and $p(\boldsymbol{Y} = i|\boldsymbol{V}, \boldsymbol{T_p}^{(i)})$ is the predicted probability of class i given images $\boldsymbol{V}$ and encoded learnable text prompt $\boldsymbol{T_p}^{(i)}$.

Second, since the context to be learned is unified among all the classes, we minimize the difference between $\boldsymbol{P_g}$ and $\boldsymbol{T}_p$ in the SCCM component to ensure that the general biomedical knowledge is properly learned:

$$\mathcal{L}_{\texttt{SCCM}} = \sum_{i=1}^{C} ||\boldsymbol{T_p}^{(i)} - \boldsymbol{P_g}^{(i)}||_2^2 \tag{9}$$

Lastly, to align the distribution of the logits from image embeddings with learnable context prompts (student logits) and the logits from image embeddings with selective LLM-generated text embeddings (teacher logits), we minimize the KL divergence between these two distributions in the KDSP component:

$$\mathcal{L}_{\texttt{KDSP}} = \sum_{i} \boldsymbol{Pr}_{\text{teacher}}(i) \log \frac{\boldsymbol{Pr}_{\text{teacher}}(i)}{\boldsymbol{Pr}_{\text{student}}(i)}, \tag{10}$$

where $\boldsymbol{Pr}_{\text{teacher}}(i)$ is the probability distribution of the logits of $\boldsymbol{V}$ and $\boldsymbol{P_s}$, and $\boldsymbol{Pr}_{\text{student}}(i)$ is the probability distribution of the logits of $\boldsymbol{V}$ and $\boldsymbol{T_p}$.

The KL divergence term, $\mathcal{L}_{\texttt{KDSP}}$ restricts the model from drifting toward embeddings that are not representative of the actual medical scans. By minimizing this KL divergence, we guide the model to stay within a meaningful embedding space closely related to the content of the medical scans. This alignment helps ensure that the learned embeddings retain essential information about the biomedical images, preventing the model from diverging into unrelated semantic spaces.

The overall learning objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\texttt{CE}} + \lambda_1 \mathcal{L}_{\texttt{SCCM}} + \lambda_2 \mathcal{L}_{\texttt{KDSP}} \tag{11}$$

where λ_1, and λ_2 are loss-balancing weights.

4. Experiments and Results

4.1. Experimental Setup

We assess the efficacy of our proposed BiomedCoOp framework across a comprehensive set of benchmark biomedical imaging datasets under multiple evaluation protocols designed to test accuracy and generalization within and across various few-shot image classification tasks.

Few-Shot Learning: To assess the model's performance under limited supervision, we conduct few-shot experiments with varying numbers of labeled examples per class (K = 1, 2, 4, 8, and 16 shots). This is critical for evaluating the model's ability to learn effectively from sparse data, a common scenario in biomedical applications, while retaining both task-specific and general domain knowledge.

Base-to-Novel Class Generalization: To assess model generalizability of our technique, each dataset is divided into base and novel classes. The model is trained on the base classes using a 16-shot setup and subsequently evaluated on both base and novel classes. This setup tests the model's ability to generalize to unseen classes within the same dataset, showcasing its potential to recognize novel disease presentations without additional fine-tuning.

Datasets: We conduct our experiments on 11 diverse medical imaging datasets covering 10 different organs and 9 imaging modalities: Computerized Tomography (CTKidney [20]), Dermatoscopy (DermaMNIST [10, 40]), Endoscopy (Kvasir [35]), Fundus Photography (RETINA [31, 36]), Histopathology (LC25000 [5], CHMNIST [22]), Magnetic Resonance Imaging (BTMRI [34]), Optical Coherence Tomography (OCTMNIST [23]), Ultrasound (BUSI [2]), and X-Ray (COVID-QU-Ex [39], KneeXray [7]). This selection includes complex datasets, such as brain MRI and ultrasound, ensuring a thorough assessment of the model's performance across a broad spectrum of biomedical imaging contexts. The detailed data split and tasks for the experiments are included in the *Supplementary Materials*.

Implementation details: We employed BiomedCLIP with a ViT-B/16 backbone, averaging results over three runs. The training was set to 100 epochs for few-shot and 50 epochs for base-to-novel benchmarking. We initialized the learnable context with the embedding vector corresponding to "`a photo of a`" and used 50 LLM prompts, a 0.0025 learning rate, a batch size of 4, and an SGD optimizer across datasets. Optimal values for λ_1, λ_2, and ζ_s were selected

Table 1. **Evaluation against state-of-the-art techniques.** This table presents the average classification accuracy (%) obtained from 11 benchmarks, along with the standard deviation derived from 3 sampled support sets for each dataset. Results show mean±std. The top-performing results are in bold.

Method	$K = 1$	$K = 2$	$K = 4$	$K = 8$	$K = 16$
			Zero-shot Methods		
BiomedCLIP [48]			42.05		
BiomedCLIP [48] + Ensemble			52.27		
BiomedCLIP [48] + Selective Ensemble			53.72		
			CLIP-based Adapter Methods		
CLIP-Adapter [16]	44.66 ± 2.97	43.91 ± 2.48	44.36 ± 1.94	45.42 ± 2.38	46.69 ± 1.71
Tip-Adapter [47]	49.19 ± 4.84	52.36 ± 6.57	57.33 ± 5.07	61.98 ± 5.76	67.15 ± 4.25
Tip-Adapter-F [47]	51.17 ± 8.33	52.74 ± 5.88	61.23 ± 6.22	65.91 ± 3.64	70.91 ± 2.65
			Linear Probing Methods		
Standard LP [37]	47.25 ± 8.65	54.21 ± 7.80	61.00 ± 6.81	65.85 ± 4.89	69.40 ± 2.91
LP++ [19]	47.24 ± 7.68	53.18 ± 7.29	59.02 ± 6.93	63.69 ± 4.68	68.35 ± 3.59
			Prompt Learning Methods		
CoOp [51]	50.16 ± 6.93	54.18 ± 4.31	59.75 ± 3.72	65.84 ± 3.66	69.62 ± 2.83
CoCoOp [50]	48.49 ± 4.39	51.28 ± 5.06	54.69 ± 4.79	61.08 ± 3.49	65.09 ± 2.87
KgCoOp [44]	50.85 ± 5.59	53.18 ± 4.33	57.82 ± 4.50	62.08 ± 2.59	62.84 ± 1.72
ProGrad [52]	51.88 ± 6.39	54.71 ± 4.46	60.42 ± 4.78	65.61 ± 3.02	67.13 ± 3.00
BiomedCoOp (Ours)	**57.03** ± **2.80**	**59.13** ± **3.64**	**63.95** ± **2.42**	**68.32** ± **2.65**	**72.42** ± **1.69**

using the validation set (detailed values are reported in the *Supplementary Materials*). All experiments ran on a single NVIDIA A100 GPU (40GB RAM).

4.2. Few-shot Evaluation

We compared BiomedCoOp against four text-prompt learning methods (CoOp, CoCoOp, ProGrad, KgCoOp), three CLIP-based adapter methods (CLIP-Adapter, Tip-Adapter, Tip-Adapter-F), and two linear probing methods (Standard Linear Probe, LP++). In this study, we focus on text-modal prompt learning techniques that solely optimize text prompts, excluding methods that learn both text and image prompts (i.e. MaPLe, DCPL). Additionally, we limit our comparison to shallow prompt learning techniques rather than deep prompt approaches that require additional parameters for a fair comparison. Additionally, zero-shot and LLM-prompted zero-shot BiomedCLIP configurations were tested. All models used BiomedCLIP as the backbone and were tuned to their optimal settings. As shown in Table 1, BiomedCoOp consistently outperformed these baselines, especially in low-shot settings, where it surpassed the second-best performing method- ProGrad- by 5.2% and 4.6% in 1- and 2-shot scenarios. This improvement is attributed to BiomedCoOp's use of class-specific, selectively ensembled LLM prompts, enhancing generalizability and sensitivity to diverse biomedical conditions, even with minimal labeled data. BiomedCoOp continued to lead as the number of shots increased, proving its robustness across different data availability scenarios. Its consistent performance across all K-shot settings supports BiomedCoOp's effectiveness in prompt-based biomedical adaptation, ensuring reliable cross-dataset accuracy gains.

4.3. Base-to-Novel Generalization

We evaluated BiomedCoOp's base-to-novel generalization against the state-of-the-art prompt learning methods (CoOp, CoCoOp, ProGrad, and KgCoOp) by measuring accuracy across both base and novel classes, using harmonic mean (HM) for balanced generalization. Note that *the BUSI dataset was excluded due to insufficient class diversity for the experiment*. Results reported in Table 2 show that BiomedCoOp consistently outperformed others, with 5-10% higher accuracy on challenging datasets like CTKidney and Kvasir. This indicates BiomedCoOp's strong capacity to generalize to unseen biomedical categories—an essential feature in medical contexts with evolving disease patterns. Its LLM-augmented prompts further aid in retaining base knowledge while effectively adapting to novel classes, reducing forgetting.

4.4. Ablation Experiments

4.4.1. Effect of Different Components

Table 3 demonstrates the critical contributions from each of BiomedCoOp's components. The baseline BiomedCLIP model, without ensembling or BiomedCoOp components, showed limited accuracy, particularly for novel classes. Adding SCCM alone significantly improved adaptability by embedding contextual medical information, enhancing low-shot performance. KDSP, applied alone, boosted accuracy in novel classes by filtering out low-quality and

Dataset		BiomedCLIP [48]	CoOp [51]	CoCoOp [50]	KgCoOp [43]	ProGrad [53]	BiomedCoOp (Ours)
Average on 10 datasets	Base	47.84	73.85	72.26	68.36	71.67	**76.26**
	Novel	65.42	64.75	67.03	64.08	66.93	**73.92**
	HM	53.81	67.23	67.22	64.61	67.43	**75.07**
BTMRI	Base	40.88	82.25	77.88	78.03	82.13	**82.42**
	Novel	96.18	94.51	94.84	95.05	94.98	**96.84**
	HM	57.37	87.95	85.53	85.69	88.09	**89.05**
COVID-QU-Ex	Base	53.96	75.92	**77.28**	75.42	75.19	75.91
	Novel	89.43	90.07	87.61	89.61	90.34	**91.63**
	HM	67.31	82.39	82.12	81.90	82.07	**83.03**
CTKIDNEY	Base	38.55	82.24	81.96	81.67	83.86	**86.93**
	Novel	52.99	67.92	56.56	58.45	63.01	**78.94**
	HM	44.63	74.40	66.93	68.14	71.96	**82.74**
DermaMNIST	Base	34.95	48.06	42.88	36.41	35.52	**54.86**
	Novel	49.59	59.41	60.66	47.31	63.28	**74.1**
	HM	41.00	53.14	50.24	41.15	45.50	**63.04**
Kvasir	Base	75.00	86.22	85.94	81.56	82.89	**86.50**
	Novel	60.50	58.06	53.95	59.00	60.45	**61.83**
	HM	66.97	69.39	66.29	68.47	69.91	**72.11**
CHMNIST	Base	37.63	**89.41**	87.77	75.45	82.98	88.87
	Novel	40.69	35.11	42.51	38.70	**44.19**	42.73
	HM	39.10	50.42	57.28	51.16	57.67	**57.71**
LC25000	Base	59.73	90.12	88.33	88.13	90.29	**93.77**
	Novel	87.60	87.55	95.02	86.44	85.47	**97.00**
	HM	71.03	88.82	91.55	87.28	87.81	**95.36**
RETINA	Base	45.18	**70.98**	66.88	60.77	68.77	68.46
	Novel	55.28	56.90	65.56	54.91	58.43	**67.72**
	HM	49.72	63.16	66.21	57.69	63.18	**68.09**
KneeXray	Base	35.89	38.28	34.08	37.94	40.88	**44.23**
	Novel	71.90	47.69	63.14	61.19	59.12	**78.35**
	HM	47.88	42.47	44.27	46.84	48.34	**56.54**
OCTMNIST	Base	56.60	75.00	79.6	68.20	74.20	**80.33**
	Novel	50.00	50.23	**50.47**	50.13	50.02	50.07
	HM	53.10	60.17	**61.77**	57.79	59.76	61.69

Table 2. Accuracy comparison (%) on Base-to-novel generalization of BiomedCoOp with SOTA prompt learning methods. HM = harmonic mean of the classification accuracy between base and novel classes.

case-specific prompts. However, using KDSP with a CLIP-only setup hindered generalization, as the model lacked the domain knowledge needed to effectively exclude outliers (gray highlight). The combined use of SCCM and KDSP yielded optimal results (green highlight), balancing generalization and adaptability, especially in few-shot and base-to-novel tasks, affirming both components' essential roles in handling limited-data biomedical scenarios.

4.4.2. Effect of Number of LLM Prompts

Prompt diversity in relation to the number of LLM prompts may affect the quality of context mapping. To investigate this, Table 4 shows the effect of increasing LLM-generated prompts on BiomedCoOp's performance across few-shot settings (K = 0, 1, 2, 4, 8, and 16 shots). At lower shot levels (K = 0 and 1), a higher prompt count noticeably boosts accuracy, improving by 5-6% as prompts increase from 10 to 50. This indicates that prompt diversity is key for generalization with minimal labeled data. In intermediate shot

Components			Base-to-Novel			Few-shot				
BiomedCLIP	SCCM	KDSP	Base	Novel	HM	1	2	4	8	16
✗	✗	✗	71.50	45.06	55.28	41.46	45.69	51.41	54.80	61.86
✗	✓	✗	71.91	43.04	53.85	43.72	44.77	52.76	60.08	64.04
✗	✗	✓	73.21	39.95	51.69	42.14	44.27	53.75	56.67	62.14
✗	✓	✓	72.95	39.04	50.86	42.11	44.52	53.48	56.36	62.37
✓	✗	✗	73.85	64.75	67.23	50.16	54.18	59.75	65.84	69.62
✓	✓	✗	75.21	65.79	70.19	51.62	54.99	61.43	65.93	69.99
✓	✗	✓	75.74	72.91	74.30	56.78	58.76	63.68	67.68	71.79
✓	✓	✓	**76.11**	**73.22**	**74.64**	**57.03**	**59.13**	**63.95**	**68.32**	**72.42**

Table 3. Impact of each component of the proposed BiomedCoOp method on the accuracy (%) of few-shot and Base-to-Novel benchmarks, including BiomedCLIP vs. CLIP, SCCM, and KDSP.

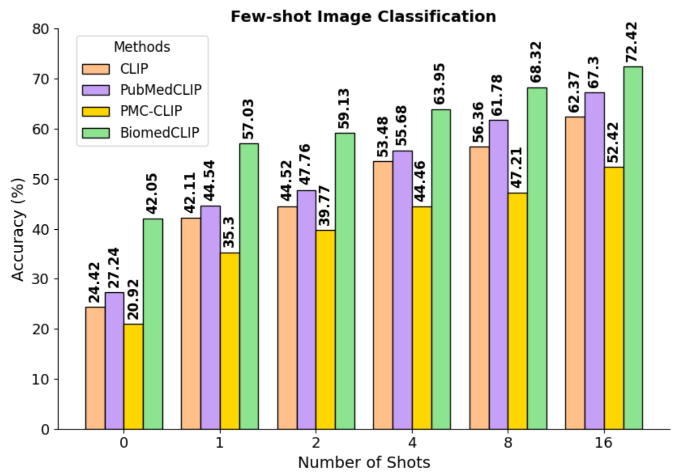

Figure 2. Barplot to compare classification accuracies (%) of different CLIP-based backbone models in BiomedCoOP across various few-shot settings.

Prompts #	$K=0$	$K=1$	$K=2$	$K=4$	$K=8$	$K=16$
10	47.55	52.43	54.61	60.69	64.81	67.66
20	50.51	55.27	57.65	62.85	66.92	70.96
30	51.88	55.91	58.52	63.89	68.05	71.97
40	52.20	56.59	59.05	63.92	68.24	72.20
50	52.27	57.03	59.13	63.95	68.32	72.42

Table 4. BiomedCoOp's classification accuracy (%) across different K-shot settings with varying numbers of LLM prompts.

settings (K = 2 and 4), performance continues to improve with more prompts but at a slower rate, showing diminishing returns as more labeled information becomes available. At higher shot levels (8 and 16), accuracy stabilizes, with only minor gains from additional prompts. As shown in other studies [45], prompt diversity strengthens model performance by providing varied semantic cues that guide the model toward essential biomedical features. Each prompt introduces a unique contextual perspective, enabling the model to build a more robust, flexible understanding of medical concepts. This diversity helps the model focus on shared, critical features, enhancing its ability to recognize subtle variations and generalize effectively, particularly in

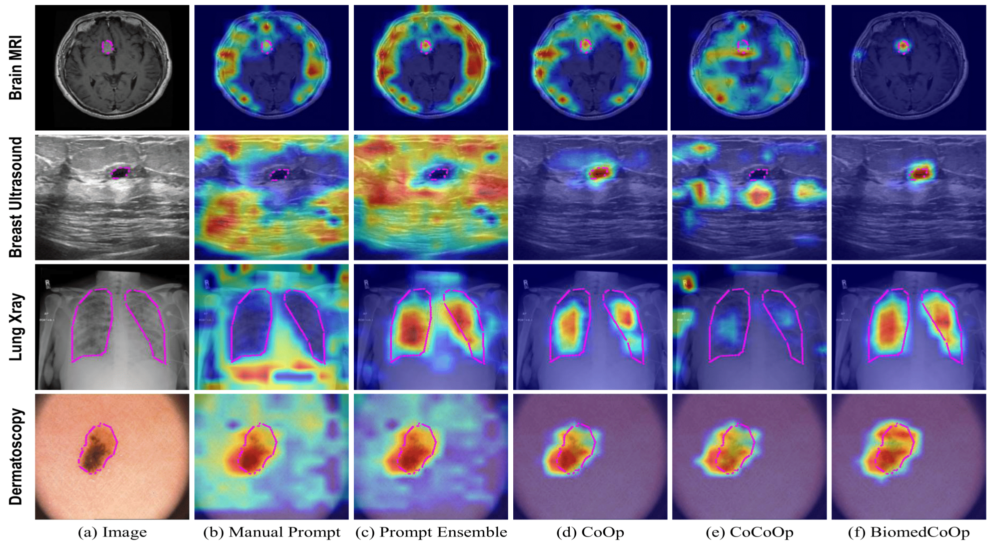

Figure 3. Effect of various text prompt techniques on visual saliency maps. Columns (b)-(f) represent different prompt methods.

low-data scenarios.

4.4.3. Effect of Different CLIP-based Models

To better understand the impact of backbone CLIP-based models on the proposed BiomedCoOp methods, we examined the performance with four CLIP-based models, including CLIP (ViT-B/16), PubMedCLIP (ViT-B/32), PMC-CLIP (RN50), and BiomedCLIP (ViT-B/16). The bar plot in Fig. 2 compares the few-shot classification accuracy of the variants of BiomedCoOp with these CLIP backbones. Notably, BiomedCLIP consistently achieves the highest accuracy across all shot settings, showcasing its ability to capture medical domain-specific features effectively. As the number of shots increases, all models improve, with BiomedCLIP reaching 72.42% accuracy at 16 shots, significantly outperforming others. CLIP and PMC-CLIP show similar results, while PubMedCLIP remains competitive, but trails BiomedCLIP. These results validate our selection of BiomedCLIP as the VLM backbone and highlight the importance of specialized biomedical VLMs for improving few-shot biomedical image classification performance.

4.4.4. Visual Interpretability

In this experiment, gScoreCAM [8] was used to assess how different text prompts influence the visual saliency maps of biomedical images. Each column in Fig. 3 represents a distinct prompt type: the "manual prompt" (column b) uses `a photo of a [CLASS]`, which tends to make the model focus on global features (i.e. scan modality), rather than the region of interest; the "prompt ensemble" (column c) averages multiple LLM-generated prompts, potentially introducing conflicting signals and causing the model to focus on the background; and CoOp, CoCoOp, and BiomedCoOp (columns d, e, f) use optimized, learned contextual prompts with the `[CLASS]` label. Tested on BUSI (breast tumor ultrasound), COVID-QU-Ex (lung X-ray), brain tumor [9], and ISIC [11, 12, 41] (dermatology) datasets, BiomedCoOp (column f) consistently aligns best with ground truth regions across modalities, accurately highlighting clinically relevant areas, especially in complex modalities like MRI and ultrasound, with notable yet smaller improvements in dermatology due to its similarity to natural images, leading to fewer false positives and negatives. This precise localization enhances interpretability, making it particularly valuable for medical applications [28, 29], where explainability is essential. This observation further validates the effectiveness of our proposed method.

5. Conclusion

Our proposed BiomedCoOp framework represents the first large-scale exploration of prompt learning tailored for few-shot adaptation across diverse biomedical datasets. By distilling domain-specific insights from LLMs, BiomedCoOp enriches prompt representations and enhances the model's ability to generalize across a wide range of biomedical imaging contexts. Our approach combines generalizable biomedical knowledge with selective prompt refinement to improve classification accuracy and generalizability. This study highlights the potential of integrating LLM-derived prompts with BiomedCLIP to achieve accurate, data-efficient biomedical diagnosis, serving as a foundational step towards the broader adoption of adaptable VLMs in clinical applications.

Acknowledgment We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

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

[2] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 5

[3] Yequan Bie, Luyang Luo, Zhixuan Chen, and Hao Chen. Xcoop: Explainable prompt learning for computer-aided diagnosis via concept-guided context optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 773–783. Springer, 2024. 2

[4] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022. 2

[5] Andrew A. Borkowski, Marilyn M. Bui, L. Brannon Thomas, Catherine P. Wilson, Lauren A. DeLand, and Stephen M. Mastorides. Lung and colon cancer histopathological image dataset (lc25000), 2019. 5

[6] Qinglong Cao, Zhengqin Xu, Yuntian Chen, Chao Ma, and Xiaokang Yang. Domain-controlled prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 936–944, 2024. 2

[7] Pingjun Chen. Knee osteoarthritis severity grading dataset, 2018. 5

[8] Peijie Chen, Qi Li, Saad Biaz, Trung Bui, and Anh Nguyen. gscorecam: What objects is clip looking at? In *Proceedings of the Asian Conference on Computer Vision*, pages 1959–1975, 2022. 8

[9] Jun Cheng. brain tumor dataset, 2017. 8

[10] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 5

[11] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018. 8

[12] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019. 8

[13] Yuexi Du, Brian Chang, and Nicha C Dvornek. Cleft: Language-image contrastive learning with efficient large language model and prompt fine-tuning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 465–475. Springer, 2024. 2

[14] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain?, 2021. 2

[15] Xiao Fang, Yi Lin, Dong Zhang, Kwang-Ting Cheng, and Hao Chen. Aligning medical images with general knowledge from large language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 57–67. Springer, 2024. 2

[16] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 1, 3, 6

[17] Roman Johannes Gertz, Thomas Dratsch, Alexander Christian Bunck, Simon Lennartz, Andra-Iza Iuga, Martin Gunnar Hellmich, Thorsten Persigehl, Lenhard Pennig, Carsten Herbert Gietzen, Philipp Fervers, et al. Potential of gpt-4 for detecting errors in radiology reports: Implications for reporting accuracy. *Radiology*, 311(1):e232714, 2024. 3, 4

[18] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 1

[19] Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. Lp++: A surprisingly strong linear probe for few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23773–23782, 2024. 1, 3, 6

[20] Md Nazmul Islam, Mehedi Hasan, Md Kabir Hossain, Md Golam Rabiul Alam, Md Zia Uddin, and Ahmet Soylu. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ct-radiography. *Scientific Reports*, 12(1):1–14, 2022. 5

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[22] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6(1): 1–11, 2016. 5

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

[24] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple:

Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2

[25] Muhammad Uzair khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[26] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 2

[27] Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Muzammal Naseer, Luc Van Gool, and Federico Tombari. Learning to prompt with text only supervision for vision-language models. *arXiv preprint arXiv:2401.02418*, 2024. 2, 3

[28] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-sam: Bridging text and image towards universal medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 643–653. Springer, 2024. 8

[29] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-samv2: Towards universal text-driven medical image segmentation. *arXiv preprint arXiv:2409.19483*, 2024. 8

[30] Satheesh Krishna, Nishaant Bhambra, Robert Bleakney, and Rajesh Bhayana. Evaluation of reliability, repeatability, robustness, and confidence of gpt-3.5 and gpt-4 on a radiology board–style examination. *Radiology*, 311(2):e232715, 2024. 3, 4

[31] Thomas Köhler, Attila Budai, Martin Kraus, Jan Odstrcilik, Georg Michelson, and Joachim Hornegger. Automatic no-reference quality assessment for retinal fundus images using vessel segmentation, 2013. 5

[32] Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Xiaotang Gai, Yang Feng, and Zuozhu Liu. A chatgpt aided explainable framework for zero-shot medical image diagnosis. *arXiv preprint arXiv:2307.01981*, 2023. 4

[33] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 2

[34] Msoud Nickparvar. Brain tumor mri dataset, 2021. 5

[35] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169, New York, NY, USA, 2017. ACM. 5

[36] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid), 2018. 5

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6

[38] Julio Silva-Rodriguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23681–23690, 2024. 3

[39] Anas M. Tahir, Muhammad E.H. Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M. Sohel Rahman, Somaya Al-Maadeed, Sakib Mahmud, Maymouna Ezeddin, Khaled Hameed, and Tahir Hamid. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in Biology and Medicine*, 139:105002, 2021. 5

[40] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, page 180161, 2018. 5

[41] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 8

[42] Yan Xu, Rixiang Quan, Weiting Xu, Yi Huang, Xiaolong Chen, and Fengyuan Liu. Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches. *Bioengineering*, 11(10): 1034, 2024. 2

[43] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization, 2023. 2, 7

[44] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023. 6

[45] Jamil Zaghir, Marco Naguib, Mina Bjelogrlic, Aurélie Névéol, Xavier Tannier, and Christian Lovis. Prompt engineering paradigms for medical applications: scoping review and recommendations for better practices. *arXiv preprint arXiv:2405.01249*, 2024. 7

[46] Chenrui Zhang, Lin Liu, Chuyuan Wang, Xiao Sun, Hongyu Wang, Jinpeng Wang, and Mingchen Cai. Prefer: Prompt ensemble learning via feedback-reflect-refine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19525–19532, 2024. 4

[47] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 1, 3, 6

[48] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu

Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2024. 1, 2, 6, 7

[49] Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A comprehensive survey. *arXiv preprint arXiv:2312.07353*, 2023. 2

[50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 1, 2, 6, 7

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 1, 2, 6, 7

[52] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023. 6

[53] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning, 2024. 2, 7