This is the Pre-Published Version.
The following publication Y. Du, N. Yu and S. Wang, "Medical Knowledge Intervention Prompt Tuning for Medical Image Classification," in IEEE Transactions on Medical Imaging, vol. 44, no. 12, pp. 4945-4959, Dec. 2025 is available at https://doi.org/10.1109/TMI.2025.3584841.

Medical Knowledge Intervention Prompt Tuning for Medical Image Classification

Ye Du, Nanxi Yu, and Shujun Wang, *Member, IEEE*

***Abstract*— Vision-language foundation models (VLMs) have shown great potential in feature transfer and generalization across a wide spectrum of medical-related downstream tasks. However, fine-tuning these models is resource-intensive due to their large number of parameters. Prompt tuning has emerged as a viable solution to mitigate memory usage and reduce training time while maintaining competitive performance. Nevertheless, the challenge is that existing prompt tuning methods cannot precisely distinguish different kinds of medical concepts, which miss essentially specific disease-related features across various medical imaging modalities in medical image classification tasks. We found that Large language models (LLMs), trained on extensive text corpora, are particularly adept at providing this specialized medical knowledge. Motivated by this, we propose incorporating LLMs into the prompt tuning process. Specifically, we introduce the CILMP, Conditional Intervention of Large Language Models for Prompt Tuning, a method that bridges LLMs and VLMs to facilitate the transfer of medical knowledge into VLM prompts. CILMP extracts disease-specific representations from LLMs, intervenes within a low-rank linear subspace and utilizes them to create disease-specific prompts. Additionally, a conditional mechanism is incorporated to condition the intervention process on each individual medical image, generating instance-adaptive prompts and thus enhancing adaptability. Extensive experiments across diverse medical image datasets demonstrate that CILMP consistently outperforms state-of-the-art prompt tuning methods, demonstrating its effectiveness. Code will be available at https://usr922.github.io/cilmp.**

***Index Terms*— Prompt Tuning, Vision Language Foundation Model, Representation Fine-tuning, Conditional Intervention, Medical Image Classification**

I. Introduction

With the rapid advancement of deep learning technologies, numerous studies have been proposed to enhance computer-aided diagnosis within medical data analysis. Recently, vision-language foundation models (VLMs) [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], which are trained on extremely large-scale data encompassing multiple domains and data distributions, have demonstrated significant potential in feature transfer and generalization across a wide array of downstream tasks. However, the fine-tuning of these foundation models is often prohibitively expensive due to the substantial number of model parameters involved.

To address the above challenge, parameter efficient fine-tuning (PEFT) methods [11, 12, 13, 14, 15, 16, 17] have emerged as a promising alternative by updating only a small subset of the model's weights while achieving performance levels comparable to full fine-tuning in many scenarios [18, 19, 20, 21, 22, 23, 24]. One of the most effective PEFT methods for adapting VLMs to downstream tasks is prompt tuning [11, 13]. This technique, which has its origins in the field of natural language processing [25, 26, 27, 28, 29], involves maintaining the fixed parameters of the large model while training a set of learnable tokens, referred to as prompt contexts [11]. These prompt contexts are concatenated with class names to serve as the input to the text encoder within VLMs. Following optimization, these prompts are employed to generate cosine classifiers [30, 11] for each category, thereby facilitating a range of tasks such as natural image classification [31, 32, 33, 34, 35, 36, 18, 20, 21] and semantic segmentation [22, 37, 23]. The straightforward nature of this paradigm highlights its potential to exploit the capabilities of pre-trained VLMs while preserving flexibility and achieving high performance.

Despite its success in natural image domains, the application of prompt tuning in medical image analysis still faces challenges. One primary challenge in medical imaging tasks is the requirement for precise differentiation of various medical concepts to accurately understand diseases, which underscores the necessity of integrating knowledge of the specialized medical domain [38, 39, 8, 40]. However, when adapting a pre-trained VLM to recognize "basal cell carcinoma" from dermatoscopy images, conventional prompt tuning methods employ disease-agnostic learnable prompt context tokens. These tokens lack the specific information necessary for the accurate identification of basal cell carcinoma, thereby constraining the full potential of the pre-trained VLM. Furthermore, existing prompt tuning methods typically construct shared prompt contexts across different categories. [11, 13, 31, 32, 36, 18, 41, 35] Such a design struggles with the fine-grained nature of medical image classification, as category names alone often do not provide sufficient information for recognition. For example, when distinguishing between "viral pneumonia" and "bacterial pneumonia" in X-ray images, the class names can barely

Ye Du and Nanxi Yu are with the Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China. (E-mail: {duyee.du, nx-nancy.yu}@connect.polyu.hk)

Shujun Wang is with the Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China. Shujun Wang is also affiliated with Research Institute for Smart Ageing, The Hong Kong Polytechnic University, Hong Kong SAR, China. (E-mail: shujun.wang@polyu.edu.hk)

Shujun Wang is the corresponding author.

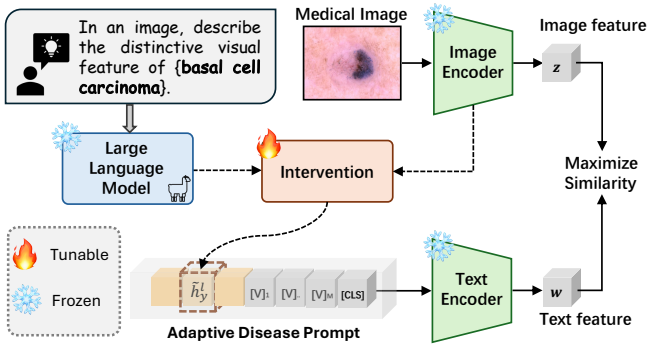


Fig. 1: Concept illustration of our CILMP method. CILMP first extracts concept-aware representations from a frozen large language model. It then intervenes in these representations with the guidance of image features to generate the adaptive disease prompts for the VLM text encoder.

provide adequate discriminative information to differentiate these two diseases due to their textual similarity.

Based on the preceding analysis, it is evident that incorporating medical domain knowledge into the prompt tuning process is essential. Large language models (LLMs) [42, 43, 44, 45, 46, 47, 48], which are trained on extensive text corpora, are particularly effective in providing this specialized knowledge [20, 21, 49]. These models possess a comprehensive understanding of various diseases, enabling them to generate context-specific information, discern subtle distinctions between similar medical conditions, and provide diverse medical knowledge. For instance, when queried about "basal cell carcinoma", an LLM can offer detailed descriptions of its distinctive visual features like "telangiectasia and rolled borders". This professional and nuanced information can significantly enhance the richness and accuracy of the prompts, as it provides discriminative representations for each disease.

Motivated by these capabilities, this paper aims to integrate large language models into the prompt tuning process to enhance the adaptation of vision-language foundation models for medical image analysis. Specifically, we propose Conditional Intervention of Large Language Models for Prompt Tuning (CILMP), a framework designed to bridge the gap between LLMs and VLMs by facilitating the transfer of medical domain knowledge from LLMs into VLM prompts. As shown in Fig. 1, for a certain disease, our CILMP framework first extracts concept-aware representations from the LLM. Given that LLMs and VLMs are usually pre-trained on different data modalities and distributions, the CILMP framework mitigates this gap by intervening in the LLM representations within a low-rank linear subspace. The corresponding intervention functions are trained to steer the framework's behavior towards accurate disease diagnosis, which is inspired by representation fine-tuning (ReFT) [16]. Moreover, we propose learning the intervention functions guided by input medical images to generate instance-adaptive prompts. This is achieved by conditioning the subspace intervention process on a relationship descriptor [22, 23] between the image feature and the unintervened LLM representations. Such a conditional mechanism allows for the incorporation of image-text matching prior into the prompts before making the final decision. Notably, the CILMP framework requires only the learning of intervention functions applied to the LLM representations, without necessitating gradient flow through the LLM itself, thus preserving the efficiency of the prompt tuning process.

To evaluate the effectiveness of the proposed CILMP framework, we conduct extensive experiments across 11 diverse medical image datasets. These datasets encompass a wide range of imaging modalities, including dermatoscope, fundus, ultrasound, histopathology, and X-ray images, etc. In our evaluation, we compare CILMP against 15 recently proposed state-of-the-art prompt tuning methods using four commonly used metrics: Accuracy, F1-score, Area Under Curve (AUC), and Kappa score. The experimental results demonstrate that CILMP consistently surpasses existing prompt tuning methods by considerable margins on 10 out of 11 datasets, thereby validating the effectiveness of our approach.

Our contributions are summarized as follows.

- We introduce the application of large language models to address the deficiencies in medical domain knowledge inherent in traditional prompt tuning methods for computer-aided diagnosis.
- To this end, we propose CILMP as an intermediary framework that bridges LLMs and vision-language foundation models, facilitating the generation of class-specific and instance-adaptive prompts.
- We perform extensive experiments to evaluate the performance of CILMP across a diverse set of datasets. Comparisons with recently proposed state-of-the-art prompt tuning methods demonstrate the effectiveness of our approach.

The rest of this paper is arranged as follows. We delve into related works in Section II. We then elaborate on the technical details of our method in Section III. Experimental results and discussions are presented in Section IV and Section V. Finally, we summarize the conclusions of this paper in Section VI.

II. Related Work

A. Large Language Models

Recent advancements in pre-training LLMs [42, 43, 44, 45, 46, 47, 48] have significantly enhanced their ability to understand and generate human language, as evidenced by models such as ChatGPT [45], GPT-4 [47], and LLaMA [42]. These models, pre-trained on extensive corpora and comprising billions of parameters, excel in capturing linguistic patterns and contextual relationships. Notably, LLaMA [42] and its iterations, LLaMA-2 [43] and LLaMA-3 [44], have demonstrated improvements in scale, efficiency, and reasoning capabilities. Various studies have leveraged LLMs for downstream tasks, such as generating predefined prompts for image classification [21], proposing LLM-guided concept bottlenecks for interpretable classification [49], and jointly training LLMs and VLMs for classification tasks [20]. Despite their high performance, the training workload for these large models remains substantial.

B. Vision Language Foundation Models

Recent advancements have significantly bridged visual signals with linguistic semantics, particularly through the pre-training of VLMs [1, 2, 3, 50, 4, 5, 6, 7]. Notably, the CLIP model [1] has demonstrated effective zero-shot prediction capabilities using a contrastive learning objective by learning the correlation between image and text embeddings [51]. Following this paradigm, Jia *et al.* [2] achieved robust zero-shot inference with exascale noisy image alt-text data, while Alayrac *et al.* [3] bridge powerful pre-trained vision-only and language-only models to enable powerful in-context few-shot learning capabilities. Li *et al.* [4] unified image-text understanding and generation, enhancing functionalities such as image captioning and visual question answering. Yang *et al.* [6] introduced an attentive masking mechanism to improve VLM pre-training efficiency.

In the medical imaging domain, notable efforts [8, 9, 10] have focused on pre-training VLMs with large-scale medical data. For instance, Zhang *et al.* [8] incorporated medical domain knowledge into pre-training for chest X-ray images using a knowledge graph. Christensen *et al.* [10] developed a VLM for echocardiography, learning relationships between ultrasound images and expert interpretations. Additionally, Kim *et al.* [9] highlighted VLMs' suitability for trustworthy and transparent medical AI systems. For further details on VLMs, refer to [52, 53].

C. Prompt Tuning

Prompt tuning [25, 26, 27, 28, 29, 40, 11] is a parameter-efficient fine-tuning method, which keeps the VLM parameters fixed and only trains additional prompt tokens, achieving comparable or superior performance to full fine-tuning while maintaining high efficiency [22, 37, 23]. Building on this paradigm, numerous studies [31, 32, 54, 36, 18, 33, 20, 34, 35, 21] have sought to enhance the adaptation performance of VLMs. Specifically, Zhou *et al.* [13] propose conditional context optimization by training a meta-network to generate image-condition prompt tokens, Yao *et al.* [31] introduce knowledge-guided context optimization, and Lee *et al.* [32] present a read-only prompt optimization framework by masked attention. Other notable contributions include text-to-text optimization strategy [36], self-regularization strategies [18], and class-aware prompt tuning approach [33]. Additionally, Chen *et al.* integrate optimal transport into prompt tuning [55], and Zhang *et al.* propose decoupled prompt tuning [35]. Roy *et al*. emphasize consistency in the prompt tuning process to prevent overfitting [21]. Moreover, recent studies [56, 40, 41] also explore integrating domain-specific knowledge into the prompt tuning process, utilizing GPT-4 [47] or pre-trained MedSAM [57]. For a comprehensive overview of prompt tuning methods, readers are referred to recent surveys [58, 59].

D. Representation Fine-Tuning

Recently, Wu *et al.* [16] introduced the Representation Fine-Tuning (ReFT) paradigm as an alternative to parameter-efficient fine-tuning methods, such as prompt tuning [25], LoRA [12], and model adapters [14]. ReFT is motivated by the interchange intervention studies [60, 61] that aim to interpret deep learning models by establishing the causal role of a representation or testing the concepts it encapsulates. Specifically, ReFT freezes the model parameters and intervenes in the model representations to steer the model's behavior toward adaptation on the downstream task. This process is operated in a linear subspace using the learnable intervention function, thus it is highly efficient. In this paper, we are motivated by this adaptation technique, by utilizing it to incorporate medical domain knowledge into the prompts for the VLM, in order to construct disease-specific and instance-adaptive prompts to enhance the transferability of VLM on the medical image classification task.

III. Methodology

In this section, we begin with a brief review of the vision-language foundation model and the prompt tuning paradigm. Subsequently, we introduce the overall pipeline of the proposed CILMP framework. We then provide a comprehensive explanation of the conditional intervention mechanism within the CILMP framework, which adaptively modulates the intervention on LLM representations based on the input image, thereby generating class-specific and instance-adaptive prompts. Finally, we summarize the training loss and the inference process of our method.

A. Preliminary

1) Vision-Language Foundation Model: We first briefly introduce the Vision-Language Foundation Model, using the groundbreaking CLIP model [1] as an illustrative example. CLIP consists of an image encoder $f(\cdot)$ and a text encoder $g(\cdot)$, both trained via contrastive learning [51] on a massive number of image-text pairs. Typically, the image and text encoders are instantiated as a Vision Transformer [63] and a standard Language Transformer [64], respectively. During training, CLIP considers each image $\boldsymbol{x}$ and its associated text $\boldsymbol{t}$ as a positive pair and utilizes the InfoNCE loss [51] for instance discrimination. Specifically, CLIP first extracts the image and text embeddings via

$$\boldsymbol{z} = \text{Normalize}\left(f(\boldsymbol{x})\right), \tag{1}$$

$$\boldsymbol{w} = \text{Normalize}\left(g(\boldsymbol{t})\right), \tag{2}$$

where $\boldsymbol{z}$ is derived from a learnable classification token prepended to the image tokens and $\boldsymbol{w}$ is drawn from the $[EOS]$ token appended after the text description [63, 64]. The InfoNCE loss is then applied to these normalized embeddings for pre-training:

$$L_{\text{CLIP}} = 0.5 \times (L_v + L_t), \tag{3}$$

$$\text{where} \quad L_v = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{w}_i/\tau)}{\sum_{j=1}^{N}\exp(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{w}_j/\tau)}, \tag{4}$$

$$L_t = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{z}_i/\tau)}{\sum_{j=1}^{N}\exp(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{z}_j/\tau)}, \tag{5}$$

where N denotes the batch size during training and τ is a learnable temperature parameter used to scale the logits.

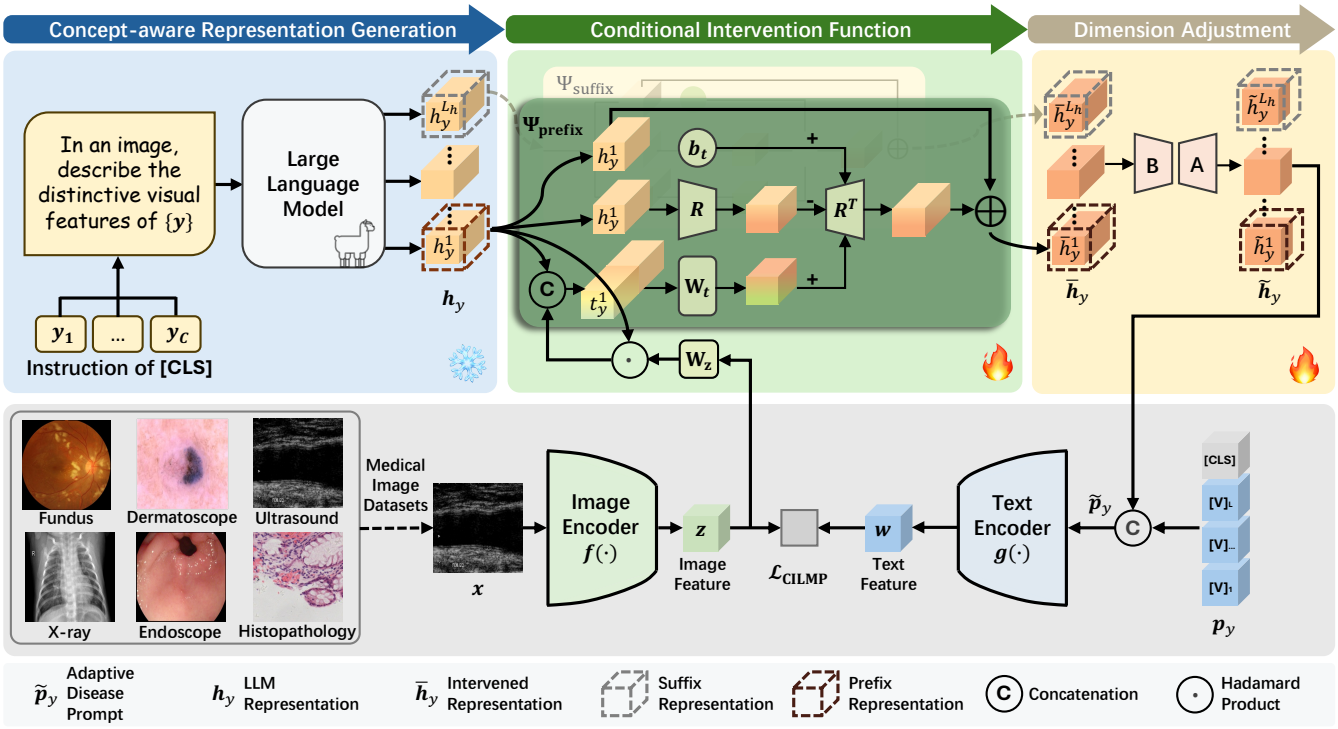


Fig. 2: Illustration of the CILMP framework. CILMP first extracts concept-aware representations $\boldsymbol{h}_y$ from an LLM. Then, a conditional intervention function is introduced to adapt these representations towards accurate disease label prediction, producing intervened representations $\bar{\boldsymbol{h}}_y$. After dimension adjustment, $\tilde{\boldsymbol{h}}_y$ are concatenated with the original prompts $\boldsymbol{p}_y$ to generate the adaptive disease prompts $\tilde{\boldsymbol{p}}_y$. Finally, $\mathcal{L}_{\text{CILMP}}$ is used to guide the prompt tuning process for the VLM.

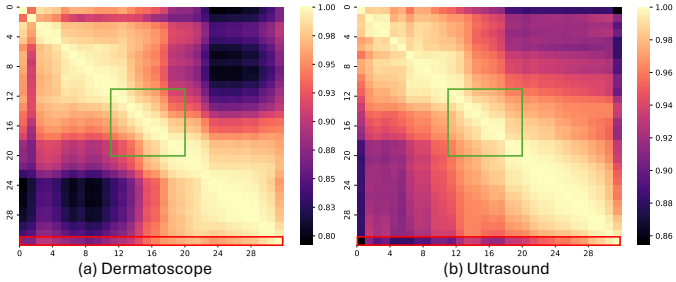


Fig. 3: Centered kernel alignment heatmap [62] between representations from different layers of the LLaMA3-8B [44]. The red box (last row) displays the similarity between representation from the last layer and those from other layers, while the green box highlights the similarity between adjacent layers.

2) Prompt Tuning: Prompt tuning could adapt CLIP to specific downstream tasks without the need to fine-tune all the model parameters [11]. Specifically, the prompt of each class is designed by

$$\boldsymbol{p} = [V]_1[V]_2[...][V]_L[CLS], \tag{6}$$

where each $[V]_l, (l \in \{1, ..., L\})$ is a learnable context vector, $[CLS]$ is the embedding of each class name, and L is a hyper-parameter specifying the number of learnable context tokens. During fine-tuning, the parameters of the CLIP encoders are frozen, and only the context vectors are optimized using the contrastive training objective:

$$L_{\text{prompt}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{w}_{y_i}/\tau)}{\sum_{c=1}^{C}\exp(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{w}_c/\tau)}, \tag{7}$$

where $\boldsymbol{y}_i$ is the class index of $\boldsymbol{z}_i$, C is the total number of classes.

Although the existing prompt tuning paradigm has achieved remarkable performance with minimal training overhead, it exhibits two significant limitations. First, it lacks the specialized medical domain knowledge crucial for computer-aided diagnosis tasks. Second, the learnable context vectors are generally shared across categories, which prevents them from serving as discriminative features for specific diseases. In this paper, we address these challenges by proposing a CILMP approach.

B. The CILMP Framework

Recent studies [49, 21, 20] have demonstrated that LLMs are effective candidates for providing concept-specific rich knowledge for various purposes. Therefore, we propose the CILMP framework (Fig. 2), which leverages the encyclopedic knowledge embedded in pre-trained large language models to address the limitations of existing prompt tuning methods.

1) Concept-aware Representation Generation: Given a dataset $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, where $y \in \{1, ..., C\}$ denotes the disease types corresponding to each image $\boldsymbol{x}$ in a specific modality (*e.g.*, dermatoscopy images), we generate concept-aware representations $\boldsymbol{h}_y$ by querying the LLM for class y (*e.g.* basal cell carcinom). Specifically, we prompt the LLM with the question: "In an image, describe the distinctive visual features of $\{y\}$". Then LLM responds with a sequence representation for y, denoted by $\boldsymbol{h}_y \in \mathbb{R}^{L_h \times D_h}$, L_h represents the sequence length and D_h denotes the hidden dimensionality of the LLM.

Each element $\boldsymbol{h}_y^l$ ($l \in \{1, ..., L_h\}$) in $\boldsymbol{h}_y$ is drawn from the last token (*i.e.*, the EOS token) of each independent layer of the LLM, as it encapsulates the information for the entire response. Notably, these elements contain different semantic information. To support this assertion, we compare the representations from different layers of the LLM using Centered Kernel Alignment [62], a powerful tool for determining the correspondence between the hidden layers of neural networks. As shown in Fig. 3, the heatmap reveals a substantial dissimilarity between the last layer and the preceding layers, as well as noticeable differences between adjacent layers. This underscores the importance of utilizing these elements, each containing distinct semantic information, to enhance the process of concept-aware representation generation. As such, we derive $\boldsymbol{h}_y$, a concept-aware representation that encapsulates rich medical knowledge about disease y, because it implies the LLM's comprehensive understanding of the queried disease type.

2) Adaptive Disease Prompts Construction: The inherent domain gap between the representation spaces of LLMs and VLMs hinders direct knowledge transfer between them. To address this, we introduce an additional adaptation function, a meta-function $\Psi_\theta(\cdot)$ parameterized by θ. $\Psi_\theta(\cdot)$ maps $\boldsymbol{h}_y$ into the VLM feature space by optimizing towards the true disease label. Formally, the adapted LLM representation $\overline{\boldsymbol{h}}_y$ for each class y is obtained via

$$\overline{\boldsymbol{h}}_y = \Psi_\theta(\boldsymbol{h}_y). \tag{8}$$

Then we concatenate $\overline{\boldsymbol{h}}_y$ with the original prompts $\boldsymbol{p}_y$ (obtained via Eq. (6)) to form the adaptive disease prompt for the text encoder. To deal with the dimension mismatch problem between $\overline{\boldsymbol{h}}_y$ and $\boldsymbol{p}_y$, we introduce a linear projection layer $\boldsymbol{W}$ to adjust the dimensions. To maintain parameter efficiency, we employ a low-rank decomposition [12] for $\boldsymbol{W}$, generating $\widetilde{\boldsymbol{h}}_y \in \mathbb{R}^{L_h \times D_p}$:

$$\widetilde{\boldsymbol{h}}_y = \boldsymbol{W}\overline{\boldsymbol{h}}_y = \boldsymbol{B}\boldsymbol{A}\overline{\boldsymbol{h}}_y, \tag{9}$$

where $\boldsymbol{A} \in \mathbb{R}^{r \times D_h}$ and $\boldsymbol{B} \in \mathbb{R}^{D_p \times r}$. Here, r denotes the low-rank dimensionality and D_p represents the dimensionality of original prompt tokens. Finally, following [20], the adaptive disease prompts for y, denoted by $\widetilde{\boldsymbol{p}}_y \in \mathbb{R}^{(L_h+L) \times D_p}$, is defined as

$$\widetilde{\boldsymbol{p}}_y := \text{concat}[\widetilde{\boldsymbol{h}}_y,\ \boldsymbol{p}_y], \tag{10}$$

C. Conditional Intervention Mechanism

1) Unconditional Intervention: The goal of CILMP is to control the transformation of LLM representations through targeted interventions, directing the entire pipeline toward predicting disease labels. To this end, an unconditional intervention mechanism [16] can be introduced. Specifically, given $\boldsymbol{h}_y^l \in \mathbb{R}^{D_h}$ at the l-th ($l \in \{1, ..., L_h\}$) position of the sequential LLM representation $\boldsymbol{h}_y$, the adaptation process can be formulated as a standard interchange-based intervention function [16] applied on top of $\boldsymbol{h}_y^l$. Formally, $\Psi : \mathbb{R}^{D_h} \to \mathbb{R}^{D_h}$ is defined as

$$\Psi_\theta(\boldsymbol{h}_y^l) = \boldsymbol{h}_y^l + \mathbf{R}^\mathrm{T}(\mathbf{W}_h\boldsymbol{h}_y^l + \mathbf{b}_h - \mathbf{R}\boldsymbol{h}_y^l), \tag{11}$$

where $\mathbf{R} \in \mathbb{R}^{r \times D_h}$ is a low-rank projection matrix, $\mathbf{W}_h \in \mathbb{R}^{r \times D_h}$, $\mathbf{b}_h \in \mathbb{R}^r$, and r is the dimensionality of the subspace being intervened upon. The learnable parameters are thus composed as $\theta = \{\mathbf{R}, \mathbf{W}_h, \mathbf{b}_h\}$.

In this formulation, $\mathbf{W}_h\boldsymbol{h}_y^l + \mathbf{b}_h$ represents a learned projected term that modifies the original $\boldsymbol{h}_y^l$ within a linear subspace spanned by the rows of $\mathbf{R}$. In other words, during supervision, Eq. 11 learns to "edit" the LLM representation by embedding the concept of the target disease label into it. The editing process operates within a linear subspace, where $\mathbf{R}$ is trained to project the original LLM representation into this subspace. Therefore, this can be explained as finding the subspace that maximizes the probability of the desired output after the intervention. Notably, this process is highly efficient, as the learnable parameters are represented by low-rank matrices.

2) Conditional Intervention: The aforementioned process utilizes an unconstrained intervention mechanism, which means the learned projected term $\mathbf{W}_h\boldsymbol{h}_y^l + \mathbf{b}_h$ remains unrestricted. However, we seek to enable the generation of adaptive prompts that not only incorporate medical knowledge specific to each category, but are also tailored to each individual image, in order to enhance the flexibility and adaptability [13, 54]. As such, we further incorporate the matching prior [22] into the intervention function. The matching prior quantifies modality compatibility between image and text representations before the final cross-modal alignment, which aims at preconditioning the adaptive prompt generation on each individual image example. To achieve this, we further propose a conditional intervention mechanism, which conditions the intervention process on a Relationship Descriptor (RD) between the image representation $\boldsymbol{z}$ and each unintervened representation $\boldsymbol{h}_y^l$. Following [22, 23], the RD, denoted by $\boldsymbol{t}_y^l \in \mathbb{R}^{2D_h}$, is calculated by

$$\boldsymbol{t}_y^l = \text{concat}[\boldsymbol{h}_y^l,\ \boldsymbol{h}_y^l \odot \mathbf{W}_z\boldsymbol{z}], \tag{12}$$

where $\odot$ is the Hadamard product, $\mathbf{W}_z \in \mathbb{R}^{D_h \times D_p}$ is introduced for size adjustment. To maintain parameter efficiency, we also decompose $\mathbf{W}_z$ with two low-rank matrices [12]. Subsequently, we leverage RD to intervene on the LLM representation within the r-dimensional linear subspace, as given by

$$\Psi_\theta(\boldsymbol{h}_y^l, \boldsymbol{z}) = \boldsymbol{h}_y^l + \mathbf{R}^\mathrm{T}(\mathbf{W}_t\boldsymbol{t}_y^l + \mathbf{b}_t - \mathbf{R}\boldsymbol{h}_y^l), \tag{13}$$

where $\mathbf{W}_t \in \mathbb{R}^{r \times 2D_h}$ and $\mathbf{b}_t \in \mathbb{R}^r$. The set of learnable parameters is therefore updated to $\theta = \{\mathbf{R}, \mathbf{W}_t, \mathbf{W}_z, \mathbf{b}_t\}$.

In addition, given that the large language model generates a sequence of representations, addressing how to intervene effectively across the entire sequence presents another problem. Unlike PEFT methods, representation fine-tuning operates on a sequence of representations, necessitating interventions on each element in the sequence. However, intervening on all elements can lead to excessive computational burden, especially when dealing with lengthy sequences. This excessive intervention may impede extendability and risk the loss of valuable medical knowledge embedded in the original LLM representation [16]. To strike a balance between adaptation and knowledge retention, a compromise strategy should be

devised to determine efficient interventions across the entire sequence. Drawing inspiration from ReFT, a practical approach involves intervening on a subset of them, where the length of intervention acts as a hyperparameter. This not only maintains efficiency but also introduces flexibility. In light of this, our CILMP framework adopts a simple bilateral intervention strategy [16]. Specifically, we propose the development of separate intervention functions for the prefix and suffix segments of the LLM representations. In this context, the intervention can be encapsulated as $\langle \Psi, P \rangle$, where P denotes the set of positions within the LLM representation at which Ψ is applied. Hence, the prefix intervention $\langle \Psi_{\text{prefix}}, P_{\text{prefix}} \rangle$ targets positions $P_{\text{prefix}} = \{1, ..., L_{\text{prefix}}\}$, while the suffix intervention $\langle \Psi_{\text{suffix}}, P_{\text{suffix}} \rangle$ focuses on positions $P_{\text{suffix}} = \{L_h - L_{\text{suffix}}, ..., L_h\}$, where L_{prefix} and L_{suffix} are hyper-parameters that specify the intervention lengths.

D. Model Optimization and Inference

After obtaining the adaptive disease prompts $\widetilde{\boldsymbol{p}}$ via Eq. (10) for each category, we proceed by freezing the parameters of the VLM. Then, we utilize the supervised contrastive loss [65] to guide the prompt tuning process as

$$\mathcal{L}_{\text{CILMP}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\boldsymbol{z}_i^{\mathrm{T}} g(\widetilde{\boldsymbol{p}}_{y_i})/\tau)}{\sum_{c=1}^{C}\exp(\boldsymbol{z}_i^{\mathrm{T}} g(\widetilde{\boldsymbol{p}}_c)/\tau)}, \tag{14}$$

where N denotes the number of samples in a training batch, τ is the learned temperature parametered by the VLM.

During inference, the CILMP framework predicts the disease label for a given image $\boldsymbol{x}$ by comparing the cosine similarity between the image embedding and the embeddings of each enhanced prompt. In particular, the probability that the image $\boldsymbol{x}$ belongs to class c is computed by

$$p(y=c|\boldsymbol{x}) = \frac{\exp(f(\boldsymbol{x})^{\mathrm{T}} g(\widetilde{\boldsymbol{p}}_c)/\tau)}{\sum_{j=1}^{C}\exp(f(\boldsymbol{x})^{\mathrm{T}} g(\widetilde{\boldsymbol{p}}_j)/\tau)}. \tag{15}$$

IV. Experiments

A. Datasets

There are 11 datasets utilized for evaluation, which encompass six different data modalities. Specifically, we conduct experiments on the dermatoscope modality using the DermaMNIST [66], Derm7pt [67], and ISIC 2018 [68] datasets. For the fundus modality, we use the ADAM [69], APTOS 2019 [70], and ODIR [71] datasets. Additionally, ultrasound images are evaluated using the Fetal-US [72] dataset, histopathology images are evaluated with the Chaoyang [73] dataset, and endoscope images are assessed using the Kvasir [74] dataset. X-ray images are incorporated using the CPN-X-ray [75] and the Pneumonia [76] datasets. TABLE I summarizes the number of classes, training images, and test images for each dataset employed in this paper.

B. Implementation Details

For a fair comparison, all experiments are conducted using the pre-trained CLIP model [1]. Specifically, the image encoder employs the ViT-B/16 backbone [63], which has an output dimensionality of 512. The text encoder utilizes the Transformer backbone, incorporating architectural refinements as detailed in [77]. In our CILMP framework, we utilize the LLaMA-3-8B [44] as the LLM, which is one of the most capable openly available LLM to date. LLAMA-3-8B is a decoder-only LLM, which has 32 layers with a representation dimension of 4096. The length of contexts for image prompts and text prompts is set to 4. In alignment with existing methods [18, 35, 55, 21], we employ a deep prompting strategy for both encoders, injecting the prompts into intermediate layers. Prompts are initialized with a zero-mean Gaussian distribution with a standard deviation of 0.02, except for the text prompts in the first layer, which are initialized with the word embeddings of "a photo of a". The model is trained for 100 epochs with a batch size of 64. During training, an SGD optimizer is employed with an initial learning rate of 0.0025. Training hyper-parameters are kept consistent across datasets unless otherwise specified. To ensure fair comparisons, we train all the compared methods with the same configurations, including the training epochs, batch size, and model backbones. All models are trained using one NVIDIA RTX 4090 GPU. We report the performance of the model from the last epoch, and no validation set is used during the training process. For the image inputs, although data augmentation is very useful to improve performance in medical image analysis, we simply follow [11, 13, 35, 18] to apply random resized crop and random flip as the data augmentations to maintain fair comparisons. The input size of training images is 224×224. During the inference stage, the size of test images is also set to 224×224. For the evaluation of the methods, we employ four widely recognized metrics, including the Accuracy, F1-score, area under the ROC curve (AUROC), and Kappa score. These metrics are chosen to provide a comprehensive assessment of the model performance across different aspects. The results are reported as the mean and standard deviation over three independent runs with different random seeds.

TABLE I: The statistics of 11 datasets over 6 modalities.

Dataset	Modality	# Classes	# Training	# Test
ADAM [69]	Fundus	2	400	400
APTOS 2019 [70]	Fundus	5	2,930	732
Chaoyang [73]	Histopathology	4	4,021	2,139
CPN-X-ray [75]	X-ray	3	3,659	1,569
DermaMNIST [66]	Dermatoscope	7	5,600	2,400
Derm7pt [67]	Dermatoscope	5	413	395
Fetal-US [72]	Ultrasound	6	7,129	5,271
ISIC 2018 [68]	Dermatoscope	7	10,015	1,512
Kvasir [74]	Endoscope	8	5,600	2,400
ODIR [71]	Fundus	8	5,113	1,279
Pneumonia [76]	X-ray	3	5,216	624

C. Comparison with the State-of-the-art (SOTA) Methods

In TABLEs II and III, we compare our method with 15 SOTA prompt tuning methods over 11 datasets. The compared benchmarking methods include CoOp [11], CoCoOp [13], KgCoOp [8], CLIP-Adapter [14], PRO [32], LASP [36], PromptSRC [18], TCP [33], AAPL [34], PLOT++ [55], DCPL [41], ViP [56], XCoOp [40], DePT [35] and Co-Prompt [21]. Notably, DCPL introduces MedSAM as a knowl-

TABLE II: Comparison with state-of-the-art methods on 6 datasets including ADAM, APTOS 2019, Chaoyang, CPN-X-ray, Derm7pt, and DermaMNIST. The best and the second best results are highlighted in red and blue, respectively.

Method	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
CoOp [11]	84.0 ± 1.5	77.5 ± 2.4	86.3 ± 0.8	55.0 ± 4.7
CoCoOp [13]	81.5 ± 0.6	74.3 ± 0.9	84.3 ± 1.6	48.6 ± 1.7
KgCoOp [31]	79.8 ± 0.0	56.3 ± 1.2	81.5 ± 1.0	18.2 ± 1.6
CLIP-Adapter [14]	82.4 ± 1.0	74.9 ± 1.1	84.8 ± 0.5	49.8 ± 2.2
RPO [32]	87.8 ± 1.2	82.7 ± 1.6	92.7 ± 0.4	65.5 ± 3.2
LASP [36]	86.8 ± 0.6	81.3 ± 0.6	90.5 ± 0.6	62.6 ± 1.2
PromptSRC [18]	88.2 ± 0.3	83.2 ± 1.2	91.8 ± 1.6	66.5 ± 2.3
TCP [33]	80.5 ± 0.0	60.5 ± 0.6	83.4 ± 0.3	24.7 ± 0.9
AAPL [34]	82.0 ± 0.9	75.6 ± 0.9	85.3 ± 0.6	51.4 ± 1.8
PLOT++ [55]	87.6 ± 1.2	82.4 ± 1.4	92.3 ± 0.2	64.8 ± 2.9
DCPL [41]	86.3 ± 0.8	80.7 ± 0.9	89.8 ± 0.5	61.4 ± 1.8
ViP [56]	83.9 ± 0.1	72.6 ± 0.4	83.9 ± 1.0	45.9 ± 0.8
XCoOp [40]	84.7 ± 0.5	79.6 ± 0.6	89.1 ± 0.1	59.3 ± 1.3
DePT [35]	88.1 ± 0.9	83.9 ± 1.1	93.0 ± 0.2	70.0 ± 2.1
CoPrompt [21]	87.6 ± 0.2	82.3 ± 0.4	90.7 ± 0.6	64.6 ± 0.9
CILMP (Ours)	90.1 ± 0.4	85.4 ± 0.6	93.5 ± 0.3	70.9 ± 1.2

(a) ADAM

Method	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
CoOp [11]	76.8 ± 0.2	40.2 ± 0.6	91.1 ± 0.2	62.3 ± 0.2
CoCoOp [13]	79.2 ± 0.6	51.8 ± 1.4	92.2 ± 0.1	66.8 ± 0.9
KgCoOp [31]	74.1 ± 0.3	32.7 ± 0.1	87.0 ± 0.4	57.8 ± 0.4
CLIP-Adapter [14]	73.2 ± 0.1	32.3 ± 0.1	84.2 ± 0.1	56.3 ± 0.2
RPO [32]	81.9 ± 0.4	61.4 ± 1.4	93.6 ± 0.1	71.7 ± 0.7
LASP [36]	76.0 ± 0.2	38.9 ± 0.7	88.7 ± 0.2	60.8 ± 0.3
PromptSRC [18]	76.8 ± 0.6	38.3 ± 2.3	92.7 ± 0.4	62.3 ± 1.0
TCP [33]	74.4 ± 0.2	32.7 ± 0.1	87.9 ± 0.1	58.2 ± 0.2
AAPL [34]	78.4 ± 1.0	48.9 ± 3.6	91.8 ± 0.2	65.4 ± 1.7
PLOT++ [55]	83.5 ± 0.4	65.5 ± 0.3	94.3 ± 0.1	74.5 ± 0.6
DCPL [41]	83.6 ± 0.3	64.7 ± 0.1	94.4 ± 0.1	74.7 ± 0.4
ViP [56]	78.8 ± 0.5	50.9 ± 1.4	90.9 ± 0.4	66.3 ± 0.8
XCoOp [40]	80.9 ± 0.6	59.6 ± 1.2	93.3 ± 0.1	70.1 ± 0.9
DePT [35]	81.7 ± 0.4	56.0 ± 2.2	94.1 ± 0.2	71.0 ± 0.8
CoPrompt [21]	82.7 ± 0.9	66.0 ± 1.8	94.2 ± 0.2	73.3 ± 1.5
CILMP (Ours)	84.5 ± 0.5	66.7 ± 0.8	94.6 ± 0.2	75.7 ± 0.7

(b) APTOS 2019

Method	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
CoOp [11]	74.1 ± 0.4	65.5 ± 0.3	89.4 ± 0.2	62.3 ± 0.6
CoCoOp [13]	76.5 ± 0.4	68.5 ± 0.4	90.6 ± 0.2	65.8 ± 0.4
KgCoOp [31]	65.0 ± 0.5	55.6 ± 0.3	83.4 ± 0.2	48.2 ± 0.6
CLIP-Adapter [14]	73.0 ± 0.3	63.6 ± 0.3	88.7 ± 0.2	60.4 ± 0.5
RPO [32]	78.4 ± 0.6	71.0 ± 0.2	91.4 ± 0.4	68.6 ± 0.9
LASP [36]	78.8 ± 0.3	71.4 ± 0.3	92.1 ± 0.3	69.3 ± 0.4
PromptSRC [18]	78.8 ± 1.0	69.8 ± 1.2	92.9 ± 0.4	69.0 ± 1.5
TCP [33]	67.6 ± 0.4	58.1 ± 0.6	85.4 ± 0.2	52.0 ± 0.5
AAPL [34]	75.3 ± 0.4	67.0 ± 0.5	89.9 ± 0.3	64.1 ± 0.4
PLOT++ [55]	79.9 ± 0.1	73.7 ± 0.1	92.7 ± 0.1	71.0 ± 0.2
DCPL [41]	82.3 ± 0.3	76.0 ± 0.3	94.0 ± 0.1	74.5 ± 0.5
ViP [56]	76.2 ± 0.3	66.8 ± 1.4	89.7 ± 0.2	65.2 ± 0.4
XCoOp [40]	79.7 ± 0.1	72.6 ± 0.2	93.0 ± 0.1	70.7 ± 0.2
DePT [35]	79.6 ± 0.7	71.4 ± 0.4	93.0 ± 0.1	70.3 ± 1.1
CoPrompt [21]	80.8 ± 0.2	74.8 ± 0.4	93.8 ± 0.1	72.4 ± 0.3
CILMP (Ours)	82.3 ± 0.3	76.1 ± 0.5	94.1 ± 0.1	74.4 ± 0.4

(c) Chaoyang

Method	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
CoOp [11]	93.7 ± 0.3	93.9 ± 0.3	99.0 ± 0.1	90.6 ± 0.5
CoCoOp [13]	94.9 ± 0.1	95.0 ± 0.1	99.2 ± 0.1	92.2 ± 0.1
KgCoOp [31]	83.7 ± 0.5	84.0 ± 0.5	96.4 ± 0.0	75.3 ± 0.8
CLIP-Adapter [14]	92.2 ± 0.2	92.3 ± 0.1	98.7 ± 0.0	88.2 ± 0.2
RPO [32]	96.8 ± 0.1	96.9 ± 0.1	99.6 ± 0.0	95.1 ± 0.2
LASP [36]	95.5 ± 0.2	95.6 ± 0.2	99.4 ± 0.0	93.2 ± 0.3
PromptSRC [18]	92.7 ± 0.1	92.8 ± 0.1	99.2 ± 0.1	89.0 ± 0.1
TCP [33]	86.4 ± 0.2	86.7 ± 0.2	97.3 ± 0.1	79.5 ± 0.4
AAPL [34]	94.5 ± 0.4	94.6 ± 0.3	99.2 ± 0.0	91.7 ± 0.5
PLOT++ [55]	95.5 ± 0.0	95.7 ± 0.0	99.4 ± 0.1	93.3 ± 0.1
DCPL [41]	96.6 ± 0.1	96.7 ± 0.1	99.6 ± 0.0	94.9 ± 0.1
ViP [56]	94.6 ± 0.2	94.7 ± 0.2	99.0 ± 0.1	91.8 ± 0.3
XCoOp [40]	96.0 ± 0.1	96.2 ± 0.1	99.6 ± 0.0	94.1 ± 0.1
DePT [35]	95.2 ± 0.7	95.4 ± 0.7	99.3 ± 0.0	92.9 ± 1.0
CoPrompt [21]	97.7 ± 0.1	97.8 ± 0.0	99.8 ± 0.0	96.5 ± 0.1
CILMP (Ours)	96.2 ± 0.3	96.3 ± 0.3	99.6 ± 0.0	94.3 ± 0.5

(d) CPN-X-ray

Method	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
CoOp [11]	77.4 ± 0.6	50.5 ± 1.3	93.3 ± 0.5	55.8 ± 1.3
CoCoOp [13]	79.1 ± 0.6	57.6 ± 1.8	94.6 ± 0.3	60.6 ± 0.9
KgCoOp [31]	70.7 ± 0.1	26.4 ± 0.3	83.3 ± 1.1	32.8 ± 0.8
CLIP-Adapter [14]	76.7 ± 0.2	43.0 ± 0.2	90.2 ± 0.2	53.9 ± 0.3
RPO [32]	82.0 ± 0.6	60.8 ± 3.1	95.6 ± 0.1	65.2 ± 1.4
LASP [36]	82.4 ± 0.2	65.9 ± 0.5	95.8 ± 0.2	66.4 ± 0.3
PromptSRC [18]	83.8 ± 0.4	64.1 ± 2.0	95.5 ± 0.5	68.1 ± 0.8
TCP [33]	72.2 ± 0.2	29.2 ± 1.0	87.1 ± 0.2	37.3 ± 0.5
AAPL [34]	78.9 ± 0.1	54.0 ± 1.1	93.4 ± 0.2	57.7 ± 0.1
PLOT++ [55]	82.9 ± 0.3	66.6 ± 1.3	95.9 ± 0.1	67.4 ± 0.6
DCPL [41]	86.8 ± 0.5	75.9 ± 0.6	97.9 ± 0.1	74.8 ± 0.7
ViP [56]	80.0 ± 0.5	55.1 ± 3.1	94.4 ± 0.1	59.9 ± 0.8
XCoOp [40]	84.3 ± 0.1	71.1 ± 0.7	96.7 ± 0.1	69.9 ± 0.3
DePT [35]	83.8 ± 0.6	64.4 ± 0.3	95.9 ± 0.1	68.1 ± 0.9
CoPrompt [21]	84.6 ± 0.4	69.3 ± 0.3	97.0 ± 0.1	70.0 ± 0.5
CILMP (Ours)	87.4 ± 0.3	77.7 ± 1.0	98.0 ± 0.0	76.0 ± 0.4

(e) DermaMNIST

Method	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
CoOp [11]	67.2 ± 1.8	43.9 ± 5.1	86.8 ± 0.9	41.3 ± 2.9
CoCoOp [13]	64.6 ± 1.6	33.4 ± 2.5	81.6 ± 1.3	31.5 ± 3.5
KgCoOp [31]	63.9 ± 0.3	35.1 ± 0.8	80.3 ± 1.2	28.9 ± 0.8
CLIP-Adapter [14]	69.5 ± 0.1	46.6 ± 0.3	84.8 ± 0.2	44.9 ± 0.4
RPO [32]	70.7 ± 0.9	48.9 ± 2.9	87.9 ± 0.6	47.2 ± 2.1
LASP [36]	69.6 ± 1.4	49.4 ± 2.0	83.6 ± 0.7	44.3 ± 3.6
PromptSRC [18]	70.1 ± 0.3	46.5 ± 1.9	87.6 ± 0.8	46.3 ± 0.7
TCP [33]	66.8 ± 0.6	39.1 ± 0.4	82.6 ± 0.4	36.6 ± 0.9
AAPL [34]	67.4 ± 0.8	43.0 ± 2.2	84.7 ± 0.1	39.8 ± 2.0
PLOT++ [55]	69.5 ± 0.5	45.3 ± 1.9	86.6 ± 0.5	44.3 ± 1.0
DCPL [41]	70.4 ± 1.8	49.6 ± 1.2	87.9 ± 1.7	48.6 ± 2.7
ViP [56]	69.1 ± 1.5	43.0 ± 2.8	83.9 ± 1.8	43.5 ± 4.1
XCoOp [40]	71.4 ± 0.2	53.4 ± 1.5	89.0 ± 0.0	49.1 ± 0.8
DePT [35]	73.0 ± 0.4	50.7 ± 1.1	90.2 ± 0.3	52.3 ± 1.0
CoPrompt [21]	71.9 ± 2.0	49.8 ± 4.1	89.0 ± 0.4	51.8 ± 4.1
CILMP (Ours)	76.4 ± 1.8	58.3 ± 0.8	91.9 ± 0.5	59.6 ± 3.5

(f) Derm7pt

edge provider, KgCoOp introduces hand-crafted knowledge into prompt tuning, and ViP, XCoOp and CoPrompt employ LLM to provide fixed knowledge.

Our method significantly outperforms milestone prompt tuning methods such as CoOp, CoCoOp, and CLIP-Adapter across 11 datasets. In the fundus modality, our method surpasses all recently proposed advanced prompt tuning methods, such as PromptSRC, DePT, and CoPrompt. Notably, on the challenging ODIR dataset, our method outperforms the CoPrompt method, which already exhibits high performance, by 0.9% in accuracy, 2.3% in F1 score, 1.0% in AUC score, and 1.9% in kappa score. A similar trend is observed in the dermatoscope modality with the DermaMNIST, Derm7pt, and ISIC datasets. Across these three datasets, our method exceeds the DePT method, which is built on top of the soild PromptSRC method, by 3.5% in accuracy and 2.0% in AUC score on average. Moreover, in the ultrasound modality on the Fetal-US dataset, our method shows a performance improvement of 1.4%, 2.0%, 0.2%, and 1.8% in each metric over the second-best DePT method. Furthermore, on the category-balanced Kvasir dataset of the endoscope modality, our method outperforms the CoPrompt method by 1.6% in accuracy, 1.6% in F1-score, 0.1% in AUC score, and 1.8% in kappa score. X-ray images present a particular challenge for our method. While

TABLE III: Comparison with state-of-the-art methods on 5 datasets including Fetal-US, ISIC 2018, Kvasir, ODIR, and Pneumonia. The best and second best results are highlighted in red and blue . "Param." is the number of trainable parameters.

Method	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
CoOp [11]	81.4 ± 0.9	76.9 ± 1.0	96.9 ± 0.3	76.0 ± 1.1
CoCoOp [13]	85.4 ± 0.4	82.3 ± 0.8	97.8 ± 0.1	81.4 ± 0.5
KgCoOp [31]	64.0 ± 0.2	49.4 ± 0.5	89.3 ± 0.2	52.4 ± 0.4
CLIP-Adapter [14]	78.2 ± 0.2	66.3 ± 0.4	95.3 ± 0.1	71.7 ± 0.2
RPO [32]	88.1 ± 0.8	86.1 ± 0.8	98.5 ± 0.1	84.8 ± 1.0
LASP [36]	88.4 ± 0.3	86.0 ± 0.3	98.6 ± 0.1	85.2 ± 0.3
PromptSRC [18]	91.4 ± 0.4	90.4 ± 0.3	99.2 ± 0.1	89.2 ± 0.5
TCP [33]	66.8 ± 0.4	53.0 ± 0.4	91.5 ± 0.1	56.3 ± 0.5
AAPL [34]	83.9 ± 0.7	79.5 ± 0.8	97.4 ± 0.2	79.3 ± 0.9
PLOT++ [55]	86.6 ± 0.3	82.4 ± 0.4	98.3 ± 0.0	82.8 ± 0.3
DCPL [41]	91.9 ± 0.3	91.9 ± 0.4	99.3 ± 0.1	90.8 ± 0.4
ViP [56]	83.8 ± 0.2	80.5 ± 0.6	97.3 ± 0.0	79.2 ± 0.3
XCoOp [40]	90.3 ± 0.6	89.0 ± 0.6	98.9 ± 0.0	87.7 ± 0.7
DePT [35]	92.0 ± 0.5	90.7 ± 0.9	99.2 ± 0.0	89.8 ± 0.6
CoPrompt [21]	88.9 ± 1.2	87.4 ± 1.4	98.8 ± 0.1	85.8 ± 1.6
CILMP (Ours)	93.4 ± 0.2	92.7 ± 0.2	99.4 ± 0.0	91.6 ± 0.2

(a) Fetal-US

Method	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
CoOp [11]	75.5 ± 0.5	53.8 ± 1.2	94.0 ± 0.3	54.7 ± 1.0
CoCoOp [13]	77.8 ± 0.2	59.8 ± 0.2	95.1 ± 0.2	60.2 ± 0.3
KgCoOp [31]	66.9 ± 0.3	27.0 ± 0.5	83.4 ± 0.4	31.2 ± 0.9
CLIP-Adapter [14]	73.3 ± 0.3	41.6 ± 0.6	91.7 ± 0.2	50.3 ± 0.6
RPO [32]	80.4 ± 0.5	62.4 ± 2.4	95.5 ± 0.2	65.1 ± 0.9
LASP [36]	79.9 ± 0.4	62.8 ± 1.5	95.4 ± 0.2	64.1 ± 0.9
PromptSRC [18]	81.3 ± 0.5	63.6 ± 2.2	95.8 ± 0.1	66.0 ± 1.1
TCP [33]	69.4 ± 0.1	30.5 ± 0.4	86.7 ± 0.3	38.5 ± 0.1
AAPL [34]	76.4 ± 0.5	56.7 ± 0.2	93.5 ± 0.7	56.9 ± 0.3
PLOT++ [55]	80.7 ± 0.3	64.8 ± 1.3	95.9 ± 0.0	65.8 ± 0.5
DCPL [41]	85.0 ± 0.8	74.2 ± 1.7	97.4 ± 0.2	74.0 ± 1.4
ViP [56]	77.2 ± 0.2	58.3 ± 0.8	94.5 ± 0.1	58.8 ± 0.4
XCoOp [40]	82.9 ± 0.2	71.3 ± 0.8	96.8 ± 0.0	70.2 ± 0.3
DePT [35]	83.6 ± 0.5	65.7 ± 1.2	95.9 ± 0.3	70.6 ± 0.2
CoPrompt [21]	79.0 ± 0.3	62.0 ± 0.7	96.0 ± 0.1	62.3 ± 1.0
CILMP (Ours)	86.9 ± 0.4	77.4 ± 1.0	97.9 ± 0.0	77.5 ± 0.8

(b) ISIC 2018

Method	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
CoOp [11]	84.3 ± 0.8	84.2 ± 0.8	98.7 ± 0.0	82.0 ± 0.9
CoCoOp [13]	86.5 ± 0.2	86.5 ± 0.2	99.0 ± 0.0	84.5 ± 0.3
KgCoOp [31]	74.3 ± 0.1	74.5 ± 0.1	97.0 ± 0.1	70.6 ± 0.1
CLIP-Adapter [14]	83.6 ± 0.3	83.5 ± 0.3	98.6 ± 0.0	81.2 ± 0.4
RPO [32]	87.3 ± 0.4	87.3 ± 0.4	99.1 ± 0.1	85.5 ± 0.5
LASP [36]	89.1 ± 0.2	89.1 ± 0.2	99.2 ± 0.0	87.6 ± 0.2
PromptSRC [18]	91.6 ± 0.2	91.6 ± 0.2	99.4 ± 0.1	90.5 ± 0.2
TCP [33]	78.5 ± 0.5	78.6 ± 0.5	97.7 ± 0.1	75.4 ± 0.6
AAPL [34]	85.8 ± 0.5	85.8 ± 0.5	98.9 ± 0.0	83.8 ± 0.5
PLOT++ [55]	89.2 ± 0.2	89.2 ± 0.2	99.3 ± 0.0	87.7 ± 0.2
DCPL [41]	92.4 ± 0.3	92.4 ± 0.3	99.6 ± 0.0	91.4 ± 0.3
ViP [56]	86.5 ± 0.3	86.5 ± 0.2	98.9 ± 0.0	84.5 ± 0.3
XCoOp [40]	90.9 ± 0.3	90.9 ± 0.4	99.5 ± 0.0	89.6 ± 0.4
DePT [35]	91.0 ± 0.5	90.9 ± 0.5	99.4 ± 0.0	89.7 ± 0.6
CoPrompt [21]	91.6 ± 0.4	91.6 ± 0.4	99.5 ± 0.0	90.4 ± 0.5
CILMP (Ours)	93.2 ± 0.1	93.2 ± 0.1	99.6 ± 0.0	92.2 ± 0.1

(c) Kvasir

Method	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
CoOp [11]	56.6 ± 0.3	37.2 ± 1.3	81.3 ± 0.5	29.0 ± 0.4
CoCoOp [13]	56.0 ± 0.6	37.3 ± 2.4	80.6 ± 0.7	27.8 ± 1.9
KgCoOp [31]	49.6 ± 0.3	21.2 ± 1.0	69.5 ± 0.1	11.8 ± 0.3
CLIP-Adapter [14]	49.9 ± 0.4	23.5 ± 1.1	73.2 ± 0.7	13.5 ± 0.9
RPO [32]	59.2 ± 0.6	44.9 ± 0.5	83.5 ± 0.3	35.8 ± 0.3
LASP [36]	60.6 ± 0.5	45.6 ± 1.2	83.6 ± 0.7	38.9 ± 1.2
PromptSRC [18]	60.3 ± 1.9	41.5 ± 4.2	83.9 ± 0.5	36.3 ± 3.7
TCP [33]	51.4 ± 0.3	25.7 ± 0.7	73.7 ± 0.3	16.2 ± 0.6
AAPL [34]	56.1 ± 0.6	37.1 ± 0.5	80.3 ± 0.1	28.5 ± 1.0
PLOT++ [55]	58.7 ± 0.6	44.9 ± 0.9	84.2 ± 0.2	37.3 ± 0.9
DCPL [41]	66.6 ± 0.8	55.7 ± 1.3	89.7 ± 0.4	49.5 ± 1.2
ViP [56]	57.4 ± 0.1	38.3 ± 0.8	81.3 ± 0.5	30.8 ± 0.5
XCoOp [40]	62.9 ± 0.5	50.0 ± 0.3	87.2 ± 0.2	42.7 ± 0.6
DePT [35]	62.5 ± 0.5	44.2 ± 3.2	85.2 ± 0.7	42.2 ± 1.8
CoPrompt [21]	67.6 ± 0.7	55.6 ± 0.3	89.7 ± 0.6	51.5 ± 0.8
CILMP (Ours)	68.5 ± 0.9	57.9 ± 1.3	90.7 ± 0.3	53.4 ± 1.4

(d) ODIR

Method	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
CoOp [11]	75.9 ± 0.6	74.6 ± 1.0	90.4 ± 0.2	62.6 ± 0.9
CoCoOp [13]	77.8 ± 1.3	76.3 ± 2.0	91.4 ± 0.6	65.6 ± 2.2
KgCoOp [31]	63.6 ± 0.8	61.7 ± 0.6	79.0 ± 0.5	42.9 ± 1.2
CLIP-Adapter [14]	72.1 ± 0.5	69.3 ± 0.8	88.3 ± 0.1	56.2 ± 0.9
RPO [32]	85.1 ± 0.5	83.8 ± 0.7	95.0 ± 0.3	77.2 ± 0.8
LASP [36]	84.5 ± 0.3	83.3 ± 0.2	94.5 ± 0.1	76.2 ± 0.4
PromptSRC [18]	84.5 ± 1.0	83.6 ± 1.0	95.2 ± 0.4	76.6 ± 1.4
TCP [33]	67.2 ± 0.1	65.2 ± 0.4	82.4 ± 0.3	48.5 ± 0.3
AAPL [34]	77.2 ± 1.4	75.6 ± 1.6	91.1 ± 0.5	64.7 ± 2.3
PLOT++ [55]	86.9 ± 0.4	85.6 ± 0.4	95.8 ± 0.1	79.8 ± 0.5
DCPL [41]	87.5 ± 0.2	86.1 ± 0.3	96.0 ± 0.3	80.8 ± 0.4
ViP [56]	79.0 ± 0.5	78.6 ± 0.5	89.7 ± 0.5	68.0 ± 0.7
XCoOp [40]	84.4 ± 0.7	83.2 ± 0.8	94.9 ± 0.3	76.0 ± 1.0
DePT [35]	85.0 ± 0.5	84.1 ± 0.5	95.5 ± 0.7	77.4 ± 0.8
CoPrompt [21]	80.1 ± 0.3	79.1 ± 0.5	93.2 ± 0.5	70.0 ± 0.4
CILMP (Ours)	89.5 ± 0.7	88.7 ± 0.7	96.7 ± 0.3	84.0 ± 1.0

(e) Pneumonia

Method	Param.	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
CoOp [11]	2K	77.0	63.5	91.6	61.1
CoCoOp [13]	35K	78.1	65.7	91.5	62.3
KgCoOp [31]	2K	68.7	47.6	84.6	42.7
CLIP-Adapter [14]	131K	74.9	57.9	89.0	56.9
RPO [32]	5K	81.6	71.5	93.9	69.2
LASP [36]	39K	81.1	69.9	92.9	68.1
PromptSRC [18]	46K	81.8	69.6	93.9	69.1
TCP [33]	332K	71.1	50.8	86.9	47.6
AAPL [34]	35K	77.8	65.3	91.2	62.1
PLOT++ [55]	14K	81.9	72.4	94.1	69.9
DCPL [41]	3.8M	84.5	76.7	95.1	74.1
ViP [56]	41.5M	78.8	65.9	91.2	63.1
XCoOp [40]	94K	82.6	74.3	94.4	70.8
DePT [35]	48K	83.2	72.5	94.6	72.2
CoPrompt [21]	4.7M	83.0	74.2	94.7	71.7
CILMP (Ours)	3.8M	86.2	79.1	96.0	77.2

(f) **Average over 11 datasets**

it surpasses other methods on the Pneumonia dataset by over 2.0% in accuracy, 2.6% in F1-score, 0.7% in AUC score, and 3.2% in kappa score, it lags behind on the CPN-X-ray dataset.

On CPN-X-ray, the CoPrompt method achieves the best performance with an accuracy of 97.7%, an F1-score of 97.8%, an AUC of 99.8%, and a kappa score of 96.5%. In comparison, our method underperforms CoPrompt by 1.5% in accuracy, 1.5% in F1-score, 0.2% in AUC score, and 1.2% in kappa score. This underperformance is attributed to CoPrompt's use of more data augmentations and the combination of different PEFT strategies, including both prompt tuning and adapter, which provide increased tuning flexibility. Nevertheless, these results highlight the competitiveness of our proposed method compared to recent prompt tuning methods.

We also compare CILMP with three prompt tuning methods specifically designed for the medical domain, including DCPL, ViP and XCoOp. Notably, for the DCPL method, MedSAM [57] is employed as the domain-specific feature encoder to align with the original paper. As shown in II and III, our CILMP outperforms these methods by significant margins across almost all datasets. For instance, on the ISIC dataset, CILMP exceeds DCPL, the second-best performing method in this modality, by 1.9%, 3.2%, 0.5%, and 3.5% in terms of accuracy, F1 score, AUC score, and Kappa score,

TABLE IV: Comparison with medical vision language foundation models. "Zero-Shot" means zero-shot performance, "FFT" means fully fine-tuning paradigm, which fine-tunes all parameters of the VLM. "PT" denotes the parameter-efficient fine-tuning. "Param." is the number of trainable parameters.

Setup	Method	Param.	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
***Dermatoscopy*:**						
Zero-Shot	PubMedCLIP [78]	-	10.1	6.4	57.7	2.3
	BioMedCLIP [79]	-	43.8	16.7	67.8	10.3
	MONET [9]	-	54.2	29.4	80.9	28.3
FFT	PubMedCLIP [78]	151M	85.7 ± 0.3	55.4 ± 0.5	94.8 ± 0.1	72.3 ± 0.9
	BioMedCLIP [79]	196M	87.5 ± 0.2	80.1 ± 1.1	98.0 ± 0.1	76.0 ± 0.5
	MONET [9]	428M	88.3 ± 1.1	79.3 ± 1.8	98.1 ± 0.2	77.6 ± 1.8
PT	**CILMP (Ours)**	3.8M	87.4 ± 0.3	77.7 ± 1.0	98.0 ± 0.0	76.0 ± 0.4
***Chest X-ray*:**						
Zero-Shot	PubMedCLIP [78]	-	39.4	27.3	61.7	8.4
	BioMedCLIP [79]	-	51.6	45.6	75.6	22.2
	MONET [9]	-	51.8	39.4	59.8	22.7
FFT	PubMedCLIP [78]	151M	90.5 ± 0.7	89.7 ± 0.7	97.1 ± 0.2	85.5 ± 1.0
	BioMedCLIP [79]	196M	90.4 ± 0.1	89.4 ± 0.1	97.2 ± 0.1	85.2 ± 0.2
	MONET [9]	428M	88.8 ± 0.5	87.8 ± 0.6	97.0 ± 0.3	82.8 ± 0.7
PT	**CILMP (Ours)**	3.8M	89.5 ± 0.7	88.7 ± 0.7	96.7 ± 0.3	84.0 ± 1.0

respectively, Furthermore, in the fundus modality, CILMP surpasses XCoOp by 5.6%, 7.9%, 3.5%, and 10.7% for each metric. These comparisons further highlight the effectiveness of the CILMP method in the field of medical vision-language model research.

In a nutshell, TABLE IIIf presents a more comprehensive comparison, which includes the average results across 11 datasets. As illustrated, our CILMP method consistently outperforms the second-best approach DCPL by 1.7%, 2.4%, 0.9%, and 3.1% in terms of accuracy, F1-score, AUC score, and Kappa score, respectively. The comparison with other knowledge-enhanced methods, such as KgCoOp, XCoOp and CoPrompt, also shows our method's superiority in the medical image classification tasks. Overall, these results indicate that our method demonstrates superior performance across diverse data modalities and distributions, highlighting its enhanced capability to transfer pre-trained vision language models for medical image classification over existing state-of-the-art prompt tuning methods. Consequently, the experimental results affirm the effectiveness of the proposed method.

D. Comparison with Medical Vision Language Models

TABLE IV compares our method with more medical vision language models. Notably, the focus of our work is to design effective adaptation techniques to adapt VLM for downstream datasets. In contrast, existing medical VLMs aim to achieve strong transfer capabilities through extensive pre-training on large-scale datasets. Thus, these represent two orthogonal lines of study. The experiments show that even state-of-the-art VLMs pre-trained on massive medical datasets exhibit limited zero-shot performance on dermatoscopy and chest X-ray tasks (*e.g.*, MONET [9] achieves only 54.2% ACC and 80.9% AUC on dermatoscopy despite in-domain pretraining), underscoring the necessity of task-specific adaptation in clinical applications. Second, when fully fine-tuned, these VLMs can achieve strong performance (*e.g.*, 88.3% ACC for MONET on dermatoscopy), but at the cost of updating a large number of parameters (*e.g.*, 428M for MONET). In contrast, our method with merely 3.8M trainable parameters delivers competitive results. For example, on chest X-ray, CILMP attains 89.5% ACC, trailing fully fine-tuned PubMedCLIP [78] and BioMedCLIP [79] by only about 1% while reducing the trainable parameters by 40–50x. Overall, CILMP not only outperforms zero-shot medical VLMs by large margins but also narrows the gap to the fully fine-tuning paradigm with extreme parameter efficiency, demonstrating its effectiveness in bridging general-purpose medical VLMs to domain-specific clinical tasks.

TABLE V: Ablation study on the effectiveness of each component in CILMP on the dermatoscopy modality. "RD" means the relationship descriptor.

Dataset	Strategy	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
DermaMN.	CILMP	87.43±0.33	77.73±0.97	98.03±0.05	76.03±0.45
	w/o RD	87.00±0.16	76.70±0.93	97.77±0.12	75.10±0.29
	w/o Conditional	86.63±0.29	76.30±1.27	97.80±0.16	74.70±0.57
	w/o Intervention	86.00±0.49	74.57±0.17	97.40±0.22	72.93±1.07
Derm7pt	CILMP	76.40±1.76	58.27±0.76	91.93±0.49	59.63±3.48
	w/o RD	75.43±0.56	56.83±1.11	91.40±0.29	58.67±0.82
	w/o Conditional	73.00±0.85	51.83±1.32	90.93±0.09	53.03±2.15
	w/o Intervention	70.13±1.43	47.23±1.96	87.03±0.39	47.83±3.53
ISIC 2018	CILMP	86.90±0.42	77.40±0.99	97.90±0.00	77.47±0.75
	w/o RD	86.27±0.75	75.80±1.13	97.83±0.09	76.00±1.35
	w/o Conditional	85.67±0.21	75.57±0.78	97.47±0.09	75.03±0.25
	w/o Intervention	85.33±0.66	74.93±1.51	97.47±0.09	74.33±1.23
Average	CILMP	83.58	71.13	95.95	71.04
	w/o RD	82.90	69.78	95.67	69.92
	w/o Conditional	81.77	67.90	95.40	67.59
	w/o Intervention	80.49	65.58	93.97	65.03

E. Ablation Study

1) Ablation study of CILMP components: We investigate the effectiveness of three key components in CILMP: the relationship descriptor (*w/o* RD), the conditional mechanism (*w/o* Conditional), and the intervention function (*w/o* Intervention). The experimental results are presented in TABLE V. Experiments are conducted on the dermatoscope image modality using three datasets: DermaMNIST, Derm7pt, and ISIC 2018, chosen for their representativeness and challenging classification tasks due to multiple categories and varying training set scales. Firstly, removing the relationship descriptor (*w/o* RD in TABLE V) means the model merely conditions the intervention on the input image without injecting the matching prior. This leads to notable performance drops across each dataset. On average, the accuracy, F1 score, AUC, and Kappa decrease by 0.70%, 1.32%, 0.28%, and 1.12%, respectively. Removing the conditional mechanism (*w/o* Conditional in TABLE V) reverts to using the original intervention function (Eq. (11)) proposed in [16]. The exclusion of the conditional mechanism results in accuracy decreases of 0.80%, 3.40%, and 1.23% across the datasets. Similar declines are observed in the F1 and Kappa scores, each dropping by over 2%. Although the AUC score remains relatively stable, likely due to its already high value. Moreover, removing the intervention function Ψ (*w/o* Intervention) reduces the framework to merely projecting LLM representations into the prompt dimension and concatenating them to form class-specific prompts for the VLM. The

TABLE VI: Ablation study on the effectiveness of each component in CILMP on the fundus modality.

Dataset	Strategy	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
ADAM	CILMP	90.17±0.47	85.47±0.62	93.53±0.34	70.97±1.21
	w/o RD	89.67±0.47	85.17±0.47	94.13±0.75	70.40±0.99
	w/o Conditional	89.23±0.29	85.03±0.25	93.50±0.14	70.07±0.50
	w/o Intervention	88.15±1.35	83.07±1.58	92.30±0.21	66.15±3.15
APTOS	CILMP	84.47±0.46	66.67±0.76	94.63±0.21	75.73±0.71
	w/o RD	83.87±0.61	65.67±1.52	94.43±0.05	74.83±1.11
	w/o Conditional	83.27±0.37	62.53±0.42	94.63±0.05	73.90±0.57
	w/o Intervention	82.17±0.62	59.73±1.48	94.17±0.12	71.90±0.99
ODIR	CILMP	68.47±0.87	57.93±1.32	90.67±0.31	53.43±1.41
	w/o RD	67.80±1.30	57.45±1.15	89.70±0.30	52.65±2.05
	w/o Conditional	67.37±0.57	56.67±0.76	89.83±0.29	51.27±0.58
	w/o Intervention	65.97±1.71	55.37±2.98	89.23±1.16	48.63±3.50
Average	CILMP	81.04	70.02	93.61	66.71
	w/o RD	80.45	69.43	92.75	65.96
	w/o Conditional	79.96	66.74	92.65	65.08
	w/o Intervention	78.76	66.06	91.90	62.23

results show that this degradation underperforms our CILMP method by an average of 3.09%, 5.55%, 1.98%, and 5.01% in each metric across three datasets, respectively. To further demonstrate the generalizability of the proposed components across diverse medical imaging domains, we then conduct ablation study on the fundus modality, which comprises 3 datasets including ADAM, APTOS and ODIR. The results are shown in TABLE VI. As can be seen, the omission of the conditional mechanism results in a performance decrease of 3.28%, whereas the absence of the proposed intervention function leads to an average performance decline of 3.96% across the three datasets in terms of F1 score. In summary, these findings underscore the critical importance of the intervention function Ψ, the conditional mechanism, and the relationship descriptor in adapting LLM representations to the VLM space and generating class-specific, instance-adaptive prompts, thereby significantly enhancing the performance of the CILMP framework.

2) Ablation on the usage of LLM: Different from our CILMP method using LLM representations to incorporate medical domain knowledge during prompt tuning, previous studies [31, 21] commonly utilize plain text descriptions. In this plain text setting, the LLM generates a text description for each disease, which is then combined with the disease name to create a prompt for the VLM. To further explore the effectiveness of these two strategies, we conduct comparable experiments and report the results in TABLE VII. On average across three datasets, our representation-based method outperforms text format strategy by 1.80% in accuracy and 4.24% in F1 score. These comparisons demonstrate that using LLM representations in conjunction with our proposed intervention mechanism is a more effective way to incorporate medical domain knowledge into the prompt-tuning process.

3) Ablation of the prefix/suffix length of intervention: We conducted an ablation study to examine the sensitivity of our method to the prefix and suffix lengths (from 2 to 16) for LLM representations, setting $L_{\text{prefix}} = L_{\text{suffix}}$ for simplicity. As shown in TABLE VIII, optimal performance is achieved with an intervention length of 4. Consequently, we set $L_{\text{prefix}} = L_{\text{suffix}} = 4$ across all datasets. However, it is important to note

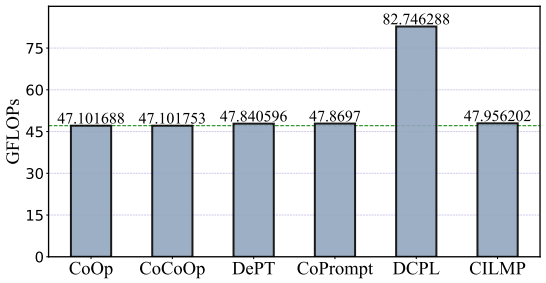


Fig. 4: Comparison of our CILMP and the other competitive prompt tuning methods in floating point of operations (FLOPs). Comparison is conducted on the ADAM dataset.

that this may not be optimal for every dataset and could be further tuned to improve performance.

4) Ablation of the dimensionality of the linear subspace: We conducted experiments on various dimensionalities of the linear subspace (1, 4, 8, and 16 dimensions) and found that a dimensionality of 8 achieves the best performance, as shown in TABLE IX. Consequently, we set the linear subspace dimensionality to 8 for all other settings.

5) Ablation of the prompt length: We also conducted experiments on the prompt length (1, 4, 8, and 16 context tokens). As shown in TABLE X, a length of 4 yields optimal performance with low sensitivity to parameter changes. Consequently, we set the prompt length to 4 for all other settings.

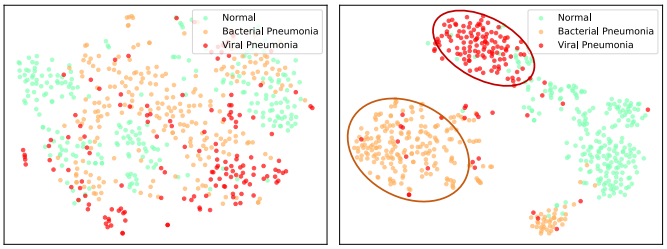
F. Efficiency Analysis

1) Trainable Parameter Statistics: TABLE IIIf presents the number of trainable parameters for each prompt tuning method. In alignment with established practices in the literature, these methods employ the ViT-B/16 backbone of CLIP. CoPrompt contributes an additional 4.7M parameters, DCPL adds 3.8M, and ViP incorporates 41.5M parameters. In contrast, our proposed CILMP method introduces 3.8M parameters, yet it achieves better performance compared to these approaches. Moreover, compared to the whole CLIP that consists of 149.6M parameters, our method introduces approximately 2.5% of trainable parameters relative to the complete VLM. Compared to the LLM used, it introduces only about 0.0475% additional parameters.

2) FLOPs Statistics: We also compare the computing efficiency between CILMP and other competitive prompt tuning methods in terms of floating point of operations (FLOPs). The results are shown in Fig. 4. Note that CoOp does not introduce computational burden compared to the original CLIP, so it can be viewed as a baseline for other methods. Compared to CoOp, DCPL incurs a significant increase of 35.6445 GFLOPs, primarily because it requires forwarding MedSAM during both training and inference. In contrast, our CILMP only adds 0.8545 GFLOPs, which is on par with other methods like CoCoOp, DePT, and CoPrompt, but it delivers significantly better performance. Although CILMP involves jointly fine-tuning an LLM and a VLM, it does not require any forward or backward calculations through the LLM during training and inference, because CILMP operates solely on the extracted fixed representations from the LLM. Therefore, these comparisons highlight the efficiency of our approach.

TABLE VII: Ablation study on the usage of large language model. "DermaMN" means the DermaMNIST dataset. "Average" means the average results across three datasets of the dermatoscope modality.

Dataset	Strategy	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
DermaMN	Text	86.33±0.68	74.47±1.01	97.67±0.17	73.87±1.31
	Representation	87.43±0.33	77.73±0.97	98.03±0.05	76.03±0.45
Derm7pt	Text	73.83±0.80	51.93±1.27	91.33±0.83	54.60±1.90
	Representation	76.40±1.76	58.27±0.76	91.93±0.49	59.63±3.48
ISIC 2018	Text	85.17±0.21	74.27±0.39	94.93±0.88	74.10±0.29
	Representation	86.90±0.42	77.40±0.99	97.90±0.00	77.47±0.75
Average	Text	81.78	66.89	95.31	67.52
	Representation	83.58	71.13	95.95	71.04



(a) CoOp (b) CILMP (Ours)

Fig. 5: Qualitive analysis based on T-SNE [80] visualization. Compared to conventional prompt tuning, CILMP generates features that are more discriminative across classes.

G. Qualitive Analysis

Fig. 5 presents a qualitative analysis of the feature space using t-SNE visualization [80]. We specifically examine the subtle differences in visual features between two challenging disease types: bacterial pneumonia and viral pneumonia. The results indicate that traditional prompt tuning leads to ambiguous feature representation for these categories. This ambiguity arises because traditional methods rely solely on textual class names for differentiation, while bacterial and viral pneumonia share significant textual similarities. In contrast, our CILMP method demonstrates enhanced class-wise discrimination between the two categories, due to the incorporation of disease-adaptive prompts. This underscores the superiority of our approach. However, the figure also reveals the presence of some outliers. For instance, several examples from the viral pneumonia class are found within the cluster of bacterial pneumonia. This shows a potential limitation of our method. Modeling at the image feature level could be a viable solution to address the failure cases, and this lies beyond the scope of our current work and will be explored in future research.

V. Discussions

A. Employing more Large Language Models

As large language models are crucial for integrating disease-specific knowledge into VLMs, we assess the performance of more open-source LLMs, including LLaMA-2-7B [43], LLaMA-3-8B [44], OpenBioLLM-8B [81] and MedLLaMA3-V20-8B [82], with results presented in TABLE XI. OpenBioLLM-8B and MedLLaMA3-V20-8B are esteemed as among the most powerful open-source medical LLMs, according to the Open Medical LLM Leaderboard [83]. The results indicate that LLaMA3-8B consistently outperforms LLaMA2-7B across all metrics, with improvements of approximately 0.4% in accuracy, 0.9% in F1 score, 0.1% in AUC score, and 0.8% in Kappa score. This suggests that more advanced LLMs enhance performance due to their sophisticated architecture and comprehensive pre-training.

Then, we discuss how an LLM pre-trained specifically on medical data could enhance the performance of our framework. The results in TABLE XI indicate that: **(i)** Medical domain LLMs generally enhance performance on both the ISIC and DermaMNIST datasets. This improvement stems from their specialized training on medical data, fostering a deeper comprehension of diseases and offering more pertinent prior knowledge in contrast to general LLMs, such as LLaMA3. Consequently, they generate more distinctive disease representations that aid in refining the visual language model prompt tuning. **(ii)** Integration of these medical domain LLMs elevates CILMP's performance on the CPN-X-ray dataset, where CILMP initially lags behind other methods by a considerable margin, boosting accuracy from 96.20% to 97.47% and F1 score from 96.30% to 97.53%. Furthermore, on the challenging ODIR dataset, incorporating MedLLaMA3-V20-8B leads to a notable 2.17% enhancement in F1 score. These findings underscore the capability of medical domain LLMs to enhance the quality of disease-specific representations, thereby augmenting the discriminability of prompts and diagnostic performance.

Additionally, these results demonstrate the flexibility of our CILMP framework, which allows seamless integration with various LLMs without requiring modifications to the architectures. This adaptability positions our approach to accommodate future advancements in LLMs, thus ensuring continuous improvements in applicability.

TABLE VIII: Ablation study on the length of prefix/suffix of the intervention process. We simply set the prefix length equal to the suffix length.

Length	2	4	8	16
ACC [%]	86.33±0.38	86.90±0.42	85.97±0.52	86.10±0.14
F1 [%]	77.07±0.05	77.40±0.99	77.53±0.21	75.70±0.28
AUC [%]	97.73±0.09	97.90±0.00	97.73±0.12	97.63±0.05
Kappa [%]	76.83±0.80	77.47±0.75	76.20±0.78	76.27±0.05

TABLE IX: Ablation study on the dimensionality of the linear subspace of the intervention function.

Dimension	1	4	8	16
ACC [%]	86.53±0.62	86.40±0.65	86.90±0.42	86.30±0.92
F1 [%]	77.03±0.50	76.00±1.94	77.40±0.99	77.60±0.75
AUC [%]	97.90±0.08	97.83±0.05	97.90±0.00	97.80±0.14
Kappa [%]	76.87±0.98	76.70±1.10	77.47±0.75	77.00±1.22

B. Sensitivity to data-efficient scenarios

In many real-world applications, computing resources and the availability of data or labels are often limited. Thus, we examine the impact of data-efficient training scenarios on the performance of our CILMP method by testing its sensitivity

TABLE X: Ablation study on the prompt length.

Length	1	4	8	16
ACC [%]	86.07±0.48	86.90±0.42	86.40±0.14	86.40±0.28
F1 [%]	75.83±1.23	77.40±0.99	77.17±0.33	76.63±0.19
AUC [%]	97.63±0.21	97.90±0.00	97.83±0.05	97.93±0.24
Kappa [%]	76.03±0.96	77.47±0.75	76.67±0.33	76.70±0.42

TABLE XI: Experimental comparison based on various Large Language Models (general & medical).

LLM	ACC [%]	F1 [%]	AUC [%]	Kappa [%]
ISIC:				
LLaMA2-7B [43]	86.47±0.54	76.53±1.13	97.83±0.09	76.70±0.96
LLaMA3-8B [44]	86.90±0.42	77.40±0.99	97.90±0.00	77.47±0.75
OpenBioLLM-8B [81]	87.37±0.48	77.50±1.45	97.93±0.09	78.37±0.93
MedLLaMA3-V20-8B [82]	87.33±0.09	78.63±0.24	97.80±0.14	78.33±0.25
DermaMNIST:				
LLaMA2-7B [43]	87.00±0.24	77.80±1.42	98.13±0.12	75.13±0.24
LLaMA3-8B [44]	87.43±0.33	77.73±0.97	98.03±0.05	76.03±0.45
OpenBioLLM-8B [81]	88.13±0.71	81.47±1.24	98.40±0.16	77.77±1.13
MedLLaMA3-V20-8B [82]	88.63±0.34	81.67±1.11	98.37±0.21	78.30±0.86
CPN-X-ray:				
LLaMA3-8B [44]	96.20±0.33	96.30±0.33	99.57±0.05	94.30±0.49
MedLLaMA3-V20-8B [82]	97.47±0.12	97.53±0.17	99.77±0.05	96.20±0.22
ODIR:				
LLaMA3-8B [44]	68.47±0.87	57.93±1.32	90.67±0.31	53.43±1.41
MedLLaMA3-V20-8B [82]	69.03±0.75	60.10±1.45	91.43±0.31	54.73±1.11

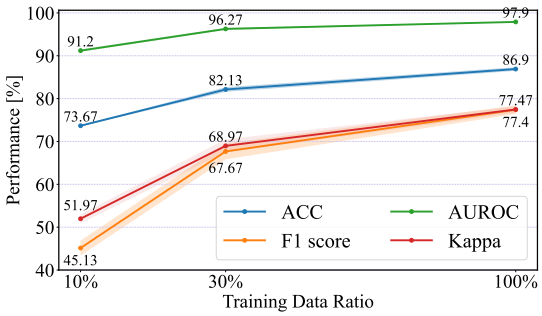


Fig. 6: Discussion on the sensitivity to data-efficient scenario.

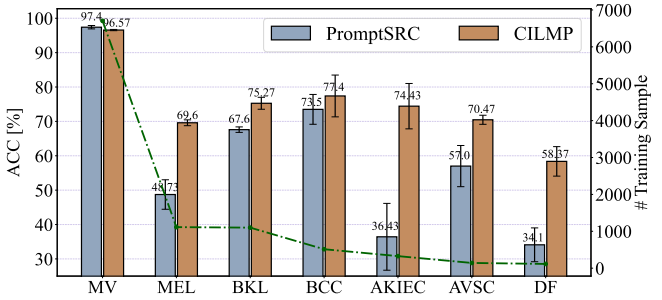


Fig. 7: Discussion on the sensitivity to the class imbalance issue. The accuracy [%] over three random runs is reported.

to different training data ratios. We conduct experiments using 10%, 30%, and 100% of the available labeled training data with results presented in Fig. 6. At a 10% label ratio, the framework achieves an accuracy of 73.67%, an F1 score of 45.13%, an AUC of 91.20%, and a Kappa score of 51.97%, indicating reasonable performance even with limited labeled data. Increasing the label ratio to 30% significantly enhances performance across all metrics. At the maximum label ratio of 100%, the framework achieves its highest performance. These results demonstrate that while the framework is robust in data-limited scenarios, it scales effectively with additional labeled data, achieving optimal performance when fully trained. This adaptability is crucial for practical applications where the availability of labeled data may vary, underscoring the framework's potential for deployment in real-world settings.

C. Sensitivity to class imbalance

Class imbalance is a prevalent issue in many real-world datasets, leading to biased models that perform poorly on minority classes. To evaluate the robustness of our CILMP framework to label imbalance, we assess its performance on the ISIC 2018 dataset, which has varying numbers of training samples per class. The results, presented in Fig. 7, compare our CILMP method against the PromptSRC method across seven classes: melanoma (MV), benign keratosis (MEL), benign keratosis (BKL), basal cell carcinoma (BCC), actinic keratosis (AKIEC), vascular lesion (AVSC), and dermatofibroma (DF). Despite the imbalance observed, CILMP consistently shows robust performance, particularly excelling in minority classes. For instance, CILMP achieves an accuracy of 77.40% for BCC, 74.43% for AKIEC, 70.47% for AVSC, and 58.37% for DF, outperforming PromptSRC in these categories. This robustness is attributed to CILMP's adaptive learning mechanisms, which generate instance-adaptive prompts that enhance learning on tail classes. In summary, the CILMP framework effectively handles class imbalance and limited labeled data, making it a versatile and reliable tool for real-world applications with uneven data distributions.

D. Limitations and Future Studies

Despite the potential advantages offered by the proposed CILMP methodology, it still has some limitations. While CILMP does exhibit a reduction in training parameters compared to the fully fine-tuning paradigm, its implementation necessitates a non-trivial allocation of computational resources to facilitate the inference of a large language model for knowledge extraction. This requirement may present formidable obstacles in contexts characterized by severe computational constraints. Additionally, the current research has primarily focused on the application of CILMP in 2D medical image classification, leaving its potential in other domains—such as 3D imaging or video-based diagnostics—largely unexplored. Investigating the framework's efficacy in these areas could unlock its broader applicability. Moreover, integrating advanced multi-modal fusion techniques, such as combining imaging data with patient and clinical information, presents an opportunity to improve the framework's adaptability and diagnostic performance in real-world clinical settings. For instance, leveraging patient histories, laboratory results, or genomic data alongside imaging could enable a more comprehensive diagnostic approach. Future studies should address these limitations to refine the CILMP framework, explore its applicability across diverse medical domains and modalities, and assess its performance in varied clinical settings to ensure its broader impact and generalizability.

VI. Conclusion

In this paper, we propose an LLM-based prompt tuning method for medical image analysis, leveraging medical knowledge in large language models to create disease-specific prompts for vision-language foundation models. An intervention function introduced bridges the LLM and VLM for knowledge transfer, while a conditional mechanism integrates the matching prior to generate instance-adaptive prompts. Extensive experiments across 11 diverse medical datasets demonstrate that our CILMP method outperforms recent state-of-the-art prompt tuning methods, showcasing its effectiveness.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021.

[2] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*, 2021.

[3] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, 2022.

[4] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, 2022.

[5] J. Yu, Z. Wang, V. Vasudevan *et al.*, "Coca: Contrastive captioners are image-text foundation models," *TMLR*, 2022.

[6] Y. Yang, W. Huang, Y. Wei, H. Peng, X. Jiang, H. Jiang, F. Wei, Y. Wang, H. Hu, L. Qiu *et al.*, "Attentive mask clip," in *IEEE InternationalConference on Computer Vision*, 2023.

[7] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, "A visual–language foundation model for pathology image analysis using medical twitter," *Nature medicine*, vol. 29, no. 9, pp. 2307–2316, 2023.

[8] X. Zhang, C. Wu, Y. Zhang, W. Xie, and Y. Wang, "Knowledge-enhanced visual-language pre-training on chest radiology images," *Nature Communications*, vol. 14, no. 1, p. 4542, 2023.

[9] C. Kim, S. U. Gadgil, A. J. DeGrave *et al.*, "Transparent medical image ai via an image–text foundation model grounded in medical literature," *Nature Medicine*, 2024.

[10] M. Christensen, M. Vukadinovic, N. Yuan, and D. Ouyang, "Vision–language foundation model for echocardiogram interpretation," *Nature Medicine*, pp. 1–8, 2024.

[11] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *IJCV*, vol. 130, no. 9, pp. 2337–2348, 2022.

[12] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[13] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[14] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *IJCV*, vol. 132, no. 2, pp. 581–595, 2024.

[15] N. Ding, Y. Qin, G. Yang *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, 2023.

[16] Z. Wu, A. Arora, Z. Wang *et al.*, "Reft: Representation finetuning for language models," *Advances in Neural Information Processing Systems*, 2024.

[17] Z. Han, C. Gao, J. Liu, S. Q. Zhang *et al.*, "Parameter-efficient fine-tuning for large models: A comprehensive survey," *arXiv preprint arXiv:2403.14608*, 2024.

[18] M. U. Khattak, S. T. Wasim, M. Naseer, S. Khan, M.-H. Yang, and F. S. Khan, "Self-regulating prompts: Foundational model adaptation without forgetting," in *IEEE InternationalConference on Computer Vision*, 2023, pp. 15 190–15 200.

[19] R. Dutt, L. Ericsson, P. Sanchez, S. A. Tsaftaris, and T. Hospedales, "Parameter-efficient fine-tuning for media: The missed opportunity," in *Medical Imaging with Deep Learning*, 2024.

[20] Z. Zheng, J. Wei, X. Hu, H. Zhu, and R. Nevatia, "Large language models are good prompt learners for low-shot image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28 453–28 462.

[21] S. Roy and A. Etemad, "Consistency-guided prompt learning for vision-language models," in *ICLR*, 2024.

[22] Z. Zhou, Y. Lei, B. Zhang, L. Liu, and Y. Liu, "Zegclip: Towards adapting clip for zero-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[23] Y. Li, Z.-Y. Li, Q.-S. Zeng, Q. Hou, and M.-M. Cheng, "Cascade-CLIP: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation," in *International Conference on Machine Learning*, 2024.

[24] C. Lian, H.-Y. Zhou, Y. Yu, and L. Wang, "Less could be better: Parameter-efficient fine-tuning advances medical vision foundation models," in *Medical Imaging with Deep Learning*, 2024.

[25] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Annual Meeting of the Association for Computational Linguistics*, 2021.

[26] Z. Zhong, D. Friedman, and D. Chen, "Factual probing is [mask]: Learning vs. learning to recall," in *NAACL*, 2021.

[27] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *EMNLP*, 2021.

[28] Y. He, S. Zheng, Y. Tay, J. Gupta, Y. Du, V. Aribandi, Z. Zhao, Y. Li, Z. Chen, D. Metzler *et al.*, "Hyperprompt: Prompt-based task-conditioning of transformers," in *International Conference on Machine Learning*, 2022.

[29] P. Liu, W. Yuan, J. Fu *et al.*, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, 2023.

[30] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4367–4375.

[31] H. Yao, R. Zhang, and C. Xu, "Visual-language prompt tuning with knowledge-guided context optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[32] D. Lee, S. Song, J. Suh *et al.*, "Read-only prompt optimization for vision-language few-shot learning," in *IEEE InternationalConference on Computer Vision*, 2023.

[33] H. Yao, R. Zhang, and C. Xu, "Tcp: Textual-based class-aware prompt tuning for visual-language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 438–23 448.

[34] G. Kim, S. Kim, and S. Lee, "Aapl: Adding attributes to prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1572–1582.

[35] J. Zhang, S. Wu, L. Gao, H. T. Shen, and J. Song, "Dept: Decoupled prompt tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 924–12 933.

[36] A. Bulat and G. Tzimiropoulos, "Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 232–23 241.

[37] M. Zhu, H. Li, H. Chen *et al.*, "Segprompt: Boosting open-world segmentation via category-level prompt learning," in *IEEE InternationalConference on Computer Vision*, 2023.

[38] X. Xie, J. Niu, X. Liu *et al.*, "A survey on incorporating domain knowledge into deep learning for media," *MedIA*, 2021.

[39] W. Yang, J. Zhao, Y. Qiang *et al.*, "DScGANS: Integrate Domain Knowledge in Training Dual-Path Semi-supervised Conditional Generative Adversarial Networks and S3VM for Ultrasonography Thyroid Nodules Classification," in *MICCAI*, 2019.

[40] Y. Bie, L. Luo, Z. Chen, and H. 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*. Springer, 2024, pp. 773–783.

[41] Q. Cao, Z. Xu, Y. Chen, C. Ma, and X. Yang, "Domain-controlled prompt learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 936–944.

[42] H. Touvron, T. Lavril, G. Izacard *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[43] H. Touvron, L. Martin, K. Stone *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[44] AI@Meta, "Llama 3 model card," 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[45] OpenAI, "Chatgpt," 2023. [Online]. Available: https://openai.com/chatgpt

[46] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen *et al.*, "Palm 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023.

[47] OpenAI, "Gpt-4 technical report," 2024. [Online]. Available: https://openai.com/gpt-4

[48] W.-L. Chiang, Z. Li, Z. Lin *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna.lmsys.org*, 2023.

[49] Y. Yang, A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch, and M. Yatskar, "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[50] J. Yang, C. Li, P. Zhang, B. Xiao, C. Liu, L. Yuan, and J. Gao, "Unified contrastive learning in image-text-label space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[51] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[52] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *TPAMI*, 2024.

[53] P. Shrestha, S. Amgain, B. Khanal *et al.*, "Medical vision language pretraining: A survey," *arXiv preprint arXiv:2312.06224*, 2023.

[54] M. M. Derakhshani, E. Sanchez, A. Bulat *et al.*, "Bayesian prompt learning for image-language model generalization," in *ICCV*, 2023.

[55] G. Chen *et al.*, "Plot: Prompt learning with optimal transport for vision-language models," in *International Conference on Learning Representations*, 2024.

[56] X. Fang, Y. Lin, D. Zhang, K.-T. Cheng, and H. Chen, "Aligning medical images with general knowledge from large language models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 57–67.

[57] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[58] J. Gu *et al.*, "A systematic survey of prompt engineering on vision-language foundation models," *arXiv preprint arXiv:2307.12980*, 2023.

[59] J. Xing, J. Liu, J. Wang *et al.*, "A survey of efficient fine-tuning methods for vision-language models—prompt and adapter," *Computers & Graphics*, 2024.

[60] A. Geiger, H. Lu, T. Icard, and C. Potts, "Causal abstractions of neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9574–9586, 2021.

[61] A. Geiger, Z. Wu, C. Potts, T. Icard, and N. Goodman, "Finding alignments between interpretable causal variables and distributed neural representations," in *Causal Learning and Reasoning*. PMLR, 2024, pp. 160–187.

[62] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *International conference on machine learning*. PMLR, 2019, pp. 3519–3529.

[63] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[64] J. Devlin *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Annual Meeting of the Association for Computational Linguistics*, 2019.

[65] P. Khosla, P. Teterwak, C. Wang *et al.*, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, 2020.

[66] J. Yang, R. Shi, D. Wei *et al.*, "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, 2023.

[67] J. Kawahara, S. Daneshvar, G. Argenziano, G. Hamarneh, and S.-P. Checklist, "Skin lesion classification using multitask multimodal neural nets," *IEEE JBHI*, 2019.

[68] N. Codella, V. Rotemberg, P. Tschandl *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration," *arXiv preprint arXiv:1902.03368*, 2019.

[69] H. Fang, F. Li, H. Fu *et al.*, "Adam challenge: Detecting age-related macular degeneration from fundus images," *IEEE TMI*, 2022.

[70] S. D. Karthik, Maggie, "Aptos 2019 blindness detection," 2019. [Online]. Available: https://kaggle.com/competitions/aptos2019-blindness-detection

[71] Larxel, "Ocular disease recognition," 2020. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k

[72] X. Burgos-Artizzu, D. Coronado-Gutiérrez, B. Valenzuela-Alcaraz *et al.*, "Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes," *Nature Scientific Reports*, 2020.

[73] C. Zhu, W. Chen, T. Peng *et al.*, "Hard sample aware noise robust learning for histopathology image classification," *IEEE TMI*, 2021.

[74] K. Pogorelov, K. R. Randel, C. Griwodz *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *ACM MM*, 2017.

[75] S. Shastri, I. Kansal, S. Kumar, K. Singh, R. Popli, and V. Mansotra, "Cheximagenet: a novel architecture for accurate classification of covid-19 with chest x-ray digital images using deep convolutional neural networks," *Health and technology*, vol. 12, no. 1, pp. 193–204, 2022.

[76] D. S. Kermany, M. Goldbaum, W. Cai *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, 2018.

[77] A. Radford, J. Wu, R. Child *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, 2019.

[78] S. Eslami, C. Meinel, and G. De Melo, "Pubmedclip: How much does clip benefit visual question answering in the medical domain?" in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 1151–1163.

[79] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri *et al.*, "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023.

[80] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[81] M. S. Ankit Pal, "Openbiollms: Advancing open-source large language models for healthcare and life sciences," https://huggingface.co/aaditya/Llama3-OpenBioLLM-8B, 2024.

[82] YonseiMAILab, "Medllama3-v20," https://huggingface.co/ProbeMedicalYonseiMAILab/medllama3-v20, 2024.

[83] OpenLifeScienceAI, "Open medical-llm leaderboard," https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard, 2025.