Disease-informed Adaptation of Vision-Language Models

Jiajin Zhang, Ge Wang *Fellow, IEEE*, Mannudeep K. Kalra, Pingkun Yan* *Senior Member, IEEE*

***Abstract*— Expertise scarcity and high cost of data annotation hinder the development of artificial intelligence (AI) foundation models for medical image analysis. Transfer learning provides a way to utilize the off-the-shelf foundation models to address the clinical challenges. However, such models encounter difficulties when adapting to new diseases not presented in their original pre-training datasets. Compounding this challenge is the limited availability of example cases for a *new* disease, which further leads to the poor performance of the existing transfer learning techniques. This paper proposes a novel method for transfer learning of foundation Vision-Language Models (VLMs) to efficiently adapt them to a *new* disease with only a few examples. Such an effective adaptation of VLMs hinges on learning the nuanced representation of *new* disease concepts. By capitalizing on the joint visual-linguistic capabilities of VLMs, we introduce disease-informed contextual prompting in a novel disease prototype learning framework, which enables VLMs to quickly grasp the concept of the *new* disease, even with limited data. Extensive experiments across multiple pre-trained medical VLMs and multiple tasks showcase the notable enhancements in performance compared to other existing adaptation techniques. The code will be made publicly available at `https://github.com/RPIDIAL/Disease-informed-VLM-Adaptation`.**

***Index Terms*— Vision-Language Model, Foundation Model, Transfer Learning, Model Adaptation, New Disease, COVID-19.**

I. Introduction

IN medical image analysis, developing large artificial intelligence (AI) models at local clinical sites is often impeded by the scarcity of expertise and the high costs associated with data annotation [1], [2]. Against this backdrop, transfer learning [3], [4] has emerged as a crucial strategy. It leverages existing pre-trained foundation models, developed on large, publicly available datasets, to adapt to the specific needs of data-limited local clinical sites. This scheme effectively addresses the challenges of limited local expertise and the cost of annotating medical data, enabling more efficient deployment of advanced diagnostic technologies [5]–[7]. Among the various foundation models, Vision-Language Models (VLMs) stand out as a distinct category integrating both visual and linguistic information. This multimodal nature leads to good generalizability of VLMs, making them particularly valuable in scenarios marked with data scarcity, a frequent challenge in medical settings, especially at local clinical sites. Recent advancements in pre-training Vision-Language Models (VLMs), such as Contrastive LanguageImage Pre-training (CLIP) [8]–[14], have demonstrated impressive domain adaptation capabilities. Integrating natural language processing allows foundation models to learn better visual representations with an appropriate alignment with textual concepts than using image data alone. However, efficiently improving the generalizability of medical VLMs on unseen diseases with few image samples still requires the development and implementation of novel techniques in transfer learning.

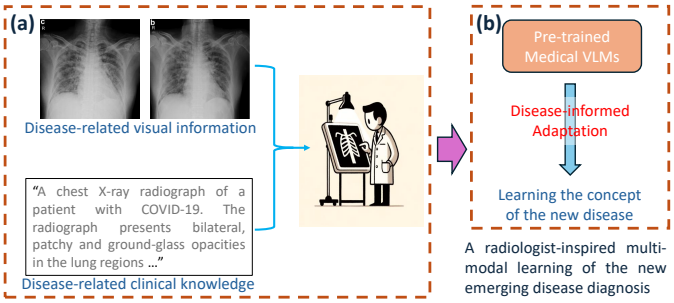

Fig. 1. The framework overview. **(a)** DiCoP produces prompts informed by specific diseases. **(b)** DPL enables learning disease representations with limited data.

In this work, we aim to tackle the challenge of adapting VLMs for computer-assisted diagnosis using medical images. Specifically, we focus on enabling pre-trained medical VLMs to understand a new disease that is entirely absent from the pre-training dataset. Several methods have been proposed for transfer learning of VLMs. Adapter-based approaches, such as linear probing [8], CLIP-Adapter [15], and Tip-Adapter [16], adopt a classifier to tailor pre-trained visual encoders on new tasks, by only tweaking the final layers. Alternatively, prompting-based methods, like CoOp [17], CoCoOp [6], and KgCoOp [18], focus on optimizing learnable prompts without actual text inputs, prioritizing performance over the acquisition of meaningful concepts. The recent method MaPLe [7] utilizes dual-modality prompts to adapt CLIP to new domains, yet it demands extensive data for fine-tuning and is not suitable for diseases with limited data. More importantly, recent works

Asterisk indicates corresponding author.
J. Zhang, G. Wang and P. Yan are with the Department of Biomedical Engineering and Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180, USA.
M. K. Kalra is with the Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.
This research was partially supported by the National Science Foundation (NSF) under the CAREER award OAC 2046708.

on transfer learning on VLMs [19], [20] showed that large-scale VLMs excel at recognizing common objects but may struggle with visual concepts that rarely appeared in their pre-training data. This observation indicates that the efficacy of VLMs could diminish when the relevant categories are only sparsely represented in the training data. This problem can often occur in medical imaging and image analysis, especially when adapting existing models to a *newly identified* disease with only few samples, for instance, diagnosing COVID-19 with existing models during the early stage of the outbreak.

Drawing inspiration from how clinicians assimilate new medical discoveries with only limited clinical cases, we introduce a novel framework for disease-informed adaptation of pre-trained VLMs. As depicted in Figure 1, radiologists typically acquire knowledge about new diseases by analyzing information in two steps. They first observe disease-specific visual patterns in medical images. After that, they integrate these observations with clinical knowledge obtained from radiology reports and relevant literature. Our multimodal adaptation strategy seeks to replicate the learning process of clinicians by integrating both visual and linguistic information with disease-specific insights. More specifically, our research contends that effective representation learning of disease concepts is central to the success of adapting VLMs. The contributions of our work are four-fold as follows.

1) We consider a clinically significant adaptation task that leverages the visual-linguistic capabilities of VLMs to prepare for the diagnosis of a *newly identified* disease. In our work, we use the emergence of COVID-19 pneumonia as an exemplar use case to illustrate how the proposed scheme can enhance preparedness and response capabilities for future health crises.

2) We propose *Disease-informed Contextual Prompting* (DiCoP), a method that harnesses the clinical knowledge to craft prompts for representing the concept of the new disease, COVID-19 pneumonia. The prompts highlight the disease characteristics with descriptive attributes, such as texture, shape, and location. To overcome the problem that crafted prompts lack instance specifics, we further propose to enrich the textual prompts with instance-specific features derived from image context.

3) We introduce *Disease Prototype Learning* (DPL) to address the lack of structural regulation in the latent space of CLIP-based VLMs [21], which is critical for recognizing the new target diseases. The DPL framework fine-tunes the image encoder to actively learn the prototypes of diseases, and regularizes the geometric structure of the learned representations for the downstream visual recognition tasks.

4) Extensive experiments across multiple evaluation tasks demonstrate the effectiveness, efficiency, and generalizability of the proposed VLM adaptation framework.

A preliminary version of this work [5] was accepted for MICCAI 2024. This paper presents significant enhancements over our initial submission, including new technical developments, evaluations on additional datasets, and further numerical experiments across various evaluation tasks. These comprehensive additions contribute substantially to advancing the field of VLM adaptations in medical image analysis.

II. Related Works

A. Vision-Language Models

The convergence of computer vision and natural language processing has given rise to a new group of foundation models, namely Vision-Language Models (VLMs). VLMs marry visual and linguistic models to achieve cross-modal comprehension and reasoning capabilities. This integration has been pivotal in advancing tasks that require both visual understanding [22] and language reasoning [23], [24]. Groundbreaking models such as CLIP [8] and ALIGN [25] have further bridged the gap between language models and vision tasks, showcasing the feasibility of cross-modal applications. Since the introduction of CLIP in 2021, medical VLMs have also been developed, including but not limited to BioViL [11], MedCLIP [13], MGCA [12], CheXzero [14], LLaVA-Med [26], PLIP [10], and CONCH [27]. All these methods significantly outperform the corresponding vision-only models in medical domain, demonstrating the great potential of vision-language foundation models.

B. Adaptation of VLMs to Downstream Tasks

Two main strategies for adapting pre-trained Vision-Language Models (VLMs) to downstream tasks are adapter-based methods and prompt tuning methods. Adapter-based approaches like linear probing [8], CLIP-Adapter [15], and Tip-Adapter [16] utilize a classifier to adapt pre-trained visual encoders to new tasks by adjusting only the final layers. While this method of selective modification allows for efficient adaptation, it also introduces significant limitations. Primarily, by focusing changes on the final layers, these approaches may fail to leverage lower level nuanced features that could be crucial for complex tasks. This limitation may lead to suboptimal outcomes, requiring more comprehensive task-specific modifications. Furthermore, while minimal changes preserve the core architectures developed during initial training, they might not provide sufficient adaptability to accurately represent and capture the unique features of new tasks, potentially compromising the model's performance across various settings.

Prompt tuning techniques use task-related tokens to enhance task-specific knowledge. For example, filling various class names into the template of "`a photo of a [CLS]`" in CLIP [8] creates textual embeddings for zero-shot predictions. However, these static, hand-crafted prompts often fail to capture the complexities of specific tasks. To address this problem, CoOp [17] introduces learnable soft prompts from few-shot samples, which, however, do not vary across different instances of the same task. To remedy this, CoCoOp [6] provides image-conditional contexts for each image, enhancing textual prompts with visual contexts. Similarly, KgCoOp [18] adds standard language template ("`a photo of a [CLS]`") into the global learning of soft prompts. Despite these advancements, CoOp-based methods primarily focus on optimizing prompts rather than incorporating clinical knowledge for diagnosing *newly identified* diseases. This approach often prioritizes performance over meaningful conceptual understanding.

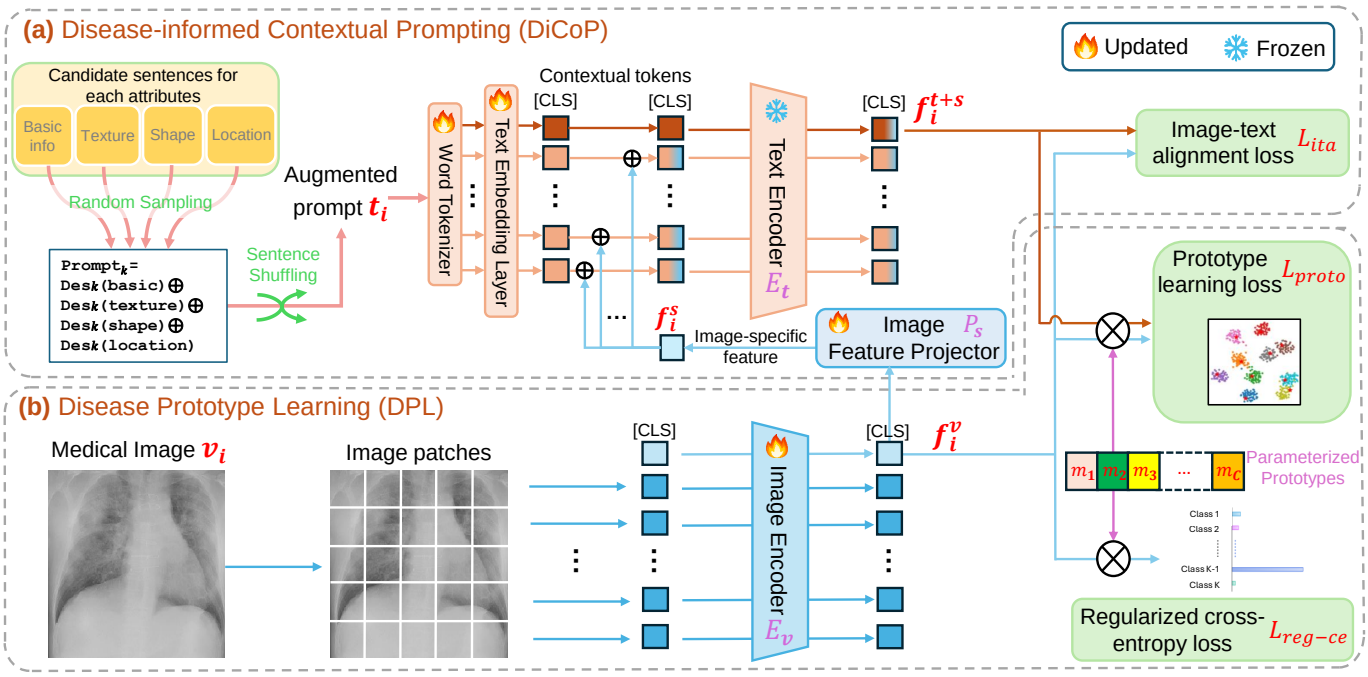

Fig. 2. Framework overview. **(a)** Disease-informed Contextual Prompting (DiCoP) produces prompts informed by specific diseases. **(b)** Disease Prototype Learning (DPL) enables learning disease representations with limited data.

A recently developed method, MaPLe [7], combines adapter-based and prompt tuning methods to enhance the adaptation of CLIP to new domains through dual-modality prompt tuning. While this approach introduces significant advancements, it requires substantial amount of data for fine-tuning. This poses a challenge for diseases with limited available data, restricting the applicability of such methods in scenarios where collecting extensive datasets is impractical or impossible.

C. Prototype Learning

The presented work in this paper leverages prototype learning to facilitate the representation learning of new diseases. The aim of prototype learning is to group similar items into a unified embedding as a prototype. van den Oord et al. illustrated that images can be represented in discrete forms [28], and Ramesh et al. established the efficacy of prototype learning for tasks involving cross-modality interactions [29]. Chen et al. developed a vision-language memory bank specifically for radiology report generation [30]. Unlike report generation models, VLMs must preserve detailed visual information during the pre-training phase to ensure precise representation of boundaries and lesions in subsequent tasks. We utilize prototype learning for sentence representation in medical reports, converting a continuous embedding space into a categorical one. Our method ensures the retention of both overarching concepts and intricate details, thereby maintaining structural integrity and fidelity at various levels.

D. Transfer Learning in Medical Image Analysis

Transfer learning has become a foundational element in the field of medical image analysis [3], [4], addressing the challenge of limited labeled data, which is a common obstacle in medical settings [1], [2]. This technique involves leveraging knowledge gained from one task or data domain and applying it to another related task or domain. In medical image analysis, where acquiring and annotating large datasets can be prohibitively expensive and time-consuming, transfer learning proves particularly beneficial [31], [32]. Models pre-trained on large datasets of general images can be fine-tuned with a relatively small set of local medical images to achieve substantial improvement in accuracy. This approach not only speeds up the training process but also enhances model generalizability, especially in tasks like disease diagnosis [33], organ segmentation [34], and tumor detection [35]. Recent studies have demonstrated the effectiveness of transfer learning in various applications [6], [7], [36], showcasing its versatility and potential to bridge the gap between data scarcity in medical domains and the need for highly accurate medical diagnostic tools.

III. Methods

Our method utilizes medical VLMs pre-trained with CLIP to diagnose *newly identified* diseases. Effective representation learning is key for VLMs to understand new or unseen diseases. Our proposed approach is inspired by the process depicted in Fig. 1, where radiologists learn about new diseases through aligning the disease-specific visual patterns with existing clinical knowledge from literature. Similarly, our method approaches the representation learning of new disease concepts in two steps. First, we introduce *Disease-informed Contextual Prompting* (DiCoP), which connects the concepts of new diseases with established clinical knowledge through textual prompts. These prompts emphasize the characteristics of a

disease using descriptive attributes, including texture, shape, and location. Then, we employ *Disease Prototype Learning* (DPL) to facilitate the representation learning of new disease categories. The goal of DPL is to consolidate cases with similar medical findings into a unified prototype embedding. This prototype embodies the abstract concept of a new disease, effectively representing its distinctive characteristics.

As shown in Fig. 2, a CLIP model comprises a text encoder $\boldsymbol{E}_t$ and a vision encoder $\boldsymbol{E}_v$ for image encoding. Let $\{(\boldsymbol{v}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ denote a labeled image dataset of target diseases. The images $\boldsymbol{v}_i$ are grouped into C categories, each for a distinct disease or condition. Our proposed DiCoP method first utilizes clinical knowledge, specifically from manually crafted short text prompts by radiologists with hands-on experience with the new disease. The image and text encoders, $\boldsymbol{E}_v$ and $\boldsymbol{E}_t$, encode each $(\boldsymbol{v}_i, \boldsymbol{t}_i)$ pair, using the `[CLS]` token to derive global visual and linguistic representations $\boldsymbol{f}_i^t, \boldsymbol{f}_i^v \in \mathbb{R}^{h\times 1}$, respectively. We **freeze** the pre-trained text encoder to prevent it from overfitting to the small set of hand-crafted prompts. The rest of this section presents the technical details.

A. Disease-informed Contextual Prompting (DiCoP)

Drawing on insights from the prior research on general image analysis [6], [19], [20], [37], we believe incorporating both *clinical knowledge* and *image-specific features* into contextual prompts is critical for enhancing the transferability of VLMs. The key of adapting pre-trained VLMs to new/unseen diseases is to link the concepts of new diseases with the existing clinical knowledge. For instance, radiological descriptions of texture attributes can aid in illustrating pneumonia. Through this approach, the concept of pneumonia is connected to the pre-existing radiology corpus via linguistic semantics by the text encoder $\boldsymbol{E}_t$. We thus propose creating contextual prompts with notable attributes of the target categories. In particular, we develop descriptive contextual prompts using the template

$$\begin{aligned}\texttt{Prompt}_k\big|_{k=1}^{C} = \texttt{Des}_k(\texttt{basic}) \oplus \texttt{Des}_k(\texttt{texture}) \\ \oplus \texttt{Des}_k(\texttt{location}) \oplus \texttt{Des}_k(\texttt{shape}),\end{aligned} \tag{1}$$

where $\oplus$ symbolizes the concatenation of descriptions $\texttt{Des}_k(\cdot)$ for the k_{th} disease category, covering each attribute (texture, location, shape). The three attributes recommended by radiologists are the most common and significant attributes for comprehensively establishing new medical concepts with clinical knowledge. These attributes selected are fundamental to the description and understanding of pathological features in radiographs. $\texttt{Des}_k(\texttt{basic})$ desribes critical identifiers like the type of disease or condition, which establishes the context. Texture and shape provide specific visual cues that are crucial for distinguishing different pathologies, while location helps pinpoint the exact area of interest within an organ, thus aiding in precise diagnostics. These are crucial for distinguishing between disease manifestations, essential for accurate diagnosis. We initialized the attribute descriptions $\texttt{Des}_k(\cdot)$ with the assistance of a radiologist who has over 15 years of experience in thoracic imaging. This step ensures that the generated prompts accurately reflect relevant clinical knowledge.

To preserve the linguistic capability of VLMs to process free-form text, and enhance contextual understanding [38], we prepared three candidate prompts with the same semantic meaning for each attribute by leveraging the prior text augmentation research [38], [39]. In this work, GPT-4 was employed to rephrase the manually crafted $\texttt{Des}_k(\cdot)$ into two additional augmented prompt variants, which were subsequently verified for correctness by the radiologist. All prompt candidates are reported in Appendix-A. To formulate $\texttt{Prompt}_k$, one $\texttt{Des}_k(\cdot)$ was randomly selected from the three candidates for each attribute. We further randomly shuffle the order of the selected elements $\texttt{Des}_k(\cdot)$, to increase the diversity of $\texttt{Prompt}_k$. This process significantly varies the text prompts for each sample within the same category, *i.e.*, $\boldsymbol{t}_i = \textit{Augmentation}(\texttt{Prompt}_k)$ if $i \in \boldsymbol{S}_k$.

Considering the visual diversity of images within the same category, we combine general clinical knowledge-based prompts with specific image variability to create more informative contextual prompts. As depicted in Fig. 2, an Image Feature Projector $\boldsymbol{P}_s$ is introduced to project a visual `[CLS]` token $\boldsymbol{f}_i^v$ to an image-specific representation $\boldsymbol{f}_i^s = \boldsymbol{P}_s(\boldsymbol{f}_i^v)$. The dimensionality of image-specific representation $\boldsymbol{f}_i^s$ matches the contextual tokens from the Text Embedding Layer as in Fig. 2. $\boldsymbol{f}_i^s$ is then added to each contextual token embedding, except for the `[CLS]` token. We denote the representation of text-only prompts as $\{\boldsymbol{f}_k^t\}_{i=1}^{N}$ and the representations enriched with image-specific features as $\{\boldsymbol{f}_i^{t+s}\}_{i=1}^{N}$. An image-text alignment loss L_{ita} is defined to encourage the model to maximize the cosine similarity between the matched images and prompts, while reducing the similarity between unmatched pairs

$$\begin{aligned}L_{\text{ita}} = \frac{1}{2N}\sum_{i=1}^{N}\Bigg[\frac{\exp(\boldsymbol{f}_i^v \cdot \boldsymbol{f}_i^{t+s}/\tau_1)}{\log\left(\sum_{j=1}^{N}\exp(\boldsymbol{f}_j^v \cdot \boldsymbol{f}_i^{t+s}/\tau_1)\right)} \\ + \frac{\exp(\boldsymbol{f}_i^v \cdot \boldsymbol{f}_i^{t+s}/\tau_1)}{\log\left(\sum_{j=1}^{N}\exp(\boldsymbol{f}_i^v \cdot \boldsymbol{f}_j^{t+s}/\tau_1)\right)}\Bigg],\end{aligned} \tag{2}$$

where τ_1 is a temperature hyperparameter.

B. Word Embedding for a New Disease

Our proposed adaptation approach is motivated by the need to handle *newly identified* diseases, a more challenging problem than previous VLM adaptation works [6], [7], [17], [18], [40], since neither the vision model nor the language model has encountered these diseases before. For example, considering the term of "`COVID-19`" disease, it should be recognized as a singular term that identifies the Coronavirus Disease 2019. Yet, not only is the language model unfamiliar with the concept of this new disease, but its tokenizer also splits "`COVID-19`" into four tokens: "`cov`", "`##id`", "`-`", and "`19`".

To address this challenge, we update both Word Tokenizer and Text Embedding Layer of the language model for VLMs as shown in Fig. 3. Specifically, for the word tokenizer, we treat "`COVID-19`" as a singular term, assigning it a new word token and ID. Taking ClinicalBERT tokenizer [41] as

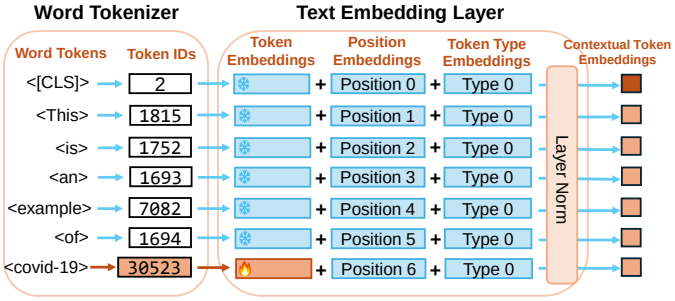

Fig. 3. Update of the Word Tokenizer and Text Embedding Layer of the language model in a given VLM.

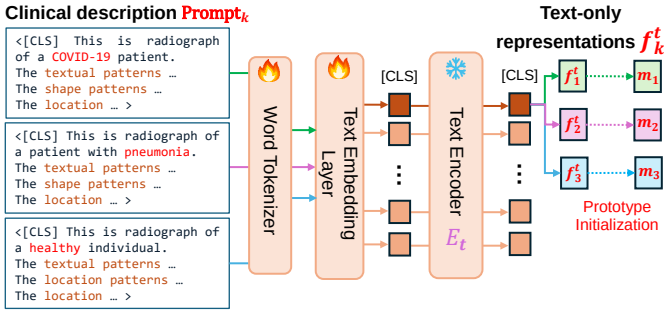

Fig. 4. The computation of text-only representations f_k^t and the initialization of the prototypes representations m_k.

an example, of which the original token IDs are between "1" and "30522", the new token ID of "COVID-19" would be "30523". In addition, we introduce a trainable text embedding vector to learn a representation of the newly defined "COVID-19" token ID. Throughout training, all other token embeddings remain fixed, except for the newly introduced trainable text embedding vector. This approach preserves the original linguistic semantics learned during VLM pre-training, while acquiring the word embeddings of a new disease.

C. Disease Prototype Learning (DPL)

With both DiCoP generated prompts and a limited number of new image samples, we are able to finetune medical VLMs to learn representations of new diseases. However, the original CLIP-based VLMs do not impose geometric constraints on the representations within each modality. For example, mismatched samples from the same category could not be explicitly differentiated in the latent space. This intra-modality inconsistency [21] may limit the transfer of VLMs to downstream classification tasks, such as disease diagnosis. To address the challenge, we propose to learn explicit representations of each disease category.

More specifically, we employ a set of trainable vectors $\{\boldsymbol{m}_k\}_{i=1}^{C}$ to represent C distinct disease prototypes. As presented in Fig. 4, each prototype $\boldsymbol{m}_k$ is initialized by the text-only representation $\boldsymbol{f}_k^t$, encoded from clinical description $\texttt{Prompt}_k$ without any image-specific features. We then simultaneously finetune the model and learn the prototypes by minimizing the cosine similarity between samples grouped under the same category and their respective prototype $\boldsymbol{m}_k$. Meanwhile, we maximize the separability between different disease prototypes. The total prototype learning loss L_{proto} is defined as

$$L_{proto} = \frac{1}{C}\sum_{k=1}^{c}\frac{1}{|2S_k|}\sum_{\boldsymbol{i}\in S_k}\left[\exp\left(\frac{\boldsymbol{f}_i^v\cdot\boldsymbol{m}_k}{\tau_2}\right)+\exp\left(\frac{\boldsymbol{f}_i^{t+s}\cdot\boldsymbol{m}_k}{\tau_2}\right)\right] - \lambda_1\cdot\frac{2}{k(k-1)}\sum_{k\neq j}\exp\left(\frac{\boldsymbol{m}_k\cdot\boldsymbol{m}_j}{\tau_2}\right), \tag{3}$$

where τ_2 is another temperature hyperparameter and λ_1 is a weighting factor.

Given the limited samples for adaptation, the learned disease prototypes $\boldsymbol{m}_k$ may overfit to these scarce samples. To mitigate this issue, we utilize the text-only representation $\boldsymbol{f}_k^t$, encoded from the clinical description $\texttt{Prompt}_k$, as concept anchors to regularize the model's embedding $\boldsymbol{m}_k$. This regularization ensures that the learned prototypes align more closely with clinical knowledge rather than being overly tailored to the limited data. We update $\boldsymbol{f}_k^t$, the representations of the concept anchors, at the end of every training epoch to reflect the changes in the text embedding layer during training. We designed a regularization term for the cross-entropy loss to minimize ℓ_2-norm distance between the disease prototypes $\boldsymbol{m}_k$ and the representations $\boldsymbol{f}_k^t$ of the disease-informed textual prompts as

$$L_{reg-ce} = -\frac{1}{N}\sum_{i=1}^{N}\log(\boldsymbol{p}_i)\cdot\boldsymbol{y}_i + \frac{\lambda_2}{C}\sum_{k=1}^{C}||\boldsymbol{m}_k - \boldsymbol{f}_k^t||_2\,, \tag{4}$$

where λ_2 is a weighting factor and $\boldsymbol{p}_i$ denotes the prediction probabilities by projecting the image feature to each prototype

$$\boldsymbol{p}_i = \texttt{Softmax}(\left[\boldsymbol{f}_i^v\cdot\boldsymbol{m}_1, \boldsymbol{f}_i^v\cdot\boldsymbol{m}_2, ..., \boldsymbol{f}_i^v\cdot\boldsymbol{m}_C\right]). \tag{5}$$

D. Overall Model Training and Inference

The vision model $\boldsymbol{E}_v$, image feature projector $\boldsymbol{P}_s$, the updated text embedding layers of the language model and the representations $\{\boldsymbol{m}_k\}_{i=1}^{C}$ of all disease prototypes are optimized simultaneously by minimizing the total loss

$$L_{total} = L_{ita} + L_{proto} + L_{reg-ce}. \tag{6}$$

During inference, diagnosis relies solely on the vision model $\boldsymbol{E}_v$ and the concatenation of prototype representations $\boldsymbol{M}$, by computing the probability of each category according to Eq. 5.

IV. Experiments and Results

A. Experimental Setup

1) Network Architectures: To demonstrate the effectiveness of our proposed method, we evaluated the adaptation performance of two representative large scale pre-trained medical VLMs, BioViL [11] and MGCA [12]. Both models adopt BERT as their language model architecture. BioViL and MGCA were selected as they represent the two most common types of CLIP-based medical VLMs within the realm of chest X-ray diagnosis. BioViL exemplifies medical VLMs with a CNN-based architecture for global alignment of image-text

TABLE I

10-RUN AVERAGE RESULTS OF COVID FINDING AND ALL FINDINGS ON COVID-X WITH 1% TRAINING DATA. THE RESULTS ARE SHOWN FOR BOTH BIOVIL (LEFT) AND MGCA (RIGHT) MODELS. THE BEST PERFORMANCE IS IN **BOLD** AND THE SECOND BEST IS UNDERLINED.

Category	Method	BioViL				MGCA			
		COVID-19 finding		All findings		COVID-19 finding		All findings	
		Precision	Recall	Acc	Weighted F1	Precision	Recall	Acc	Weighted F1
Linear probing of vision encoders	ImageNet Init. [42]	0.636(0.035)*	0.510(0.041)*	0.586(0.040)*	0.588(0.040)*	0.679(0.035)*	0.590(0.038)*	0.648(0.036)*	0.647(0.036)*
	CheXpert Init. [43]	0.643(0.036)*	0.557(0.040)*	0.602(0.040)*	0.603(0.039)*	0.682(0.036)*	0.608(0.037)*	0.651(0.037)*	0.652(0.037)*
Adapters	Linear Probing [8]	0.660(0.033)*	0.601(0.037)*	0.623(0.037)*	0.621(0.038)*	0.688(0.033)*	0.641(0.037)*	0.647(0.036)*	0.647(0.036)*
	CLIP-Adapter [15]	0.625(0.034)*	0.645(0.032)*	0.638(0.033)*	0.640(0.033)*	0.692(0.031)*	0.668(0.032)*	0.670(0.032)*	0.670(0.032)*
	Tip-Adapter [16]	0.628(0.037)*	0.644(0.035)*	0.637(0.034)*	0.641(0.034)*	0.696(0.033)*	0.663(0.031)*	0.675(0.031)*	0.676(0.031)*
Prompt tuning	CoOp [17]	0.665(0.035)*	0.628(0.036)*	0.638(0.036)*	0.636(0.037)*	0.703(0.031)*	0.666(0.032)*	0.681(0.027)*	0.682(0.024)*
	CoCoOp [6]	0.643(0.038)*	0.703(0.030)*	0.662(0.035)*	0.658(0.038)*	0.665(0.033)*	0.707(0.017)*	0.688(0.033)*	0.688(0.033)*
	KgCoOp [18]	0.674(0.037)*	0.697(0.035)*	0.674(0.036)*	0.680(0.036)*	0.747(0.030)*	0.714(0.028)	0.726(0.031)*	0.728(0.031)*
	MaPLe† [7]	-	-	-	-	0.862(0.015)*	0.694(0.033)*	0.772(0.014)*	0.772(0.014)*
Disease-informed	DiCoP + DPL (ours)	**0.737**(0.030)	**0.781**(0.026)	**0.758**(0.030)	**0.762**(0.028)	**0.890**(0.009)	**0.735**(0.033)	**0.800**(0.014)	**0.805**(0.014)

* $p < 0.05$ in the one-tailed paired Student's t-test with our method.
† MaPLe is designed for transformer-based VLMs, thus, is not applicable to BioViL with CNN-based vision encoders.

representations. In contrast, MGCA employs a transformer-based architecture with both global and local alignment, enabling detailed analysis of specific image regions. All the existing medical VLMs, including PLIP [10], Med-CLIP [13], CXR-CLIP [44], and CheX-zero [14], belong to one of the two categories. Therefore, demonstrating the effectiveness of our proposed adaptation method on these two models sheds light on the broader applicability of our method.

2) Datasets: In this paper, we demonstrate our adaptation approach on diagnosing COVID-19 pneumonia using VLMs pre-trained without any knowledge of COVID-19. The experiments in our work involve the following three public datasets.

MIMIC-CXR Dataset [45] consists of 227,943 image-report pairs and serves as the primary dataset for pre-training chest X-ray foundation models, including BioViL and MGCA. Developed prior to the COVID-19 pandemic, MIMIC-CXR dataset does not contain any COVID-19 pneumonia cases.

COVID-x(v6) Dataset [46] plays a crucial role in adapting pre-trained models such as BioViL and MGCA by facilitating the learning of COVID-19 disease concepts. This open-access benchmark dataset consists of 13,975 chest X-ray images, which lack accompanying text information, from 13,870 patient cases. As a multi-national chest X-ray dataset, it is specifically curated for a 3-way classification task that distinguishes COVID-19 pneumonia, non-COVID-19 pneumonia, and normal cases.

COVID-sev [47] comprises 580 COVID-19 and 784 no-finding images, sourced from [48]–[51]. Additional annotations by two expert radiologists include binary masks of COVID-19 infected regions and COVID-19 severity scores on a scale of 0 (no symptoms) to 6 (most severe).

CheXpert [52] is a 2D chest X-ray dataset containing 191,229 frontal chest radiographs. We only utilized the frontal chest radiographs for a multi-label classification task including five individual binary labels: *atelectasis, cardiomegaly, consolidation, edema, and pleural effusion*.

RSNA (v2) [53] dataset comprises 29,700 frontal view chest radiographs. The primary task associated with this dataset is binary classification, which involves categorizing each chest image as either normal or pneumothorax positive.

3) Data Split and Preprocessing: Due to the overlap between COVID-x [46] and COVID-sev [47], we excluded 503 COVID-19 images from COVID-x. These removed images are from the Radiography Database [49], [54] and COVID Chest X-Ray Dataset [50]. The remaining 29,131 images of COVID-x were randomly split into training and validation subsets in a 7 : 1 ratio. The original 400 validation images of COVID-x were repurposed as the test set. All images were resized to 512×512 pixels for BioViL and 224×224 pixels for MGCA. The training and validation sets were used to adapt medical VLMs to COVID-19 using the proposed methods. We then evaluated the diagnosis performance of the adapted models on the test set of COVID-x. For COVID-sev, we kept its original training, validation, and test subsets as described in [47]. The training and validation sets are utilized for the linear probing of the adapted medical VLMs on the COVID-19 severity estimation task. Since no further training is required, the entire COVID-sev dataset is used to assess the performance of the adapted VLMs on COVID-19 visual grounding. For the CheXpert dataset, following the previous work [55], we reserve the expert-labeled validation set to serve as our test data and randomly select 5,000 radiographs from the training dataset for validation purposes. For the RSNA dataset, in accordance with reference [55], we divide the dataset into training, validation, and test sets, in the ratio of 70:15:15, respectively.

4) Disease Description and Prompt Preparation: We crafted prompts corresponding to the three diagnostic categories, *i.e.*, *COVID-19 pneumonia*, *non-COVID pneumonia*, and *healthy individuals* in COVID-x, using the template outlined in Eq. 1. Alongside these templates, three trainable disease prototypes were developed to accurately represent the characteristics of these categories. This setup enhances the adapted VLMs' ability to learn and diagnose diseases effectively. Detailed examples of the prompts are provided in Appendix-A. To leverage medical knowledge from a large language model for automatically drafting approximately correct clinical prompts for each descriptive attribute (texture, shape, and location), we use the following sample query to prompt GPT-4: *"Please detail the radiographic features of COVID-19, normal pneumonia, and*

healthy patients, focusing on their texture, shape, and location. Structure your response using bullet points for clarity." This automated generation of descriptive attributes can significantly reduce the time and effort required for our radiologists to create prompts from scratch. To generate concept anchors $\boldsymbol{f}_k^t$ in Eq. 4, we randomly selected the candidate $\texttt{Des}_k(\cdot)$ for each attribute and concatenated the prompts according to Eq. 1.

B. Training Details

When adapting the VLMs, we keep their text encoders frozen. Due to the computational expense of directly fine-tuning the large vision encoders, we adopted two distinct finetuning frameworks: LoRA [56] (rank $r = 4$ and scaling $\alpha = 4$) and layer-wise learning rate decay [57] (decay factor $\beta = 0.9$) to finetune MGCA and BioViL on COVID-x training set, respectively. LoRA, known for its efficiency in finetuning transformer-based models, was selected for refining the MGCA model due to its ability to adapt parameters effectively without extensive retraining. Conversely, weight decay fine-tuning, which is traditionally preferred for CNN-based encoders, was used for the BioViL model. This approach leverages the strengths of each tuning method to optimize the respective architectures of the VLMs.

All the trainable model parameters, including the updated text embedding layers, the image feature projector, the image encoder, and the prototypes, are trained for 100 epochs with a batch size of 128, using a learning rate of 0.0005 and the AdamW optimizer for both BioViL and MGCA. We empirically set the temperature parameters $\tau_1 = \tau_2 = 0.07$ and weighting factors $\lambda_1 = \lambda_2 = 0.1$ in all the experiments. Detailed analysis on the effects of these hyperparameters on the model diagnosis performance is presented in Sec. IV-H. All experimental settings in this paper can be accommodated on a single NVIDIA A100 GPU.

C. Method Effectiveness

We first demonstrate the effectiveness of our method by comparing it with three kinds of methods. In our comparison, we include adapter-based methods, such as linear probing [8], CLIP-Adapter [15], and Tip-Adapter [16], as well as prompting-based approaches like CoOp [17], CoCoOp [6], KgCoOp [18], and MaPLe [7]. Additionally, to highlight the potential of VLMs over traditional single-modal models, we also evaluate the performance of vision-only models [42], [43] pre-trained on datasets like ImageNet [58] and CheXpert [52]. For a fair comparison, these vision-only models utilize the same architectures as the vision encoders in the compared VLMs, specifically the ResNet-50 in BioViL and the ViT-B/16 in MGCA. All baseline methods are trained following their original settings.

To mitigate the impact of randomness, we reported the mean metric values and standard deviation across 10 runs on the test set. In each run, we randomly selected 1% of the COVID-x training data to fine-tune the models. The diagnosis performance evaluated on the COVID-x test set is presented in Table I. First, the linear probing of vision-only models generally performed the worst among all methods. This demonstrates the significant advantage of multimodal foundation models over traditional single-modal encoders. This finding indicates that pre-trained VLMs can be quickly adapted for diagnosing *newly identified* diseases. The superior performance of VLMs could attribute to the use of both visual and textual data during training, which enhances their ability to understand visual contexts with semantic language associations. In addition, among the compared adaptation methods, the adapter-based methods generally perform the worst, indicating that simply tweaking the additional classifier may not effectively leverage the capabilities of VLMs in the downstream tasks.

Our approach significantly ($p < 0.05$ with Students t-test) outperformed other prompting-based methods in almost all the cases. This underscores DiCoP's advantage over optimizing prompts without clinical knowledge and demonstrates DPL's benefit in tailoring VLMs to less represented diseases, thereby avoiding overfitting on scarce data. Additionally, we observed that the choice of pre-trained medical VLMs has a substantial impact on the final results, with MGCA typically outperforming BioViL. This discrepancy can be attributed to the differences in their architectural designs and alignment strategies. Unlike BioViL, which utilizes a CNN-based vision encoder focusing on conventional image-text feature alignment, MGCA incorporates a more advanced transformer-based architecture. In addition, MGCA performs both global and local image-text feature alignments, enabling a more detailed analysis of specific image regions. As a result, MGCA not only excels in downstream tasks [11], [12] but also shows significant improvements in our newly introduced disease adaptation tasks.

D. Representation Visualization

To gain more insights into the capabilities of the methods, we visualized the geometric structures of the learned representations for our method and three baselines using t-SNE [59] on both the validation and test sets of COVID-x. The results are shown in Fig. 5. In this study, we take BioViL as the backbone VLM for instance. Similar results on MGCA model can be found on our Github project page. All methods used only 1% of the training set to adapt BioViL to COVID-19 in this study. Linear probing **(a)** and Tip-Adapter **(b)** show less effective separation between non-COVID-19 and COVID-19 pneumonia samples, highlighting the inadequacies of basic final layer tuning in adapting pre-trained medical VLMs to new diseases. In contrast, our method **(d)** achieved significantly better contrastive separation compared to other three baseline methods on both the validation and test sets, corroborating the outstanding results reported in Table I.

Fig. 5 also shows the alignment between the learned visual and linguistic representations. To ensure a fair comparison, we standardized the text input for the language encoder. The text prompts were formatted as "`This is a chest X-ray image of [class name]`" for each class. CoCoOp **(c)** not only struggled to distinguish disease categories but also failed to align the visual representations (○) of each medical finding with the corresponding linguistic representations (△) of the prompt. This indicates its limitation in taking up

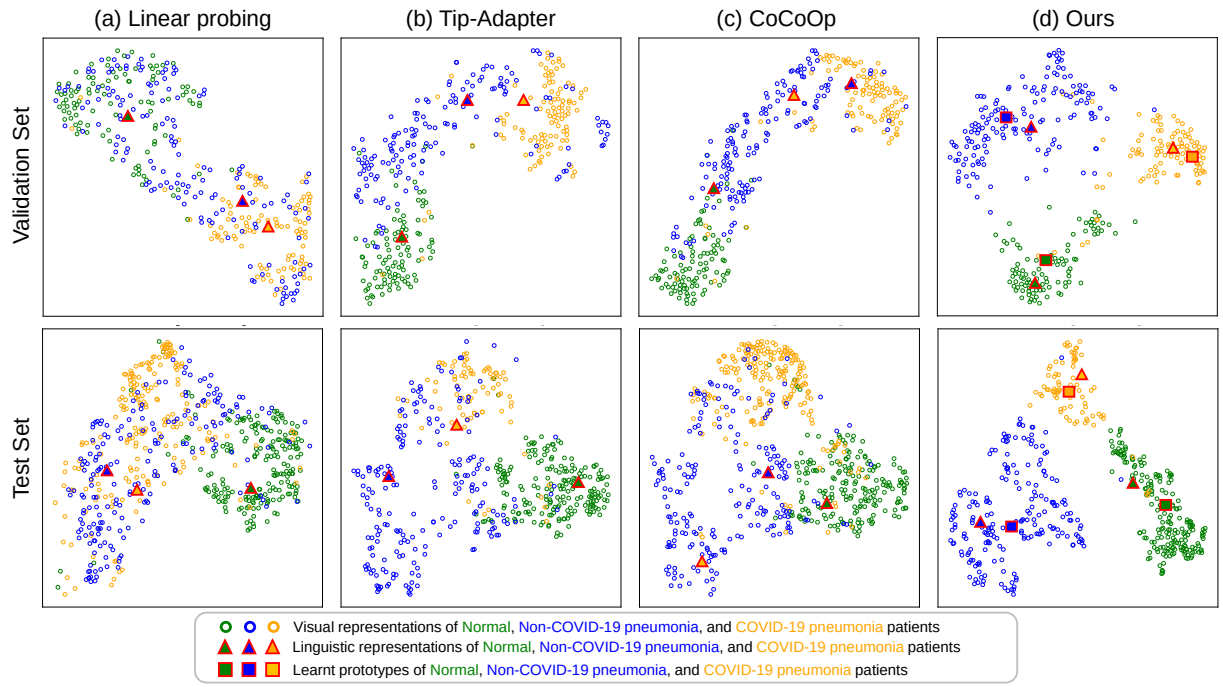

Fig. 5. t-SNE visualization of the latent space from a random batch of COVID-x samples from the validation set (first row) and test set (second row), showing the feature distribution from **(a)** Linear probing [8], **(b)** Tip-Adapter [16], **(c)** KgCoOp [18], and **(d)** our method.

clinical knowledge from textual inputs and lacking geometric regularization during the VLM adaptation. Additionally, as shown in Fig. 5**(d)**, the learned prototypes (□) of our method not only effectively separate different medical findings but also remain close to the correct text input representations. This validates the effectiveness of the overfitting regularization term in Eq. 4.

E. Data Efficiency

At the onset of the pandemic, the scarcity of data for *newly identified* diseases posed a significant challenge. To address this, we evaluated the data efficiency of the adapted VLMs by testing their diagnostic performance with varying proportions of training data. We adapted pre-trained medical VLMs using only a specific percentage of samples randomly selected from the COVID-x training set. The diagnostic performance on the COVID-x test set is shown in Fig. 6. Notably, our method outperforms other baselines in all metrics, particularly when the number of training data samples is small. That can be very useful for preparing pre-trained VLMs for diagnosing *newly identified* diseases even at the early days. This success is attributed to the integration of the DiCoP on incorporating clinical knowledge of new diseases into pre-trained VLMs and DPL on discovering the patterns of new diseases through latent space regularization.

F. Phrase Grounding

To illustrate the effectiveness of locating clinical findings described in human language on chest X-ray images, we conducted a phrase grounding evaluation of the adapted BioViL and MGCA models. Both VLMs were trained using 1% of the COVID-x training set and evaluated on all COVID-19 samples in the COVID-sev dataset. We tried to locate the text prompt formatted as "`This is an chest X-ray image of a patient with COVID-19.`" on each COVID-19 sample via visualizing the correlation between text embedding and vision patch representations.

More specifically, for each image-prompt pair, the image is passed to the image encoder to obtain a grid of P patch representations $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_P]$. For MGCA, P equals the number of input image patches to ViT, *i.e.* number of visual tokens. For BioViL, we adhered to the approach described in the original paper [11], setting a grid of $P = 16 \times 16$ on the output feature map produced by ResNet-50 vision encoder. Similarly, the textual prompt is embedded via the text encoder and projected to the joint space to obtain $\boldsymbol{f}^t$. The cosine similarity score between $\boldsymbol{f}^t$ and entries of $\boldsymbol{V}$ produces a similarity map $\boldsymbol{S} = \left[\cos(\boldsymbol{v}_1^T \cdot \boldsymbol{f}^t), \cos(\boldsymbol{v}_2^T \cdot \boldsymbol{f}^t), ..., \cos(\boldsymbol{v}_P^T \cdot \boldsymbol{f}^t)\right]$. The similarity map is normalized to the range of $[0, 1]$, resized to the dimensions of the input image using Bicubic interpolation, and then thresholded to produce visual grounding contours. These contours are depicted as red outlines in Fig. 7. The visual grounding contours are then evaluated against the radiologist annotations, marked as the blue contours in Fig. 7.

To quantitatively assess the visual grounding of our method compared to baselines, we present the mean intersection over union (mIoU) and contrast-to-noise ratio (CNR) metrics in Table II. The mIoU for the visual grounding contours and radiologist annotations is calculated by averaging across thresholds $[0.5, 0.6, 0.7]$. The CNR measures the ratio of the summed similarity scores within and outside radiologist annotations, independently of thresholds. Our method outperforms others across both BioViL and MGCA models for COVID-19 findings, demonstrating enhanced alignment and correctness

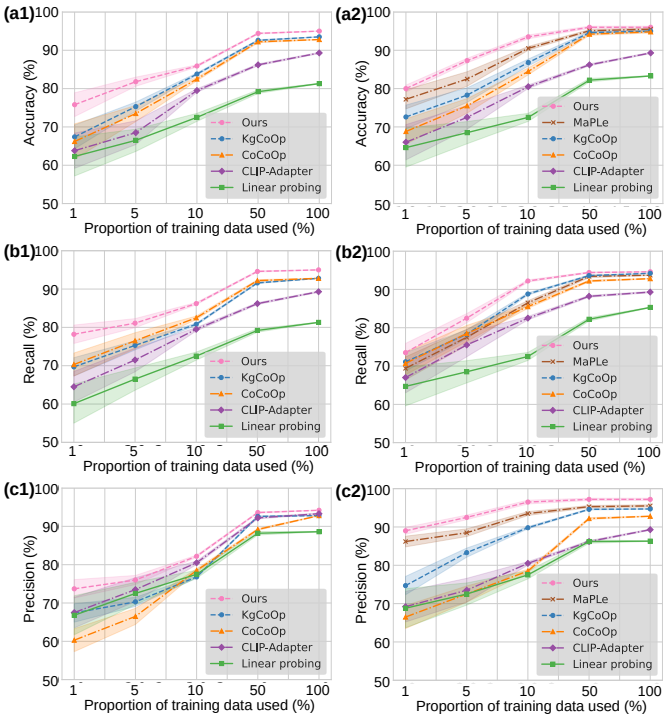

Fig. 6. Data efficiency evaluation on BioViL (left) and MGCA (right). Panels **(a1)**, **(b1)**, and **(c1)** display the all-class classification accuracy, recall, and precision of diagnosing COVID-19, respectively, using BioViL. Panels **(a2)**, **(b2)**, and **(c2)** present the same metrics for MGCA.

of learned visual-linguistic representations, facilitated by the DiCoP module.

TABLE II
10-RUN AVERAGE RESULTS OF COVID-19 PHRASE GROUNDING EVALUATION ON COVID-SEV. THE BEST PERFORMANCE IS IN **BOLD** AND THE SECOND BEST IS UNDERLINED.

Method	BioViL		MGCA	
	CNR	mIoU	CNR	mIoU
ImageNet Init.	0.732(0.011)*	0.143(0.006)*	0.737(0.013)*	0.152(0.006)*
CheXpert Init.	0.744(0.012)*	0.146(0.005)*	0.746(0.012)*	0.159(0.005)*
Linear Probing	0.831(0.011)*	0.193(0.005)*	0.850(0.012)*	0.197(0.005)*
CLIP-Adapter	0.832(0.011)*	0.202(0.005)*	0.852(0.012)*	0.211(0.004)*
Tip-Adapter	0.865(0.012)*	0.224(0.004)*	0.872(0.011)*	0.233(0.005)*
CoOp	0.929(0.012)*	0.213(0.004)*	0.966(0.011)*	0.243(0.006)*
CoCoOp	0.942(0.010)*	0.228(0.005)*	0.980(0.010)*	0.257(0.005)*
KgCoOp	0.947(0.013)*	0.256(0.004)	0.998(0.011)*	0.270(0.002)
MaPLe†	-	-	1.010(0.010)*	0.267(0.002)*
Ours	**1.025**(0.010)	**0.273**(0.003)	**1.098**(0.010)	**0.301**(0.002)

* $p < 0.05$ in the one-tailed paired Student's t-test with our method.
† MaPLe is not directly applicable to VLMs that use CNN-based vision encoders.

G. Severity Estimation

To further assess the capabilities of disease information-guided visual representations and their applicability beyond diagnosis, we incorporated a disease severity risk estimation task. We utilized the vision encoders of the BioViL and MGCA models from Sec. IV-C, adapted using 1% of the COVID-x training set. We conducted linear probing on the output of the vision encoders using the COVID-sev training set. Given that severity estimation is an ordinal regression task,

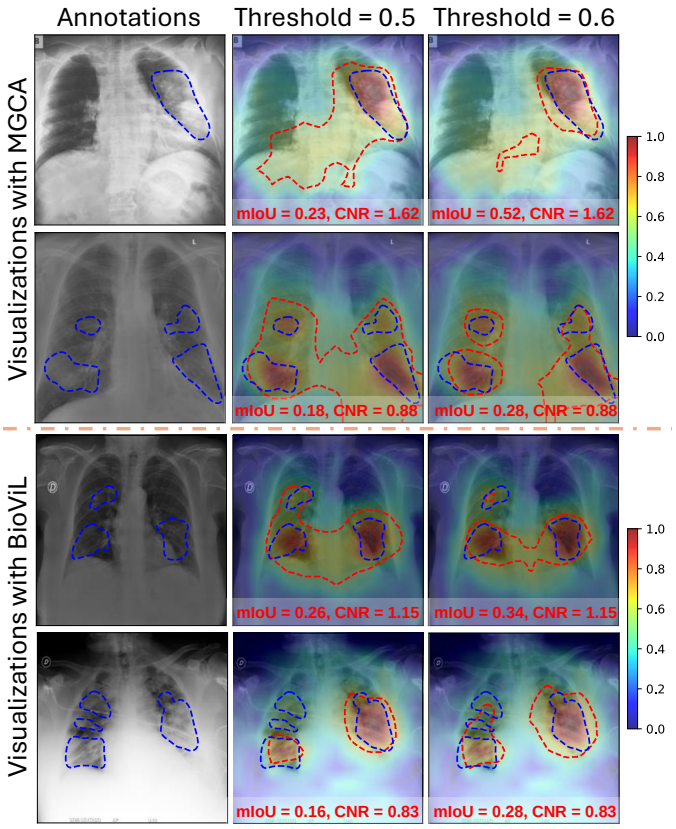

Fig. 7. Visual grounding visualization of the adapted MGCA (rows 1 and 2) and BioViL (rows 3 and 4) on four randomly selected COVID-19 examples. The first column is the radiologist's annotation of the COVID-19 infected regions. The second and third columns show the overlay of the Bicubic interpolated similarity score map, with the red contours thresholded at 0.5 and 0.6, respectively.

we employed the consistent ordinal rank loss (CORAL) [60] to fine-tune the linear layer and assessed performance using mean absolute error (MAE) and R-squared score (R^2) on the COVID-sev test set. The results, presented in Table III, show that our method surpasses others in COVID-19 severity estimation across both BioViL and MGCA models. This success underscores the significant potential of disease information-guided visual representations not only in diagnosing but also in capturing quantifiable information about disease progression.

H. Hyper-parameter Sensitivity Analysis

To validate the significance of two loss weighting factors λ_1 in the L_{proto} and λ_2 in L_{reg-ce}, we conducted a sensitivity analysis on the BioViL model, which were trained using various proportions of the COVID-x training dataset and subsequently evaluated on the COVID-x test set. The results are presented in Fig. 8. In the experiments, we initially set $\lambda_1 = \lambda_2 = 0.1$. When analyzing the sensitivity to one parameter, the other parameter is fixed.

In Fig. 8(**a**), when $\lambda_1 \in [0.05, 0.50]$ regularizes the strong separability between different disease prototypes, our method stably outperforms other baselines with different proportion of training data. Fig. 8(**b**) presents the inconsistent behavior of models adapted with different proportion of samples

TABLE III
10-RUN AVERAGE RESULTS OF COVID-19 SEVERITY ESTIMATION ON COVID-SEV. THE BEST PERFORMANCE IS IN **BOLD** AND THE SECOND BEST IS UNDERLINED.

Method	BioViL		MGCA	
	MAE(↓)	R^2(↑)	MAE(↓)	R^2(↑)
ImageNet Init.	0.667*(0.005)	0.852*(0.005)	0.652*(0.004)	0.875*(0.004)
CheXpert Init.	0.625*(0.005)	0.886*(0.005)	0.579*(0.004)	0.903*(0.005)
Linear Probing	0.492*(0.005)	0.897*(0.005)	0.489*(0.004)	0.906*(0.005)
CLIP-Adapter	0.447*(0.004)	0.901*(0.004)	0.422*(0.003)	0.907*(0.005)
Tip-Adapter	0.447*(0.004)	0.905*(0.004)	0.425*(0.003)	0.912*(0.005)
CoOp	0.386*(0.002)	0.909*(0.004)	0.374*(0.003)	0.915*(0.005)
CoCoOp	0.358*(0.002)	0.916*(0.003)	0.338*(0.002)	0.930*(0.003)
KgCoOp	0.342*(0.002)	0.927*(0.002)	0.330*(0.001)	0.933*(0.003)
MaPLe†	-	-	0.314*(0.001)	0.947*(0.003)
Ours	**0.303**(0.001)	**0.934**(0.004)	**0.265**(0.001)	**0.975**(0.003)

* $p < 0.05$ in the one-tailed paired Student's t-test with our method.
† MaPLe is not directly applicable to VLMs that use CNN-based vision encoders.

when increasing the debiasing weight λ_2. For models adapted with more training samples (100% or 50% training data), too strong debiasing ($\lambda_2 \geq 0.20$) can degrade the model performance. However, in data-scarce scenarios (1% or 10% training data), model performance improves gradually as the debiasing weight λ_2 increases within $[0.01, 0.20]$. This observation indicate the effect of the regularization term in Eq. 4 for helping align the models learning closer to clinical knowledge rather than being narrowly tailored to the limited data. Considering the trade-off debiasing effect of λ_2, we fixed $\lambda_2 = 0.01$ for all experiments for overall decent performance. Fig. 8**(c)** shows that the model performance is initially increase and then get stable if the LoRA rank $r \geq 4$. To balance computational efficiency with model performance, we selected $r = 4$ for our settings, optimizing the number of fine-tuning parameters. Fig. 8**(d)** illustrates that the performance initially improves as the scaling factor α increases, but subsequently declines when α becomes overly large. Fig. 8**(e)** presents that the performance initially improves as the decay rate β increases, but subsequently gets stable for $\beta \in [0.9, 0.99]$. The phenomenon is more pronounced in the data-scarce scenarios (training proportions of 1% or 10%). This can be explained by the use of a large decay rate, which allows us to focus tuning on the final layers of the vision model while keeping the earlier layers relatively unchanged. This strategy helps enhance model performance and prevents overfitting, especially when the number of training samples is limited. Fig. 8**(f)** and **(g)** show that the model performance increases initially, and then stays stable as the temperature parameters increase. It is because the early-stop acceleration is adopted to control the perturbation strength.

I. Generalizability of the Adapted Medical VLMs

It is paramount to ensure the model generalizability after model adaptation, especially when VLMs are adapted with limited samples. In this study, our image encoder was fine-tuned using a dataset containing a small number of new disease samples. We have thus measured the effectiveness of our model on two downstream tasks to evaluate its generalizability. The two tasks are the five-class multi-label classification on the

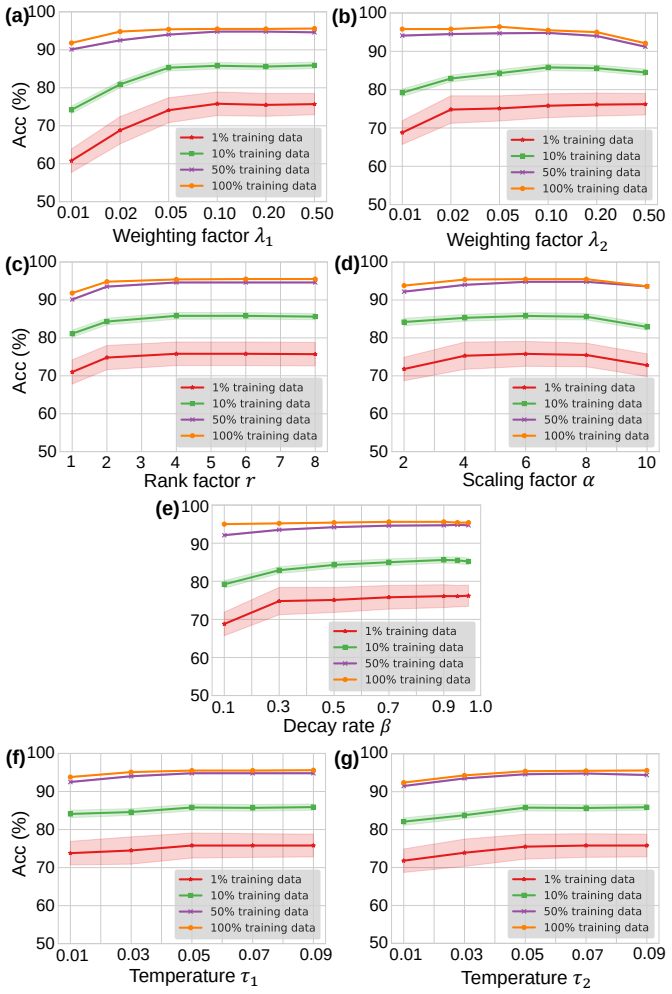

Fig. 8. Sensitivity analysis of the two loss weighting factors **(a)** λ_1 and **(b)** λ_2 on the BioViL model. **(c)** is the LoRA rank factor r, and **(d)** is the scaling factor α in LoRA. **(e)** is the rate decay used in BioViL vision model adaptation. **(f)** and **(g)** are the temperature coefficients τ_1 and τ_2.

CheXpert dataset and the binary pneumonia classification task on the RSNA dataset, which do not involve the new COVID-19 diseases. We employed zero-shot classification and few-shot linear probing (with 1%, 10% and 100% of training set) to test both the original and adapted VLMs. We evaluated the mean Area Under the Receiver Operating characteristic Curve (AUROC) and standard deviation across 10 runs on the test set. The results are presented in Table IV.

Compared with the original BioViL and MGCA, their adapted versions present similar performance without statistical significance (none of cases having $p < 0.05$) on CheXpert and the RSNA datasets. These results reaffirm the good generalizability of our adapted medical VLMs. By successfully navigating the challenges posed by limited samples, our VLM adaptation method remains broadly applicable and effective across the diverse downstream tasks of medical imaging diagnostics.

J. Ablation Studies

To evaluate the effectiveness of each component in our method, we conducted an ablation study using the BioViL

TABLE IV
10-RUN AVERAGE PERFORMANCE OF MODEL GENERALIZABILITY ANALYSIS ON CHEXPERT AND RSNA DATASETS.

Ablated modules		CheXpert (AUROC)				RSNA (AUROC)			
		zero-shot	1%	10%	100%	zero-shot	1%	10%	100%
DiCoP	Original	0.811(0.002)	0.862(0.002)	0.868(0.002)	0.876(0.002)	0.831(0.002)	0.881(0.002)	0.884(0.002)	0.891(0.002)
	Adapted *	0.806(0.003)	0.860(0.002)	0.866(0.002)	0.876(0.003)	0.835(0.002)	0.883(0.002)	0.885(0.002)	0.892(0.002)
DPL	Original	0.829(0.002)	0.888(0.003)	0.891(0.002)	0.897(0.002)	0.836(0.002)	0.891(0.003)	0.899(0.030)	0.908(0.003)
	Adapted *	0.825(0.003)	0.885(0.002)	0.890(0.002)	0.896(0.002)	0.841(0.002)	0.891(0.002)	0.901(0.003)	0.908(0.002)

* Two-sided paired Students t-test was applied to test the statistical difference between the original model and our adapted model. None of the test has $p < 0.05$. Thus, the null hypothesis, "there is no difference between the original and adapted VLMs should be accepted in all the comparisons.

model adapted with just 1% of the COVID-x training data. This approach illustrates how each component contributes to the overall performance under scenarios of data scarcity. We evaluate the diagnostic performance (Acc) and phrase grounding performance (mIoU) of ablated models. The results, displayed in Table V, demonstrate the efficacy of the components within the proposed DiCoP and DPL modules.

TABLE V
10-RUN AVERAGE ABLATION STUDY OF BIOVIL MODEL ADAPTED WITH 1% COVID-X TRAINING SET.

Ablated modules		Diagnosis (Acc)	Phrase Grounding (mIoU)
full model		0.758(0.030)	0.273(0.003)
DiCoP	no txt token & emb.	0.744(0.032)*	0.253(0.003)*
	no txt aug.	0.747(0.035)*	0.256(0.004)*
	no txt of COVID-19	0.732(0.033)*	0.245(0.004)*
	no txt of all diseases	0.724(0.035)*	0.233(0.004)*
	no E_s	0.727(0.036)*	0.218(0.004)*
DPL	w/o L_{ita}	0.684(0.035)*	0.219(0.005)*
	w/o L_{proto}	0.697(0.036)*	0.216(0.005)*
	$L_{reg-ce} \rightarrow L_{ce}$	0.688(0.037)*	0.197(0.005)*

* $p < 0.05$ in the one-tailed paired Student's *t*-test with *full model*.

1) Effectiveness of DiCoP: We evaluated the effectiveness of each proposed modules in DiCoP. Each module was removed from the *full model* in a leave-one-out fashion. The ablated modules include **1)** no update on the word tokenizer and embeddings layer (*no txt token & emb.*) in Fig. 3, **2)** no the text augmentations (*no txt aug.*), **3)** removing the descriptive attributes from Eq. 1 of the COVID-19 class (*no txt of COVID-19*), **4)** removing the descriptive attributes of all classes (*no txt of all diseases*), and **5)** no image-specific features (*no* E_s). The columns 3-7 in Table V demonstrates performance declines relative to the *full model* upon removing each component, indicating the efficacy of all components. The updated text tokenizer and embedding layer represent significant technical advancements over our MICCAI work (Acc=0.741) [5], markedly improving the overall performance of the full model. We observed that the absence of textual descriptions for all diseases (*no txt of all diseases*) had the most significant impact on overall performance. This underscores the importance of integrating disease-informed clinical knowledge into the prompt generation. In addition, we noticed that removing the text encoder (*i.e.*, without the loss of L_{ita}) results in a degradation of model performance from 0.758 to 0.684, which indicates that image-text alignment loss carries additional information. The `[CLS]` token of the language encoder fused both image specific features and disease-related textual information from prompts, instead of simply distilling the image-specific features.

2) Efficacy of the loss terms in DPL: We analyzed the efficacy of the loss terms in the DPL framework. Specifically, we disabled **1)** the image-text alignment loss *w/o* L_{ita}, **2)** the prototype learning loss *w/o* L_{proto}, and **3)** the debias term by replacing L_{reg-ce} with regular cross-entropy loss ($L_{reg-ce} \rightarrow L_{ce}$). Each loss term is excluded in a leave-one-out manner. Columns 8-10 in Table V show significant declines in model performance compared to the *full model*, underscoring the contribution of each loss term. We noted that the most significant decline in overall performance occurred when L_{proto} and the debias term in L_{reg-ce} were removed. This demonstrates the critical role these components play in regularizing the latent space of the VLMs, facilitating the learning of new disease concepts efficiently while preventing overfitting to limited samples.

V. DISCUSSIONS AND CONCLUSIONS

In this paper, we present an adaptation framework to leverage the visual-linguistic capabilities of VLMs to prepare for the diagnosis of a *newly identified* disease, using the emergence of COVID-19 pneumonia as an exemplar use case. To address this challenge, we introduced a disease-informed adaptation method with two key contributions. Firstly, **DiCoP** leverages clinical knowledge to craft prompts that effectively represent the concepts of the *newly identified* disease. Secondly, we propose **DPL** to tackle the lack of structural regulation in the latent space of CLIP-based VLMs. Empirical analyses confirm the effectiveness and efficiency of our method across various VLMs and tasks, including the disease diagnosis, phrase grounding, and severity estimation. Our method particularly excels under conditions of limited data availability, which mirrors the early stages of a pandemic when data for model adaptation is scarce.

One limitation of our study is its focus on validating our method with CLIP-based VLMs. Nonetheless, the proposed DiCoP and DPL modules can be applied to other VLMs. Another limitation of our work is that we only considered COVID-19 pneumonia as a new disease due to the limited availability of medical imaging datasets. Our study focused on three COVID-related datasets, representing a range of challenges and tasks associated with this widely known disease. The integration of DiCoP with DPL could pave the way for advancements in disease-informed computer-aided diagnosis, potentially expanding its applications to other new disease-related medical image analysis tasks in the future.

TABLE VI
THE EXAMPLE MANUALLY-CRAFTED PROMPTS FOR COVID-X DATASET.

Findings (Condition categories)	Attributes	Prompt Candidate 1	Prompt Candidate 2	Prompt Candidate 3
COVID-19 pneumonia	Des_k(basic)[1]	"A chest X-ray image of a patient with COVID-19."	"A radiograph of a COVID-19 patient."	"An X-ray image showing a patient diagnosed with COVID-19."
	Des_k(texture)	"Texture Patterns include bilateral, patchy and ground-glass opacities (GGO) in the lungs. These opacities can vary in density and distribution."	"Texture patterns feature bilateral, patchy, and ground-glass opacities in the lungs, which may differ in density and distribution."	"Texture patterns exhibit bilateral, patchy, and ground-glass opacities (GGO) in the lungs, varying in density and distribution."
	Des_k(shape)	"The opacities can have irregular shapes, appearing as hazy areas with fuzzy borders."	"The opacities may exhibit irregular contours, manifesting as hazy regions with indistinct edges."	"The opacities often feature irregular forms, presenting as blurred areas with indeterminate boundaries."
	Des_k(location)	"The opacities are commonly located in the peripheral regions of the lungs, particularly in the lower lobes. They may involve multiple lung segments of both chest sides."	"The opacities typically appear in the peripheral areas of the lungs, especially in the lower lobes, and can affect multiple segments of both sides of the chest."	"Opacities are often found in the peripheral parts of the lungs, mainly within the lower lobes, and may affect several lung segments on both sides of the chest."
Non-COVID-19 pneumonia	Des_k(basic)[1]	"A chest X-ray image of a patient with pneumonia."	"A radiograph displaying the lung condition of a patient diagnosed with pneumonia."	"An X-ray image of a pneumonia patient."
	Des_k(texture)	"Textual Patterns can include areas of increased lung density due to inflammatory infiltrates."	"Textual patterns may feature regions of heightened lung density resulting from inflammatory infiltrates."	"Textual patterns can display areas of elevated lung density caused by inflammatory infiltrates."
	Des_k(shape)	"In non-COVID pneumonia, opacities may have a lobar or segmental distribution, depending on the type of pneumonia."	"Opacities can present with a lobar or segmental distribution, varying according to the specific type of pneumonia.	"The distribution of opacities can be either lobar or segmental, based on the type of non-COVID pneumonia.
	Des_k(location)	"The location of pneumonia opacities can vary but is often seen in specific lobes or segments of the lung."	"The position of pneumonia opacities varies, usually observed in particular lobes or segments of the lung."	"Pneumonia opacities can appear in various locations but commonly manifest in specific lobes or segments of the lung."
Healthy individuals	Des_k(basic)[1]	"A chest X-ray image of normal healthy individual."	"A chest X-ray showing the lungs of a normal, healthy individual."	"An X-ray image of the chest from a healthy individual"
	Des_k(texture)	"No respiratory symptoms or underlying lung conditions, chest typically show clear lung fields with no areas of abnormal opacities."	"In the absence of respiratory symptoms or pre-existing lung conditions, a chest X-ray generally reveals clear lung fields free from any abnormal opacities."	"Without respiratory symptoms or pre-existing lung conditions, a chest X-ray typically shows clear lung fields without any abnormal opacities."
	Des_k(shape)	"No hazy areas with fuzzy borders."	"There are no unclear regions with blurred boundaries."	"There are no indistinct areas with blurred edges."
	Des_k(location)	"The whole lung fields appear homogeneous and translucent without any irregularities or opacities."	"The entire lungs seem uniform and translucent, devoid of any irregularities or areas of opacity."	"The whole lung fields presents as homogeneous and translucent, lacking any irregularities or opacities."

[1] The default prompt template setting in the previous work.

APPENDIX

A. Disease-informed Prompts

Using the GPT-4 input queries mentioned above, we automatically generated three candidates for each descriptive attribute (texture, shape, and location) based on our predefined template. These prompts were then manually revised with the help of a radiologist, ensuring that they are medically accurate. The complete set of prompt candidates for each attribute of every medical finding category is listed in Table VI.

REFERENCES

[1] S. Wang, C. Li, R. Wang, Z. Liu, M. Wang, H. Tan, Y. Wu, X. Liu, H. Sun, R. Yang, X. Liu, H. Zhou, B. I. Ayed, and H. Zheng, "Annotation-efficient deep learning for automatic medical image segmentation," Nature Communications, vol. 12, no. 1, p. 5915, 2021.

[2] S. Reiß, C. Seibold, A. Freytag, E. Rodner, and R. Stiefelhagen, "Every annotation counts: Multi-label deep supervision for medical image segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9532–9542.

[3] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2009.

[4] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27. Springer, 2018, pp. 270–279.

[5] J. Zhang, G. Wang, M. K. Kalra, and P. Yan, "Disease-informed adaptation of vision-language models," arXiv preprint arXiv:2405.15728, 2024.

[6] K. Zhou, J. Yang, C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16 816–16 825.

[7] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "MaPLe: Multi-modal prompt learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19 113–19 122.

[8] A. Radford, J. W. Kim, C. Hallacy et al., "Learning transferable visual models from natural language supervision," in International Conference on Machine Learning. PMLR, 2021, pp. 8748–8763.

[9] M. Hafner, M. Katsantoni, T. Köster, J. Marks, J. Mukherjee, D. Staiger, J. Ule, and M. Zavolan, "Clip and complementary methods," Nature Reviews Methods Primers, vol. 1, no. 1, p. 20, 2021.

[10] 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.

[11] B. Boecking, N. Usuyama, S. Bannur et al., "Making the most of text semantics to improve biomedical vision–language processing," in European Conference on Computer Vision. Springer, 2022, pp. 1–21.

[12] F. Wang, Y. Zhou, S. Wang, V. Vardhanabhuti, and L. Yu, "Multi-granularity cross-modal alignment for generalized medical visual representation learning," Advances in Neural Information Processing Systems, vol. 35, pp. 33 536–33 549, 2022.

[13] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Med-clip: Contrastive learning from unpaired medical images and text," arXiv preprint arXiv:2210.10163, 2022.

[14] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, "Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning," Nature Biomedical Engineering,

vol. 6, no. 12, pp. 1399–1406, 2022.

[15] P. Gao, S. Geng, R. Zhang et al., "Clip-adapter: Better vision-language models with feature adapters," International Journal of Computer Vision, vol. 132, no. 2, pp. 581–595, 2024.

[16] R. Zhang, R. Fang, W. Zhang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free clip-adapter for better vision-language modeling," arXiv preprint arXiv:2111.03930, 2021.

[17] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," International Journal of Computer Vision, vol. 130, no. 9, pp. 2337–2348, 2022.

[18] 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, pp. 6757–6767.

[19] S. Shen, C. Li, X. Hu et al., "K-lite: Learning transferable visual models with external knowledge," Advances in Neural Information Processing Systems, vol. 35, pp. 15 558–15 573, 2022.

[20] Z. Qin, H. Yi, Q. Lao, and K. Li, "Medical image understanding with pretrained vision language models: A comprehensive study," arXiv preprint arXiv:2209.15517, 2022.

[21] S. Goel, H. Bansal, S. Bhatia, R. Rossi, V. Vinay, and A. Grover, "Cyclip: Cyclic contrastive language-image pretraining," Advances in Neural Information Processing Systems, vol. 35, pp. 6704–6719, 2022.

[22] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," arXiv preprint arXiv:1504.00325, 2015.

[23] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," Advances in Neural Information Processing Systems, vol. 35, pp. 2507–2521, 2022.

[24] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," arXiv preprint arXiv:2308.00692, 2023.

[25] 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. PMLR, 2021, pp. 4904–4916.

[26] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," Advances in Neural Information Processing Systems, vol. 36, 2024.

[27] M. Y. Lu, B. Chen, D. F. Williamson et al., "A visual-language foundation model for computational pathology," Nature Medicine, vol. 30, no. 3, pp. 863–874, 2024.

[28] A. Van Den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," Advances in Neural Information Processing Systems, vol. 30, 2017.

[29] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," arXiv preprint arXiv:2204.06125, vol. 1, no. 2, p. 3, 2022.

[30] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," arXiv preprint arXiv:2204.13258, 2022.

[31] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," BMC medical imaging, vol. 22, no. 1, p. 69, 2022.

[32] J. Zhang, H. Chao, A. Dhurandhar, P.-Y. Chen, A. Tajer, Y. Xu, and P. Yan, "Spectral adversarial mixup for few-shot unsupervised domain adaptation," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2023, pp. 728–738.

[33] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," IEEE Transactions on Industrial Informatics, vol. 15, no. 4, pp. 2446–2455, 2018.

[34] J. Lee and R. M. Nishikawa, "Cross-organ, cross-modality transfer learning: feasibility study for segmentation and classification," IEEE Access, vol. 8, pp. 210 194–210 205, 2020.

[35] A. Anaya-Isaza and L. Mera-Jiménez, "Data augmentation and transfer learning for brain tumor detection in magnetic resonance imaging," IEEE Access, vol. 10, pp. 23 217–23 233, 2022.

[36] J. Zhang, H. Chao, and P. Yan, "Toward adversarial robustness in unlabeled target domains," IEEE Transactions on Image Processing, vol. 32, pp. 1272–1284, 2023.

[37] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, and J. Gao, "Grounded language-image pre-training," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10 965–10 975.

[38] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," arXiv preprint arXiv:1901.11196, 2019.

[39] H. Dai, Z. Liu, W. Liao et al., "Auggpt: Leveraging chatgpt for text data augmentation," arXiv preprint arXiv:2302.13007, 2023.

[40] J. Li and H. Sun, "Lift: Transfer learning in vision-language models for downstream adaptation and generalization," in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 4678–4687.

[41] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," arXiv preprint arXiv:1904.03323, 2019.

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[43] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3942–3951.

[44] K. You, J. Gu, J. Ham, B. Park, J. Kim, E. K. Hong, W. Baek, and B. Roh, "Cxr-clip: Toward large scale chest x-ray language-image pre-training," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2023, pp. 101–111.

[45] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," Scientific Data, vol. 6, no. 1, p. 317, 2019.

[46] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," Scientific Reports, vol. 10, no. 1, p. 19549, 2020.

[47] V. V. Danilov, D. Litmanovich, A. Proutski, A. Kirpich, D. Nefaridze, A. Karpovsky, and Y. Gankin, "Automatic scoring of covid-19 severity in x-ray imaging based on a novel deep learning workflow," Scientific reports, vol. 12, no. 1, p. 12791, 2022.

[48] L. Wang, A. Wong, Z. Qiu, P. McInnis, A. Chung, and H. Gunraj, "Actualmed covid-19 chest x-ray dataset initiative," 2020. [Online]. Available: https://github.com/agchung/Actualmed-COVID-chestxray-dataset

[49] M. E. Chowdhury, T. Rahman, A. Khandakar et al., "Can ai help in screening viral and covid-19 pneumonia?" Ieee Access, vol. 8, pp. 132 665–132 676, 2020.

[50] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," arXiv 2006.11988, 2020.

[51] L. Wang, A. Wong, Z. Qiu, P. McInnis, and A. Chung, "Figure 1 covid-19 chest x-ray dataset initiative," 2020. [Online]. Available: https://github.com/agchung/Figure1-COVID-chestxray-dataset

[52] J. Irvin, P. Rajpurkar, M. Ko et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019, pp. 590–597.

[53] G. Shih, C. C. Wu, S. S. Halabi, M. D. Kohli, L. M. Prevedello, T. S. Cook, A. Sharma, J. K. Amorosa, V. Arteaga, M. Galperin-Aizenberg et al., "Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia," Radiology: Artificial Intelligence, vol. 1, no. 1, p. e180041, 2019.

[54] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. Abul Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughaier, M. S. Khan, and M. E. Chowdhury, "Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images," Computers in Biology and Medicine, vol. 132, p. 104319, 2021.

[55] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3942–3951.

[56] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," arXiv preprint arXiv:2106.09685, 2021.

[57] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," arXiv preprint arXiv:1801.06146, 2018.

[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[59] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." Journal of Machine Learning Research, vol. 9, no. 11, 2008.

[60] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," Pattern Recognition Letters, vol. 140, pp. 325–331, 2020.