This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

BiomedCCPL: Causal Conditional Prompt Learning for Biomedical Vision-Language Models

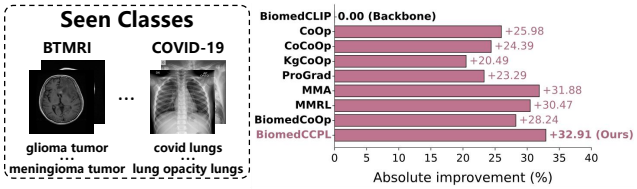
Xueliang Cui[1,2] Juncai Zhang[3,4] Jiacheng Hou[5] Dan Lu[6] Hao Zhang[3,*] Ruxin Wang[3,2,*]
[1] Southern University of Science and Technology, China
[2] Shenzhen University of Advanced Technology, China
[3] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[4] Shantou University, China
[5] Shenzhen International Graduate School, Tsinghua University, China
[6] West China Hospital, Sichuan University, China.
12433289@mail.sustech.edu.cn, {h.zhang10, rx.wang}@siat.ac.cn

Abstract

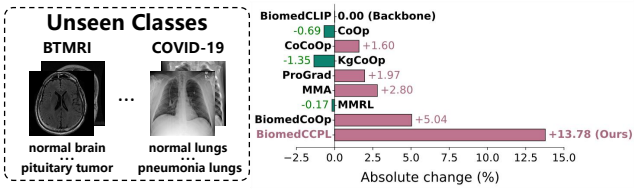
Vision-language models (VLMs) have demonstrated strong potential for adapting to downstream biomedical tasks with limited training samples. However, their generalization to unseen classes within the same dataset remains limited, as the image–text alignment semantics often rely on spurious cues present in seen classes that do not transfer. To tackle this, we propose ***BiomedCCPL (Causal Conditional Prompt Learning)****, a framework that uses VGAP (Visual Grounder with Adaptive Prototype) to generate image-conditional prompts from multi-scale adaptive prototypes and employs SCD (Synergistic Causal Disentanglement) to regularize the generation of image-conditional prompts. Guided by insights from a causal analysis of generalization to unseen classes, SCD leverages multiple synergistic learning objectives to perform front-door adjustment, ensuring that the dynamically generated image-conditional prompts focus on underlying diagnostic image features shared across seen and unseen classes. Experiments on 11 datasets across 9 modalities demonstrate that BiomedCCPL effectively enhances the model's data efficiency and generalization ability. In particular, on the Base-to-Novel task, BiomedCCPL achieves an average HM of 79.98%, surpassing the previous state-of-the-art by 6.45%. Code is available at* https://github.com/burgers0708/BiomedCCPL.

1. Introduction

Vision-language models (VLMs) such as CLIP [30] have demonstrated impressive zero-shot generalization across diverse visual tasks. Motivated by this success, some adaptation strategies (e.g., CoOp [44], MMA [36], and MMRL [13]) have been proposed on natural image datasets to efficiently adapt large-scale VLMs to downstream tasks by introducing learnable components while freezing the pre-trained encoders. Nevertheless, as illustrated in Fig. 1, these methods exhibit distinct shortcomings on biomedical datasets. **First**, several approaches such as CoOp [44], KgCoOp [37], and MMRL tend to overfit: after 16-shot adaptation, their performance on unseen classes falls even below that of the zero-shot BiomedCLIP [40] baseline. **Second**, some methods like CoCoOp [43] and ProGrad [45] alleviate this overfitting but sacrifice performance on seen classes, revealing a trade-off between adaptation and gener-



(a) BiomedCCPL achieves leading performance on seen classes.



(b) BiomedCCPL demonstrates superior generalization to unseen classes.

Figure 1. Comprehensive Comparison: Various 16-shot Adaptation Methods vs. Zero-shot Backbone in the Base-to-Novel Generalization Setting. The results are averaged across 10 datasets.

* Corresponding author

alization. **Third**, although approaches such as MMA avoid generalization degradation, their performance gain on unseen classes remains far smaller than that on seen classes.

We attribute these observations to two main factors in biomedical scenarios. **First**, most existing methods learn an image-unconditional (static) prompt for each seen class, that is, a fixed text prompt matching the common visual patterns shared among training images of that class. Such static prompts often fail to align with the visual features of unseen classes and therefore cannot effectively distinguish images from different unseen categories, leading to poor generalization. **Second**, although some approaches (e.g., CoCoOp) generate image-conditional (dynamic) prompts, they remain susceptible to shortcut learning [12]: the learned prompts may still correlate with non-diagnostic visual features that are discriminative only among seen classes—such as imaging artifacts or device-specific markers that coincidentally co-occur with certain class labels. Inspired by causal inference theory [26], we define *causal knowledge* as the associations—shared and transferable across seen and unseen classes—between text prompts and underlying diagnostic visual features (e.g., fundamental lesion morphology, tissue texture, and abnormal-density regions), whereas other associations are regarded as *non-causal knowledge*. These issues highlight an urgent need for methods capable of dynamically generating image-conditional text prompts aligned with underlying diagnostic features shared across seen and unseen classes.

To address these issues, we propose **BiomedCCPL**, a causal conditional prompt learning framework that tackles both challenges through two synergistic components. **First**, to overcome the limitations of static prompts, we introduce a Visual Grounder with Adaptive Prototype (VGAP) module that dynamically generates image-conditional text prompts. Unlike existing image-conditional methods such as CoCoOp, which generate prompts based on global image features, VGAP dynamically extracts multi-scale adaptive visual prototypes from input images (across shallow, middle, and deep layers) and uses cross-attention to modulate text prompt representations. This allows the text prompts to align with multi-scale semantics in medical images, ranging from coarse anatomical structures to fine-grained pathological textures. **Second**, guided by insights from a Structural Causal Model (SCM) [26] that analyzes the generalization process of the adapted model to unseen classes, we propose a Synergistic Causal Disentanglement (SCD) module implementing the front-door adjustment principle to reduce spurious associations between dynamically generated prompts and non-diagnostic visual features. SCD explicitly separates prompt representations into causal and non-causal components and employs a dual-pathway architecture with synergistic objectives: the causal pathway is trained for classification while being constrained to remain orthogonal to the non-causal pathway, guiding it toward underlying diagnostic features; the non-causal pathway is encouraged to produce uniform predictions, limiting the impact of irrelevant information. The two modules work synergistically: VGAP provides multi-scale fine-grained visual grounding that guides the prompt generation process, while SCD regularizes this process to emphasize causally relevant information and suppress spurious correlations.

Our main contributions are summarized as follows:

- **Causal Conditional Prompt Learning Framework.** We present BiomedCCPL, a framework combining VGAP, which dynamically generates prompts based on multi-scale adaptive visual prototypes, and SCD, which disentangles them into causal and non-causal components for front-door adjustment via multiple synergistic objectives.
- **Dual-Module Collaborative Design.** In BiomedCCPL, we propose a collaborative design where VGAP conditions text prompts on multi-scale, fine-grained visual prototypes, providing visual cues that SCD uses to separate prompt representations into causal and non-causal components, while SCD regularizes conditioning to suppress non-causal correlations and enhance generalization.
- **Extensive Empirical Validation.** On few-shot tasks, BiomedCCPL achieves better performance using fewer training samples. In Base-to-Novel evaluation, it attains a Harmonic Mean (HM) of 79.98%, surpassing the next-best method by 6.45%. Furthermore, saliency maps show that BiomedCCPL consistently focuses on clinically relevant regions, enhancing model interpretability.

2. Related Work

2.1. Biomedical Vision-Language Models

Pre-trained on about 400 million image–text pairs, the vision–language model CLIP [30] has become a widely adopted foundation model for zero-shot and few-shot vision–language tasks [22, 33, 42]. Following CLIP's success, several biomedical VLMs have emerged, including PMC-CLIP [21], PubMedCLIP [10], and BiomedCLIP [40]. Unlike CLIP, BiomedCLIP adopts BERT [8] as its text encoder instead of a standard Transformer, while keeping the Vision Transformer [9] as the visual encoder. Pre-trained on PMC-15M, a dataset of 15 million biomedical image-text pairs, BiomedCLIP has achieved remarkable performance across diverse biomedical vision–language tasks. Although these VLMs achieve strong zero-shot performance, task-specific few-shot adaptation is often necessary to further improve downstream performance.

2.2. Few-shot Adaptation of VLMs

Recent studies have explored parameter-efficient strategies for adapting VLMs to downstream tasks with limited labeled data. Adapter-based approaches [11, 36, 39] in-

sert lightweight adapters into encoders to fine-tune feature representations with minimal parameter updates. Prompt-based approaches [17, 18, 37, 43–45] inject learnable tokens to provide task-specific context that steers the model's predictions while keeping the backbone frozen. Among them, CoCoOp [43] is most relevant to our work, since both methods condition prompts on image features. However, CoCoOp uses global image features, which blurs local diagnostic details through excessive averaging and is susceptible to being dominated by spurious signals. In the biomedical domain, XCoOp [3] and BiomedCoOp [20] adapt VLMs to specialized medical tasks by integrating domain-specific prior knowledge. In contrast, our BiomedCCPL requires no external knowledge and learns task-relevant representations purely through end-to-end training.

2.3. Causal Inference

Recently, causal inference [26] has been widely adopted in representation learning [2, 31]. Existing studies mainly explore two directions: counterfactual reasoning [5, 38] and causal intervention [29, 41]. Among causal intervention approaches, ICL-MSR [29] applies back-door adjustment by using observable semantic attributes as proxies for confounders, while CDC [41] adopts front-door adjustment to suppress spurious correlations. Unlike CDC, which disentangles semantics based on multiple global image features obtained via different augmentation strategies, our BiomedCCPL disentangles causal and non-causal semantics using multi-scale local image features and synergistic learning.

3. Methodology

The overall architecture of BiomedCCPL is illustrated in Fig. 2. Built upon the BiomedCLIP backbone, BiomedCCPL comprises two synergistic components designed to learn disentangled causal prompt representations that align with fine-grained, multi-scale image semantics. On the one hand, the VGAP module generates image-conditional prompts from multi-scale visual prototypes, capturing features from shallow to deep layers. This allows the text prompts to align with multi-scale semantics in medical images, ranging from coarse anatomical structures to fine-grained pathological textures. On the other hand, the SCD module regularizes prompt generation, guiding focus toward diagnostic features transferable across seen and unseen classes while reducing spurious correlations.

3.1. Preliminary

CLIP's hand-crafted prompts are fixed and lack adaptability to domain-specific visual concepts. To overcome this limitation, prompt learning methods replace static templates with learnable context vectors that can be optimized for a specific dataset. Specifically, for the c-th class, a continuous prompt $\mathbf{T}_c$ is formed by prepending shared context vectors to the class name embedding. The probability that an image $\mathbf{I}_j$ belongs to the c-th class is computed as:

$$p(\boldsymbol{y}_j = c|\mathbf{F}_I^j) = \frac{\exp(\cos(\mathbf{F}_T^c, \mathbf{F}_I^j)/\tau)}{\sum_{i=1}^{C} \exp(\cos(\mathbf{F}_T^i, \mathbf{F}_I^j)/\tau)}, \quad (1)$$

where $\mathbf{F}_I^j$ denotes the image feature of $\mathbf{I}_j$, $\mathbf{F}_T^c$ is the text feature of $\mathbf{T}_c$, and τ is a temperature parameter. During few-shot adaptation, only the context vectors are updated while the pretrained encoders remain frozen.

3.2. Visual Grounder with Adaptive Prototype

Most existing methods learn an image-unconditional (static) prompt for each seen class, i.e., a fixed text embedding aligned with the common visual patterns of the training images. Such static prompts fail to align with the visual features of unseen classes, hindering discrimination between different unseen classes and resulting in poor generalization. To address this limitation, some methods generate image-conditional prompts based on global image features (e.g., CoCoOp). However, relying on a single global token leads to excessive averaging, blurring local diagnostic features (e.g., a tiny lung nodule) [24]. Moreover, the global image representation is also easily dominated by spurious signals in the image (e.g., watermarks or background artifacts), amplifying shortcut learning [25]. To address these issues, we propose VGAP, generating prompts based on multi-scale visual prototypes, preserving fine-grained and hierarchically structured diagnostic features.

The VGAP module operates through a prototype-based cross-attention mechanism. For clarity, we describe the interaction process between the prompt for the i-th class and the j-th image. Performing attention on every patch token individually can be noisy and may fail to capture coherent local patterns. To address this, VGAP introduces adaptive visual prototypes. Unlike traditional clustering-based prototypes, which are derived solely from visual similarity and are independent of the downstream task, VGAP's prototypes are produced by a lightweight learnable module and refined through backpropagation. This end-to-end optimization makes the prototypes adaptive to the classification objective, yielding task-relevant representations rather than purely geometric ones. Specifically, VGAP reduces M patch tokens from l-th layer, $\mathcal{V}_{j,l} = \{\boldsymbol{v}_1^{j,l}, \boldsymbol{v}_2^{j,l}, ..., \boldsymbol{v}_M^{j,l}\}$, into N adaptive visual prototypes $\mathbf{P}_l = \{\boldsymbol{p}_1^l, \boldsymbol{p}_2^l, ..., \boldsymbol{p}_N^l\}$ ($N \ll M$). This is achieved by computing a contribution matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ through a lightweight network:

$$\mathbf{A} = \mathrm{SoftMax}(\mathrm{ReLU}(\mathrm{LayerNorm}(\mathcal{V}_{j,l}\mathbf{W}_1))\mathbf{W}_2), \quad (2)$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_v \times d_h}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_h \times N}$, with d_h being the dimension of an intermediate hidden layer. Each element A_{ij} quantifies the contribution of the i-th patch token

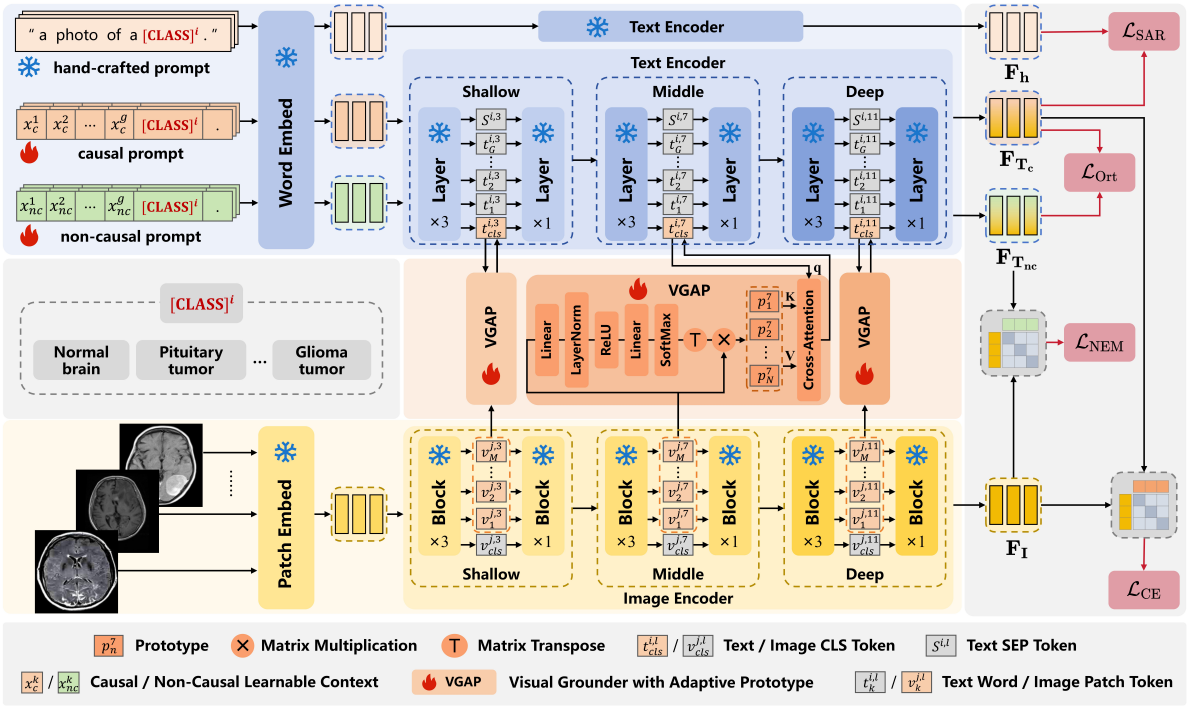


Figure 2. **Overview of the BiomedCCPL framework, built on the BiomedCLIP backbone.** The VGAP (Visual Grounder with Adaptive Prototype) module generates image-conditional prompts from multi-scale adaptive visual prototypes, capturing hierarchical visual patterns from shallow to deep layers. The SCD (Synergistic Causal Disentanglement) module regularizes these prompts via cooperative objectives, encouraging focus on diagnostic features transferable across seen and unseen classes while suppressing spurious correlations.

to the formation of the j-th prototype. The prototypes $\mathbf{P}_l$ are derived as a weighted aggregation of the patch tokens:

$$\mathbf{P}_l = \mathbf{A}^\top \mathcal{V}_{j,l}. \tag{3}$$

Subsequently, the text [CLS] token $\boldsymbol{t}_{cls}^{i,l}$ from the text encoder's l-th layer queries $\mathbf{P}_l$ via cross-attention to produce a visually-grounded representation $\boldsymbol{z}^{i,l}$. The query ($\boldsymbol{q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) are formed as:

$$\boldsymbol{q} = \boldsymbol{t}_{cls}^{i,l}\mathbf{W}_q, \quad \mathbf{K} = \mathbf{P}_l\mathbf{W}_k, \quad \mathbf{V} = \mathbf{P}_l\mathbf{W}_v, \tag{4}$$

and the fused representation ($\boldsymbol{z}^{i,l}$) is computed as:

$$\boldsymbol{z}^{i,l} = \text{SoftMax}\left(\frac{\boldsymbol{q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \tag{5}$$

The fused representation $\boldsymbol{z}^{i,l}$ is then smoothly integrated into $\boldsymbol{t}_{cls}^{i,l}$ via a momentum-based update:

$$\boldsymbol{t}_{cls}^{i,l,*} = \alpha \cdot \boldsymbol{t}_{cls}^{i,l} + (1-\alpha)\cdot \boldsymbol{z}^{i,l}. \tag{6}$$

This updated token $\boldsymbol{t}_{cls}^{i,l,*}$ is then passed to the subsequent layer. The momentum coefficient α is a hyperparameter that controls the degree of visual information infusion. The VGAP process is applied at a specific layer within each of the encoder's three scales (shallow, middle, and deep), using indices $l \in \{3, 7, 11\}$ as justified by our ablation studies (see Supplementary Material). This hierarchical fusion ensures that the final text representation is robustly grounded in diagnostic details across all scales, from coarse anatomical structures to fine-grained pathological textures.

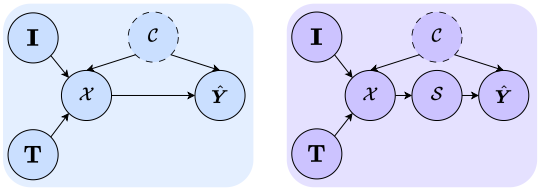


Figure 3. Structural causal models for seen-to-unseen generalization. Solid and dashed circles represent observable and unobservable variables, respectively.

3.3. Synergistic Causal Disentanglement

Although dynamic prompts are more effective than static ones, they can still be influenced by non-diagnostic visual features, such as device-specific markers or imaging artifacts. To analyze this phenomenon rigorously, as illus-

trated in Fig. 3, we construct an SCM [26] that models the causal relationships among variables to revisit CLIP's generalization from seen to unseen classes. To formalize our causal analysis, we define the variables in Fig. 3 as follows. Herein, $\mathbf{I}$ denotes a mini-batch of unseen images, and $\mathbf{T}$ the learned text prompts. We define the *Image–Text Alignment Semantics* $\mathcal{X}$ as the semantic information shared between the visual input $\mathbf{I}$ and the text prompt $\mathbf{T}$. We conceptualize the confounder $\mathcal{C}$ as the non-diagnostic visual features present in the seen classes, such as device-specific markers or imaging artifacts. This confounder acts as a common cause that influences both the formation of $\mathcal{X}$ (path $\mathcal{C} \rightarrow \mathcal{X}$) and the predicted labels $\hat{\boldsymbol{Y}}$ (path $\mathcal{C} \rightarrow \hat{\boldsymbol{Y}}$), opening a back-door path $\mathcal{X} \leftarrow \mathcal{C} \rightarrow \hat{\boldsymbol{Y}}$ that introduces spurious correlations and impairs generalization to unseen classes. Since the confounder $\mathcal{C}$ is unobserved, directly applying back-door adjustment to remove its influence is infeasible. Therefore, we instead adopt the front-door criterion by introducing an intermediate variable, the *Image–Text Alignment Causal Semantics* $\mathcal{S}$, which mediates the causal effect between $(\mathbf{I}, \mathbf{T})$ and $\hat{\boldsymbol{Y}}$. $\mathcal{S}$ captures the associations between text prompts and underlying diagnostic visual features (e.g., fundamental lesion morphology, tissue texture patterns, and abnormal-density regions) that are shared and transferable across seen and unseen classes. Specifically, for the j-th image $\mathbf{I}_j$, the causal effect can be estimated using the front-door formula:

$$P(\hat{\boldsymbol{y}}_j = c \mid \mathrm{do}(\mathbf{I}_j, \mathbf{T})) = \sum_{s \in \mathcal{S}} P(s \mid \mathbf{I}_j, \mathbf{T}) P(c \mid s). \quad (7)$$

In practice, $\mathcal{S}$ is estimated by the collaboration of VGAP and SCD, and the similarity scores derived from $\mathcal{S}$ are then used as differentiable logits to compute $P(c \mid \mathcal{S})$.

To estimate $\mathcal{S}$, the text prompts are disentangled into causal and non-causal components using three synergistic objectives: $\mathcal{L}_{\text{CE}}$, $\mathcal{L}_{\text{NEM}}$, and $\mathcal{L}_{\text{Ort}}$. Herein, $\mathcal{L}_{\text{NEM}}$ prevents the non-causal pathway from producing deterministic discriminative signals, $\mathcal{L}_{\text{Ort}}$ enforces orthogonality between causal and non-causal pathways, and $\mathcal{L}_{\text{CE}}$ guides the causal pathway to make accurate predictions within this constrained space. Given a mini-batch of images $\mathbf{I}$ and the corresponding causal and non-causal prompts for C classes—e.g., for the i-th class, the causal prompt $\mathbf{T_c^i} = \{\boldsymbol{x}_c^1, \boldsymbol{x}_c^2, ..., \boldsymbol{x}_c^g, [\text{CLASS}]^i, \text{'.'}\}$ and the non-causal prompt $\mathbf{T_{nc}^i} = \{\boldsymbol{x}_{nc}^1, \boldsymbol{x}_{nc}^2, ..., \boldsymbol{x}_{nc}^g, [\text{CLASS}]^i, \text{'.'}\}$—the encoders produce the visual features $\mathbf{F_I} \in \mathbb{R}^{B \times d_{vt}}$, the causal prompt features $\mathbf{F_{T_c}} \in \mathbb{R}^{C \times d_{vt}}$, and the non-causal prompt features $\mathbf{F_{T_{nc}}} \in \mathbb{R}^{C \times d_{vt}}$. The following paragraphs detail the formulation of each objective.

Cross-Entropy Loss. The causal pathway is optimized using the standard cross-entropy objective $\mathcal{L}_{\text{CE}}$, which encourages the causal prompts to capture discriminative diagnostic information. The class probability distribution predicted by the causal pathway is:

$$\mathbf{Q}_c = \mathrm{SoftMax}(\mathbf{F_I} \mathbf{F}_{\mathbf{T_c}}^{\top} / \tau), \quad (8)$$

where τ is the temperature. The loss is computed as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{B} \sum (\mathbf{Y} \odot \log \mathbf{Q}_c), \quad (9)$$

where $\odot$ dennotes the element-wise product.

Non-causal Entropy Maximization. To render the non-causal prompt representations uninformative, we employ the loss $\mathcal{L}_{\text{NEM}}$ to drive the non-causal pathway toward producing a uniform probability distribution. The predicted distribution $\mathbf{Q}_{nc}$ is:

$$\mathbf{Q}_{nc} = \mathrm{SoftMax}(\mathbf{F_I} \mathbf{F}_{\mathbf{T_{nc}}}^{\top} / \tau), \quad (10)$$

where τ is a temperature parameter. The $\mathcal{L}_{\text{NEM}}$ is then defined as the KL divergence between the predicted distribution $\mathbf{Q}_{nc}$ and a discrete uniform distribution $\mathcal{U}$:

$$\mathcal{L}_{\text{NEM}} = D_{\text{KL}}(\mathbf{Q}_{nc} || \mathcal{U}) = \sum_{b=1}^{B} \sum_{i=1}^{C} q_{nc,b,i} \log \frac{q_{nc,b,i}}{u_i}, \quad (11)$$

where $q_{nc,b,i}$ is the predicted probability for the i-th class of the b-th sample, and $u_i = 1/C$ is the uniform probability. Minimizing this loss maximizes the prediction entropy, making the non-causal pathway non-predictive.

Orthogonality Constraint. To separate the representational roles of the two pathways, we enforce intra-class orthogonality between causal and non-causal prompts. The orthogonality loss $\mathcal{L}_{\text{Ort}}$ is given by:

$$\mathcal{L}_{\text{Ort}} = \frac{1}{C} || \operatorname{diag}(\mathbf{F_{T_c}} \mathbf{F}_{\mathbf{T_{nc}}}^{\top}) ||_2^2, \quad (12)$$

This constraint separates the two prompt subspaces and regularizes the causal pathway by suppressing features associated with non-causal information.

3.4. Overall Learning Objective

Semantic Anchor Regularization. We introduce the loss $\mathcal{L}_{\text{SAR}}$ to stabilize BiomedCCPL training by using the encoded features $\mathbf{T_h}$ of hand-crafted class-specific prompts (e.g., "a photo of a [CLASS]") as semantic anchors:

$$\mathcal{L}_{\text{SAR}} = 1 - \frac{1}{C} \mathrm{Tr}(\mathbf{F_{T_c}} \mathbf{F}_{\mathbf{T_h}}^{\top}), \quad (13)$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of the matrix.

Overall Loss. Finally, the overall learning objective for our BiomedCCPL framework is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{SAR}} + \lambda_1 \mathcal{L}_{\text{NEM}} + \lambda_2 \mathcal{L}_{\text{Ort}}, \quad (14)$$

where λ_1 and λ_2 are balancing hyperparameters. During testing, predictions are made solely using the outputs from the causal pathway, discarding the non-causal ones.

4. Experiments

4.1. Experimental Settings

We perform a systematic evaluation of the few-shot adaptation and generalization capabilities of BiomedCCPL on 11 biomedical datasets across 9 modalities.

Baselines. We compare BiomedCCPL with seven competitive methods, including prompt-based techniques (i.e., CoOp [44], CoCoOp [43], KgCoOp [37], ProGrad [45], MMRL [13], BiomedCoOp [20]) and adapter-based approach (i.e., MMA [36]). Baselines are grouped by cross-modal interaction: with (CoCoOp, MMA, MMRL) or without (CoOp, KgCoOp, ProGrad, and BiomedCoOp).

Datasets. To thoroughly assess our proposed method, we employ the 11 diverse medical imaging datasets introduced in [20], which cover 10 distinct organs and 9 modalities. Specifically, these datasets are categorized by modality as follows: Computerized Tomography (CTKidney [14]), Dermatoscopy (DermaMNIST [7, 34]), Endoscopy (Kvasir [27]), Fundus Photography (RETINA [19, 28]), Histopathology (LC25000 [4], CHMNIST [15]), Magnetic Resonance Imaging (BTMRI [23]), Optical Coherence Tomography (OCTMNIST [16]), Ultrasound (BUSI [1]), and X-Ray (COVID-QU-Ex [32], KneeXray [6]).

Implementation details. Our BiomedCCPL model, built upon a ViT-B/16 backbone, was trained for 50 epochs on both few-shot and base-to-novel tasks. We initialized the learnable context with the prompt embedding of "a photo of a" and set the number of prototypes to 14 (the square root of 196 image patches). The model was trained using an SGD optimizer, with a learning rate of 0.0025 and a batch size of 1. The hyperparameters λ_1, λ_2, and α were tuned on a validation set (see Supplementary Material for details). For a fair comparison, all methods share the same BiomedCLIP backbone, and results are averaged over three runs with different seeds. All experiments were performed on a single NVIDIA GeForce RTX 4090 GPU.

4.2. Few-shot Evaluation

Considering the common challenge of data scarcity in biomedical applications, we conduct few-shot experiments with support sets of $K \in \{1, 2, 4, 8, 16\}$ examples per class to evaluate the model's adaptation capabilities.

As illustrated by the **AVERAGE** subplot in Fig. 4, our BiomedCCPL consistently demonstrates a performance advantage across all shot settings. Furthermore, BiomedCCPL exhibits remarkable data efficiency, achieving higher performance than competing methods with fewer training samples. For example, even with only 1 training sample per class, our model achieves 62.17%, surpassing the 2-shot performance of BiomedCoOp (58.55%). Similarly, our model trained with 8 samples per class (77.22%) outperforms MMA's 16-shot baseline (75.24%). Under various K-shot settings, the performance advantage of BiomedCCPL over competing methods becomes more pronounced as K increases. Such data efficiency and model scalability are particularly valuable for rare diseases or scenarios with limited annotated data.

4.3. Base-to-Novel Generalization

In this section, we evaluate BiomedCCPL's generalization from base (i.e., seen) to novel (i.e., unseen) classes within a single dataset. Consistent with previous studies, all models are trained on base classes under a 16-shot setting and evaluated separately on base and novel classes.

The average performance comparison of BiomedCCPL and other methods across 10 datasets (excluding BUSI due to limited class diversity) is presented in Fig. 5, and the detailed results are reported in Table 1. On average, BiomedCCPL achieves a Harmonic Mean of 79.98%, outperforming the next-best method, MMA, by 6.45%. BiomedCCPL not only achieves the best results on base classes but also demonstrates remarkable generalization to novel classes, increasing novel class accuracy by 8.74% over BiomedCoOp. These results suggest that BiomedCCPL effectively captures causal knowledge shared between base and novel classes rather than spurious correlations specific to the base classes. This robust generalization is particularly important for clinical applications, where models must reliably handle unseen pathology variations.

4.4. Ablation Experiments

4.4.1. Effect of Different Components

We conducted an ablation study to evaluate the contributions of BiomedCCPL's core modules: SAR (Semantic Anchor Regularization), VGAP (Visual Grounder with Adaptive Prototype), and SCD (Synergistic Causal Disentanglement). The results are presented in Tab. 2, which are averaged across all 11 datasets. For the Base-to-Novel task, we started from the baseline CoOp and incrementally added each module to examine its contribution. Adding SCD improves the novel class accuracy from 64.72% to 75.69%, followed by VGAP (to 73.26%) and SAR (to 65.68%), emphasizing the importance of causal disentanglement in miti-

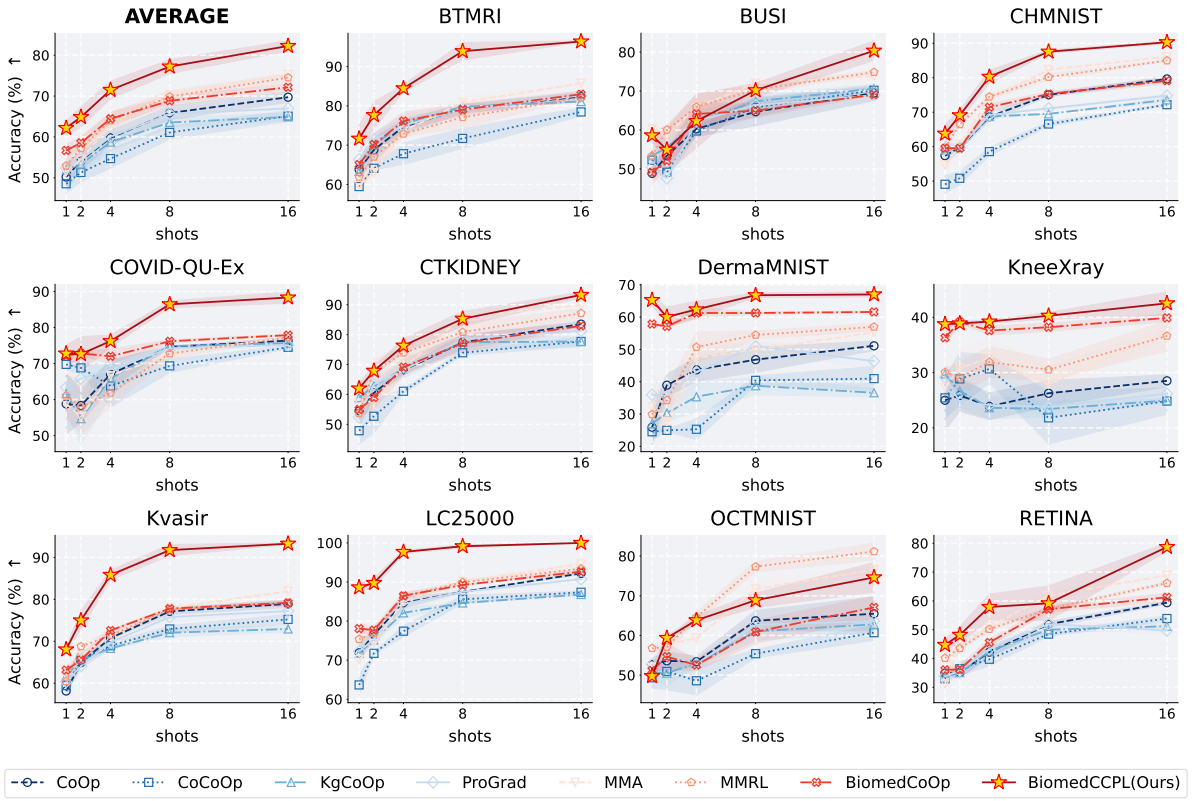


Figure 4. Few-shot performance of BiomedCCPL against state-of-the-art methods across 11 datasets. The top-left **AVERAGE** subplot, which summarizes the average performance, shows that BiomedCCPL consistently achieves superior accuracy across all shot settings. Each point represents the mean accuracy over 3 random seeds, with the shaded error bands representing the standard error.

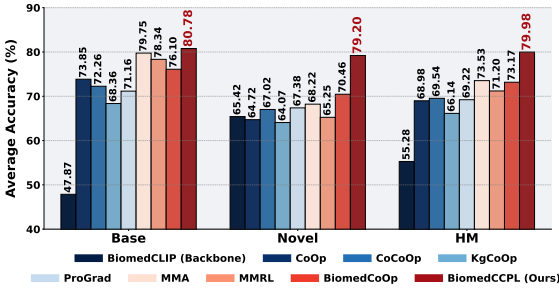


Figure 5. Base-to-Novel generalization performance, showing average results over 10 datasets. Harmonic Mean (HM) measures balanced performance across base and novel classes.

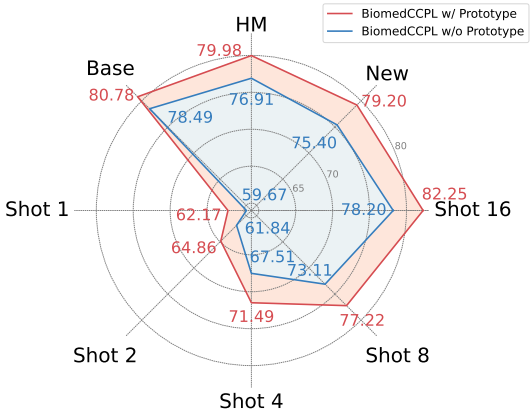


Figure 6. Ablation study of the prototype in BiomedCCPL on Base-to-Novel and Few-shot tasks, where the red and blue lines show performance with and without the prototype, respectively.

gating spurious correlations from base classes. For the Few-shot task, we started from the full BiomedCCPL model and progressively removed each module to assess its contribution. Removing VGAP decreases the accuracy from 82.25% to 71.13% at 16 shots, followed by SCD (to 79.80%) and SAR (to 77.48%), highlighting the importance of adaptive visual grounding for leveraging limited supervised data. In summary, the results indicate that SAR stabilizes learning, VGAP enhances learning from limited examples, and SCD drives generalization to unseen classes. Moreover, combining any two components consistently yields higher performance than using either one alone, demonstrating that the modules offer complementary benefits when integrated.

Table 1. Base-to-Novel generalization performance, showing detailed results on 10 datasets. Harmonic Mean (HM) measures balanced performance across base and novel classes. Primary and secondary results are highlighted in **blue** and gray, respectively.

Method	BTMRI			CHMNIST			COVID-QU-Ex			CTKIDNEY			DermaMNIST		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
BiomedCLIP $_{\text{Backbone}}$	40.88	96.18	57.37	37.63	40.69	39.10	54.00	89.43	67.34	38.55	52.99	44.63	35.19	49.59	41.17
CoOp $_{\text{IJCV'22}}$	82.25	94.51	87.95	89.41	35.11	50.42	75.92	90.06	82.39	82.26	67.92	74.41	48.06	59.41	53.14
CoCoOp $_{\text{CVPR'22}}$	77.88	94.84	85.53	87.77	42.51	57.28	77.27	87.63	82.12	81.96	56.56	66.93	42.88	60.64	50.24
KgCoOp $_{\text{CVPR'23}}$	78.04	95.05	85.71	75.49	38.70	51.17	75.42	89.61	81.90	81.67	58.45	68.14	36.41	47.33	41.16
ProGrad $_{\text{ICCV'23}}$	81.67	93.89	87.35	82.89	44.46	57.88	75.56	**90.68**	82.43	82.50	66.18	73.44	36.41	64.70	46.60
MMA $_{\text{CVPR'24}}$	**86.84**	95.46	**90.95**	93.57	33.20	49.01	79.84	90.57	84.87	87.90	54.84	67.54	53.88	45.80	49.51
MMRL $_{\text{CVPR'25}}$	83.00	96.04	89.05	91.93	36.53	52.28	79.21	89.58	84.08	87.48	45.04	59.46	49.76	42.39	45.78
BiomedCoOp $_{\text{CVPR'25}}$	82.21	96.44	88.76	89.09	**45.21**	59.98	74.88	90.38	81.90	86.21	77.16	81.43	**58.66**	55.72	57.15
BiomedCCPL $_{\text{Ours}}$	85.83	**96.58**	90.89	**94.77**	44.95	**60.98**	**84.71**	89.81	**87.19**	**92.06**	**77.35**	**84.07**	47.25	**81.00**	**59.68**

Method	KneeXray			Kvasir			LC25000			OCTMNIST			RETINA		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
BiomedCLIP $_{\text{Backbone}}$	35.89	71.90	47.88	75.00	60.50	66.97	59.73	87.60	71.03	56.60	50.00	53.10	45.18	55.28	49.72
CoOp $_{\text{IJCV'22}}$	38.25	47.69	42.45	86.22	58.06	69.39	90.12	87.57	88.83	75.00	50.00	60.00	70.98	56.90	63.16
CoCoOp $_{\text{CVPR'22}}$	34.08	63.02	44.24	85.94	53.95	66.29	88.33	95.02	91.55	79.60	50.47	61.77	66.88	65.56	66.21
KgCoOp $_{\text{CVPR'23}}$	37.94	61.19	46.84	81.56	59.00	68.47	88.13	86.43	87.27	68.13	50.00	57.67	60.77	54.91	57.69
ProGrad $_{\text{ICCV'23}}$	40.57	58.76	48.00	82.72	60.28	69.74	89.34	87.57	88.45	73.40	50.00	59.48	66.51	57.32	61.57
MMA $_{\text{CVPR'24}}$	**43.20**	75.43	**54.94**	89.00	58.39	70.52	95.76	91.60	93.63	86.73	50.00	63.43	80.73	86.88	83.69
MMRL $_{\text{CVPR'25}}$	39.51	72.75	51.21	86.28	**63.22**	**72.97**	95.61	92.97	94.27	**88.40**	50.00	63.87	82.20	63.94	71.93
BiomedCoOp $_{\text{CVPR'25}}$	42.65	72.63	53.74	86.83	59.11	70.34	92.90	94.10	93.50	78.93	50.40	61.52	68.67	63.41	65.94
BiomedCCPL $_{\text{Ours}}$	40.06	**81.39**	53.69	**92.67**	59.39	72.39	**99.43**	**98.03**	**98.73**	83.80	**73.67**	**78.41**	**87.20**	**89.82**	**88.49**

Table 2. Ablation study of our BiomedCCPL model's components (SAR, VGAP, and SCD). The table presents the average accuracy (%) across all 11 datasets on Few-shot and Base-to-Novel tasks.

Components			Base-to-Novel			Few-shot				
SAR	VGAP	SCD	Base	Novel	HM	1	2	4	8	16
✗	✗	✗	73.85	64.72	68.98	50.18	54.17	59.77	65.85	69.72
✓	✗	✗	75.78	65.68	70.37	51.97	54.98	59.84	66.09	68.64
✗	✓	✗	79.08	73.26	76.06	56.76	59.87	67.96	74.27	79.95
✗	✗	✓	75.56	75.69	75.62	54.69	57.26	62.51	68.17	71.30
✓	✓	✗	79.12	73.87	76.40	56.67	61.90	67.55	74.05	79.80
✓	✗	✓	75.74	76.20	75.97	54.71	57.21	62.56	68.34	71.13
✗	✓	✓	80.83	77.88	79.33	60.42	63.94	66.40	68.85	77.48
✓	✓	✓	**80.78**	**79.20**	**79.98**	**62.17**	**64.86**	**71.49**	**77.22**	**82.25**

4.4.2. Effect of Prototype

As illustrated in Fig. 6, we conducted an ablation study on the prototype mechanism within the VGAP (Visual Grounder with Adaptive Prototype) module. The results show that the inclusion of the prototype mechanism comprehensively boosts model performance across the board. This validates that the prototype mechanism is crucial in enhancing the model's few-shot learning capabilities and its cross-category generalization ability. We believe this is because the prototypes help prevent the model from overfitting to excessively detailed and noisy factors in the images, effectively serving as both a semantic aggregator and a noise filter. Furthermore, we also tested the impact of varying the number of prototypes on model performance, with detailed results available in the Supplementary Material.

4.4.3. Visual Interpretability

Using ScoreCAM [35], we generated saliency maps from three perspectives: comparing adaptation strategies, evaluating decoupled pathways, and analyzing specific disease classes. Results demonstrate our causal pathway consistently and precisely localizes lesions (see Supplementary Material Figs. 9 to 11). This precision crucially enhances model interpretability and fosters clinical acceptance.

5. Conclusion

In this work, we revisit the generalization challenge of adapting biomedical vision–language models from a causal perspective using an SCM and show that conventional prompt-tuning tends to encode spurious correlations limited to seen classes. To address this, we propose BiomedCCPL, a causal conditional prompt learning framework that dynamically generates image-conditional prompts based on adaptive visual prototypes via the VGAP module and regularizes them with the SCD module to emphasize transferable diagnostic features while limiting spurious associations. Extensive experiments demonstrate that BiomedCCPL improves both few-shot adaptation and cross-class generalization, offering a practical solution for building reliable vision-language models under data-scarce scenarios.

6. Acknowledgements

This paper was supported by the National Key R&D Program of China (2022YFA1008300), the National Natural Science Foundation of China (12471308, 62472415), the Guangdong Basic and Applied Basic Research Foundation (2026B1515020017), and the Excellent Young Scholars of Shenzhen (RCYX20231211090247060).

References

[1] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. 6, 4

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 3

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

[4] Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019. 6, 4

[5] Long Chen, Yuhang Zheng, Yulei Niu, Hanwang Zhang, and Jun Xiao. Counterfactual samples synthesizing and training for robust visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13218–13234, 2023. 3

[6] Pingjun Chen. Knee osteoarthritis severity grading dataset. *Mendeley Data*, 1(10.17632):30784984, 2018. 6, 4

[7] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 6, 4

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 2

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[10] Sedigheh Eslami, Christoph Meinel, and Gerard 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*, pages 1181–1193, 2023. 2

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

[12] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2

[13] Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25015–25025, 2025. 1, 6

[14] Md Nazmul Islam, Mehedi Hasan, Md Kabir Hossain, Md Golam Rabiul Alam, Md Zia Uddin, and Ahmet Soylu. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ct-radiography. *Scientific Reports*, 12(1):11440, 2022. 6, 4

[15] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific Reports*, 6(1): 1–11, 2016. 6, 4

[16] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018. 6, 4

[17] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 3

[18] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 3

[19] Thomas Köhler, Attila Budai, Martin F Kraus, Jan Odstrčilik, Georg Michelson, and Joachim Hornegger. Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In *Proceedings of the IEEE International Symposium on Computer-based Medical Systems*, pages 95–100, 2013. 6, 4

[20] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Biomedcoop: Learning to prompt for biomedical vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14766–14776, 2025. 3, 6

[21] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536, 2023. 2

[22] Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4744–4754, 2025. 2

[23] Msoud Nickparvar. Brain tumor mri dataset, 2021. 6, 4

[24] Rajendran Nirthika, Siyamalan Manivannan, Amirthalingam Ramanan, and Ruixuan Wang. Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study. *Neural Computing and Applications*, 34(7): 5321–5347, 2022. 3

[25] Frederik Pahde, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Ensuring medical ai safety: interpretability-driven detection and mitigation of spurious model behavior and associated data. *Machine learning*, 114 (9):206, 2025. 3

[26] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(none):96 – 146, 2009. 2, 3, 5

[27] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir: A multiclass image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the ACM on Multimedia Systems Conference*, pages 164–169, 2017. 6, 4

[28] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018. 6, 4

[29] Wenwen Qiang, Jiangmeng Li, Changwen Zheng, Bing Su, and Hui Xiong. Interventional contrastive learning with meta semantic regularizer. In *Proceedings of the International Conference on Machine Learning*, pages 18018–18030. PMLR, 2022. 3

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021. 1, 2

[31] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. 3

[32] Anas M Tahir, Muhammad EH Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M Sohel Rahman, Somaya Al-Maadeed, et al. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in Biology and Medicine*, 139:105002, 2021. 6, 4

[33] Haoran Tang, Meng Cao, Jinfa Huang, Ruyang Liu, Peng Jin, Ge Li, and Xiaodan Liang. Muse: Mamba is efficient multi-scale learner for text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7238–7246, 2025. 2

[34] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):1–9, 2018. 6, 4

[35] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020. 8, 2

[36] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23826–23837, 2024. 1, 2, 6

[37] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023. 1, 3, 6

[38] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15404–15414, 2021. 3

[39] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *Proceedings of the European Conference on Computer Vision*, pages 493–510, 2022. 2

[40] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 1, 2

[41] Yanan Zhang, Jiangmeng Li, Lixiang Liu, and Wenwen Qiang. Rethinking misalignment in vision-language model adaptation from a causal perspective. *Advances in Neural Information Processing Systems*, 37:39224–39248, 2024. 3

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

[43] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1, 3, 6

[44] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 6

[45] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023. 1, 3, 6