TriCLIP: Tri-stage Calibration of Prompts, Evidence and Fusion for Robust Biomedical Vision–Language Models

Fuxian Sui[1,2], **Yongping Du**[1,2,*] and **Deyi Li**[1,2]

[1]College of Computer Science, Beijing University of Technology, Beijing, China
[2]Beijing Artificial Intelligence Institute, Beijing University of Technology, Beijing, China
fuxiansui@emails.bjut.edu.cn, ypdu@bjut.edu.cn, ldeyi@emails.bjut.edu.cn

Abstract

Large visual-language models demonstrate exceptional transferability in biomedical image analysis, yet their robustness remains challenging in generalization scenarios such as few-shot learning. This limitation stems from an overly entangled prompt space, causing instability in the conditional distribution, coupled with saliency bias that drives attention to dominate key regions, thereby restricting the coverage of visual representations. Furthermore, cross-modal fusion often struggles to extract critical information signals from redundant alignments. To address these structural bottlenecks, we propose TriCLIP, a decomposition-based framework designed to reconstruct the entire information flow across prompt conditioning, visual evidence extraction, and cross-modal fusion. Prompt Distribution Regularization enhances robustness to linguistic variations by optimizing class-conditional random context perturbations for continuous prompts. Feedback-driven Counterfactual Masking constructs a feedback-driven reverse saliency view that suppresses dominant evidence to promote complementary cue learning and mitigate attention bias. Finally, Selective Regulation Fusion regulates fusion selectivity through paired reinforcement-suppression principles, amplifying information correspondence while suppressing redundant channels. Extensive experiments on diverse biomedical benchmarks indicate that the proposed TriCLIP consistently outperforms existing competitive methods.

1 Introduction

Biomedical imaging is a cornerstone of clinical diagnosis, yet it remains substantially harder to model than natural-image recognition due to sparse semantic distributions, subtle discriminative cues, and the reliance on strong biomedical priors. Recently, large-scale vision–language models (VLMs) have introduced a unified paradigm for biomedical image analysis by leveraging cross-modal alignment and strong transferability [Radford *et al.*, 2021; Li *et al.*, 2022; Huang *et al.*, 2024]. These models have shown promise for multimodal diagnosis, report generation, and disease classification. However, when transferred to biomedical scenarios, general-purpose VLMs often exhibit pronounced performance degradation, especially under high-generalization-demand settings such as base-to-novel generalization (training on base classes and evaluating on unseen novel classes with limited or no additional supervision) and few-shot learning (adapting with only a few labeled examples per category).

This degradation is not only due to data-driven factors but also stems from structural bottlenecks along the prompt–vision–fusion pipeline. First, prompt fragility, biomedical diagnostic prompts are semantically dense and highly structured, such that minor paraphrasing or template changes can induce large variance in the language-conditioned representation, leading to unstable predictions and poor robustness to linguistic perturbations. Second, saliency-dominated visual attention, biomedical images often contain small lesions with ambiguous boundaries, and VLMs tend to over-focus on a few dominant regions, yielding shortcut-like attention patterns and insufficient contextual coverage (e.g., surrounding tissue structures), which is particularly harmful when transferring across domains or classes. Third, selectivity failure in cross-modal fusion, existing cross-modal attention mechanisms may amplify redundant or noisy channels and fail to separate informative signals from distractors consistently, resulting in fragile alignment and reduced discriminability. These issues become more severe in base-to-novel and few-shot regimes, where distribution shift and limited supervision amplify prompt variance, attention bias, and cross-modal noise sensitivity.

Prior efforts have attempted to improve individual components of biomedical VLMs, including prompt tuning and prompt learning [Zhou *et al.*, 2022b; Zhou *et al.*, 2022a; Yao *et al.*, 2023; Peng *et al.*, 2025], saliency-guided training and attention distillation [Wu *et al.*, 2025], and refined fusion architectures [Li *et al.*, 2023]. While effective in specific settings, these approaches are largely component-wise and are often designed in isolation, lacking a systematic principle to jointly address the interdependencies among prompting, visual representation learning, and cross-modal fusion. As a result, improving one component can leave other failure modes intact (or even exacerbate them), limiting robustness

*Corresponding author.

and generalization in challenging transfer scenarios.

To overcome these limitations, TriCLIP is proposed as a decomposition-based cross-modal learning framework that restructures information flow across the entire pipeline through three synergistic stages: Prompt Distribution Regularization (PDR), Feedback-driven Counterfactual Masking (FCM), and Selective Regulation Fusion (SRF). PDR regularizes the prompt space by optimizing under class-conditioned stochastic context perturbations, promoting prompt-invariant conditioning and reducing sensitivity to linguistic variations. FCM alleviates saliency collapse by applying feedback-driven reverse-saliency masking on dominant regions (derived from model attribution with gradient isolation), encouraging complementary cue discovery and improving contextual coverage. Finally, SRF enhances cross-modal discriminability by regulating fusion that suppresses redundant interactions while reinforcing informative cross-modal correspondences. Jointly, these stages form a closed-loop design that stabilizes language conditioning, diversifies visual evidence, and improves fusion selectivity, leading to stronger generalization in base-to-novel transfer and few-shot adaptation. Our main contributions are summarized as follows:

- Proposal of Prompt Distribution Regularization, a class-conditioned training strategy that samples two independent stochastic perturbations of the learnable context tokens per class and enforces cross-view representation consistency, thereby learning prompt-invariant text embeddings and improving robustness to biomedical language variations.
- Development of Feedback-driven Counterfactual Masking, a mechanism that suppresses dominant saliency regions to alleviate attention collapse and encourage complementary evidence mining for improving contextual coverage.
- Development of Selective Regulation Fusion, a cross attention module based on regulation that separates informative alignments from redundant channels, improving the stability and discriminability of cross-modal fusion.

2 Related Work

2.1 Vision-Language Models

Large-scale vision–language models (VLMs) learn cross-modal consistency from massive image–text pairs and form the basis of unified visual understanding frameworks. Foundational models such as CLIP [Radford *et al.*, 2021] have driven this development, while subsequent pre-training paradigms, including BLIP [Li *et al.*, 2022], further strengthen alignment and generative capabilities for improved downstream transferability. In the biomedical domain, VLM research has expanded rapidly. MedCLIP [Wang *et al.*, 2022b] leverages medical image–text contrastive learning for effective zero-shot and fine-tuned performance. BiomedCLIP [Zhang *et al.*, 2023] constructs large-scale biomedical representations, and methods such as MedKLIP [Wu *et al.*, 2023] inject domain knowledge to improve alignment and grounding. BioViL-T [Bannur *et al.*, 2023] models temporal and multi-image contexts, while MedBLIP [Chen and Hong, 2024b] introduces lightweight 3D–text alignment for diagnosis and report generation. Medi-VLM [Goswami *et al.*, 2025] advances cross-attention fusion and medical report synthesis, and additional radiology and structure-focused approaches, including RadCLIP [Lu *et al.*, 2025] and IMITATE [Liu *et al.*, 2024], further enhance classification, retrieval, and generative tasks. Despite this progress, both general-purpose and biomedical-domain VLMs still face structural limitations in clinical transfer, including prompt sensitivity, visually biased saliency, and unstable cross-modal alignment during fusion.

2.2 Prompt Learning

Prompt learning has become a key mechanism for improving the adaptability and data efficiency of vision–language models. Traditional biomedical VLMs often rely on manually crafted discrete prompts, which lack flexibility and are highly sensitive to clinical semantic variability and distribution shifts. Continuous prompt learning thus emerges as a more robust alternative. CoOp [Zhou *et al.*, 2022b] learns continuous prompt vectors to replace manual design, while CoCoOp [Zhou *et al.*, 2022a] and ProGrad [Zhu *et al.*, 2023] mitigate overfitting through conditional modulation or gradient constraints. In the biomedical domain, XCoOp [Bie *et al.*, 2024] incorporates disease-specific knowledge, and DCPL [Cao *et al.*, 2024] improves clinical robustness by cross-modal domain modeling. With large language models achieving near-human performance in clinical report generation [Krishna *et al.*, 2024], recent work employs LLM-generated biomedical phrases or structured terminology as fixed semantic templates. In this setting, the LLM is frozen, and continuous prompts are learned on top of its template, forming a layered structure that enhances semantic stability and mitigates prompt overfitting across classes. Despite these advances, prompt learning still suffers from semantic collapse, where learned prompts overfit to specific template formulations and degrade under sparse data, unseen classes, or phrasing perturbations.

2.3 Cross-modal Discriminant Alignment

Attention mechanisms are fundamental to biomedical visual decoding and cross-modal fusion but exhibit persistent structural limitations. On the visual side, both convolutional and Transformer-based models rely heavily on saliency-driven attention to extract discriminative lesion features [Touvron *et al.*, 2021]. Due to the sparse and subtle nature of biomedical abnormalities, such mechanisms often suffer from saliency collapse, leading to over-focus on prominent local regions while neglecting broader structural context, thereby hindering generalization. At the cross-modal level, classical vision–language frameworks such as UNITER [Chen *et al.*, 2020] and BLIP [Li *et al.*, 2022] employ cross-attention for image–text alignment. However, their uniform weighting schemes frequently absorb noise and induce misalignment. Later studies improve selectivity through refined co-attention designs [Varol Arısoy *et al.*, 2025] or external knowledge integration [Chen and Hong, 2024a], yet many approaches still rely on implicit or static gating without explicitly separating informative signals from redundant activations. Re-

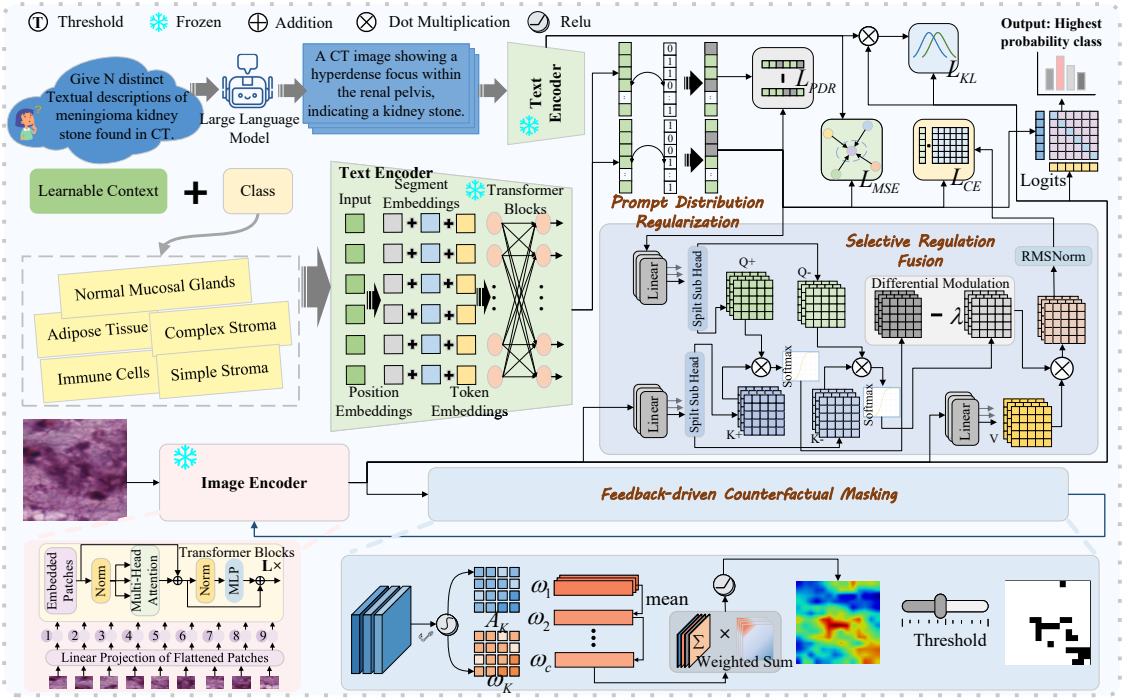

Figure 1: Overview of TriCLIP, a decomposition-driven and closed-loop calibration framework that stabilizes prompt conditioning, broadens visual evidence coverage, and enforces selective cross-modal fusion for robust biomedical vision–language classification.

cent multimodal research explores stronger structural regulation via multi-branch attention, contrastive channel separation, and dual-path alignment [Wang *et al.*, 2022a]. These methods highlight the value of constructing both reinforcing and inhibitory pathways to enhance discriminability and stability. Given the complex noise patterns in biomedical imaging, such explicit control is particularly crucial for biomedical VLMs. Motivated by these observations, the alignment mechanism introduced in this work adopts complementary attention branches that enhance informative cues while suppressing noise. Learnable modulation further enables dynamic semantic selectivity, offering structurally grounded improvements in biomedical image–text alignment.

3 Methodology

To address these challenges, TriCLIP is introduced as a decomposition-based framework that calibrates information flow across prompt conditioning, visual evidence, and cross-modal fusion, improving robustness in base-to-novel generalization and few-shot biomedical classification tasks.

3.1 Prompt Distribution Regularization

Biomedical prompts are semantically dense and highly sensitive to minor paraphrasing. Conventional prompt learning optimizes a fixed set of context tokens, which often overfits to training templates and exhibits high variance under contextual perturbations, leading to unstable language conditioning in base-to-novel transfer and few-shot regimes. To address this issue, we propose PDR, a training strategy that (i) stochastically perturbs learnable context tokens and (ii) enforces semantic consistency across perturbed views, thereby encouraging prompt-invariant text representations.

For each class i, let $\mathbf{C}_i \in \mathbb{R}^{L\times d}$ denote the learnable context token features, where L is the number of context tokens and d is the token embedding dimension. PDR applies context-token dropout with a class-wise dropout probability $\delta_i \in [0, 1]$, which is broadcast to all context tokens of that class. We set the retention probability to $p_i = 1 - \delta_i$ and choose $\delta_{\max} < 1$ to avoid degenerate all-dropout; in implementation, a small constant ϵ is used for numerical stability when needed. We sample a token mask $\mathbf{M}_i$ according to:

$$\mathbf{M}_i \sim \text{Bernoulli}(p_i\mathbf{1}), \tag{1}$$

where $\mathbf{1} \in \mathbb{R}^L$ is the all-ones vector and $\mathbf{M}_i \in \{0, 1\}^L$ indicates which context tokens are retained. The perturbed context features are computed using inverted-dropout normalization:

$$\tilde{\mathbf{C}}_i = \frac{\mathbf{C}_i \odot \mathbf{M}_i}{p_i}, \tag{2}$$

where $\odot$ denotes element-wise multiplication (broadcast along the feature dimension). This preserves the expected feature magnitude under different dropout rates. The final prompt sequence is constructed by concatenating the fixed prefix/suffix tokens with the perturbed context:

$$\mathbf{T}_i = [\mathbf{P}_{\text{prefix}},\ \tilde{\mathbf{C}}_i,\ \mathbf{P}_{\text{suffix}}]. \tag{3}$$

PDR adopts a confidence-driven schedule to determine the class-wise dropout probability δ_i. We maintain an Exponential Moving Average (EMA) confidence score $\alpha_i \in [0, 1]$ for

each class i, updated from the model's predicted probability on samples of that class:

$$\alpha_i \leftarrow m\,\alpha_i + (1-m)\,\hat{\alpha}_i, \tag{4}$$

$$\hat{\alpha}_i = \mathbb{E}_{(x,y=i)}\Big[\mathrm{StopGrad}\big(p(y \mid x)\big)\Big], \tag{5}$$

where $p(y \mid x) = \mathrm{Softmax}(s(x))_y$ denotes the predicted probability of the ground-truth label y for input x, and $s(x)$ is the corresponding logit vector. Here m is the EMA momentum, and $\mathrm{StopGrad}(\cdot)$ prevents gradients from flowing through the confidence estimator.

Intuitively, classes that the model currently predicts with lower confidence are assigned stronger prompt perturbations. We map the confidence to a bounded dropout probability via

$$\delta_i = \delta_{\min} + (1-\alpha_i)^{\gamma}\,(\delta_{\max} - \delta_{\min}), \tag{6}$$

where $\delta_{\min}$ and $\delta_{\max}$ control the dropout range and γ controls the nonlinearity. This design increases perturbation for low-confidence classes while preserving semantic cores for high-confidence classes.

To encourage prompt-invariant representations, for each class i, we generate two independently perturbed prompt views by sampling two masks, producing $\mathbf{f}_i^{(a)}$ and $\mathbf{f}_i^{(b)}$. We impose a consistency loss across the two views:

$$\mathcal{L}_{\mathrm{PDR}} = \frac{1}{n_{\mathrm{cls}}} \sum_{i=1}^{n_{\mathrm{cls}}} \left\| \mathbf{f}_i^{(a)} - \mathbf{f}_i^{(b)} \right\|_2^2, \tag{7}$$

where n_{cls} is the number of classes.

3.2 Feedback-driven Counterfactual Masking

While PDR strengthens robustness on the language side, biomedical VLMs also suffer from saliency-dominated attention, where the visual branch over-relies on a few highly discriminative regions and neglects contextual cues. To alleviate this attention bias, we design a Feedback-driven Counterfactual Masking Approach that applies reverse-saliency masking to suppress dominant areas and encourage the mining of complementary evidence.

To generate class-specific saliency maps, we adapt Grad-CAM by mapping the patch embeddings to spatial features. Specifically, the output sequence from the final normalization layer is reshaped into a 2D feature map $\mathbf{A} \in \mathbb{R}^{D\times H\times W}$. The saliency map for class c is then defined as:

$$\mathbf{Hot}^c = \mathrm{ReLU}\left(\sum_{k=1}^{D} w_k^c \mathbf{A}_k\right), \tag{8}$$

where $\mathbf{A}_k \in \mathbb{R}^{H'\times W'}$ is the k-th feature map corresponding to the embedding dimension k, and the weight w_k^c is obtained via global average pooling of the gradients:

$$w_k^c = \frac{1}{H'W'} \sum_{h=1}^{H'} \sum_{w=1}^{W'} \frac{\partial y^c}{\partial \mathbf{A}_k(h,w)}, \tag{9}$$

where y^c denotes the score for class c. To stabilize training, the saliency computation is treated as a feedback signal, and gradients are not propagated through the saliency generation path. We normalize $\mathbf{Hot}^c$ into $[0,1]$ and generate a suppression mask that zeros out the most salient regions:

$$\mathbf{M}_\tau(h,w) = \begin{cases} 0, & \text{if } \mathbf{Hot}^c(h,w) \geq \tau \cdot \max(\mathbf{Hot}^c), \\ 1, & \text{otherwise,} \end{cases} \tag{10}$$

Where τ is a constant, with a default value of 0.7. The masked image is as follows:

$$\mathbf{I}' = \mathbf{I} \odot \mathbf{M}_\tau, \tag{11}$$

By training on $\mathbf{I}'$, the model is forced to rely less on dominant evidence and to capture secondary cues, thereby mitigating saliency collapse and improving generalization under distribution shift.

3.3 Selective Regulation Fusion

When biomedical classes are fine-grained, naive fusion can amplify redundant or pseudo-correlated cross-modal signals, leading to fragile alignment under noise and domain shift. To mitigate this issue, we propose Selective Regulation Fusion(SRF), a differential regulation module that explicitly pairs reinforcing and suppressing attention flows and combines them via a learnable coefficient.

Let the image and text representations used for fusion be $\mathbf{X} \in \mathbb{R}^{B\times N\times D}$ and $\mathbf{Y} \in \mathbb{R}^{B\times N\times D}$, where B is the batch size, N denotes the number of fused embeddings, and D is the embedding dimension. We compute linear projections $\mathbf{Q} = \mathcal{Q}(\mathbf{Y})$, $\mathbf{K} = \mathcal{K}(\mathbf{X})$, and $\mathbf{V} = \mathcal{V}(\mathbf{X})$, and adopt an even number of heads H with per-head dimension $d_h = D/H$. We define $H_e = H/2$ head groups, where each group contains two sub-heads: a reinforcing sub-head and a suppressing sub-head. Accordingly, we reshape $\mathbf{Q}, \mathbf{K}$ into $2H_e$ sub-heads and reshape $\mathbf{V}$ into H_e groups with channel width $2d_h$:

$$\mathbf{Q}, \mathbf{K} \rightarrow \mathbb{R}^{B\times(2H_e)\times N\times d_h}, \mathbf{V} \rightarrow \mathbb{R}^{B\times H_e\times N\times(2d_h)}, \tag{12}$$

We compute attention probabilities in the standard way:

$$\mathbf{A} = \mathrm{Softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\Big) \in \mathbb{R}^{B\times(2H_e)\times N\times N}, \tag{13}$$

and reorganize them into paired sub-heads:

$$\mathbf{A} \rightarrow \mathbb{R}^{B\times H_e\times 2\times N\times N}, \tag{14}$$

where index 0 corresponds to the reinforcing sub-head and index 1 to the suppressing sub-head. SRF forms a regulated weighting map by subtracting the suppressing flow:

$$\mathbf{A}_{\mathrm{SRF}} = \mathbf{A}_{(:,:,0,:,:)} - \lambda\,\mathbf{A}_{(:,:,1,:,:)}, \tag{15}$$

where λ is a learnable global regulation coefficient defined as

$$\lambda = \exp(\langle \boldsymbol{\lambda}_{q1}, \boldsymbol{\lambda}_{k1} \rangle) - \exp(\langle \boldsymbol{\lambda}_{q2}, \boldsymbol{\lambda}_{k2} \rangle) + \lambda_{\mathrm{init}}, \tag{16}$$

with learnable vectors $\boldsymbol{\lambda}_{q1}, \boldsymbol{\lambda}_{k1}, \boldsymbol{\lambda}_{q2}, \boldsymbol{\lambda}_{k2} \in \mathbb{R}^{d_h}$ and initialization bias λ_{init}. The fused representation is obtained by applying the differential weighting to the values:

$$\mathbf{U} = \mathbf{A}_{\mathrm{SRF}}\,\mathbf{V} \in \mathbb{R}^{B\times H_e\times N\times(2d_h)}. \tag{17}$$

To stabilize training under differential weighting, we apply RMS normalization over the channel dimension, followed by a fixed rescaling:

$$\widetilde{\mathbf{U}} = \mathrm{RMSNorm}(\mathbf{U}) \cdot (1-\lambda_{\mathrm{init}}). \tag{18}$$

SRF improves fusion stability by explicitly regulating and suppressing signals via λ, thereby reducing redundant cross-modal interference while preserving informative alignment cues in base-to-novel and few-shot settings.

Method & Metric		BTMRI	CHMNIST	COVID	CTKidney	DermaMNIST	KneeXray	Kvasir	LC25000	OCTMNIST	RETINA	**Average**
CoCoOp	Base	77.88	87.77	77.27	81.96	42.88	34.08	85.94	88.33	79.60	66.88	72.26
	Novel	94.84	42.51	87.63	56.56	60.64	63.02	53.95	95.02	50.47	65.56	67.02
	HM	85.53	57.28	82.12	66.93	50.24	44.24	66.29	91.55	61.77	66.21	69.54
KgCoOp	Base	78.04	75.49	75.42	81.67	36.41	37.94	81.56	88.13	68.13	60.77	68.36
	Novel	95.05	38.70	89.61	58.45	47.33	61.19	59.00	86.43	50.00	54.91	64.07
	HM	85.71	51.17	81.90	68.14	41.16	46.84	68.47	87.27	57.67	57.69	66.14
Coop	Base	82.25	89.41	75.92	82.26	48.06	38.25	86.22	90.12	75.00	70.98	73.85
	Novel	94.51	35.11	90.06	67.92	59.41	47.69	58.06	87.57	50.00	56.90	64.72
	HM	87.95	50.42	82.39	74.41	53.14	42.45	69.39	88.83	60.00	63.16	68.98
ProGrad	Base	81.17	83.07	75.43	82.09	35.11	41.34	82.67	90.36	72.53	69.19	71.30
	Novel	94.29	44.32	90.83	51.78	61.92	60.10	60.39	87.89	50.00	58.48	66.00
	HM	87.24	57.80	82.42	63.50	44.81	48.99	69.80	89.11	59.19	63.39	68.55
BiomedDPT	Base	82.79	90.56	76.78	84.07	63.92	**51.01**	86.61	94.19	79.60	**73.04**	78.26
	Novel	94.00	48.27	91.44	76.36	41.31	**81.63**	57.94	92.69	52.87	57.64	69.42
	HM	88.04	62.97	83.47	80.03	50.19	**62.79**	69.43	93.43	63.54	64.43	73.58
BiomedCoOp	Base	81.04	88.17	75.94	85.75	51.86	43.32	86.78	93.29	74.07	70.35	75.08
	Novel	95.43	47.00	91.62	72.07	66.35	66.67	**61.56**	97.01	50.13	61.95	70.98
	HM	87.65	61.32	83.05	78.32	58.22	52.52	**72.03**	95.11	59.79	65.88	72.97
TriCLIP(Ours)	Base	**84.96**	**92.38**	**77.67**	**88.64**	**64.48**	48.60	**87.44**	**95.30**	**80.00**	72.83	**79.23**
	Novel	**95.78**	**49.38**	**91.85**	**82.50**	**82.30**	74.82	57.61	**97.69**	**54.07**	**75.54**	**76.15**
	HM	**90.05**	**64.36**	**84.17**	**85.46**	**72.31**	58.92	69.46	**96.48**	**64.53**	**74.16**	**77.66**

Table 1: The performance comparison with competitive methods on the Base-to-novel generalization.

3.4 Loss Function

We jointly optimize text-side robustness, vision-side representation, and cross-modal fusion with a weighted multi-term objective. The classification loss is

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{n_{\text{cls}}} y_c \log p_c, \tag{19}$$

where y_c and p_c are the one-hot label and predicted probability, n_{cls} is the number of classes classified. To maintain semantic alignment with fixed class prototypes, we use a cosine consistency term:

$$\mathcal{L}_{\text{MSE}} = 1 - \frac{\langle \mathbf{f}_v, \mathbf{E}^c_{\text{fixed}} \rangle}{\|\mathbf{f}_v\|_2 \|\mathbf{E}^c_{\text{fixed}}\|_2}, \tag{20}$$

where $\mathbf{f}_v$ is the visual feature and $\mathbf{E}^{y_i}_{\text{fixed}}$ is the fixed semantic vector of the ground-truth class. To preserve the zero-shot semantic distribution for better generalization, we distill from a frozen teacher:

$$\mathcal{L}_{KL} = \sum_{i=1}^{n_{\text{cls}}} p_{zs}(i) \cdot \log \left(\frac{p_{zs}(i)}{p_{model}(i)} \right) \tag{21}$$

where p_{zs} is the prediction probability of zero-shot, and p_{model} is the prediction probability after the network is trained normally. The overall loss is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{PDR}}. \tag{22}$$

where $\mathcal{L}_{\text{PM}}$ is the Prompt Distribution Regularization consistency loss.

4 Experiments and Results

4.1 Datasets and Setting

To comprehensively evaluate the proposed method, we assess its effectiveness under multiple evaluation protocols, primarily including base-to-novel generalization and few-shot learning. Experiments are conducted on eleven diverse biomedical imaging datasets, including BTMRI [Nickparvar, 2021], CHMNIST [Kather *et al.*, 2016], COVID [Tahir *et al.*, 2021], CTKidney [Islam *et al.*, 2022], DermaMNIST [Codella *et al.*, 2019; Tschandl *et al.*, 2018], KneeXray [Chen, 2018], Kvasir [Pogorelov *et al.*, 2017], LC25000 [Borkowski *et al.*, 2019], OCTMNIST [Kermany *et al.*, 2018], RETINA [Köhler *et al.*, 2013; Porwal *et al.*, 2018], and BUSI [Al-Dhabyani *et al.*, 2020]. These datasets cover ten different organs and nine imaging modalities. Detailed information about each dataset is provided in the Appendix.

Base-to-Novel Generalization: The image classes of the dataset are divided into base and novel classes. The model is trained only on the base classes but tested on both base and novel classes.

Few-Shot Learning: To assess the model's performance under limited supervision, we conduct few-shot learning experiments with different numbers of labeled samples per class (k = 1, 2, 4, 8, and 16).

4.2 Implementation Details

For all experiments except Few-Shot Learning, we adopt a 16-shot setting, where each class contains only 16 training samples. We use the ViT-B/16 from the BiomedCLIP model as the backbone for all experimental configurations. To ensure reliability and fairness of the results, we report the final accuracy averaged over three different random seeds. For Base-to-Novel Generalization, the number of training epochs is set to 100, and for Few-Shot Learning, it is set to 50. The batch size for all experiments is 4, and the learning rate is 0.0025. We use the SGD optimizer for training. All experiments are conducted on RTX 4090.

4.3 Performance Comparisons

Base-to-Novel Generalization. As shown in Table 1, we conduct experiments on ten public biomedical imaging datasets. Training is performed only on base classes, and performance is measured using three metrics: Base accuracy on seen classes, Novel accuracy on unseen classes, and HM, the harmonic mean of Base and Novel, reflecting the balance

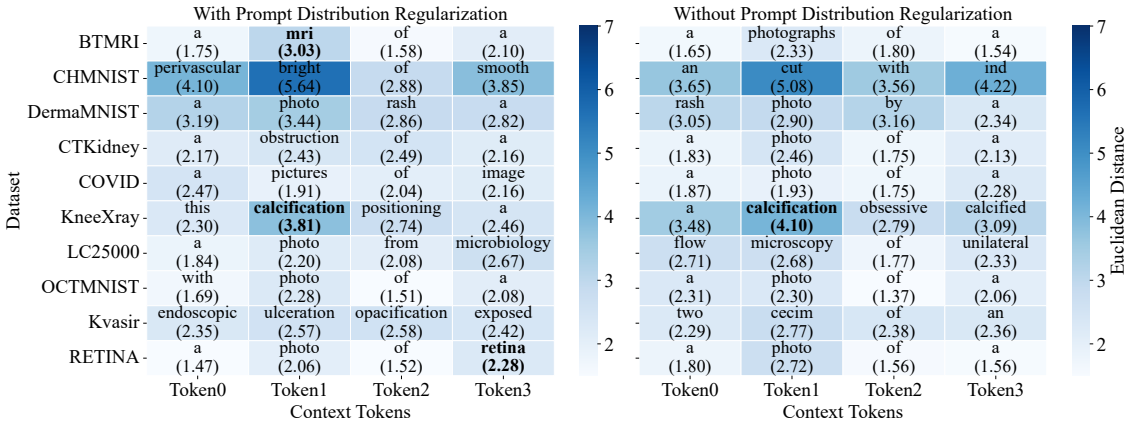

Figure 2: Visualization of the four nearest words for learned context tokens, with Euclidean distances in the embedding space.

between retention and generalization. Across most datasets, our method achieves the highest HM scores, demonstrating broad applicability and stability. It particularly improves the recognition ability of novel classes. On average, TriCLIP achieves 79.23% Base accuracy, 76.15% Novel accuracy, and 77.66% HM, outperforming all baselines, demonstrating superior generalization and robustness.

Datasets	k=1	k=2	k=4	k=8	k=16
Standard LP	50.98	53.94	60.77	67.60	68.59
LP++	52.72	53.82	57.24	64.88	68.70
CoCoOp	48.53	51.28	54.69	61.09	65.10
KgCoOp	51.89	53.34	58.71	63.53	64.90
Coop	50.18	54.17	59.77	65.85	69.72
ProGrad	51.58	54.11	60.40	65.29	67.05
BiomedDPT	56.92	60.15	64.18	68.59	71.01
BiomedCoOp	55.04	57.97	62.60	67.73	70.83
TriCLIP(ours)	**61.09**	**63.11**	**66.38**	**71.18**	**73.93**

Table 2: The performance comparison with other competitive methods on few-shot Learning.

Few-Shot Learning. To assess the efficiency of our model under limited supervision, we conduct Few-Shot learning experiments with k = 1, 2, 4, 8, 16 labeled samples per class. Several related approaches are compared, and the average accuracies on eleven datasets are reported in Table 2. More detailed data is provided in the appendix. Across different values of k, our method TriCLIP achieves the highest accuracy consistently. Specifically, it reaches 61.09% under one-shot setting, outperforming the second-best method by 4.17%.

4.4 Ablation Studies

Ablation of the component of the model. We conduct systematic ablation studies to analyze the contribution of each component, with results shown in Table 3. Each module provides consistent performance gains: PDR enhances the model's robustness to contextual perturbations, FCM improves discriminative focus, and SRF further strengthens cross-modal alignment. When combined, these strategies achieve the best overall performance, demonstrating their strong complementarity.

Variants			Few-Shot					B-to-N
SRF	FCM	PDR	k=1	k=2	k=4	k=8	k=16	HM
			55.04	57.97	62.60	67.73	70.83	72.97
		✓	57.35	59.93	64.05	68.93	72.41	74.34
	✓		55.50	58.54	63.24	68.39	71.84	74.28
✓			56.91	59.66	64.06	68.98	72.31	73.75
✓	✓		58.33	62.41	65.61	70.28	73.21	75.88
✓		✓	58.79	62.10	65.81	70.09	73.21	76.89
	✓	✓	58.81	62.06	65.54	70.23	73.16	76.84
✓	✓	✓	**61.09**	**63.11**	**66.38**	**71.18**	**73.93**	**77.66**

Table 3: Component ablation results of TriCLIP. B-to-N is the abbreviation of Base-to-Novel.

Impact of the Prompt Distribution Regularization Module. The heatmap in Figure 2 illustrates how PDR enhances prompt regularization. With the four learnable tokens, PDR drives the convergence of cue word embeddings towards domain-specific biomedical concepts (e.g., MRI, retina, calcification), which significantly brings them closer to real medical terms in the vector space (i.e., improves semantic similarity, as shown by the bolded entries in the figure. Thus, generic ambiguity of the original tokens or ineffective semantic drift is avoided. Token-level inspection shows that PDR induces more focused and clinically aligned semantics. The lighter cells indicate closer alignment to target concepts. Overall, PDR improves semantic precision, increases diversity, and yields more robust prompts across datasets by mitigating overreliance on a single handcrafted template.

Impact of Selective Regulation Fusion Module. Figure 3 presents a performance comparison among the full SRF mechanism, the reinforcing heads, and the suppressing heads. Trained only on base classes, the full mechanism consistently outperforms the two ablated settings. Reinforcing heads yield only marginal gains on base classes, while suppressing heads alone degrade base performance obviously, indicating that decorrelation without reinforcement weakens core semantics. In contrast, SRF, jointly modeling reinforcing and suppressing heads, improves novel class performance and overall generalization substantially, achieving the HM of 77.66%. Over-

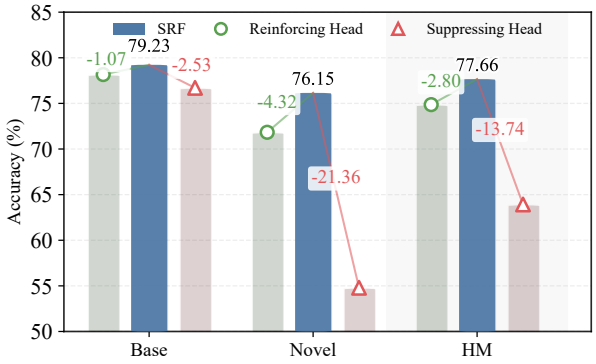

Figure 3: Study on Internal Ablation of SRF.

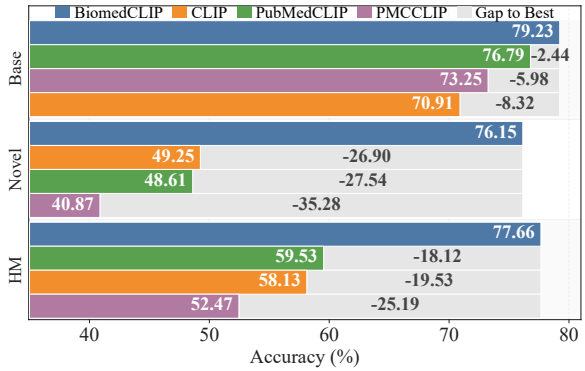

Figure 5: Different Backbone Ablation Studies.

all, the results demonstrate that the full mechanism preserves strong base-class performance while effectively enhancing transfer to novel classes.

Research on text ablation with different lengths. Figure 4 shows the impact of different initial text prompt lengths (2, 4, 8, 16, 32, and 64) on the classification performance of the model. The model is trained only on the base class and evaluated on both the base and novel classes. The results indicate that when the text prompt length is set to 4, the best performance is achieved in all indicators. In contrast, overly short prompts have slightly insufficient generalization ability on the novel class, As the cue length increases further, especially from 32 to 64, the model performance tends to decrease significantly, especially in the novel metrics. The performance performs worst when the length is 64. These results indicate that moderate length of text prompts can provide effective semantic information while avoiding redundant interference, which is more conducive to the model's generalization towards unseen classes.

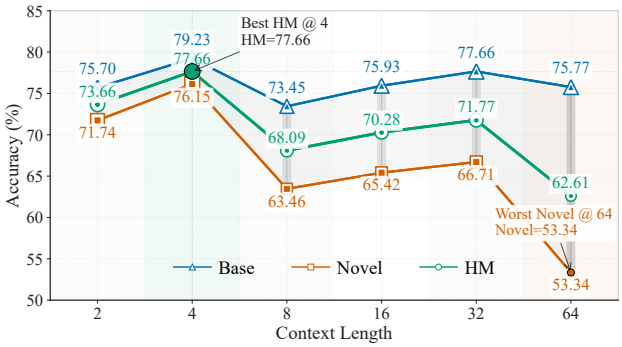

Figure 4: Research on text ablation with different lengths.

Different Backbone Ablation Studies. This study compares four representative pretrained VLM backbones, including CLIP pretrained on natural images, PMC-CLIP [Lin *et al.*, 2023] pretrained on the PMC-OA dataset, PubMedCLIP [Eslami *et al.*, 2021] adapted to the biomedical domain via domain-adaptive fine-tuning on PubMed, and BiomedCLIP. The detailed results are presented in Figure 5. BiomedCLIP consistently achieves the best performance, and therefore, it is selected as the default backbone in our experiments.

Mask visualization of Feedback-driven Counterfactual Masking Module. As shown in Figure 6, the visualization results of the FCM sequentially present the saliency heatmap (Heatmap), the suppression mask (Mask), and the masked image (Last). The Heatmap illustrates the model's response intensity to different regions of the input image during forward inference, revealing the locations of salient features that the model attends to. The Mask is generated based on the Heatmap, and it is used to suppress high-response regions during subsequent training, thereby reducing the model's over-reliance on locally salient areas.

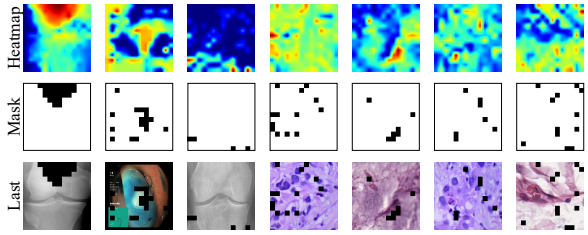

Figure 6: Mask visualization of FCM.

5 Conclusion

This work introduces a unified framework to improve the robustness of biomedical vision-language models by jointly addressing prompt regularization, saliency-aware visual rebalancing, and selective cross-modal fusion. PDR enhances robustness to linguistic variation by encouraging prompt-invariant text representations, while the FCM alleviates saliency-dominated learning through complementary visual evidence mining. Further, SRF stabilizes cross-modal fusion by suppressing redundant alignments and preserving informative cross-modal correspondences. As a whole, this approach provides a practical attempt and verification for improving cross-distribution generalization in biomedical multimodal systems, particularly under low-resource and heterogeneous clinical conditions. Despite these encouraging results, the current study primarily targets classification benchmarks, and broader validation on more tasks, institutions, and imaging modalities is needed to fully assess real-world clinical robustness.

Acknowledgments

This work is supported by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123703) and Beijing Natural Science Foundation (L253010).

References

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

[Bannur *et al.*, 2023] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023.

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

[Borkowski *et al.*, 2019] 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.

[Cao *et al.*, 2024] Qinglong Cao, Zhengqin Xu, Yuntian Chen, Chao Ma, and Xiaokang Yang. Domain-controlled prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 936–944, 2024.

[Chen and Hong, 2024a] Qiuhui Chen and Yi Hong. Alifuse: Aligning and fusing multimodal medical data for computer-aided diagnosis. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3082–3087. IEEE, 2024.

[Chen and Hong, 2024b] Qiuhui Chen and Yi Hong. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. In *Proceedings of the Asian conference on computer vision*, pages 2404–2420, 2024.

[Chen *et al.*, 2020] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

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

[Codella *et al.*, 2019] 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.

[Eslami *et al.*, 2021] Sedigheh Eslami, Gerard De Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021.

[Goswami *et al.*, 2025] Debanjan Goswami, Ronast Subedi, and Shayok Chakraborty. Medivlm: A vision language model for radiology report generation from medical images. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10287–10304, 2025.

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

[Islam *et al.*, 2022] 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.

[Kather *et al.*, 2016] 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.

[Kermany *et al.*, 2018] 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.

[Köhler *et al.*, 2013] 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 26th IEEE international symposium on computer-based medical systems*, pages 95–100. IEEE, 2013.

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

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-

training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[Lin *et al.*, 2023] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.

[Liu *et al.*, 2024] Che Liu, Sibo Cheng, Miaojing Shi, Anand Shah, Wenjia Bai, and Rossella Arcucci. Imitate: Clinical prior guided hierarchical vision-language pre-training. *IEEE Transactions on Medical Imaging*, 2024.

[Lu *et al.*, 2025] Zhixiu Lu, Hailong Li, Nehal A Parikh, Jonathan R Dillman, and Lili He. Radclip: Enhancing radiologic image analysis through contrastive language–image pretraining. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

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

[Peng *et al.*, 2025] Wei Peng, Kang Liu, Jianchen Hu, and Meng Zhang. Biomed-dpt: Dual modality prompt tuning for biomedical vision-language models. *arXiv preprint arXiv:2505.05189*, 2025.

[Pogorelov *et al.*, 2017] 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 multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169, 2017.

[Porwal *et al.*, 2018] 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.

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

[Tahir *et al.*, 2021] 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.

[Touvron *et al.*, 2021] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[Tschandl *et al.*, 2018] 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.

[Varol Arısoy *et al.*, 2025] Merve Varol Arısoy, Ayhan Arısoy, and İlhan Uysal. A vision attention driven language framework for medical report generation. *Scientific Reports*, 15(1):10704, 2025.

[Wang *et al.*, 2022a] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in neural information processing systems*, 35:33536–33549, 2022.

[Wang *et al.*, 2022b] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876, 2022.

[Wu *et al.*, 2023] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21372–21383, 2023.

[Wu *et al.*, 2025] Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102:103547, 2025.

[Yao *et al.*, 2023] 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.

[Zhang *et al.*, 2023] 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.

[Zhou *et al.*, 2022a] 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.

[Zhou *et al.*, 2022b] 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.

[Zhu *et al.*, 2023] 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.