NEARL: Interacted Query Adaptation with Orthogonal Regularization for Medical Vision-Language Understanding

Zelin Peng[†1], Yichen Zhao[†1], Yu Huang[1], Piao Yang[2], Feilong Tang[3], Zhengqin Xu[4], Xiaokang Yang[1], and Wei Shen[✉1]

[1] MoE Key Lab of Artificial Intelligence, AI Institute, School of Computer Science, Shanghai Jiao Tong University
[2] Department of Radiology, The First Affiliated Hospital, School of Medicine, Zhejiang University
[3] Mohamed bin Zayed University of Artificial Intelligence
[4] State Key Laboratory of Infrared Physics, Shanghai Institute of Technical Physics, Chinese Academy of Science

Abstract. Computer-aided medical image analysis is crucial for disease diagnosis and treatment planning. While vision-language models (VLMs) such as CLIP exhibit strong generalization ability, their direct application to medical imaging remains hindered by a substantial domain gap. Existing methods for bridging this gap, including prompt learning and unidirectional modality interaction, typically introduce domain knowledge into only one modality. However, such approaches fail to fully exploit CLIP's inherent dual-modality structure and overlook the synergistic effect of bidirectional cross-modal interaction, resulting in persistent modality misalignment. In this paper, we propose **NEARL** (iNteracted quEry Adaptation with oRthogonaL Regularization), a novel parameter-efficient VLM framework for bidirectional cross-modal interaction. NEARL consists of two key components: (1) the Unified Synergy Embedding Transformer (USEformer), which dynamically generates compact cross-modal queries to facilitate interaction; and (2) the Orthogonal Cross-Attention Adapter (OCA), which decouples new knowledge into truly novel and incremental components through orthogonal regularization. This design reduces interference from incremental components, enabling more focused learning of novel information and improving modality interaction in VLMs. Notably, NEARL introduces only **1.46M** learnable parameters. Extensive experiments on three medical imaging modalities demonstrate state-of-the-art performance (e.g., a **2.3%** relative improvement on the pneumonia dataset), along with fast inference and low memory overhead, highlighting its effectiveness for real-world medical vision-language understanding.

Keywords: Medical Vision-Language Adaptation · Bidirectional Modality Interaction · Orthogonal Feature Decoupling

[†] Equal contribution.
[✉] Corresponding author: wei.shen@sjtu.edu.cn

1 Introduction

Medical imaging analysis (e.g., disease classification) is a cornerstone of the biomedical field [2,14,3,19,4,15,22], enabling critical clinical applications such as disease diagnosis and treatment planning. Given the scarcity of large-scale annotated datasets for supervised learning in medical imaging, researchers are increasingly turning to Vision-Language Models (VLMs) pre-trained on vast amounts of image-text data, e.g., CLIP [12]. However, their direct application to medical image analysis is impeded by a domain gap [2]. This gap exists because these models are often pre-trained on natural images and thus fail to account for the unique properties of medical images, e.g., their distinct anatomical structures, varied imaging modalities, and specific disease manifestations.

To solve this issue, existing approaches can be broadly categorized into two main strategies: (*i*) **Prompt learning**. These techniques [21,2,6] focus on refining the prompt descriptions for the textual modality. The objective is to design prompts that are semantically aligned with the medical domain, thereby creating a better counterpart for the visual features. Nevertheless, the limited flexibility of the prompt space in a single modality often results in only modest performance gains. (*ii*) **Unidirectional modality interaction**. These methods [20,9,17] facilitate cross-modal interactions in a unidirectional style. This unidirectional flow fails to fully exploit the intrinsic duality of CLIP, allowing any error from a dominant modality to cascade and accumulate in the subordinate modality without any mechanism for correction. This cascading error effect ultimately hinders modality alignment and fails to unlock the full potential of VLMs.

In this paper, we introduce **NEARL** (i<u>N</u>teracted qu<u>E</u>ry <u>A</u>daptation with o<u>R</u>thogona<u>L</u> Regularization), a novel framework to unlock the full potential of VLMs for medical vision-language understanding. NEARL is built upon two core modules: USEformer and OCA. (*i*) **Unified Synergy Embedding Transformer (USEformer**). USEformer enables bidirectional cross-modal enrichment: a set of learnable queries extracts and summarizes relevant information from one modality to produce compact and complementary representations for the other, enriching text with visual context and visuals with textual meaning, thereby achieving deep mutual alignment. (*ii*) **Orthogonal Cross-Attention Adapter (OCA**). Following the cross-modal interaction in USEformer, OCA is employed to decompose the resulting features. Using Gram-Schmidt orthogonalization [13], OCA decouples this new knowledge into two orthogonal components: (1) truly novel, domain-specific information, and (2) incremental updates relative to the frozen pre-trained weights. This decomposition is critical as it prevents feature interference with the model's pre-trained generalization capabilities, thereby enabling a more focused acquisition of new knowledge.

We extensively evaluate NEARL on three medical imaging datasets spanning three imaging modalities and demonstrate state-of-the-art performance against existing methods. Our experiments show significant improvements in commonly used metrics, such as classification accuracy (up to **2.3%** relative ACC gain on the pneumonia dataset [8]), with only **1.46M** additional learnable parameters.

2 Methodology

Figure 1 illustrates the overall architecture of our proposed NEARL. We first introduce the overall pipeline (Sec. 2.1) and then describe its two core components, USEformer (Sec. 2.2) and OCA (Sec. 2.3).

2.1 Overall Pipeline

NEARL builds upon CLIP [12] and retains its pre-trained image and text encoders. Given an input image $\mathbf{I}$ and disease classes $\mathcal{C} = \{1, \ldots, C\}$, we construct textual prompts using the template "An [modality] of [CLASS]", yielding $\mathbf{T} = \{T_1, \ldots, T_C\}$, where [modality] specifies the imaging modality and [CLASS] denotes the disease category. The image encoder $\mathcal{I}$ extracts global and local visual features, while the text encoder $\mathcal{T}$ generates class-specific textual features:

$$\mathbf{V} = \mathcal{I}(\mathbf{I}; \mathbf{W}^v, \boldsymbol{\theta}^v), \tag{1}$$

$$\mathbf{S} = \mathcal{T}(\mathbf{T}; \mathbf{W}^t, \boldsymbol{\theta}^t), \tag{2}$$

where $\mathbf{W}^v$ and $\mathbf{W}^t$ denote the frozen pre-trained weights of the image and text encoders, and $\boldsymbol{\theta}^v$ and $\boldsymbol{\theta}^t$ denote the trainable parameters introduced by USEformer and OCA. The visual output is denoted as $\mathbf{V} = [\mathbf{v}^g; \mathbf{V}^l]$, where $\mathbf{v}^g \in \mathbb{R}^{D^v}$ is the global visual feature and $\mathbf{V}^l \in \mathbb{R}^{N^v \times D^v}$ contains the local visual features. $\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_C\}$ denotes the textual features of all class prompts.

Training objectives. Following CLIP, NEARL uses separate projectors for the image and text branches to obtain the projected image feature $\mathbf{v}$ and class-level text features $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_C\}$ from the global image feature $\mathbf{v}^g$ and the end-of-sequence (EOS) token feature of each class prompt. The projectors map the two modalities into a shared embedding space, where image-text similarities are computed using dot products between normalized features. Since each image is associated with a single ground-truth disease category, NEARL is optimized using the cross-entropy loss:

$$\mathcal{L}_{\text{ce}} = -\log \frac{\exp(\mathbf{v}^\top \mathbf{s}_y / \tau)}{\sum_{i=1}^{C} \exp(\mathbf{v}^\top \mathbf{s}_i / \tau)}, \tag{3}$$

where $\mathbf{s}_y$ denotes the text feature corresponding to the ground-truth disease category, and τ is a temperature parameter.

2.2 Unified Synergy Embedding Transformer

USEformer is a lightweight module designed to facilitate bidirectional cross-modal interaction between the visual and textual branches. As shown in Figure 1, USEformer receives intermediate image and text features from the frozen encoders, together with learnable image and text query tokens. It consists of

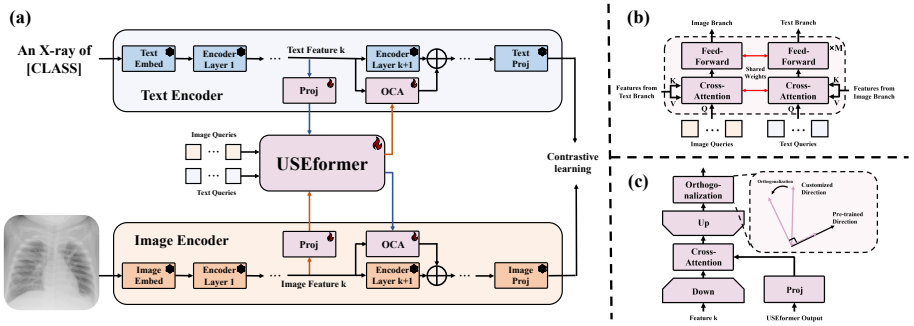

Fig. 1. Architecture of NEARL. NEARL consists of two core components: USEformer, which enables bidirectional interaction between visual and textual features for cross-modal knowledge fusion; and OCA, which injects the cross-modal residual into the frozen pre-trained encoders through an orthogonal projection, thereby mitigating feature interference during adaptation.

M stacked layers, where each layer uses cross-attention and a feed-forward network (FFN) to generate cross-modal representations conditioned on the other modality.

Given the intermediate image features $\mathbf{f}_k^v \in \mathbb{R}^{N^v \times D^v}$ and text features $\mathbf{f}_k^t \in \mathbb{R}^{N^t \times D^t}$ from the k-th encoder layer, we first project them into a shared low-dimensional space using separate linear projections:

$$\mathbf{h}_k^v = P_k^v(\mathbf{f}_k^v), \qquad \mathbf{h}_k^t = P_k^t(\mathbf{f}_k^t), \tag{4}$$

where $\mathbf{h}_k^v \in \mathbb{R}^{N^v \times D^q}$ and $\mathbf{h}_k^t \in \mathbb{R}^{N^t \times D^q}$. The learnable image queries $\mathbf{q}^v \in \mathbb{R}^{N^q \times D^q}$ and text queries $\mathbf{q}^t \in \mathbb{R}^{N^q \times D^q}$ then interact with the projected features through cross-attention. Specifically, visual information is injected into the textual query tokens by:

$$\mathrm{Attn}_{V \to T}(\mathbf{q}^t, \mathbf{h}_k^v) = \mathrm{softmax}\left(\frac{(\mathbf{q}^t W_Q)(\mathbf{h}_k^v W_K)^\top}{\sqrt{D^q}}\right)(\mathbf{h}_k^v W_V), \tag{5}$$

while textual information is injected into the visual query tokens by:

$$\mathrm{Attn}_{T \to V}(\mathbf{q}^v, \mathbf{h}_k^t) = \mathrm{softmax}\left(\frac{(\mathbf{q}^v W_Q)(\mathbf{h}_k^t W_K)^\top}{\sqrt{D^q}}\right)(\mathbf{h}_k^t W_V). \tag{6}$$

Here, W_Q, W_K, and W_V are learnable projection matrices. For parameter efficiency, they are shared across the two cross-attention branches. The outputs of the cross-attention blocks are further processed by FFNs to obtain the image-conditioned textual representation $\mathbf{z}_k^t$ and the text-conditioned visual representation $\mathbf{z}_k^v$. By stacking this block M times, USEformer iteratively refines the cross-modal representations before they are injected into the corresponding encoder layers through OCA.

2.3 Orthogonal Cross-Attention Adapter

To inject the cross-modal representations produced by USEformer into the frozen CLIP encoders, we introduce OCA, a lightweight adapter module placed at the $(k+1)$-th layer of each modality branch. OCA takes two inputs: the intermediate feature from the preceding encoder layer and the cross-modal representation generated by USEformer. These two inputs are projected into a common feature space and fused via cross-attention, allowing the adapter to incorporate cross-modal context while keeping the original pre-trained encoder weights frozen.

For the image or text branch, this operation is formulated as:

$$\Delta\mathbf{f}_{k+1}^{v/t} = W_{u,k+1}^{v/t}\,\mathrm{Attn}\big(W_{d,k+1}^{v/t}\mathbf{f}_{k}^{v/t}, W_{p,k+1}^{v/t}\mathbf{z}_{k}^{v/t}\big), \tag{7}$$

where $W_{d,k+1}^{v/t}$, $W_{p,k+1}^{v/t}$, and $W_{u,k+1}^{v/t}$ are learnable projection layers. In this formulation, the first argument of $\mathrm{Attn}(\cdot,\cdot)$ serves as the query, while the second argument provides the key and value features. The resulting $\Delta\mathbf{f}_{k+1}^{v/t}$ is an adaptive residual that carries cross-modal information for the corresponding modality branch.

Orthogonal Projection. For simplicity, we omit the modality superscript v/t in the following formulation. Directly adding the adaptive residual to the pre-trained feature may interfere with the original representation learned by CLIP. To mitigate this issue, OCA introduces an orthogonal projection operation derived from the Gram-Schmidt process [13]. Specifically, instead of directly adding the adaptive residual to the pre-trained feature, we project the residual onto the orthogonal complement of the corresponding pre-trained feature. This projection is applied token-wise:

$$\Delta\mathbf{f}_{k+1,n}^{\perp} = \Delta\mathbf{f}_{k+1,n} - \frac{\langle\Delta\mathbf{f}_{k+1,n}, \mathbf{f}_{k+1,n}\rangle}{\langle\mathbf{f}_{k+1,n}, \mathbf{f}_{k+1,n}\rangle + \epsilon}\mathbf{f}_{k+1,n}, \tag{8}$$

where n denotes the token index, $\langle\cdot,\cdot\rangle$ denotes the vector inner product, and ϵ is a small constant for numerical stability. This orthogonal projection encourages the adaptive residual to encode complementary information while preserving the pre-trained representation as the main feature backbone.

The final output feature of the $(k+1)$-th layer is obtained by adding the orthogonal adaptive residual to the frozen pre-trained feature:

$$\tilde{\mathbf{f}}_{k+1}^{v/t} = \mathbf{f}_{k+1}^{v/t} + \Delta\mathbf{f}_{k+1}^{v/t,\perp}. \tag{9}$$

The updated feature $\tilde{\mathbf{f}}_{k+1}^{v/t}$ is then passed to the subsequent encoder layers, enabling NEARL to perform cross-modal adaptation while maintaining the stability of the frozen pre-trained encoders.

3 Experiment

In this section, we first describe the experimental setup and implementation details (Sec. 3.1). We then present a comparative analysis of NEARL against state-of-the-art methods (Sec. 3.2) and evaluate its computational efficiency (Sec. 3.3).

Table 1. Performance comparison of state-of-the-art methods in terms of ACC (%) and F1 (%) on the three medical datasets. We re-implemented all compared methods using their official code. **Bold** denotes the best result, and underline indicates the second-best.

Method	Backbone	Pneumonia		Alzheimer		Retina	
		ACC	F1	ACC	F1	ACC	F1
Unimodal Methods							
ViT [5]	Transformer	$87.20_{0.6}$	$85.50_{0.7}$	$87.00_{0.3}$	$86.90_{0.3}$	$94.50_{0.2}$	$93.20_{0.2}$
MedMamba [18]	Mamba	$89.50_{1.3}$	$88.10_{1.6}$	$\underline{90.70_{1.7}}$	$\underline{90.70_{1.7}}$	$97.50_{0.6}$	$96.90_{0.7}$
Prompt Learning Methods							
CoOp [21]	CLIP	$83.40_{2.8}$	$81.30_{3.7}$	$85.80_{0.0}$	$85.80_{0.1}$	$94.50_{0.1}$	$93.10_{0.3}$
ViP [6]	CLIP	$84.60_{2.2}$	$82.60_{3.1}$	$85.60_{0.4}$	$85.60_{0.4}$	$95.00_{0.2}$	$93.90_{0.3}$
XCoOp [2]	CLIP	$88.10_{1.5}$	$86.60_{1.8}$	$90.40_{0.4}$	$90.40_{0.4}$	$\underline{97.70_{0.1}}$	$\underline{97.30_{0.0}}$
Unidirectional Modality Interaction Methods							
CoCoOp [20]	CLIP	$82.60_{0.9}$	$80.00_{1.1}$	$85.60_{0.2}$	$85.50_{0.2}$	$95.00_{0.4}$	$93.90_{0.5}$
MaPLe [9]	CLIP	$\underline{92.60_{3.2}}$	$\underline{91.70_{3.8}}$	$\underline{90.70_{1.2}}$	$90.60_{0.2}$	$97.60_{0.4}$	$97.10_{0.5}$
FATE [17]	CLIP	$82.60_{0.7}$	$80.40_{0.9}$	$84.10_{0.8}$	$84.00_{0.8}$	$92.30_{0.5}$	$90.90_{0.6}$
TextRefiner [16]	CLIP	$84.00_{0.3}$	$81.80_{0.5}$	$85.10_{0.5}$	$85.00_{0.5}$	$94.10_{0.3}$	$92.70_{0.4}$
Bidirectional Modality Interaction Methods							
NEARL ours	**CLIP**	$\mathbf{94.70_{0.2}}$	$\mathbf{94.20_{0.3}}$	$\mathbf{92.60_{0.6}}$	$\mathbf{92.60_{0.6}}$	$\mathbf{98.50_{0.2}}$	$\mathbf{98.20_{0.2}}$

We also conduct ablation studies (Sec. 3.4) and perform visual analyses of the learned feature spaces to explain why fine-tuning only the language branch is insufficient in medical domains (Sec. 3.5).

3.1 Experimental Settings

Dataset. We comprehensively evaluate the proposed method on three public medical datasets: Pneumonia [8], Alzheimer [11], and Retina [8]. These datasets vary in two key aspects: (*i*) lesion sites, covering pneumonia, Alzheimer's disease, and diabetic macular edema; and (*ii*) imaging modalities, including X-ray, MRI, and OCT. **Pneumonia** contains 5,856 chest radiograph images categorized as normal or pneumonia. We split the data according to the division protocol of [6], resulting in a 4,710/522/624 split for training, validation, and testing. **Alzheimer** comprises 80,000 brain MR images across four stages: non-demented, very mild dementia, mild dementia, and moderate dementia. To balance positive and negative samples, we select mild and moderate dementia cases together with a random subset of non-demented images, resulting in a 7,033/1,758/2,199 split for training, validation, and testing. **Retina** contains 76,607 OCT images that cover four conditions: choroidal neovascularization (CNV), diabetic macular edema (DME), drusen, and normal eyes. We select DME and normal samples for evaluation. The final dataset is partitioned into 11,887 training, 1,857 validation, and 1,859 test images.

Implementation Details. We adopt ViT-B/16 as the visual backbone and set the text embedding dimension to 512, following the original CLIP setting.

Table 2. Ablation study on key modules of NEARL.

Methods	USEformer	OR	ACC	F1
LoRA [7]	-	-	93.3	92.6
NEARL	w/o using	✓	93.7	93.0
	✓	w/o using	93.8	93.1
	✓	✓	**94.7**	**94.2**

We use natural-domain CLIP mainly for fair comparison with prior medical CLIP adaptation methods [2,6]. All results are averaged over three random seeds. The same USEformer is shared across encoder layers and is repeatedly used to produce cross-modal representations from layer-specific image and text features. OCA is inserted into each layer of both the image and text encoders to inject the corresponding cross-modal residuals. The model is trained on a single NVIDIA RTX 3090 GPU for 50 epochs. We set the number of cross-attention blocks in USEformer M to 6, the number of query tokens N^q to 32, the query token dimension D^q to 128, the adapter rank to 8, and the temperature parameter τ to 1×10^{-2}. All methods are evaluated under a supervised parameter-efficient adaptation setting.

Evaluation Metrics. We use Accuracy (ACC) and Macro F1-score (F1) to evaluate classification performance on the test set of the three benchmarks.

3.2 Comparisons with the State-of-the-Arts

We compare NEARL with state-of-the-art CLIP-based and non-CLIP-based models on three medical datasets covering X-ray, MRI, and OCT modalities, as shown in Table 1. The compared CLIP-based methods include prompt-learning approaches, i.e., ViP [6], CoOp [21], and XCoOp [2], as well as methods based on modality interaction, i.e., CoCoOp [20], MaPLe [9], FATE [17], and TextRefiner [16]. We also include two non-CLIP-based unimodal models, ViT [5] and MedMamba [18]. Among these methods, MedMamba, ViP, and XCoOp are specifically designed for medical image classification.

Overall, NEARL achieves the best performance across all three datasets, with accuracies of 94.7%, 92.6%, and 98.5%, and F1-scores of 94.2%, 92.6%, and 98.2%, respectively. These results show that NEARL generalizes effectively across different medical imaging modalities and classification tasks.

3.3 Evaluation of Real-Time Clinical Application

On a single NVIDIA RTX 3090 GPU with a batch size of 100, NEARL achieves 500 FPS at 1170 GFLOPs per batch, using only 1.46M trainable parameters while achieving the best accuracy, compared with existing PEFT baselines that use 8.19K to 26.11M parameters and run at 470 to 540 FPS. These results suggest that NEARL has low computational overhead and efficient inference performance, supporting its potential use in rapid clinical screening scenarios.

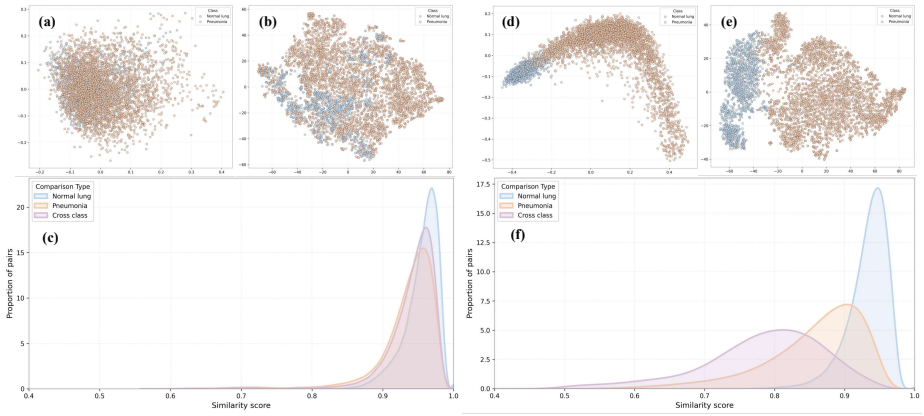

Fig. 2. Visualization of CLIP (a—c) vs. NEARL (d—f) image features on the Pneumonia dataset. PCA (a,d), t-SNE (b,e), and cosine similarity scores (c,f) show a large domain gap between pre-trained CLIP and medical images, leading to suboptimal generalization when only fine-tuning the text encoder. NEARL bridges this gap via bidirectional modality interaction.

3.4 Ablation Study

We conduct ablation studies on the Pneumonia dataset. The results are reported in Table 2. Removing USEformer leads to a clear performance drop, indicating that the shared cross-modal transformer is important for effective interaction between the image and text encoders. Compared with a simple linear projection, USEformer can better aggregate modality-specific features through learned cross-attention, thereby providing more informative cross-modal representations for subsequent adaptation. Removing orthogonal regularization (OR) also degrades performance, suggesting that explicitly separating the adaptive residual from the frozen pre-trained representation helps reduce feature interference during adaptation. When both USEformer and OR are removed, the framework degenerates into a LoRA-style adaptation baseline [7], which achieves the lowest performance. This result indicates that both cross-modal interaction and orthogonal residual adaptation are necessary for the effectiveness of NEARL.

3.5 Visual Analysis of Learned Feature Spaces

Existing prompt-learning methods [6] typically adapt the language branch while keeping CLIP's image encoder frozen, which may be insufficient for medical image classification due to the domain gap between natural and medical images. To examine this issue, we visualize the learned feature spaces using PCA [1] and t-SNE [10]. As shown in Fig. 2(a)–(c), pneumonia and normal chest X-ray features extracted by the frozen CLIP image encoder are highly overlapped, with weak intra-/inter-class cosine divergence, indicating limited class separability. This partly explains the limited performance of methods that keep the image

encoder frozen, such as FATE [17] and TextRefiner [16]. In contrast, NEARL jointly adapts the image and text branches through bidirectional cross-modal interaction, producing more separable feature distributions and clearer similarity patterns in Fig. 2(d)–(f). These results suggest that USEformer and OCA help learn medical-domain-aware representations while preserving the stability of the pre-trained CLIP backbone.

4 Conclusion

In this work, we propose **NEARL**, a parameter-efficient bidirectional framework for enhancing CLIP's cross-modal alignment in medical vision-language adaptation. By keeping CLIP's original parameters frozen and optimizing only **1.46M** additional trainable parameters, NEARL achieves substantial performance improvements with minimal computational overhead. These results suggest that NEARL provides an efficient and scalable solution for adapting vision-language foundation models to diverse medical imaging scenarios.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant Nos. 62322604, 62576207, and 62401361.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdi, H., Williams, L.J.: Principal component analysis. Wiley interdisciplinary reviews: computational statistics **2**(4), 433–459 (2010)
2. Bie, Y., Luo, L., Chen, Z., Chen, H.: Xcoop: Explainable prompt learning for computer-aided diagnosis via concept-guided context optimization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 773–783. Springer (2024)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
4. Chen, Z., Du, Y., Hu, J., Liu, Y., Li, G., Wan, X., Chang, T.H.: Multi-modal masked autoencoders for medical vision-and-language pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 679–689. Springer (2022)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Fang, X., Lin, Y., Zhang, D., Cheng, K.T., Chen, H.: Aligning medical images with general knowledge from large language models (2024)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)

8. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5), 1122–1131.e9 (2018)
9. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19113–19122 (2023)
10. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
11. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of Cognitive Neuroscience **19**(9), 1498–1507 (09 2007)
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
13. Schmidt, E.: Über die auflösung linearer gleichungen mit unendlich vielen unbekannten. Rendiconti del Circolo Matematico di Palermo (1884-1940) **25**(1), 53–77 (1908)
14. Wang, Z., Sun, Q., Zhang, B., Su, W., Wang, P., Zhang, J., Zhang, Q.: Pm 2: A new prompting multi-modal model paradigm for few-shot medical image classification. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3799–3804. IEEE (2024)
15. Wu, L., Zhuang, J., Chen, H.: Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22873–22882 (2024)
16. Xie, J., Zhang, Y., Peng, J., Huang, Z., Cao, L.: Textrefiner: Internal visual feature as efficient refiner for vision-language models prompt tuning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8718–8726 (2025)
17. Xu, Z., Peng, Z., Yang, X., Shen, W.: Fate: Feature-adapted parameter tuning for vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9014–9022 (2025)
18. Yue, Y., Li, Z.: Medmamba: Vision mamba for medical image classification. arXiv preprint arXiv:2403.03849 (2024)
19. Zhao, Y., Peng, Z., Yang, P., Yang, X., Shen, W.: Thinking like a radiologist: A dataset for anatomy-guided interleaved vision language reasoning in chest x-ray interpretation. arXiv preprint arXiv:2602.12843 (2026)
20. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
21. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision (IJCV) (2022)
22. Zhuang, J., Wu, L., Wang, Q., Fei, P., Vardhanabhuti, V., Luo, L., Chen, H.: Mim: Mask in mask self-supervised pre-training for 3d medical image analysis. IEEE Transactions on Medical Imaging (2025)