Can LLMs' Tuning Methods Work in Medical Multimodal Domain?

Jiawei Chen[1,2*] Yue Jiang[1,2*] Dingkang Yang[1,2] Mingcheng Li[1,2]
Jinjie Wei[1,2] Ziyun Qian[1,2] Lihua Zhang[1,2,3,4✉]

[1] Academy for Engineering and Technology, Fudan University
[2] Cognition and Intelligent Technology Laboratory (CIT Lab), Institute of Meta-Medical, Fudan University
[3] Engineering Research Center of AI and Robotics, Ministry of Education, China
[4] Jilin Provincial Key Laboratory of Intelligence Science and Engineering, Changchun, China
{chenjiawei22,jiangyue23}@m.fudan.edu.cn
lihuazhang@fudan.edu.cn

Abstract. While Large Language Models (LLMs) excel in world knowledge understanding, adapting them to specific subfields requires precise adjustments. Due to the model's vast scale, traditional global fine-tuning methods for large models can be computationally expensive and impact generalization. To address this challenge, a range of innovative Parameters-Efficient Fine-Tuning (PEFT) methods have emerged and achieved remarkable success in both LLMs and Large Vision-Language Models (LVLMs). In the medical domain, fine-tuning a medical Vision-Language Pretrained (VLP) model is essential for adapting it to specific tasks. Can the fine-tuning methods for large models be transferred to the medical field to enhance transfer learning efficiency? In this paper, we delve into the fine-tuning methods of LLMs and conduct extensive experiments to investigate the impact of fine-tuning methods for large models on the existing multimodal model in the medical domain from the training data level and the model structure level. We show the different impacts of fine-tuning methods for large models on medical VLMs and develop the most efficient ways to fine-tune medical VLP models. We hope this research can guide medical domain researchers in optimizing VLMs' training costs, fostering the broader application of VLMs in healthcare fields. The code and dataset have been released at `https://github.com/TIMMY-CHAN/MILE`.

Keywords: Vision-language model · Parameters-efficient fine-tuning · Visual question answering.

1 Introduction

The rise of ChatGPT has ignited significant interest in Large Language Models (LLMs). However, LLMs often require fine-tuning to adapt to specific domains

[*] Co-first author. [✉] Corresponding author.

such as medicine due to their general-purpose nature. Global fine-tuning methods are computationally expensive and may compromise model generalization capabilities. Therefore, numerous studies have begun exploring Parameter-Efficient Fine-Tuning (PEFT) techniques [1, 9, 15, 21] aimed at enabling more efficient fine-tuning and achieved remarkable success.

Some works [2,18] have attempted to apply PEFT methods to Large Vision-Language Models (LVLMs). Compared to Natural Language Processing (NLP) tasks, visual language tasks introduce visual inputs, leading to more diverse content and requiring more challenging fine-tuning. These endeavours have demonstrated that PEFT methods successful in LLMs can enhance the few-shot and zero-shot capabilities of LVLMs or achieve comparable results to global fine-tuning approaches.

However, despite advancements [12,25,26] in reducing computational requirements for fine-tuning LVLMs, for many researchers, particularly those in interdisciplinary fields outside of Computer Science (CS), such as biomedicine, accessing the computational resources required for large-scale model fine-tuning remains a significant challenge. The lack of access to server-level GPUs, crucial for effective model fine-tuning, poses a considerable challenge. Consequently, there is a pressing need for small-scale VLMs (which are called basic or fundamental VLMs), particularly given the privacy concerns surrounding medical images. The legality and ethical concerns surrounding the upload of private medical images to publicly available LVLMs further emphasize the necessity of investigating whether PEFT methods, successful in LLMs and LVLMs, can achieve comparable results when applied to basic VLMs.

To empower researchers in the medical domain with limited computational resources to effectively fine-tune multi-modal models for practical applications, we embark on experimental research to investigate the applicability of the LLMs' tuning methods in the realm of medical multimodal (vision-language) learning. In this paper, we design a **M**odularized med**I**cal vision-**L**anguage fine-tuning mod**E**l (**MILE**) that builds upon a medical Vision-Language Pretrained (VLP) model and incorporates various PEFT modules through modular design. Specifically, from the model structure perspective, we conduct a systematic investigation of the PEFT methods in the LLMs and develop the corresponding modules which can be integrated into a generative vision-language baseline model. From the training data perspective, we propose an instruction-format medical multi-modal dataset for applying instruction-tuning on different MILE variants. We conduct in-depth ablation studies on those LLMs' tuning methods and validate them on two radiographic image benchmarks. We believe that these empirical analyses will catalyze the development of fine-tuned medical multimodal models.

Our main contributions are as follows: **(i)** We systematically explored how trainable parameters in different medical VLM modules affect overall performance, revealing strategies for achieving competitive results akin to global fine-tuning. **(ii)** Through extensive experiments, we conducted a novel comparison of the PEFT methods tailored for small-scale medical VLM based on a baseline model, offering insights distinct from large-scale models. **(iii)** We conducted a thorough analysis of the impact of instruction-tuning on fine-tuning basic VLP models and released an instruction-format medical image-text dataset. Our investigation revealed both positive and negative effects of instruction-tuning, offering a nuanced understanding of its implications for the fine-tuning process of the small-scale VLP models.

2 Related Work

PEFT Techniques of LLMs: Fine-tuning large pretrained language models (PLMs) is resource-intensive, often requiring substantial computational resources and training data. To address this challenge, PEFT techniques have emerged, aiming to enhance PLMs' performance on specific tasks with minimal changes to model parameters. Various methods have been developed in this regard. Adapter Tuning [8], Dora [19], LoRA [9] etc. [17, 20] add small components to PLMs or the input embedding [1, 11] to realize PEFT. Besides methods that add small components, some PEFT techniques focus on data manipulation to minimize or eliminate changes to the original model weights. OpenAI [1] and Google [24] have independently introduced instruction-tuning methods that modify original data into instruction pairs for fine-tuning models, achieving better generative results compared to multi-task training. While these PEFT methods have been successfully applied to LLMs, their impact on small-scale VLMs remains underexplored, especially in the medical domain.
Medical Vision-Language Models: Recent advancements in the pretraining-finetuning paradigm have led to the emergence of medical VLMs [3, 4, 14] based on VLP models. However, their pretraining and fine-tuning require data scales of more than 100,000 image-text pairs and the number of trainable parameters for global fine-tuning is not much different from that of PEFT in LVLMs. Therefore, under the dual factors of large-scale training data and high training parameters, small-scale VLMs' training costs remain unaffordable for many researchers. Thus, in this work, we systematically review LLMs' tuning methods and discuss their applicability to medical VLMs.

3 Method

3.1 Architecture of MILE

Baseline model: Most VLMs architecture are based on CLIP [22] or BLIP [13]. In this paper, we use MISS [3], a generative multimodal medical VLM as our baseline model, the architecture has been shown in Figure 1a. MISS has an image encoder and a Joint Text-Multimodal (JTM) encoder, the former for image feature extraction and the latter for text feature extraction and multimodal feature interaction. A text decoder is appended after the JTM decoder for causal reasoning and text generation. The image encoder of the baseline model is a ViT-base [7] model; The JTM encoder is designed based on Bert with 12 transformer-based [23] layers where a cross-attention layer is inserted between the bi-self attention layer and the feed-forward layer; the architecture of the text decoder is similar to the JTM encoder and the bi-self attention layer is replaced by a causal attention layer.

The Unified Model: To validate the effectiveness of the above PEFT methods on small-scale medical VLMs, we construct MILE and equip it with 4 modules of commonly used PEFT methods [9, 15, 17, 20], obtaining four variants: MILE-LoRA, MILE-Prefix, MILE-IA3, and MILE-PTv2.

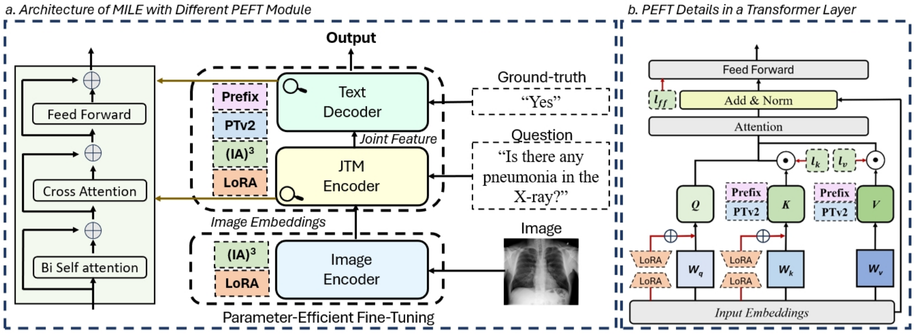


Fig. 1: Architecture of MILE with different PEFT modules (a) and PEFT details in a Transformer layer (b).

- MILE-LoRA: As shown in Figure 1b, low-rank matrixes are selectively injected into all the attention layers' parameter matrix *query* and *key* in the image encoder, JTM encoder or text decoder for LoRA-Tuning.

- MILE-$(IA)^3$: For $(IA)^3$-Tuning (IA3) [17], learnable vectors l_k, l_v and l_{ff} which rescale the *key*, *value* and the inner activation are respectively injected into all the attention layers and feed-forward layers in the image encoder, JTM encoder or text decoder of the baseline model during tuning, as shown in Figure 1b.

- MILE-Prefix: For MILE-Prefix, prefix vectors are selectively attached before the input of the JTM encoder or text decoder, while tuning the image encoder. When the prefix vectors are only attached before the input of the text decoder, the input embeddings are defined as $z = [PREFIX, x]$ [15], while the prefix vectors are attached by both the input of the JTM encoder and text decoder the input is defined as $z = [PREFIX, x, PREFIX']$.

- MILE-PTv2: As shown in Figure 1a, prompt tokens are selectively attached before the input of the JTM encoder or text decoder. Different from prefix-tuning, all prefixes in the input of each attention layer are derived from a trainable matrix when we apply P-Tuning v2 [20].

3.2 Instruction-format Data Generation

To analyze the impact of instruction-tuning on basic VLMs, we curated a medical image-text dataset using Instruction templates (details in the Appendix). 'Closed' templates suit closed-ended questions, and 'Opened' templates are for open-ended ones. During training, each QA pair randomly incorporates a template. For answer options, inspired by [3], we categorized question attributes and created diverse candidate pools. In opened instructions, incorrect answers from the same attribute are randomly included with the correct answer.

3.3 Training

We use the Slake Dataset [16] and the VQA-RAD Dataset [10] for training, testing and validating. To ensure a fair comparison of the impact of instruction-tuning and other PEFT methods on basic VLMs, the setting of the training hyper-parameter is the same with [3] and the dataset splits are identical to those used in most current works [3, 4, 14]. The training loss $\mathcal{L}$ is the language modeling loss [6]. For each MILE employing PEFT, the PEFT parameters are varied during training (if applicable) to investigate the effect of different percentage parameter changes on model performance. Each MILE is trained with both instruction-format data we make and original data provided by the dataset, getting two different models. Testing is conducted under the original benchmark and the task is generative, with no candidate answers provided to the model.

4 Experiment Results and Analysis

We initially trained a series of MILE equipped with PEFT modules using the original data. Tables 1 to 4 present the accuracy (ACC(%)) of four MILE variants training with the origin data on the Slake benchmark. 'F' denotes the freezing of all parameters within a given module, 'T' signifies that all parameters are trainable, and the acronyms 'LoRA', 'Prefix', 'IA3', and 'PTV2' identify the specific PEFT methods applied. 'Memory' represents the GPU memory (GB) required for training. And '#Params' indicates the weight of trainable parameters over all parameters.

Table 1: Results of MILE-LoRA(origin data).

ViT	JTM	Dec	Rank	#Params	Memory	Opened	Closed	Gobal
F	LoRA	LoRA	4	0.163%	5.19	3.57	50.70	20.34
		LoRA	8	0.325%	5.21	3.57	50.70	20.34
LoRA	LoRA	LoRA	4	0.327%	26.63	48.65	50.70	49.34
			8	0.652%	26.75	48.93	50.70	49.57
F	T	LoRA	4	38.022%	7.26	47.76	70.70	55.53
		LoRA	8	38.072%	7.45	50.21	70.99	57.18
T	LoRA	LoRA	4	24.009%	26.96	68.14	50.70	62.29
		LoRA	8	24.133%	27.29	68.28	50.70	62.38
T	T	LoRA	4	61.887%	27.60	78.52	79.44	78.83
			8	61.919%	28.11	78.66	80.56	79.30

The results demonstrate that a fully frozen visual encoder (ViT) within the VLM significantly hampers the model's ability to correctly interpret texts and images, as observed in the MILE-IA3 and MILE-PTV2, where global ACC plummets to 0.57% and 0%, respectively. Conversely, a modest increase in tunable parameters by 0.16% in MILE-LoRA leads to notable improvements of 45% and 29% in open-ended and global ACC, respectively. When all parameters of the

ViT are set to trainable, the global ACC of the four models increases by 42%, 21%, 17%, and 19% compared to when they are completely frozen.

Table 2: Results of MILE-Prefix.

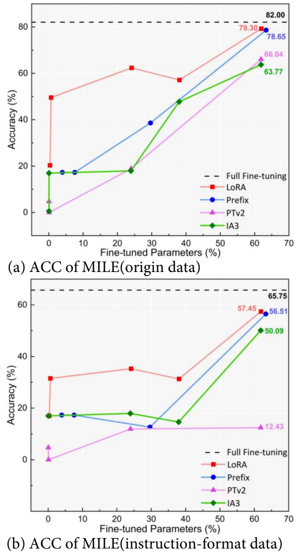
ViT	JTM	Dec	#Params	Memory	Opened	Closed	Global
F	F	Prefix	3.926%	4.62	0	50.7	17.3
F	Prefix	Prefix	7.556%	4.67	0	50.7	17.3
T	Prefix	Prefix	29.636%	26.41	41.50	32.95	38.61
T	T	Prefix	63.354%	27.97	76.82	**82.25**	78.65

Table 3: Results of MILE-IA3.

ViT	JTM	Dec	#Params	Memory	Opened	Closed	Global
F	IA3	IA3	0.051%	6.35	0	1.69	0.57
IA3	IA3	IA3	0.061%	23.01	0	50.70	16.98
T	IA3	IA3	23.924%	26.83	12.77	28.17	17.92
F	T	IA3	37.987%	7.52	46.24	50.70	47.74
T	T	IA3	61.866%	27.90	72.20	47.04	63.77

Table 4: Results of MILE-PTV2.

ViT	JTM	Dec	#Params	Memory	Opened	Closed	Global
F	PTV2	PTV2	0.102%	4.52	0	0	0
F	F	PTV2	0.051%	4.57	7.10	0	4.72
T	PTV2	PTV2	23.963%	25.41	13.62	29.30	18.87
T	T	PTV2	61.876%	27.46	74.18	49.86	66.04



(a) ACC of MILE(origin data)

(b) ACC of MILE(instruction-format data)

Fig. 2: ACC of MILE trained with different data.

On the other hand, in MILE-Prefix and MILE-IA3, when the ViT is frozen, even converting the JTM encoder from frozen to fully trainable barely improves global ACC. This underscores the pivotal role of the visual encoder in a VLP model for downstream task adaptation, where even minimal adjustments via PEFT can significantly enhance performance.

When the parameters of the ViT are updatable, increasing the proportion of parameter updates for the JTM encoder can also lead to significant improvements. Notably, in MILE-Prefix, shifting from Prefix-Tuning to full parameter updates for the JTM encoder boosts global ACC by 40%, with closed-ended question ACC surpassing that of the baseline model employing global fine-tuning. In MILE-IA3 and MILE-PTV2, elevating the update ratio for the JTM encoder markedly improves open-ended question ACC by 67% and 70%. However, this comes with the cost of increasing the training parameter ratio to 61% to 64%.

It is also worth mentioning that full parameters updating of both the visual encoder and JTM encoder, alongside PEFT application to the decoder, can reduce the parameter count by 40% while maintaining performance on par with global fine-tuning.

4.1 Performance Differences Among Different PEFT Methods

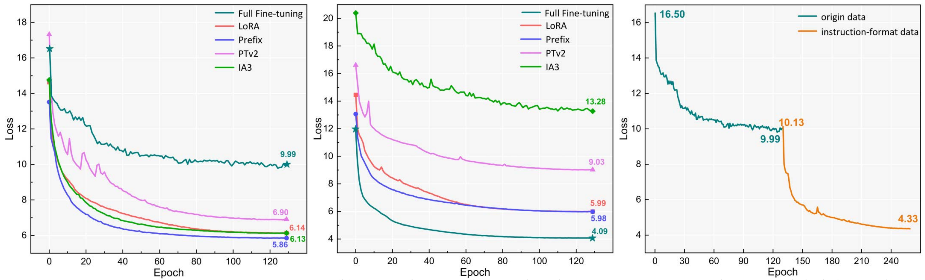
Although the aforementioned PEFT methods have been compared in their respective papers within the LM domain, our experiments reveal differing efficacies of them within the medical VLM domain.

LoRA-Tuning exhibits the most competitive performance in this domain, effective for PEFT both language modeling Transformers and visual modeling Transformers. Compared with IA3, when all model parameters are subject to PEFT, MILE-IA3 consistently answers 'no' (ACC 50.70%) to all questions. While MILE-LoRA also responds 'no' to closed-ended questions, it demonstrates a better understanding of the semantic information from images and text, achieving a 48.65% ACC on open-ended questions, outperforming MILE-Prefix, MILE-IA3, and MILE-PTV2, which have about 20%-30% trainable parameters.

Furthermore, we conducted an ablation study on the rank of the LoRA unit. As shown in Table 1, within the same tuning paradigm, a doubling of the number of parameters in the LoRA matrix brings about a minor improvement in model performance. A LoRA rank of 8 (ViT froze) and LoRA-tuned (rank=4) visual encoders, adjusting a similar parameter fraction (about 0.32%), differed by 45% in open-ended questions' accuracy.

MILE-Prefix also demonstrates promising results. When both the ViT and the JTM are fully trainable, and the decoder employs PEFT, the performance of MILE-Prefix is comparable to MILE-LoRA. When the JTM encoder also utilizes PEFT, MILE-Prefix lags, indicating that simply adding a prefix to vectors does not effectively promote the alignment of features from different modalities in the cross-attention layer, consistent with the principle of Prefix-Tuning.

Compared to the baseline model, the performance of MILE-IA3 and MILE-PTV2 is inferior to MILE-LoRA and MILE-Prefix. Within fundamental VLM, updating a negligible fraction (less than 0.1%) of the decoder's parameters substantially impairs its generative task performance. This impact is markedly less pronounced in LLMs, underscoring the critical role of the number of parameters proportionality in tuning efficacy.



(a) Loss of MILE (origin data)

(b) Loss of MILE (instruction-format data)

(c) Loss of MILE (global+instrucion-tuning)

Fig. 3: Loss of MILE variants under the training with origin data (a), instruction-format data (b) and an instruction-tuning MILE after fine-tuning by origin data (c).

4.2 The Impact of Data on Training Effectiveness

We employed instruction-format image-text pairs to conduct both global fine-tuning and PEFT of the model but did not parallel the positive effects seen in LLMs when applied to a basic VLP model. As shown in Table 5, the use of instruction-format data during full fine-tuning led to more than 20% decrease in the global ACC. Moreover, Figure 2a&b shows that the synergy of structure-level PEFT with data-level instruction-tuning significantly reduced the performance of various MILE variants by over 20%, diverging from outcomes with raw data (more precise data are presented in tables in the Appendix).

This disparity underscores the potential model and task-specific sensitivity of instruction-format data's benefits. For basic VLP models, such data application might not yield the expected advantages.

4.3 Is Instruction-Tuning Ineffective for Basic VLMs?

While LLMs and LVLMs can benefit from instruction-tuning to enhance their generalization across different types of tasks, this approach may not be effective for basic VLMs. This is because base VLMs are typically fine-tuned for specific downstream tasks. Furthermore, instruction-format text inputs provide candidate answers to questions (see in Appendix), which do not exist in real-world scenarios during inference and practical applications of generative models.

As shown in Figure 3a&b, global fine-tuning with instruction-based data showed lower losses, hinting at an improved task format comprehension but potentially oversimplifying the difficulty of training tasks, thus diminishing real-world inferencing efficacy.

However, given that instruction-format data can enhance the model's understanding of the target task and result in lower training losses, could it potentially improve a global fine-tuned model that has already converged on the original data? Figure 3c demonstrates the loss reduction when a converged model is further trained on instruction-format data, the loss of the converged model continues to decrease and finally reaches about 4.33 after global fine-tuning. As shown in Table 5, MILE ultimately achieves 86.7% open-ended ACC, 81.42% closed-ended ACC, and 83.02% overall ACC, surpassing the baseline model and demonstrating state-of-the-art performance among generative VLMs.

Table 5: Comparsion with other medical VLMs which have different tuning paradigms, '♣' means global fine-tuning with ordinary data (no instruction-format) and '♠' means instruction-tuning with all parameters updating.

Methods	Pretrain # images	Tuning paradigm	Type of task	VQA-RAD			SLALKE		
				CLOSED	OPENED	OVERALL	CLOSED	OPENED	OVERALL
MTL [5]	87,952	♣	classification	79.8	69.8	75.8	86.1	80.2	82.5
M3AE [4]	298,000	♣	classification	83.4	67.2	77	87.8	80.3	83.2
MUMC [14]	387,000	♣	ranking	84.2	71.5	79.2	-	-	84.9
MISS [3]	38,800	♣	**generating**	80.35	**71.81**	76.05	82.91	**81.47**	82
MILE	38,800	♠	**generating**	2.68	45.58	24.22	59.72	68.79	65.75
MILE	38,800	♣ + ♠	**generating**	76.34	**73.45**	74.89	**86.70**	81.42	**83.02**

5 Conclusion

In this paper, we comprehensively investigate whether fine-tuning methods for LLMs can be applied to the medical multimodal domain, aiming to ease the training burden on resource-constrained practitioners. We developed a suite of MILE models incorporating various fine-tuning strategies atop generative VLP frameworks, delving into the effects of structural and parametric modifications on performance. From a series of experiments, we observe that updating the parameters of the visual encoder is crucial for VLMs. Furthermore, updating the parameters of the JTM encoder which is responsible for text feature extraction and multimodal feature fusion can significantly enhance model performance. Through a comparison of different PEFT methods, we find that LoRA-Tuning and Prefix-Tuning exhibit the best tuning effects, achieving comparable performance to global fine-tuning models while reducing training costs by 40%.

Additionally, we explore the impact of data-level fine-tuning, specifically instruction-tuning, on model performance. Although directly fine-tuning with instruction-format data simplifies the training task, it leads to suboptimal performance for basic VLMs in practical tasks. Nonetheless, instruction-tuning on top of models already optimized on original datasets demonstrated notable performance gains. We hope that our work can inspire researchers in the medical field who aim to reduce the training costs of multimodal models and promote the application of VLMs in the medical domain.

Acknowledgments. This work is supported in part by the National Key R&D Program of China (2021ZD0113502).

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

References

1. Brown, T., Mann, B., Ryder, et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
2. Chen, J., Yang, D., Jiang, Y., Li, M., Wei, J., Hou, X., Zhang, L.: Efficiency in focus: Layernorm as a catalyst for fine-tuning medical visual language pre-trained models. arXiv preprint arXiv:2404.16385 (2024)
3. Chen, J., Yang, D., Jiang, Y., et al.: Miss: A generative pretraining and finetuning approach for med-vqa. arXiv preprint arXiv:2401.05163 (2024)
4. Chen, Z., Du, Y., Hu, et al.: Multi-modal masked autoencoders for medical vision-and-language pre-training. In: MICCAI. pp. 679–689. Springer (2022)
5. Cong, F., Xu, S., et al.: Caption-aware medical vqa via semantic focusing and progressive cross-modality comprehension. In: ACM MM. pp. 3569–3577 (2022)
6. Devlin, J., Chang, M.W., Lee, et al.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. 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)
8. Houlsby, N., Giurgiu, A., Jastrzebski, et al.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)
9. Hu, E.J., Shen, et al.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
10. Lau, J.J., Gayen, et al.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
11. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)
12. Li, C., Wong, C., Zhang, et al.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890 (2023)
13. Li, J., Li, D., Xiong, et al.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICCV. pp. 12888–12900 (2022)
14. Li, P., Liu, G., He, et al.: Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In: MICCAI. pp. 374–383. Springer (2023)
15. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021)
16. Liu, B., Zhan, L.M., Xu, et al.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 ISBI. pp. 1650–1654 (2021)
17. Liu, H., Tam, D., Muqeeth, et al.: Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. Advances in Neural Information Processing Systems **35**, 1950–1965 (2022)
18. Liu, H., Li, C., Wu, et al.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
19. Liu, S.Y., Wang, C.Y., Yin, H., Molchanov, P., Wang, Y.C.F., Cheng, K.T., Chen, M.H.: Dora: Weight-decomposed low-rank adaptation. arXiv preprint arXiv:2402.09353 (2024)
20. Liu, X., Ji, K., Fu, Y., Tam, et al.: P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602 (2021)

21. Liu, X., Zheng, Y., Du, Z., Ding, et al.: Gpt understands, too. AI Open (2023)
22. Radford, A., Kim, J.W., Hallacy, et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
23. Vaswani, A., Shazeer, N., Parmar, et al.: Attention is all you need. NIPS **30** (2017)
24. Wei, J., Bosma, M., Zhao, V.Y., Guu, et al.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021)
25. Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023)
26. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)