General or Medical CLIP, Which one Shall We Choose?

Haiting Huang[1] HAITING.HUANG@FAU.DE
Dario Zanca[1] DARIO.ZANCA@FAU.DE
Bjoern Eskofier[1,2,3] BJOERN.ESKOFIER@FAU.DE
Emmanuelle Salin[1] EMMANUELLE.SALIN@FAU.DE
[1] *Department Artificial Intelligence in Biomedical Engineering (AIBE), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany*
[2] *Chair of AI-supported Therapy Decisions, LMU München, Munich, Germany*
[3] *Institute of AI for Health, Helmholtz Zentrum München, Neuherberg, Germany*

Editors: Under Review for MIDL 2026

Abstract

Despite growing interest in multimodal deep learning for medical imaging, researchers and clinicians still lack a systematic understanding of when medical-domain vision–language models outperform their general-domain counterparts. While several medical CLIP variants have been proposed, they are typically evaluated in isolation and on narrow tasks, leaving open questions about how pre-training data, downstream task, and fine-tuning strategy jointly affect performance. We systematically compare four vision–language CLIP-based models on three representative tasks: image classification, image-to-text retrieval, and visual question answering. Across tasks, zero-shot performance is generally insufficient for clinical use, even for medically pre-trained models, confirming the need for task-specific fine-tuning. Medical-domain pre-training offers clear benefits in low-data regimes and for in-distribution modalities, but can underperform CLIP when downstream data deviates from the pre-training distribution. When sufficient labeled data is available, and especially under LoRA-based tuning, general-domain CLIP systematically matches or surpasses specialized medical models. VQA remains notably challenging, with none of the evaluated models achieving competitive results even after fine-tuning, suggesting that more advanced multimodal reasoning approaches are needed. Based on these findings, we provide recommendations for selecting and adapting vision-language models in clinical settings.

Keywords: Foundation Model, Classification, Visual-question answering, Image-to-text retrieval, medical applications

1. Introduction

Patient care generates diverse multimodal data, including medical images and clinical notes. Models capable of synthesizing information from these modalities could provide valuable support for clinicians in documentation (e.g., radiology report generation (Chen et al., 2024b)), education (Li et al., 2023), or clinical decision-making. However, privacy concerns limit the development of large-scale multimodal datasets for the development of specialized task-specific models. The recent advent of multimodal foundation models has enabled joint analysis of visual and textual data for a range of tasks and domains. Among them, contrastive language–image pretraining (CLIP) (Radford et al., 2021) has become a de facto standard backbone for vision–language representation learning because of its scalable training, zero-shot performance, and flexibility as a general-purpose multimodal encoder.

However, the unique characteristics of medical images (e.g., modality-specific appearance, subtle and localized abnormalities, protocol, and device variability) and clinical text (e.g., domain-specific terminology, abbreviations, and report-style phrasing) pose challenges for directly adapting CLIP-based models to healthcare data. This has led to the development of medical-specific foundation models, including Biomed-CLIP (Zhang et al., 2023), PMC-CLIP (Lin et al., 2023), and CXR-CLIP (You et al., 2023). These models have shown promise on specific tasks, but they have usually been evaluated in narrow contexts that do not account for the diversity of medical data and the variety of transfer learning scenarios encountered in real-world clinical settings (Marzullo and Ranzini, 2024).

In this study, we systematically evaluate vision–language models for medical applications to determine how the model domain (general vs. medical), pre-training data, and fine-tuning strategy jointly affect downstream performance and to identify the most effective configurations for different scenarios. We release the full implementation of our evaluation to support further research.[1] Our contributions are as follows:

- We quantify how the pre-training data distribution affects performance on medical tasks and demonstrate that task-specific fine-tuning is essential for achieving strong performance, regardless of pre-training domain. While medical-domain pre-training often leads to superior performance, it can be outperformed by CLIP when the downstream data diverges significantly from the pre-training distribution. With sufficient fine-tuning data, CLIP can match or surpass specialized medical models.
- We compare model performance across three downstream tasks: image classification, Visual Question Answering (VQA), and Image-Text Retrieval (ITR). Fine-tuned models consistently and significantly outperform their zero-shot counterparts, which are generally unreliable on medical applications. Our findings show that, even after fine-tuning, VQA performance remains substantially below state-of-the-art, highlighting critical limitations in the multimodal reasoning capabilities of CLIP-based models.
- Based on our analysis, we establish a list of recommendations for the application of CLIP-based models in the medical domain.

2. Related Work

Multimodal machine learning has become a prominent research direction, with the development of foundation models showing strong cross-modal understanding capabilities (Xu et al., 2024). One notable example is the CLIP architecture (Radford et al., 2021), which employs two separate encoders that process texts and images respectively. A contrastive loss function is used to align semantically similar text-image pairs within a shared embedding space. This approach has shown strong generalizability and has achieved state-of-the-art performance in diverse applications (Tankala et al., 2024; Radford et al., 2021).

CLIP has demonstrated its capability in medical image recognition across various imaging modalities, especially for x-ray data (Zhao et al., 2023). For example, CLIP achieves an accuracy of 87.4% on a multi-label classification task on the VinDr-CXR dataset (Nguyen et al., 2022; Mishra et al., 2023). However, it can also struggle to generalize effectively to

1. https://github.com/98haiting/General-or-Medical-CLIP-Which-one-Shall-We-Choose

other domain-specific applications (Chen et al., 2024a). For instance, while CLIP achieves a zero-shot accuracy of 70.1% on object recognition tasks for natural-domain images (Radford et al., 2021), its performance significantly degrades on tasks involving medical images (Zhang et al., 2023; Lin et al., 2023; You et al., 2023).

Yet, the general-domain pre-training of CLIP could limit its performance in the healthcare domain. Empirical evidence suggests that pre-training from scratch on domain-specific data outperforms transferring a pre-trained model to a specific domain using large-scale datasets (Gu et al., 2021). Thus, researchers have developed medical vision-language models for generic applications (Lin et al., 2023; Eslami et al., 2023; Zhang et al., 2023) and specific imaging modalities (You et al., 2023; Hamamci et al., 2024; Salentin et al., 2015).

However, studies comparing the performance of general and medical-domain vision-language models remain limited. PubMedClip (Eslami et al., 2023) shows an improvement of 1% compared to CLIP on medical VQA tasks under similar conditions (Eslami et al., 2023). Biomed-CLIP (Zhang et al., 2023) reports a significant performance increase compared to a fine-tuned CLIP model on several medical applications, including histopathological (Veeling et al., 2018; Borkowski et al., 2019; Saltz et al., 2018) and radiological image classification (Shih et al., 2019), and cross-modal retrieval (Zhang et al., 2023).

Recent works have also validated Biomed-CLIP for other medical applications, including scoliosis detection (Polis et al., 2025) and hematological recognition (Patel et al., 2024). PMC-CLIP also shows competitive performance across various medical downstream tasks (Lin et al., 2023), including image classification, VQA, and ITR. However, it has also shown notable limitations with respect to its generalization capability, in particular for zero-shot classification or domain identification tasks (Zhao et al., 2023).

3. Experimental Setup

In this work, we evaluate the performance of four vision-language models: CLIP (Radford et al., 2021), pre-trained on general-domain data; Biomed-CLIP (Zhang et al., 2023) and PMC-CLIP (Lin et al., 2023), both pre-trained in the universal biomedical domain; and CXR-CLIP (You et al., 2023), a model specialized for chest X-ray images. We assess their performance and generalization abilities for image classification, VQA and ITR tasks in the medical domain. Each model is first evaluated in a zero-shot setup for the image classification and ITR datasets. Then, models are fine-tuned with task-specific datasets for all downstream tasks. In this paper, we compare the transferability of three fine-tuning methods, including full, partial (Kumar et al., 2022), and LoRA (Hu et al., 2022). The experimental pipeline is detailed in Figure 1.

3.1. Datasets

We select biomedical datasets across various medical domains for medical image classification, ITR and VQA downstream tasks. Additional details are summarized in Appendix A.

Image classification: We select four subsets of MedMNIST (Yang et al., 2023). We include PneumoniaMNIST (i.e., chest X-rays), BreastMNIST (Al-Dhabyani et al., 2020) (i.e., breast ultrasound), OrganAMNIST (Bilic et al., 2019), and OrganCMNIST (Xu et al., 2019) (i.e., abdominal CT scans) in this evaluation. These datasets represent both in-domain and out-of-domain tasks for the specialized CXR-CLIP model.

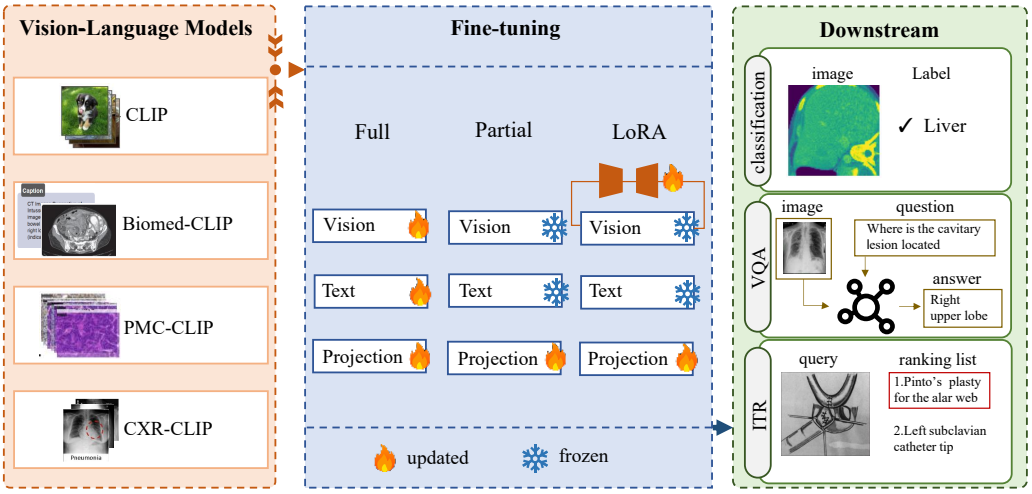


Figure 1: **Experiment pipeline**. In this work, we evaluated four selected vision-language models (CLIP (Radford et al., 2021), Biomed-CLIP (Zhang et al., 2023), PMC-CLIP (Lin et al., 2023), CXR-CLIP (Johnson et al., 2019; Irvin et al., 2019; Wang et al., 2017)) on classification (Yang et al., 2023), VQA (Lau et al., 2018), and ITR (Pelka et al., 2018) tasks with full, partial, and LoRA fine-tuning methods.

Image-to-text retrieval (ITR): ROCO (Pelka et al., 2018) comprises non-compound images extracted from articles in the PMC[2], and is automatically detected and classified as either radiology or non-radiology (Pelka et al., 2018), creating two distinct subsets.

Visual Question Answering (VQA): VQA-RAD (Lau et al., 2018) contains radiological image-text pairs from different modalities (Lau et al., 2018). SLAKE (Liu et al., 2021) is a bilingual (English/Chinese) dataset designed to support medical VQA. We utilize only the English version during training and inference.

3.2. Implementation Details

We select the ViT-B/32 and ViT-B/16 backbones for CLIP and Biomed-CLIP respectively, and ResNet for PMC-CLIP and CXR-CLIP. All models are trained and evaluated using an input resolution of 224×224 regardless of their visual backbone architecture. Following medical VQA models (Nguyen et al., 2019; Zhan et al., 2020) where a bilinear attention network is used to enhance feature fusion, we incorporated the same module into the vision-language models (Kim et al., 2018). For the classification and ITR tasks, no additional layer is added. The models are fine-tuned for 200 epochs in a mixed-precision manner, early stopped by validation loss, and optimized by Adam (Kingma and Ba, 2014). For the classification task, we adopt a learning rate of 5.0×10^{-5} for PMC-CLIP, CLIP, and CXR-CLIP, and a learning rate of 1.0×10^{-5} for Biomed-CLIP with a weight decay of 1.0×10^{-3} for encoders, and evaluate with AUC, accuracy and F1-score. VQA tasks employ a learning

2. https://pubmed.ncbi.nlm.nih.gov/

	BreastMNIST			PneumoniaMNIST			OrganAMNIST			OrganCMNIST		
	AUC	Acc.	F1	AUC	Acc.	F1	AUC	Acc.	F1	AUC	Acc.	F1
CLIP	50.0	26.9	0	50.0	62.5	76.9	49.0	6.9	5.2	48.0	5.8	4.6
Biomed-CLIP	**67.0**	54.5	56.4	**67.0**	60.1	54.3	**59.0**	**28.5**	**24.36**	**57.0**	24.0	**22.0**
PMC-CLIP	52.0	**69.9**	**81.6**	50.0	37.5	0	49.0	7.8	4.9	50.0	8.5	4.0
CXR-CLIP	42.0	57.1	71.7	54.0	**65.4**	**78.1**	53.0	21.1	8.6	53.0	**24.6**	13.9

Table 1: **Zero-shot performance on selected classification tasks of the MedMNIST dataset**. The best results for each metric are highlighted in bold.

	Radiology [%]			Non-radiology [%]			Mixed [%]		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	0.7	2.5	4.01	18.4	36.1	45.6	1.4	3.5	5.4
Biomed-CLIP	**33.3**	**66.1**	**79.2**	**74.3**	**94.6**	**97.5**	**35.0**	**67.4**	**80.1**
PMC-CLIP	0.0	0.1	0.2	0.5	1.65	2.1	0.0	0.1	0.2
CXR-CLIP	0.0	0.1	0.2	0.3	1.6	2.1	0.0	0.1	0.2

Table 2: **Zero-shot performance on image-to-text retrieval task on the ROCO dataset**. The best performance for each metric is highlighted in bold.

rate of 5.0×10^{-5} with a weight decay of 0.1, and are measured with accuracy and F1-score. The learning rate of ITR is set at 1.0×10^{-5}, and the task is assessed with recall@k (k=1, 5, 10). All experiments are implemented using PyTorch on a single NVIDIA V100 GPU within one day. We conduct all the experiments under the same single random seed to ensure reproducibility.

4. Results and Discussion

4.1. Zero-shot capability

We evaluate the zero-shot performance of both general- and medical-domain CLIPs to determine how pre-training impacts generalizability in medical applications. Table 1 presents the results for medical imaging classification tasks. CLIP, PMC-CLIP, and CXR-CLIP achieve an AUC of around 50% for all subsets, showing no zero-shot capability. While Biomed-CLIP shows classification performance above the chance level, they remain far below state-of-the-art performances. In clinical scenario, we argue that fine-tuning models, either CLIP or its medical variants, is still necessary to achieve reasonable performance.

Table 2 presents zero-shot image-to-text retrieval results. Biomed-CLIP demonstrates superior performance compared to other CLIP models, likely attributable to the high similarity between its pre-training corpus and the downstream evaluation data, both of which are derived from PubMed. However, this overlap raises concerns about potential data leakage between the pre-training corpus and the ROCO dataset. In contrast, other medical CLIP models exhibit limited zero-shot performance. CXR-CLIP's reliance on templated

		CLIP			Biomed-CLIP			PMC-CLIP			CXR-CLIP		
		Part	Full	LoRA	Part	Full	LoRA	Part	Full	LoRA	Part	Full	LoRA
Breast	AUC	78.0	50.0	81.0	80.0	84.0	84.0	78.0	**85.0**	78.0	31.0	20.0	25.0
	Acc.	78.9	26.9	84.0	83.3	87.8	87.2	81.4	**89.7**	80.1	29.5	23.7	25.6
	F1.	73.1	0.0	88.9	88.4	91.7	91.2	87.0	93.2	**95.8**	36.1	35.0	34.1
Pneu.	AUC	85.0	50.0	85.0	94.0	89.0	93.0	**95.0**	88.0	**95.0**	**95.0**	90.0	94.0
	Acc.	88.3	62.5	88.6	94.4	91.4	95.0	95.5	91.0	96.0	**96.2**	92.8	94.7
	F1.	91.2	72.0	91.6	95.6	93.5	96.2	96.4	93.3	96.9	**97.0**	94.5	95.9
OrganA	AUC	96.0	**98.0**	**98.0**	96.0	**98.0**	**98.0**	96.0	**98.0**	95.0	92.0	79.0	96.0
	Acc.	93.0	96.6	**96.8**	93.4	96.5	96.7	92.2	95.8	91.9	87.8	57.7	93.0
	F1.	93.0	96.6	**96.8**	93.3	96.5	96.6	92.0	95.8	91.9	87.2	54.3	92.9
OrganC	AUC	92.0	94.0	95.0	94.0	**97.0**	96.0	93.0	96.0	91.0	90.0	73.0	94.0
	Acc.	86.8	90.8	92.5	91.2	**94.4**	**94.4**	88.8	94.2	86.4	82.8	53.4	90.1
	F1.	86.9	80.7	92.5	91.1	**94.3**	**94.3**	88.7	94.1	86.3	82.8	50.0	90.2

Table 3: **Classification performance on selected MedMNIST datasets for full, partial, and LoRa fine-tuning methods**. "Pneu." denotes PneumoniaMNIST. Bold values and underlined scores are the best and the second best performance.

pre-training prompts with limited vocabulary may restrict its expressiveness in retrieval tasks. The performance of general-domain CLIP on the non-radiology subset, with a recall@10 score of 45.6%, suggests potential for zero-shot generalization capabilities in the medical domain beyond radiology images. These findings align with previous observations that PMC-CLIP often underperforms relative to the original CLIP in zero-shot tasks (Zhao et al., 2023).

4.2. Fine-tuning performance

We present the performance of each CLIP model after task-specific fine-tuning in Tables 3, 4, and 5, for classification, ITR, and VQA, respectively.

Classification: There is a significant increase in performance after fine-tuning across most models and fine-tuning methods, except for CXR-CLIP on ultrasound images. While no single model outperforms all others, Biomed-CLIP and PMC-CLIP consistently exhibit strong classification capabilities across all subsets. CLIP exhibits lower performance on the BreastMNIST and PneumoniaMNIST subsets, potentially attributable to their smaller dataset sizes (see Table 7 in Appendix A). However, while fine-tuned CLIP-based models achieve competitive performance on the MedMNIST benchmark, they still fall short of surpassing current state-of-the-art results (see Appendix B.1).

Image-to-Text Retrieval: While Biomed-CLIP still outperforms the other models, its results show very limited improvement over the zero-shot baseline, which supports the hypothesis that data leakage may compromise the fairness of the comparison. PMC-CLIP reaches the second-best performance on radiology data with a recall@1 of 13.9% with LoRA fine-tuning, but reaches lower performances than the general CLIP on non-radiology data, which shows strong performance on this subset. However, its recall@10 of 18.8% in the

ROCO		CLIP			Biomed-CLIP			PMC-CLIP			CXR-CLIP		
		Part	Full	LoRA	Part	Full	LoRA	Part	Full	LoRA	Part	Full	LoRA
Radiology	R@1	2.0	4.1	1.0	**33.7**	24.3	<u>33.3</u>	12.9	13.4	13.9	0.1	0.4	0.2
	R@5	6.4	12.5	4.7	**66.6**	52.1	<u>65.6</u>	32.0	33.8	33.4	0.3	0.7	0.3
	R@10	10.7	18.8	8.4	**79.5**	66.1	<u>79.2</u>	43.3	46.0	44.8	0.4	0.8	0.4
Non-radiology	R@1	19.7	21.3	5.6	**76.7**	<u>75.1</u>	<u>75.1</u>	9.7	9.8	9.5	0.2	1.2	0.2
	R@5	44.3	43.0	15.1	**95.4**	<u>94.8</u>	94.6	31.6	26.7	31.5	0.7	3.6	1.5
	R@10	55.3	55.7	24.8	**98.4**	**98.4**	<u>98.2</u>	41.3	37.1	40.7	2.1	5.3	2.0

Table 4: **Image-to-text retrieval performance fine-tuned on ROCO dataset for full, partial, and LoRa fine-tuning methods**. Bolded and underlined values are the best and the second best performance.

		CLIP			Biomed-CLIP			PMC-CLIP			CXR-CLIP		
		Part	Full	LoRA	Part	Full	LoRA	Part	Full	LoRA	Part	Full	LoRA
VQA-RAD	Acc.	50.8	52.0	52.7	40.4	43.2	42.0	48.0	<u>53.7</u>	48.0	49.6	**53.9**	46.8
	F1	46.4	52.3	51.6	36.9	36.4	38.1	41.6	<u>52.6</u>	42.7	50.5	**55.1**	43.5
SLAKE	Acc.	72.9	66.3	73.4	17.5	53.8	19.1	**76.6**	75.1	<u>75.5</u>	70.7	68.4	67.3
	F1	73.3	66.4	74.1	19.8	51.4	17.5	<u>76.7</u>	74.2	**78.1**	70.5	67.6	66.8

Table 5: **Visual-question answering performance on medical VQA datasets**. The best performance are bolded, and the second best values are underlined.

radiology subset shows limited generalizability to this setting. CXR-CLIP performs poorly across both subsets, with scores of recall metrics approaching zero. The performance of CXR-CLIP shows that the narrow diversity of specialized medical pre-training datasets may limit the generalizability and transferability of vision-language models. This is likely due to its restricted pre-training text diversity, which could explain the worse performance compared to medical image classification tasks. Indeed, CXR-CLIP is pre-trained on image-label data which are prompted with template sentences (You et al., 2023).

Visual Question Answering: Despite the good performance of Biomed-CLIP on classification and ITR, it does not transfer to VQA. PMC-CLIP surpasses other models on the Slake dataset, and CLIP also reaches overall good performance, achieving an accuracy of 74.1%. On the VQA-RAD dataset, fine-tuned CXR-CLIP reaches performance slightly above the others, with an accuracy of 53.9%. Overall, the poorer performance of models on VQA could be due to the fact that CLIP-based models process image and text separately, leading to limited modality interactions. PMC-CLIP incorporates a transformer-based fusion module, which strengthens cross-modal interaction and likely contributes to its superior performance on the VQA task. On the VQA-RAD dataset, models perform better on closed-ended than on open-ended questions, reaching an accuracy of 65.0% and 9.5% respectively for the fine-tuned CXR-CLIP. This could be due to the more limited answer choices, and the reduced need for medical reasoning and understanding of object relationships.

		OrganA					OrganC				
Size	ft.	5%	10%	25%	50%	100%	5%	10%	25%	50%	100%
CLIP	full	75.1	87.3	94.8	**96.5**	96.6	16.1	27.8	80.7	85.8	90.8
	LoRA	85.9	92.1	94.6	96.0	**96.8**	73.4	81.4	88.0	90.9	92.5
Biomed	full	**94.4**	**94.7**	**95.3**	96.3	96.5	**89.9**	**91.9**	**93.5**	92.7	94.4
	LoRA	90.6	93.4	**95.3**	96.3	96.7	74.2	84.6	91.9	**93.5**	**94.4**

Table 6: **Fine-tuning accuracy under varying proportions of OrganAMNIST and OrganCMNIST dataset**. "ft." suffix indicates the fine-tuning method used. Bolded accuracy values are the best performance for each portion.

Impact of pre-training dataset: The results show that medical-domain pre-training usually provides advantages compared to general-domain pre-training across fine-tuning methods, especially for smaller downstream datasets, which is a common challenge in real-world medical applications. This is especially true for the radiological domain, as the size of the fine-tuning dataset may be too small to adapt the visual encoder of the general domain CLIP. On the other hand, CLIP shows good performance on non-specialized medical downstream tasks, such as the non-radiology ROCO and the Slake dataset. In addition, the general-domain CLIP demonstrates strong transferability on specialized downstream tasks for larger datasets in medical image classification, such as the OrganA and OrganC subsets.

To confirm this hypothesis, we evaluate how the size of the fine-tuning dataset impacts the performance of general-domain and medical-domain CLIP. To that end, we specifically consider the OrganCMNIST and OrganAMNIST datasets. We evaluate CLIP and Biomed-CLIP on portions of these datasets, ranging from 5% to 100%, and present the results in Table 6. In both cases, the models perform increasingly better with a larger amount of data. In the case of the OrganC dataset, smaller than OrganA, Biomed-CLIP consistently outperforms the CLIP model, though the gap between the models continues to shrink as the fine-tuning dataset size increases. In the case of OrganA, while the Biomed-CLIP outperforms the general-domain CLIP for smaller proportions of data, it reaches comparable or better results than Biomed-CLIP when at least 50% of the data is used. This amounts to a fine-tuning dataset size of more than 15,000 instances. These results underscore that the general-domain CLIP requires substantially larger downstream datasets to match the performance of BiomedCLIP on medical images. This shows the limits of the visual encoder of the general-domain CLIP for fine-grained medical image analysis.

Fine-tuning methods: Fine-tuning strategies can have a significant impact on model performance. For example, Table 3 shows that full fine-tuning can be efficient in the case of medical vision-language models for classification tasks, but can lead to catastrophic forgetting for the general-domain CLIP fine-tuned on smaller datasets (e.g., BreastMNIST, PneumoniaMNIST). Parameter-efficient approaches such as LoRA offer a promising alternative by significantly reducing the number of trainable parameters while enabling the adaptation of the image encoder to other medical modalities. By applying LoRA to the vision encoder of the vision-language models, the models are able to integrate task-specific features into the pre-trained representations. In addition, these methods tend to produce

more stable outcomes across different runs. Indeed, full fine-tuning of the models is often less stable, requiring careful calibration of the hyperparameters. This is especially the case for the ITR and VQA tasks. Figure 2 in the Appendix shows how slight variation in hyperparameters impacts the results of each fine-tuning approach, with full fine-tuning consistently resulting in the highest variance. LoRA and partial fine-tuning achieve more consistent performance. We also show that standard deviation is high for VQA fine-tuning, which could be due to the necessity of training the bilinear attention network. However, no single fine-tuning method consistently outperforms the others, underscoring the need to tailor the fine-tuning approach to the task and selected model.

Qualitative evaluation of VQA: To understand the capabilities of vision-language models, we classify their errors on VQA tasks into three distinct categories: language errors, multimodal errors, and reasoning errors. Appendix B.3 provides illustrative examples.

Language errors include three sub-categories. First, models may produce answers that are inappropriate for the question type, such as responding with a 'yes/no' answer to a non-binary question. Our observations suggest that this can stem from the distribution of the training dataset, leading to an over-reliance on the question's phrasing (e.g., answering yes/no to a question starting with 'Is'). Additionally, the model's answer may exhibit the wrong granularity compared to the reference ground truth. This mismatch can stem from question ambiguity (e.g., a question refers to a singular object, but the expected answer references several). Finally, some errors are paraphrases, where the answer is semantically correct but phrased differently from the reference response.

Beyond language-related errors, we categorize cross-modal errors as predictions that are linguistically valid but visually irrelevant. These errors highlight the model's inability to connect textual and visual information effectively, even when no in-depth medical knowledge is required. An example is when an answer refers to an object that is not present in the image. Spatial reasoning errors are another example, where the model fails to locate an object or attribute correctly within the image.

Finally, we define medical reasoning errors as predictions that are linguistically and visually relevant but incorrect, due to a lack of precise medical knowledge. Figure 10 shows how CXR-CLIP correctly identifies an organ visible in the image but does not understand which one is abnormal. While such reasoning errors would provide valuable insights, they are relatively underrepresented compared to the other error types, especially in the VQA-Rad dataset. Indeed, each error type reveals distinct insights into the performance of vision-language models, which aggregated metrics might obscure. Future research should prioritize the development of targeted datasets or metrics for each of these error types. This would enable a more transparent understanding of foundation models' capabilities in clinical applications.

4.3. Recommendation on the use of CLIP models for medical applications

Based on our results, we recommend the following practical guidelines for selecting and adapting foundation models in clinical applications.

Zero-shot applicability: Both general-domain and specialized CLIP models struggle with downstream tasks. While Biomed-CLIP shows strong results, data leakage concerns warrant further evaluation of its reliability. However, general-domain CLIP's zero-shot

performance in image-to-text retrieval for non-radiology images could support its use for human-in-the-loop applications. For example, it could assist in selecting medical images for dataset creation by leveraging information such as image type, anatomical location, or the absence of identifiable information (e.g., text, faces).

Model fine-tuning: Task-specific fine-tuning on representative data is typically required. If large downstream task datasets are available (e.g., more than 10,000 instances), general-domain CLIP is suitable, as fine-tuning can inject sufficient in-domain knowledge to fine-tune the visual encoder. With limited data, either universal medical-domain models or domain-aligned medical models are preferable for fine-grained analysis of medical images. However, fine-tuning a specialized medical CLIP on off-domain medical images is not recommended.

Modality considerations: For predominantly image-based tasks on medical images (e.g., classification, detection, segmentation), universal medical foundation models typically achieve good performance with fewer resources. For tasks with substantial textual components or leveraging diverse natural language (e.g., image-to-text retrieval, reporting), general-domain models are often more effective due to the higher pre-training text variety.

Multimodal reasoning: In our experimental setting, fine-tuning CLIP for fine-grained medical reasoning proved challenging with tested models. For these tasks, architectures with explicit image–text feature fusion tend to perform better, and are suggested by the stronger results of PMC-CLIP on VQA. Incorporating question-based pre-training or cross-attention modules could improve the results. Beyond model design, there is also a need to create tasks and metrics that can accurately assess reasoning capabilities, as existing ones can blur distinctions between linguistic, multimodal, and medical reasoning errors.

Training a medical CLIP model: Some studies have explored pre-training custom CLIP models for medical applications (Yan et al., 2025). In such cases, leveraging a general-purpose language model would be advantageous, given that medical vision-language models often suffer from limited textual diversity, as evidenced by CXR-CLIP results.

5. Conclusion

This study aims to provide a thorough overview toward the choice of vision-language foundation models in the healthcare domain. Our experiments reveal several important findings regarding the behavior and adaptability of vision-language models depending on datasets and tasks. The chance-level zero-shot performance of models on most tasks demonstrate the necessity of task-specific fine-tuning even with medical pre-training. However, the pre-training dataset significantly influences downstream performance, particularly in how well models generalize to new data distributions, the efficiency of fine-tuning with limited data, and the adaptability to different task types. Medical-domain models typically show better performance on fine-grained medical imaging, especially when the fine-tuning data is limited. When the downstream image modality differs from the pre-training distribution, the models exhibit worse performance than the general-domain model. Additionally, large fine-tuning data or non-medical imaging tasks enable general-domain CLIP to match or outperform medical-domain models. Therefore, universal medical models should generally be preferred, and a general model can be considered when downstream data is sufficient. Domain-specific medical model do not transfer beyond the specific modality domain.

References

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

Patrick Bilic, Patrick Ferdinand Christ, et al. The liver tumor segmentation benchmark (lits). *CoRR*, abs/1901.04056, 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.

Haolong Chen, Hanzhi Chen, Zijian Zhao, Kaifeng Han, Guangxu Zhu, Yichen Zhao, Ying Du, Wei Xu, and Qingjiang Shi. An overview of domain-specific foundation model: key technologies, applications and challenges. *arXiv preprint arXiv:2409.04267*, 2024a.

Yaxiong Chen, Chuang Du, Chunlei Li, Jingliang Hu, Yilei Shi, Shengwu Xiong, Xiao Xiang Zhu, and Lichao Mou. Unicrossadapter: Multimodal adaptation of clip for radiology report generation. In *International Workshop on Foundation Models for General Medical AI*, pages 113–123. Springer, 2024b.

Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, 2023.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *CoRR*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems*, 31, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 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.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

Gang Liu, Jinlong He, Pengfei Li, Genrong He, Zhaolin Chen, and Shenjun Zhong. Pefomed: Parameter efficient fine-tuning of multimodal large language models for medical imaging. *arXiv preprint arXiv:2401.02797*, 2024.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023.

Aldo Marzullo and Marta Bianca Maria Ranzini. Exploring zero-shot anomaly detection with clip in medical imaging: Are we there yet? *arXiv preprint arXiv:2411.09310*, 2024.

Aakash Mishra, Rajat Mittal, Christy Jestin, Kostas Tingos, and Pranav Rajpurkar. Improving zero-shot detection of low prevalence chest pathologies using domain pre-trained language models. *arXiv preprint arXiv:2306.08000*, 2023.

Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 522–530. Springer, 2019.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

Tanviben Patel, Hoda El-Sayed, and Md Kamruzzaman Sarker. Microscopic hematological image classification with captions using few-shot learning in data-scarce environments. In *2024 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, pages 184–190. IEEE, 2024.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and C. Friedrich. Radiology objects in context (roco): A multimodal image dataset. In *CVII-STENT/LABELS@MICCAI*, 2018. URL https://api.semanticscholar.org/CorpusID:53087891.

Bartosz Polis, Agnieszka Zawadzka-Fabijan, Robert Fabijan, Róża Kosińska, Emilia Nowosławska, and Artur Fabijan. Exploring biomedclip's capabilities in medical image analysis: A focus on scoliosis detection and severity assessment. *Applied Sciences (2076-3417)*, 15(1), 2025.

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.

Sebastian Salentin, Sven Schreiber, V Joachim Haupt, Melissa F Adasme, and Michael Schroeder. Plip: fully automated protein–ligand interaction profiler. *Nucleic acids research*, 43(W1):W443–W447, 2015.

J Saltz, R Gupta, L Hou, T Kurc, P Singh, V Nguyen, D Samaras, KR Shroyer, T Zhao, R Batiste, et al. Tumor-infiltrating lymphocytes maps from tcga h&e whole slide pathology images [data set]. *Cancer Imaging Arch*, 4, 2018.

George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

Pavan Kalyan Tankala, Piyush Pasi, Sahil Dharod, Azeem Motiwala, Preethi Jyothi, Aditi Chaudhary, and Krishna Srinivasan. Wikido: A new benchmark evaluating cross-modal retrieval for vision-language models. *Advances in Neural Information Processing Systems*, 37:140812–140827, 2024.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical image computing and computer assisted intervention–mICCAI 2018: 21st international conference, granada, Spain, September 16-20, 2018, proceedings, part II 11*, pages 210–218. Springer, 2018.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Mengwei Xu, Wangsong Yin, Dongqi Cai, Rongjie Yi, Daliang Xu, Qipeng Wang, Bingyang Wu, Yihao Zhao, Chen Yang, Shihe Wang, et al. A survey of resource-efficient llm and multimodal foundation models. *arXiv preprint arXiv:2401.08092*, 2024.

X. Xu, F. Zhou, et al. Efficient multiple organ localization in ct image using 3d region proposal network. *IEEE Transactions on Medical Imaging*, 38(8):1885–1898, 2019.

Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Lie Ju, Gin Tan, et al. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, pages 1–12, 2025.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer, 2023.

Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345–2354, 2020.

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.

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

Appendix A. Dataset Details

Table 7 lists the size of each dataset used for our evaluation.

The radiological subset of ROCO covers modalities such as CT, ultrasound, X-ray, and PET (Pelka et al., 2018), whereas non-radiology includes annotated diagrams, digital illustrations, and portraits of medical professionals (Pelka et al., 2018).

For the VQA-RAD dataset, we implement an image-based split to prevent data leakage. The resulting splits contain 1473 items with 201 images for training, 354 question-answer pairs with 51 images for validation. The resulting test set includes 421 question-answer pairs (337 closed-ended and 84 open-ended) associated with 62 samples. Notably, 3.09% of answers in the testset are not presented in the training or validation sets, labeled as 'unknown'.

The Slake dataset includes 140 head CTs or MRIs, 41 nect CTs, 219 chest X-rays or CTs, 201 abdomen CTs or MRIs and 41 pelvic cavity CTs (Liu et al., 2021). The ratio of "unknown" answers in the test set is 0.18%.

	Dataset	Training	Validation	Test
Image Classification	BreastMNIST	546	78	156
	PneumoniaMNIST	4708	524	624
	OrganAMNIST	34561	6491	17778
	OrganCMNIST	12975	2392	8216
Image-to-Text Retrieval	ROCO-Rad.	65414	8171	8176
	ROCO-Non.	4888	610	610
Visual Question Answering	VQA-RAD	1473	354	421
	SLAKE	4919	1053	1061

Table 7: **Size of fine-tuning datasets**. The "ROCO-Rad." indicates the radiology subset of the ROCO, and "ROCO-Non." presents the non-radiology subset.

Appendix B. Additional Results

B.1. Comparison to State-of-the-Art

We compare the selected vision-language models to models of the MedMNIST benchmark [3] under the same fine-tuning datasets as in Table 8. The best vision-language models reach similar performances to the visual models after fine-tuning on these standard medical datasets.

In addition, we referenced these models on VQA tasks using the two best-performing VQA methods in Table 9 at the time of writing, according to the VQA-RAD and SLAKE benchmarks[4][5]. It is important to note that we manually split the VQA-RAD by image.

3. https://medmnist.com/
4. https://paperswithcode.com/sota/medical-visual-question-answering-on-vqa-rad
5. https://paperswithcode.com/sota/medical-visual-question-answering-on-vqa

	Breast		Pneu.		OrganA		OrganC	
Models	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc
CLIP	81.0	84.0	85.0	88.6	98.0	**96.8**	95.0	92.5
Biomed	84.0	87.8	94.0	94.4	98.0	96.7	97.0	**94.4**
PMC	85.0	**89.7**	95.0	96.0	98.0	95.8	96.0	94.2
CXR	31.0	29.5	95.0	**96.2**	96.0	93.0	94.0	90.1
Google	**91.9**	86.1	**99.1**	94.6	99.0	88.6	98.8	87.7
ResNet	90.1	86.3	95.6	86.4	**99.8**	95.1	**99.4**	92.0

Table 8: **Comparison to benchmarking performance of MedMNIST2D on classification task**. 'Pneu.' stands for Pneumonia, Google for Google AutoML Vision, and ResNet for ResNet-18. Bolded values are the best performance for each subset.

	Selected Models				Benchmarking	
Dataset	CLIP	Biomed	PMC	CXR	Model	Acc.
RAD	52.7	47.7	55.1	69.9	PeFoMed (Liu et al., 2024)	**81.9**
SLAKE	75.7	69.3	78.1	71.8	B-GPT (Luo et al., 2023)	**86.1**

Table 9: **Reference to benchmarking performance VQA task**. "RAD" refers to dataset VQA-RAD, and "B-GPT" indicates BiomedGPT. Scores are presented in accuracy. Best results are shown in bold.

B.2. Impact of fine-tuning method on Result Stability

Full fine-tuning of the models is often less stable, requiring careful calibration of the hyperparameters. This is especially the case for the ITR and VQA tasks. Figure 2 shows how slight variation in hyperparameters impacts the results of each fine-tuning approach, for the considered vision-language models. This shows that full fine-tuning consistently results in the highest variance.

B.3. Qualitative Analysis of VQA errors

Table 10 shows some examples of VQA errors from VQA-RAD and SLAKE datasets. The listed five image-question pairs come from the VQA-RAD dataset and the medical reasoning error only exists in SLAKE dataset.

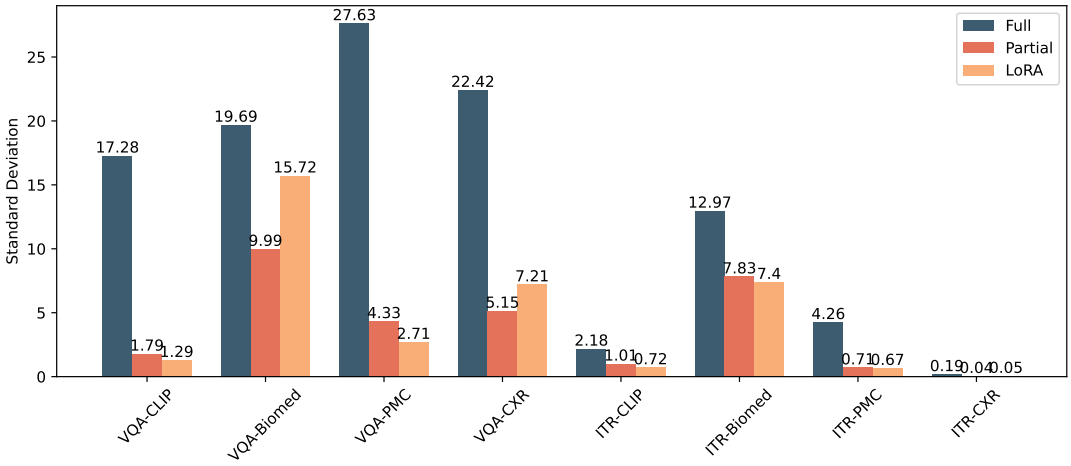


Figure 2: **Variation of model performance with different models and tasks**. Standard deviations of performance are reported to assess the stability of model performance on visual-question answering (metric: accuracy) over seven runs and image-to-text retrieval tasks (metric: recall@1) over nine runs.

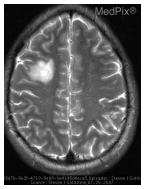
Image	Question	Answer	Prediction	Error Type	Model
	Is the consistency of the abscess located in the left upper quadrant homogeneous or heterogeneous?	heterogeneous	yes	Inappropriate for question type	Biomed
	What organ contains multiple lesions in the above image?	kidneys	left kidney	Wrong granularity	PMC
	Where is there evidence of a pleural effusion?	right side	right lung	Paraphrase	PMC
	What organ is enlarged?	pancreas	brain	Cross-modal error	CLIP
	What lobe of the brain is the lesion located in?	Right frontal lobe	right temporal lobe	Spatial reasoning	PMC
	Which organ is abnormal, heart or lung?	heart	lung	Medical reasoning	CXR

Table 10: Example prediction errors on VQA task. The first five examples come from the VQA-RAD dataset, and the last one comes from the SLAKE dataset.