Localization Lens for Improving Medical Vision-Language Models

Hasan Farooq[1][0009–0005–3782–8105], Murtaza Taj[1][0000–0003–2353–4462],
Mehwish Nasim[2][0000–0003–0683–9125], and Arif Mahmood[3][0000–0001–5986–9876]

[1] Lahore University of Management Sciences, Lahore, Pakistan,
[2] The University of Western Australia, Perth, Australia
[3] Information Technology University of the Punjab, Lahore, Pakistan.
hasan.farooq@lhr.nu.edu.pk, murtaza.taj@lums.edu.pk,
mehwish.nasim@uwa.edu.au, arif.mahmood@itu.edu.pk

Abstract. Medical Vision-Language Models (Med-VLMs) have demonstrated strong capabilities in clinical tasks. However, they often struggle to understand anatomical structures and spatial positioning, which are crucial for medical reasoning. To address this, we propose a **localization-aware** enhancement to the Med-VLM pipeline, introducing improvements at three levels: data, architecture, and alignment. First, we introduce **localization lens**, a set of expert-validated representations that provide richer anatomical and positional context. However, as these representations increase input complexity, we integrate pixel shuffle within the model architecture to filter and refine representations, enhancing spatial information processing while preserving anatomical continuity. Lastly, to effectively align the localization lens representations with textual features, we incorporate decoupled contrastive loss (DCL) alongside the standard loss function. This ensures better feature discrimination and robustness, particularly in data-limited medical settings. Through extensive evaluations on medical visual question answering (Med-VQA) datasets, we show that our methodology improves localization-driven performance across different Med-VLM architectures. Our analysis of localization-based questions further reveals that improvements in anatomy and spatial reasoning directly enhance the overall accuracy of Med-VQA upto 6.2%. The proposed approach is **model-agnostic** and can be seamlessly integrated into existing Med-VLM pipelines. The dataset, code, and trained models will be made publicly available at https://github.com/CVLABLUMS/localizationlens.

Keywords: Med-VLMs · VLMs · Visual Question Answering (VQA)

1 Introduction

Medical Vision-Language Models (Med-VLMs) have improved medical image interpretation by integrating vision and text understanding into a unified framework [5, 12]. These models support various clinical applications, including automated report generation, medical visual question answering (Med-VQA), and clinical decision support [5, 12]. Many approaches adopt contrastive learning frameworks, such as CLIP, where image and text encoders are trained jointly on large-scale datasets [4, 21, 29]. More recently, large language models (LLMs) have been integrated into Med-VLMs to enable broader multimodal reasoning [6, 21]. LLaVA-Med [11] fine-tunes a general-domain vision-language model on a figure-caption dataset with GPT-4-generated instruction data, while Med-Flamingo [18] is pre-trained on interleaved medical image-text pairs to improve few-shot generative VQA.

These contrastive and LLM-driven approaches demonstrate an increasing role of Med-VLMs in clinical reasoning and decision-making [26]. While these advancements have led to improvements in medical AI, Med-VLMs often rely on large architectures, requiring extensive computation and large datasets, thus limiting their accessibility [21]. An alternative involves developing smaller Med-VLMs that retain key functionalities while reducing training and inference costs [14,23]. However, since anatomical structures and spatial relationships are essential for accurate clinical reasoning [2, 8, 16], smaller models often have difficulty encoding fine-grained localization cues, which are necessary for tasks such as disease detection, lesion identification, and reporting [8, 14, 22, 27].

Several strategies have been explored to address these limitations, including data-centric methods, parameter-efficient scaling, advanced tokenization, region-based interpretations, and mixture-of-experts (MoE) approaches [8, 14, 22, 27]. For example, small language models (SLMs) have been proposed to improve reliability and accessibility in chronic disease management [23], while Med-MoE [8] uses multiple small Med-VLMs as part of an MoE framework to improve Med-VQA performance. In the region-based interpretations, methodologies have been proposed to improve the localization capability of existing models by integrating bounding boxes within VLMs [19, 25].

In this work, we explore the localization aspect in Med-VLMs by introducing a localization-aware enhancement that improves anatomical and spatial reasoning. Our approach consists of clinically meaningful representations, architectural modifications, and alignment refinements, making it model-agnostic and easily integrable into existing Med-VLMs. Our **contributions** are as follows:

(a) We propose clinically meaningful representations, validated by experts at a Hospital & Research Center, which act as **localization lens** to enhance anatomical and positional understanding in medical images.

(b) We integrate a **pixel-shuffle mechanism** within the model architecture to effectively handle the increased input complexity from the localization lens, improving the capture and refinement of spatial and anatomical details.

(c) We propose a **vision-language alignment** pipeline that first aligns the localization lens representations using decoupled contrastive learning, followed by their integration into Med-VQA tasks.

(d) We conduct a **localization analysis** to evaluate how the proposed localization-aware enhancements impact Med-VQA performance, demonstrating significant improvements in spatial reasoning and anatomical understanding across different Med-VLM architectures.

2 Methodology

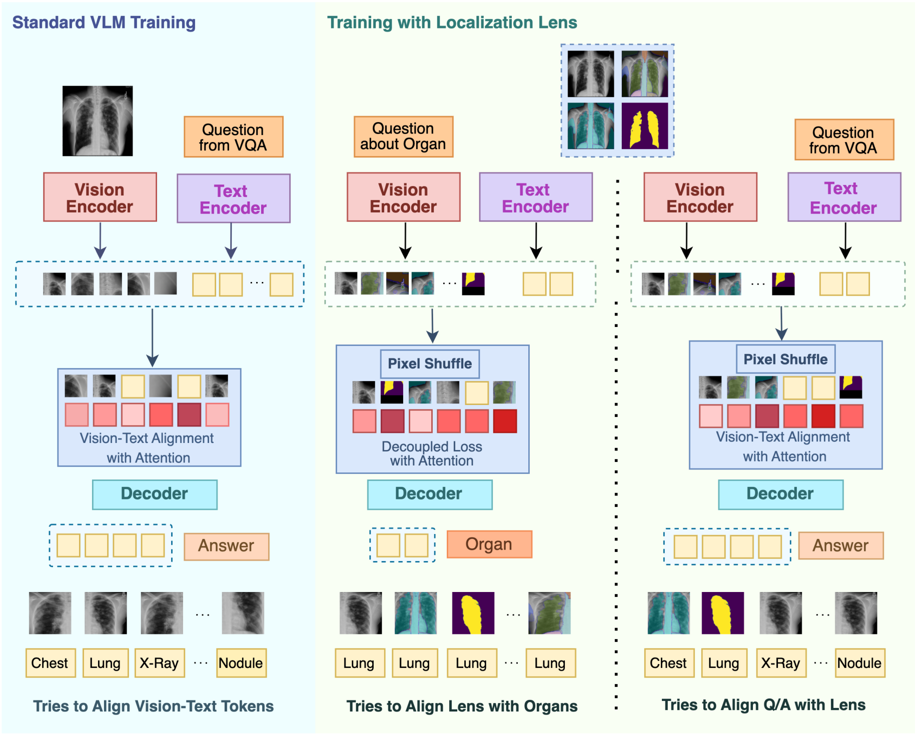
Standard VLM Training
Training with Localization Lens
Question from VQA
Question about Organ
Vision Encoder
Text Encoder
Pixel Shuffle
Vision-Text Alignment with Attention
Decoupled Loss with Attention
Decoder
Answer
Organ
Chest
Lung
X-Ray
Nodule
Tries to Align Vision-Text Tokens
Tries to Align Lens with Organs
Tries to Align Q/A with Lens

Fig. 1. Comparison of the existing standard Med-VLM training with the proposed training utilizing localization lens.

2.1 Preparation of Localization Lens

The standard medical visual question answering (Med-VQA) datasets (such as VQA-RAD [10] and SLAKE [13]) come with a medical image, questions related to the image, answers to the questions, and metadata. In this work, on top of the public datasets, we constructed a clinically meaningful dataset by augmenting each clinical image with several complementary representations. These representations include (a) original image, (b) single-color segmented representation of original image, (c) multi-color coded segmented representation of original image, and (d) masked representation of the original image. We refer to these as **localization lens** as they enhance the localization context.

The quality and clinical relevance of these representations have been validated by experts at a Hospital and Research Center. The complete dataset will be made publicly available at `https://github.com/CVLABLUMS/localizationlens`.

Uni and Multi-color Segmented Representations The segmented representations are obtained from the *Segment Anything Model* (SAM) [9] and *Med-SAM* [15]. The representations from SAM produce coarser segmentation maps that may not be medically accurate and precise [15, 24]. However, these representations help in the generalization of the model as observed in our ablation studies (see Tab. **??**). The representations from Med-SAM produce precise segmentation maps that capture anatomical structures and help the model learn the granular details of the medical image [15]. Let I be the original image, $f(\cdot)$ be the segmentation function, and M_i be a segmentation mask where $1 \leq i \leq n$.

$$\{M_1, M_2, \ldots, M_n\} = f(I). \tag{1}$$

Union of all segmentation masks results in full map:

$$\mathbf{M} = \{M_1 \cup M_2 \cup \ldots \cup M_n\}. \tag{2}$$

After obtaining these segmentation masks, we blend these with the original image using single or multi color-coding. The blending is formulated as:

$$I_{\text{blend}} = \alpha I + (1 - \alpha)\mathbf{C}, \quad \text{where} \quad \mathbf{C} = \begin{cases} \mathbf{c} \sum_{i=1}^{n} M_i & \text{(unicolor)} \\ \sum_{i=1}^{n} M_i \mathbf{c}_i & \text{(multicolor)}, \end{cases} \tag{3}$$

where α samples values from a normal distribution 0.50 ± 0.10 for balanced blending and $\mathbf{c}_i \in \mathbb{R}^3$ represents color for $i^t h$ mask M_i.

Masked Representations Each segment in the map represents an instance in the medical image. For example, in a chest X-Ray image, the question-answer pairs can be related to lungs. Masking the background regions assists the model to focus on learning the relevant features. We obtain these masked representations from segmentation maps with the help of experts. However, these masks can also be automatically obtained from medical segmentation models or medical vision language models. As a proof-of-concept, in our code, we share a VLM based promptable pipeline for automatically extracting these masks on other datasets (see Fig. 1).

2.2 Architectural Changes

As shown in Fig. 1, the standard VLM training is optimized by aligning the image and text tokens. However, as we introduce context complexity by incorporating multiple representations, we optimize the architecture in two ways: (a) **input representation** and (b) **filtering representations**.

Instead of relying solely on an original image-text pair, our approach constructs an input that aggregates multiple visual representations along with the text. For each original image-text pair, we randomly select two representations from (a) single-color segmentation maps, (b) multi-color segmentation maps, and (c) masked representations. These modalities are independently encoded by the vision encoder, and their embeddings are fused to form a unified representation.

$$E_{\text{fused}} = v(E_{\text{orig}}, E_{\text{rep1}}, E_{\text{rep2}}) \tag{4}$$

where E_{orig} is the embedding of the original image, and $E_{\text{rep1}}, E_{\text{rep2}}$ are the embeddings of the selected representations. The function v represents the attention-based vision encoder mechanism.

Once the input is fused and partitioned into patches for processing, a **pixel shuffle mechanism** [7] is applied to refine the visual tokens. Unlike standard patch partitioning, which divides an image into fixed-size non-overlapping patches, pixel shuffle takes an alternative approach by restructuring the spatial information into a more compact form. Given an input of shape $(H, W, D \times r^2)$, pixel shuffle rearranges it into:

$$T_{\text{out}}(h', w', d) = T\left(\frac{h'}{r}, \frac{w'}{r}, d \times r^2 + (h' \mod r) \times r + (w' \mod r)\right) \tag{5}$$

which effectively reduces the number of patches while preserving local dependencies. This method provides two key advantages over conventional partitioning:

- Instead of treating adjacent pixels independently, pixel shuffle redistributes sub-pixel information, reducing redundant tokenization. This helps the model filter relevant tokens for attention.
- Unlike uniform patch extraction, which can break meaningful structures, pixel shuffle retains continuity by preserving fine-grained details within each patch. This helps the training as the model focuses on patches from different representations.

By filtering representations more effectively, pixel shuffle enables the model to focus on semantically rich tokens, improving both computational efficiency and vision-text alignment.

2.3 Training Details

Since we introduce complexity by adding **localization lens**, our training is divided into two sequential phases designed to improve the vision-text alignment.

Alignment of Visual Representations The first phase serves as pre-training where we train the model for visual representations only. The organ represented in the image is used as text prompt, ensuring the encoder adapts to organ-specific features without interference from additional modalities. We employ **decoupled contrastive loss (DCL)** [28], which differs from standard InfoNCE by eliminating explicit negative repulsion. The InfoNCE loss is:

$$L_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_k \exp(\text{sim}(z_i, z_k)/\tau)} \tag{6}$$

where z_i, z_j are a positive pair, and negatives z_k force repulsion. In medical imaging, this can suppress meaningful variations, specifically when the data is scarce. DCL removes this constraint:

$$L_{DC} = -\log \frac{\exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau)}{\exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau) + \sum_k \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_k \rangle / \tau)} \tag{7}$$

where $\mathbf{z}_i^{(1)}$ and $\mathbf{z}_i^{(2)}$ are two augmented views of the same sample i.

The model aims to maximize their similarity while handling negatives separately, avoiding repulsion. This improves: (i) *representation alignment*, as similar medical transformations remain close; (ii) *gradient stability*, avoiding conflicting updates; and (iii) *low-batch efficiency*, making it suitable for small datasets. This pre-training step enhances domain adaptation and prepares the encoder for multimodal fusion.

Alignment of Question-Answer Pairs In the second phase, we train the model for Med-VQA alignment. We use the actual question-answer pairs as text prompt along with the **localization lens**. Since, we are training on the question-answer pairs, the standard **InfoNCE** contrastive loss (equation 6) is used for vision-text alignment. This penalizes the negative text tokens as well and ensures the alignment of visual representations with the question-answer (QA) pairs.

2.4 Inference

The localization lens are medically sound representations that enhance the anatomical and positional details whereas the proposed architectural design and alignment help the model extract patterns from the localization lens. After the model has been trained with the localization lens, we can perform vision-language tasks on the **original image** only.

3 Experiments

3.1 Medical VLMs

We select 4 existing SOTA medical VLMs (RaDialog [20], Med-Flamingo [18], LLaVA-Med [11], and Med-MoE (Phi) [8]) and 2 non-medical VLMs (Phi [30] and SmolVLM [1]]) for our experiments. We follow the training pipeline in Sec. 2.3 and use the VQA-RAD [10] and SLAKE [13] **datasets** for Med-VQA. To ensure a fair comparison, we follow a 70:30 train-test split.

In our experiments, we use 16-bit precision, learning rate of 0.001, weight decay of 0.01, and optimize using AdamW. For pixel-shuffle, we use an upscaling factor of 4. We report accuracy for closed-category questions and recall for open-category questions. Our results show that the performance of the Med-VLMs improves significantly with the proposed methodology (see Table 1).

From our results, we infer that the proposed methodology improves the performance of both the medical and non-medical VLMs. Interestingly, training models from scratch (i.e. non-medical models) shows more improvement. The possible explanation for this is the pre-training of VLMs that align the localization lens with organ labels more effectively [3]. Moreover, the results show that the small VLMs are capable of performance improvement with effective training pipelines and architectural design.

Table 1. Comparison of different Med-VQA models on the VQA-RAD and SLAKE.

Model	**VQA-RAD** [10]		**SLAKE** [13]		**Model Size**
	Open	Closed	Open	Closed	
Medical Vision Language Models					
RaDialog [20]	54.6	57.9	51.4	56.3	7B
RaDialog (with Lens)	55.3↑0.7	60.0↑2.1	53.4↑2.0	61.1↑4.8	7B
Med-Flamingo [18]	64.1	70.7	61.2	68.2	7B
Med-Flamingo (with Lens)	65.8↑1.7	72.1↑1.4	65.0↑3.8	71.6↑3.4	7B
LLaVA-Med [11]	61.2	76.2	70.4	75.0	7B
LLaVA-Med (with Lens)	62.4↑1.2	79.8↑3.6	74.0↑3.6	79.2↑4.2	7B
Small Medical Vision Language Models					
Phi-MoE [8]	36.7	61.8	43.9	57.0	3.6B
Phi* (with Lens)	48.3	56.7	51.3	58.4	2.7B
SmolVLM [1]	53.8	57.6	46.5	55.9	1.7B
SmolVLM (with Lens)	58.0↑4.2	63.5↑5.9	54.2↑7.7	59.0↑3.1	1.7B
SmolVLM [1]	48.3	51.4	43.2	49.8	0.5B
SmolVLM (with Lens)	53.6↑5.3	59.2↑7.8	49.3↑6.1	54.7↑4.9	0.5B

***** The results could not be reproduced, and we resorted to the base model.

3.2 Localization Analysis

The Med-VQA datasets come with different types of Q/A pairs (i.e. disease identification, anatomical representation, positioning, modality, severity) [10,13]. We conduct an extensive analysis on the Q/A pairs involving localization. We define localization as questions that (a) focus on presence of an anatomical structure or (b) focus on the positioning. We filter the questions that focus on the localization. Since there are thousands of Q/A pairs, we use the Llama-Vision (11B) [17] for filtering localization pairs.

Table 2. Localization improvement results on combined dataset.

Model	**Anatomical**	**Positioning**
LLAVA-Med [11]	4.7% ↑	5.6% ↑
Med-Flamingo [18]	3.3% ↑	5.2% ↑
Phi* [8, 30]	2.8% ↑	3.6% ↑
SmolVLM [1] (0.5B)	4.2% ↑	5.8% ↑
SmolVLM [1] (1.7B)	3.7% ↑	6.1% ↑

***** The results could not be reproduced, and we resorted to the base model.

Table 2 shows the results of our analysis on the combined VQA-RAD [10] and SLAKE [13] datasets with recall as the evaluation metric. Our results show that the proposed methodology significantly (upto 8.0%) improves the performance

Table 3. Ablation study on VQA-RAD [10] with SmolVLM (1.7B).

Lens			Loss		Shuffle	Metrics	
SAM	**Med-SAM**	**Mask**	**InfoNCE**	**DCL**	**Pixel**	**Overall**	**Localization**
✗	✗	✗	✓	✗	✗	55.4%	-
✓	✗	✗	✗	✓	✓	57.2%	4.9% ↑
✗	✓	✗	✗	✓	✓	56.8%	3.6% ↑
✓	✓	✗	✗	✓	✓	59.4%	7.4% ↑
✓	✓	✓	✓	✗	✗	57.3%	5.2% ↑
✓	✓	✓	✓	✗	✓	58.8%	6.2% ↑
✓	✓	✓	✗	✓	✗	57.7%	5.6% ↑
✓	✓	✓	✗	✓	✓	61.6%	8.0% ↑

of Med-VLMs on the localization Q/A pairs. These results suggest that improvements in localization-based Q/A pairs contribute to the overall improvement in performance across different Med-VLMs.

3.3 Ablation Studies

To determine the contribution of each visual representation, we performed a series of ablation experiments. We compared the performance of models trained with only original images, original images with SAM [9] segmentation maps, original images with Med-SAM [15] segmentation maps, original images with both segmentation maps, and original images with both segmentation sources plus masked images. Our ablation study indicates that while both SAM [9] and Med-SAM [15] improve results independently, their combined representations improves results by up to 6.2%. Also, although the (binary) masked representations alone do not represent meaningful structures, using these masks with segmentation maps achieves the best results.

4 Conclusion

In this work, we integrate localization context as **localization lens**, a set of clinically meaningful representations, on the top of publicly available Med-VQA datasets. This localization lens adds both useful context and complexity for training the Med-VLMs. To address this, we propose integrating (a) **pixel-shuffle mechanism** within the architecture for filtering relevant context and (b) vision-text alignment with **decoupled contrastive loss**. These changes in the architecture design and vision-text alignment enable the Med-VLMs capture the context using localization lens effectively. Our experiments on Med-VQA datasets illustrate how the proposed methodology improves the localization context and subsequently, improves performance across Med-VLMs. The proposed methodology is **model-agnostic** and can be integrated into existing Med-VLM pipelines to improve their performance.

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

References

1. SmolLM: Everything about the SmolLM2 and SmolVLM family of models (2024), `https://github.com/huggingface/smollm`
2. Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal Biomedical AI. Nature Medicine **28**(9), 1773–1784 (2022)
3. Christophe, C., Raha, T., Maslenkova, S., Salman, M.U., Kanithi, P., Pimentel, M.A., Khan, S.: Beyond fine-tuning: Unleashing the potential of continuous pre-training for clinical LLMs. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 10549–10561 (Nov 2024)
4. Eslami, S., Meinel, C., De Melo, G.: PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In: Findings of the Association for Computational Linguistics. pp. 1151–1163 (2023)
5. Hartsock, I., Rasool, G.: Vision-language models for medical report generation and visual question answering: a review. Frontiers in Artificial Intelligence (2024)
6. Huix, J.P., Ganeshan, A.R., Haslum, J.F., Söderberg, M., Matsoukas, C., Smith, K.: Are natural domain foundation models useful for medical image classification? In: Proc. of the Winter Conference on Applications of Computer Vision (2024)
7. Ibrahem, H., Salem, A., Kang, H.S.: Pixel shuffling is all you need: spatially aware convmixer for dense prediction tasks. Pattern Recognition **158**, 111068 (2025)
8. Jiang, S., Zheng, T., Zhang, Y., Jin, Y., Yuan, L., Liu, Z.: Med-MoE: Mixture of domain-specific experts for lightweight medical vision-language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 3843–3860 (2024)
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P.: Segment anything. In: In. Proc. of Int. Conf. on Computer Vision. pp. 4015–4026 (2023)
10. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data **5**(1), 1–0 (2018)
11. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems. vol. 36 (2024)
12. Li, X., et al.: Vision-language models in medical image analysis: From simple fusion to general large models. Information Fusion (2025)
13. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: Int. Symposium on Biomedical Imaging. pp. 1650–1654 (2021)
14. Liu, D., et al.: SPHINX-x: Scaling data and parameters for a family of multi-modal large language models. In: Proc. of the Int. Conf. on Machine Learning. pp. 32400–32420. PMLR (2024), `https://Proc..mlr.press/v235/liu24cc.html`
15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15** (2024)
16. Magnini, M., Aguzzi, G., Montagna, S.: Open-source small language models for personal medical assistant chatbots. Intelligence-Based Medicine p. 100197 (2025)

17. Meta AI: Llama 3.2 vision model card (2024), https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD_VISION.md
18. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-Flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health. pp. 353–367. PMLR (2023)
19. Müller, P., Kaissis, G., Rueckert, D.: ChEX: Interactive Localization and Region Description in Chest X-Rays. In: European Conference on Computer Vision. Springer, Cham (2025)
20. Pellegrini, C., Özsoy, E., Busam, B., Navab, N., Keicher, M.: Radialog: A large vision-language model for radiology report generation and conversational assistance. arXiv preprint arXiv:2311.18681 (2023)
21. Shakeri, F., et al.: Few-shot adaptation of medical vision-language models. In: Medical Image Computing and Computer Assisted Intervention. pp. 553–563 (2024)
22. Shen, Z., Tao, T., Ma, L., Neiswanger, W., Liu, Z., Wang, H., Tan, B., Hestness, J., Vassilieva, N., Soboleva, D., Xing, E.: SlimPajama-DC: Understanding data combinations for LLM training (2024), https://arxiv.org/abs/2309.10818
23. Shi, B., Wu, Z., Mao, M., Wang, X., Darrell, T.: When do we not need larger vision models? In: European Conference on Computer Vision. pp. 444–462. Springer, Cham (2025)
24. Sun, J., Chen, K., He, Z., Ren, S., He, X., Liu, X., Peng, C.: Medical image analysis using improved SAM-Med2D: Segmentation and classification perspectives. BMC Medical Imaging **24** (2024)
25. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: In Proc. of Int. Conf. on Computer Vision and Pattern Recognition. pp. 7433–7442 (2023)
26. Thirunavukarasu, A.J., Ting, D.S., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.: Large language models in medicine. Nature Medicine pp. 1930–1940 (2023)
27. Xue, F., Fu, Y., Zhou, W., Zheng, Z., You, Y.: To repeat or not to repeat: Insights from scaling llm under token-crisis. In: Advances in Neural Information Processing Systems. vol. 36, pp. 59304–59322 (2023)
28. Yeh, C.H., Hong, C.Y., Hsu, Y.C., Liu, T.L., Chen, Y., LeCun, Y.: Decoupled contrastive learning. In: European Conf. on Computer Vision. pp. 668–684 (2022)
29. Zhang, S., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI p. AIoa2400640 (2025)
30. Zhu, M., Zhu, Y., Liu, X., Liu, N., Xu, Z., Shen, C., Peng, Y., Ou, Z., Feng, F., Tang, J.: A comprehensive overhaul of multimodal assistant with small language models. In: Proc. of the AAAI Conf. on Artificial Intelligence (2024)