Localization-Grounded Supervision: Revisiting Vanilla SFT of Large Vision-Language Models for Medical Image Analysis

Yiming Shi[1†], Hengyu Zhang[1†], Jiawei Chen[4†], Shaoshuai Yang[1], Xi Chen[1], Xun Zhu[1], Kaiwen Wang[1], Haolin Li[1], Xiuheng Zhao[1], Shuheng Zhang[5], Miao Li[1✉], and Ji Wu[1,2,3✉]

[1] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
sym23@mails.tsinghua.edu.cn
{miao-li, wuji_ee}@tsinghua.edu.cn
[2] College of AI, Tsinghua University, Beijing 100084, China
[3] Beijing National Research Center for Information Science and Technology, Beijing 100084, China
[4] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China
[5] Shanghai Medical Image Insights Intelligent Technology Co., Ltd., Shanghai 200232, China

Abstract. Large vision-language models (LVLMs) achieve strong performance on general-domain tasks. Supervised fine-tuning (SFT) is commonly used for domain adaptation. However, unlike traditional pretrain-finetune paradigms, the benefits of large-scale pretraining do not reliably transfer to LVLMs in medical image analysis. Using the proposed DePass-VL framework, we quantitatively analyze the failure of LVLMs. Together with attention-based qualitative analysis, we find that LVLMs struggle to effectively focus on task-relevant regions, indicating a lack of fine-grained vision-language semantic alignment. Although pretrained LVLMs exhibit strong 2D spatial alignment, this capability is not effectively activated during SFT. To address this gap, we propose Localization-Grounded Supervision (LGS), which introduces explicit localization signals to accelerate semantic alignment and improve both performance and interpretability. For 3D medical imaging tasks, pretrained LVLMs lack intrinsic 3D spatial alignment due to limited volumetric exposure. We therefore incorporate 3D anatomical localization data through multi-task learning to jointly build spatial and semantic alignment in 3D space. Experiments across anatomical regions, imaging modalities, model architectures, and parameter scales demonstrate that LGS is a simple yet effective paradigm for LVLM adaptation in medical image analysis. The code is available at: https://github.com/MSIIP/LGS

Keywords: Medical Image Analysis · Large Vision-Language Models · Supervised Fine-Tuning · Localization-Grounded Supervision.

† Equal contribution
✉ Corresponding authors: Ji Wu and Miao Li

1 Introduction

The rapid progress of large language models (LLMs) have accelerated the development of large vision-language models (LVLMs) [15, 2]. Both proprietary [1, 19] and open-source LVLMs [3, 22] have demonstrated strong zero-shot generalization across multimodal tasks [24, 23]. However, for complex and domain-specific tasks such as medical image analysis, zero-shot performance often remains insufficient for reliable deployment [20, 14]. In practice, supervised fine-tuning (SFT) remains a fundamental approach for domain adaptation.

In traditional deep learning (DL), pretrained encoders consistently improve downstream performance (Tab. 1) compared to training from scratch [17, 26, 5]. However, this empirical pattern does not reliably extend to LVLMs. In fine-grained diagnostic tasks, such as small-lesion lung nodule detection, LVLMs pretrained on large-scale multimodal corpora often exhibit limited performance gains after SFT. The advantages of pretraining are not fully translated into downstream improvements, suggesting that transfer dynamics in LVLMs may differ fundamentally from those in conventional DL models.

To investigate this discrepancy, we analyze LVLM behavior from both qualitative and quantitative perspectives. Attention visualization reveals that under vanilla SFT, attention distributions are diffuse and fail to consistently concentrate on task-relevant regions. To quantify this phenomenon, we propose DePass-VL, the first adaptation of DePass [12] to LVLMs, enabling regional contribution analysis for LVLMs. By combining attention inspection with decomposed forward attribution, we identify insufficient fine-grained vision-language semantic alignment as the primary factor limiting SFT performance.

Motivated by this finding, and recognizing that large-scale pretraining has endowed LVLMs with well-established 2D vision-language spatial alignment [13, 25], we introduce Localization-Grounded Supervision (LGS). LGS incorporates explicit localization signals into task data, leveraging pretrained spatial alignment to accelerate the formation of fine-grained semantic alignment during SFT. This complementary supervision mechanism not only improves downstream performance but also enhances spatial interpretability.

We note that localization coordinate prediction and grounded medical VLMs have been explored in prior works [4, 16, 6, 27]. Our goal is not to claim coordinate prediction itself as a new capability. Instead, we use localization as a simple supervision interface to revisit vanilla SFT of off-the-shelf LVLMs in fine-grained medical image analysis. In this setting, LGS is designed to test whether explicit spatial supervision can help LVLMs better leverage their pretrained spatial alignment and form task-relevant semantic alignment during SFT.

Beyond 2D tasks, 3D medical image analysis presents additional challenges. Most existing LVLMs have limited exposure to volumetric data during pretraining and therefore lack intrinsic 3D vision-language spatial alignment. Since semantic alignment depends on spatial grounding, directly applying LGS to 3D tasks may yield limited gains. To address this issue, we introduce 3D anatomical localization data and adopt a multi-task learning strategy to jointly construct

spatial and semantic alignment in volumetric space, thereby extending LGS to 3D medical imaging scenarios.

Our main contributions are summarized as follows:

1. We conduct systematic quantitative and qualitative analyses of LVLM behavior based on the proposed **DePass-VL** framework and attention visualization, indicating that insufficient fine-grained semantic alignment is an important factor in fine-grained medical tasks.
2. We introduce **LGS**, a simple and architecture-agnostic training strategy that incorporates explicit localization signals to leverage pretrained spatial alignment and accelerate semantic alignment during SFT.
3. For **3D medical imaging tasks** where pretrained LVLMs lack intrinsic 3D spatial alignment, we integrate 3D anatomical localization data via multi-task learning to jointly construct spatial and semantic alignment in 3D space.

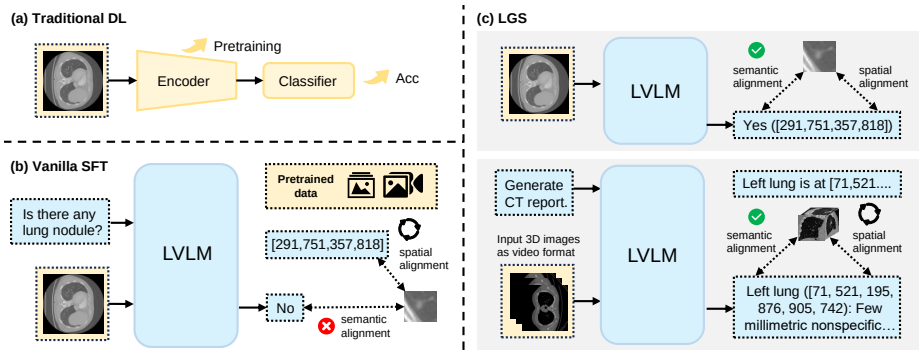

Fig. 1. (a) Traditional DL benefits significantly from pretrained encoders. (b) Vanilla SFT fails to establish fine-grained vision-language semantic alignment and does not effectively leverage the pretrained spatial alignment of LVLMs. (c) LGS introduces explicit localization signals to leverage pretrained spatial alignment and accelerate semantic alignment. In 3D scenarios, incorporating anatomical localization data with joint training enables the construction of both spatial and semantic alignment.

2 Why Vanilla SFT Fails?

2.1 DePass-VL Analysis

For comprehensive attribution analysis, we propose DePass-VL, the first adaptation of DePass [12] to LVLMs, supporting region-level contribution analysis. Following the notation in Sec. 2.2, let the Transformer [21] decoder hidden states of the LVLM at layer l be

$$\mathbf{h}^{(l)} \in \mathbb{R}^{N \times d}. \tag{1}$$

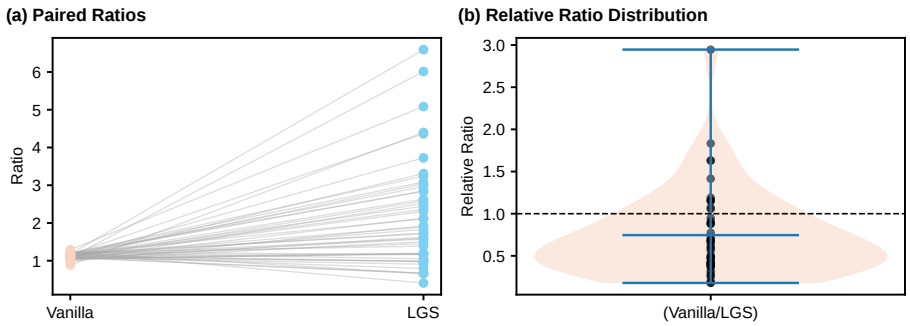

Fig. 2. (a) The y-axis shows the ratio between the contribution from task-relevant image regions and that from non-relevant regions. A value greater than 1 indicates that the relevant regions contribute more to the final answer. Gray lines connect the same samples under vanilla SFT and LGS. (b) The y-axis shows the per-sample relative ratio of contribution scores between Vanilla SFT and LGS. A value below 1 indicates that LGS increases the LVLM's focus on task-relevant regions for that sample.

We introduce a decomposed representation $\mathbf{h}_{\text{dec}}^{(l)}$, such that the hidden states remain additively decomposed:

$$\mathbf{h}^{(l)}[i,:] = \sum_{m=0}^{M-1} \mathbf{h}_{\text{dec}}^{(l)}[i,m,:]. \tag{2}$$

In our setting, we set $M = 3$, corresponding to task-relevant image regions $\mathcal{R}_{V,\text{tr}}$, non-relevant image regions $\mathcal{R}_{V,\text{nr}}$, and other tokens $\mathcal{R}_O$. At the embedding layer ($l = 0$), the decomposition is initialized by masking:

$$\mathbf{h}_{\text{dec}}^{(0)}[i,m,:] = \begin{cases} \mathbf{h}^{(0)}[i,:], & \text{if } m = 0,\ i \in \mathcal{R}_{V,\text{tr}}, \\ \mathbf{h}^{(0)}[i,:], & \text{if } m = 1,\ i \in \mathcal{R}_{V,\text{nr}}, \\ \mathbf{h}^{(0)}[i,:], & \text{if } m = 2,\ i \in \mathcal{R}_O, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{3}$$

As shown in the DePass paper [12], due to the linearity of MHSA, residual connections, and linear projections within each decoder layer, the decomposed components remain linearly additive throughout forward propagation. Let $\mathbf{w}_y$ denote the LM head vector corresponding to target token y. For the answer position index $t = N_v + N_q + N_a$, the component-wise contribution scores are

$$\text{score}_y^{V,\text{tr}} = \mathbf{w}_y^\top \mathbf{h}_{\text{dec}}^{(L)}[t,0,:], \tag{4}$$

$$\text{score}_y^{V,\text{nr}} = \mathbf{w}_y^\top \mathbf{h}_{\text{dec}}^{(L)}[t,1,:], \tag{5}$$

$$\text{score}_y^{O} = \mathbf{w}_y^\top \mathbf{h}_{\text{dec}}^{(L)}[t,2,:]. \tag{6}$$

These scores quantify the exact contribution from task-relevant regions, non-relevant regions, and other tokens to the final prediction.

As shown in Fig. 2, under vanilla SFT, the contribution ratio between task-relevant and non-relevant regions is close to 1, indicating insufficient fine-grained vision-language semantic alignment and weak reliance on relevant regions.

2.2 Attention-Based Analysis

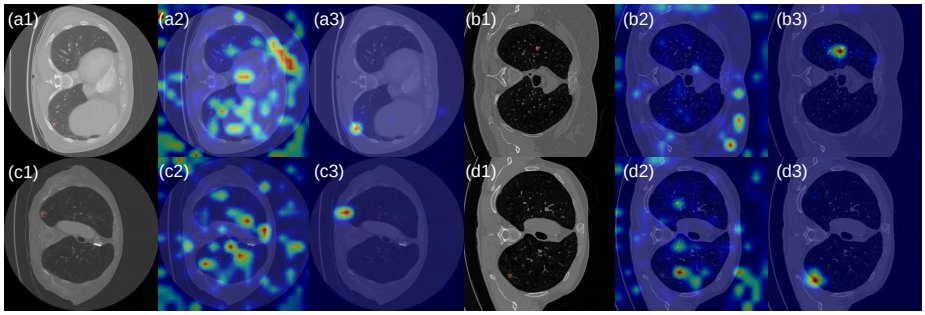

Fig. 3. (a1)(b1)(c1)(d1) show the original images of four different samples, with lesions highlighted by red bounding boxes. (a2)(b2)(c2)(d2) present the corresponding attention maps obtained after vanilla SFT. (a3)(b3)(c3)(d3) show the attention distributions produced by LVLMs trained with LGS.

Since the language decoder of LVLMs is based on the Transformer [21] architecture, we analyze self-attention weights to measure how answer tokens attend to image regions. Let the input to a decoder layer be

$$\mathbf{h} = [\mathbf{h}_v; \mathbf{h}_q; \mathbf{h}_a] \in \mathbb{R}^{N \times d}, \tag{7}$$

where $N = N_v + N_q + N_a$, and $\mathbf{h}_v$, $\mathbf{h}_q$, $\mathbf{h}_a$ denote image, question, and answer tokens, respectively.

For the l-th layer and m-th head, the scaled dot-product attention is

$$\mathbf{A}^{(l,m)} = \text{Softmax}\left(\frac{(\mathbf{h}\mathbf{W}_Q^{(l,m)})(\mathbf{h}\mathbf{W}_K^{(l,m)})^\top}{\sqrt{d_h}}\right), \tag{8}$$

where d_h is the head dimension and $\mathbf{A}^{(l,m)} \in \mathbb{R}^{N \times N}$.

We focus on the attention from answer tokens to image tokens. For the i-th answer token,

$$\mathbf{a}_{i \to v}^{(l,m)} = \mathbf{A}_{N_v+N_q+i-1,\, 0:N_v-1}^{(l,m)}, \tag{9}$$

which represents its attention distribution over image tokens. Since image tokens correspond to patches, $\mathbf{a}_{i \to v}^{(l,m)}$ can be reshaped to the spatial layout of the image.

As shown in Fig. 3, the LVLM trained with vanilla SFT exhibits diffuse attention patterns and fails to consistently focus on task-relevant regions.

3 Localization-Guided Supervision

Under the vanilla SFT setting, disease diagnosis tasks typically require LVLMs to directly predict the diagnostic outcome (see Fig. 1). For complex cases, particularly when lesions are small, such outcome-only supervision provides limited spatial constraints, making it difficult for LVLMs to consistently focus on diagnostically relevant regions during training.

To strengthen the supervision signal, LGS adopts a complementary output format in which LVLMs predict both the diagnostic result and the corresponding localization coordinates, e.g., "Yes, $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$." Training follows the standard autoregressive next-token cross-entropy objective. Specifically, given an input image-text pair X and a target response sequence Y, the training loss is

$$\mathcal{L}_{\mathrm{LGS}} = -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, X), \tag{10}$$

where Y contains both the answer and textual localization coordinates. By explicitly incorporating spatial information into the target sequence, this design promotes the establishment of fine-grained vision-language semantic alignment.

Importantly, the coordinate format shown in Fig. 1 is aligned with the official demonstration format [3], allowing the model to better leverage its pretrained 2D vision-language spatial alignment. In contrast, pretrained LVLMs lack intrinsic spatial alignment for 3D medical images due to limited volumetric exposure. To address this gap, we introduce 3D localization tasks and employ multi-task learning to jointly construct spatial and semantic alignment during SFT.

4 Experiment

4.1 Implementation Details

DePass-VL is implemented based on the official DePass [12] repository. Model training is conducted using ms-swift [29], with Qwen3-VL-2B-Instruct [3] as the base model. The batch size is set to 128. We use AdamW optimizer with an initial learning rate of 2×10^{-5} and a warm-up ratio of 0.03.

4.2 2D Disease Diagnosis

To mitigate pretraining-data overlap, we derive 2D slice-level lesion-presence classification datasets from 3D segmentation data, rather than original 3D segmentation tasks. For example, in LUNA16 [18], we select CT scans with a single lung nodule, use the middle slice of the nodule mask as the positive sample, and use an adjacent slice just outside the mask boundary as the negative sample. This design creates lesion-near negatives and reduces coarse anatomical shortcuts. Following this strategy, we construct a lung nodule diagnosis dataset based on LUNA16 [18] and a vestibular schwannoma diagnosis dataset based on CrossMoDA2021 [7], both with a balanced 1:1 ratio of positive and negative samples.

Table 1. Fine-tuning performance on 2D disease diagnosis.

Method	Dataset	Accuracy	F1
CNN [11] (From scratch)	LUNA16	0.5530	0.5755
CNN [11] (Pretrained)	LUNA16	0.6439	0.6116
ViT [8] (From scratch)	LUNA16	0.5227	0.3226
ViT [8] (Pretrained)	LUNA16	0.5530	0.5203
CLIP [17]	LUNA16	0.5909	0.5714
SigLIP [26]	LUNA16	0.5606	0.5151
DINO [5]	LUNA16	0.5833	0.6154
Qwen3-VL [3] (Vanilla SFT)	LUNA16	0.5227	0.6702
Qwen3-VL [3] (LGS)	LUNA16	**0.7652**	**0.7559**
Qwen3-VL [3] (Vanilla SFT)	CrossMoDA2021	0.5938	0.6176
Qwen3-VL [3] (LGS)	CrossMoDA2021	**0.9844**	**0.9846**

Vanilla SFT and LGS are compared under the same base model, data split, learning rate, batch size, optimizer, training epochs, parameter-freezing strategy, and prompt setting, with the only intended difference being that LGS adds localization coordinates to the answer target. As shown in Tab. 1, LGS consistently outperforms vanilla SFT on disease diagnosis datasets across different anatomical regions, demonstrating its effectiveness.

4.3 3D Report Generation

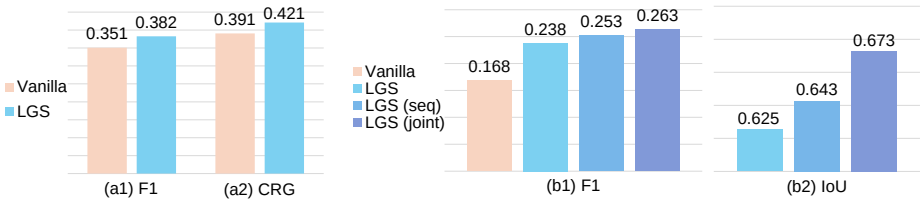

Fig. 4. (a1)(a2) Fine-tuning performance on 3D report generation using RadGenome-Chest CT. (b1) Fine-tuning performance on **filtered** RadGenome-Chest CT, retaining only samples with both reports and segmentation masks. (b2) 3D localization performance of LGS-trained LVLMs on the filtered dataset.

We construct a chest CT report generation benchmark based on RadGenome-Chest CT [28]. Generated reports are evaluated using the official VLM3D Challenge framework [10], with emphasis on clinically relevant metrics, including F1 Score and CRG Score [9]. To control token length for 3D data, CT volumes are resized to $128 \times 256 \times 256$ and provided to the LVLM in a video-like format.

As shown in Fig. 4 (a), LGS significantly outperforms vanilla SFT in report generation. Since not all CT cases in the original dataset contain both reports

and segmentation masks, we further train on a filtered subset with matched annotations, where LGS achieves larger performance gains (Fig. 4 (b1)).

Given the limited 3D spatial alignment of pretrained LVLMs, we construct a 3D anatomical localization VQA dataset using masks from RadGenome-Chest CT. This 3D localization task is integrated with LGS in both sequential and multi-task joint training settings. Results (Fig. 4 (b1)) indicate that incorporating 3D localization further improves report generation performance, with multi-task learning achieving the best results. The 3D localization outcomes (Fig. 4 (b2)) demonstrate accurate region-level grounding, confirming that LGS enhances spatial alignment in 3D scenarios.

4.4 Robustness Across Architectures and Scales

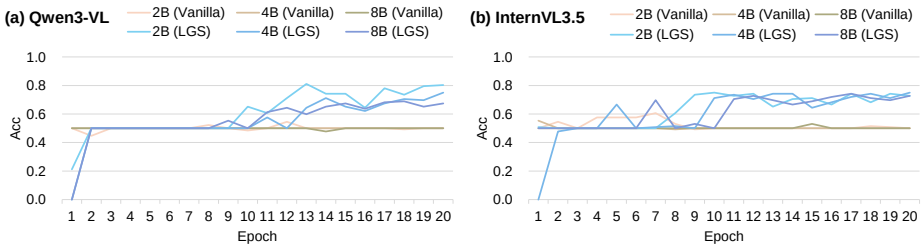

Fig. 5. Fine-tuning performance of disease diagnosis (LUNA16) for (a) Qwen3-VL and (b) InternVL3.5 across different parameter scales.

To evaluate the robustness of LGS, we conduct experiments on LVLMs with different architectures and parameter scales, including the 2B, 4B, and 8B variants of Qwen3-VL [3] and InternVL3.5 [22]. As shown in Fig. 5, vanilla SFT exhibits performance limitations across architectures and model scales, whereas LGS consistently yields performance gains. Results suggest that insufficient fine-grained vision-language semantic alignment is a common issue during LVLM fine-tuning, and that LGS provides a robust and architecture-agnostic solution.

4.5 Alignment Behavior and Attention Analysis

As shown in Fig. 2 and 3, vanilla SFT exhibits diffuse attention for task-relevant regions, indicating insufficient fine-grained vision-language semantic alignment. In contrast, LGS concentrates answer-token attention on relevant image regions, reflecting stronger spatial grounding and more stable semantic alignment.

5 Conclusion

We investigate the limitations of SFT for LVLMs in medical image analysis through qualitative attention analysis and the proposed quantitative framework

DePass-VL. Our study reveals that insufficient fine-grained vision-language semantic alignment is the primary cause of performance degradation. To address this issue, we introduce LGS, which leverages pretrained spatial alignment by incorporating explicit localization signals into task data, improving both performance and interpretability. For 3D scenarios, we further combine LGS with multi-task learning to explicitly construct volumetric spatial alignment. Overall, LGS provides a simple, scalable, and architecture-agnostic paradigm for effective domain adaptation of LVLMs.

Acknowledgments. This study was funded by Tsinghua University-Beijing Electronics Holding Corporation, Ltd. Joint Research Center for Chip-Display Fusion and System Integration Technology (JCCDFSIT) (grant NO.2026CDF001), Beijing Natural Science Foundation NO.L251072, Beijing Natural Science Foundation NO.4252046. This study was supported by HUAWEI's Al Hundred Schools Program and was carried out using the Ascend AI technology stack.

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

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
6. Deperrois, N., Matsuo, H., Ruipérez-Campillo, S., Vandenhirtz, M., Laguna, S., Ryser, A., Fujimoto, K., Nishio, M., Sutter, T.M., Vogt, J.E., et al.: Radvlm: a multitask conversational vision-language model for radiology. arXiv preprint arXiv:2502.03333 (2025)
7. Dorent, R., Kujawa, A., Ivory, M., Bakas, S., Rieke, N., Joutard, S., Glocker, B., Cardoso, J., Modat, M., Batmanghelich, K., et al.: Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. Medical Image Analysis **83**, 102628 (2023)
8. 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)

9. Hamamci, I.E., Er, S., Shit, S., Reynaud, H., Kainz, B., Menze, B.: Crg score: A distribution-aware clinical metric for radiology report generation. arXiv preprint arXiv:2505.17167 (2025)
10. Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Durugol, O.F., Hou, B., Shit, S., et al.: Generalist foundation models from a multimodal dataset for 3d computed tomography. Nature Biomedical Engineering pp. 1–19 (2026)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hong, X., Jiang, C., Tian, K., Qi, B., Sun, Y., Ding, N., Zhou, B.: Depass: Unified feature attributing by simple decomposed forward pass. arXiv preprint arXiv:2510.18462 (2025)
13. Jiang, Q., Huo, J., Chen, X., Xiong, Y., Zeng, Z., Chen, Y., Ren, T., Yu, J., Zhang, L.: Detect anything via next point prediction. arXiv preprint arXiv:2510.12798 (2025)
14. 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. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
15. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
16. Luo, L., Tang, B., Chen, X., Han, R., Chen, T.: Vividmed: Vision language model with versatile visual grounding for medicine. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 1800–1821 (2025)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
18. Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., Van Den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical image analysis **42**, 1–13 (2017)
19. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
20. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. Nature medicine **29**(8), 1930–1940 (2023)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
22. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
23. Wu, J., Gan, W., Chen, Z., Wan, S., Yu, P.S.: Multimodal large language models: A survey. In: 2023 IEEE International Conference on Big Data (BigData). pp. 2247–2256. IEEE (2023)

24. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024)
25. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. arXiv preprint arXiv:2310.07704 (2023)
26. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
27. Zhang, M., Wu, X., Luo, H., Wang, F., Lv, Y.: Medground: Bridging the evidence gap in medical vision-language models with verified grounding data. arXiv preprint arXiv:2601.06847 (2026)
28. Zhang, X., Wu, C., Zhao, Z., Lei, J., Tian, W., Zhang, Y., Xie, W., Wang, Y.: Development of a large-scale grounded vision language dataset for chest ct analysis. Scientific Data **12**(1), 1636 (2025)
29. Zhao, Y., Huang, J., Hu, J., Wang, X., Mao, Y., Zhang, D., Jiang, Z., Wu, Z., Ai, B., Wang, A., et al.: Swift: a scalable lightweight infrastructure for fine-tuning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 29733–29735 (2025)