Lesion-Centric Structured Report Generation for 3D CT with Multimodal Large Language Models via Hierarchical Evidence Tokens

Xiuheng Zhao[1], Yiming Shi[1], Xun Zhu[1], Chen Wang[2], Fuze Cong[2], Jingzhi Yang[5], Ji Wu[1,3,4], Yonglan He[2*], and Miao Li[1*]

[1] Department of Electronic Engineering, Tsinghua University, Beijing, China
miao-li@tsinghua.edu.cn
[2] Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
heyonglan@pumch.cn
[3] College of AI, Tsinghua University, Beijing, China
[4] Beijing National Research Center for Information Science and Technology, Beijing, China
[5] School of Computer Science, National Pilot Software Engineering School, Beijing University of Posts and Telecommunications, Beijing, China

Abstract. Structured clinical report generation for 3D CT requires reasoning over long visual contexts while grounding predictions in lesion-relevant evidence. In practice, multimodal large language models (MLLM) often suffer from attention dispersion under long visual sequences, where critical cues are overwhelmed by background tokens, especially in low-data and imbalanced settings. We present a lesion-aware multimodal framework that addresses this challenge through two complementary mechanisms. First, we introduce a lesion-guided attention loss that regularizes cross-modal attention during training to encourage focus on lesion-relevant evidence without altering the inference interface. Second, we propose a compact set of Global Evidence Slots (GES) that summarize long visual sequences via cross-attention pooling and act as an explicit evidence bottleneck injected into the MLLM input, facilitating task-dependent routing of visual evidence. Together, these components form a hierarchical evidence token interface that combines fine-grained image tokens with compact evidence slots for robust lesion-centric reasoning. We evaluate our method on an ovarian tumor CT cohort with 12 structured attributes and validate its generalization on LIDC-IDRI. Experiments demonstrate consistent improvements over strong baselines.

Keywords: 3D CT · Structured Report Generation · Multimodal Large Language Models · Lesion-aware Modeling · Hierarchical Evidence Aggregation

* Corresponding authors: Miao Li and Yonglan He

1 Introduction

Accurate and standardized radiology reporting is essential for cancer diagnosis, treatment planning, and longitudinal management, and structured reporting has been advocated to encode lesion characteristics into predefined, clinically actionable attributes [8]. However, producing such reports for 3D CT is time-consuming and cognitively demanding: diagnostically relevant evidence is localized to a lesion, while clinical decisions require integrating multiple attribute-level judgments across volumetric scans. This setting calls for automatic systems that can reason over long 3D visual contexts in a lesion-centric manner.

Recent multimodal large language models (MLLMs) bridge pretrained LLMs with visual encoders and demonstrate strong vision–language reasoning [1, 13, 14], with further extensions to medical visual question answering and report generation [16, 12, 20]. However, most existing MLLMs are designed for 2D images or short clips. Directly unfolding a CT volume into multi-slice inputs yields extremely long visual token sequences, where subtle but decisive lesion cues are easily overwhelmed by irrelevant tokens, leading to an *attention dispersion* effect under long visual sequences [5, 22, 7, 10]. Moreover, although lesion masks or region annotations are often available [17, 23, 9], existing pipelines typically treat them only as input-level hints rather than constraints that explicitly shape cross-modal reasoning, making attention prone to drifting toward statistically dominant but diagnostically irrelevant regions. This issue is further exacerbated under limited data and severe class imbalance, where models tend to exhibit shortcut learning and unstable attribute prediction.

In this work, we target structured clinical attribute prediction for 3D CT with explicit lesion priors. Concretely, given a CT volume and a lesion mask, the model answers a structured template of 12 ovarian tumor attributes, and we further evaluate our method on LIDC-IDRI [2] as an external validation benchmark to assess cross-dataset generalization and robustness. This problem setting exposes three intertwined challenges: (i) attention dispersion under long visual sequences, (ii) ineffective utilization of lesion priors in cross-modal reasoning, and (iii) unstable multi-task learning in low-data and imbalanced regimes.

To address these challenges, we use a lightweight 3D-to-MLLM visual interface that represents a volumetric CT scan as a fixed-length set of image tokens, a design also adopted in recent medical LVLM pipelines such as MedM-VL [19]. On top of this interface, we introduce two complementary mechanisms for robust lesion-centric multimodal reasoning. First, we propose a *lesion-guided attention loss*, a training-time regularization term that constrains cross-modal attention over image tokens toward lesion-relevant evidence, mitigating attention dispersion under long visual sequences and suppressing spurious background correlations, in the spirit of attention-based visual explanation and region-sensitivity modeling [18]. Second, we introduce a compact set of *Global Evidence Slots (GES)*, which summarize long visual sequences via cross-attention pooling with learnable queries and are injected into the MLLM input as an explicit evidence bottleneck, conceptually related to latent-set and iterative attention representations [11, 15, 10, 6]. Together, fine-grained image tokens and GES form a *hier-*

archical evidence token interface that supports efficient and lesion-centric multimodal reasoning. Experiments on ovarian CT and LIDC-IDRI demonstrate consistent gains across data scales and model backbones.

The contributions of this work are threefold: (1) We develop a lesion-aware multimodal framework for structured reporting for 3D CT based on a **hierarchical evidence token interface**, which combines fine-grained image tokens with compact evidence slots to support lesion-centric reasoning under long visual contexts. (2) We propose a **lesion-guided attention loss** that regularizes cross-modal attention during training, encouraging the model to focus on lesion-relevant evidence and stabilizing optimization in low-data and severely imbalanced regimes. (3) We introduce **GES** obtained via cross-attention pooling and injected into the MLLM input as an explicit evidence bottleneck, facilitating task-dependent routing of visual evidence and alleviating attention dispersion under long visual sequences.

2 Method

2.1 Problem Formulation and Overall Pipeline

Given a 3D CT volume $\mathbf{V} \in \mathbb{R}^{H\times W\times D}$ and its corresponding lesion mask $\mathbf{M}$, our goal is to predict a structured set of discrete clinical attributes $\{y_k\}_{k=1}^{K}$ (with K=12 for the ovarian cohort). We formulate the task in an instruction-following manner, where each attribute corresponds to a templated query and the MLLM generates a categorical answer under a unified prompting protocol.

To enable efficient multimodal reasoning without relying on heavy 3D encoders, we use a lightweight 3D-to-MLLM visual interface that represents a volumetric CT scan as a fixed-length set of image tokens. Concretely, slices are selected using the lesion mask $\mathbf{M}$ (MaskZ). For explicit marking, we compute a 2D minimum enclosing circle from the slice-wise lesion mask (with a small dilation margin) and overlay it as a white circular contour on the image, providing a coarse localization cue without modifying the backbone or input channels. The slices are then encoded by a shared 2D backbone and aggregated into a fixed number of image tokens via attention pooling. These tokens are projected into the language embedding space and inserted into the MLLM input as visual context.

On top of this interface, we introduce two mechanisms to address attention dispersion under long visual sequences and to enable robust lesion-centric reasoning. First, we propose a *lesion-guided attention loss* that regularizes cross-modal attention during training, encouraging the model to focus on lesion-relevant evidence while suppressing spurious background correlations. Second, we introduce a compact set of GES that summarize long visual sequences via cross-attention pooling and are injected into the MLLM input as an explicit evidence bottleneck, facilitating task-dependent routing of visual evidence. Together, fine-grained image tokens and GES constitute a *hierarchical evidence token interface* that balances detailed visual perception with compact evidence aggregation for multimodal reasoning.

An overview of the proposed framework is illustrated in Fig. 1. The following sections detail the lesion-guided attention loss and the GES.

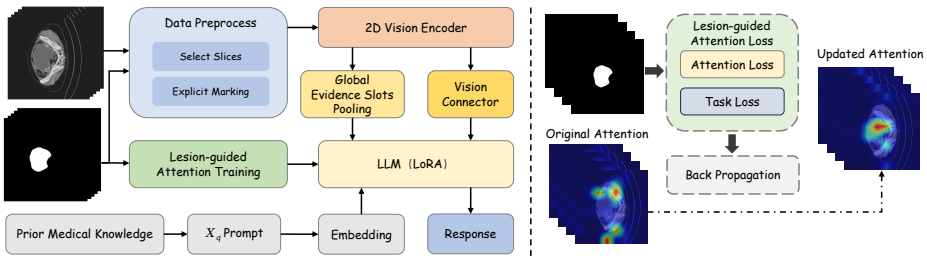

Fig. 1. The overall framework of the proposed method.

2.2 Lesion-Guided Attention Loss (Training-Time Regularization)

Mask-guided slice selection and marking bias the *input* toward lesion evidence but do not guarantee lesion-focused *reasoning*. We therefore introduce a *lesion-guided attention loss* as a training-time regularizer that constrains cross-modal attention from text to *image tokens*.

Let $\mathbf{A}^{(\ell,h)} \in \mathbb{R}^{T\times q}$ denote the cross-attention weights from T text tokens to q image tokens. We derive a binary indicator $\mathbf{m} \in \{0,1\}^q$ by mapping the lesion mask $\mathbf{M}$ to the image-token grid through the pooling alignment. We encourage attention to concentrate on lesion-relevant tokens by minimizing

$$\mathcal{L}_{\text{attn}} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{q}\mathbf{m}_i \log\left(\bar{\mathbf{A}}_{t,i} + \epsilon\right), \tag{1}$$

where $\bar{\mathbf{A}}$ denotes the aggregated cross-attention and ϵ is a small constant. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda\mathcal{L}_{\text{attn}}, \tag{2}$$

with λ controlling the strength of the regularization. This loss is applied during training and only to image tokens, leaving the inference interface unchanged.

2.3 Global Evidence Slots and Hierarchical Evidence Token Interface

Even with fixed-length image tokens, clinically critical cues may still be diluted when the model reasons over long visual contexts. We therefore introduce *GES* as a small set of learnable evidence slots that summarize the visual sequence and provide an explicit bottleneck for task-dependent information routing.

Cross-attention pooling. Let $\mathbf{Z} \in \mathbb{R}^{L \times d_{\text{enc}}}$ denote the concatenated patch tokens from all selected slices after 2D encoding. We introduce G learnable queries $\mathbf{Q}_g \in \mathbb{R}^{G \times d_{\text{enc}}}$ and obtain G evidence slots by cross-attention pooling:

$$\mathbf{H}_g = \mathrm{MHA}(\mathbf{Q}_g, \mathbf{Z}, \mathbf{Z}) \in \mathbb{R}^{G \times d_{\text{enc}}}, \qquad \mathbf{U}_g = \mathbf{H}_g \mathbf{W}_g \in \mathbb{R}^{G \times d_{\text{llm}}}. \tag{3}$$

The resulting $\mathbf{U}_g$ are learned end-to-end from the downstream objective and act as compact, content-adaptive summaries of the long visual sequence.

Token injection. Let $\mathbf{U}_{\text{txt}}$ denote the text tokens and $\mathbf{U}_{\text{img}} \in \mathbb{R}^{q \times d_{\text{llm}}}$ the fixed-length image tokens obtained by attention pooling. We inject the evidence slots and form the multimodal input

$$\mathbf{X} = [\, \mathbf{U}_{\text{txt}} \,;\, \mathbf{U}_{\text{img}} \,;\, \mathbf{U}_g \,] \in \mathbb{R}^{(T+q+G) \times d_{\text{llm}}}, \tag{4}$$

where $G \ll q$. The sequence $\mathbf{X}$ is fed into the MLLM to predict structured attributes autoregressively:

$$p(\mathbf{y} \mid \mathbf{V}) = \prod_t p(y_t \mid \mathbf{X}, y_{<t}). \tag{5}$$

In this formulation, fine-grained image tokens and GES jointly constitute a *hierarchical evidence token interface*, where the evidence slots serve as a compact bottleneck that preserves detailed visual tokens while enabling task-dependent routing through summarized evidence.

3 Experiments

3.1 Datasets and Task Definition

Ovarian tumor cohort (contrast-enhanced CT). We curate a real-world contrast-enhanced pelvic CT cohort with 750 cases, each with a 3D lesion mask and a radiologist-annotated structured template of 12 lesion-centric attributes covering morphology, internal composition, and malignancy-related assessment. The attributes include boundary, shape, composition, cyst wall, locularity, septation, cystic density homogeneity, mural nodules, solid component, intralesional fat, intralesional hemorrhage, and an overall malignancy decision (with conditional N.A. labels for inapplicable cases). The label distribution is highly imbalanced, with dominant categories reaching $\sim$80% and rare ones below $\sim$5%.

LIDC-IDRI (nodule-level attributes). We build a nodule-level benchmark from LIDC-IDRI with reliable masks and annotations, using a structured template of 9 attributes (subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, texture, and malignancy). Continuous scores are discretized into three-level categories, and malignancy is mapped to benign, malignant, or indeterminate, providing a cross-dataset generalization test distinct from ovarian CT.

Problem formulation. For each volume, the model predicts a structured set of answers by treating each attribute as an instruction-following query. Performance is reported on the held-out test set(s) under identical inference prompts.

3.2 Implementation Details

All experiments are conducted on 2×A800 GPUs. We follow the standard fine-tuning recipe of the underlying backbone (e.g., M3D) and use a base learning rate of 2×10^{-5}. The batch size is 32 unless otherwise stated. We freeze the vision encoder and optimize the connector along with LoRA parameters of the LLM. For each setting, we select the checkpoint with the best validation MacroF1 within 11–15 epochs and report test performance using the same decoding configuration across methods. All experiments use patient-level splits to prevent information leakage across sets.

3.3 Evaluation Metric

We use macro F1-score as the primary metric due to severe class imbalance. Concretely, for each task t, we compute the macro F1 across its categories, and report the overall score as the unweighted mean across all tasks:

$$\mathrm{MacroF1}_{\mathrm{overall}} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{MacroF1}(t) \tag{6}$$

where $T = 12$ for the ovarian cohort. This directly reflects robust multi-task performance rather than being dominated by frequent labels.

3.4 Baselines and Compared Methods

Evaluation protocol. All compared methods are evaluated under the same lesion-prior settings to ensure fair comparison: (i) no prior, (ii) lesion-centered slice selection (MaskZ), and (iii) MaskZ with explicit marking. This allows us to disentangle the effect of lesion priors from backbone and modeling choices.

Compared backbones. We consider four representative baselines: (i) **MedM-VL** [19], a lightweight 3D-to-MLLM pipeline using a 2D encoder with attention pooling; (ii) **M3D** [3], a 3D ViT encoder that models volumetric context natively; (iii) **cross-backbone MLLMs** including InternVL3.5 [21] and Qwen2.5-VL [4]; and (iv) a **3D ViT multi-label classifier** with one-vs-rest heads, serving as a purely discriminative baseline without language generation.

Ours. Starting from the strongest lesion-aware MedM-VL setting (MaskZ + explicit marking), we add (i) lesion-guided attention applied to cross-modal attention over image tokens and (ii) GES obtained by cross-attention pooling and injected into the LLM input. We report each component and their combination to verify complementarity.

3.5 Main Results

Tab. 1–3 and Fig. 2 provide three lines of evidence: (i) the necessity of lesion priors for establishing strong baselines, (ii) the distinct and complementary roles

Table 1. Baseline comparison on the ovarian cohort. MacroF1 is averaged over 12 tasks. Mark denotes explicit lesion marking (circle overlay).

Method	MaskZ	Mark	MacroF1
3D ViT multi-label classifier	–	–	0.4304
InternVL3.5 (finetune)	–	–	0.4066
InternVL3.5 (finetune)	✓	–	0.4121
InternVL3.5 (finetune)	✓	✓	0.4182
Qwen2.5-VL (finetune)	–	–	0.3615
Qwen2.5-VL (finetune)	✓	–	0.3734
Qwen2.5-VL (finetune)	✓	✓	0.3751
M3D (3D ViT encoder)	–	–	0.3399
M3D (3D ViT encoder)	✓	–	0.3981
M3D (3D ViT encoder)	✓	✓	0.4064
MedM-VL	–	–	0.4674
MedM-VL	✓	–	0.4867
MedM-VL	✓	✓	0.5113
Ours	✓	✓	**0.6529**

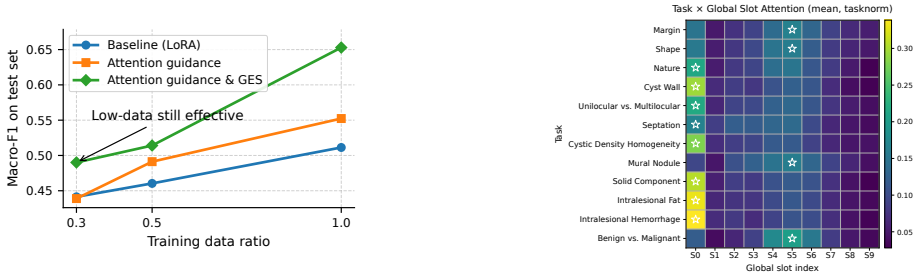

Fig. 2. (a) Data efficiency under reduced training fractions. (b) Task-by-slot attention over GES, showing task-dependent routing.

of the lesion-guided attention loss and GES, and (iii) the transferability and robustness of the proposed mechanisms across datasets and data regimes.

First, Tab. 1 establishes a strong lesion-aware baseline. Across backbones, MaskZ consistently improves performance, and explicit marking yields further gains, identifying MaskZ + explicit marking as a unified reference setting. Among all compared backbones, MedM-VL achieves the best overall performance under this setting; therefore, we adopt it as the primary backbone for subsequent ablations and analyses.

Second, Tab. 2 directly evaluates the two proposed components on top of this baseline. The lesion-guided attention loss provides consistent yet moderate improvements across backbones, supporting its role as a training-time regularizer that stabilizes optimization under attention dispersion and severe class imbalance. In contrast, GES yields substantially larger gains, highlighting its effect as

Table 2. Component analysis on the ovarian cohort under explicit marking.

Variant	Attn. loss	GES	MacroF1
Qwen2.5-VL	–	–	0.3751
Qwen2.5-VL	✓	–	0.4000
M3D	–	–	0.4064
M3D	✓	–	0.4200
M3D	–	✓	0.4411
M3D	✓	✓	0.4891
MedM-VL	–	–	0.5113
MedM-VL	✓	–	0.5523
MedM-VL	–	✓	0.6240
MedM-VL(zero)	✓	✓	0.4264
MedM-VL(shuffle)	✓	✓	0.4727
MedM-VL	✓	✓	**0.6529**

Table 3. Generalization on LIDC-IDRI with nodule-specific explicit marking.

Variant	Attn. loss	GES	MacroF1
MedM-VL	–	–	0.3903
+ Attention guidance	✓	–	0.4079
+ GES	–	✓	0.4201
+ both	✓	✓	**0.4292**

a representation-level evidence bottleneck that improves robustness to long visual contexts. Importantly, **combining both achieves the best performance across all settings**, indicating that the two mechanisms address different failure modes—optimization stability versus evidence aggregation—rather than providing redundant benefits.

Third, the same trend holds on LIDC-IDRI (Tab. 3), confirming that neither the attention regularization nor GES is dataset-specific, but instead captures transferable principles for lesion-centric reasoning for 3D CT. Beyond overall MacroF1, Fig. 2(a) shows that **the proposed method consistently outperforms the baseline under reduced training fractions**, consistent with the gains brought by both attention regularization and GES in Tab. 2, indicating improved sample efficiency in low-data regimes. Fig. 2(b) further reveals structured, task-dependent routing over the G=10 evidence slots, rather than reliance on a single dominant slot. Finally, inference-time interventions by zeroing or shuffling GES (Tab. 2, bottom) cause substantial performance drops, providing causal evidence that the gains stem from the semantic content encoded in these slots rather than from additional parameters or longer context.

4 Conclusion

We presented a lesion-aware multimodal framework for structured clinical attribute prediction for 3D CT built upon a *hierarchical evidence token interface*. The proposed approach introduces two complementary mechanisms: a *lesion-guided attention loss* that regularizes cross-modal attention during training to stabilize optimization under attention dispersion and severe class imbalance, and *GES* that serve as a compact evidence bottleneck for robust aggregation and task-dependent routing of visual information. Together, these components instantiate the hierarchical evidence token interface, combining fine-grained image tokens with compact evidence slots for lesion-centric reasoning. Extensive experiments on an ovarian tumor cohort and LIDC-IDRI demonstrate consistent improvements over strong lesion-aware baselines across data scales, backbone capacities, and evaluation settings. A limitation is the assumption of an available lesion prior; here, masks provide coarse localization for slice selection, marking, and attention regularization rather than precise boundary supervision. Robustness to imperfect or automatically generated masks remains future work.

Acknowledgements This work was supported 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), National High-Level Hospital Clinical Research Funding (grant no. 2025-PUMCH-A-023), and CAMS Innovation Fund for Medical Sciences (CIFMS) (grant no. 2024-I2M-C&T-B-032).

Disclosure of Interests

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

References

1. 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)
2. Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., Van Beeke, E.J., Yankelevitz, D., Biancardi, A.M., Bland, P.H., Brown, M.S., Engelmann, R.M., Laderach, G.E., Max, D., Pais, R.C., Qing, D., Roberts, R.Y., Smith, A.R., Starkey, A., Batra, P., Caligiuri, P., Farooqi, A., Gladish, G.W., Jude, C.M., Munden, R.F., Petkovska, I., Quint, L.E., Schwartz, L.H., Sundaram, B., Dodd, G.D., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Casteele, V., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. Medical Physics **38**(2), 915–931 (2011)

3. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020)
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
7. Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al.: Rethinking attention with performers. arXiv preprint arXiv:2009.14794 (2020)
8. Ganeshan, D., Duong, P.A.T., Probyn, L., Lenchik, L., McArthur, T.A., Retrouvey, M., Ghobadi, E.H., Desouches, S.L., Pastel, D., Francis, I.R.: Structured reporting in radiology. Academic radiology **25**(1), 66–73 (2018)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
10. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International conference on machine learning. pp. 4651–4664. PMLR (2021)
11. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: International conference on machine learning. pp. 3744–3753. PMLR (2019)
12. 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)
13. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
14. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
15. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Advances in neural information processing systems **33**, 11525–11538 (2020)
16. 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 (ML4H). pp. 353–367. PMLR (2023)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)

19. Shi, Y., Yang, S., Zhu, X., Wang, H., Fu, X., Li, M., Wu, J.: Medm-vl: What makes a good medical lvlm? In: International Workshop on Agentic AI for Medicine. pp. 290–299. Springer (2025)
20. Thawakar, O.C., Shaker, A.M., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.: Xraygpt: Chest radiographs summarization using large medical vision-language models. In: Proceedings of the 23rd workshop on biomedical natural language processing. pp. 440–448 (2024)
21. 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)
22. Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al.: Big bird: Transformers for longer sequences. Advances in neural information processing systems **33**, 17283–17297 (2020)
23. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: International workshop on deep learning in medical image analysis. pp. 3–11. Springer (2018)