Reformulating Chest X-ray Report Generation as Fine-Grained Visual Question Answering with Structured Clinical Report Templates

Ziqin Xiao[1,2⋆], Zehui Liao[1*], Zilin Lu[1], Shishuai Hu[1], Shaoyi Leng[3], Jingfeng Zhang[3], and Yong Xia[1,2](✉)

[1] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[2] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China
[3] Ningbo No.2 Hospital, Wenzhou Medical University, Ningbo 315000, China
yxia@nwpu.edu.cn

Abstract. Automated chest X-ray (CXR) report generation is essential for reducing radiologists' workload and enhancing diagnostic efficiency in pulmonary screening. While vision-language models (VLMs) have shown promise, most existing methods rely on holistic free-text supervision, which often provides highly concise, incomplete clinical signals and leads to factual hallucinations. To address this issue, we propose **TGST**, a **T**emplate-**G**uided **S**tructured **T**raining framework that reformulates report generation as a fine-grained, clinically-aligned visual question answering (VQA) task. Specifically, we first construct a tree-structured clinical template for pulmonary findings, decomposing complex reports into anatomical entities and their specific attributes. We then introduce a two-stage pipeline that maps original reports to the template and leverages a medical VLM to infer missing visual details, creating dense supervision signals. By supervising the model on these localized clinical queries, TGST enforces precise visual-textual alignment and penalizes reliance on linguistic priors. Extensive experiments on the IU-Xray dataset demonstrate that TGST significantly outperforms state-of-the-art methods. Notably, our framework achieves substantial gains in factual correctness (GREEN score) and node-level observation accuracy across multiple VLM backbones, highlighting its potential for reliable and systematic automated pulmonary screening.

Keywords: Chest X-ray report generation · Vision-language model · Structured template.

1 Introduction

Medical image interpretation and standardized report generation are essential for clinical diagnosis and patient management. Chest X-ray (CXR), the most

⋆ Z. Xiao and Z. Liao contributed equally. Corresponding author: Y. Xia.

widely performed imaging examination worldwide, serves as a primary tool for pulmonary disease screening due to its low cost and broad accessibility. However, the rapid growth in CXR volume has imposed heavy workload pressure on radiologists [13]. Manual interpretation and reporting are still labor-intensive and error-prone, and subtle abnormalities are frequently missed under time constraints [4].

To address this clinical challenge, automated radiology report generation has emerged as a high-priority research direction, aiming to extract clinically relevant findings from medical images and generate reliable, standardized textual reports, thereby reducing radiologists' workload and improving diagnostic workflow efficiency [26]. Automated CXR report generation has evolved from CNN–RNN frameworks [16,17] to Transformer-based models [10,22,24], reinforcement learning approaches for enhancing factual correctness [21], and more recent vision-language models (VLMs) [2,3,7,12,18,20,27], achieving notable improvements in fluency and reasoning. However, most existing methods rely on holistic long-text supervision and one-step generation. In contrast, radiologists follow a structured, step-by-step process when reviewing CXR images, systematically examining each tissue or organ in a spatial sequence before summarizing their findings in a concise free-form report. As such, the report generated by radiologists for a given CXR does not serve as a comprehensive supervision signal. Using such reports to train VLMs inevitably limits the model's ability to fully understand the image, thereby hindering the accuracy and completeness of report generation.

To overcome this issue, we propose TGST, a Template-Guided Structured Training framework for VLM-based CXR report generation. TGST reformulates holistic report generation as a clinically guided, fine-grained VQA task. First, we construct a radiology textbook-aligned, tree-structured clinical template for pulmonary findings, where the root node represents the lung, the intermediate nodes correspond to organs or lesions, and each leaf node represents a clinically actionable visual attribute, such as specific observable features in the chest X-ray that are critical for clinical decision-making. Second, we introduce a two-stage pipeline: the free-text reports are mapped into the structured template, after which the missing parts within the template are filled in using a powerful medical VLM to generate explicit supervision signals with clear semantic targets. This transformation replaces one-step holistic generation with multi-step, detail-oriented reasoning, enforcing localized visual evidence learning and precise visual-text alignment. Finally, the structured reports generated from the template are used to train medical VLM. We conduct extensive experiments on the IU-Xray dataset [11]. The results show that TGST significantly outperforms conventional long-text supervision paradigms and recent report generation methods, with particularly notable improvements in factual correctness.

The main contributions are three-fold: (1) We propose TGST, a new framework that reformulates CXR report generation as a fine-grained, clinically guided VQA task, introducing a two-stage pipeline for high-confidence supervision to reduce noise and hallucinations. (2) We develop a tree-structured template for CXR pulmonary findings, enabling systematic decomposition of free-text reports

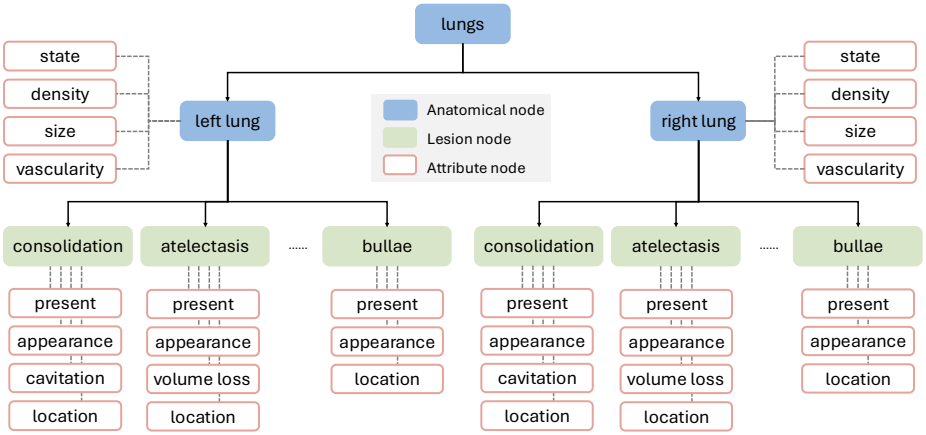

Fig. 1. Hierarchical clinical template for pulmonary CXR findings. The template consists of anatomical nodes, lesion nodes, and attribute nodes, forming a structured observation space for fine-grained VQA supervision.

into semantically clear, clinically relevant supervision signals. (3) Extensive experiments on IU-Xray demonstrate that TGST achieves state-of-the-art factual correctness, highlighting its potential for reliable report generation.

2 Method

The proposed TGST framework reformulates holistic CXR report generation as a template-guided, fine-grained VQA task to mimic the systematic reviewing process of radiologists in routine clinical practice. To support this reformulation, we first construct a clinical knowledge-driven tree-structured template that formalizes the structured observation space for pulmonary CXR findings, as shown in Fig. 1. Based on this template, as illustrated in Fig. 2, TGST introduces a two-stage structuring pipeline that transforms sparse free-text reports into dense, high-confidence supervision signals by systematically mapping existing findings and inferring missing attributes. Finally, the model learns to generate reports via template-guided VQA, enforcing precise visual-text alignment through localized clinical queries across the hierarchical nodes.

2.1 Clinical Knowledge-Driven Structured Report Template

To mimic the systematic diagnostic process of radiologists and ensure clinical validity, we construct a tree-structured clinical template $\mathcal{T}$ that formalizes the observation space for CXR pulmonary findings. As illustrated in Fig. 1, $\mathcal{T}$ is designed as a hierarchical backbone enriched with node-specific attributes, encompassing three types of components: (1) **Anatomical Nodes:** The root node

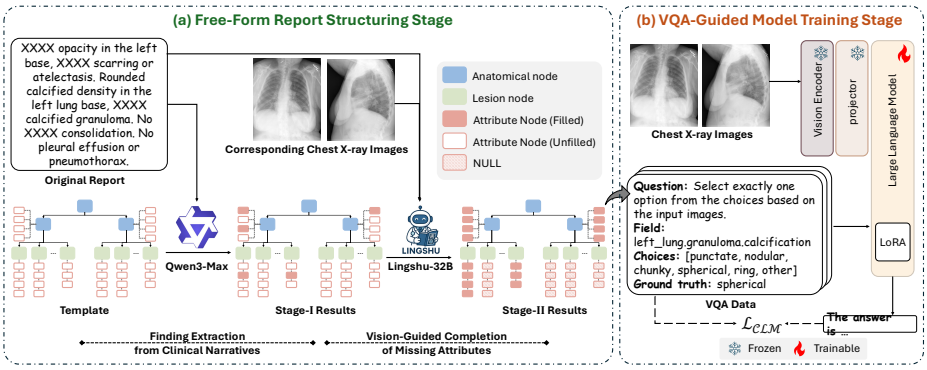

Fig. 2. Overview of the proposed TGST framework. We first construct a tree-structured clinical template for CXR pulmonary findings, then use a two-stage pipeline to structure free-form reports, finally fine-tune the VLM with the constructed VQA supervision.

(*Lungs*) branches into *Left Lung* and *Right Lung*, establishing the spatial reference for subsequent findings. (2) **Lesion Nodes:** Under each lung node, we define seven clinically prevalent parenchymal abnormalities (e.g., *consolidation, atelectasis, bullae, granuloma, mass, nodule,* and *scarring*). These nodes represent the categorical identification of pathological entities within the anatomical context. (3) **Attribute Nodes:** Each anatomical or lesion node is linked to multiple fine-grained attribute nodes. Crucially, these attributes are heterogeneous and tailored to the specific node. For instance, a *consolidation* node triggers attributes like *appearance, cavitation,* and *location*, while the *left lung* node is associated with *state, density, size,* and *vascularity.*

Formally, each attribute node corresponds to a clinically actionable question $q \in \mathcal{Q}$, defining a closed-set VQA space. To handle out-of-template expressions, an *Other* option is provided for each attribute. This structured design decomposes holistic reports into deterministic, localized semantic targets, providing precise visual-textual alignment for model training. For example, consider an original report stating: *"There is minimal left basilar opacity compatible with scarring or atelectasis."* During the filling stage, the visual evidence may support the interpretation of *atelectasis* rather than *scarring.* Accordingly, the *present* attribute node under the *Left Lung* $\rightarrow$ *Atelectasis* node is set to *present.* Subsequently, the model proceeds to infer fine-grained attribute node such as *volume loss.* If, after both the text-guided initialization stage and the image-refined reasoning stage, none of the predefined options (e.g., *hilum displacement, rib approximation*) can be selected with high confidence, the corresponding *volume loss* attribute is assigned to *Other*. This mechanism ensures that the structured template remains clinically faithful while preserving robustness to ambiguous or weak visual cues.

2.2 Two-Stage Pipeline for Free-Form Report Structuring

To transform sparse, holistic free-text reports into dense, structured supervision signals, we develop a two-stage pipeline. To prioritize pulmonary screening, we utilize Qwen3-Max [29] to extract lung-specific findings from the original free-form reports, effectively filtering out irrelevant clinical information.

Stage I: Finding Extraction from Clinical Narratives. In this stage, the lung-specific report is mapped onto the template $\mathcal{T}$ using Qwen3-Max with temperature set to 0.0. We impose a *strict-evidence constraint*: an attribute node is completed only if the report provides explicit and unambiguous evidence. Attributes omitted in the concise clinical narratives (e.g., laterality or specific morphology) remain unfilled to avoid erroneous assignments, with their inference deferred to the subsequent stage. Furthermore, *medical dependency rules* are incorporated to maintain logical coherence. For example, the identification of any lesion automatically triggers a state transition of the parent `lung.state` attribute node to *abnormal*.

Stage II: Vision-Guided Completion of Missing Attributes. To recover the visual details missing from the clinical narratives, we introduce a completion stage using Lingshu-32B [28], a domain-specific VLM, with temperature set to 0.0. Taking the CXR image and the partially completed template from Stage I as input, the VLM performs node-wise inference to populate the remaining unfilled attribute nodes. Specifically, each missing attribute node is reformulated as a visual query, enabling the model to supplement clinical details that were visually apparent but absent from the original textual narrative. To mitigate hallucination and redundancy, we introduce: (1) **Logical Pruning:** If a lesion (e.g., *atelectasis*) is confirmed absent, all its sub-attribute nodes are deterministically set to *null*, bypassing redundant VLM inference. (2) **Uncertainty Handling:** Attributes that remain clinically ambiguous after multimodal reasoning, or those that fall outside predefined categories, are assigned to the *Other* class. Finally, only node with valid assignments (excluding *null*) along with the CXR image is converted into image-question-answer triplet (x_i, q_i, a_i). All triplets together form our training data $\mathcal{D} = \{(x_i, q_i, a_i)\}_{i=1}^{N}$ for the next step, where N is the number of training samples. This filtering ensures that the VQA-based fine-tuning is guided by high-density, low-noise supervision signals, effectively reducing model hallucinations.

2.3 Learning to Report via Template-guided VQA

After converting the lung-specific reports into corresponding fine-grained VQA samples, we leverage these samples $\mathcal{D} = \{(x_i, q_i, a_i)\}_{i=1}^{N}$ to fine-tune a VLM for CXR report generation to enhance pulmonary screening capabilities. Here, $x_i = \{x_i^f, x_i^l\}$ denotes the paired frontal and lateral CXR images as multi-view visual context, q_i is the node-specific clinical question serving as the textual

instruction, and a_i is the corresponding high-confidence node value as the target output.

During training, we optimize the model parameters using the standard causal language modeling (CLM) loss. For sample (x_i, q_i, a_i), its loss $\mathcal{L}_i$ is calculated as follows:

$$\mathcal{L}_i = -\frac{1}{T}\sum_{t=1}^{T}\log p_\theta\left(y_t|y_{<t}, x_i, q_i\right), \tag{1}$$

where y_t is the t-th token of the target answer a_i, T is the total length of the target sequence, and θ denotes the learnable parameters of the VLM.

Compared with conventional long-form report supervision, this VQA-based training paradigm provides more explicit, localized training signals, forcing the VLM to attend to region visual evidence rather than relying primarily on linguistic priors. This fine-tuning strategy explicitly enforces fine-grained visual-text alignment, ensuring high consistency between the generated report and the underlying imaging evidence, and substantially improving the factual accuracy and descriptive granularity of the final report generation.

3 Experiments

3.1 Experimental Setup

Dataset. We conduct experiments on the IU-Xray dataset, a widely used benchmark for CXR report generation. The official release contains 7,470 CXR images paired with 3,955 radiology reports annotated by board-certified radiologists. To ensure view consistency, we select cases containing exactly one frontal view and one lateral view, resulting in 5,560 images and 2,780 reports. The subset is split into training, validation, and test sets with a ratio of 7:1:2. All methods are trained and evaluated on the lung-specific reports extracted using Qwen3-Max.
Evaluation Metrics. We evaluate the performance of the VLM fine-tuned via TGST based on two distinct output formats: (1) *Free-form Report Generation:* We assess the model's ability to generate descriptive reports by measuring the factual consistency between the generated text and ground-truth using GREEN [23]. (2) *Structured Template Completion:* We evaluate the model's fine-grained observation by treating each node in $\mathcal{T}$ as a VQA task. The accuracy of the filled attributes is measured via *Accuracy (Acc)*, *Precision (Pre)*, *Recall(Rec)*, and *F1-score*, aggregated through both macro and weighted averaging.
Implementation Details. For model fine-tuning, we evaluate three multimodal LLMs, including InternVL2.5-4B [8], Qwen2.5-VL-3B-Instruct [1], and HuatuoGPT-Vision-7B-Qwen2.5VL [5]. We fine-tune the language decoder using Low-Rank Adaptation, abbreviated as LoRA [14]. The rank is set to 8, and α is set to 16. Training is implemented using the TRL SFT framework based on HuggingFace Transformers. All models are trained for 10 epochs using the AdamW optimizer, with a per-device batch size of 16, gradient accumulation of 2 steps, and a learning rate of 2×10^{-5}. All experiments are conducted on two NVIDIA RTX 4090 GPUs.

Table 1. Performance of three VLMs fine-tuned using different supervision formats in free-form report generation (FRG) and structured template completion (STC). The best results are in **bold**.

Model	Training	Data	FRG	STC			
			GREEN	Acc.	F1 (M/W)	Prec. (M/W)	Rec. (M/W)
InternVL	Zero-shot	/	0.521	0.431	0.191/0.217	0.200/0.242	0.318/0.356
	LoRA	Report	0.502	0.607	0.247/0.311	0.300/0.338	0.340/0.387
		Ours	**0.756**	**0.817**	**0.339/0.416**	**0.401/0.497**	**0.370/0.410**
Qwen	Zero-shot	/	0.567	0.131	0.127/0.089	0.135/0.154	0.302/0.331
	LoRA	Report	0.629	0.250	0.184/0.154	0.218/0.309	0.325/0.352
		Ours	**0.828**	**0.830**	**0.329/0.400**	**0.351/0.468**	**0.356/0.399**
HuatuoGPT	Zero-shot	/	0.727	0.462	0.165/0.220	0.195/0.304	0.291/0.347
	LoRA	Report	0.774	0.302	0.164/0.151	0.167/0.209	0.307/0.328
		Ours	**0.815**	**0.831**	**0.306/0.341**	**0.336/0.386**	**0.342/0.354**

3.2 Comparison Results of the Fine-tuning Paradigm

Table 1 provides a comprehensive comparison of model performance in free-form report generation (FRG) and structured template completion (STC) across three distinct supervision paradigms and three backbone VLMs: (1) *Zero Shot*: the pre-trained VLMs are evaluated directly without any task-specific fine-tuning using the IU-Xray dataset, with their performance serving as the lower bound, (2) *Report-level Fine-tuning*: VLMs are fine-tuned using LoRA with conventional holistic free-text radiology reports; (3) *Ours*: VLMs are fine-tuned using the same LoRA strategy but under our proposed TGST supervision, reformulating the task into fine-grained VQA. The results demonstrate that our TGST significantly enhances VLMs' report generation capabilities, indicating that it offers a more comprehensive supervision signal. Across all three backbones, our TGST consistently improves factual correctness measured by GREEN. For example, by replacing traditional report supervision with our structured VQA data, GREEN increases from 0.629 to 0.828 on Qwen2.5-VL-3B-Instruct, resulting in an improvement of nearly 20%. Similar trends are observed on InternVL and HuatuoGPT, suggesting that the benefits of structured supervision generalize across architectures.

Beyond FRG evaluation, our TGST also yields notable improvements in STC, including Accuracy and Macro/Weighted F1. This demonstrates that decomposing long-form reports into semantically explicit supervision units provides clearer optimization signals and facilitates more effective learning of structured clinical attributes. In contrast, report-level supervision requires the model to implicitly capture multiple anatomical findings from a single holistic target, which may dilute supervision signals. By explicitly modeling each clinical attribute as an independent learning objective, structured supervision promotes more stable optimization and leads to consistent gains in both report-level and fine-grained evaluation metrics.

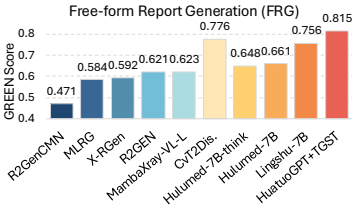

Fig. 3. FRG performance of our TGST+HuatuoGPT and other competing methods on the IU-Xray test set.

Table 2. Structured template completion performance of our proposed TGST using HuatuoGPT and other medical VLMs (Macro metrics) on the IU-Xray test set. The best results are in **bold**.

Model	Acc.	M-F1	M-Prec.	M-Rec.
Lingshu-7B	0.707	0.241	0.252	0.343
Hulumed-7B[15]	0.720	0.291	0.321	**0.350**
Hulumed-7B-think	0.713	0.273	0.304	0.335
Huatuo+TGST	**0.831**	**0.306**	**0.336**	0.342

3.3 Comparison with Other Report Generation Methods

Table 2 presents a comprehensive comparison between HuatuoGPT + TGST and several representative report generation approaches, including conventional encoder–decoder architectures [6,19,25] as well as recent medical VLMs. The GREEN scores for free-form report generation are further visualized in Fig. 3.

The results demonstrate that TGST substantially improves report-level factual correctness. Among traditional methods, CvT2Dis. [22] achieves the best FRG performance with a GREEN score of 0.776, outperforming earlier methods such as R2GEN [10] and R2GenCMN [9]. However, HuatuoGPT + TGST further increases the GREEN score to 0.815, establishing new state-of-the-art performance under the same evaluation protocol.

Beyond FRG evaluation, TGST also achieves superior performance in structured template completion. HuatuoGPT + TGST achieves the highest STC Accuracy of 0.831 and competitive Macro F1 scores, compared with recent medical VLMs such as Lingshu-7B and Hulumed-7B. This indicates that structured supervision not only improves holistic report generation but also enhances fine-grained attribute reasoning. In contrast, conventional report generation methods rely primarily on sequence-level optimization, which does not explicitly enforce alignment between localized visual evidence and structured clinical attributes. By modeling each attribute node as an independent supervision signal, TGST provides more precise optimization targets, leading to consistent gains across both report-level and fine-grained evaluation metrics.

4 Conclusion

In this paper, we present TGST, a new Template-Guided Structured Training framework that reformulates chest X-ray report generation as a fine-grained, clinically-aligned VQA task. By introducing a tree-structured clinical template and a two-stage supervision construction pipeline, TGST effectively addresses the limitations of sparse, holistic long-text supervision. This approach enables the model to bridge the gap between abstract clinical narratives and localized

visual evidence, mimicking the systematic diagnostic reasoning of expert radiologists. Extensive experiments on the IU-Xray dataset demonstrate that TGST significantly outperforms existing methods across multiple VLMs, achieving superior performance in both free-form report generation and structured template completion. We believe that this structured training paradigm provides a scalable and reliable pathway for enhancing automated pulmonary screening, ultimately contributing to more accurate diagnostic workflows in clinical practice.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 92470101, in part by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang, China, under Grant 2025C01201(SD2), in part by the Project of National Key Clinical Specialty (Department of Medical Imaging, No. 2024017), in part by the Ningbo Key Laboratory of Digital Imaging and Medical-Engineering Interdisciplinarity (No. 2024031). Zilin Lu was supported by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grant CX2026XXX.

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

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. 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)
3. Bu, S., Li, T., Yang, Y., Dai, Z.: Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14194–14204 (2024)
4. Carter, A.J., Davis, K.A., Evans, L.V., Cone, D.C.: Information loss in emergency medical services handover of trauma patients. Prehospital Emergency Care **13**(3), 280–285 (2009)
5. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Cai, Z., Ji, K., Wan, X., et al.: Towards injecting medical visual knowledge into multimodal llms at scale. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 7346–7370 (2024)
6. Chen, Q., Xie, Y., Wu, B., Chen, X., Ang, J., To, M.S., Chang, X., Wu, Q.: Act like a radiologist: radiology report generation across anatomical regions. In: Proceedings of the Asian Conference on Computer Vision. pp. 1–17 (2024)
7. Chen, W., Shen, L., Lin, J., Luo, J., Li, X., Yuan, Y.: Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9494–9509 (2024)
8. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal

models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
9. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 5904–5914 (2021)
10. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. pp. 1439–1449 (2020)
11. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
12. Gu, T., Yang, K., An, X., Feng, Z., Liu, D., Cai, W.: Orid: Organ-regional information driven framework for radiology report generation. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 378–387 (2025)
13. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J.: Artificial intelligence in radiology. Nature Reviews Cancer **18**(8), 500–510 (2018)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations (2022)
15. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025)
16. Jing, B., Wang, Z., Xing, E.: Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6570–6580 (2019)
17. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers). pp. 2577–2586 (2018)
18. Liu, C., Tian, Y., Chen, W., Song, Y., Zhang, Y.: Bootstrapping large language models for radiology report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18635–18643 (2024)
19. Liu, K., Ma, Z., Kang, X., Li, Y., Xie, K., Jiao, Z., Miao, Q.: Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10348–10359 (2025)
20. Liu, T., Wang, J., Hu, Y., Li, M., Yi, J., Chang, X., Gao, J., Yin, B.: Hc-llm: Historical-constrained large language models for radiology report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5595–5603 (2025)
21. Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5288–5304 (2021)
22. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial Intelligence in Medicine **144**, 102633 (2023)
23. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Md, A.E.M., Moseley, M., Langlotz, C., Chaudhari, A.S., et al.: Green: Generative radiology report evaluation and error notation. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 374–390 (2024)

24. 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)
25. Wang, X., Wang, F., Li, Y., Ma, Q., Wang, S., Jiang, B., Tang, J.: Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5123–5133 (2025)
26. Wang, X., Figueredo, G., Li, R., Zhang, W.E., Chen, W., Chen, X.: A survey of deep-learning-based radiology report generation using multimodal inputs. Medical Image Analysis **103**, 103627 (2025)
27. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. Meta-Radiology **1**(3), 100033 (2023)
28. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
29. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)