RadHiera: Semantic Hierarchical Reinforcement Learning for Medical Report Generation

Bodong Du[1], Honglong Yang[1,2], and Xiaomeng Li[1,2⋆]

[1]The Hong Kong University of Science and Technology, Hong Kong SAR, China
[2]Shenzhen Loop Area Institute, Shenzhen, China
eexmli@ust.hk

Abstract. Vision-language models have shown promising results in radiology report generation. However, most existing methods generate reports as flat text and do not explicitly model the semantic dependency between the *Findings* and *Impression* sections, which can lead to inconsistencies between clinical observations and diagnostic conclusions. In this paper, we propose **RadHiera**, a semantic hierarchical reinforcement learning framework for radiology report generation. **RadHiera** follows the semantic organization of radiology reports by first optimizing overall report quality, then improving the diagnostic accuracy of the *Impression* section, and finally enforcing consistency between *Findings* and *Impression* so that diagnostic conclusions are supported by clinical evidence. Specifically, we begin with a base reward that combines linguistic quality and medical factuality to provide supervision on the whole report. On this basis, we introduce a severity-aware reward for the *Impression* section that places greater emphasis on errors involving clinically critical conditions, thereby reducing both missed diagnoses and overstatement. We further enforce cross-section consistency using Expert Model-derived label sets, with subset constraints and hallucination penalties to ensure that impressions remain faithful to the findings. Experiments on three public chest X-ray benchmarks show that **RadHiera** consistently improves diagnostic accuracy and inter-section consistency over state-of-the-art methods, while also demonstrating good adaptability to ultrasound report generation. The source code is available at https://github.com/xmed-lab/RadHiera.

Keywords: Medical Report Generation · Reinforcement Learning · Semantic Hierarchy.

1 Introduction

Radiology reports adhere to a hierarchical structure where Findings describe observations and Impression synthesizes conclusions, ensuring clinical reliability [15, 6]. However, existing MRG systems underexplore this dependency. Early frameworks [10, 12, 17, 4, 8, 18, 21] neglect the Impression section due to its higher

⋆ Corresponding author

semantic complexity and optimization difficulty. In recent years, MLLMs [19, 16, 26] generate whole reports including the Findings and Impression sections but fail to explicitly model the inter-section hierarchy. Concurrent RL-based work such as UniRG [13] improves report-level optimization, but does not specifically target this Findings–Impression dependency. As a result, generated reports may exhibit diagnostic inaccuracies or inconsistencies, such as missing critical pathologies or introducing unsupported conclusions (see the "MedVersa" results in Figure 2).

To address this structural mismatch and improve clinical reliability, we propose **RadHiera**, a semantic hierarchical reinforcement learning framework that directly optimizes policy behavior toward clinically structured reporting. Central to this approach is a multi-level reward mechanism that explicitly models the semantic dependency between Findings and Impression sections. First, we apply a base reward that integrates linguistic fluency and medical factual correctness, ensuring coherent and clinically valid report generation. Building on this, we introduce a **severity-aware Impression reward** for the Impression section. It compares the generated Impression with the reference and assigns higher weights to clinically critical conditions (e.g., pneumothorax, pulmonary edema). This reward penalizes omissions of high-risk findings and discourages overstatements, thus prioritizing critical diagnoses and improving clinical fidelity. Finally, we introduce a **cross-sectional consistency reward** between the *Findings* and *Impression* sections, using CheXbert-derived label sets. The reward enforces consistency by applying subset constraints and hallucination penalties, reducing diagnostic contradictions and promoting structured alignment across sections.

Extensive experiments across three public chest X-ray benchmarks [5, 11, 24] demonstrate that RadHiera consistently enhances diagnostic precision and inter-section coherence relative to state-of-the-art baselines. Furthermore, the framework exhibits remarkable cross-modal generalizability, extending its performance gains to the domain of ultrasound report generation in an in-house CarotidUS-MRG dataset.

To summarize, this work makes the following contributions:

- We propose **RadHiera**, a semantic hierarchical reinforcement learning framework that incorporates clinical constraints into the policy optimization process, explicitly modeling the hierarchical structure of MRG.
- We design a **multi-level reward mechanism** composed of a base reward, a severity-aware Impression reward for diagnostic prioritization, and a cross-sectional consistency reward that enforces alignment between report sections.
- We conduct comprehensive experiments on diverse MRG datasets, demonstrating that our method achieves state-of-the-art performance on MRG both quantitatively and qualitatively.

2 Method

In this section, we first formulate the medical report generation task and introduce the reinforcement learning algorithm we build upon. Then we present

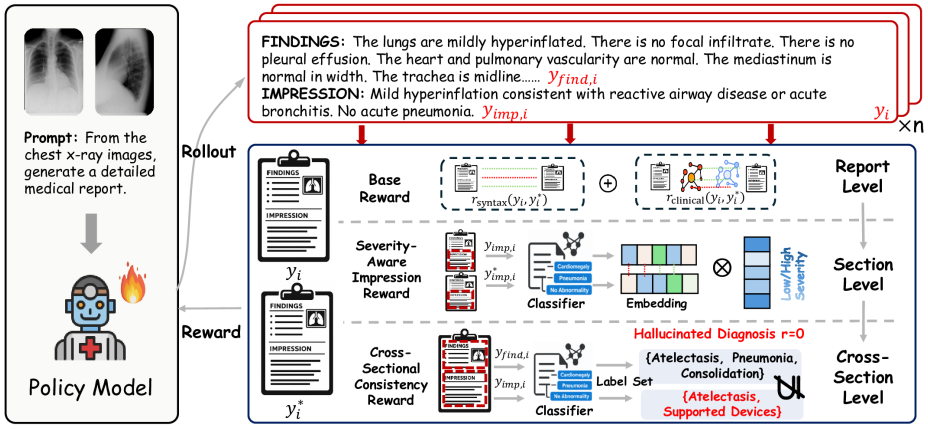

Fig. 1. Overall architecture of **RadHiera**. The policy model generates n rollouts from queries through GRPO optimization with the multi-level rewards: (1) base reward for policy optimization stability at the whole-report level, (2) Severity-Aware Impression Reward that leverages expert model embeddings to assess diagnostic severity levels, and (3) Cross-Sectional Consistency Reward that ensures alignment between Findings and Impression sections by detecting hallucinated diagnoses. This hierarchical reward enables generation of clinically effective and coherent radiology reports.

the overall design of **RadHiera**, a semantic hierarchical reinforcement learning framework as shown in Figure 1.

2.1 Problem Formulation and Preliminaries

Medical Report Generation (MRG) aims to generate structured and clinically coherent reports from medical images. A typical radiology report comprises two key sections: the *Findings* section, which describes observable visual cues, and the *Impression* section, which summarizes diagnostic conclusions.

Given an input medical image x and its reference report $y^* = \{y_1^*, \ldots, y_T^*\}$, a vision-language model π_θ models the conditional distribution of tokens:

$$p(y_t \mid x, y_{<t}),$$

where y_t denotes the t-th generated token. Standard supervised fine-tuning (SFT) maximizes the likelihood of the reference report:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log p(y_t^* \mid x, y_{<t}^*),$$

While SFT ensures fluency and domain alignment, it fails to explicitly enforce clinical consistency between Findings and Impression or optimize clinical metrics, necessitating structured reinforcement learning.

To align model behavior with clinically grounded signals, we adopt Group Relative Policy Optimization (GRPO), a reinforcement learning algorithm that evaluates groups of candidate completions without a value network. For input x, the policy π_{old} samples G candidate reports $\{s_i\}_{i=1}^{G}$, each receiving a scalar reward r_i from a reward function $R(s_i, x)$.

Within each group, rewards are normalized to compute relative advantage:

$$A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^{G})}{\text{std}(\{r_j\}_{j=1}^{G}) + \epsilon}, \tag{1}$$

where ϵ is a constant. For each token $s_{i,t}$ in candidate s_i, the policy ratio is:

$$r_{i,t}(\theta) = \frac{\pi_\theta(s_{i,t} \mid x, s_{i,<t})}{\pi_{\text{old}}(s_{i,t} \mid x, s_{i,<t})}. \tag{2}$$

The GRPO objective, incorporating clipping and KL regularization against a reference model π_{ref}, is:

$$\begin{aligned}\mathcal{J}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|s_i|}\sum_{t=1}^{|s_i|}\min\Big(r_{i,t}(\theta)A_i,\ \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)A_i\Big) \\ -\beta D_{KL}\big(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)\big)\Bigg],\end{aligned} \tag{3}$$

where β weights the KL penalty. This formulation increases the probability of above-average reports while preserving pretrained behavior. To optimize policy behavior toward clinically structured reporting, we design a reward across different report semantic levels:

$$r(y, y^*) = r_{\text{base}}(y, y^*) + r_{\text{imp}}(y, y^*) + r_{\text{consist}}(y). \tag{4}$$

Each reward targets a specific semantic granularity, detailed as follows.

2.2 Report-Level Base Reward

The base reward maintains general language quality and medical validity during policy optimization, while the hierarchical rewards reshape model behavior toward workflow-aligned reporting. The base reward is defined as

$$r_{\text{base}}(y, y^*) = r_{\text{syntax}}(y, y^*) + r_{\text{clinical}}(y, y^*). \tag{5}$$

r_{syntax} evaluates surface-level fluency and semantic similarity and r_{clinical} measures factual correctness by extracting clinical entities from y and comparing them with y^*. This base reward ensures that policy updates preserve medical validity and prevents significant deviations from the initial policy model, avoiding drift in language style and medical terminology. While the base reward maintains general quality, it does not explicitly encode the report structure. We therefore introduce two additional semantic-level rewards that shape the policy toward clinically meaningful structured reports.

2.3 Severity-Aware Impression Reward

Clinically, the Impression section reflects a high-level diagnostic decision. Errors on critical conditions (e.g., pneumothorax) carry higher clinical risk than minor lexical deviations. To model this asymmetry, we design a severity-aware reward on the Impression section. Let y^{imp} denote the generated Impression and $y^{\text{imp}*}$ the reference Impression. A frozen classifier `fcls` extracts binarized clinical labels:

$$z^{\text{imp}} = \texttt{fcls}(y^{\text{imp}}), \quad z^{\text{imp}*} = \texttt{fcls}(y^{\text{imp}*}). \tag{6}$$

We compute a severity-weighted matching reward:

$$r_{\text{imp}}(y, y^*) = \sum_c w_c \cdot \left(\mathbb{I}_{\text{TP}}^{(c)} - \mathbb{I}_{\text{FN}}^{(c)} - \mathbb{I}_{\text{FP}}^{(c)} \right), \tag{7}$$

where $\mathbb{I}_{\text{TP}}^{(c)} = \mathbb{I}[z_c^{\text{imp}} = 1 \wedge z_c^{\text{imp}*} = 1]$ (and similarly for FN/FP), and w_c assigns larger weights to clinically critical conditions [2, 7]. This reward shapes the policy toward risk-sensitive diagnostic alignment and improves diagnostic accuracy.

2.4 Cross-Sectional Consistency Reward

Radiology reports require *Impression* conclusions to be grounded in *Findings*. However, recent methods do not enforce this dependency, risking hallucinated or contradictory diagnoses. To address this, we extract clinical label sets from both sections:

$$z^{\text{find}} = \texttt{fcls}(y^{\text{find}}), \quad z^{\text{imp}} = \texttt{fcls}(y^{\text{imp}}). \tag{8}$$

Let $\mathcal{F}$ and $\mathcal{I}$ denote the sets of positive labels corresponding to z^{find} and z^{imp}, respectively. The consistency reward is defined as:

$$r_{\text{consist}}(y) = \begin{cases} 1, & \text{if } \mathcal{I} \subseteq \mathcal{F} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

Here, $\mathcal{I} \subseteq \mathcal{F}$ checks if the *Impression* set is a subset of the *Findings* set, and the reward is 0 for any *Impression* labels not found in the *Findings* section (hallucinated diagnoses). This reward term shapes the policy to reduce cross-sectional contradictions and discourage unsupported diagnoses, ensuring that diagnostic conclusions in the *Impression* are aligned with the *Findings* section. In summary, while the subset formulation in the consistency reward enforces rigorous consistency by grounding the *Impression* in the *Findings* to suppress unsupported conclusions, it may increase the risk of clinically significant omissions. The severity-aware reward addresses this trade-off by penalizing omissions of critical findings, thereby maintaining clinical sensitivity.

3 Experiments

Datasets We evaluate our model on four datasets covering different imaging modalities: CarotidUS-MRG: A private dataset of 12,000 carotid ultrasound studies with expert-annotated reports, split by patient (70/15/15%) to prevent leakage. Reports follow a standardized format for structured clinical documentation. MIMIC-CXR [11]: A public chest X-ray dataset with 377,110 frontal images and 227,827 associated reports. IU-Xray [5]: A public chest X-ray dataset with 8,121 images and 7,470 reports from 3,955 patients. ReXGradient-160K [24]: The largest publicly available multi-site chest X-ray dataset, containing 273,004 images from 160,000 radiological studies across 109,487 patients from 79 medical sites. We adopt the official split for consistency with prior studies [25], enabling direct comparison with existing work.

Evaluation Metrics. We assess performance using natural language generation (NLG) and clinically grounded metrics. BLEU [14], BERTScore [23], and SembScore [17] evaluate report-level fluency and semantic similarity. For MIMIC-CXR, IU-Xray, and ReXGradient, we follow the RexRank [25] protocol with RadGraph [9], 1/RadCliQ-v1 [22], CheXbert-F1 [17], and consistency (CheXbert labeler subset check). For CarotidUS-MRG, where RadGraph and RadCliQ-v1 are inapplicable, KeyWordMatch uses a modality-specific labeler trained on privately annotated reports to extract and overlap key medical terms.

Baselines and Implementation. We evaluate our approach, which employs Supervised Fine-Tuning on Qwen2.5-VL-3B [1] followed by GRPO with our semantic hierarchical rewards, against five state-of-the-art methods: R2GenGPT [19], MedGemma-4B [16], MedVersa [26], CheXpertPlus [3], and RadFM [20]. All models are trained under identical settings: learning rate 1×10^{-5}, LoRA rank 8, four generations per case and a batch size of 2. Training took approximately 672 GPU-hours on 8× MetaX C500 GPUs (64 GB VRAM each). For reproducibility, r_{syntax} utilizes BLEU, ROUGE-L, BERTScore, and SembScore; r_{clinical} uses RadGraph micro-F1. r_{imp} and r_{consist} employ CheXbert (CXR) and a private ultrasound labeler. For fair comparison, MedGemma and MedVersa are fine-tuned on IU-Xray and ReXGradient.

3.1 Main Results

Cross-Dataset Performance Superiority. RadHiera consistently outperforms others across all four datasets in Tables 1 and 2, leading in six of seven MIMIC-CXR metrics and sweeping IU-Xray, ReXGradient, and the ultrasound CarotidUS-MRG. On MIMIC-CXR, while MedVersa slightly leads in BERTScore (0.430 vs. 0.421), RadHiera substantially improves semantic alignment with a SembScore of 0.442 vs. 0.315, alongside gains in clinical efficacy metrics like RadGraph (0.282 vs. 0.273) and Consistency (0.851 vs. 0.776). This gap widens on IU-Xray and ReXGradient, where RadHiera exceeds the runner-up by 0.473 in 1/RadCliQ-v1 (2.325 vs. 1.852), suggesting stronger generalization to other datasets. In CarotidUS-MRG, consistent improvements in all metrics, including an increase in F1 from 0.469 to 0.519, demonstrate robustness to the modality shift.

Table 1. Performance comparison on MIMIC-CXR, IU-Xray, and ReXGradient. NLG metrics evaluate textual similarity at the report level. Clinical metrics assess factual and diagnostic correctness. Bold = best, underline = second-best.

Dataset	Model	NLG Metrics			CE Metrics			
		BLEU-2 ↑	BERTScore ↑	SembScore ↑	RadGraph ↑	1/RadCliQ-v1 ↑	F1 ↑	Consistency ↑
MIMIC-CXR	RadFM [20]	0.081	0.281	0.245	0.111	0.625	0.245	0.741
	CheXpertPlus [3]	0.196	0.389	0.429	0.166	0.838	0.322	0.729
	R2GenGPT [19]	0.203	0.407	0.305	0.243	0.909	0.348	0.753
	MedGemma-4B [16]	0.185	0.417	0.301	0.258	0.905	0.359	0.771
	MedVersa [26]	<u>0.193</u>	**0.430**	<u>0.315</u>	<u>0.273</u>	<u>0.919</u>	<u>0.376</u>	<u>0.776</u>
	RadHiera (Ours)	**0.206**	<u>0.421</u>	**0.442**	**0.282**	**0.939**	**0.394**	**0.851**
IU-Xray	RadFM [20]	0.228	0.491	0.589	0.254	1.320	0.421	0.841
	CheXpertPlus [3]	0.264	0.516	0.608	0.257	1.619	0.469	0.828
	R2GenGPT [19]	0.271	0.558	0.601	0.269	1.751	0.483	0.852
	MedGemma-4B [16]	<u>0.259</u>	0.571	0.603	0.268	1.821	<u>0.502</u>	0.891
	MedVersa [26]	0.251	<u>0.579</u>	<u>0.611</u>	<u>0.274</u>	<u>1.852</u>	0.494	<u>0.898</u>
	RadHiera (Ours)	**0.284**	**0.607**	**0.626**	**0.313**	**2.325**	**0.519**	**0.915**
ReXGradient	RadFM [20]	0.176	0.368	0.409	0.151	0.837	0.279	0.827
	CheXpertPlus [3]	0.238	0.437	0.463	0.208	1.038	0.313	0.809
	R2GenGPT [19]	0.251	0.478	0.479	0.231	1.050	0.334	0.812
	MedGemma-4B [16]	0.258	<u>0.486</u>	0.482	0.242	1.120	0.341	<u>0.882</u>
	MedVersa [26]	<u>0.266</u>	0.481	<u>0.488</u>	<u>0.249</u>	<u>1.170</u>	<u>0.349</u>	0.874
	RadHiera (Ours)	**0.298**	**0.544**	**0.492**	**0.299**	**1.450**	**0.376**	**0.891**

Table 2. Performance comparison on the CarotidUS-MRG dataset. F1 is computed on the Impression section over three observations (stenosis, plaque, and IMT thickness). RadGraph and RadCliQ-v1 are inapplicable, so we introduce KeyWordMatch, a metric based on a proprietary ultrasound labeler trained on expert-annotated data.

Model	BLEU-2 ↑	BERTScore ↑	SembScore ↑	Keyword ↑	F1 ↑	Consistency ↑
R2GenGPT [19]	0.453	0.844	0.815	0.589	0.449	0.823
MedGemma-4B [16]	0.458	0.846	0.812	0.571	0.461	0.873
MedVersa [26]	<u>0.473</u>	<u>0.868</u>	<u>0.837</u>	<u>0.593</u>	<u>0.469</u>	<u>0.885</u>
RadHiera (Ours)	**0.522**	**0.915**	**0.891**	**0.742**	**0.519**	**0.915**

Diagnostic Consistency and Cross-Sectional Alignment. RadHiera excels at maintaining agreement between Findings and Impression sections, as reflected by the Consistency metric. Across MIMIC-CXR, IU-Xray, and ReXGradient, RadHiera scores 0.851, 0.915, and 0.891, surpassing second-best methods by 0.075, 0.017, and 0.009. These gains correlate with higher CheXbert-F1 on the Impression section (0.394 on MIMIC-CXR), linking enhanced cross-sectional coherence to more accurate diagnostic labeling. This consistent improvement extends to CarotidUS-MRG, as evidenced by the Consistency metric (0.915 versus 0.885).

Case Study. Figure 2 compares MedVersa, RadHiera, and expert reports on MIMIC-CXR and CarotidUS-MRG. MedVersa often produces template-like descriptions focusing on superficial phrasing (e.g., pulmonary congestion, intima-media thickness, or plaque) while overlooking clinically critical indicators like ICD, stenosis severity, or plaque characteristics. Its Findings and Impression sections are sometimes weakly connected, limiting diagnostic coherence. In contrast,

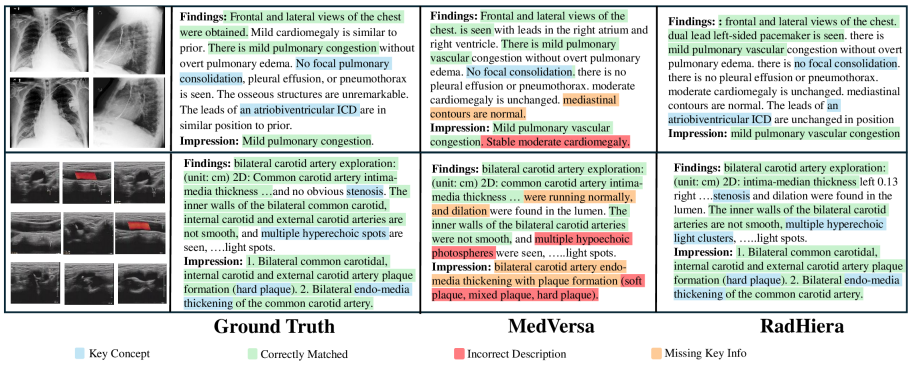

Fig. 2. Comparison of MedVersa and RadHiera outputs against ground truth. We observe that MedVersa sometimes misses clinical indicators and exhibits weak coherence between Findings and Impression sections. In contrast, RadHiera generates clinically grounded reports with higher cross-sectional consistency and diagnostic accuracy.

RadHiera generates clinically grounded reports with strong cross-sectional consistency. It links imaging evidence to the diagnostic process, quantifies stenosis severity and characterizes plaque morphology. The resulting Impression sections are concise yet logically traceable to Findings, closely resembling expert-level summaries in structure and clinical reliability.

3.2 Ablation Study

Table 3. Ablation study of reward components in RadHiera on MIMIC-CXR. We compare the full hierarchical reward against the SFT baseline, standard GRPO with base reward, and partial reward combinations.

Model Configuration	NLG Metrics			Clinical Efficacy (CE) Metrics			
	BLEU-2 ↑	BERTScore ↑	SembScore ↑	RadGraph ↑	1/RadCliQ-v1 ↑	F1 ↑	Cons. ↑
SFT (Qwen2.5-VL-3B)	0.175	0.395	0.402	0.241	0.901	0.332	0.747
GRPO (Base Reward Only)	0.188	0.402	0.411	0.255	0.915	0.368	0.772
RadHiera w/o Imp. Reward	0.202	0.417	0.426	0.261	0.912	0.368	0.789
RadHiera w/o Cons. Reward	0.204	0.419	0.433	0.274	0.927	0.359	0.821
RadHiera (Full)	**0.206**	**0.421**	**0.442**	**0.282**	**0.939**	**0.394**	**0.851**

Table 3 demonstrates that RadHiera consistently outperforms the SFT baseline and standard GRPO. GRPO improves SembScore from 0.402 to 0.411 and F1 from 0.332 to 0.368, whereas RadHiera further elevates these to 0.442 and 0.394. Ablation studies reveal synergistic effects: removing consistency rewards reduces Consistency from 0.851 to 0.821 and RadGraph from 0.282 to 0.274, while ablating Impression rewards causes a sharper Consistency drop to 0.789.

These results confirm that joint hierarchical optimization effectively enhances both linguistic fluency and clinical alignment.

4 Conclusion

In this paper, we introduce RadHiera, a semantic hierarchical reinforcement learning framework that improves medical report generation by enforcing structural and clinical consistency. By integrating severity-aware rewards and cross-sectional constraints, RadHiera ensures coherent and accurate diagnostics. Experiments across four datasets demonstrate that RadHiera consistently outperforms state-of-the-art methods in diagnostic accuracy and clinical alignment.

Acknowledgments. This work was supported by a research grant from the Joint Research Scheme (JRS) under the National Natural Science Foundation of China (NSFC) and the Research Grants Council (RGC) of Hong Kong (Project No. N_HKUST654/24); a research grant from the Innovation and Technology Commission (ITC) (Project No. GHP/124/22); and research grants from the RGC (Project Nos. 16500825, R6005-24, and C6088-25Y).

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., 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 (2025), https://arxiv.org/abs/2502.13923
2. Cardinale, L., et al.: Revisiting signs, strengths and weaknesses of standard chest x-ray in acute dyspnoea. J Thorac Dis **4**, 398–407 (2012). https://doi.org/10.3978/j.issn.2072-1439.2012.05.05
3. Chambon, P., et al.: Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. arXiv preprint arXiv:2405.19538 (2024)
4. Chen, Z., Shen, Y., Song, Y., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proc. 59th Annu. Meeting Assoc. Comput. Linguistics and 11th Int. Joint Conf. Natural Lang. Process. (ACL-IJCNLP). pp. 1439–1449 (Nov 2021)
5. Demner-Fushman, D., et al.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
6. Ganeshan, D., et al.: Structured reporting in radiology. Acad. Radiol. **25**(1), 66–73 (2018)
7. Horng, S., Liao, R., Schwab, P., Dalal, S., Golland, P.: Deep learning to quantify pulmonary edema in chest radiographs (2020)
8. Hou, W., Cheng, Y., Xu, K., Li, W., Liu, J.: RECAP: Towards precise radiology report generation via dynamic disease progression reasoning. In: Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP). pp. 2134–2147 (Dec 2023)

9. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463 (2021)
10. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL). pp. 2577–2586 (Jul 2018)
11. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
12. Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. Adv. Neural Inf. Process. Syst. **31** (2018)
13. Liu, Q., et al.: Scaling medical imaging report generation with multimodal reinforcement learning. arXiv preprint arXiv:2601.17151 (2026)
14. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
15. Schwartz, L.H., Panicek, D.M., Berk, A.R., Li, Y., Hricak, H.: Improving communication of diagnostic radiology findings through structured reporting. Radiology **260**(1), 174–181 (2011)
16. Sellergren, A., Kazemzadeh, S., Jaroensri, T., et al.: MedGemma technical report (2026), https://arxiv.org/abs/2507.05201
17. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A., Lungren, M.: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Proc. Empirical Methods Natural Lang. Process. (EMNLP). pp. 1500–1519 (Nov 2020)
18. Wang, Z., Han, H., Wang, L., Li, X., Zhou, L.: Automated radiographic report generation purely on transformer: A multicriteria supervised approach. IEEE Trans. Med. Imag. **41**(10), 2803–2813 (Oct 2022)
19. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. Meta-Radiology **1**(3), 100033 (2023)
20. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. Nature Communications **16**(1), 7866 (2025)
21. Yan, S., Cheung, W.K., Chiu, K.W., Tong, T.M., Cheung, K.C., See, S.: Attributed abnormality graph embedding for clinically accurate x-ray report generation. IEEE Trans. Med. Imag. **42**(8), 2211–2222 (aug 2023). https://doi.org/10.1109/tmi.2023.3245608
22. Yu, F., et al.: Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation. Tech. rep., Radiology and Imaging (Aug 2022). https://doi.org/10.1101/2022.08.30.22279318
23. Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2020)
24. Zhang, X., Acosta, J.N., Miller, J., Huang, O., Rajpurkar, P.: Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. arXiv preprint arXiv:2505.00228 (2025)
25. Zhang, X., Zhou, H.Y., Yang, X., Banerjee, O., Acosta, J.N., Miller, J., Huang, O., Rajpurkar, P.: Rexrank: A public leaderboard for ai-powered radiology report generation. arXiv preprint arXiv:2411.15122 (2024)

26. Zhou, H.Y., Acosta, J.N., Adithan, S., Datta, S., Topol, E.J., Rajpurkar, P.: Medversa: A generalist foundation model for medical image interpretation. arXiv preprint arXiv:2405.07988 (2024)