Wavelet-Enhanced Coarse-to-Fine Radiology Report Generation with Large Language Models

Shuanghong Yang[1], Yan Deng[1], Qibing Qin[2], Pan Zeng[1], Henggui Zhang[3], Wei Hu[1]✉, and Wenfeng Zhang[1]✉

[1] Chongqing Normal University, Chongqing, China
wei369.workstation@cqnu.edu.cn, itzhangwf@cqnu.edu.cn
[2] Weifang University, Weifang, China
[3] The University of Manchester, Manchester, United Kingdom

Abstract. Automatic chest X-ray report generation requires models to capture fine-grained visual evidence and describe findings within clinical context. Existing approaches often entangle global structure and local details, making subtle cues prone to attenuation during compression and cross-modal alignment. Most frameworks also rely on single-image inputs and underuse patient priors, limiting their ability to reflect diagnostic intent and disease evolution. We propose Wavelet-Enhanced Radiology Report Generation (WERG), a clinically motivated coarse-to-fine framework built on a multimodal large language model that supports joint conditioning on clinical context and prior images. WERG generates an initial draft, retrieves observation-aligned report snippets as textual priors, and refines the report with enhanced visual evidence. Moreover, we introduce a wavelet-based frequency-domain detail enhancement module to emphasize subtle abnormalities while preserving global semantics, and a diagnosis-guided calibration branch trained with weak observation labels. Experiments on MIMIC-CXR and IU X-Ray show consistent gains on NLG and clinical metrics, producing more detailed and clinically consistent reports.

Keywords: Radiology Report Generation · Large Language Models · Wavelet Transform.

1 Introduction

Automatic chest X-ray report generation can reduce radiologists' documentation burden and improve reporting consistency [1–4]. However, clinical utility demands more than fluent text: models should reliably capture subtle abnormalities and describe them within the available clinical context (e.g., indication, history, comparison, and prior studies). Many existing methods compress global anatomy and local cues into a unified representation, where high-frequency details can be attenuated during feature pooling and cross-modal alignment, leading to reports that are structurally complete but lack fine-grained findings [3, 5, 6]. In addition, most pipelines still generate from a single study and do not

fully exploit patient-level priors [2, 4]. Recent multimodal large language models (LLMs), from general systems to radiology-oriented models [7–11], have improved generation flexibility, yet factuality and clinical consistency remain sensitive to the quality and diagnostic focus of visual evidence [12].

To address these issues, we propose Wavelet-Enhanced Report Generation (WERG), a multimodal-LLM-based coarse-to-fine framework that strengthens the utility of visual evidence while incorporating patient priors. WERG introduces Frequency-domain Detail Enhancement (FDE), which performs wavelet decomposition on intermediate features to separate low-frequency structure and high-frequency details and applies controllable enhancement to emphasize subtle cues while preserving global semantics. We further add Diagnosis-Guided Visual Feature Calibration (DGVFC) with weak supervision from report-derived observation labels, encouraging the enhanced features to align with diagnostic semantics. Based on these components, WERG first produces a coarse draft and then refines it using clinical context and retrieved similar report snippets, improving detail coverage and consistency.

The main contributions are summarized as follows:

- We develop WERG as a coarse-to-fine multimodal LLM framework that combines clinical context, retrieved report priors, and enhanced visual evidence for report refinement.
- We introduce FDE and DGVFC as visual-side components, using wavelet-based detail enhancement and weakly supervised diagnostic calibration to improve fine-grained cue utilization and diagnostic semantic alignment.
- Experiments on MIMIC-CXR and IU X-Ray show improvements on multiple NLG and clinical metrics, producing reports with richer detail coverage and stronger clinical consistency.

2 Method

2.1 Overall Framework

The proposed Wavelet-Enhanced Radiology Report Generation (WERG) improves detail coverage and clinical consistency by leveraging patient priors, retrieved report priors, and fine-grained visual cues within a multimodal large language model. As shown in Fig. 1, WERG generates a coarse draft and then refines it by injecting observationally aligned textual priors through Report Retrieval and Knowledge Extraction (RRKE), while enhancing and calibrating visual features through Frequency-domain Detail Enhancement (FDE) and Diagnosis-Guided Visual Feature Calibration (DGVFC).

WERG first produces a coarse draft report, denoted $y^{(0)}$, which serves as a stable scaffold for subsequent refinement. Based on BLIP-3 [13], we employ a vision encoder coupled with a Perceiver Resampler [13] to transform each image into a set of visual tokens. These tokens, concatenated with the textual prompt, are fed to a decoder-only LLM to generate the report autoregressively.

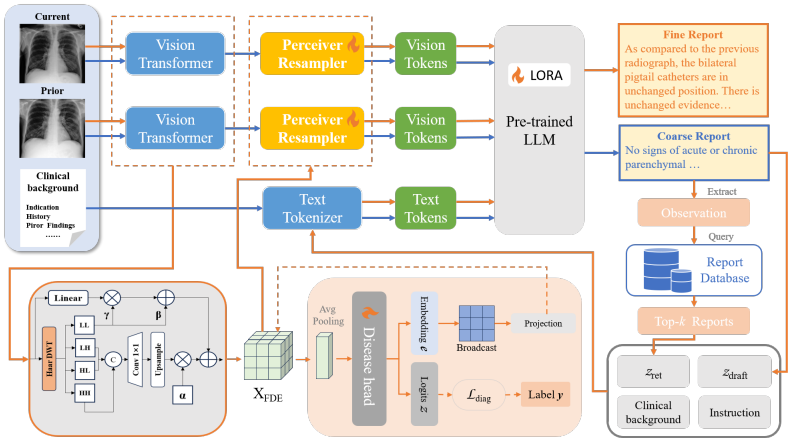

Fig. 1. Overview of WERG. The model conditions on the current CXR with optional clinical context and prior images. RRKE injects retrieved report priors, and visual evidence is enhanced and calibrated by FDE and DGVFC for report generation.

Given the current radiograph $I^{(0)}$, the model may additionally take patient priors, including clinical context text c and an optional set of prior radiographs $\mathcal{I}^{(\text{prior})} = \{I^{(1)}, \ldots, I^{(m)}\}$. We convert c into a structured textual prompt x and use (x, V) to condition report generation, where V denotes the visual context assembled from the available radiographs.

For each radiograph $I^{(j)}$ (including $j = 0$), the vision encoder $f_{\mathrm{v}}(\cdot)$ outputs patch-level features, which are compressed by a Perceiver Resampler $r(\cdot)$ into a fixed number of visual tokens. Concretely, $r(\cdot)$ aggregates the variable-length patch sequence into M latent tokens via cross-attention, yielding:

$$\mathrm{V}^{(j)} = r\big(f_{\mathrm{v}}(I^{(j)})\big) \in \mathbb{R}^{M \times d}, \tag{1}$$

where M is fixed across images. We concatenate the per-image tokens to form the overall visual context $\mathrm{V} = \mathrm{Concat}\big(\mathrm{V}^{(0)}, \mathrm{V}^{(1)}, \ldots, \mathrm{V}^{(m)}\big)$, omitting missing priors when unavailable. The coarse report is generated autoregressively conditioned on (x, V) under the standard negative log-likelihood objective. While the coarse report provides a coherent structure, it may still miss subtle but clinically important findings. The following modules refine the draft by injecting observation-aligned text priors and strengthening visual features.

2.2 Report Retrieval and Knowledge Extraction (RRKE)

Disease-logit similarity retrieval We construct a report database $\mathcal{D} = \{R_i\}_{i=1}^{N}$ from the training split, where each entry is a study-level findings report with sentence boundaries for extraction. To retrieve semantically related reports beyond surface word overlap, we represent each report by a disease-probability profile produced by a predictor $g(\cdot)$ that outputs K_r-dimensional disease logits. Given

the query text q formed from the available patient context, we obtain disease distributions $p_q = \mathrm{Softmax}(g(q))$ and $p_i = \mathrm{Softmax}(g(R_i))$ for each database entry, and rank candidates by the KL divergence:

$$d_i = D_{\mathrm{KL}}(p_q \,\|\, p_i). \tag{2}$$

We keep the top-k reports with the smallest distances as retrieval candidates for subsequent sentence-level prior extraction.

Observation-guided extraction and prior injection To mitigate noise from directly appending retrieved reports, we construct two short snippets for refinement. We extract z_{draft} from the coarse report as a case-specific anchor, and z_{ret} from the top-k retrieved reports by selecting sentences using observation tags.

Concretely, we start from the coarse report $y^{(0)}$ with sentence-level observation annotations. From this, we derive an image-side observation set $\mathcal{O}_{\mathrm{img}}$ using the same fixed observation predictor employed for indexing the report database. This ensures a consistent tag space across both processes. We form z_{draft} by keeping sentences whose tags overlap with $\mathcal{O}_{\mathrm{img}}$ as a lightweight filter for potentially unsupported statements. For each retrieved report, we then select sentences whose tags are relatively less emphasized in z_{draft}, and concatenate them into z_{ret}. Next, the clinical context c together with z_{draft} and z_{ret} is formatted as the refinement prompt:

$$x^{+} = \mathrm{Format}(c,\, z_{\mathrm{draft}},\, z_{\mathrm{ret}},\, \texttt{Instruction}), \tag{3}$$

where $\mathrm{Format}(\cdot)$ is a fixed template and `Instruction` prompts the model to organize observations before generating the final report.

2.3 Frequency-domain Detail Enhancement (FDE)

RRKE supplies observation-aligned textual priors, yet detailed generation still relies on informative visual features. FDE enhances intermediate visual representations in the frequency domain (between the vision encoder and tokenization) to emphasize subtle cues, without altering the language-model decoding mechanism.

Given the patch feature sequence output by the vision encoder, we remove the global token and reshape the remaining features into a 2D feature map:

$$\mathrm{X} \in \mathbb{R}^{B\times C\times H\times W}. \tag{4}$$

FDE applies a 2D orthogonal Haar wavelet transform channel-wise [14, 15], decomposing features into low-frequency structure and high-frequency details:

$$(\mathrm{X}_{LL}, \mathrm{X}_{LH}, \mathrm{X}_{HL}, \mathrm{X}_{HH}) = \mathrm{DWT}(\mathrm{X}). \tag{5}$$

We concatenate the three high-frequency bands, fuse them, and upsample to the original resolution to form a detail residual:

$$\mathrm{H} = \mathrm{Up}(\phi([\mathrm{X}_{LH}, \mathrm{X}_{HL}, \mathrm{X}_{HH}])), \tag{6}$$

where $[\cdot]$ denotes channel concatenation, $\phi(\cdot)$ is a 1×1 convolution, and $\mathrm{Up}(\cdot)$ is bilinear upsampling.

To preserve structural semantics, we extract global statistics from the low-frequency component and generate channel-wise affine modulation parameters, which softly adjust the original features. To emphasize subtle cues, we inject H as a residual controlled by a global scalar. The final FDE output is:

$$\mathrm{X}_{\mathrm{FDE}} = (1 + \boldsymbol{\gamma}) \odot \mathrm{X} + \boldsymbol{\beta} + \alpha \cdot \mathrm{H}, \tag{7}$$

where $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^{B \times C \times 1 \times 1}$ are channel-wise modulation parameters, α is a global scalar, and $\odot$ denotes element-wise multiplication. This design provides controllability: the low-frequency branch stabilizes global semantics, while the high-frequency branch highlights details without "overwriting" the original representation. We reshape $\mathrm{X}_{\mathrm{FDE}}$ back to the patch sequence for tokenization.

2.4 Diagnosis-Guided Visual Feature Calibration (DGVFC)

Beyond detail enhancement, we further calibrate the visual representation toward diagnostically relevant cues. DGVFC applies weak supervision from report-derived observation labels and outputs a compact disease embedding for conditioning (Fig. 1), without modifying the LLM decoder.

Given the FDE-enhanced feature map $\mathrm{X}_{\mathrm{FDE}} \in \mathbb{R}^{B \times C \times H \times W}$, we apply global average pooling to obtain an image-level representation h, which is then passed to a disease prediction head $f_{\mathrm{dis}}(\cdot)$ to produce multi-label observation logits and a compact disease embedding:

$$h = \mathrm{GAP}(\mathrm{X}_{\mathrm{FDE}}), \qquad (z, e) = f_{\mathrm{dis}}(h), \tag{8}$$

where $z \in \mathbb{R}^{B \times K}$ are logits for K observation classes and $e \in \mathbb{R}^{B \times D}$ is the disease embedding. We use observation labels extracted from reference reports as weak supervision [16] and apply a multi-label classification loss to z:

$$\mathcal{L}_{\mathrm{diag}} = \mathrm{BCEWithLogits}(z,\, y), \tag{9}$$

where $y \in \{0, 1\}^{B \times K}$ is the corresponding label vector. This supervision encourages the visual representation to remain sensitive to diagnostic semantics before compression and cross-modal interaction, and makes e a compact semantic summary of the case. For calibration, we spatially broadcast e and fuse it with $\mathrm{X}_{\mathrm{FDE}}$ via a lightweight projection to obtain a disease-conditioned modulation on the visual feature map.

3 Experiments

3.1 Datasets and Experimental Settings

We fine-tune on MIMIC-CXR [17] using frontal views and evaluate cross-dataset performance on IU X-Ray [18] using all available frontal images. We report

Table 1. Comparison on the MIMIC-CXR dataset. The highest scores are highlighted in bold, while the second-highest scores are indicated with an underline.

MIMIC-CXR								
Methods	Venue	B-1	B-4	MTR	R-L	P	R	F1
PPKED [3]	CVPR'21	0.360	0.106	0.149	0.284	–	–	–
R2GenCMN [4]	ACL'21	0.353	0.106	0.142	0.278	0.334	0.275	0.278
M2KT [22]	MedIA'23	0.386	0.111	–	0.274	0.420	0.339	0.352
KiUT [23]	CVPR'23	0.393	0.113	0.160	0.285	0.371	0.318	0.321
METrans [24]	CVPR'23	0.386	0.124	0.152	0.291	0.364	0.309	0.311
DCL [5]	CVPR'23	–	0.109	0.150	0.284	0.471	0.352	0.373
EKAGen [6]	CVPR'24	**0.419**	0.119	0.157	0.287	<u>0.517</u>	0.483	<u>0.499</u>
HSA [25]	MICCAI'24	0.386	0.120	0.163	0.288	0.480	0.357	0.379
MPO [26]	AAAI'25	0.416	0.139	0.162	**0.309**	0.436	0.376	0.353
DDaTR [27]	TMI'25	0.401	0.113	0.162	0.272	–	–	–
RGRG [28]	CVPR'23	0.373	0.126	0.168	0.264	0.461	0.475	0.447
KARGEN [12]	MICCAI'24	0.417	<u>0.140</u>	0.165	<u>0.305</u>	–	–	–
R2-LLM [29]	AAAI'24	0.402	0.128	<u>0.175</u>	0.291	0.465	0.482	0.473
PromptMRG [30]	AAAI'24	0.398	0.112	0.157	0.268	0.501	<u>0.509</u>	0.476
KACL [31]	MICCAI'25	0.414	0.136	0.169	0.303	0.503	0.442	0.469
WERG (Ours)	MICCAI'26	<u>0.418</u>	**0.146**	**0.347**	0.289	**0.583**	**0.610**	**0.596**

NLG metrics for text similarity, including BLEU (B-1/B-4) [19], METEOR (MTR) [20], and ROUGE-L (R-L) [21]. On MIMIC-CXR, we additionally report CE metrics, including Precision/Recall/F1 (P/R/F1), which are evaluated by converting reports into 14 disease classification labels using CheXbert [16]. We omit CE on IU X-Ray due to potential label-space mismatch.

We train with the AdamW optimizer using a learning rate of 5×10^{-5} and cosine learning-rate scheduling. Mixed precision uses bf16. At inference, we apply beam search with num beams=5. For retrieval augmentation, we keep the top-$k = 2$ most similar report snippets as textual priors and insert them into the prompt to guide refinement.

3.2 Comparison with the Baselines

We compare WERG on MIMIC-CXR with two groups of baselines: expert non-LLM models and recent LLM-based generators (Table 1). We report both NLG and CE metrics. WERG achieves the best CE metrics, outperforming EKAGen and the LLM-based R2-LLM. It also obtains the highest MTR. These improvements are consistent with combining diagnosis-aware visual features enhancement and observation-aligned textual priors.

For NLG metrics, WERG achieves the highest B-4 and MTR, and remains competitive on B-1, which is close to the best-performing EKAGen. Meanwhile, it does not dominate every surface-overlap metric, which can be affected by report-writing variability and the fact that multiple clinically valid phrasings

Table 2. Comparison on the IU X-Ray dataset. The highest scores are highlighted in bold, while the second-highest scores are indicated with an underline.

Methods	Venue	B-4	MTR	R-L
M2KT	MedIA'23	0.078	0.153	0.261
DCL	CVPR'23	0.074	0.152	0.267
RGRG	CVPR'23	0.063	0.146	0.180
PromptMRG	AAAI'24	<u>0.098</u>	<u>0.160</u>	0.281
DDaTR	TMI'25	**0.103**	<u>0.160</u>	<u>0.307</u>
WERG (Ours)	MICCAI'26	0.086	**0.309**	**0.323**

Table 3. Ablation studies on MIMIC-CXR. All variants start from the same fine-tuned backbone. The highest scores are highlighted in bold, while the second-highest scores are indicated with an underline.

RRKE	FDE	DGVFC	B-4	MTR	R-L	P	R	F1
			0.123	0.304	0.264	0.576	0.464	0.514
✓			0.130	<u>0.317</u>	0.265	0.576	<u>0.608</u>	<u>0.591</u>
	✓		0.126	0.307	0.268	0.583	0.484	0.529
	✓	✓	<u>0.131</u>	<u>0.317</u>	<u>0.267</u>	**0.586**	0.592	0.589
✓	✓		<u>0.131</u>	0.314	0.266	0.576	<u>0.608</u>	<u>0.591</u>
✓	✓	✓	**0.146**	**0.347**	**0.289**	<u>0.583</u>	**0.610**	**0.596**

may describe the same observations. Thus, gains in content coverage and phrasing choices may not be fully reflected by n-gram overlap with a single reference.

Following prior zero-shot IU X-Ray evaluation protocols [30, 27], we report only NLG metrics on IU X-Ray because the standard IU test split may not be well suited to disease-aware metrics. Under this setting, WERG achieves the best MTR and R-L, but its B-4 is lower than the strongest baselines while still exceeding several prior methods. The lower B-4 can arise from wording variability despite correct diagnostic predictions, since prior-guided refinement may introduce clinically consistent paraphrases, sentence reordering, and contextual or negative statements; the higher MTR and R-L indicate stronger semantic coverage and robustness to distribution shift.

3.3 Ablation Studies

We ablate RRKE, FDE, and DGVFC on MIMIC-CXR (Table 3) using the same fine-tuned backbone and decoding settings. DGVFC is evaluated together with FDE since it is defined on the FDE-enhanced feature space.

As shown in Table 3, RRKE obtains the largest gain on clinical coverage, mainly improving Recall and F1. FDE alone remains comparable to the fine-tuned baseline, suggesting that wavelet-based enhancement is conservative without an explicit diagnostic constraint. Adding DGVFC on top of FDE improves

Image	Clinical background	Ground Truth	Coarse report	Fine report
	"**Bilateral pigtails** and possible left **pneumothorax**. Known right **pneumothorax**."	There is a mild-to-moderate left **pneumothorax** with rightward **mediastinal shift**. The small right **pneumothorax** is **stable**. There is also a moderate left **pleural effusion**. **Bilateral pigtail** catheters are **in place** …	No signs of acute or chronic parenchymal infiltrates are present, and the lateral and posterior pleural sinuses are free. ~~**No pneumothorax in the apical area** ...~~	**As compared to** the previous radiograph, the **bilateral pigtail** catheters are in **unchanged** position. There is **unchanged** evidence of a left apical **pneumothorax**. The extent of the left **pleural effusion** is also **unchanged** ...
	"Right lung empyema; post **chest tube** management."	Right-sided **chest tube** has been **removed**. There is a **hydropneumothorax** in the inferior right chest. The amount of fluid has increased. There is no focal consolidation.	~~**...the right chest tube has been removed and there is no evidence of pneumothorax.**~~ Little change in the appearance of the loculated right **pleural effusion** ...	**As compared to** the prior study, the right **chest tube** has been **removed**. The extent of the loculated right basal **pneumothorax** is **unchanged**. Also **unchanged** is the extent of the pleural fluid collection ...
	"Right-sided chest pain, assess pneumonia, **pneumothorax**, or other abnormalities."	Multiple right rib deformities are noted, consistent with prior fractures. Areas of pleuroparenchymal **scarring** are present. **No pneumothorax**.	The lungs are clear without focal consolidation. No **pleural effusion** or **pneumothorax**. The osseous structures are unremarkable.	...Blunting of the right costophrenic angle is compatible with chronic **pleural thickening**. Old right-sided **rib fractures** are present. **No pneumothorax**.
	"Alcoholic cirrhosis, in withdrawal, and encephalopathy, **interval change**."	The right apical opacity is **stable** and concerning for either a **nodule** or infection. There are low lung volumes with bibasilar **atelectasis**. **No pneumothorax**.	No focal parenchymal opacity suggesting pneumonia. No pulmonary edema. **No pneumothorax**.	The known right lung **nodule** is again noted. There is minimal **atelectasis** at the right lung bases. **No pneumothorax**

Fig. 2. Qualitative examples on MIMIC-CXR. Rows show image, clinical background, ground truth, coarse report, and refined report. The prior image (if available) appears at the bottom-right. Colors highlight key categories and strikethrough marks incorrect negations in the coarse report.

diagnosis consistency (notably Precision), indicating that weak supervision helps translate enhanced cues into observation-aligned evidence. Combining RRKE with FDE+DGVFC achieves the best overall performance, supporting the complementarity between textual priors and diagnosis-aware visual enhancement.

3.4 Qualitative Analysis

As shown in Fig. 2, the coarse report may omit or incorrectly negate clinically important statements (e.g., devices and abnormalities), whereas WERG more consistently preserves case-relevant findings and comparison language. Across cases, WERG reduces incorrect negations and improves coverage of device-related and subtle abnormalities compared with the coarse drafts, while maintaining coherent trend descriptions when comparison information is available.

4 Discussion and Conclusion

We presented WERG, a wavelet-enhanced coarse-to-fine radiology report generation framework based on a multimodal large language model. By integrating retrieved textual priors and enhanced visual evidence, WERG improves text quality and clinical consistency on MIMIC-CXR and IU X-Ray.

Nevertheless, WERG remains sensitive to retrieval quality and automatically extracted observation labels, and clinical deployment requires privacy-preserving indexing and local validation.

Acknowledgments. This work was supported by the Special Foundation for State Major Basic Research Program of China under Grant 2025YFC2708700, in part by the National Natural Science Foundation of China under Grant 62302338, in part by the

Natural Science Foundation of Chongqing under Grants CSTB2025NSCQ-GPX1037 and CSTB2023NSCQ-MSX0407, in part by the Natural Science Foundation Innovation and Development Joint Fund Project of Chongqing under Grant CSTB2023NSCQ-LZX0148, in part by the Brain-Gain Plan of New Chongqing Foundation under Grant CSTB2024YCJH-KYXM0108, in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grants KJQN202500532 and KJQN202500511, in part by the Chongqing Normal University Foundation under Grant 24XLB018, and in part by the Shandong Province Natural Science Foundation under Grants ZR2022QF046 and ZR2025MS1067.

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

References

1. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proc. 56th Annual Meeting of the Association for Computational Linguistics, pp. 2577–2586 (2018).
2. Chen, Z., Song, Y., Chang, T.-H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proc. Conf. on Empirical Methods in Natural Language Processing, pp. 1439–1449 (2020).
3. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 13753–13762 (2021).
4. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. In: Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 5904–5914 (2021).
5. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest X-ray report generation. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 3334–3343 (2023).
6. Bu, S., Li, T., Yang, Y., Dai, Z.: Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 14194–14204 (2024).
7. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., et al.: LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems, vol. 36 (2023).
8. Hyland, S.L., Bannur, S., Bouzid, K., Castro, D.C., Ranjit, M., Schwaighofer, A., et al.: MAIRA-1: a specialised large multimodal model for radiology report generation. arXiv preprint arXiv:2311.13668 (2023).
9. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., et al.: MAIRA-2: grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024).
10. Zhang, X., Meng, Z., Lever, J., Ho, E.S.L.: Libra: leveraging temporal images for biomedical radiology analysis. In: Findings of the Association for Computational Linguistics: ACL 2025, pp. 17275–17303 (2025).
11. Hou, W., Cheng, Y., Xu, K., Li, H., Hu, Y., Li, W., et al.: RADAR: enhancing radiology report generation with supplementary knowledge injection. In: Proc. 63rd Annual Meeting of the Association for Computational Linguistics, pp. 26366–26381 (2025).

12. Li, Y., Wang, Z., Liu, Y., Wang, L., Liu, L., Zhou, L.: KARGEN: knowledge-enhanced automated radiology report generation using large language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 382–392. Springer (2024).
13. Xue, L., Shu, M., Awadalla, A., Wang, J., Yan, A., Purushwalkam, S., et al.: xGen-MM (BLIP-3): a family of open large multimodal models. arXiv preprint arXiv:2408.08872 (2024).
14. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. 11(7), 674–693 (1989).
15. Yao, T., Pan, Y., Li, Y., Ngo, C.-W., Mei, T.: Wave-ViT: unifying wavelet and transformers for visual representation learning. In: Computer Vision – ECCV 2022. LNCS, vol. 13685, pp. 328–345. Springer (2022).
16. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.P.: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Proc. Conf. on Empirical Methods in Natural Language Processing, pp. 1500–1519 (2020).
17. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-Y., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data 6(1), 317 (2019).
18. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., et al.: Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc. 23(2), 304–310 (2016).
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: Proc. 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318 (2002).
20. Denkowski, M., Lavie, A.: Meteor universal: language specific translation evaluation for any target language. In: Proc. Workshop on Statistical Machine Translation, pp. 376–380 (2014).
21. Lin, C.-Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004).
22. Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S.K., Xiao, L.: Radiology report generation with a learned knowledge base and multi-modal alignment. Med. Image Anal. 86, 102798 (2023).
23. Huang, Z., Zhang, X., Zhang, S.: KiUT: knowledge-injected U-transformer for radiology report generation. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 19809–19818 (2023).
24. Wang, Z., Liu, L., Wang, L., Zhou, L.: METransformer: radiology report generation by transformer with multiple learnable expert tokens. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 11558–11567 (2023).
25. Zhu, Z., Cheng, X., Zhang, Y., Chen, Z., Long, Q., Li, H., et al.: Multivariate cooperative game for image-report pairs: hierarchical semantic alignment for medical report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 303–313. Springer (2024).
26. Xiao, T., Shi, L., Liu, P., Wang, Z., Bai, C.: Radiology report generation via multi-objective preference optimization. Proc. AAAI Conf. on Artificial Intelligence 39(8), 8664–8672 (2025).
27. Song, S., Tang, H., Yang, H., Li, X.: DDaTR: dynamic difference-aware temporal residual network for longitudinal radiology report generation. IEEE Trans. Med. Imaging 44(12), 5345–5357 (2025).

28. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 7433–7442 (2023).
29. Liu, C., Tian, Y., Chen, W., Song, Y., Zhang, Y.: Bootstrapping large language models for radiology report generation. Proc. AAAI Conf. on Artificial Intelligence 38(17), 18635–18643 (2024).
30. Jin, H., Che, H., Lin, Y., Chen, H.: PromptMRG: diagnosis-driven prompts for medical report generation. Proc. AAAI Conf. on Artificial Intelligence 38(3), 2607–2615 (2024).
31. Sha, Y., Pan, H., Meng, W., Li, K.: Contrastive knowledge-guided large language models for medical report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 111–120. Springer (2026).