Hierarchical Memory for Radiology Report Generation with Rich Context

Qingyue Jiao[1][0009−0006−4188−0579], Jiahao Zheng[1][0009−0008−0996−3774], Jun Zhuang[2][0000−0002−7142−2193], and Yiyu Shi[1][0000−0002−6788−9823]⋆

[1] University of Notre Dame, Notre Dame IN 46556, USA
[2] Boise State University, Boise ID 83725, USA
{qjiao, jzheng7, yshi4}@nd.edu; junzhuang@boisestate.edu

Abstract. Radiology report generation (RRG) increasingly relies on rich evidence, most commonly retrieved similar reports from other patients and longitudinal history from the same patient. Prior approaches often concatenate evidence into a long prompt, which can degrade performance. Such long contexts are typically noisy and only partially relevant, making it challenging for the backbone decoder to reliably select the most relevant evidence for the current query. We propose **HM-RRG**, a hierarchical-memory framework that organizes and selects evidence instead of treating all context as a flat sequence. HM-RRG builds an image-conditioned retrieval memory bank by compressing each retrieved report into a compact memory vector grounded on the current image. It then processes the patient's prior reports as a chronological stream, maintains segment-level longitudinal memories, and performs hierarchical retrieval over the union of cross-patient and within-patient memories to form an evidence prompt for each segment during decoding. This design enables selective recall of relevant information while suppressing irrelevant or conflicting details under long contexts. Experiments on MIMIC-CXR with Longitudinal-MIMIC show that HM-RRG improves both text quality and clinical accuracy. In particular, HM-RRG achieves the best factuality metrics among state-of-the-art RRG models, demonstrating that structured memory is effective for radiology report generation. Our code is available at https://github.com/QingyueJ-nd/HM-RRG.

Keywords: Radiology Report Generation · Longitudinal Data · LLM

1 Introduction

Automated radiology report generation with large language models (LLMs) has been extensively studied [16], and recent systems increasingly rely on external evidence to improve factuality and clinical completeness. In practice, two evidence sources are used most often: (i) retrieved similar reports from other patients and (ii) longitudinal history from the same patient. Retrieval-augmented approaches mainly improve the retriever and filtering stage, then condition the

⋆ Corresponding author.

decoder on a small set of retrieved evidence. For example, Prefilling [23] integrates multimodal history with cross-attention and a memory-driven transformer to reduce hallucinations, while Diff-RRG [20] uses disease-wise patch differences to guide change-aware generation. MLRG [9] further models longitudinal disease progression by fusing current multi-view chest X-rays with prior-visit information in a multi-view longitudinal fusion network and pretraining via multi-view longitudinal contrastive learning aligned to the corresponding report.

However, despite these advancements, a key gap remains: how to use both evidence sources together during decoding. Cross-patient retrieval can provide strong priors for rare findings, but it may miss patient-specific trends without longitudinal context. Longitudinal history is often most relevant to the current patient, as temporal information in radiology reports captures disease progression across examinations and supports clinical decision-making [18], yet it can bias the model toward copying prior statements and overlooking new changes. To the best of our knowledge, there is still no unified report-generation pipeline that effectively leverages both retrieved similar reports and longitudinal history within a single, evidence-driven decoding framework.

Motivated by real clinical workflows, we aim to combine both evidence sources in a way that remains robust in long and noisy contexts. Radiologists do not treat prior notes and similar cases as a flat sequence of text; instead, they organize, retain, and recall information selectively: they consult a patient's history to identify changes, and they draw on similar cases to resolve uncertainty and refine the diagnosis. In contrast, most LLM-based report-generation systems use external resources as a single flat, concatenated prompt, forcing the backbone to perform evidence selection implicitly. This "flat-memory" setup is brittle: as context grows, models are more likely to overlook critical evidence, over-weight irrelevant snippets, or fail to resolve conflicts.

To address this, we propose Hierarchical Memory Radiology Report Generation (HM-RRG), a rich-context framework that explicitly structures retrieved evidence and longitudinal history into a hierarchical memory. Our method is inspired by the Hierarchical Memory Transformer (HMT) [2], which improves long-context processing by maintaining segment-level memory recurrence and enabling selective recall of previously seen information. Concretely, we segment input evidence into manageable units, construct memory-augmented representations that persist across segments, and condition memory formation and retrieval on the current image so that the model can prioritize clinically relevant content while suppressing noise and contradictions. This design enables HM-RRG to leverage large evidence sets without relying on the backbone to solve long-context selection purely through prompt concatenation. Our contributions are:

1. We propose **HM-RRG**, a rich-context radiology report generation framework that mirrors real clinical reasoning by integrating both cross-patient retrieved reports and longitudinal patient history to generate more fact-grounded reports.

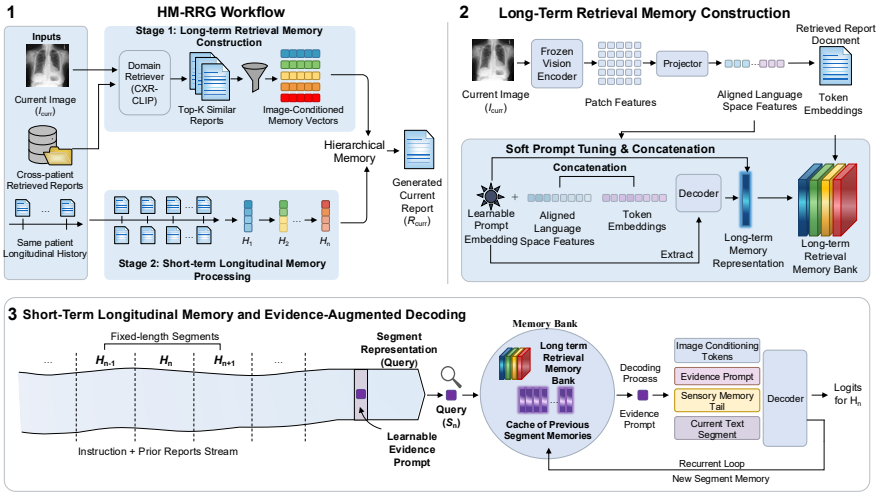

Fig. 1. Overall architecture of HM-RRG. The upper-left panel provides an overview of the HM-RRG workflow. The upper-right panel illustrates long-term memory construction, and the bottom panel depicts short-term longitudinal memory and evidence-augmented decoding.

2. We introduce current-image-conditioned hierarchical memory tokens over retrieved evidence and longitudinal context, enabling robust evidence selection under long contexts.
3. We evaluate HM-RRG on the Longitudinal-MIMIC dataset, demonstrating consistent improvements in report quality over state-of-the-art methods when leveraging rich context.

2 Method

In this section, we present the overall workflow of HM-RRG, a hierarchical-memory framework for radiology report generation that integrates retrieved evidence from other patients and longitudinal history from the same patient. Our design is motivated by the observation that these two information sources exhibit fundamentally different structures: retrieved reports form an unordered and often noisy set, whereas longitudinal priors constitute an ordered temporal stream. Treating both sources with a single uniform context mechanism can waste context budget on irrelevant retrieved details and dilute the clinically meaningful temporal signal.

HM-RRG decomposes report generation into two stages. The first stage is the construction of long-term retrieval memory. Given the current image I_{curr}, we use a domain retriever to obtain the top-K similar reports $\{R_i^{\text{retr}}\}_{i=1}^{K}$. We then transform each retrieved report into an image-conditioned memory

representation, forming a retrieval memory bank that distills "what is similar to the current case" while filtering irrelevant content. The second stage is short-term longitudinal memory processing. We process the patient's prior reports $\left\{R_t^{\text{prior}}\right\}_{t=1}^{T-1}$ in chronological order as a sequential stream, and at each step construct compact segment representations that are adaptively augmented by both the retrieval memory bank and recent longitudinal context. The resulting hierarchical memory state conditions the decoder to generate the current report R_{curr} for I_{curr}, with an explicit mechanism to focus on clinically salient evidence and temporal changes.

Long-term Retrieval Memory Given the current study image I_{curr}, we first construct an image-conditioned long-term retrieval memory bank that summarizes evidence from similar cases. This design explicitly addresses two challenges of retrieval-augmented report generation: retrieved reports form an unordered set and often contain irrelevant details, and injecting the full retrieved text into the generation context is costly and can dilute clinically salient signals. Instead, we compress each retrieved report into a compact memory vector in the decoder's hidden embedding space, conditioned on I_{curr}, yielding a lightweight evidence block that can be queried repeatedly during longitudinal decoding.

We employ a domain-specific chest X-ray retriever CXR-CLIP [19] to retrieve the top-K most similar reports based on the current image I_{curr}. We denote the retrieved reports as $\{R_i^{\text{retr}}\}_{i=1}^{K}$. We then encode the input image I_{curr} with a pre-trained frozen vision encoder to obtain patch features $\tilde{V}(I_{\text{curr}}) = \text{VisionEnc}(I_{\text{curr}})$. We then project them to the language space via $V(I_{\text{curr}}) = \text{Proj}(\tilde{V}(I_{\text{curr}}))$. We adopt the CXR-CLIP vision encoder and train a lightweight projector to match the decoder's hidden size d.

For each retrieved report R_i^{retr}, we obtain its token embeddings via the language embedding layer $E(R_i^{\text{retr}})$. We then summarize the retrieved report conditioned on the current image by soft prompt tuning [10]. We augment the retrieved report embedding with t_{mem}, a summarization prompt embedding. Specifically, it is a learnable parameter embedding to prompt the backbone decoder to summarize the current embedding. We input the report embedding with t_{mem} into the backbone decoder, denoted as LM: $H_i = \text{LM}\Big(t_{\text{mem}} \,\|\, V(I_{\text{curr}}) \,\|\, E(R_i^{\text{retr}}) \,\|\, t_{\text{mem}}\Big)$ The long-term memory representation for the i-th retrieved report is taken as the hidden state at the memory token position: $M_i^{\text{retr}} = H_i[-1]$. Finally, we stack all retrieved memories to form the long-term retrieval memory bank $M^{\text{retr}} = \left[M_1^{\text{retr}}; \ldots; M_K^{\text{retr}}\right]$, which provides a compact, image-grounded representation of retrieved evidence that can be efficiently queried during longitudinal decoding.

Short-term Longitudinal Memory Next, HM-RRG processes a patient's longitudinal history as a short-term, temporally ordered memory stream and integrates it with the long-term retrieval memory bank. The core adaptation

from HMT [2] for radiology report generation is to preserve chronological structure in prior reports and enable selective access to retrieved evidence without concatenating full retrieved text into the decoder context.

Stream Construction and Segmentation Given the current image I_{curr} and the set of prior reports $\{R_t^{\text{prior}}\}$ ordered by time, we form a single textual stream $\mathcal{X} = [\,\text{Instr} \,\|\, R_1^{\text{prior}} \,\|\, \cdots \,\|\, R_{T-1}^{\text{prior}}]$, where Instr denotes the instruction prompt. We tokenize $\mathcal{X}$ and map token IDs to embeddings using the language embedding layer, then partition the resulting embedding sequence into N fixed-length segments of L tokens $H_1, \ldots, H_N$. To avoid splitting visual tokens across segments, we treat image embeddings as segment-level conditioning (prefixed during decoding) rather than inserting them into the segmented text stream.

Segment Representation for Evidence Querying For each segment H_n, we compute a compact segment representation S_n that serves as a query for selecting relevant evidence. Concretely, we attach a small learnable summary token t_{sum} to the segment prefix and use the decoder to produce a summary vector, similar to the long-term memory construction, with S_n extracted from the summary-token position. This design provides a stable, fixed-size descriptor of the current local context, even when segments cut across report boundaries.

Hierarchical Memory Retrieval In addition to the long-term retrieval memory bank M^{retr}, we maintain a cache of segment-level longitudinal memories produced from earlier segments of the same patient, $\{M_1^{\text{seg}}, \ldots, M_{n-1}^{\text{seg}}\}$. At segment n, we form a unified hierarchical memory set by concatenating cross-patient evidence and within-patient summaries: $M_n^{\text{all}} = \big[\, M^{\text{retr}} \,;\, M_{1:n-1}^{\text{seg}} \,\big]$. We then compute an evidence prompt P_{mem} by attention over M_n^{all} using the segment query S_n: $Q_n = S_n W_q$, $K_n = M_n^{\text{all}} W_k$, $\alpha_n = \text{softmax}\Big(\frac{Q_n K_n^\top}{\sqrt{d_h}}\Big)$, and $P_{\text{mem}} = \alpha_n M_n^{\text{all}}$. This construction yields a single, compact prompt vector that adaptively mixes cross-patient evidence distilled from retrieved reports and salient cues accumulated from earlier longitudinal segments.

Local Sensory Memory Longitudinal narratives often span several segments. To preserve local information, we carry a short-range contextual tail from the previous segment. We denote the last k token embeddings of segment $n{-}1$ as $C_n = H_{n-1}[L-k : L]$ with C_1 initialized to zeros. In all experiments we set k=32, which provides local continuity while keeping overhead small [2].

Evidence-augmented Decoding and Memory Update For segment n, the decoder receives an augmented input that combines the evidence prompt P_{mem}, the sensory memory C_n, the current text segment H_n, and the current-image conditioning tokens $V(I_{\text{curr}})$: $\mathcal{X}_n = [\,V(I_{\text{curr}}) \,\|\, P_{\text{mem}} \,\|\, C_n \,\|\, H_n \,\|\, P_{\text{mem}}\,]$. Compared with flat concatenation, this formulation avoids feeding all retrieved and longitudinal reports into the decoder as one long sequence. Instead, the decoder processes the current segment together with compact memory prompts, while

evidence selection is performed over the compressed memory set, reducing the self-attention cost of long-context evidence. The decoder produces hidden states Y_n, from which we compute logits for the current segment and extract a compact segment memory M_n^{seg}. We cache $\{M_n^{\text{seg}}\}$ as short-term longitudinal memory for subsequent segments, enabling the model to keep clinically relevant information as generation progresses. Unlike the original HMT formulation, which is designed for a long (over 30k tokens) and potentially infinite input stream and thus employs a fixed-size sliding window over memory states, we have a bounded clinical setting with a capped number of longitudinal reports and retrieved cases. Therefore, we retain all segment memories within each example by default to avoid removing clinically important temporal changes and to simplify training and evaluation. Importantly, this is an implementation choice rather than a limitation: the same retrieval formulation naturally supports an optional sliding-window cache when scaling to substantially longer histories.

3 Experiments

Datasets and Implementation We conduct experiments on MIMIC-CXR-JPG [5] and Longitudinal-MIMIC [23]. The MIMIC-CXR dataset contains 377,110 chest X-ray images corresponding to 227,835 radiographic reports from 65,379 patients. Longitudinal-MIMIC is derived from MIMIC-CXR and includes 26,625 patients and 94,169 samples with at least two visits. We use MIMIC-CXR as the retrieval source and Longitudinal-MIMIC to construct longitudinal report generation samples with prior visits. We follow the official train, validation, and test splits as Longitudinal-MIMIC and ensure patient-level separation across splits to prevent leakage. For retrieval augmentation, we additionally restrict the retrieval source to the training split only. Each Longitudinal-MIMIC example provides a current study and its most recent prior. To better reflect realistic clinical history, we aggregate all available prior studies for the same patient within the split, order them chronologically, and form a single training instance per current study. The model input includes the current image, an instruction prompt, and longitudinal prior reports. The target is the current report.

We use CXR-CLIP [19] as the domain retriever to retrieve top-K similar reports from other patients, with K=5 in all experiments. We follow Diff-RRG [20] by using the same metrics and the BioMistral-7B decoder backbone [6], paired with the CXR-CLIP vision encoder [19]. We train a lightweight projector that maps CXR-CLIP vision features into the decoder embedding space, and fine-tune the decoder using LoRA [3] with rank 16 and α=16. We train using an AdamW optimizer [11] with peak learning rate 2×10^{-4}, linear decay, warmup ratio 0.03, and gradient clipping with max grad norm 1.0. Unless otherwise stated, we use gradient accumulation to reach an effective batch size of 16 per GPU. To train the hierarchical memory, we adopt a two-stage training. We set the segment length $L = 128$, the segment-summary prefix length $j = L/2 = 64$, and the sensory memory length $k = 32$. In Stage 1, we train the long-term memory construction for 2,000 steps with a learning rate of 5×10^{-6}. In Stage 2, we initialize from the

Table 1. Evaluation results on the MIMIC-CXR dataset. Baseline evaluations are cited from the original publications and MLRG [9]. Bold and underlined values indicate the highest and second-highest scores, respectively. The last row reports the percentage difference between HM-RRG and the best prior method.

Input	Method	NLG Metrics↑			CE Metrics↑			
		BLEU	**MTR**	**R-L**	**RG**	**P**	**R**	**F1**
SI	R2GenGPT [17]	0.191	0.125	0.240	-	0.331	0.253	0.287
SI	B-LLM [8]	0.243	0.175	0.291	-	0.465	<u>0.482</u>	0.473
SI	LLaVA-Rad [21]	0.168	-	<u>0.292</u>	0.208	-	-	0.475
Long	Prefilling [23]	0.198	0.137	0.274	-	0.506	0.364	0.423
Long	HERGen [15]	0.234	0.156	0.285	-	0.421	0.289	0.343
Long	Diff-RRG [20]	0.236	0.164	0.276	-	0.528	0.430	0.474
MVL	MLRG [9]	<u>0.262</u>	<u>0.176</u>	**0.320**	<u>0.291</u>	<u>0.549</u>	0.468	<u>0.505</u>
R&L	HM-RRG	**0.263**	**0.177**	<u>0.292</u>	**0.299**	**0.561**	**0.600**	**0.580**
-	Δ (%)↑	+0.4	+0.6	-8.8	+2.7	+2.2	+24.4	+14.9

Stage-1 checkpoint and continue training with hierarchical retrieval enabled. The decoder LoRA weights, image projector, learnable memory and summary tokens, and memory-retrieval projections are trainable, while the base BioMistral-7B decoder and CXR-CLIP vision encoder remain frozen. We train Stage 2 with backpropagation through time (BPTT) [12] to a depth of 2, use a micro-batch size of 1, and optimize the hierarchical-memory parameters for 8,000 steps with a learning rate of 5×10^{-6}. All hierarchical-memory-related hyperparameters follow the sensitivity analysis in [2], unless otherwise noted. All experiments are run on a single NVIDIA A10 GPU.

We report natural language generation (NLG) similarity and clinical efficacy (CE) metrics. For the main comparison, we report average BLEU [13], ROUGE-L [7], and METEOR [1] between generated and reference reports. For clinical accuracy, we compute the micro-average of Precision (P), Recall (R), and F1-score (F1) of CheXbert labelling [14], F1-score of RadGraph [4], which is the overlap in radiological entities and clinical relations.

Results We compare HM-RRG with recent MIMIC-CXR report generation methods across three input types: single-image (SI), longitudinal history (Long), and multi-view longitudinal (MVL). Our HM-RRG models use both retrieved reports and longitudinal history (R&L). As shown in Table 1, HM-RRG outperforms prior methods across these input settings on clinical accuracy metrics, demonstrating its ability to generate clinically meaningful reports by effectively leveraging both retrieved and longitudinal context via hierarchical memory. Figure 2 shows the HM-RRG generated report with one prior report and one retrieved report used as evidence. This example demonstrates that HM-RRG captures disease progression from the prior report and selectively extracts relevant information while filtering noise from the retrieved report.

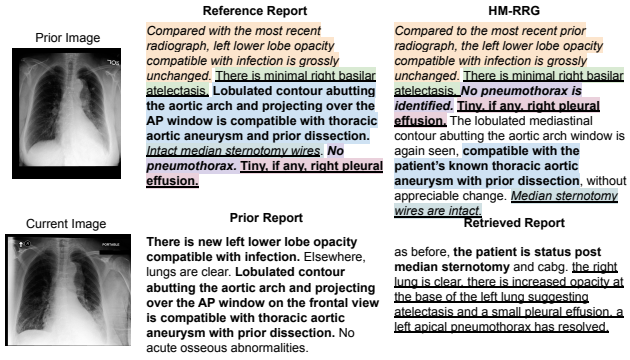

Fig. 2. HM-RRG generated report. Matching findings between the reference and generated reports are indicated with the same colors and text styles. Relevant and irrelevant evidence are marked in bold and underlined, respectively.

Table 2. Ablation results on the MIMIC-CXR dataset.

Model	BLEU	ROUGE-L	BERTScore	F1-RadGraph	F1-CheXbert
Baseline	0.175	0.215	0.399	0.214	0.463
+ Prior Reports	0.227	0.267	0.434	0.225	0.507
+ top-1 Retr.	0.171	0.221	0.413	0.217	0.539
+ top-5 Retr.	0.176	0.224	0.415	0.242	0.500
+ Prior & top-5 Retr.	0.177	0.220	0.416	0.211	0.483
w/ HM-RRG	**0.263**	**0.292**	**0.521**	**0.299**	**0.580**

Ablation Results To test whether hierarchical memory is necessary for using rich context, we conduct ablation studies and report the results in Table 2. For the ablation study, we additionally report BERTScore [22] to assess semantic similarity between generated and reference reports. We compare a baseline that uses only the current image with an instruction prompt, adding longitudinal context, adding top-1 or top-5 retrieved reports, a flat combination of priors and retrieval, and the full HM-RRG. Adding prior reports consistently improves NLG and clinical accuracy, demonstrating that longitudinal history provides useful cues for disease progression. In contrast, retrieval-only inputs may introduce irrelevant clinical context, leading to mixed results across metrics. A key finding is that simply concatenating longitudinal and retrieved context does not reliably improve performance. This suggests that when the context becomes long and noisy, the model struggles to focus on the right evidence, and important signals can be diluted. HM-RRG resolves this issue by converting unordered retrieved reports into current-image-conditioned memory vectors and allowing each longitudinal segment to query a compact unified memory, making evidence selection explicit rather than relying on the decoder to resolve a long, noisy sequence. Overall, these results confirm the need for hierarchical memory: when both longitudinal

and retrieved contexts are available, hierarchical memory can selectively leverage useful evidence while suppressing irrelevant content.

4 Conclusion

In this paper, we propose HM-RRG, a hierarchical memory framework for radiology report generation that stores, organizes, and selects key evidence from both cross-patient retrieval and within-patient longitudinal history. Unlike prior approaches that rely on flat context concatenation, HM-RRG leverages structured memory and adaptive evidence selection to better leverage rich clinical context and generate clinically meaningful reports. The hierarchical memory structure and decoding mechanism are model-independent and can be readily applied to other LLM and VLM decoders with lightweight training overhead. Future work can extend HM-RRG to richer retrieval settings, longer longitudinal histories, external clinical datasets, and additional decoder backbones, further expanding its applicability to diverse medical settings.

Acknowledgments. The authors thank Professor Jason Cong at the University of California, Los Angeles, for his valuable advice.

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

References

1. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein, J., Lavie, A., Lin, C.Y., Voss, C. (eds.) Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005), https://aclanthology.org/W05-0909/
2. He, Z., Cao, Y., Qin, Z., Prakriya, N., Sun, Y., Cong, J.: HMT: Hierarchical memory transformer for efficient long context language processing. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 8068–8089. Association for Computational Linguistics (2025)
3. 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: International Conference on Learning Representations (ICLR) (2022)
4. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Nguyen Duong, D., Bui, T., Chambon, P., Lungren, M., Ng, A., Langlotz, C., Rajpurkar, P.: RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. PhysioNet (2021)
5. Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., Horng, S.: MIMIC-CXR-JPG - chest radiographs with structured labels. PhysioNet (2024)

6. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M., Dufour, R.: BioMistral: A collection of open-source pretrained large language models for medical domains. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 5848–5864. Association for Computational Linguistics, Bangkok, Thailand (2024)
7. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics (2004)
8. Liu, C., Tian, Y., Chen, W., Song, Y., Zhang, Y.: Bootstrapping large language models for radiology report generation. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'24/IAAI'24/EAAI'24, AAAI Press (2024). https://doi.org/10.1609/aaai.v38i17.29826
9. 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: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10348–10359 (2025)
10. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Comput. Surv. (2023)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
12. Mozer, M.C.: A focused backpropagation algorithm for temporal pattern recognition, p. 137–169. L. Erlbaum Associates Inc., USA (1995)
13. 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 on Association for Computational Linguistics. p. 311–318. Association for Computational Linguistics, USA (2002)
14. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A., Lungren, M.: Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1500–1519. Association for Computational Linguistics, Online (2020)
15. Wang, F., Du, S., Yu, L.: Hergen: Elevating radiology report generation with longitudinal data. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LV. p. 183–200. Springer-Verlag, Berlin, Heidelberg (2024)
16. 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 p. 103627 (2025)
17. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. Meta-Radiology p. 100033 (2023)
18. Xu, G., Li, X., Chen, Y., Duan, Y., Wu, S., Yu, H., Chiu, C.H., Ni, J., Tang, N., Li, T.J.J., Yuille, A., Jin, W., Shi, Y.: A comprehensive survey of ai agents in healthcare. Journal of Biomedical Informatics **179**, 105045 (2026). https://doi.org/https://doi.org/10.1016/j.jbi.2026.105045, https://www.sciencedirect.com/science/article/pii/S1532046426000699
19. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: proceedings

of Medical Image Computing and Computer Assisted Intervention. p. 101–111. Springer-Verlag (2023)
20. Yun, H., Maeng, J., Kang, E., Suk, H.I.: Diff-RRG: Longitudinal Disease-wise Patch Difference as Guidance for LLM-based Radiology Report Generation . In: proceedings of Medical Image Computing and Computer Assisted Intervention (2025)
21. Zambrano Chaves, J.M., Huang, S.C., et al.: A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. Nature Communications p. 3108 (2025)
22. Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2020)
23. Zhu, Q., Mathai, T.S., Mukherjee, P., Peng, Y., Summers, R.M., Lu, Z.: Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. proceedings of Medical Image Computing and Computer Assisted Intervention pp. 189–198 (2023)