

A Survey Of AI-Based Radiology Report Generation and Clinical Consistency Verification

Nithya R¹, Dhanu Sri S²

¹Associate Professor, Department of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Tamilnadu, India

²PG Scholar, Department of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Tamilnadu, India

Abstract—Radiology report generation (RRG) aims to translate medical images into fluent, clinically meaningful narrative reports, easing radiologist workload and shortening reporting delays. Progress from CNN–RNN captioning to memory-driven transformers, knowledge-enhanced networks and radiology-specific multimodal large language models has steadily improved lexical overlap with reference reports. However, generated reports still contain hallucinated findings, omitted abnormalities, wrong laterality or severity, and invented references to prior studies, errors that n-gram metrics barely register. This survey examines AI-driven radiology report generation (RRG) and explores techniques designed to improve the factual reliability of generated reports by evaluating their alignment with medical images and reference reports. We propose a taxonomy of generation paradigms and verification mechanisms, summarise public datasets (IU X-ray, MIMIC-CXR, CheXpert Plus, PadChest, CT-RATE) and evaluation metrics, and compare published results on MIMIC-CXR and IU X-ray. We close with open challenges, including dataset bias, evaluation validity, hallucination, longitudinal reasoning and clinical deployment, and outline research directions toward verifiable, radiologist-in-the-loop report generation.

Index Terms—Radiology Report Generation, Clinical Consistency Verification, Hallucination, Vision–Language Models, Large Language Models, Chest X-ray, Evaluation Metrics.

I. INTRODUCTION

Chest radiography is among the most frequently performed imaging examinations, and every examination ends with a free-text report describing findings and an impression. Writing these reports is

slow and repetitive, and growing imaging volumes with radiologist shortages can delay reporting. Radiology report generation (RRG) produces a draft report from one or more images for a radiologist to review and sign. It builds on image captioning [1], [2] but is harder: reports are long and domain-specific, most text describes normal anatomy so abnormal findings are rare, and one wrong word (“left” for “right”, or a dropped “no”) can reverse the clinical meaning.

Early encoder–decoder systems [3], [4] were followed by memory-driven transformers [5], [6], knowledge-enhanced models [7], [8] and radiology-specific multimodal large language models (LLMs) [9], [10], [11]. Yet rising lexical scores do not guarantee clinical safety: overlap metrics correlate only partially with radiologist-counted errors [12], and fluent reports may assert findings the image does not show (hallucination [13]) or omit critical ones. This has made clinical consistency verification, checking that a report is supported by the image and consistent with the ground truth, an active area spanning RL rewards [14], [15], labeler and graph metrics [16], [17], LLM judges [18], [10] and region grounding [19], [10].

Earlier surveys [20], [21] focus on architectures, datasets and explainability, and treat factual-consistency verification as a side topic. Our contributions are: (1) a taxonomy coupling generation and verification; (2) a structured review of 2015–2025 methods; (3) a view of datasets and metrics that stresses what each can miss; and (4) a comparison of published results with challenges and future directions. Section II describes the methodology, III the background, IV the literature, V datasets and

metrics, VI the comparative analysis, VII challenges, VIII future directions and IX concludes.

II. SURVEY METHODOLOGY

The study adopts a narrative review approach supported by a predefined search strategy; therefore, PRISMA-based study selection statistics are not reported. It addresses four questions: RQ1 which model families exist for RRG; RQ2 how clinical consistency is verified; RQ3 which datasets and metrics are used and what they miss; RQ4 what published results show and what is unsolved.

We searched Google Scholar, IEEE Xplore, PubMed, ACM Digital Library, arXiv, ACL Anthology and CVF Open Access for 2015–2025 using combinations of “radiology/medical report generation”, “chest X-ray report”, “vision–language model”, “factual consistency”, “hallucination”, “clinical efficacy” and “report evaluation metric”, then followed reference lists of earlier surveys [20], [21]. We included papers and widely cited preprints that generate reports from images or propose a metric, reward, dataset or procedure for judging report factuality, and excluded work limited to classification or visual question answering. Scores in Section VI are quoted from original papers, which use different splits, report sections and label sets, so they are not a controlled benchmark. Chest radiography is over-represented in the literature and hence here.

III. BACKGROUND AND PRELIMINARIES

Reports and task.

A chest radiography report typically includes the clinical reason for the examination, a review of previous images when available, observations of relevant anatomical structures, and a concluding summary of the main findings. Most work targets the findings section; some also use the indication and prior report [10], [22].

Let X be the images and C optional context. Teacher-forced training minimises

$$L_{CE} = -\sum_t \log P(y_t | y_{<t}, X, C; \theta) \quad (1)$$

while RL variants maximise a reward $R(\hat{Y}, Y)$ [4], [14]. A verifier $V(\hat{Y}, X, Y)$ returns a score, an error list or a corrected report, either as a training reward or an inference-time check.

Building blocks.

Visual feature extraction has evolved from convolutional neural networks (CNNs) toward vision transformer architectures [23], with several models leveraging pre-training techniques based on image–text contrastive learning [24], [25], [26]. Decoders moved from LSTMs to Transformers [27] and LLMs receiving visual tokens [28], [9]. Add-ons include memory banks [5], [6], knowledge graphs [29], [7], retrieval [4], [30] and region detectors [19].

Consistency and error types.

Factual completeness measures the extent to which a generated report covers the findings documented in the reference, whereas factual consistency evaluates whether the statements in the report are supported by the available evidence [14]. Errors include (i) false-positive findings (hallucination), (ii) omissions, (iii) attribute errors (location, laterality, severity), (iv) negation or certainty errors, (v) comparison errors such as invented priors [31], and (vi) contradictions between findings and impression.

IV. LITERATURE REVIEW AND CLASSIFICATION OF EXISTING METHODS

A. Retrieval, Template and Encoder–Decoder Models
HRGR-Agent [4] uses RL to choose between retrieving a template sentence and generating a new one; KERP [30] retrieves and paraphrases text for abnormalities encoded in a graph; Liu et al. [32] retrieve and customise similar reports from finding patterns. Jing et al. [3] adapted captioning with tag prediction, co-attention and a hierarchical LSTM. These produce fluent normal text but handle rare findings poorly.

B. Transformer and Memory-Driven Models

R2Gen [5] adds relational memory and memory-driven normalisation to a Transformer; CMN [6] shares a memory matrix across modalities; METransformer [33] votes over learnable expert tokens; and CvT2DistilGPT2 [34] shows that warm-starting from pre-trained vision and language models beats training from scratch.

Figure 1 presents a classification of the reviewed studies based on report generation strategies (A–E) and the methods used to assess consistency (F).

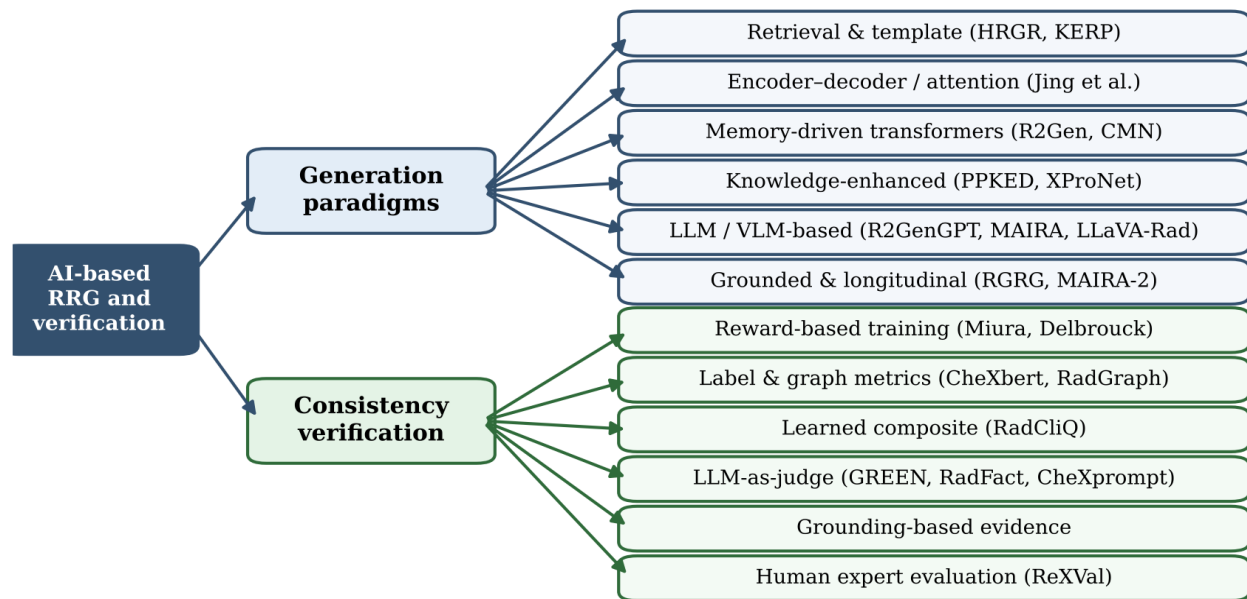


Fig. 1. Taxonomy of AI-based radiology report generation (blue) and clinical consistency verification (green).

C. Knowledge-Enhanced Models

Zhang et al. Knowledge-guided approaches have been explored to improve radiology report generation. The method in [29] incorporates disease relationships using a structured medical knowledge graph. PPKED [7] integrates information about detected abnormalities with prior knowledge from retrieved reports and medical sources, whereas XProNet [8] develops representative features to establish connections between visual and textual information. Scene graphs such as Chest ImaGenome [35] supply region-level labels.

D. LLM- and VLM-Based Models

R2GenGPT [28] aligns a frozen LLM to visual features. MAIRA-1 [9] pairs a radiology image encoder with Vicuna. Med-PaLM M [36] and RadFM [37] are generalist models, while CheXagent [38] and LLaVA-Rad [11] specialise in chest X-rays, the latter showing that a 7B open model is competitive. PromptMRG [39] adds diagnosis-driven prompts, and retrieval-augmented methods [40] give the LLM similar reports as evidence.

E. Grounded and Longitudinal Models

RGRG [19] generates a sentence per detected anatomical region, making each statement traceable;

MAIRA-2 [10] outputs sentence with bounding boxes. BioViL-T [41] and Dalla Serra et al. [22] use prior studies, which makes reports realistic but invites invented comparisons [31].

F. Clinical Consistency Verification

Reward-based methods optimise entity-overlap and entailment rewards [14] or RadGraph-based semantic rewards [15]. *Labeler and graph checks* compare 14 CheXpert-style observations [42], [16] or RadGraph entities and relations [17]. *Composite* RadCliQ [12] regresses four metrics onto radiologist error counts. *LLM judges* such as GREEN [18], RadFact [10], CheXprompt [11] and FineRadScore [43] list errors or test each sentence against a reference. *Grounding* [19], [10] makes consistency visually checkable. *Data cleaning* removes references to non-existent priors from training targets [31]. *Human evaluation* remains the gold standard but is costly.

V. DATASETS AND EVALUATION METRICS

MIMIC-CXR [44] is the de facto benchmark and IU X-ray [45] a small, widely used second check (Table I). CheXpert [42] and ChestX-ray14 [46] provide report-mined labels, CheXpert Plus [47] adds report text, and CT-RATE [48] offers 3D CT with reports.

TABLE I. PUBLIC DATASETS (SIZES AS REPORTED IN [21])

Dataset	Modality	Size	Notes
IU X-ray [45]	Chest X-ray	7,470 images; 3,955 reports	Small; frontal and lateral
MIMIC-CXR [44]	Chest X-ray	377,110 images; ~227,800 reports; 65,379 patients	Single hospital; main benchmark
CheXpert [42]	Chest X-ray	224,316 images	14 uncertainty labels; extended by [47]
PadChest [49]	Chest X-ray	160,868 images; 109,931 studies	Spanish reports
ChestX-ray14 [46]	Chest X-ray	112,120 images; 30,805 patients	NLP-mined labels only
CT-RATE [48]	Chest CT (3D)	Paired volumes and reports	First large volumetric set

Lexical metrics.

BLEU [50], ROUGE-L [51], METEOR [52] and CIDEr [53] reward word overlaps with one reference and are blind to negation; BERTScore [54] uses embedding similarity but is not clinically aware. Clinical efficacy applies a labeler [42], [16] to both reports and computes precision P , recall R and

$$F1 = 2PR / (P + R) \quad (2)$$

as micro, macro or example averages. RadGraph F1 [17] compares entities and relations; RaTEScore [55] weights entities by clinical importance; RadCliQ [12] and the LLM judges above are described in Section IV-F. The ReXrank leaderboard [56] scores systems

on eight metrics (BLEU-2, BERTScore, SembScore, RadGraph F1, 1/RadCliQ-v1, RaTEScore, GREEN, 1/FineRadScore) over ReXGradient, MIMIC-CXR, IU X-ray and CheXpert Plus.

VI. COMPARATIVE ANALYSIS

Table II quotes lexical scores reported by the original authors. From R2Gen to R2GenGPT, MIMIC-CXR BLEU-4 rises from 0.103 to 0.134 and METEOR from 0.142 to 0.160: steady but modest progress. IU X-ray is small with differing splits, so gaps of a point or two should not be over-interpreted

TABLE II. REPORTED LEXICAL RESULTS (B-1/B-4: BLEU; R-L: ROUGE-L; M: METEOR; “–” NOT REPORTED)

Model	MIMIC-CXR B-1	B-4	R-L	M	IU X-ray B-1	B-4	R-L	M
R2Gen [5]	0.353	0.103	0.277	0.142	0.470	0.165	0.371	0.187
CMN [6]	0.353	0.106	0.278	0.142	0.475	0.170	0.375	0.191
PPKED [7]	0.360	0.106	0.284	0.149	0.483	0.168	0.376	0.190
METransformer [33]	0.386	0.124	0.291	0.152	0.483	0.172	0.380	0.192
CvT2DistilGPT2 [34]	0.393	0.127	–	0.155	0.473	0.175	0.376	0.200
R2GenGPT [28]	0.411	0.134	0.297	0.160	0.488	0.173	0.377	0.211

Clinical metrics for recent multimodal models on MIMIC-CXR Findings, as compiled in [40], rank them differently: LLaVA-Rad [11] reaches BLEU-4 15.4 and ROUGE-L 30.6 and MAIRA-1 [9] CE micro-F1 55.7, but RA-RRG [40] has the best CE micro-F1 (58.5) and macro-F1 (41.7) despite a BLEU-4 of only 8.0, a clear case of lexical and clinical rankings disagreeing. Med-PaLM M [36] reports CE micro-F1 53.6 and RadGraph F1 26.7.

On the uniform ReXrank leaderboard [56], MedVersa ranks first on MIMIC-CXR (Findings plus Impression) with 1/RadCliQ-v1 of 0.92, RadGraph F1

of 0.27 and GREEN of 0.42, ahead of the next models (0.83). Overall: (1) gains have shifted from lexical overlap to clinical content, but the two are not interchangeable;

(2) evidence from knowledge, retrieval, priors or grounding generally helps clinical metrics; (3) no single verifier suffices, since labelers miss location and severity, graph metrics depend on extraction, and LLM judges are costly and may be biased; (4) cross-paper comparison is unreliable, which is why shared leaderboards matter.

VII. RESEARCH CHALLENGES AND LIMITATIONS

Data. Normal text dominates, so models emit high-probability normal sentences; data is mostly English, single-institution and chest X-ray, and labelers add noise [16]. Evaluation. Overlap metrics are weakly tied to clinical risk, CE metrics cover few conditions, LLM judges are hard to reproduce and often rely on commercial APIs, and metric–radiologist correlations were measured on small samples [12]. Hallucination and omission. Fluent LLM output need not be image-supported [13]; omissions are harder to measure than false claims, and models invent prior comparisons [31]. Grounding. Region labels are costly or silver-standard [35], and a box does not prove the model reasoned from that region. Longitudinal, multi-view and non-chest data are handled inconsistently [41], [22], [48]. Deployment. Confidence is rarely calibrated, clinicians may over-trust fluent drafts, and privacy, fairness, regulation and compute remain open; reader studies are encouraging [38] but prospective trials are scarce.

VIII. FUTURE RESEARCH DIRECTIONS

Generate–verify–revise loops: decompose drafts into atomic claims, check each against image, priors and knowledge, and rewrite unsupported claims, using the verifier as both training reward [14], [15] and inference-time filter. Evidence-linked reporting: attach regions, retrieved cases or priors to every sentence [19], [10]. Calibrated uncertainty and abstention in conventional hedging language. Shared, clinically validated benchmarks [56] with multi-site data, radiologist error labels, confidence intervals and per-finding results. Longitudinal, multimodal and multilingual models, including CT and MRI [48]. Learning from radiologist edits via human-feedback RL. Privacy-preserving, fair and compact models [11], and prospective human–AI workflow studies measuring accuracy, time and automation bias.

IX. CONCLUSION

Report generators have evolved from CNN–RNN captioners to memory-driven transformers, knowledge-enhanced networks and radiology-specific multimodal LLMs, with grounded and longitudinal

variants supplying evidence and context. Published lexical scores have improved modestly, and several results show lexical and clinical rankings diverging. Various techniques have been proposed to assess the reliability of AI-generated radiology reports, including entity-based evaluation, textual entailment analysis, clinical label comparison, knowledge-graph assessment, combined scoring methods, evaluations by large language models, and image-grounding techniques. Although each approach can identify certain types of errors, expert radiologist review remains essential for ensuring clinical accuracy. For safe and effective clinical implementation, radiology report generation systems should incorporate iterative report refinement, evidence-based verification, and reliable uncertainty estimation. Their performance should also be assessed using standardized benchmarks validated for clinical use and prospective studies involving radiologists in the evaluation process.

REFERENCES

- [1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3156–3164.
- [2] K. Xu *et al.*, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2048–2057.
- [3] B. Jing, P. Xie, and E. Xing, “On the automatic generation of medical imaging reports,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2018, pp. 2577–2586.
- [4] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, “Hybrid retrieval-generation reinforced agent for medical image report generation,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 1530–1540.
- [5] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, “Generating radiology reports via memory-driven transformer,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 1439–1449.
- [6] Z. Chen, Y. Shen, Y. Song, and X. Wan, “Cross-modal memory networks for radiology report generation,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics–Int. Joint Conf. Natural*

- Lang. Process. (ACL-IJCNLP), 2021, pp. 5904–5914.
- [7] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, “Exploring and distilling posterior and prior knowledge for radiology report generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 13753–13762.
- [8] J. Wang, A. Bhalerao, and Y. He, “Cross-modal prototype driven network for radiology report generation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2022, pp. 563–579.
- [9] S. L. Hyland et al., “MAIRA-1: A specialised large multimodal model for radiology report generation,” arXiv preprint arXiv:2311.13668, 2023.
- [10] S. Bannur et al., “MAIRA-2: Grounded radiology report generation,” arXiv preprint arXiv:2406.04449, 2024.
- [11] J. M. Zambrano Chaves et al., “Towards a clinically accessible radiology foundation model: Open-access and lightweight, with automated evaluation,” arXiv preprint arXiv:2403.08002, 2024, doi: 10.1038/s41467-025-58344-x.
- [12] F. Yu et al., “Evaluating progress in automatic chest X-ray radiology report generation,” Patterns, vol. 4, no. 9, Art. no. 100802, 2023.
- [13] Z. Ji et al., “Survey of hallucination in natural language generation,” ACM Comput. Surv., vol. 55, no. 12, Art. no. 248, 2023.
- [14] Y. Miura, Y. Zhang, E. Tsai, C. Langlotz, and D. Jurafsky, “Improving factual completeness and consistency of image-to-text radiology report generation,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT), 2021, pp. 5288–5304.
- [15] J.-B. Delbrouck et al., “Improving the factual correctness of radiology report generation with semantic rewards,” in Findings of the Association for Computational Linguistics: EMNLP, 2022, pp. 4348–4360.
- [16] A. Smit et al., “CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 1500–1519.
- [17] S. Jain et al., “RadGraph: Extracting clinical entities and relations from radiology reports,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Datasets and Benchmarks Track, 2021.
- [18] S. Ostmeier et al., “GREEN: Generative radiology report evaluation and error notation,” in Findings of the Association for Computational Linguistics: EMNLP, 2024, arXiv:2405.03595.
- [19] E. Tanida, P. Müller, G. Kaissis, and D. Rueckert, “Interactive and explainable region-guided radiology report generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 7433–7442.
- [20] P. Messina et al., “A survey on deep learning and explainability for automatic report generation from medical images,” ACM Comput. Surv., vol. 54, no. 10s, Art. no. 203, 2022.
- [21] P. Sloan, P. Clatworthy, E. Simpson, and M. Mirmehdi, “Automated radiology report generation: A review of recent advances,” IEEE Rev. Biomed. Eng., arXiv:2405.10842.
- [22] F. Dalla Serra, C. Wang, F. Deligianni, J. Dalton, and A. Q. O’Neil, “Controllable chest X-ray report generation from longitudinal representations,” in Findings of the Association for Computational Linguistics: EMNLP, 2023.
- [23] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
- [24] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 8748–8763.
- [25] B. Boecking et al., “Making the most of text semantics to improve biomedical vision–language processing,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2022, pp. 1–21.
- [26] E. Tiu et al., “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning,” Nature Biomedical Engineering, vol. 6, pp. 1399–1406, 2022.
- [27] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.
- [28] Z. Wang, L. Liu, L. Wang, and L. Zhou, “R2GenGPT: Radiology report generation with frozen LLMs,” Meta-Radiology, vol. 1, no. 3, Art. no. 100033, 2023.
- [29] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, “When radiology report generation meets knowledge graph,” in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 7, 2020, pp. 12910–12917.

- [30] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in Proc. AAAI Conf. Artif. Intell., 2019, pp. 6666–6673.
- [31] V. Ramesh, N. A. Chi, and P. Rajpurkar, "Improving radiology report generation systems by removing hallucinated references to non-existent priors," in Proc. Mach. Learn. Health (ML4H), PMLR, vol. 193, 2022.
- [32] G. Liu et al., "Clinically accurate chest X-ray report generation," in Proc. Mach. Learn. Healthcare Conf. (MLHC), PMLR, vol. 106, 2019, pp. 249–269.
- [33] Z. Wang, L. Liu, L. Wang, and L. Zhou, "METransformer: Radiology report generation by transformer with multiple learnable expert tokens," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 11558–11567.
- [34] A. Nicolson, J. Dowling, and B. Koopman, "Improving chest X-ray report generation by leveraging warm starting," Artificial Intelligence in Medicine, vol. 144, Art. no. 102633, 2023.
- [35] J. T. Wu et al., "Chest ImaGenome dataset for clinical reasoning," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Datasets and Benchmarks Track, 2021.
- [36] T. Tu et al., "Towards generalist biomedical AI," NEJM AI, vol. 1, no. 3, 2024.
- [37] C. Wu et al., "Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data," arXiv preprint arXiv:2308.02463, 2023.
- [38] Z. Chen et al., "A vision-language foundation model to enhance efficiency of chest X-ray interpretation," arXiv preprint arXiv:2401.12208, 2024.
- [39] H. Jin et al., "PromptMRG: Diagnosis-driven prompts for medical report generation," in Proc. AAAI Conf. Artif. Intell., vol. 38, no. 3, 2024, pp. 2607–2615.
- [40] S. Park et al., "RA-RRG: Multimodal retrieval-augmented radiology report generation with key phrase extraction," arXiv preprint arXiv:2504.07415, 2025.
- [41] S. Bannur et al., "Learning to exploit temporal structure for biomedical vision-language processing," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 15016–15027.
- [42] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proc. AAAI Conf. Artif. Intell., 2019, pp. 590–597.
- [43] A. Huang et al., "FineRadScore: A radiology report line-by-line evaluation technique generating corrections with severity scores," arXiv preprint arXiv:2405.20613, 2024.
- [44] A. E. W. Johnson et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," Scientific Data, vol. 6, Art. no. 317, 2019.
- [45] D. Demner-Fushman et al., "Preparing a collection of radiology examinations for distribution and retrieval," Journal of the American Medical Informatics Association, vol. 23, no. 2, pp. 304–310, 2016.
- [46] X. Wang et al., "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 2097–2106.
- [47] P. Chambon 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.
- [48] I. E. Hamamci et al., "Developing generalist foundation models from a multimodal dataset for 3D computed tomography," arXiv preprint arXiv:2403.17834, 2024.
- [49] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest X-ray image dataset with multi-label annotated reports," Medical Image Analysis, vol. 66, Art. no. 101797, 2020.
- [50] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL), 2002, pp. 311–318.
- [51] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Text Summarization Branches Out, 2004, pp. 74–81.
- [52] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with

- improved correlation with human judgments,” in Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, pp. 65–72.
- [53] R. Vedantam, C. L. Zitnick, and D. Parikh, “CIDEr: Consensus-based image description evaluation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 4566–4575.
- [54] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating text generation with BERT,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.
- [55] W. Zhao *et al.*, “RaTEScore: A metric for radiology report generation,” *arXiv preprint arXiv:2406.16845*, 2024.
- [56] X. Zhang *et al.*, “ReXrank: A public leaderboard for AI-powered radiology report generation,” *arXiv preprint arXiv:2411.15122*, 2024.