# Precision Recall Controllable Radiology Report Generation via Hybrid Natural Language and Clinical Reward Learning

Ling Chen[1], Ruinan Jin[1], Jun Luo[1], Hanliang Chen[1], Quirin Strotzer[2], Rongkai Yan[1], Yuan Xue[1], Luciano Prevedello[1], and Dufan Wu[1]

[1] The Ohio State University, Columbus, OH 43221, USA
[2] University Medical Center Regensburg, Regensburg, Germany
lingchen.chen@osumc.edu, dufan.wu@osumc.edu

**Abstract.** Automated radiology report generation (RRG) has gained increasing attention because it can reduce the heavy workload of clinical report writing. However, most existing methods mainly optimize for natural language generation (NLG) metrics that focus on language fluency, while providing little control over clinically important factors such as precision and recall. As consequence, generated reports may be fluent but not well aligned with different clinical needs. To address this challenge, we propose a reinforcement learning framework for precision recall controllable RRG, where a control parameter explicitly adjusts the trade-off between clinical precision and recall during inference. This design allows the model to flexibly generate reports according to different clinical requirements. To ensure clinical correctness, we introduce a clinical reward into the training objective, which helps improve clinical efficacy (CE) beyond standard language-based optimization. In addition, we apply a group-relative training strategy that normalizes rewards within each training group, reducing reward variance and improving training stability. Extensive experiments on the MIMIC-CXR dataset show that our method consistently outperforms state-of-the-art approaches in both NLG and CE evaluation metrics, while providing reliable control over the CE precision recall trade-off.

**Keywords:** Radiology Report Generation · Receiver Operation Curve · Reinforcement Learning · Group-Relative Training · Chest X-ray.

## 1 Introduction

Automated radiology report generation (RRG) aims to produce free-text clinical reports directly from medical images, such as chest X-rays, and has received growing attention in recent years. By reducing the time and effort required for manual report writing, these systems have the potential to improve clinical workflow efficiency and support large-scale medical imaging analysis. Most existing approaches treat report generation as a sequence-to-sequence learning problem and are commonly trained using maximum likelihood estimation (MLE) with word-level supervision [1, 2].

While MLE-based methods can generate fluent and grammatically correct reports, they often suffer from exposure bias and show limited alignment with clinical goals [3] [4]. To address these issues, recent studies have introduced reinforcement learning (RL)–based optimization strategies that directly optimize sentence-level evaluation metrics [5–9]. However, most prior work focuses on improving natural language generation (NLG) scores, such as BLEU [10], METEOR [11], and ROUGE-L [12], which mainly measure text similarity and linguistic quality. As a result, generated reports may appear fluent but still contain clinically incorrect, incomplete, or unbalanced findings. More importantly, existing methods provide little control over clinically critical trade-offs during report generation. In real-world practice, different clinical scenarios may require different emphasis on precision and recall [13]. For example, screening exams often favor high recall to avoid missing abnormalities, while diagnostic tests may prioritize high precision to reduce false positives. Current report generation models lack explicit mechanisms to adjust such trade-offs, limiting their practical usability in diverse clinical settings.

In this work, we propose a reinforcement learning framework for precision recall controllable RRG for chest X-rays, in which a continuous control parameter explicitly adjusts the balance between clinical precision and recall at inference time. This design allows the same model to flexibly generate reports that align with different clinical requirements, without retraining or changing decoding strategies. To improve clinical performance, we further introduce a clinical reward explicitly supervises the presence and absence of clinically relevant findings, enhancing clinical efficacy beyond language-only optimization. In addition, reinforcement learning for report generation often suffers from high reward variance and unstable training [14]. To address this challenge, we adopt a group-relative training strategy [15] that normalizes rewards within each sample group, effectively reducing variance and stabilizing policy optimization. This strategy improves learning efficiency and leads to more consistent performance gains across datasets. We evaluate the proposed method on MIMIC-CXR [16] dataset. Experimental results show that our approach consistently outperforms state-of-the-art methods in both language quality and clinically oriented evaluation metrics, while providing reliable and interpretable control over the precision recall trade-off.

## 2 Method

The overall framework is illustrated in Fig. 1, which consists of the overall architecture (Fig. 1(a)), the reinforcement learning process (Fig. 1(b)), and the transformer encoder details (Fig. 1(c)).

### 2.1 Precision Recall Controllable Reinforcement Learning

We formulate RRG as a reinforcement learning problem, where the model generates a report sequence $Y = (y_1, \ldots, y_T)$ conditioned on an input image $X$.

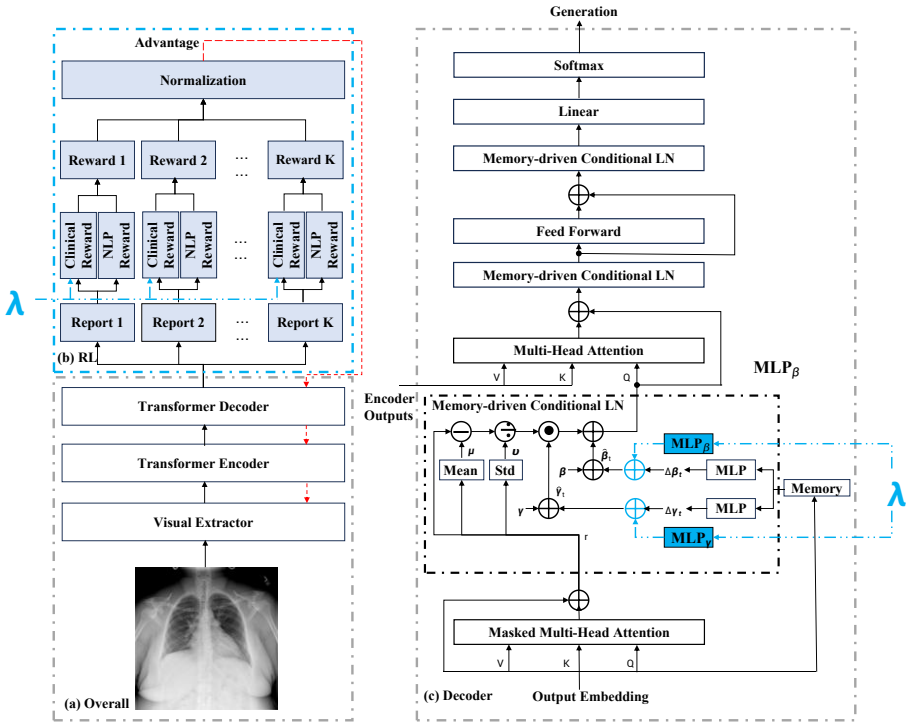



**Fig. 1.** Overall framework of the proposed precision recall controllable RRG model. (a) Overall architecture. (b) Reinforcement learning process. (c) Transformer decoder. The overall framework and decoder are shown in gray dashed boxes with details omitted. Red dotted lines denote gradient back-propagation during training. Blue dotted lines indicate the precision recall control parameter $\lambda$ applied at both the representation and reward levels. The group-relative training component is highlighted in blue dashed boxes.

To explicitly control the precision recall trade-off, we introduce a continuous control parameter $\lambda \in [0, 1]$ that is consistently applied at both the representation level and the reward level. This unified design ensures coherent controllability throughout the generation and optimization process.

At the representation level, inspired by conditional normalization techniques such as FiLM [17], we introduce a Precision Recall Conditioned Adaptive Layer Normalization (PRC-AdaLN) module to incorporate the clinical control parameter $\lambda$ into the decoder representations. Specifically,

$$\text{PRC-AdaLN}(x, \lambda) = \big(\gamma + \Delta\gamma + \text{MLP}_\gamma(\lambda)\big) \odot \frac{x - \mu(x)}{\sigma(x)} + \big(\beta + \Delta\beta + \text{MLP}_\beta(\lambda)\big), \quad (1)$$

where $x$ denotes the decoder feature, $\mu(x)$ and $\sigma(x)$ are the mean and standard deviation computed along the feature dimension, and $\gamma$ and $\beta$ are the learnable affine parameters in standard LayerNorm. The terms $\Delta\gamma$ and $\Delta\beta$ denote

additional residual modulation parameters, while two lightweight MLPs, $\mathrm{MLP}_\gamma$ and $\mathrm{MLP}_\beta$, map the clinical control parameter $\lambda$ to separate conditioning offsets for the normalization scale and bias, respectively. As illustrated in Fig. 1(c), this representation level modulation dynamically adjusts feature statistics, enabling the decoder to explicitly steer generation along the clinically meaningful precision recall axis.

At the reward level, $\lambda$ further controls the optimization objective by balancing precision- and recall-oriented clinical efficacy (CE) rewards:

$$R_{\mathrm{CE}}(\lambda) = \lambda\, R_{\mathrm{prec}} + (1-\lambda)\, R_{\mathrm{rec}}. \tag{2}$$

During training, for each training sample, the representation modulation and reward optimization shared a random $\lambda$ sampled uniformly from $[0, 1]$. Thus the model learns a family of conditional policies with consistent precision recall characteristics. At inference time, adjusting $\lambda$ allows a single trained model to flexibly generate reports that emphasize either higher precision or higher recall, without retraining or modifying the decoding strategy.

### 2.2 Hybrid Natural Language and Clinical Reward Learning

We optimize the report generation model using a hybrid training objective that combines reinforcement learning with standard cross-entropy loss:

$$\mathcal{L}_{\mathrm{total}} = \alpha \mathcal{L}_{\mathrm{rl}} + (1-\alpha)\mathcal{L}_{\mathrm{ce}}, \tag{3}$$

where $\alpha$ balances reinforcement learning loss and cross-entropy loss.

For the reinforcement learning component, we minimize the negative expected reward:

$$\mathcal{L}_{\mathrm{rl}} = -\mathbb{E}_{Y \sim p_\theta}[R(Y)], \tag{4}$$

where $\theta$ refers to the model parameters and the reward function $R(Y)$ integrates both natural language generation (NLG) quality and CE:

$$R(Y) = \rho R_{\mathrm{NLG}}(Y) + (1-\rho) R_{\mathrm{CE}}(Y). \tag{5}$$

Here, $\rho \in [0, 1]$ is a fixed hyperparameter that controls the trade-off between natural language quality and clinical efficacy. The clinical reward is based on the clinical efficacy metric, with $R_{\mathrm{CE}}(Y)$ computed according to Eq. 2. The natural language generation reward $R_{\mathrm{NLG}}(Y)$ is defined as:

$$R_{\mathrm{NLG}}(Y) = \sum_i \omega_i r_i(Y), \tag{6}$$

where $r_i(Y)$ represents multiple NLG evaluation metrics, including BLEU-4, METEOR, and ROUGE-L [7], and $\omega_i$ are non-negative weighting coefficients with $\sum_i \omega_i = 1$.

### 2.3 Group-Relative Training

Reinforcement learning for report generation often suffers from high reward variance, which can lead to unstable training, especially when multiple reward components are involved. To alleviate this issue, we adopt a group-relative self-critical sequence training strategy [15].

For each input image $x$, we perform group sampling by drawing $K$ candidate reports $\{Y^{(1)}, Y^{(2)}, \ldots, Y^{(K)}\}$ from the current policy. Each sampled report is evaluated using the task-specific reward function, yielding a set of rewards $\{R^{(k)}\}_{k=1}^{K}$.

Instead of relying on absolute reward values, we compute a group-relative baseline by normalizing rewards within each group:

$$\hat{A}^{(k)} = \frac{R^{(k)} - \text{mean}(\{R^{(i)}\}_{i=1}^{K})}{\text{std}(\{R^{(i)}\}_{i=1}^{K}) + \epsilon}, \tag{7}$$

where $\hat{A}^{(k)}$ denotes the standardized advantage of the $k$-th sampled report and $\epsilon$ is a small constant for numerical stability. The reinforcement learning loss is computed according to Eq. 4 by replacing $R(Y)$ with $\hat{A}$ for each sample. We then use the REINFORCE algorithm [18] to compute policy gradients and update the model parameters.

By using the group-relative standardized advantage as the learning signal, the model focuses on the relative quality of candidate reports generated for the same image. This strategy effectively reduces reward variance and stabilizes policy optimization.

## 3 Experiments

**Datasets and Evaluation Metrics** We evaluate our model on the MIMIC-CXR dataset [16], a large-scale chest X-ray benchmark containing 473,057 radiographs paired with 206,563 reports. Following the preprocessing protocol in [2], we adopt the official patient-level split with 270,790 training, 2,130 validation, and 3,858 testing samples. Each study is paired with its corresponding report as ground truth. We only used the findings section for each report.

Following prior work [7, 19], we evaluate model performance using a combination of natural language generation (NLG) metrics and CE metrics. The NLG metrics include BLEU [10], METEOR [11], and ROUGE-L [12]. For CE evaluation, generated and ground-truth reports are first converted into 14 CheXbert labels [20]. Following our training setup, uncertain findings are considered positive, while negative and blank labels are considered negative. We then report micro-averaged precision, recall, and F1 score across all labels and test samples.

**Implementation Details** We used $\alpha = 0.99$ in Eq. 3 that balances RL and cross-entropy loss, $\rho = 0.1$ in Eq. 5 that balances NLG and clinical efficacy. Following [7], the weighting coefficients for the NLG reward in Eq. 6 were set to

$\omega_1 = 5/11$, $\omega_2 = 5/11$, and $\omega_3 = 1/11$ for BLEU-4, METEOR, and ROUGE-L, respectively. For the group-relative training, we set the group size $K = 5$. All remaining hyperparameters are kept consistent with [7]. A pretrained model [2] was fine-tuned on the MIMIC-CXR dataset for 5 epochs with a batch size of 4 using 16 NVIDIA A100 GPUs provided by the Ohio Supercomputer Center [21]. We further applied post-processing to all generated reports to remove repeating sentences and sentences that were shorter than three words. Our code is available at: https://github.com/98lingchen/MICCAI2026.

## 4 Results

**Parameter Performance** As shown in Fig. 2, the precision recall control parameter $\lambda$ provides effective and continuous regulation over clinical precision and recall. We performed inference using different $\lambda$ values from 0 to 1 with an interval of 0.1. As $\lambda$ increases, clinical precision consistently improves while recall gradually decreases, demonstrating a clear trade-off between precision and recall. The F1-score exhibits a unimodal trend and reaches its maximum at $\lambda = 0.3$. These results validate that $\lambda$ enables flexible adjustment of report generation behavior to meet different clinical requirements, such as high-recall screening and high-precision confirmation, without requiring model retraining. For the results reported below, we use $\lambda = 0.3$, which achieved the best F1 scores.

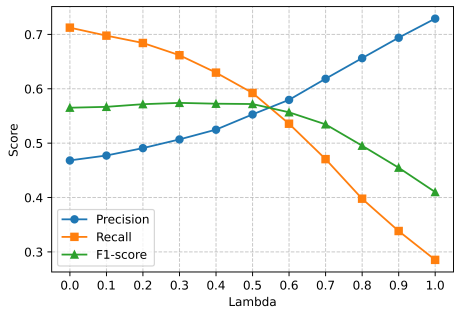



**Fig. 2.** Effect of the precision recall control parameter $\lambda$ on clinical precision, recall, and F1-score.

**Performance Comparison** We compared our method with several state-of-the-art methods on the MIMIC-CXR dataset, including R2Gen [2], METrans [22], R2GenGPT [23], BoostRRG [24], Diff-RRG [25], MLRG [26], and MedGemma 1.5 4B [27]. Results of the methods except MedGemma were reported as provided in their original papers. For MedGemma, the reports were generated as follows:

**Table 1.** Performance comparisons between the proposed method and existing approaches on the test set of the MIMIC-CXR dataset using both NLG and CE metrics.

| Method | NLG Metrics | | | | | | CE Metrics | | |
|---|---|---|---|---|---|---|---|---|---|
| | B-1 | B-2 | B-3 | B-4 | RG-L | MET. | PREC. | REC. | F1 |
| R2Gen | 0.353 | 0.218 | 0.145 | 0.103 | 0.277 | 0.142 | 0.333 | 0.273 | 0.276 |
| METrans | 0.386 | 0.250 | 0.169 | 0.124 | 0.291 | 0.152 | 0.364 | 0.309 | 0.334 |
| R2GenGPT | 0.411 | 0.267 | 0.186 | 0.134 | 0.297 | 0.160 | 0.392 | 0.387 | 0.389 |
| BoostRRG | 0.402 | 0.262 | 0.180 | 0.128 | 0.291 | 0.175 | 0.465 | 0.482 | 0.473 |
| Diff-RRG | 0.405 | 0.251 | 0.169 | 0.120 | 0.276 | 0.164 | 0.528 | 0.430 | 0.474 |
| MLRG | 0.411 | 0.277 | 0.204 | 0.158 | 0.320 | 0.176 | **0.549** | 0.468 | 0.505 |
| MedGemma | 0.200 | 0.114 | 0.071 | 0.045 | 0.160 | 0.125 | 0.520 | 0.383 | 0.441 |
| Ours | **0.451** | **0.310** | **0.219** | **0.159** | **0.333** | **0.177** | 0.505 | **0.656** | **0.571** |

The prompt for system is "You are a professional radiologists", followed by the user prompt: "Please provide a radiology report of the following chest X-ray image. Keep only the 'Findings' and 'Impression' sections in your report" and the CXR image. The findings section were then retrieved by extracting the part between "findings" and "impression" via simple text search.

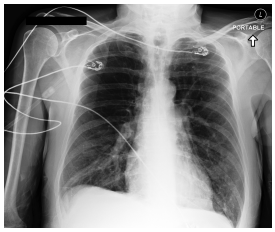
As shown in Tables 1, our method achieved state-of-the-art performance on the MIMIC-CXR dataset across both NLG and CE metrics. In particular, our model obtains notable improvements on NLP metrics, indicating stronger capability in modeling long-range dependencies. Moreover, our approach yields substantial gains in and CE metrics, demonstrating its effectiveness in generating clinically accurate reports.

**Ablation Study** Table 2 presents an ablation study on the MIMIC-CXR dataset to evaluate the individual and combined effects of the clinical reward and group-relative training. We did not separately ablate the two $\lambda$-related designs because they are coupled: $\lambda$ conditioned decoder enables test-time control, while $\lambda$ weighted rewards provide the supervision during training that makes this control meaningful. Without either component, the model achieves relatively limited performance on both NLG and CE metrics. Introducing the clinical reward alone leads to substantial improvements in CE metrics, demonstrating its effectiveness in enhancing clinical correctness. Applying group-relative training alone also yields notable gains across both NLG and CE metrics. When both components are jointly applied, the model achieves the best overall performance across all metrics, highlighting their complementary roles in improving both linguistic quality and clinical efficacy.

**Quantitative Analysis** We further present qualitative analysis of representative samples in Fig. 3 to illustrate the differences between our method and the baseline model [2]. The reports generated by our method demonstrated improved clinical accuracy and closer alignment with the ground truth. In the first case, our model correctly identified the hyperinflated lung which was missed by the baseline model. In the second case our model confirmed the enlarged car-

**Table 2.** Ablation study of clinical reward and group-relative training on the MIMIC-CXR dataset.

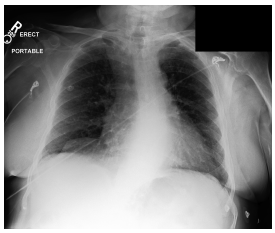
| Method | | NLG Metrics | | | | | | CE Metrics | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Clinical | Group-rel. | B-1 | B-2 | B-3 | B-4 | RG-L | MET. | PREC. | REC. | F1 |
| × | × | 0.400 | 0.253 | 0.171 | 0.120 | 0.276 | 0.154 | 0.392 | 0.335 | 0.342 |
| ✓ | × | 0.404 | 0.260 | 0.180 | 0.129 | 0.312 | 0.153 | **0.590** | 0.330 | 0.423 |
| × | ✓ | 0.433 | 0.298 | 0.212 | 0.153 | **0.339** | 0.170 | 0.546 | 0.396 | 0.459 |
| ✓ | ✓ | **0.451** | **0.310** | **0.219** | **0.159** | 0.333 | **0.177** | 0.505 | **0.656** | **0.571** |



**Ground Truth:** ap portable upright chest radiograph was provided . the lungs are hyperinflated with upper lobe lucency compatible with emphysema . no focal consolidation effusion or pneumothorax seen . cardiomediastinal silhouette is normal . bony structures are intact .

**Baseline:** ap portable upright view of the chest . overlying ekg leads are present . there is no focal consolidation effusion or pneumothorax . the cardiomediastinal silhouette is normal . imaged osseous structures are intact .

**Ours:** ap view of the chest . the lungs are hyperinflated . the lungs are clear . there is no focal consolidation . no pleural effusion or pneumothorax . the cardiomediastinal silhouette is normal . the mediastinal and hilar contours are normal .



**Ground Truth:** cardiomegaly is stable . there is no focal consolidation concerning for pneumonia . there is no pleural effusion pneumothorax or pulmonary edema . scoliosis is again noted . an old left clavicular deformity is noted .

**Baseline:** as compared to the previous radiograph there is no relevant change . low lung volumes . borderline size of the cardiac silhouette without pulmonary edema . no pleural effusions . no pneumonia no pneumothorax .

**Ours:** ap view of the chest . the lungs are clear . there is no focal consolidation . no pleural effusion or pneumothorax . the cardiac silhouette is enlarged . the mediastinal and hilar contours are normal .

**Fig. 3.** Qualitative comparison of radiology reports generated by the baseline method and our method with ground truth references. Main improvements are highlighted in blue.

diac silhouette whereas the baseline model described the size as "borderline". These examples demonstrate that the reports generated by our framework are more consistent with the ground truth and better reflect clinically meaningful findings.

## 5 Conclusion

This paper proposed a novel approach to enhance clinical consistency and controllability in RRG. Our method enables flexible precision recall control through a hybrid natural language and clinical efficacy reward mechanism, allowing the model to better capture clinically relevant findings while maintaining semantic coherence. Extensive experiments on the MIMIC-CXR dataset demonstrated the

effectiveness and robustness of our approach compared with existing methods. However, the current model is limited by the small size of the model as well as uncleaned training data, e.g., some data points were lateral CXR only, or the reports mentioned previous exam but the input had only the current CXR. In future works, we plan to extend Group reward-decoupled normalization (GDPO) [28] to our method and extend the framework to LLMs such as MedGemma and finetune it with a cleaner dataset. We also plan to perform more comprehensive evaluations of the methods with more advanced metrics [3], human reader study, and institutional data, and compare our method with recent approaches such as CURE [29] for curriculum-guided anatomy grounding.

**Acknowledgments** This work was supported in part by NIBIB under Award Number R01EB035394. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

**Disclosure of Interests** The authors have no competing interests in the paper.

## References

1. Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 2577–2586, 2018.
2. Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.
3. Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.
4. Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, 18:368–387, 2024.
5. Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D Manning, and Curtis Langlotz. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5108–5120, 2020.
6. Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31, 2018.
7. Xiulong Yi, You Fu, Jianzhi Yu, Ruiqing Liu, Hao Zhang, and Rong Hua. Lhr-rfl: Linear hybrid-reward-based reinforced focal learning for automatic radiology report generation. *IEEE transactions on medical imaging*, 44(3):1494–1504, 2025.
8. Han Qin and Yan Song. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, 2022.
9. Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. Deep communicating agents for abstractive summarization. *arXiv preprint arXiv:1803.10357*, 2018.

10. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
11. Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
12. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
13. Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.
14. Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017.
15. Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
16. Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.
17. Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
18. Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.
19. Hong Liu, Dong Wei, Zhe Xu, Xian Wu, Yefeng Zheng, and Liansheng Wang. Rrg-dpo: Direct preference optimization for clinically accurate radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–562. Springer, 2025.
20. Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.
21. Ohio Supercomputer Center. Ohio supercomputer center, 1987.
22. Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11558–11567, 2023.
23. Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023.
24. Yuyang Sha, Hongxin Pan, Weiyu Meng, and Kefeng Li. Contrastive knowledge-guided large language models for medical report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 111–120. Springer, 2025.
25. Hannah Yun, Junyeong Maeng, Eunsong Kang, and Heung-Il Suk. Diff-rrg: Longitudinal disease-wise patch difference as guidance for llm-based radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 152–161. Springer, 2025.

26. Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10348–10359, 2025.
27. Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.
28. Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026.
29. Pablo Messina, Andrés Villa, Juan León Alcázar, Karen Sánchez, Carlos Hinojosa, Denis Parra, Álvaro Soto, and Bernard Ghanem. Cure: Curriculum-guided multi-task training for reliable anatomy grounded report generation. *arXiv preprint arXiv:2601.15408*, 2026.