EditGRPO: Reinforcement Learning with Post-Rollout Edits for Clinically Accurate Chest X-Ray Report Generation

Kai Zhang[1,2]*, **Christopher Malon**[2], **Lichao Sun**[1], **Martin Renqiang Min**[2]
[1]NEC Laboratories America, [2]Lehigh University
kaz321@lehigh.edu, malon@nec-labs.com, lis221@lehigh.edu, renqiang@nec-labs.com
Code available at: https://github.com/taokz/EditGRPO

Abstract

Radiology report generation requires advanced medical image analysis, effective temporal reasoning, and accurate text generation. Although recent innovations, particularly multimodal large language models (MLLMs), have shown improved performance, their supervised fine-tuning (SFT) objective is not explicitly aligned with clinical efficacy. In this work, we introduce **EditGRPO**, a mixed-policy reinforcement learning (RL) algorithm designed specifically to optimize the generation through clinically motivated rewards. EditGRPO integrates on-policy exploration with off-policy guidance by injecting sentence-level detailed corrections during training rollouts. This mixed-policy approach addresses the exploration dilemma and sampling efficiency issues typically encountered in RL. Applied to a Qwen2.5-VL-3B MLLM initialized with supervised fine-tuning (SFT), EditGRPO outperforms both SFT and vanilla GRPO baselines, achieving an average improvement of 3.4% in CheXbert, GREEN, Radgraph, and RATEScore metrics across four major chest X-ray report generation datasets. Notably, EditGRPO also demonstrates superior out-of-domain generalization, with an average performance gain of 5.9% on unseen datasets.

1 Introduction

Automatic generation of chest X-ray reports from medical images represents a significant challenge in multi-modal artificial intelligence (Demner-Fushman et al., 2016; Irvin et al., 2019; Johnson et al., 2019; Zhang et al., 2025b). Effective AI systems in this domain can substantially reduce labor costs and help standardize radiological interpretations. Recent multi-modal large language models (MLLMs), including ChexAgent (Chen et al., 2024) and MAIRA-2 (Bannur et al., 2024), have achieved competitive performance on natural language generation metrics. However, human evaluations indicate these models often hallucinate details or omit critical clinical information (Zhang et al., 2024; Tu et al., 2024; Wu et al., 2024).

To overcome these limitations and improve clinical accuracy, reinforcement learning (RL) presents a promising alternative to supervised fine-tuning (SFT). Unlike SFT, which is sensitive to specific phrasing, RL directly optimizes clinical efficacy metrics, allowing greater flexibility in expression and prioritizing essential clinical content. Previous studies (Zhou et al., 2024b; Yang et al., 2025) have explored Proximal Policy Optimization (PPO) for report generation, but PPO suffers from low sample efficiency due to the need for a value function model. To overcome this drawback, Group Relative Policy Optimization (GRPO) was introduced (Shao et al., 2024), which avoids a value function by computing advantages based on normalized rewards relative to peer samples.

However, GRPO, as an on-policy RL method, inherently depends on the model's current capabilities, limiting effective skill expansion. Our experiments using QWen2.5-VL-3B (Bai et al., 2025) reveal that this constraint leads to low-quality outputs, frequently defaulting to generic "no finding" statements. Consequently, advantage estimates remain flat, and training performance quickly plateaus, as demonstrated in Fig. 4.

To address this exploration dilemma, we propose EditGRPO, a variant of GRPO that edits generated report candidates during rollout by injecting clinically correct information from reference reports. Edits include deleting or replacing incorrect (false-positive) findings and adding missing (false-negative) findings as shown in Fig. 1. These edits keep the model close to its existing policy while enabling off-policy guidance to bring its generations closer to the reference report.

Our work introduces three major contributions:

*Work done as an intern at NEC Laboratories America.

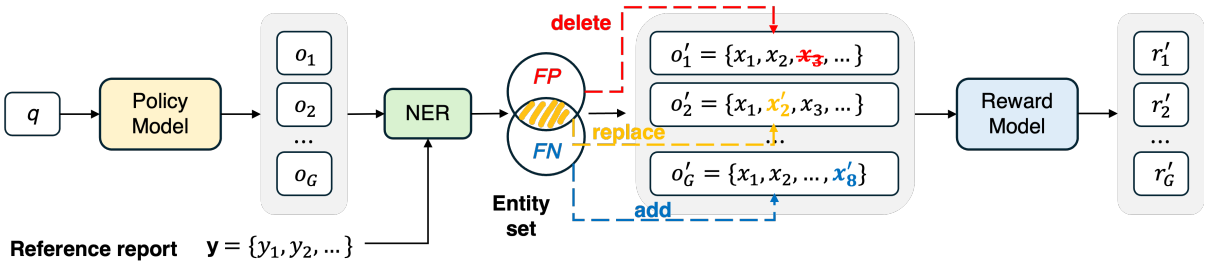

Figure 1: The graphical diagram illustrates the *post-rollout-edit* technique used in the proposed EditGRPO algorithm. For each rollout, the generated response o is edited based on the gold-standard or reference report $\mathbf{y}$ at the sentence level. This includes replacing incorrect or false positive (FP) sentences x. For example, if the reference contains "cardiomegaly" but the generated report states "the heart is within normal limits," the incorrect sentence is replaced. Additionally, missing findings, referred to as false negatives (FN), can be added based on the reference report.

- We propose EditGRPO, a mixed-policy RL algorithm that balances exploration and imitation, effectively overcoming the sampling efficiency bottleneck in report generation.
- We explore a range of editing strategies and introduce a similarity-based, sentence-level editing approach that mitigates large policy shifts by making minimal yet reward-improving modifications.
- We implement EditGRPO along with a two-stage training strategy across four multi-view and longitudinal chest X-ray report datasets, outperforming three widely used medical MLLMs. Furthermore, out-of-domain evaluation reveals that EditGRPO generalizes well to unseen data distributions.

2 Method

The formulation of GRPO is reviewed in Equation 2 in the Appendix B. Within a group of responses o_i, $i = 1, \ldots, G$ to a prompt q, GRPO generally uses a normalized advantage at every token t:

$$\hat{A}_{i,t} = \frac{R(q, o_i) - \text{mean}\left(\{R(q, o_1), \ldots, R(q, o_G)\}\right)}{\text{std}\left(\{R(q, o_1), \ldots, R(q, o_G)\}\right)} \quad (1)$$

but Dr. GRPO (Liu et al., 2025b) revealed that vanilla GRPO introduces optimization biases. For example, longer responses are penalized less due to their larger $|o_i|$, leading the policy to favor lengthier but potentially incorrect outputs. Additionally, a question-difficulty bias arises from the standard deviation term in the advantage function, where questions with low standard deviation (i.e., very easy or very hard, where rewards are mostly 1 or 0) are disproportionately weighted during policy updates. To mitigate these issues, EditGRPO adopts their proposed modification and removes this bias-inducing quotient as in Dr. GRPO, to use *unnormalized advantage*. By default, EditGRPO samples $\frac{G}{2}$ responses o_i, and fills the other half of the batch with edited versions o'_i of these.

Editing Rule. Given a generated report $\mathbf{x}$ and a reference report $\mathbf{y}$, the RaTE-NER model (Zhao et al., 2024) extracts entities from each, together with vector embeddings and labels indicating presence or absence. Fix a cosine similarity threshold τ. For each reference sentence y_j, let $E[y_j]$ to be the collection of all entities e with $\cos(e, e') > \tau$ for some $e' \in y_j$. An entity $e \in x$ is *spurious* if it is not in any $E[y_j]$ with a matching presence label. RadGraph can also identify edit candidates but is sensitive to phrasing (see Table 4 in the appendix).

(a) **Mislabeled Entities.** If there exists an entity e in x_i and y_j with conflicting presence labels, select one such conflict uniformly at random and replace x_i with y_j.

(b) **False-positive replacement.** Otherwise, if some sentence x_i contains a set of spurious entities $E_{\text{fp}} \not\subseteq E[y_j]$ for all j at threshold τ, compute

$$s^* = \arg\max_{j:\, y_j \neq x_i} \frac{1}{|E_{\text{fp}}|} \sum_{e \in E_{\text{fp}}} \max_{e' \in E[y_j]} \cos(e, e').$$

If $\frac{1}{|E_{\text{fp}}|} \sum_{e \in E_{\text{fp}}} \max_{e' \in E[y_{s^*}]} \cos(e, e') \geq \tau$, replace x_i with y_{s^*}.

(c) **False-positive deletion.** Otherwise, if no replacement was made and there is still a sentence x_i with false-positive entities, delete it.

(d) **False-negative augmentation.** Otherwise, if there is an entity in $\mathbf{y}$ missing from $\mathbf{x}$, select one reference sentence y_j containing it and *append* y_j to the end of $\mathbf{x}$.

(e) **Termination.** If none of the above applies, return $\mathbf{x}$; if it is empty, replace it with any unused false-negative sentence y_j.

3 Experiments

3.1 Experimental Setting

Data. We focus on MIMIC-CXR (Johnson et al., 2019), a well-studied large-scale dataset, for both training and evaluation. Additionally, we incorporate the newly released large-scale RexGradient (Zhang et al., 2025a) dataset to further validate the effectiveness of EditGRPO. We also study two smaller benchmark datasets, IU-XRay (Demner-Fushman et al., 2016) and PadChest-GR (de Castro et al., 2025), both for training and for out-of-domain generalization. See Appendices F and G.

Models. Qwen-VL-2.5 (3B) (Bai et al., 2025) is used for training report generation models. For a comparative baseline on MIMIC-CXR, we include results from other models supporting multi-view and longitudinal inputs, such as ChexAgent (Chen et al., 2024), MAIRA-2 (Bannur et al., 2024), and MedGemma (Sellergren et al., 2025). The performance of the baselines reported in this paper is obtained through our own inference using the curated multi-view and longitudinal datasets. The model's performance may be influenced by the prompt.

Training. Because of the time and resources required for reinforcement learning, we compare results for MIMIC-CXR at step 1000, corresponding to about 1/4 of the training data, even though performance continues to improve as shown in Fig. 4. For RexGradient, we use the checkpoint at step 600, representing roughly 15% of the data scale. For the smaller datasets, we train for one full epoch. More details are stated in Appendix D and E.

Evaluation. We evaluate the generated reports using radiology-specific metrics, following established protocols. These include RadGraph-F1 (Jain et al., 2021), CheXbert scores (Micro/Macro-F1-14/5) (Smit et al., 2020), RaTE Score (Zhao et al., 2024), and GREEN (Ostmeier et al., 2024). These clinical metrics are designed to emphasize the accuracy of medical findings, with a focus on detecting and assessing clinically relevant entities.

3.2 Main Results

SFT is necessary before RL for chest X-ray report generation. Table 1 presents detailed results for the MIMIC-CXR dataset, alongside the performance of SFT and EditGRPO methods on RexGradient. GRPO alone achieves an average score of only 17.35%, demonstrating limited effectiveness without prior SFT training. However, incorporating an initial SFT phase significantly boosts performance, yielding approximately a 25% improvement. The same conclusion is also supported by related works (Guo et al., 2025; Fan et al., 2025; Liu et al., 2025a).

EditGRPO surpasses SFT and other RL variants. The SFT+EditGRPO model (or its normalized advantage variant) lead all metrics except RadGraph F1, which is sensitive to the exact wordings prioritized by SFT training. Accounting for semantic equivalence, we see a 5.0% improvement in CheXbert Micro-F1-5, 4.5% improvement in CheXbert Micro-F1-14, and 2.8% improvement in RaTEScore with SFT+EditGRPO. For the macro variants of the CheXbert metrics, the normalized advantage variant of EditGRPO performs slightly better, although its overall performance is generally lower. An ablation where EditGRPO modifies entire G paragraphs rather than individual sentences (limited to one batch element per query) significantly reduces performance. This highlights the importance of small, localized edits closely aligned with the current policy.

EditGRPO enables general-domain models to detect abnormalities more accurately than existing domain-specific models. As shown in Table 1, applying EditGRPO to Qwen-VL-2.5 (3B) achieves state-of-the-art performance across all metrics except the GREEN score. Notably, we observe substantial gains in CheXbert-F1 scores, which are directly associated with the model's ability to detect common abnormalities. Specifically, SFT + EditGRPO yields improvements over MAIRA-2 of 4.4%, 1.2%, 5.9%, and 5.5% across four CheXbert-F1 variants, respectively.

SFT-then-EditGRPO performs effectively on small-scale datasets. Fig. 2 demonstrates the effectiveness of EditGRPO on two smaller datasets: IU-XRay and PadChest-GR. We primarily focus on RadGraph-F1, RaTE Score, CheXbert-Macro-F1-14, and GREEN Score for evaluation. On IU-XRay, the SFT+EditGRPO model achieves improvements over SFT by 4.6% (RadGraph-F1), 2.2% (RaTE Score), 1.6% (CheXbert-Macro-F1-14), and 7.1% (GREEN Score). Similarly, on PadChest-GR, SFT+EditGRPO outperforms SFT

Method	Micro-F1-14	Macro-F1-14	Micro-F1-5	Macro-F1-5	RadGraph F1	RaTE	GREEN	Avg.
MIMIC-CXR								
MAIRA-2 (7B)	0.5154	0.3557	0.5531	0.4695	0.2104	0.3042	**0.5042**	0.4161
ChexAgent (8B)	0.2981	0.1772	0.3593	0.2363	0.1544	0.4370	0.2359	0.2712
MedGemma (4B)	0.4742	0.3462	0.5222	0.4752	0.1113	0.4680	0.2253	0.3746
Qwen2.5-VL-3B								
SFT (*ep3*)	0.5144	0.3344	0.5625	0.4737	0.2937	0.5434	0.3583	0.4401
GRPO	0.1381	0.0966	0.1167	0.0981	0.0905	0.4067	0.2676	0.1735
SFT(*ep2*) + GRPO	0.4781	0.3042	0.5376	0.4527	**0.2963**	0.5411	0.3497	0.4228
SFT(*ep2*) + Dr. GRPO	0.5470	0.3481	0.6050	0.5023	0.2959	0.5460	0.3627	0.4528
SFT(*ep2*) + EditGRPO (*para*)	0.5353	0.3458	0.5946	0.4986	0.2546	0.5257	0.3114	0.4380
SFT(*ep2*) + EditGRPO (*norm*)	0.5467	**0.3789**	0.5986	**0.5388**	0.2721	0.5441	0.3500	0.4613
SFT(*ep2*) + EditGRPO	**0.5594**	0.3674	**0.6124**	0.5242	0.2841	**0.5712**	0.3825	**0.4716**
RexGradient								
Qwen2.5-VL-3B								
SFT (*ep2*)	0.4035	0.1969	0.2273	0.1748	0.3294	0.5811	0.4430	0.3366
SFT (*ep1*) + EditGRPO	**0.4061**	**0.2159**	**0.2867**	0.2168	0.2986	**0.5981**	**0.4756**	**0.3568**

Table 1: Performance of various training variants and medical *SoTA* models across clinical metrics on two large-scale RRG datasets: MIMIC-CXR and RexGradient. ***Note:*** EditGRPO is built on the Dr. GRPO and utilizes sentence-level edits by default; *norm* indicates normalized advantage are utilized; *para* means edits are performed on the paragraph (or report) level. *ep** denotes the training epoch index.

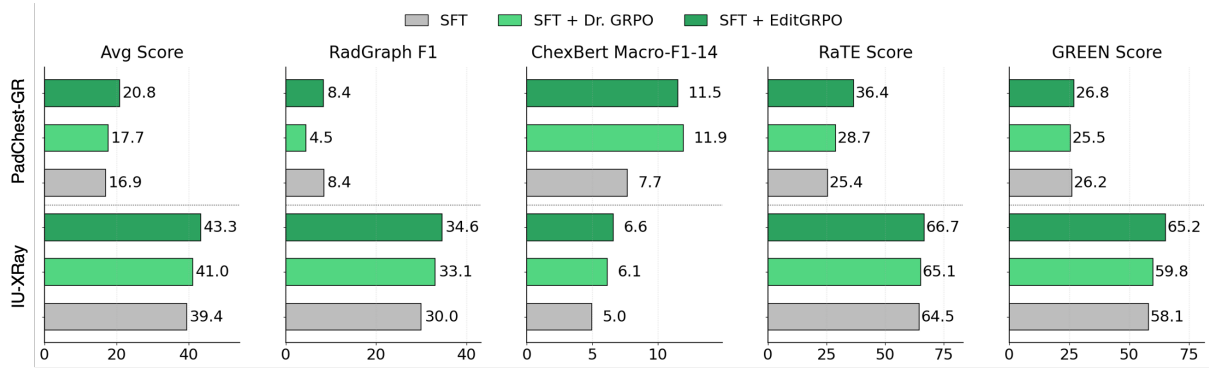

Figure 2: Performance (%) of different training strategies on two small-scale datasets: IU-XRay and PadChest-GR.

by 3.8% (ChexBert-Macro-F1-14), 11.0% (RaTE Score), and 0.6% (GREEN Score), although the improvement in CheXbert-Macro-F1-14 is marginal. Comprehensive metrics are detailed in Table 5.

RL demonstrates better generalization. Table 2 compares the average cross-metric performance (Avg. Score) of models trained on MIMIC-CXR and then evaluated on two external datasets, IU-XRay and PadChest-GR, using three training strategies. On PadChest-GR, the SFT baseline achieves an Avg. Score of 0.174, which rises to 0.193 (+11%) when Dr.GRPO is applied and further to 0.260 (+49%) with EditGRPO. A similar but less pronounced trend appears on IU-XRay: SFT starts at 0.403, Dr. GRPO improves it to 0.416 (+3%), and EditGRPO brings it up to 0.435 (+8%). Moreover, EditGRPO consistently yields the highest Avg. Score across both target domains, demonstrating its superior robustness in out-domain generalization (see Table 6 for the full breakdown of evaluation metrics).

Dataset	SFT	SFT + Dr. GRPO	SFT + EditGRPO
IU-XRay	0.4026	0.4160	**0.4348**
PadChest	0.1742	0.1926	**0.2595**

Table 2: Average score across all evaluation metrics for models trained on MIMIC-CXR.

4 Conclusion

Our EditGRPO framework makes it possible to optimize diverse clinical efficacy metrics in a GRPO-based framework, addressing the failure of pure GRPO to adequately explore. Even with a simple MLLM architecture pretrained on general-domain data, our technique achieves gains in multiple rewards with either large or small training sets. Our technique is effective even in challenging multi-view, longitudinal image settings. Furthermore,

SFT+EditGRPO models from a large training set generalize better than SFT or SFT+Dr.GRPO to small, out-of-domain data. Although our focus is on the training algorithm, we expect advances in architectures or training data could be integrated to further improve performance.

5 Limitation

Although our proposed EditGRPO method demonstrates the robust potential of reinforcement learning to improve the clinical efficacy of chest X-ray report generation, several limitations remain. First, we did not extend training to larger models (e.g., 7B or 32B parameters), primarily due to the multiplicative increase in computational and time requirements. This constraint is especially relevant in our settings involving multi-view and longitudinal data, where multiple image inputs are required and resource efficiency becomes critical. Second, while the current clinical metrics used in evaluation are generally reliable, they may not fully align with expert radiologist assessments. Additional human evaluation will be necessary to more comprehensively validate the quality and safety of reports generated by models trained with EditGRPO. Lastly, this work focuses specifically on chest X-ray report generation. The applicability of EditGRPO to other medical imaging modalities, such as CT or MRI, remains an open question and is a promising direction for future research.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, and 2 others. 2024. Maira-2: Grounded radiology report generation. *Preprint*, arXiv:2406.04449.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, Jeya Maria Jose Valanarasu, Mohamed Siddig Eltayeb Muneer, Eduardo Pontes Reis, Joseph Paul Cohen, Cameron Olsen, Tanishq Mathew Abraham, Emily B. Tsai, Christopher F. Beaulieu, Jenia Jitsev, and 4 others. 2024. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *Preprint*, arXiv:2401.12208.

Daniel Coelho de Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, and 1 others. 2025. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *NEJM AI*, 2(7):AIdbp2401120.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360.

D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonale. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.*, 23(2):304–310.

Ziqing Fan, Cheng Liang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Chestx-reasoner: Advancing radiology foundation models with reasoning through step-by-step verification. *arXiv preprint arXiv:2504.20930*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. Nguyen Duong, T. Bui, P. Chambon, M. Lungren, A. Ng, C. Langlotz, and P. Rajpurkar. 2021.

RadGraph: Extracting clinical entities and relations from radiology reports. In *PhysioNet*.

A. E. W. Johnson, T. J. Pollard, and S. J. Berkowitz et al. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*, 6(317).

Navdeep Kaur and Ajay Mittal. 2022. Cadxreport: Chest x-ray report generation using co-attention mechanism and reinforcement learning. *Computers in Biology and Medicine*, 145:105498.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. 2025. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, and 1 others. 2025a. X-reasoner: Towards generalizable reasoning across modalities and domains. *arXiv preprint arXiv:2505.03981*.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025b. Understanding r1-zero-like training: A critical perspective. *Preprint*, arXiv:2503.20783.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025c. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and 1 others. 2024. Green: Generative radiology report evaluation and error notation. *arXiv preprint arXiv:2405.03595*.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. 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)*, pages 1500–1519, Online. Association for Computational Linguistics.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.

Lai Wei, Yuting Li, Chen Wang, Yue Wang, Linghe Kong, Weiran Huang, and Lichao Sun. 2025. Unsupervised post-training for multi-modal llm reasoning via grpo. *arXiv preprint arXiv:2505.22453*, arXiv:2505.22453.

Jinge Wu, Yunsoo Kim, and Honghan Wu. 2024. Hallucination benchmark in medical visual question answering. *Preprint*, arXiv:2401.05827.

Lingrui Yang, Yuxing Zhou, Jun Qi, Xiantong Zhen, Li Sun, Shan Shi, Qinghua Su, and Xuedong Yang. 2025. Aligning large language models with radiologists by reinforcement learning from ai feedback for chest ct reports. *European Journal of Radiology*, 184:111984.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, and 1 others. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiaoman Zhang, Julián N Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. 2025a. Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. *arXiv preprint arXiv:2505.00228*.

Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. 2025b. Rexrank: A public leaderboard for ai-powered radiology report generation. In *AAAI Bridge Program on AI for Medicine and Healthcare*, pages 90–99. PMLR.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. RaTEScore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, Miami, Florida, USA. Association for Computational Linguistics.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, and 1 others. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proceedings of the VLDB Endowment*, 16(12):3848–3860.

Hong-Yu Zhou, Julián Nicolás Acosta, Subathra Adithan, Suvrankar Datta, Eric J Topol, and Pranav Rajpurkar. 2024a. Medversa: A generalist foundation model for medical image interpretation. *arXiv preprint arXiv:2405.07988*.

Zijian Zhou, Miaojing Shi, Meng Wei, Oluwatosin Alabi, Zijie Yue, and Tom Vercauteren. 2024b. Large model driven radiology report generation with clinical quality reinforcement learning. *arXiv preprint arXiv:2403.06728*.

A Related Work

A.1 Radiology Report Generation

Recent research has focused on transformer-based architectures for medical imaging analysis (Li et al., 2023). Some of these systems are differentiated by their ability to support multi-image inputs, including MAIRA-2 (Bannur et al., 2024) and CheXAgent (Chen et al., 2024). The recent leaderboard RexRank (Zhang et al., 2025b) studies only a single-image setting, and the latest leaders such as MedVersa (Zhou et al., 2024a) focus on integrating multiple modules for detection and segmentation and a large-scale data wrangling effort. In contrast, we focus on training algorithm improvements for a commonly used MLLM architecture, rather than an architecture or data contribution.

A.2 Reinforcement Learning

To further enhance clinical accuracy, some methods incorporate RL to optimize for task-specific rewards, such as capturing "clinically relevant" features (Liu et al., 2019; Kaur and Mittal, 2022; Zhou et al., 2024b) or maintaining logical consistency (Delbrouck et al., 2022; Miura et al., 2021). Rule-based reinforcement learning techniques, exemplified by Group Relative Policy Optimization (GRPO), have shown strong potential for large-scale RL applications, particularly in tasks such as mathematical reasoning and code generation, and have recently been extended to multimodal setting (Wei et al., 2025). In the medical domain, GRPO has been explored in visual question answering tasks (Pan et al., 2025; Fan et al., 2025; Lai et al., 2025). However, to the best of our knowledge, it has not yet been applied to radiology report generation.

A.3 Clinical Evaluation Metrics

Evaluating the quality of generated radiology reports is non-trivial. Early works adopted general-domain natural language processing metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and BERTScore (Zhang* et al., 2020). While these metrics are widely used for text evaluation, they treat differences in wording the same as clinically significant errors, failing to reflect medical accuracy. To address this limitation, clinically informed evaluation metrics, such as CheXbert (Smit et al., 2020), RadGraph (Jain et al., 2021), GREEN (Ostmeier et al., 2024), and RaTEScore (Zhao et al., 2024), have been proposed to better assess clinical correctness and utility. CheXbert is based on multi-label classification results for 5 or 13 diseases (along with one extra "normal" label). RadGraph considers literal entity agreement considering the positive or negative context of each entity. GREEN judges recall and precision errors by LLM prompting. RaTEScore is inspired by RadGraph but less sensitive to phrasing by an F1-like computation which allows semantic matching between entities based on a cosine similarity.

B Group Relative Policy Optimization

Let $P(Q)$ denote the distribution over questions (images and prompts) used for training, where q is a sampled question in the current iteration. Let $\pi_{\theta_{\text{old}}}$ and $\pi_{\theta_{\text{new}}}$ denote the old policy and current (new) policy, respectively, where o is a complete response sampled from a policy. Let $\pi_{\theta_{\text{ref}}}$ denote the reference policy, which in practice is the frozen base MLLM. Let G be the number of responses sampled per question in each iteration. The GRPO objective is given by:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \Bigg(\min\Bigg(\frac{\pi_{\theta_{\text{new}}}(o_{i,t}\mid q)}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q)} A_{i,t}, \text{clip}\left(\frac{\pi_{\theta_{\text{new}}}(o_{i,t}\mid q)}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q)}, 1-\epsilon, 1+\epsilon \right) A_{i,t} \Bigg) \Bigg) - \beta\, \mathbb{D}_{\text{KL}}\left(\pi_{\theta_{\text{new}}} \,\|\, \pi_{\theta_{\text{ref}}}\right) \Bigg], \tag{2}$$

where $\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ is the policy ratio, and A_i is the estimated advantage defined in Equation 1, and ϵ is the clipping threshold for policy updates.

C Motivation on Post-Rollout Edits

A key obstacle to applying reinforcement learning to chest X-ray report generation is a shortcut bias: because normal studies vastly outnumber abnormal ones, a policy can maximize its expected reward by emitting a generic "No abnormality" statement. This imbalance drives policy collapse, with the model converging to a single high-probability "normal" mode—an effect that becomes even more pronounced in low-resource settings (see Table 1, where training with pure GRPO results in low CheXbert Macro-F1 scores.). Unlike classification tasks, a radiology report is a holistic narrative that may describe an arbitrary subset and count of findings, so straightforward

class-frequency re-weighting (Lin et al., 2017) may not be practical for all findings simultaneously. To counter this collapse we introduce *post-rollout edits*: after sampling a candidate report o_i from the policy, we inject expert corrections based on the gold-standard report, recompute the edited trajectory's log-probability, and use the resulting adjusted advantage to steer the policy gradient toward clinically faithful outputs. To prevent policy drift, we apply this editing at the sentence level,, which also help the model absorb fine-grained corrections throughout the long-context report.

D Reward Design

Reasoning or no-thinking? Most existing MLLM works that utilize the GRPO algorithm aim to incentivize reasoning capabilities in the VQA setting (Lai et al., 2025; Pan et al., 2025; Liu et al., 2025c; Huang et al., 2025; Fan et al., 2025; Liu et al., 2025a). To achieve this, they incorporate a format score that encourages the model to enclose its reasoning process within <*think*> and </*think*> tags. However, for report generation, explicit self-rationalization is unnecessary: the radiology report itself inherently embodies the trackable reasoning of imaging analysis. It performs semantic mapping from image features to clinical description and lays out the inferred relationships among findings and their differential diagnoses. Therefore, we do not incorporate a format reward and instead conduct "no-thinking" GRPO setup.

Clinical efficacy rewards. We adopt a composite reward strategy, as different clinical efficacy metrics evaluate report quality from complementary perspectives. We adopt the following metrics as the primary reward signals without applying weighting, resulting in a composite reward defined as:

$$R = \text{RadGraph-F1} + \text{CheXbert-Micro-F1-14} + \text{RaTE}.$$

We choose Micro-F1 over Macro-F1 because macro-level rewards tend to be too sparse across most datasets, resulting in weak or uninformative learning signals.

While other metrics, such as GREEN, show stronger alignment with human evaluations (Ostmeier et al., 2024), they rely on LLM-based evaluations, which incur significantly higher inference latency and resource costs. The metrics are computed using their official and standardized implementations: RadGraph-F1[1], CheXbert-F1[2], RaTE Score[3], and GREEN[4].

E Training Configurations

Training begins with a general-domain MLLM with trainable parameters θ, following a two-stage approach: SFT followed by RL. This strategy allows the model to first adapt to the chest X-ray domain and subsequently perform effective sampling during the RL phase.

Stage 1: SFT for domain adaptation. We train the model using cross-entropy loss until both the loss and clinical metrics stabilize. Formally, for each *multi-view and longitudinal* chest X-ray image set $\mathbf{X}_v$, we append a text prompt $\mathbf{X}_q$ containing clinical context information such as indication, technique, and comparison. The MLLM is then trained to predict the target report tokens using its original autoregressive objective. Specifically, given a target findings section $\mathbf{X}_a$ of length L, the model is optimized to maximize the likelihood:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{i=1}^{L} \log p_\theta\big(x_i \mid \mathbf{X}_v,\, \mathbf{X}_{q,<i},\, \mathbf{X}_{a,<i}\big).$$

Stage 2: RL for clinical improvement. To prevent model collapse, we use the checkpoint from the epoch prior to convergence as the starting point for RL training. The optimization objective is described in Equation 2.

Prompts. An prompt example with two images are shown in the following:

```
<image><image>

As a radiologist assistant, your task is to
    interpret a chest X-ray study. Given
    the current <view> image, and the prior
    <view> image.

Please provide a detailed description of
    the findings from the image(s) in the
    current study. If there are prior
    studies available, please incorporate
    relevant details from those as well in
    your analysis.

INDICATION: <INDICATION>
TECHNIQUE:  <TECHNIQUE>
COMPARISON:  <COMPARISON>
```

[1] https://pypi.org/project/radgraph/0.1.2/
[2] https://pypi.org/project/f1chexbert/
[3] https://pypi.org/project/RaTEScore/0.5.0/
[4] https://pypi.org/project/green-score/0.0.8/

Dataset	# Samples		# Images		% Has Prior	
	Train	Test	Train	Test	Train	Test
MIMIC-CXR	146,893	2,231	4.17 ± 0.63	4.30 ± 0.62	42.59	68.60
RexGradient	139,884	9,992	1.69 ± 0.63	1.69 ± 0.63	0	0
IU-Xray	2,365	590	2.00 ± 0.00	2.00 ± 0.00	0	0
PadChest-GR	3,640	915	1.32 ± 0.47	1.31 ± 0.46	32.40	31.91

Table 3: The datasets used for training and evaluating EditGRPO include statistics such as the proportion of samples containing prior studies and the total number of images. For RexGradient, cases containing more than six images were excluded due to resource constraints. Note that The numbers reported in this table reflect the dataset statistics after preprocessing, rather than the raw data.

Hyperparameters. In the reinforcement learning setup, we use a rollout batch size of 32. Image inputs are constrained to a maximum of 401,408 pixels and a minimum of 262,144 pixels. We employ the AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate of 1.0 e-6 and a weight decay of 1.0 e-2. During the rollout phase, the generation temperature is set to 1.0, and we sample $\frac{G}{2} = 5$ responses per prompt. To reduce memory usage, we enable gradient checkpointing (Chen et al., 2016) and apply Fully Sharded Data Parallelism (FSDP) (Zhao et al., 2023). Additionally, vLLM (Kwon et al., 2023) is used to accelerate inference and response generation. We set the cosine similarity threshold τ to 0.6 in the editing rules based on preliminary observations, without formal hyperparameter optimization.

Computation. Training Qwen2.5-VL-3B is conducted on a computational infrastructure equipped with NVIDIA A100 GPUs (80GB memory). For MIMIC-CXR, EDITGRPO training up to step 1000 is estimated to take approximately 5 days using four GPUs. For REXGRADIENT, training is estimated at 3 days with four GPUs. In contrast, training on IU-XRAY requires roughly 5 hours using two GPUs, while PADCHEST-GR takes approximately 8 hours using two GPUs.

F Datasets

For all datasets, we follow the official train-test splits. For MIMIC-CXR and PadChest-GR, we construct multi-view and longitudinal inputs based on metadata. Specifically, we retain both frontal and lateral views and identify prior exams using chronological order. Prior reports are excluded from the input to minimize token costs—given the resource-intensive nature of multi-image settings—and to prevent shortcut learning, such as copying content from earlier reports.

It is worth noting that MIMIC-CXR includes a "Comparison" section but does not provide exact identifiers for prior exams, making precise longitudinal matching infeasible. Therefore, we use the most recent prior exam to build sequential image inputs. For IU-XRay and RexGradient, we only curate multi-view inputs due to limitations in longitudinal metadata.

The detailed data statistics is shown in Table 3.

Data license and use agreement. Our study employs four fully de-identified, publicly released chest radiograph collections—MIMIC-CXR (Johnson et al., 2019), PadChest-GR (de Castro et al., 2025) (English version), IU-Xray (Demner-Fushman et al., 2016), and RexGradient (Zhang et al., 2025a)—each governed by permissive open-use terms that together guarantee ethical compliance and reproducibility. The MIMIC-CXR dataset is accessed under its data use agreement and CITI "Data or Specimens Only Research" certification. PadChest-GR is distributed under the PADCHEST Dataset Research Use Agreement[5], which grants free access for academic research only and explicitly prohibits redistribution or commercial use. IU-Xray is released under a CC BY-NC-ND 4.0 license, which permits non-commercial redistribution with attribution but prohibits both commercial reuse and the creation of derivative works. RexGradient is available via Hugging Face[6], and is released under a Non-Commercial Data Access and Use Agreement that restricts use to non-clinical, non-commercial research while requiring proper attribution to Harvard Medical School and Gradient Health. No patient can be re-identified from any of these sources, and all secondary analyses

[5] https://bimcv.cipf.es/bimcv-projects/padchest/padchest-dataset-research-use-agreement/

[6] https://huggingface.co/datasets/rajpurkarlab/ReXGradient-160K

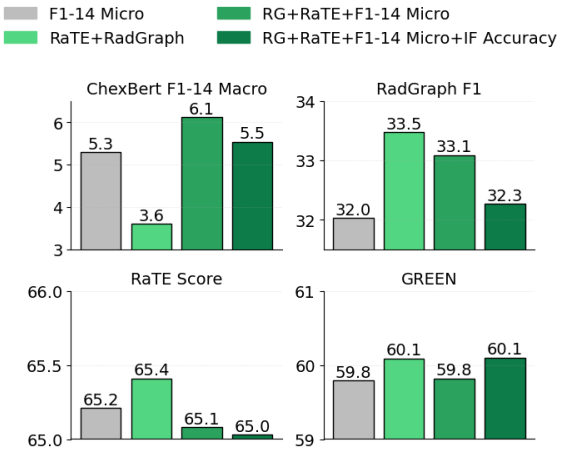

Figure 3: Influence of reward design on the IU-XRay dataset under the SFT + Dr.GRPO setting. *RG* denotes the RadGraph reward, and *IF* denotes the inverse-frequency reward, which assigns higher scores when a rare condition is hit according to the label distribution across the 14 CheXpert classes (Irvin et al., 2019) of the training data.

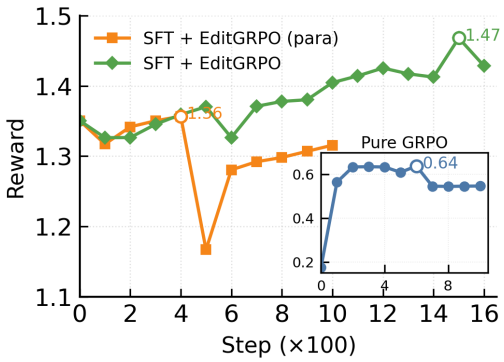

Figure 4: Reward gains (RadGraph + RaTE + Chexbert-Micro-F1-14, the maximum is 3) over training step on MIMIC-CXR.

Metrics	RaTE-NER	RadGraph
Micro-F1-14	**0.5120**	0.5009
Macro-F1-14	**0.0662**	0.0556
Micro-F1-5	**0.0839**	0.0432
Macro-F1-5	**0.0502**	0.0215
RadGraph F1	**0.3458**	0.3285
RaTEScore	**0.6672**	0.6484

Table 4: Influence of editing rules on IU-XRay dataset under the SFT + EditGRPO setting.

conform strictly to each dataset's specific attribution, share-alike, and noncommercial clauses. By restricting our work to these four open-license datasets, we uphold the highest standards of data privacy, transparency, and legal clarity in ethical AI research.

G Additional Experiments

In this section, we present comprehensive evaluation metrics on IU-XRay and PadChest-GR, as shown in Table 5. In addition, we conduct ablation studies on both reward design (see Fig. 3) and post-rollout editing strategies (see Fig. 4) to analyze their individual contributions to overall performance.

Composite reward design makes sense. Figure 3 compares four reward schemes on IU-XRay. When optimizing solely for RaTE and RadGraph, the ChexBert-Macro-F1-14 score falls sharply from 5.3% to 3.6%, indicating that these metrics alone do not drive macro-F1 improvements. Adding the F1-14 Micro component not only recovers but surpasses the baseline, boosting macro-F1 to 6.1; the subsequent inclusion of IF Accuracy yields a slight decrease to 5.5, suggesting diminishing returns from additional reward terms. The decreases in RadGraph-F1, RaTE Score, and GREEN by adding the Chexbert-F1-14 rewards to the RaTE+RadGraph are all quite small compared to the gain in the Chexbert-F1-14 metric. These results confirm that a composite reward incorporating the primary target metric, F1-14 Micro, is crucial for maximizing overall generalization.

Controllable editing is necessary for EditGRPO. Figure 4 illustrates the reward gains over training steps on MIMIC-CXR. We observe that applying paragraph-level editing leads to unstable training dynamics. In contrast, sentence-level editing results in a stable training curve with consistent improvement, without signs of performance saturation. This highlights the importance of fine-grained, controllable edits for effective reinforcement learning in report generation. In contrast, applying pure GRPO results in rapid performance saturation, achieving significantly lower reward gains, approximately half, compared to the SFT + EditGRPO setting.

RaTE-based editing outperforms RadGraph-based editing under the EditGRPO. Although RaTEScore only identifies presence or absence (as abnormality versus non-abnormality or disease versus non-disease), RadGraph identifies three presence labels (definitely present, definitely absent, or uncertain). However, RadGraph does not come with an entity embedding model, so the cosine similarity threshold reverts to exact entity matching.

Table 5: Performance of various training variants on two small-scale datasets: IU-XRay and PadChest-GR.

Method	Micro-F1-14	Macro-F1-14	Micro-F1-5	Macro-F1-5	RadGraph F1	RaTE	GREEN	Avg.
IU-XRay								
Qwen2.5-VL-3B								
SFT (*ep1*)	**0.5250**	0.0497	0	0	0.3001	0.6446	0.5806	0.3000
SFT(*ep1*) + Dr. GRPO	0.5166	0.0613	0.0143	0.0053	0.3309	0.6508	0.5982	0.3111
SFT(*ep1*) + EditGRPO	0.5120	**0.0662**	**0.0839**	**0.0502**	**0.3458**	**0.6672**	**0.6517**	**0.3396**
PadChest-GR								
Qwen2.5-VL-3B								
SFT (*ep3*)	0.4533	0.0768	0.0114	0.0056	**0.0844**	0.2541	0.2617	0.1639
SFT (*ep2*) + Dr. GRPO	0.4128	**0.1194**	0.0952	0.0559	0.0446	0.2872	0.2551	0.1815
SFT (*ep2*) + EditGRPO	**0.4608**	0.1153	**0.1017**	**0.0708**	0.0839	**0.3641**	**0.2683**	**0.2093**

Table 6: Out-of-domain generalization performance of models trained on MIMIC-CXR using various training strategies, evaluated on two small-scale datasets: IU-XRay and PadChest-GR.

Method	Micro-F1-14	Macro-F1-14	Micro-F1-5	Macro-F1-5	RadGraph F1	RaTE	GREEN	Avg.
IU-XRay								
Qwen2.5-VL-3B-MIMIC-CXR								
SFT (*ep3*)	0.5759	0.1682	0.2581	0.2050	**0.3161**	0.6529	0.6423	0.4026
SFT(*ep2*) + Dr. GRPO	0.5498	0.1848	0.3386	0.2591	0.2983	0.6363	**0.6452**	0.4160
SFT(*ep2*) + EditGRPO	**0.5833**	**0.2040**	**0.3644**	**0.2775**	0.3136	**0.6557**	0.6448	**0.4348**
PadChest-GR								
Qwen2.5-VL-3B-MIMIC-CXR								
SFT (*ep3*)	0.4128	0.1194	0.0952	0.0559	0.0446	**0.2872**	0.2041	0.1742
SFT (*ep2*) + Dr. GRPO	0.4608	0.1153	0.1017	0.0708	**0.0839**	0.2776	0.2379	0.1926
SFT (*ep2*) + EditGRPO	**0.4841**	**0.2208**	**0.3073**	**0.2563**	0.0222	0.2858	**0.2397**	**0.2595**

Results on IU-XRay, shown in Table 4, indicate that RadGraph-based editing underperforms. This likely stems from reliance on exact string matching with RadGraph, which lacks canonical concept IDs or mention-level embeddings for semantic matching, making the procedure sensitive to phrasing (Jain et al., 2021).