MedXEdit-GRPO: Reasoning-Aware Online RL for Counterfactual Medical Image Generation

Yaning Pan[1], Renwei Xiao[1], Qingqiu Li[1], Xiaoyi Li[1], Xiaoling Li[2], Rui Feng[1,2,3(✉)], and Xiaobo Zhang[2(✉)]

[1] College of Computer Science and Artificial Intelligence, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
fengrui@fudan.edu.cn

[2] Children's Hospital of Fudan University, National Children's Medical Center, Shanghai, China
zhangxiaobo0307@163.com

[3] College of Intelligent Robotics and Advanced Manufacturing, Fudan University, Shanghai, China

Abstract. Counterfactual medical image generation enables "what-if" clinical exploration through instruction-conditioned editing. However, existing methods often lack the multimodal reasoning required to align abstract directives with precise visual modifications, failing to balance instruction compliance with anatomical preservation. Reinforcement learning offers a promising solution, but it remains largely unexplored in this field. In this paper, we propose **MedXEdit-GRPO**, a reinforcement learning framework for counterfactual medical image editing. To provide stable and comprehensive feedback, we curate MedXEdit-Reward, a large-scale dataset of 30,000 samples with rationales, and develop MedXEdit-Score, a reward model enforcing a "reasoning-before-scoring" paradigm for semantically grounded optimization. We optimize our generative policy via Group Relative Policy Optimization (GRPO), which effectively harmonizes diverse clinical objectives into a unified, stable training process. Quantitative and qualitative evaluations demonstrate that MedXEdit-GRPO significantly outperforms state-of-the-art methods in both instruction compliance and anatomical consistency, highlighting its potential for robust and interpretable medical analysis.

Keywords: Counterfactual Generation · Disease Progression · Image Editing · Reinforcement Learning.

1 Introduction

Counterfactual medical image generation synthesizes hypothetical "what-if" scenarios (e.g., disease progression) through the instruction-conditioned editing of pathological features [4,13]. This methodology provides a critical instrument

✉ Corresponding authors.

for evaluating model robustness and enhancing classifier performance via targeted synthetic data augmentation [12,21]. Furthermore, it facilitates systematic anomaly detection [7] and the visual exploration of diverse clinical hypotheses [1,26], providing a controlled and interpretable mechanism for simulating complex pathological variations across distinct patient profiles.

Recent works in counterfactual medical image generation have demonstrated impressive results by leveraging autoregressive or diffusion-based architectures to synthesize plausible pathological edits [10,14,17,18]. However, current methods are predominantly constrained by the Supervised Fine-Tuning (SFT) paradigm, which primarily optimizes for pixel-level distribution matching rather than explicit semantic-pathological alignment. To bridge this gap, Reinforcement Learning (RL) offers a promising framework for direct, goal-oriented optimization. However, its application in medical image editing remains challenging by the scarcity of effective reward mechanisms. RL necessitates a reliable reward model capable of providing stable and multidimensional feedback [20,23], yet existing imaging-based reward functions, typically restricted to hand-crafted pixel heuristics (e.g., MSE) [8] or task-specific proxy classifiers, face two challenges. First, optimizing for multiple independent rewards often introduces optimization instability, where conflicting gradients can trigger "reward hacking", leading the model to exploit individual metrics at the expense of overall synthesis fidelity. Second, these isolated rewards provide a reductionist assessment that fails to capture the holistic and nuanced nature of expert clinical judgment. Consequently, a unified and reasoning-aware reward model is essential to provide stable, multidimensional feedback for RL-based medical editing.

To address these challenges, we first curate **MedXEdit-Reward**, a large-scale dataset specifically designed for reward modeling in medical CXR image editing, using a Gemini-based distillation pipeline (Fig. 1). MedXEdit-Reward contains 30,000 quintuplets, each of which includes a source image, an edited image, an edit instruction, and corresponding expert-level rationales paired with comprehensive scores. Second, leveraging this dataset, we develop **MedXEdit-Score**, a specialized reward model that enforces a "reasoning-before-scoring" paradigm to provide semantically grounded feedback for generative edits. Building upon this evaluator, we propose **MedXEdit-GRPO**, a reinforcement learning framework to optimize medical image editing through a unified optimization process (Fig. 2). By employing Group Relative Policy Optimization (GRPO), our framework harmonizes diverse clinical objectives into a single, stable training signal. This approach effectively overcomes the optimization instability and reductionist assessment inherent in fragmented reward systems, ensuring superior alignment with clinical directives. Our contributions can be summarized as follows:

1. We introduce MedXEdit-Reward, a large-scale dataset specifically designed for reward modeling in medical image editing, containing 30,000 quintuplets that pair clinical instructions and images with expert-level rationales.
2. We develop MedXEdit-Score, a logic-driven reward model that enforces a "reasoning-before-scoring" paradigm, providing interpretable and semanti-

cally grounded feedback to bridge the gap between abstract clinical directives and visual modifications.
3. We propose MedXEdit-GRPO, a reinforcement learning framework that optimizes medical image editing via GRPO, effectively harmonizing diverse clinical objectives into a unified training signal.
4. Quantitative and qualitative evaluations demonstrate that MedXEdit-GRPO significantly outperforms state-of-the-art methods in both instruction compliance and anatomical consistency, highlighting its potential for robust and reliable counterfactual medical analysis.

2 Related Work

General Image Editing. Driven by the rapid advancement of diffusion models [22,15] and unified multimodal architectures [28], general image editing has achieved remarkable progress [2,30]. Recent works [28,6] have demonstrated the capability to precisely understand instructions and efficiently accomplish image editing tasks. Building upon this, to further enhance model performance on complex editing instructions, researchers have begun incorporating reinforcement learning [24,16] into this field to achieve more precise alignment with user editing intent. Nevertheless, this highly promising technical paradigm remains in its early exploratory stages for applications in high-stakes specialized scenarios such as medical image editing.

Counterfactual Medical Image Generation. Early works such as Biomed-Journey [10] and PIE [14] established the foundation for modeling longitudinal clinical trajectories through image-editing paradigms. To refine spatial precision, InstructX2X [18] introduces region-specific editing to prevent unintended modifications to non-target anatomical or demographic attributes. ProgEmu [17] further addresses the "silent prediction" problem by jointly synthesizing counterfactual images and their corresponding textual interpretations. More recently, UniMedVL [19] has emerged as a unified medical foundation model, where counterfactual generation is treated as one of its diverse multimodal tasks. Despite these architectural advancements, existing methodologies predominantly rely on the SFT paradigm, which often fails to fully elicit the model's multimodal reasoning capabilities for instruction-conditioned editing. Our MedXEdit-GRPO bridges this gap by introducing RL for reasoning-aware policy alignment, ensuring superior generative fidelity and instruction compliance.

3 Methodology

In this section, we present the MedXEdit-GRPO framework, which integrates clinical logic into generative optimization through two main stages: (1) the construction of a reasoning-aware reward model, MedXEdit-Score, to provide semantically grounded feedback, and (2) the iterative alignment of a multimodal generative policy using GRPO.

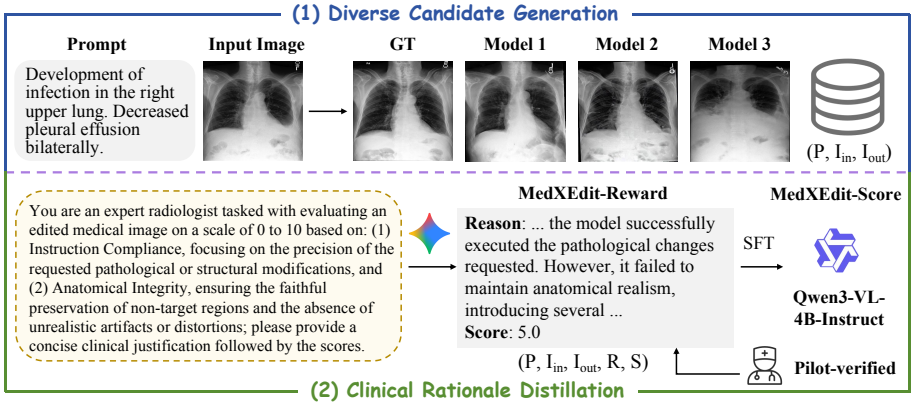

Fig. 1: The MedXEdit-Score construction pipeline. (1) Diverse Candidate Generation: Multiple generative models synthesize varied edited outputs to curate training triplets. (2) Clinical Rationale Distillation: Clinical rationales and scores are distilled from Gemini-3-Flash to form the MedXEdit-Reward dataset. Finally, the MedXEdit-Score model is developed by fine-tuning the Qwen3-VL-4B backbone on this pilot-verified data.

3.1 Reward Modeling

Problem Formulation. Inspired by [27,16], we formulate reward modeling as a conditional textual generation task. We fine-tune the Qwen3-VL-4B backbone using a standard autoregressive objective to serve as a specialized medical evaluator. Given the input triplet (P, I_{in}, I_{out}), where P is the editing instruction, I_{in} is the source medical image, and I_{out} is the edited output image, our MedXEdit-Score is trained to output a sequence consisting of a clinical rationale R followed by a scalar score S. This "reasoning-before-scoring" mechanism not only provides transparent clinical justifications but also leverages the model's internal logic to significantly enhance the sensitivity and accuracy of the final reward.

MedXEdit-Reward Curation Pipeline. MedXEdit-Reward is derived from the ICG-CXR [17] dataset, which comprises 10,679 training and 760 test samples. Each original sample in ICG-CXR is a triplet (P, I_{in}, I_{target}), where I_{target} denotes the ground-truth edited image. To broaden the variety of editing quality for reward learning, the diverse edited outputs I_{out} are generated by three medical generative models: ProgEmu [17], InstructX2X [18], and UniMedVL [19]. For each model, we systematically adjust the inference hyperparameters to produce a wide spectrum of image qualities. As shown in Fig. 1, for each candidate output, we employ Gemini-3-Flash, which has demonstrated robust clinical reasoning capabilities [3], to annotate numerical scores ranging from 0 to 10 and provide rationales considering instruction following and anatomical consistency. Next, we apply filtering based on group-wise maximum scores to exclude unachievable tasks and group standard deviation to remove cases with low discriminability, resulting in a final 30,000 training samples.

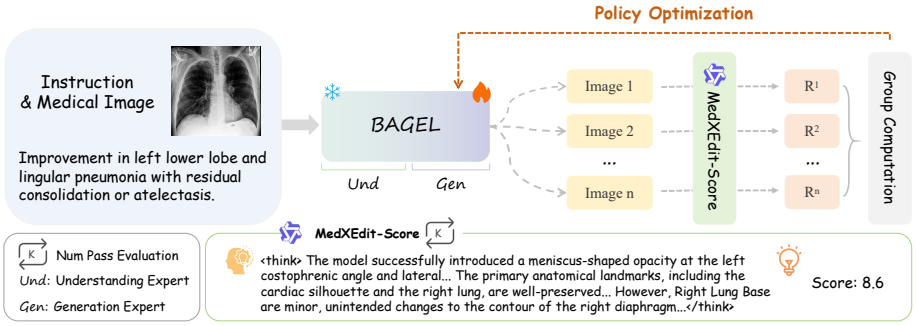

Fig. 2: Overview of the MedXEdit-GRPO. BAGEL, with its understanding expert frozen and generation expert trained, receives instructions and medical images as input to generate a group of edited images. The MedXEdit-Score model then evaluates these images against the instructions using an N-Pass strategy, assigning scores accompanied by clinical rationales. Consequently, these rewards drive the MedXEdit-GRPO policy optimization to refine the BAGEL iteratively.

Reward Model Evaluation. To evaluate the model's discriminative ability, we follow the evaluation protocol of [16] to transform the 760 test samples into a set of 2,000 pairwise preference tuples. Specifically, for a given input image and instruction, we form pairs of candidate edits (A, B), where A is preferred over B based on the oracle's assessment. To ensure the reliability of this gold standard, 200 pairs were randomly selected for verification by board-certified radiologists, achieving an inter-rater agreement of 92.41%. The evaluation is quantified by the preference prediction accuracy, calculated as the proportion of pairs where the model assigns a higher score to the preferred output, i.e., $S(A) > S(B)$.

3.2 MedXEdit-GRPO

As shown in Fig. 2, we validate the effectiveness of MedXEdit-Reward based on Bagel [6], a multimodal foundation architecture with understanding and generation capabilities. Given the tuple (P, I_{in}), the source image I_{in} is first projected into a sequence of visual tokens $V \in \mathbb{R}^{k \times d}$ using a vision encoder. These visual tokens, along with the text, simultaneously condition the generation expert to generate the edited image.

Reinforcement learning (RL) aims to improve the model's performance by maximizing expected cumulative reward. Group Relative Policy Optimization (GRPO) eliminates the value function and estimates the advantage in a group-relative manner compared to PPO [23]. For each input (P, I_{in}), we sample a group of G independent candidate outputs $\{z_0^1, \ldots, z_0^G\}$ from the current policy π_θ. The Advantage A_i is computed by normalizing the MedXEdit-Score reward r_i across the group, as follows:

$$A_i = \frac{r_i - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G)} \tag{1}$$

Table 1: Performance comparison of different methods for disease progression image editing. The best results are highlighted in **bold** and the second-best results are underlined.

Method	Instruct Alignment			Identity		FID ↓	Holistic Score↑
	AUC↑	CLIP-T↑	CLIP-I↑	Race↑	Age↑		
Real images (GT)	-	38.08	100.00	98.95	91.06	-	-
PIE [14]	77.35	34.03	-	-	-	-	-
BiomedJourney [10]	83.85	34.59	-	-	-	-	-
CXR-IRGen [25]	52.36	32.51	75.18	-	-	-	-
ProgEmu [17]	78.42	36.19	89.79	93.56	81.86	0.64	5.58
InstructX2X [18]	85.25	37.08	89.95	96.51	84.66	1.19	4.69
UniMedVL [19]	82.46	**38.48**	90.02	94.56	81.90	0.49	6.88
MedXEdit-GRPO	**86.40**	37.97	**91.60**	**97.93**	**86.74**	**0.40**	**7.84**

The GRPO objective function is formulated as:

$$\mathcal{L}_{GRPO}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\left[\min\left(\rho_i(\theta)A_i, \text{clip}(\rho_i(\theta), 1-\epsilon, 1+\epsilon)A_i\right) - \beta D_{KL}(\pi_\theta||\pi_{ref})\right], \tag{2}$$

where $\rho_i(\theta) = \frac{\pi_\theta(z^i|X)}{\pi_{\theta_{old}}(z^i|X)}$ is the importance sampling ratio. KL divergence penalty D_{KL} is incorporated to constrain the policy π_θ from deviating excessively from the reference model π_{ref}.

In RL training, we enhance reward accuracy through two strategies: (1) We require the model to first think across two dimensions before outputting a score: instruction following and consistency. (2) We perform an *N-Pass* evaluation, scoring each sample multiple times and averaging the results to derive the final reward, which helps mitigate the volatility of individual assessments.

4 Experiments

4.1 Experimental Setup

Implementation Details. For the MedXEdit-Score model, we conduct full-parameter SFT utilizing the LLaMA-Factory framework. Training was performed for 5 epochs with a batch size of 32, employing the AdamW optimizer at a learning rate of 10^{-5}. For the MedXEdit-GRPO model, we initialize its parameters with pre-trained weights from the medical foundation model UniMedVL [19] and perform RL training via LoRA with a rank of $r = 64$. All medical images are processed at a resolution of 512×512. We train our model on 24 GPUs with a group size of 8 and a learning rate of 10^{-4} using bf16 mixed precision. For the

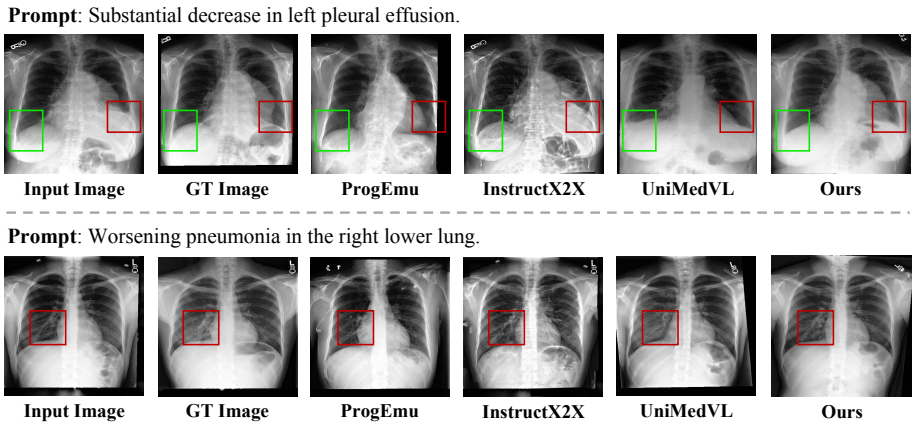

Fig. 3: Qualitative results of counterfactual images, where red and green bounding boxes highlight the targeted modifications and the preserved non-target regions, respectively.

diffusion process, we employ 25 timesteps during training and 50 steps for evaluation. During inference, the classifier-free guidance scales for text and image are set to 4.0 and 1.0, respectively, to balance editing fidelity and visual quality.

Evaluation Metrics. We evaluate our model across four key dimensions: edit-instruction alignment, demographic identity preservation, generative quality, and holistic assessment. Specifically, edit-instruction alignment is measured by AUC from a pathology classifier [4] for clinical accuracy, and CLIP-T/CLIP-I scores via BiomedCLIP [29] for semantic and image consistency. To ensure demographic identity preservation, we use the state-of-the-art race classifier [9] and age classifier [11] to evaluate identity drift. Generative synthesis quality is assessed using Fréchet Inception Distance (FID) with XRV [5] to evaluate visual realism. Finally, we report a Holistic Score from MedXEdit-Score as an integrative measure of overall editing performance and clinical plausibility.

4.2 Experimental Results

Quantitative Comparison with State-of-the-Art. We compare MedXEdit-GRPO to six state-of-the-art methods for counterfactual generation. As shown in Table 1, MedXEdit-GRPO achieves superior results across nearly all metrics. It leads in Instruct Alignment with an AUC of 86.40 and a CLIP-I of 91.60, while remaining highly competitive in CLIP-T. Notably, our model excels in Identity Preservation, reaching the highest Race (97.93) and Age (86.74) scores, effectively suppressing unintended demographic modifications. Furthermore, it attains the best Visual Fidelity with a remarkably low FID (0.40), demonstrating that our RL-based approach successfully harmonizes precise editing with robust anatomical consistency.

Table 2: Reward model performance comparison across different medical LVLMs.

Method	Lingshu-I	HuatuoGPT-V	Hulu-Med-4B	Qwen3-VL-4B	Ours (w/o reas.)	**Ours (Full)**
Score ↑	31.97	36.96	37.72	37.85	71.48	**74.04**

Qualitative Comparison of Counterfactual Images. Fig. 3 demonstrates MedXEdit-GRPO's capability to balance precise pathological editing with anatomical preservation. Compared to baselines, our model achieves a more realistic reduction of pleural effusion (red boxes) while maintaining strict consistency in unmentioned regions (green boxes). Similarly, for "worsening pneumonia", our method generates consolidations that best align with the clinical directives with superior visual fidelity.

Analysis of Reward Model Performance. As shown in Table 2, MedXEdit-Score achieves a peak of 74.04, nearly doubling its backbone (37.85) and significantly surpassing specialized models like Hulu-Med-4B. This margin suggests that general LVLMs lack the sensitivity required for complex editing evaluation. Furthermore, our ablation confirms that the "reasoning-before-scoring" paradigm further elevates performance (74.04 vs. 71.48), enhancing the model's discriminatory power and providing a more accurate reward signal for optimization.

Table 3: Ablation study for reward model.

Reward Model	**AUC**↑	**CLIP-T**↑	**CLIP-I**↑	**Race**↑	**Age**↑	**FID**↓
Hybrid Reward	83.87	37.46	89.74	92.69	79.27	0.50
MedXEdit-Score	**86.40**	**37.97**	**91.60**	**97.93**	**86.74**	**0.40**

Ablation Study. We evaluate the necessity of a unified reward model by replacing MedXEdit-Score with a heuristic baseline, which uses a weighted average of the pathological classifier's AUC, MSE, and SSIM as the reward signal. As shown in Table 3, the significant performance gap proves that conventional, disjoint metrics are insufficient for capturing the complex clinical logic and multi-faceted constraints inherent in medical image editing, underscoring the necessity of a unified evaluator for stable policy optimization.

5 Conclusion

We introduce MedXEdit-GRPO, a reinforcement learning framework for counterfactual medical image editing. To enable reasoning-aware optimization, we curate a large-scale MedXEdit-Reward dataset and develop MedXEdit-Score, a specialized reward model enforcing a "reasoning-before-scoring" paradigm. By optimizing the generative policy through GRPO, our framework effectively harmonizes diverse clinical objectives into a unified training signal. Experimental

results demonstrate that MedXEdit-GRPO significantly outperforms existing methods in both instruction compliance and anatomical consistency, highlighting its potential for robust medical analysis.

Acknowledgments. This work was supported by National Natural Science Foundation of China (No. 62576107), the decision support system for Shanghai Municipal Health Commission, early warning and auxiliary diagnosis and treatment of pediatric severe pneumonia (No. 2025ZHYL040), Shanghai Municipal Commission of Economy and Informatization, Corpus Construction for Large Language Models in Pediatric Respiratory Diseases (No. 2024-GZL-RGZN-01013), and 2025 National Major Science and Technology Project — Noncommunicable Chronic Diseases-National Science and Technology Major Project, Research on the Pathogenesis of Pancreatic Cancer and Novel Strategies for Precision Medicine (No. 2025ZD0552303). Rui Feng and Xiaobo Zhang are corresponding authors.

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

References

1. Bedel, H.A., Çukur, T.: Dreamr: Diffusion-driven counterfactual explanation for functional mri. IEEE Transactions on Medical Imaging **44**(9), 3654–3669 (2024)
2. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
3. Choi, B., Bae, S., Kweon, S., Choi, E.: Kormedmcqa-v: A multimodal benchmark for evaluating vision-language models on the korean medical licensing examination. arXiv preprint arXiv:2602.13650 (2026)
4. Cohen, J.P., Brooks, R., En, S., Zucker, E., Pareek, A., Lungren, M.P., Chaudhari, A.: Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays. In: Medical imaging with deep learning. pp. 74–104. PMLR (2021)
5. Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M., et al.: Torchxrayvision: A library of chest x-ray datasets and models. In: International Conference on Medical Imaging with Deep Learning. pp. 231–249. PMLR (2022)
6. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
7. Fontanella, A., Mair, G., Wardlaw, J., Trucco, E., Storkey, A.: Diffusion models for counterfactual generation and anomaly detection in brain images. IEEE Transactions on Medical Imaging (2024)
8. Furuta, R., Inoue, N., Yamasaki, T.: Fully convolutional network with multi-step reinforcement learning for image processing. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 3598–3605 (2019)
9. Gichoya, J.W., Banerjee, I., Bhimireddy, A.R., Burns, J.L., Celi, L.A., Chen, L.C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.C., et al.: Ai recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health **4**(6), e406–e414 (2022)

10. Gu, Y., Yang, J., Usuyama, N., Li, C., Zhang, S., Lungren, M.P., Gao, J., Poon, H.: Biomedjourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys. arXiv preprint arXiv:2310.10765 (2023)
11. Ieki, H., Ito, K., Saji, M., Kawakami, R., Nagatomo, Y., Takada, K., Kariyasu, T., Machida, H., Koyama, S., Yoshida, H., et al.: Deep learning-based age estimation from chest x-rays indicates cardiovascular prognosis. Communications Medicine **2**(1), 159 (2022)
12. Ktena, I., Wiles, O., Albuquerque, I., Rebuffi, S.A., Tanno, R., Roy, A.G., Azizi, S., Belgrave, D., Kohli, P., Cemgil, T., et al.: Generative models improve fairness of medical classifiers under distribution shifts. Nature Medicine **30**(4), 1166–1173 (2024)
13. Lee, S.I., Topol, E.J.: The clinical potential of counterfactual ai models. The Lancet **403**(10428), 717 (2024)
14. Liang, K., Cao, X., Liao, K.D., Gao, T., Ye, W., Chen, Z., Cao, J., Nama, T., Sun, J.: Pie: Simulating disease progression via progressive image editing. arXiv preprint arXiv:2309.11745 (2023)
15. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
16. Luo, X., Wang, J., Wu, C., Xiao, S., Jiang, X., Lian, D., Zhang, J., Liu, D., et al.: Editscore: Unlocking online rl for image editing via high-fidelity reward modeling. arXiv preprint arXiv:2509.23909 (2025)
17. Ma, C., Ji, Y., Ye, J., Zhang, L., Chen, Y., Li, T., Li, M., He, J., Shan, H.: Towards interpretable counterfactual generation via multimodal autoregression. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 611–620. Springer (2025)
18. Min, H., You, T., Lee, H., Cho, Y., Cho, S.: Instructx2x: An interpretable local editing model for counterfactual medical image generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 279–288. Springer (2025)
19. Ning, J., Li, W., Tang, C., Lin, J., Ma, C., Zhang, C., Liu, J., Chen, Y., Gao, S., Liu, L., et al.: Unimedvl: Unifying medical multimodal understanding and generation through observation-knowledge-analysis. arXiv preprint arXiv:2510.15710 (2025)
20. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
21. Pérez-García, F., Bond-Taylor, S., Sanchez, P.P., van Breugel, B., Castro, D.C., Sharma, H., Salvatelli, V., Wetscherek, M.T., Richardson, H., Lungren, M.P., et al.: Radedit: stress-testing biomedical vision models via diffusion image editing. In: European Conference on Computer Vision. pp. 358–376. Springer (2024)
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
23. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
24. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
25. Shentu, J., Al Moubayed, N.: Cxr-irgen: An integrated vision and language model for the generation of clinically accurate chest x-ray image-report pairs. In: Pro-

ceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5212–5221 (2024)
26. Singla, S., Eslami, M., Pollack, B., Wallace, S., Batmanghelich, K.: Explaining the black-box smoothly—a counterfactual approach. Medical image analysis **84**, 102721 (2023)
27. Wang, Y., Zang, Y., Li, H., Jin, C., Wang, J.: Unified reward model for multimodal understanding and generation. arXiv preprint arXiv:2503.05236 (2025)
28. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
29. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
30. Zhang, S., Yang, X., Feng, Y., Qin, C., Chen, C.C., Yu, N., Chen, Z., Wang, H., Savarese, S., Ermon, S., et al.: Hive: Harnessing human feedback for instructional visual editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9026–9036 (2024)