Towards Interpretable Counterfactual Generation via Multimodal Autoregression

Chenglong Ma[1,2†], Yuanfeng Ji[3†], Jin Ye[4], Lu Zhang[5], Ying Chen[4], Tianbin Li[4], Mingjie Li[3], Junjun He[2,4(✉)], and Hongming Shan[1(✉)]

[1] Institute of Science and Technology for Brain-Inspired Intelligence; MOE Frontiers Center for Brain Science; Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Ministry of Education), Fudan University, Shanghai, China
[2] Shanghai Innovation Institute, Shanghai, China
[3] School of Medicine, Stanford University, Stanford, California, USA
[4] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[5] The First Affiliated Hospital of Jinan University, Guangzhou, Guangdong, China

Abstract. Counterfactual medical image generation enables clinicians to explore clinical hypotheses, such as predicting disease progression, facilitating their decision-making. While existing methods can generate visually plausible images from disease progression prompts, they produce *silent* predictions that lack interpretation to verify how the generation reflects the hypothesized progression—a critical gap for medical applications that require traceable reasoning. In this paper, we propose **I**nterpretable **C**ounterfactual **G**eneration (ICG), a novel task requiring the joint generation of counterfactual images that reflect the clinical hypothesis and interpretation texts that outline the visual changes induced by the hypothesis. To enable ICG, we present ICG-CXR, the first dataset pairing longitudinal medical images with hypothetical progression prompts and textual interpretations. We further introduce ProgEmu, an autoregressive model that unifies the generation of counterfactual images and textual interpretations. Extensive experimental results demonstrate the superiority of ProgEmu in generating progression-aligned counterfactuals and interpretations, showing significant potential in enhancing clinical decision support and medical education. Project resources are available at: https://progemu.github.io.

Keywords: Counterfactual generation · Multimodal autoregressive model · Disease progression prediction.

1 Introduction

Counterfactual generation, which aims to synthesize hypothetical "what-if" scenarios based on real data, has emerged as a transformative tool for medical

[†]Equal contribution. [✉]Co-corresponding authors.

image analysis [4,15]. By generating realistic counterfactual images from a patient's scan under various clinical scenarios (*e.g.*, disease progression and regression), clinicians can visualize potential outcomes and customize patient-specific interventions, thereby enhancing decision-making and medical education.

Recent advances in generative models have shown promising results in counterfactual medical image generation, particularly in chest X-ray (CXR) analysis [4,11,18]. Although these models can generate visually plausible CXR images from input images and disease progression prompts (*e.g.*, "New small left pleural effusion"), their clinical application is limited by what we call *silent* predictions. In other words, the models produce counterfactual images without providing accompanying texts that explain the specific radiological changes associated with the disease progression (*e.g.*, "The left hemidiaphragm shows new blunting in the posterior pleural sulcus, indicating a new small left pleural effusion"). Consequently, these models fail to establish a transparent, causal chain to validate the counterfactual generation, undermining their utility in educational or decision-support applications that require traceable reasoning.

To this end, we introduce **I**nterpretable **C**ounterfactual **G**eneration (ICG), a novel task that requires the joint generation of a counterfactual image reflecting the input hypothesis and an interpretation text justifying the generated changes, thereby ensuring causal alignment between the produced images and texts. However, ICG presents two key challenges, which can be summarized as follows: (a) *data scarcity*–publicly available datasets rarely include longitudinal studies with paired interpretation for the changes observed between prior and subsequent images, and (b) *disjointed modeling*–existing solutions typically isolate image and text generation [11,14,17,24], hindering effective modeling of vision-language dependencies.

In this paper, we tackle ICG through two key innovations. First, we focus on CXR modality and curate the ICG-CXR dataset using a GPT4-based pipeline (Fig. 1). To our knowledge, ICG-CXR is the first large-scale longitudinal CXR dataset designed for interpretable counterfactual generation. It contains 11,439 quadruples from 7,388 unique patients, each including a prior image, a subsequent image, a progression prompt, and a corresponding textual interpretation of the observed changes. Second, we develop ProgEmu, a multimodal autoregressive model that jointly generates counterfactual images and their corresponding textual interpretations (Fig. 2). By unifying these modalities, ProgEmu ensures that the generated image aligns with the hypothesized progression while the text outlines the underlying radiological changes. Our contributions can be summarized as follows:

- We formalize ICG as a novel task requiring the joint generation of counterfactual images and interpretation texts.
- We curate the ICG-CXR dataset, addressing the critical data scarcity for the ICG task.
- We introduce ProgEmu, a multimodal autoregressive model that unifies image and text generation for interpretable counterfactual generation.

– Quantitative and qualitative results show that ProgEmu achieves state-of-the-art generation of progression-aligned counterfactual images and interpretation texts, highlighting its potential for counterfactual medical analysis.

2 Related Work

Counterfactual Medical Image Generation. Counterfactual medical image generation shows significant potential for simulating hypothetical disease scenarios [1,11,14,18,24] and improving the interpretability of AI diagnostic models [7,9,10,20]. BiomedJourney [11] frames the counterfactual image generation problem as an image-editing task and adapts an image-editing framework [3] to patient journey data for generating counterfactual images. PIE [18] uses diffusion-based [23] progressive image editing to simulate realistic disease progression, although it relies on external region masks and does not explain the generated images. CXR-IRGen [24] combines a diffusion model with a large language model [16] to produce paired CXR images and radiology reports simultaneously, yet the text only describes the counterfactual image without highlighting differences between images. In contrast, our ProgEmu unifies the generation of counterfactual images and interpretation texts within a single model.

Unified Multimodal Generative Models. Recent breakthroughs in generative AI have enabled single model to generate various data types (*e.g.*, image and text) simultaneously [6,26,27], achieving superior performance on multiple tasks. Chameleon [26] discretizes visual and textual data into tokens and trains a generic multimodal transformer from scratch in an autoregressive fashion, enabling seamless generation of images and texts. Emu3 [27] extends this approach and enhances the generation of images, texts, and videos by incorporating a more powerful vision tokenizer and an improved post-training strategy. Following this line of work, we propose an autoregressive model, ProgEmu, that leverages the strengths of unified vision-language processing and autoregressive modeling, making it naturally suitable for the proposed interpretable counterfactual generation task.

3 Methodology

For a consecutive study pair, given a prior image and a disease progression prompt, the ICG task aims to generate a progression-aligned counterfactual image and an interpretation text describing the progression-induced radiological changes between prior and counterfactual images. We address this task by first curating the ICG-CXR dataset and then developing an autoregressive multimodal generative model, ProgEmu, detailed below.

3.1 ICG-CXR Dataset Curation

CXR image-report datasets like MIMIC-CXR [13] (377,110 pairs from 227,827 studies) and CheXpertPlus [5] (223,228 pairs from 187,711 studies) are widely

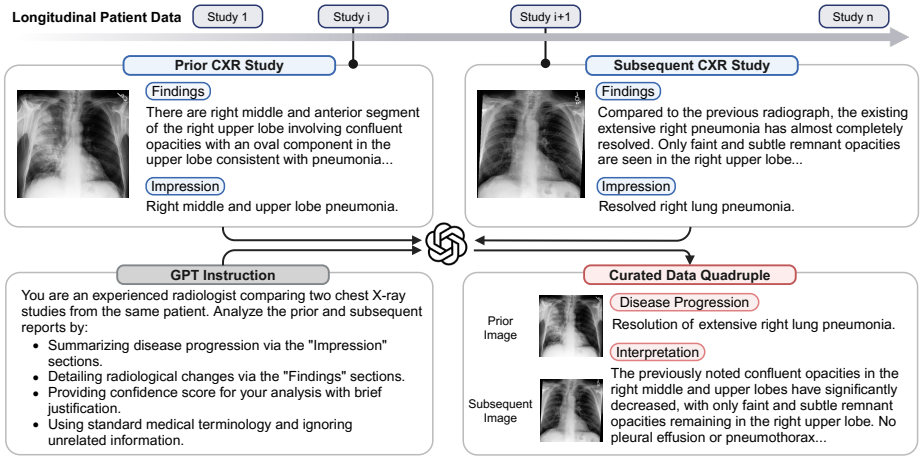

Fig. 1. Dataset curation process of ICG-CXR. Two consecutive CXR studies and their corresponding "Findings" and "Impression" sections are fed into GPT along with instructions, generating a disease progression prompt and an interpretation text. The resulting data quadruple is curated for the interpretable counterfactual generation.

used in medical imaging research. However, these datasets are not directly applicable for the interpretable counterfactual generation task, because they lack comparative descriptions between two CXR studies, and a non-trivial number of studies are of low quality, requiring further processing.

To this end, we curate ICG-CXR, a longitudinal dataset for the interpretable counterfactual generation task, derived from MIMIC-CXR and CheXpertPlus. We start the curation by keeping only those with posterior-anterior view images and discarding any with empty "Impression" or "Findings" sections. For each patient, we form study pairs by selecting a prior CXR study and a subsequent one conducted within 100 days to minimize unrelated confounding factors, and then resize the images to a 512×512 resolution and spatially align them using SimpleITK's registration.

Following this, we employ GPT4 to generate disease progression prompts and interpretation texts. As illustrated in Fig. 1, for each study pair, GPT4 compares the "Impression" sections of the prior and subsequent reports to summarize disease progression and contrasts the "Findings" sections to detail the radiological differences underlying the visual changes observed between the CXRs, while also providing a confidence score based on the clarity and completeness of the reports. As a result, each curated data quadruple in ICG-CXR includes the prior and subsequent CXRs, the disease progression prompt, and the corresponding interpretation text.

After filtering out quadruples with low confidence scores or issues such as poor image quality or spatial misalignment, the final curated dataset comprises 11,439 data quadruples from 7,388 patients, where 7,784 quadruples are sourced from MIMIC-CXR and 3,655 are from CheXpertPlus. The dataset is then split

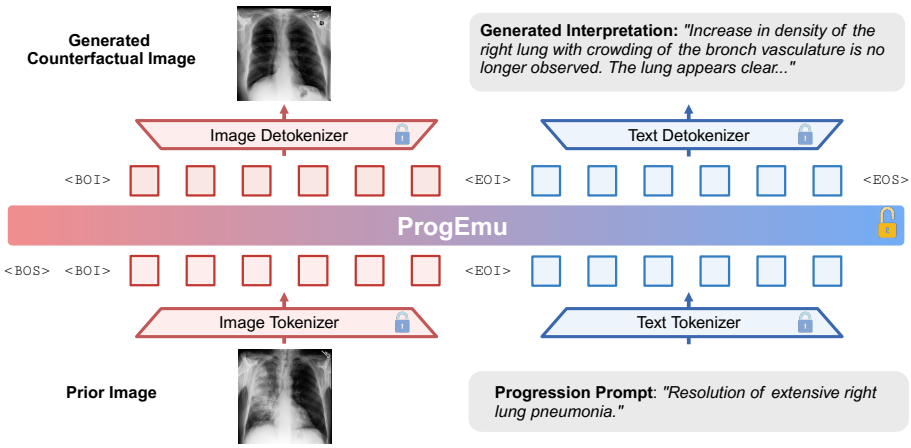

Fig. 2. Overview of ProgEmu, an autoregressive model for interpretable counterfactual generation. The prior image is tokenized along with the textual progression prompt, and the model performs next-token prediction to generate both image and text tokens. These tokens are then detokenized to produce the counterfactual image and corresponding interpretation.

into training and test sets, with 10,679 quadruples used for training and 760 quadruples reserved for testing, ensuring no overlap between patients.

3.2 Multimodal Autoregression for ICG

ICG demands causal alignment between prior and subsequent data and joint modeling of image and text modalities. Existing approaches either do not produce interpretation texts or isolate image generation and text generation [9,17,24], leading to fragmented reasoning and unexplored multi-modality synergy. To this end, we introduce a multimodal autoregressive transformer, ProgEmu.

As illustrated in Fig. 2, ProgEmu integrates both image and text modalities in a unified framework to jointly generate counterfactual images and corresponding interpretation texts. The model begins by tokenizing the prior image, the disease progression prompt, the subsequent image, and the interpretation text into respective discrete token sequences, creating a mixed-modal token sequence formatted as "`<BOS><BOI>`{prior image tokens}`<EOI>`{prompt tokens}`<BOI>`{subsequent image tokens}`<EOI>`{interpretation text tokens}`<EOS>`", where `<BOS>` and `<EOS>` are predefined special tokens indicating the beginning and end of the entire sequence, and `<BOI>` and `<EOI>` separate the modalities.

During training, this mixed-modal token sequence is fed into ProgEmu. The model observes the input part (*i.e.*, prior image tokens and progression prompt tokens) and is tasked with predicting the target part (*i.e.*, subsequent image tokens and interpretation text tokens) in an autoregressive manner, minimizing the cross-entropy loss between its generated token and the target token at each

Table 1. Quantitative evaluation for state-of-the-art methods on the ICG-CXR dataset. The best results are highlighted in **bold** and the second best results are underlined. ProgEmu^2: a cascade variant of ProgEmu for ablation study.

	Counterfactual Image				Explanatory Text		
Method	FID↓	CLIP-I↑	AUROC↑	F1↑	BLEU-3↑	METEOR↑	ROUGE-L↑
PIE [18]	37.60	34.03	0.7735	0.8894	–	–	–
BiomedJourney [11]	36.62	34.59	**0.8385**	0.8411	–	–	–
CXR-IRGen [24]	35.39	32.51	0.5236	0.7609	0.0448	0.2115	0.1846
ProgEmu^2 (**ours**)	29.51	35.00	0.7951	0.8853	**0.2492**	**0.5660**	0.2496
ProgEmu (**ours**)	**29.21**	**35.24**	0.7921	**0.8914**	0.1241	0.4097	**0.2606**

step in the sequence:

$$\mathcal{L} = -\sum_{i} \log p_\theta(v_i | u_1, ..., u_N, v_1, ..., v_{i-1}), \tag{1}$$

where θ denotes the parameters of ProgEmu, and u_* and v_* denote tokens in the input and target parts, respectively. By representing visual and textual data as discrete tokens, ProgEmu seamlessly generates counterfactual images and interpretation texts within a single model, and its autoregressive nature ensures causality between the input-counterfactual image pair and the generated interpretation, suitable for the ICG task.

4 Experiments

4.1 Experimental Setup

Implementation Details. Training a well-performing multimodal model from scratch demands significant computational costs [25]. Instead, we reuse the autoregressive multimodal model, Emu3 [27], to initialize the weights of ProgEmu, and adopt its image tokenizer that encodes a 512×512 image into $4,096$ discrete tokens. We train ProgEmu for 30 epochs with a batch size of 8 on 4 NVIDIA H100 GPUs using AdamW [21] optimizer. The base learning rate is 10^{-5}, following a cosine annealing with minimum learning rate of 10^{-6}.

Evaluation Metrics. For the generated counterfactual images, we evaluate their visual quality, prompt alignment, and preservation of pathological information. Visual quality is assessed using Fréchet Inception Distance (FID) [12] with a pathology classifier [8] pretrained on multiple CXR datasets. The alignment with the prompt is measured by CLIP-I score using BiomedCLIP [28]. For pathological information preservation, following BiomedJourney [11], we apply the pathology classifier to the counterfactual images, comparing predicted pathologies with labels from radiology reports in the subsequent CXR study, and calculate the area under the Receiver Operating Curve (AUROC) and F1 score. For the generated interpretation texts, we use BLUE-3 [22], METEOR [2] and ROUGE-L [19] scores, which are widely applied in text generation evaluation.

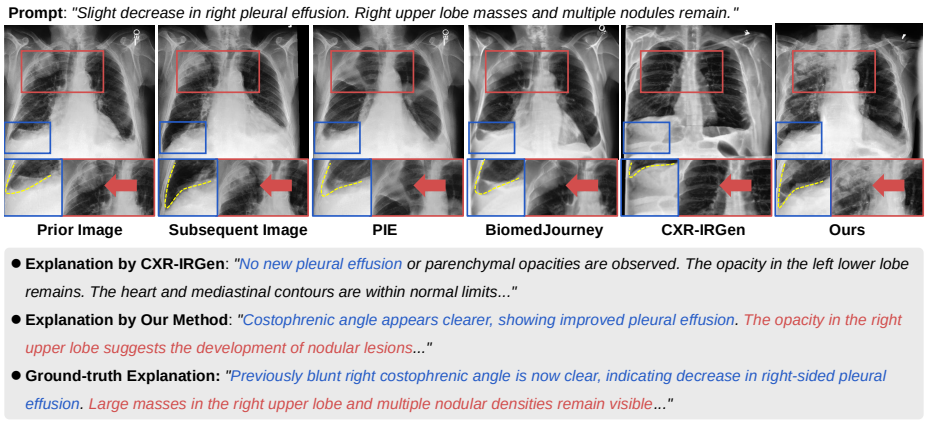

Fig. 3. Qualitative comparison of counterfactual images generated by different methods with enlarged regions of interest. For each image, the right costophrenic angle is outlined by a yellow dashed curve, and the masses and nodules in the right upper lobe are indicated by a red arrow.

4.2 Comparison with State-of-the-Art

We compare ProgEmu to three state-of-the-art methods for counterfactual generation, including PIE [18], BiomedJourney [11], and CXR-IRGen [24]. To adapt these models to the new ICG task, we finetune them on the training data of ICG-CXR dataset. The comparison results are presented in Table 1.

While BiomedJourney and PIE show competitive image generation capability, they do not provide textual interpretation for the counterfactual image, limiting their utility in clinical applications. CXR-IRGen achieves competitive FID, but the AUROC and F1 score are relatively low. This may be attributed to its architecture, where the large language model tries to generate interpretation texts from solely the counterfactual image without considering the prior one, forcing the counterfactual embedding to retain features from the prior image and ultimately leading to a loss of pathological features in the counterfactual image. In contrast, by unifying image and text generation into one single model, our proposed ProgEmu demonstrates superiority over state-of-the-art counterfactual generation methods in ICG task, providing understandable insights for the predicted disease progression.

Fig. 3 provides a qualitative example, where the prior image shows pleural effusion and nodules in the right upper lobe. Given the prompt "Slight decrease in right pleural effusion. Right upper lobe masses and multiple nodules remain", ProgEmu generates a counterfactual image that closely matches the real subsequent image, where the decrease of pleural effusion is evidenced by a sharper right costophrenic angle. The nodules indicated by the opacities in the right upper lobe, which are blurred or removed in the images generated by other methods, remain clearly visible in ProgEmu's output. Notably, the radiologi-

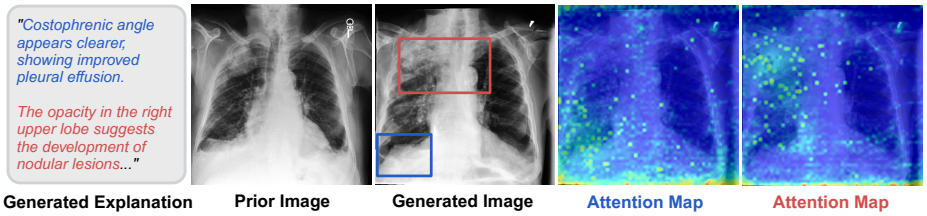

Fig. 4. Cross-attention maps between image and text tokens generated by ProgEmu. The generated interpretation (left) includes color-coded phrases ("Costophrenic angle..." in blue, "The opacity..." in red). The bounding boxes on the generated image (center) match these phrases, and the attention maps (right) show how the model aligns text tokens with the corresponding region.

cal changes suggestive of the disease progression are also well captured in the interpretation texts generated by ProgEmu.

To further investigate whether ProgEmu internally learns to explain the counterfactuals, we visualize the cross-attention maps between generated image tokens and text tokens in Fig. 4, where the region highlighted in each attention map is consistent with the generated texts. This demonstrates ProgEmu's ability to explain the radiological changes in the generated image, establishing a causal chain for interpretable counterfactual generation.

4.3 Ablation Study

To better understand the contribution of unified image and text generation, we develop ProgEmu^2, a cascade of two networks: an "image-only" network that generates only the counterfactual image without interpretation, and a "text-only" network that generates only the interpretation text. Both networks share the same architecture as ProgEmu and are trained independently with the same dataset and setting.

During inference, the image-only network predicts the counterfactual image, which will be fed into the text-only network along with the prior image and progression prompt for generating the textual interpretation. As shown in the last two rows of Table 1, ProgEmu^2 achieves only competitive performance compared to ProgEmu, despite doubling the model size. This suggests that the joint modeling of image and text (as in ProgEmu) may offer synergistic benefits and facilitate the ICG task.

5 Conclusion and Discussion

We introduce interpretable counterfactual generation (ICG), a novel task aimed at producing progression-aligned counterfactual images with textual interpretation for the progression-induced radiological changes. To enable ICG, we curate

the ICG-CXR dataset, the first longitudinal CXR dataset annotated with progression prompts and explanations. Using this dataset, we develop ProgEmu, a multimodal autoregressive model that generates interpretable counterfactuals. Experimental results demonstrate that ProgEmu outperforms existing counterfactual generation methods, highlighting its potential for advancing counterfactual medical image analysis. While current evaluations rely on quantitative metrics such as FID and BLEU-3, an important future direction is to incorporate clinically informed assessments (*e.g.*, through large language models or expert radiologist review) to better capture the relevance, accuracy, and clinical plausibility of generated outputs. Other future directions include generalizing to more imaging modalities and conditions beyond disease progression.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (No. 62471148).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Alaya, M.B., Lang, D.M., Wiestler, B., Schnabel, J.A., Bercea, C.I.: MedEdit: Counterfactual diffusion-based image editing on brain MRI. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 167–176 (2024)
2. Banerjee, S., Lavie, A.: 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. pp. 65–72 (2005)
3. 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)
4. Chambon, P., Bluethgen, C., Delbrouck, J.B., Van der Sluijs, R., Połacin, M., Chaves, J.M.Z., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A.: Roentgen: vision-language foundation model for chest x-ray generation. arXiv preprint arXiv:2211.12737 (2022)
5. Chambon, P., Delbrouck, J.B., Sounack, T., Huang, S.C., Chen, Z., Varma, M., Truong, S.Q., Chuong, C.T., Langlotz, C.P.: CheXpert Plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. arXiv preprint arXiv:2405.19538 (2024)
6. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
7. 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 (2021)
8. Cohen, J.P., Hashir, M., Brooks, R., Bertrand, H.: On the limits of cross-domain generalization in automated x-ray prediction. In: Medical Imaging with Deep Learning. pp. 136–155 (2020)

9. Fang, Y., Jin, Z., Guo, S., Liu, J., Gao, Y., Ning, J., Yue, Z., Li, Z., Walsh, S.L., Yang, G.: Decoding report generators: A cyclic vision-language adapter for counterfactual explanations. arXiv preprint arXiv:2411.05261 (2024)
10. Fang, Y., Wu, S., Jin, Z., Wang, S., Xu, C., Walsh, S., Yang, G.: Diffexplainer: Unveiling black box models via counterfactual generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 208–218 (2024)
11. 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)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in Neural Information Processing Systems **30** (2017)
13. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019)
14. Kyung, D., Kim, J., Kim, T., Choi, E.: Towards predicting temporal changes in a patient's chest x-ray images based on electronic health records. arXiv preprint arXiv:2409.07012 (2024)
15. Lee, S.I., Topol, E.J.: The clinical potential of counterfactual ai models. The Lancet **403**(10428), 717 (2024)
16. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880 (2020)
17. Li, M., Lin, H., Qiu, L., Liang, X., Chen, L., Elsaddik, A., Chang, X.: Contrastive learning with counterfactual explanations for radiology report generation. In: European Conference on Computer Vision. pp. 162–180 (2024)
18. 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)
19. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
20. Liu, S., Wang, F., Ren, Z., Lian, C., Ma, J.: Controllable counterfactual generation for interpretable medical image classification. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 143–152 (2024)
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
23. 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)
24. 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: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5212–5221 (2024)

25. Shi, W., Han, X., Zhou, C., Liang, W., Lin, X.V., Zettlemoyer, L., Yu, L.: Llamafusion: Adapting pretrained language models for multimodal generation. arXiv preprint arXiv:2412.15188 (2024)
26. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
27. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
28. 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)