InstructX2X: An Interpretable Local Editing Model for Counterfactual Medical Image Generation

Hyungi Min[1], Taeseung You[1], Hangyeul Lee[1]
Yeongjae Cho[1], and Sungzoon Cho[1†]

Seoul National University
{alsgusrl0,luca0621,mikelee,zxc5932,zoon}@snu.ac.kr

Abstract. Counterfactual medical image generation have emerged as a critical tool for enhancing AI-driven systems in medical domain by answering "what-if" questions. However, existing approaches face two fundamental limitations: First, they fail to prevent unintended modifications, resulting collateral changes in demographic attributes when only disease features should be affected. Second, they lack interpretability in their editing process, which significantly limits their utility in real-world medical applications. To address these limitations, we present *InstructX2X*, a novel interpretable local editing model for counterfactual medical image generation featuring *Region-Specific Editing*. This approach restricts modifications to specific regions, effectively preventing unintended changes while simultaneously providing a *Guidance Map* that offers inherently interpretable visual explanations of the editing process. Additionally, we introduce *MIMIC-EDIT-INSTRUCTION*, a dataset for counterfactual medical image generation derived from expert-verified medical VQA pairs. Through extensive experiments, InstructX2X achieve state-of-the-art performance across all major evaluation metrics. Our model successfully generates high-quality counterfactual chest X-ray images along with interpretable explanations, as validated by experienced radiologists. Our code and dataset are publicly available at https://github.com/hgminn/InstructX2X.

Keywords: Counterfactual Image Generation · Chest X-ray · XAI.

1 Introduction

Counterfactual medical image generation is an emerging approach that enhances AI-driven high-stakes medical decision-making. This methodology aims to answer what-if questions such as "How would this medical image change if the patient had a different disease?" [9,27]. By precisely manipulating target features in medical image while preserving unrelated attributes, this technique generates realistic edited images and helps uncover causal structures or biases in AI models. Counterfactual medical image generation offers various applications, such as

[†] Corresponding author: zoon@snu.ac.kr

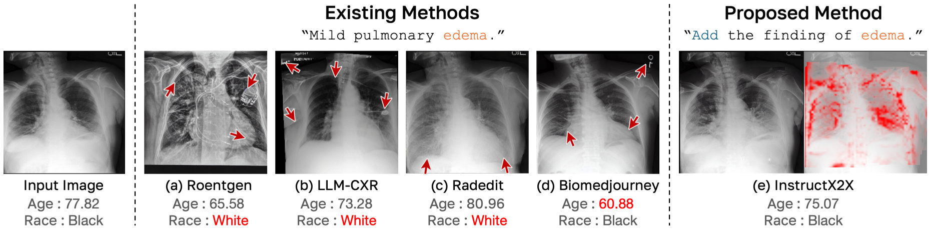

Fig. 1. Comparison of counterfactual medical image generation results between existing methods and our proposed approach. When adding edema features to an input chest X-ray image, existing methods (a–d) demonstrate unintended modifications (red arrows), causing significant variations in age and race (note the demographic predictions below each image). In contrast, InstructX2X preserves the demographic attributes while achieving precise editing and provides a visual explanation via guidance map (red overlay).

evaluating model robustness [23], providing counterfactual explanations [9,4,26], enhancing classifier performance [27,19], and detecting anomalies [11,25].

Despite the promising applications of counterfactual medical image generation, several technical challenges remain unresolved. A critical issue is the unintended modification of unrelated attributes when manipulating target features. In the context of chest radiography, Figure 1 illustrates such failure cases: when adding the edema feature to the input chest X-ray image, methods (a-c) alter the racial characteristics, while method (d) significantly changes the age attribute, despite the fact that these demographic attributes are independent of the edema features. Such unintended modifications distort the original clinical presentation and compromise the validity of the generated images [29].

Another critical challenge in counterfactual medical image generation is a lack of interpretability. Interpretability (such as visual explanation) helps users to understand the model's decision-making process and validate the appropriateness of modifications [9,2]. Current methods predominantly adopt post-hoc explanation techniques for model interpretation [5,13]. Although visually compelling, recent studies have demonstrated that these explanations frequently fail to represent the true decision mechanisms of the underlying models [14,28,24,27]. Such unreliable interpretability severely restricts the utility of counterfactual images for both clinical applications and model evaluation [7,24,2].

To address these two critical limitations, we propose **InstructX2X**, a novel interpretable local editing model for counterfactual medical image generation. Our model introduces a *Region-Specific Editing* approach that restricts editing to specific regions, preventing unintended modifications. Our targeted editing mechanism excludes potential spurious correlations outside the region of interest, resulting in highly reliable counterfactual images. Additionally, our region-specific editing methodology provides a *Guidance Map*, visualized as the red overlay in Figure 1(e), which highlights the modified areas, offering clear visual explanations of how the model processes the editing instructions. InstructX2X achieves inherent interpretability by directly revealing the decision mechanism to users, eliminating the need for post-hoc explanations of uncertain reliability.

Furthermore, the development of reliable counterfactual medical image generation has been constrained by the scarcity of datasets with reliable editing descriptions. To overcome this data deficiency, we repurpose an existing dataset from a different task domain into *MIMIC-EDIT-INSTRUCTION*, a new counterfactual medical image generation dataset. We leverage expert-verified medical VQA pairs, unlike existing approaches that depend on large language models to generate editing descriptions without clinical validations [13,8].

The key contributions of our research are:

1. We propose a novel interpretable local editing model, InstructX2X, which effectively addresses existing challenges in counterfactual medical image generation.
2. We introduce innovative region-specific editing technique to ensure precisely controlled modification and enhance interpretability by providing guidance map.
3. We release *MIMIC-EDIT-INSTRUCTION*, instruction-based editing dataset for counterfactual medical image generation derived from expert-verified medical VQA pairs.
4. InstructX2X demonstrates state-of-the-art performance through extensive experiments, as well as clinically significant interpretability validated by radiologist evaluations.

2 Method

The design of InstructX2X is outlined in Figure 2. In this section, we describe the construction of the *MIMIC-EDIT-INSTRUCTION* dataset and elaborate on the concept of *Region-Specific Editing* with the visual explanation *Guidance Map*.

2.1 Dataset preparation

InstructX2X utilizes three publicly available datasets: MIMIC-CXR [18], MIMIC-Diff-VQA [15], and MS-CXR [6]. MIMIC-CXR contains 377,110 chest X-ray images and 227,827 radiology reports from 63,478 patients, while MIMIC-Diff-VQA builds upon it with 164,324 pairs of longitudinal chest X-rays and 700,703 expert-verified question-answer pairs. MS-CXR provides phrase grounding with 1,162 radiologist-annotated image-sentence pairs across eight diseases.

From MIMIC-Diff-VQA, PA view image pairs are selected and resized to 512 × 512. Registration is performed with SimpleITK [21] following BioViL-T [3], and pairs with low scores are discarded. Only *Difference* type questions are included, excluding ‘nothing has changed’ answers. Class imbalance is addressed by undersampling CheXpert [17] ‘no finding’ cases. Using MIMIC-CXR’s official split with P19 as holdout, the final dataset comprises 11,703 training, 200 validation (official split), and 1,933 test (holdout set) samples.

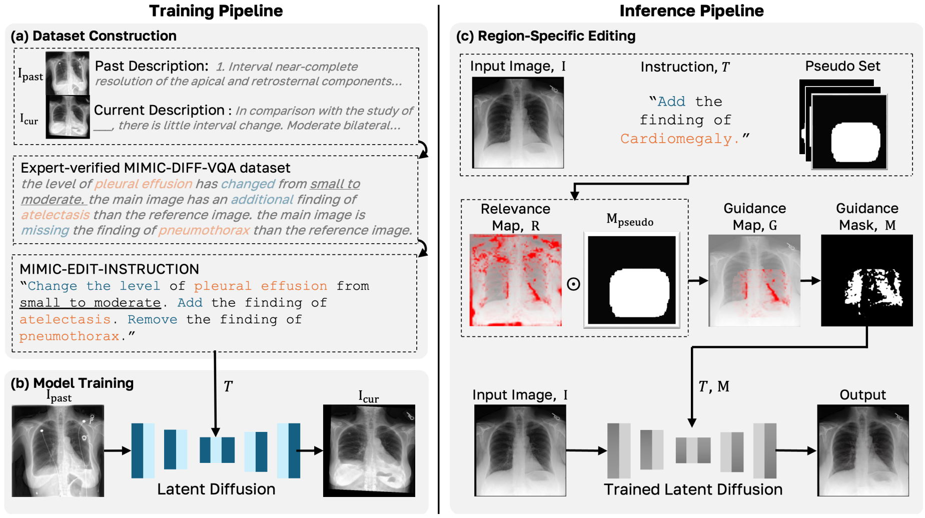
Training Pipeline
(a) Dataset Construction
I_{past}
Past Description: 1. Interval near-complete resolution of the apical and retrosternal components...
I_{cur}
Current Description : In comparison with the study of ___, there is little interval change. Moderate bilateral...
Expert-verified MIMIC-DIFF-VQA dataset
the level of pleural effusion has changed from small to moderate. the main image has an additional finding of atelectasis than the reference image. the main image is missing the finding of pneumothorax than the reference image.
MIMIC-EDIT-INSTRUCTION
"Change the level of pleural effusion from small to moderate. Add the finding of atelectasis. Remove the finding of pneumothorax."
(b) Model Training
I_{past}
T
I_{cur}
Latent Diffusion
Inference Pipeline
(c) Region-Specific Editing
Input Image, I
Instruction, T
"Add the finding of Cardiomegaly."
Pseudo Set
Relevance Map, R
⊙
M_{pseudo}
Guidance Map, G
Guidance Mask, M
T, M
Input Image, I
Trained Latent Diffusion
Output

Fig. 2. Overview of our InstructX2X framework. (a) Dataset construction process that converts expert-verified VQA pairs into *MIMIC-EDIT-INSTRUCTION* data. (b) Training pipeline where model learns to transform I_{past} to I_{cur} using constructed instructions. (c) region-specific editing approach that enables precise and interpretable image editing.

2.2 Dataset Construction

Existing approaches often use LLMs to generate editing descriptions [13,8], but the lack of expert validation may result in clinical inaccuracies. To address this issue, we repurpose the MIMIC-Diff-VQA dataset, which provides expert-verified descriptions of temporal changes in chest X-rays with a 97.33% validation rate. This dataset serves as a reliable source for constructing image-editing instructions—beyond its original VQA purpose. We identified three core operations that form an intuitive instruction of medical image modifications:

- **Add**: Introducing new findings or symptoms.
- **Remove**: Eliminating existing findings or symptoms.
- **Change the level**: Adjusting the severity level of present abnormalities.

As illustrated in Figure 2(a), this expert-verified approach eliminates the need for LLMs in instruction construction by employing a rule-based conversion of difference descriptions. By decomposing complex medical changes into these instructions, we establish *MIMIC-EDIT-INSTRUCTION*, a new dataset for counterfactual medical image generation that maintains clinical precision while providing more precise and intuitive control [8].

2.3 Region-Specific Editing

region-specific editing prevents unintended modifications by precisely editing target regions. This method provides inherent interpretability by generating a guidance map, where the explanation directly reveals the decision process [27]. As shown in Figure 2(c), our approach combines model-derived relevance maps [22] with dataset-derived pseudo masks to achieve precise and interpretable editing.

During inference, our region-specific editing process works as follows. First, given an input image I and an edit instruction T, we encode I into the latent space and inject a fixed amount of Gaussian noise at a chosen diffusion timestep t_{rel}. We then compute two noise predictions by feeding $(z_{t_{rel}}, I, T)$ and $(z_{t_{rel}}, I, T = "")$ into ϵ_θ. The relevance map (R) is obtained as the normalized absolute difference between these two predictions $(\epsilon_{I,T}(z_{t_{rel}}),\epsilon_I(z_{t_{rel}}))$, which highlights regions that require modification according to the edit instruction [22]:

$$R_{x,I,T} = \text{normalize}\Big|\epsilon_{I,T}(z_{t_{rel}}) - \epsilon_I(z_{t_{rel}})\Big|. \tag{1}$$

Next, to precisely localize anatomical regions associated with pathological findings, we incorporate expert-annotated bounding boxes from the MS-CXR dataset to create anatomically-aware pseudo mask. For each of the eight findings, we create individual pathology masks by taking the outer union of overlapping bounding box annotations, forming a pseudo set of eight masks. During inference, the final pseudo mask is generated by selecting each pathology mask from pseudo set that corresponds to the findings mentioned in the editing instruction, and then merging them into a single mask. For findings outside the eight annotated categories, we employ a 512 × 512 mask.

The final guidance map G is computed by multiplying R with the final pseudo mask M_{pseudo}. $\odot$ denotes element-wise multiplication:

$$G = M_{pseudo} \odot R_{x,I,T}, \tag{2}$$

By this multiplication, the guidance map G effectively integrates (i) the model-identified regions to modify from the relevance map with (ii) the disease-related anatomical locations from the pseudo mask. The guidance map then represents pixel-wise information about precisely localized regions that align with editing instructions, serving as a visual explanation of the editing process for users.

We apply a threshold τ on G to form a binary editing mask $M_{x,I,T} = \mathbb{1}(G \geq \tau)$. In each denoising iteration, we keep the unmasked regions identical to the input image by maintaining identical noise pixels in these areas throughout all steps. This approach prevents any modifications outside the region of interest to avoid unrelated spurious correlations. Furthermore, region-specific editing method supports user-defined mask specifications in place of M_{pseudo}, enabling flexible control over editing process and cross-domain adaptability without architectural modifications.

3 Experiment

Implementation details Our model builds upon the pretrained InstructPix2Pix [8]. The architecture employs a frozen CLIP text encoder and a U-Net backbone, processing triplet data $(I_{\text{past}}, I_{\text{cur}}, T)$ from longitudinal chest X-ray

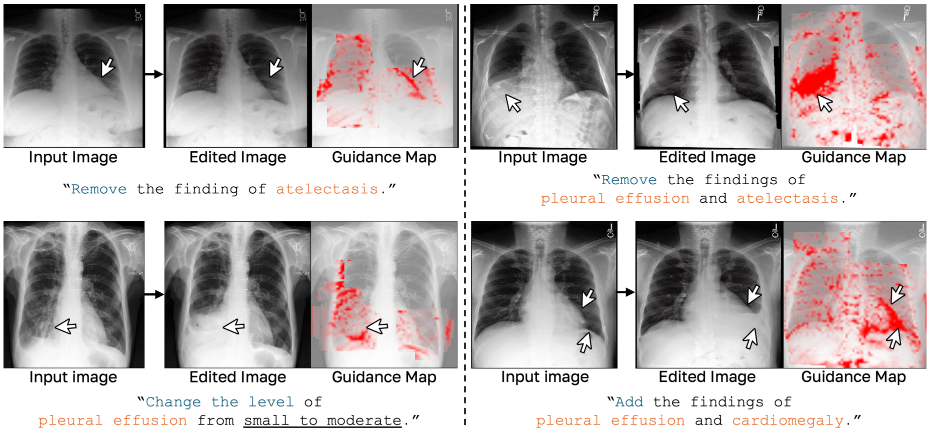

Fig. 3. Examples showing InstructX2X's editing capabilities. Left: Single-finding editing examples. Right: Multi-finding editing examples. Each case includes input image, edited result, and guidance map visualization (red overlay) showing modified regions.

Table 1. Comparison of InstructX2X with baseline methods across CMIG, KL divergence, and FID metrics. GT refers to Ground Truth. Best results are shown in bold.

Model	**CMIG** (↑)				**KL** (↓)	**FID** (↓)
	Patho (↑)	**Race** (↑)	**Age** (↑)	**CMIG** (↑)		
Real images(GT)	84.15	99.60	88.62	90.77	-	-
Roentgen	82.87	50.48	18.71	65.63	51.71	35.96
LLM-CXR	74.46	51.79	9.92	61.95	51.69	54.96
Radedit	83.26	81.82	46.56	80.81	39.01	28.40
BiomedJourney	80.44	83.23	68.95	81.47	22.18	13.77
InstructX2X	80.76	98.81	83.91	**88.03**	**9.69**	**2.88**

Table 2. Ablation study on the effects of region-specific editing components on model performance. RelMap refers to Relevance Map and PsMask refers to Pseudo Mask. Best results are shown in bold.

Model	**Guidance**		**CMIG** (↑)				**KL** (↓)	**FID** (↓)
	RelMap	**PsMask**	**Patho**	**Race**	**Age**	**CMIG**		
(a)			80.64	90.43	76.17	84.68	21.26	11.76
(b)	✓		79.93	97.58	82.50	87.06	10.92	3.95
(c)		✓	80.24	98.27	83.29	87.47	13.72	2.89
InstructX2X	✓	✓	80.76	98.81	83.91	**88.03**	**9.69**	**2.88**

pairs and editing instructions, as shown in Figure 2(b). Model trained on 8 A100 GPUs for 4,500 steps with learning rate 1×10^{-4} and batch size 576. During inference, we set $\tau = 0.1$ and fix $s_I = 1.5$ and $s_T = 7.5$.

Baselines We compare against four baseline models: RoentGen [5], a high-quality chest X-ray image generation model from descriptions; LLM-CXR [20], a text-only LLM for CXR vision tasks; BiomedJourney [13], a model that leverages GPT-4 [1] to generate disease progression; and RadEdit [23], a model using multiple masks to ensure consistency. For fair comparison, RoentGen and LLM-CXR (both originally designed for CXR generation, not editing) use provided impressions section in MIMIC-CXR reports, BiomedJourney uses GPT-4 generated descriptions following their implementation, and RadEdit uses impressions and the same pseudo mask as our region-specific editing method.

Metrics We evaluate using the CMIG score [13], KL divergence [13] and FID. CMIG score combines pathology classification accuracy (using DenseNet-121 from XRV [10]) with the preservation of demographic attributes, specifically race [12] and age [16]. These measurements are integrated through geometric means to ensure robustness across different scales. For pathology classification, we focus on five specific findings: Atelectasis, Cardiomegaly, Edema, Pleural Effusion, and Pneumothorax. KL divergence measures the distribution difference between real and generated images' pathology classifications, identifying potential evaluation bias that may inflate performance metrics. FID computed with DenseNet-121 [10] evaluates the visual quality and realism of generated images.

3.1 Results

Table 1 shows InstructX2X's performance compared to previous methods. Our model achieves a CMIG score of 88.03, approaching real test images (90.77). This demonstrates our model's ability to perform precise modifications while preserving patient-specific attributes. Our method maintains near-real image levels in both race (98.81 vs 99.60) and age (83.91 vs 88.62) preservation, while achieving competitive pathology modification (80.76). Furthermore, the significantly lower FID score (2.88) demonstrates our model's superior ability to generate high-quality, realistic images.

In contrast, other methods show significant imbalances. Despite achieving higher pathology scores (Roentgen: 82.87, RadEdit: 83.26), they exhibit substantial degradation in preserving patient characteristics, suggesting a trade-off between modification accuracy and unrelated attribute preservation. Their large KL divergence values (Roentgen: 51.71, RadEdit: 39.01) suggest that their pathology performance could be inflated. it represents a significant discrepancy in predicted distributions when applying the same classifier to both generated and real images. InstructX2X's significantly lower KL divergence (9.69) shows our method avoids such inflation while achieving competitive pathology scores.

Building on the strong quantitative performance discussed above, Figure 3 offers visual evidence of InstructX2X's capabilities in both single-finding and multi-finding modification scenarios. The examples demonstrate the model's ability to execute diverse radiological manipulations with high precision. The distinctive feature of our approach lies in its targeted editing capability, as shown by modifications that occur exclusively within the instruction-specified region. The accompanying guidance maps (red overlays) precisely highlight the regions of interest, providing transparent visual interpretations of the model's editing process—for example, targeting the lung bases for pleural effusion and the cardiac silhouette for cardiomegaly.

3.2 Radiologist Assessment

Two board-certified radiologists (with 11 and 12 years of experiences, respectively) evaluated our model using 40 diverse image pairs selected from our test set. For each pair, radiologists reviewed input, edited, and guidance-map images (120 images total) across five key findings: pleural effusion, cardiomegaly, edema, pneumothorax, and atelectasis. Using a 5-point likert scale, radiologists assessed both performance (natural disease progression and accurate modifications) and interpretability (effectiveness of guidance maps in explaining editing decisions). Both experts gave moderately favorable scores on both performance (M = 3.59, SD = 1.11) and interpretability (M = 3.45, SD = 1.17), indicating the model's ability to generate consistent and reliable modifications along with meaningful visual explanations.

3.3 Ablation studies

Table 2 presents ablation studies that analyze the effectiveness of our region-specific editing method: (a) the trained latent diffusion model without any region-specific editing components, (b) with only the relevance map, (c) with only the pseudo-mask and our full model (InstructX2X) incorporating both components. The results demonstrate the efficacy of our region-specific editing approach. Model (a) shows moderate performance across metrics, while the incorporation of relevance map (b) significantly improves feature preservation and reduces KL divergence (10.92). Model (c) with pseudo mask shows improved retention of patient characteristics (race: 98.27, age: 83.29) and image quality (FID: 2.89). Our full model synergistically combines these benefits, achieving optimal performance across metrics (CMIG: 88.03).

4 Conclusion

InstructX2X addresses two critical limitations of counterfactual medical image generation: unintended modification and insufficient interpretability. Our *Region-Specific Editing* approach achieves precise feature modification while preserving unrelated attributes, constraining the influence of spurious correlations during image generation. The *Guidance Map* offers transparent visual explanations of the modification process, providing inherent interpretability rather than post-hoc explanations of uncertain reliability. By introducing the instruction-based *MIMIC-EDIT-INSTRUCTION* dataset, we establish a more reliable foundation for future work. InstructX2X not only demonstrates state-of-the-art performance across multiple metrics, but it also confirm its clinical validity and explainability via radiologist assessments. These innovations collectively elevate counterfactual medical image generation for high-stakes clinical applications and AI model validation.

Acknowledgments

This work was supported by the BK21 FOUR Program (Education and Research Center for Industrial Innovation Analytics) funded by the Ministry of Education, Korea (No. 4120240214912). This study was approved by the Institutional Review Board of Seoul National University (IRB No. E2501/003-006; IRB No. 2502/004-004).

Disclosure of Interests. The authors have no competing interests.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., Consortium, P.: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making **20**, 1–9 (2020)
3. Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., et al.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15016–15027 (2023)
4. Bedel, H.A., Çukur, T.: Dreamr: Diffusion-driven counterfactual explanation for functional mri. IEEE Transactions on Medical Imaging (2024)
5. Bluethgen, C., Chambon, P., Delbrouck, J.B., van der Sluijs, R., Połacin, M., Zambrano Chaves, J.M., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A.S.: A vision–language foundation model for the generation of realistic chest x-ray images. Nature Biomedical Engineering pp. 1–13 (2024)
6. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
7. Borys, K., Schmitt, Y.A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C.M., Nensa, F.: Explainable ai in medical imaging: An overview for clinical practitioners–beyond saliency-based xai approaches. European journal of radiology **162**, 110786 (2023)
8. 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 (CVPR). pp. 18392–18402 (June 2023)
9. 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)
10. 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)

11. 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)
12. 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)
13. 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)
14. Han, T., Srinivas, S., Lakkaraju, H.: Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. Advances in neural information processing systems **35**, 5256–5268 (2022)
15. Hu, X., Gu, L., An, Q., Zhang, M., Liu, L., Kobayashi, K., Harada, T., Summers, R.M., Zhu, Y.: Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 4156–4165 (2023)
16. 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)
17. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
18. 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)
19. 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)
20. Lee, S., Kim, W.J., Chang, J., Ye, J.C.: Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. In: The Twelfth International Conference on Learning Representations
21. Lowekamp, B.C., Chen, D.T., Ibáñez, L., Blezek, D.: The design of simpleitk. Frontiers in neuroinformatics **7**, 45 (2013)
22. Mirzaei, A., Aumentado-Armstrong, T., Brubaker, M.A., Kelly, J., Levinshtein, A., Derpanis, K.G., Gilitschenski, I.: Watch your steps: Local image and scene editing by text instructions. In: ECCV (2024)
23. 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)
24. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)

25. Sanchez, P., Kascenas, A., Liu, X., O'Neil, A.Q., Tsaftaris, S.A.: What is healthy? generative counterfactual diffusion for lesion localization. In: MICCAI Workshop on Deep Generative Models. pp. 34–44. Springer (2022)
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. Sun, S., Woerner, S., Maier, A., Koch, L.M., Baumgartner, C.F.: Inherently interpretable multi-label classification using class-specific counterfactuals. In: Medical Imaging with Deep Learning. pp. 937–956. PMLR (2024)
28. White, A., d'Avila Garcez, A.: Measurable counterfactual local explanations for any classifier. In: ECAI 2020, pp. 2529–2535. IOS Press (2020)
29. Xia, T., Roschewitz, M., De Sousa Ribeiro, F., Jones, C., Glocker, B.: Mitigating attribute amplification in counterfactual image generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 546–556. Springer (2024)