PRISM: High-Resolution & Precise Counterfactual Medical Image Generation using Language-guided Stable Diffusion

Amar Kumar[1,2] AMAR.KUMAR@MAIL.MCGILL.CA
Anita Kriz[1,2] ANITA.KRIZ@MAIL.MCGILL.CA
Mohammad Havaei[3] MHAVAEI@GOOGLE.COM
Tal Arbel[1,2] ARBEL@CIM.MCGILL.CA
[1] *Center for Intelligent Machines, McGill University, Montreal, Canada.*
[2] *MILA (Quebec AI institute), Montreal, Canada.*
[3] *Google Research, Montreal, Canada.*

Editors: Under Review for MIDL 2025

Abstract

Developing reliable and generalizable deep learning systems for medical imaging faces significant obstacles due to spurious correlations, data imbalances, and limited text annotations in datasets. Addressing these challenges requires architectures robust to the unique complexities posed by medical imaging data. The rapid advancements in vision-language foundation models within the natural image domain prompt the question of how they can be adapted for medical imaging tasks. In this work, we present PRISM, a framework that leverages foundation models to generate high-resolution, language-guided medical image counterfactuals using Stable Diffusion. Our approach demonstrates unprecedented precision in selectively modifying spurious correlations (the medical devices) and disease features, enabling the removal and addition of specific attributes while preserving other image characteristics. Through extensive evaluation, we show how PRISM advances counterfactual generation and enables the development of more robust downstream classifiers for clinically deployable solutions. To facilitate broader adoption and research, we make our code publicly available at https://github.com/Amarkr1/PRISM.

Keywords: Counterfactual Image Synthesis, Diffusion, Foundation Models, Generative Models, Image Synthesis, Large Language Models

1. Introduction

The development of deep learning models in healthcare settings has the potential to transform current medical practices in disease diagnosis, biomarker discovery, and personalized treatment. However, clinical deployment requires robust models – a standard that remains largely unmet due to the inherent complexities of medical imaging data. Class imbalances and spurious correlations can cause models to learn misleading patterns that are not penalized when optimizing the training objective. This flawed training paradigm results in incorrect disease classification, ultimately degrading the model's generalizability to real-world clinical scenarios. To address these challenges, the field has explored counterfactual generation to expose shortcut learning and alleviate data imbalance issues by augmenting underrepresented classes. Previous work has focused on classifier-guided counterfactual (CF) image generation methods, such as using standard classifiers with robust empirical minimization techniques (Mertes et al., 2022; Singla et al., 2019) or classifiers based on distributional robust optimization (Group-DRO) (Kumar et al., 2023). An alternative approach leverages Structural Causal Models (SCMs) to explicitly model and intervene on causal relationships between attributes during the generation process; these methods also (largely) rely on classifiers to produce high-quality results (Ribeiro et al., 2023). These methods expose a paradox in their formulation – their performance is dependent on the same biased data (and classifiers) they are designed to mitigate (see Fig. 1). Moreover, end-to-end architectures face a tradeoff between competing objectives: high-quality generation demands fine-grained details, while classification relies on abstract features. Compounded by the computational burden of training high-capacity architectures from scratch, synthesizing high-resolution and precise CFs remains elusive.

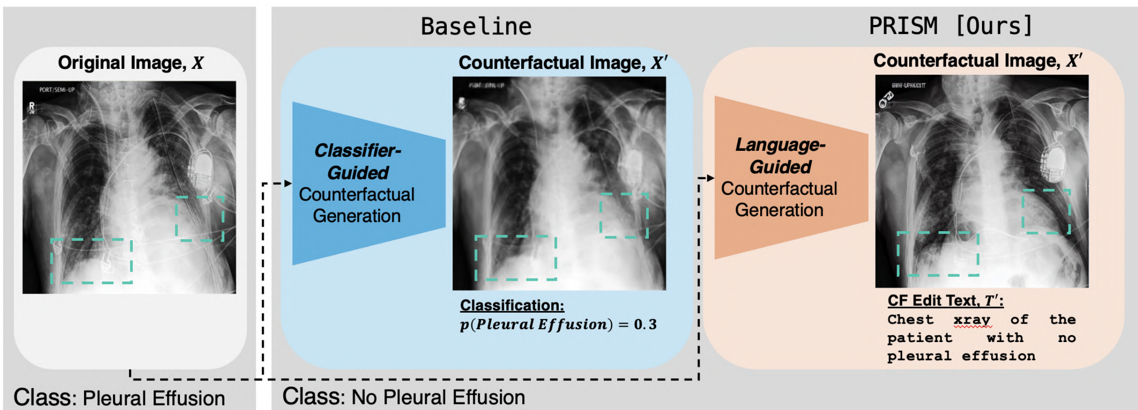


Figure 1: Counterfactual (CF) explanations for a subject with 'Pleural Effusion'. (a) Original chest radiograph of subject; (b) Classifier-guided CF image fails to show changes in the diseased area and determined the CF image is healthy. The classifier is biased and associates the disease with the medical device; (c) PRISM modifies the area of disease pathology, leaving the devices (e.g. pacemaker) unchanged.

Fine-tuning foundation models has recently emerged at the forefront of deep learning for medical image analysis (Wang et al., 2023; Dutt et al., 2023; Azad et al., 2023), outperforming existing state-of-the-art (SOTA) methods in tasks such as zero-shot classification (Yuan et al., 2021), out-of-distribution generalization (Goyal et al., 2023), and histopathology image classification (Roth et al., 2024), visual question answering (Li et al., 2024). BiomedJourney (Gu et al., 2023) was the first work to fine-tune foundation models for counterfactual medical image generation via language descriptions and achieved SOTA results. However, it does not remove confounding artifacts (e.g. medical devices), does not fully maintain faithfulness of the CF image to the original, and is constrained to low resolution images (256 $\times$ 256). This raises a natural question: *Could we leverage a vision-language foundation model pre-trained on diverse natural images and adapt it to generate precise high-resolution medical image counterfactuals?*

In this work, we introduce PRISM (**Pr**ecise counterfactual **I**mage generation using language-guided **S**table Diffusion **M**odel), a strategically fine-tuned vision-language foundation model, that leverages language guidance to generate medical image counterfactuals for novel generative tasks (see Fig. 2). Specifically, PRISM presents the first framework to generate high-resolution (512×512) medical counterfactuals that can selectively remove significant spurious artifacts, such as medical devices. Crucial for explainability in medical settings, it can isolate and modify individual disease attributes (and the spurious correlations) while preserving others. Existing approaches have relied on detailed clinician's notes to train language models (Zhang et al., 2023; Luo et al., 2024). In order to leverage the guidance of a language embedding, our framework adapts binary labels, typical for medical datasets, into text captions.

Through extensive experimentation on the publicly available CheXpert dataset (Irvin et al., 2019), we validate our approach by - (i) generating difference maps between the original and the synthesized CF image to assess the clinical plausibility of the disease, and (ii) using multi-head classifiers to confirm that the counterfactuals are correctly classified. We also show improvement over a baseline classifier-guided GAN-based model, GANterfactual (Mertes et al., 2022). As a key demonstration of PRISM's utility, we show that our counterfactuals enable improving the accuracy of an existing classifier.

2. Methodology

While state-of-the-art vision-language foundation models in computer vision utilize millions of image-text pairs to generate images, their direct application to the medical domain is hindered by two key challenges. First, patient information is stored as tabular data (e.g., numerical labels for age or sex) rather than descriptive text , limiting direct integration into existing vision-language models. Second, medical imaging datasets are significantly smaller than those in computer vision, making it impractical to train a foundation model from scratch. To address these shortcomings and enable CF generation, our methodology consists of three main steps: (i) convert patient tabular data into text format, enabling the generation of rich semantic embeddings via a pre-trained CLIP (Contrastive Language-Image Pre-training) text encoder, Section 2.1; (ii) fine-tune a Stable Diffusion model, to better adapt to a medical imaging dataset, Section 2.2; (iii) at inference, synthesize CF images guided by a text input, Section 2.3.

2.1. Tabular Data to Text Conversion

One of the key requirements of training a Stable Diffusion (specifically v1.5) (Rombach et al., 2022) model is the image-text pair. CheXpert, the medical dataset we use here, only contains binary labels for different diseases and the presence of support devices. To leverage Stable Diffusion, we create a custom template for image-text pairs based on the available tabular data. For example, if the subject's radiograph shows pleural effusion and cardiomegaly, our text caption for the image is `chest x-ray of a patient showing pleural effusion, cardiomegaly.` Additionally, for patients with no findings, we use the template text `Normal chest X-ray with no significant findings`.

2.2. Fine-Tuning the Stable Diffusion Model

The Stable Diffusion v1.5 architecture consists of three components: (i) the Variational Autoencoder (VAE), which encodes images into the latent space and subsequently decodes

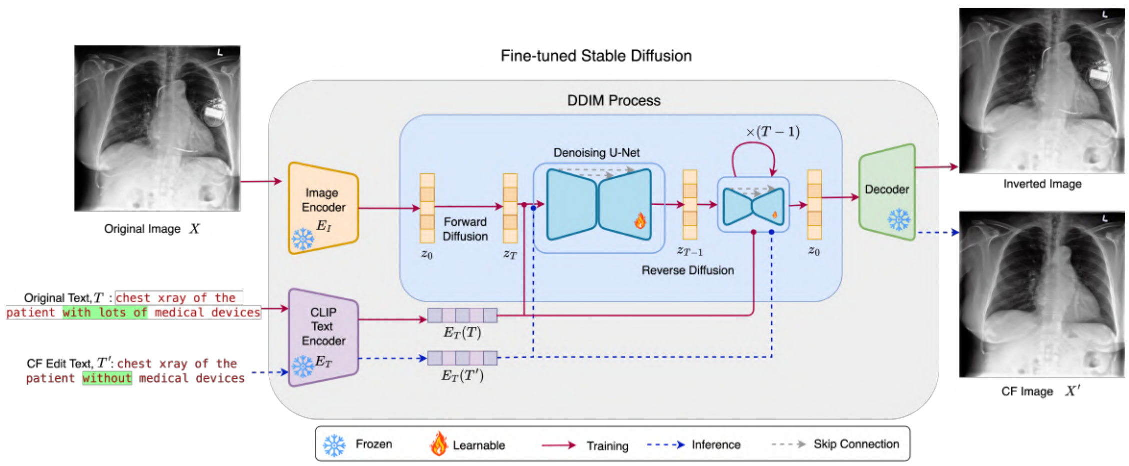


Figure 2: The framework for synthesizing counterfactual (CF) images takes an original input image (X) and its corresponding text description (T), along with an edited text prompt (T') for the CF image. It employs a frozen VAR consisting of an image encoder (E_I) and decoder, as well as a frozen CLIP text encoder (E_T). The core component of the framework is a denoising U-Net, which is the only trainable module during the fine-tuning process. During inference, the encoded text prompt ($E_T(T')$) is used to condition the U-Net, guiding the generation of a high-resolution 512×512 counterfactual image that aligns with the modified text description.

the processed latent representation back into image space; (ii) the U-Net, which operates at the latent level and is trained to predict and remove noise introduced during the forward diffusion process, enabling iterative image refinement; and (iii) the CLIP Encoder, which encodes text descriptions into a vector embedding that is used to condition the U-Net, guiding the image generation process to match the given text description. It should also be noted here that the CLIP model is already pre-trained, providing general semantic knowledge about image-text relationships.

Given our objective of generating identity-preserving counterfactuals, we adopt the Denoising Diffusion Implicit Model (DDIM) (Song et al., 2020). This method provides a deterministic mapping between the noisy and denoised latent. By conditioning the U-Net with a CLIP-encoded text, we gain direct control over the image generation process. For stability during fine-tuning, we update only the U-Net component of the model while keeping the VAE and CLIP encoder fixed. We provide full details for implementing the fine-tuning process in Appendix A.

2.3. Generating Counterfactuals at inference

After fine-tuning the Stable Diffusion model on a medical imaging dataset, generating counterfactuals requires no extra training and is done at inference. To produce a precise counterfactual image (X'), the language embeddings of the CF edited text (T') are used as contexts within the U-Net to guide the denoising process applied to the diffused latent representation of the input image (X). To quantify the alignment of the counterfactual image with the provided edited text alignment, we use an editing score, S_{CLIP} (Eq. 1), which measures the similarity between the generated image and the intended textual modification. Following a similar approach to (Prabhu et al., 2023), we compute the editing score and directional similarity (Gal et al., 2022) to filter out edited samples where $S_{\text{CLIP}} < 0.1$.

$$S_{\text{CLIP}} = \frac{\Delta X \cdot \Delta T}{\|\Delta X\| \|\Delta T\|}, \quad \text{where} \quad \begin{array}{l} \Delta X = E_I\left(X'\right) - E_I\left(X\right), \text{ and} \\ \Delta T = E_T\left(T'\right) - E_T\left(T\right) \end{array} \tag{1}$$

3. Experiments and Results

3.1. Dataset and Implementation Details

We use the publicly available CheXpert dataset (Irvin et al., 2019) that contains over 200,000 chest X-ray images, with binary labels for 14 diseases and the presence of support devices. Table 1 shows a summary of the number of subjects in each split and their distributions.

The DDIM scheduler for image editing uses a `scaled_linear` scheduler with `beta_start` and `beta_end` as `85e-5` and `12e-3` respectively. These parameters define the range of noise variance (β) added at each timestep and linearly increase from `beta_start` to `beta_end`. Text similarity is computed based on `cosine_similarity`. Additionally, for all the synthesized counterfactual images discussed in this manuscript, we use the same hyperparameters (e.g. denoising steps, DDIM scheduler) for all tasks, except the language-based command for each case. Thus, our proposed method does not need extensive hyperparameter tuning. We provide additional implementation details in Appendix A and the code along with model weights for the fine-tuned Stable Diffusion are publicly available at https://huggingface.co/amar-kr/PRISM.

Attribute → **Splits↓**	**Pleural Effusion**	**Cardiomegaly**	**No Finding**	**Support Devices**
Train	62509	21888	12222	78211
Validation	10996	3739	2161	13678
Test	12972	4515	2591	16196

Table 1: Summary of the number of samples for train, validation and test splits.

3.2. Experiments and Metrics: Evaluating the Generated CF Images

To establish baseline comparisons, we implement GANterfactual (Mertes et al., 2022), a classifier-guided CF image generation method. To this end, we trained a binary AlexNet (Krizhevsky et al., 2012) to detect the presence (1) or absence (0) of medical support devices (e.g. pacemakers, wires, tubes) in the original and synthesized CF images produced by GANterfactual. Next, we fine-tune pre-trained Efficient-Net (Tan and Le, 2019), initially trained on Image-Net, for a multi-head classification task: pleural effusion, cardiomegaly, no finding and support devices. This classifier is then used to verify the class of the CF images synthesized by our PRISM framework, ensuring that the generated CFs accurately reflect the desired modifications of the correct target class. It should be noted that the baseline method requires an image size of 224×224.

To quantitatively evaluate the quality of synthesized counterfactual images, we use the following metrics: (i) **Subject Identity Preservation** evaluates how well the subject-identifying characteristics are maintained while only modifying the targeted attribute. Following prior work (Mothilal et al., 2020; Nemirovsky et al., 2020), this is calculated through the L_1 distance between the CF and factual images. (ii) **Counterfactual Prediction Gain (CPG)** (Nemirovsky et al., 2020) measures the absolute difference in a classifier's predictions between factual and CF images. A higher CPG indicates a greater shift across the classifier's decision boundary. We use EfficientNet and AlexNet to measure CPG score for the CF images synthesized by PRISM and baseline methods respectively.

A final set of experiments is devised in order to show that the synthesized CF images focus on the defining features of each disease (such as pleural effusion occurring at the corner of the lungs or cardiomegaly surrounding the position of the heart). The training data for the original EfficientNet classifier is then augmented with these CF images. An increase in classifier accuracy suggests that synthesized counterfactual images enhance generalizability and robustness, enabling the classifier to identify defining disease features independent of potential confounding factors in the dataset. This is particularly important in the context of pleural effusion, which is correlated with the presence of medical devices.

3.3. Results

Classifiers EfficientNet has a classification accuracy of 0.8, 0.87, 0.91 and 0.86 for pleural effusion, cardiomegaly, no finding and support devices, respectively (see first row of Table 3). The accuracy and AUC of the binary AlexNet classifier on a held-out test set are 0.89 and 0.91, respectively. These classifiers are used to measure the CPG scores reported in Table 2.

	L1↓	CPG↑
Baseline (classifier-guided GAN-based CF)	0.091	0.781
PRISM [**Ours**]	**0.031**	**0.845**

Table 2: Quantitative results comparing the scores for the CF generated by a classifier-guided GAN-based method and PRISM, when asked to remove the medical devices. The high CPG indicates that PRISM synthesizes CF images that correctly change the class labels.

Qualitative Evaluations Our qualitative evaluation demonstrates two primary capabilities of our method: (i) the ability to remove and, for completeness, *add* medical devices to the original image, and (ii) the ability to emulate distinct visual pathologies of different diseases.

Chest radiographs contain a variety of medical devices (Gambato et al., 2023) such as chest tubes for draining air, blood, or fluid from the pleural space, surgical clips that are often visible after procedures like axillary node dissection, or pacemakers that regulate heart rhythm, typically seen as a small box near the clavicle (Mathew et al., 2019). These devices vary in shape, size and position in the X-ray image. Our method, PRISM, can remove medical devices, demonstrating robust performance across various device types and positions without any external classifier-based supervision or image-level mask/annotations.

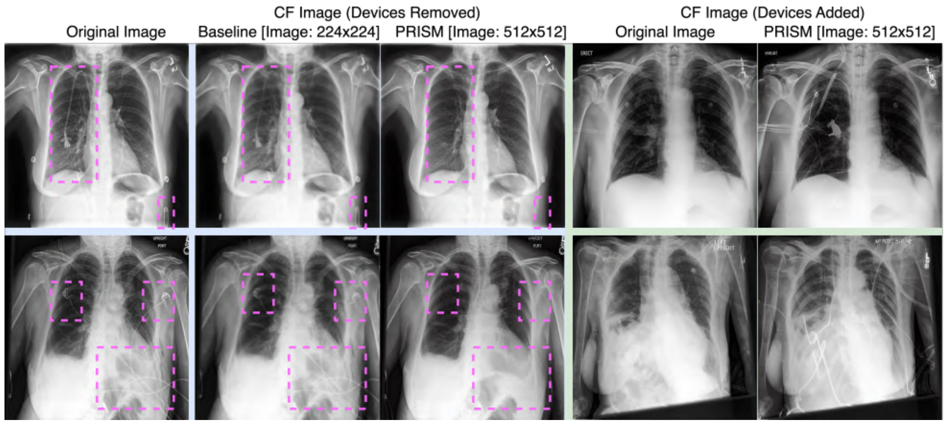


Figure 3: Sample pairs of original and CF images demonstrate the capability of PRISM to remove and add medical devices (e.g. wires, pacemaker) in high resolution. Left: CF images with medical devices removed. Language guidance is T: `chest xray of the patient with lots of medical devices`, T': `chest xray of the patient without medical devices`. Note that the baseline method cannot properly remove medical devices; Right: CF images with added medical devices. Language guidance is T: `chest xray of the patient with no support devices`, T': `chest xray of the patient with lots of support devices`.

In Fig. 3, we show how, by using language guidance, we can remove complex medical devices from the given image without altering the pathology of the disease. We also compare our framework to a baseline method, GANterfactual (Mertes et al., 2022), a classifier-guided CF generator. This method relies on the gradient from a pre-trained classifier for guidance and fails to remove devices from the image. Next, we evaluate our method's ability to effectively *differentiate* between diseases during CF image generation. Specifically, Fig. 4 demonstrates PRISM's performance in generating CFs for two diseases: Pleural Effusion and Cardiomegaly. The difference maps in Fig. 4 demonstrate that our approach can identify and remove the target disease while preserving the anatomical features of the subject, as well as the devices and other artifacts outside the regions of the expected changes.

Quantitative Evaluations To quantitatively evaluate our approach, we compare our method with GANterfactual, a classifier-guided GAN-based approach for generating counterfactuals. Table 2 shows the results for the task of removing medical devices. The counterfactual images generated by GANterfactual show similar L_1 scores to those produced by our method, indicating that the synthesized images in both cases remain close to their factual counterparts. However, counterfactuals generated by PRISM achieve higher CPG scores, suggesting that these images are more effectively converted to the opposite class (see Appendix E for additional results).

Table 3 shows the results of re-training the classifier with counterfactuals for the classes Pleural Effusion, Cardiomegaly, No Finding, and Support Devices. As shown, augmented training leads to improved classifier performance, demonstrating that incorporating counterfactuals synthesized by PRISM enhances the model's robustness.

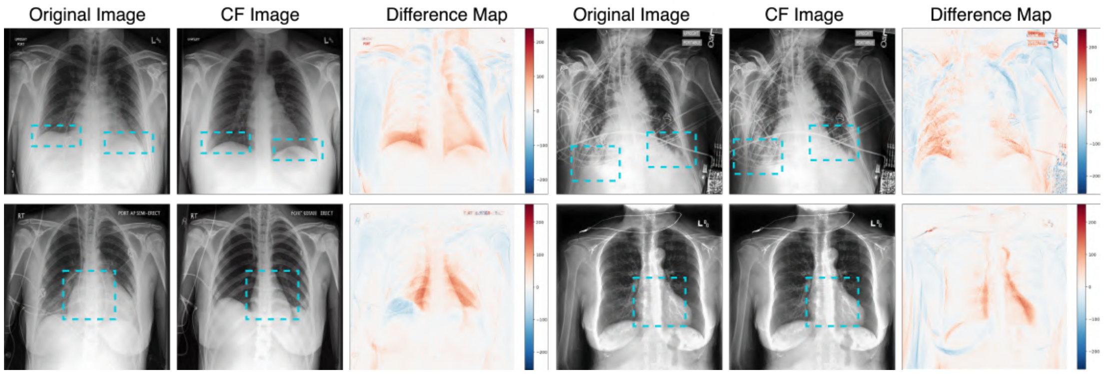


Figure 4: Sample pairs of original and edited images showcasing accurate, precise and high-resolution generated CFs for disease pathology explainability. The original (T) and edited text prompts (T') are - Row 1: T - `chest x-ray of the patient with severe pleural effusion`, T' - `chest x-ray of the patient with no finding`; Row 2: T - `chest x-ray of the patient with severe cardiomegaly`, T' - `chest x-ray of the patient with no finding`.

	Pleural Effusion	**Cardiomegaly**	**No Finding**	**Support Devices**
Efficient-Net	0.80	0.87	0.91	0.86
Efficient-Net augmented with PRISM CFs	**0.88**	**0.90**	**0.92**	**0.88**

Table 3: Augmented classifier accuracies: CF generated by our method are used to augment the training dataset. The accuracies are reported on the same held-out test set

4. Conclusion

Developing a generative model in the medical domain to produce high-quality counterfactuals requires a balance between image fidelity and controllability. In this work, we present PRISM, a fine-tuned vision-language foundation model for counterfactual medical image generation that addresses these challenges. PRISM is the first framework to use language guidance to synthesize high-resolution (512×512) medical images consistent with their factual counterparts. We demonstrate our results through extensive experiments on the CheXpert dataset. Our approach generates precise and accurate CFs representing disease states and is able to cleanly remove medical devices. We make our code and fine-tuned model weights publicly available to the medical imaging community for further development. Future work will investigate the use of synthesized counterfactual images to build robust classifiers for out-of-distribution generalization, and to assess the disentanglement capacity of language-guided foundation models.

Acknowledgements

The authors are grateful for funding provided by the Natural Sciences and Engineering Research Council of Canada, the Canadian Institute for Advanced Research (CIFAR) Artificial Intelligence Chairs program, Mila - Quebec AI Institute, Google Research, Calcul Quebec, Fonds de recherche du Québec (FRQNT), and the Digital Research Alliance of Canada.

References

Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*, 2023.

Raman Dutt, Linus Ericsson, Pedro Sanchez, Sotirios A Tsaftaris, and Timothy Hospedales. Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. *arXiv preprint arXiv:2305.08252*, 2023.

Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

Marco Gambato, Nicola Scotti, Giacomo Borsari, Jacopo Zambon Bertoja, Joseph-Domenico Gabrieli, Alessandro De Cassai, Giacomo Cester, Paolo Navalesi, Emilio Quaia, and Francesco Causin. Chest x-ray interpretation: detecting devices and device-related complications. *Diagnostics*, 13(4):599, 2023.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19338–19347, 2023.

Yu Gu, Jianwei Yang, Naoto Usuyama, Chunyuan Li, Sheng Zhang, Matthew P Lungren, Jianfeng Gao, and Hoifung Poon. Biomedjourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys. *arXiv preprint arXiv:2310.10765*, 2023.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Amar Kumar, Nima Fathi, Raghav Mehta, Brennan Nichyporuk, Jean-Pierre R Falet, Sotirios Tsaftaris, and Tal Arbel. Debiasing counterfactuals in the presence of spurious correlations. In *Workshop on Clinical Image-Based Procedures*, pages 276–286. Springer, 2023.

Ryo Kurokawa, Yuji Ohizumi, Jun Kanzawa, Mariko Kurokawa, Yuki Sonoda, Yuta Nakamura, Takao Kiguchi, Wataru Gonoi, and Osamu Abe. Diagnostic performances of claude 3 opus and claude 3.5 sonnet from patient history and key images in radiology's "diagnosis please" cases. *Japanese Journal of Radiology*, pages 1–4, 2024.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

Haozhe Luo, Ziyu Zhou, Corentin Royer, Anjany Sekuboyina, and Bjoern Menze. Devide: Faceted medical knowledge for improved medical vision-language pre-training. *arXiv preprint arXiv:2404.03618*, 2024.

Rishi P Mathew, Timothy Alexander, Vimal Patel, and Gavin Low. Chest radiographs of cardiac devices (part 1): Lines, tubes, non-cardiac medical devices and materials. *SA Journal of Radiology*, 23(1):1–9, 2019.

Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. *Frontiers in artificial intelligence*, 5:825565, 2022.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. Countergan: Generating realistic counterfactuals with residual generative adversarial nets. *arXiv preprint arXiv:2009.05199*, 2020.

Viraj Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. Lance: Stress-testing visual models by generating language-guided counterfactual images. *Advances in Neural Information Processing Systems*, 36:25165–25184, 2023.

Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High fidelity image counterfactuals with probabilistic causal models. *arXiv preprint arXiv:2306.15764*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

Benedikt Roth, Valentin Koch, Sophia J Wagner, Julia A Schnabel, Carsten Marr, and Tingying Peng. Low-resource finetuning of foundation models beats state-of-the-art in histopathology. *arXiv preprint arXiv:2401.04720*, 2024.

Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. *arXiv preprint arXiv:1911.00483*, 2019.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Dequan Wang, Xiaosong Wang, Lilong Wang, Mengzhang Li, Qian Da, Xiaoqiang Liu, Xiangyu Gao, Jun Shen, Junjun He, Tian Shen, et al. A real-world dataset and benchmark for foundation model adaptation in medical image classification. *Scientific Data*, 10(1): 574, 2023.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023.

Appendix A. Additional Implementation Details and Code Listings

We provide additional steps for our implementation.

Code Listing 1 Generating text for the images in CheXpert dataset

```
conditions = [
            'No Finding', 'Enlarged Cardiomediastinum', 'Cardiomegaly',
            'Lung Opacity', 'Lung Lesion', 'Edema', 'Consolidation',
            'Pneumonia', 'Atelectasis', 'Pneumothorax', 'Pleural Effusion',
            'Pleural Other', 'Fracture', 'Support Devices'
        ]

captions = []
for image in images:
    findings = []
    for condition in conditions:
        if image[condition] == 1:
            findings.append(condition)

    caption = "Chest X-ray showing " + ", ".join(findings) if findings else "Normal chest X-ray with no significant findings"
    captions.append(caption)
```

Algorithm 1: Fine-tuning Stable Diffusion on CheXpert

Pre-trained Stable Diffusion model components: unet, vae, textEncoder, tokenizer, noiseScheduler
CheXpert dataset: dataloader
Optimizer: optimizer **for each** *batch in dataloader* **do**

```
    latents = vae.encode(batch["image"])            ▷ encode images into latent space
    noise = sampleRandomNoise()                     ▷ add random noise to latents
    timesteps = sampleRandomTimesteps()
    noisyLatents = noiseScheduler.addNoise(latents, noise, timesteps)
    encoderHiddenStates = textEncoder(batch["inputIds"])    ▷ encode text captions
    noisePred = unet(noisyLatents, timesteps, encoderHiddenStates)    ▷ predict noise residual with U-net
    loss = mseLoss(noisePred, noise)                ▷ compute pixel wise loss
    backward(loss)                                  ▷ backpropagate
    optimizer.step()                                ▷ update weights
    optimizer.zeroGrad()
end
```

Appendix B. Image Inversion

Image inversion aims to reconstruct the original, unedited image. This verification step demonstrates the model's capacity to create faithful reproductions, confirming that the addition and removal of artifacts during editing reflects deliberate modifications rather than limitations in the model's representational capacity. Figure 5 shows the original and inverted images, with many details preserved during generation. Notably, the model struggles with the small text found within the images, which we further discuss in Appendix H. When the original and inverted images are passed through the state-of-the-art classifier, the changes in multi-class logit values are close to zero. This confirms that inversion process maintains relevant details needed for accurate image classification.

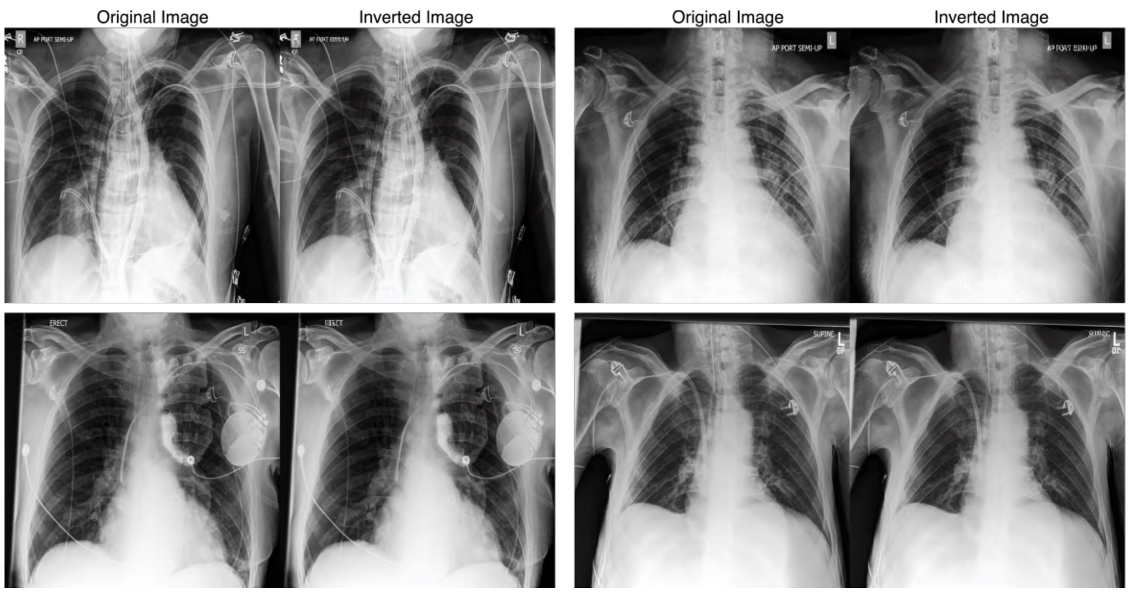


Figure 5: The inversion quality of the proposed generative model. (a) Original image, (b) Inverted image.

Appendix C. Classifier Performance on the Synthesized CF Images

We use the classifier, Efficient-Net, in Table 3 to validate the changes made when synthesizing CF images. Classifications across all heads of the classifier, along with the corresponding original and counterfactual images, are presented in Fig. 6. As shown, the intervened-upon attribute is successfully pushed across the decision boundary, while all other attributes retain their original classification. Notably, even when multiple attributes are present in the original image, only the targeted attribute undergoes a shift across the decision boundary, which is verified by the resulting counterfactual image. This demonstrates our model's ability to precisely distinguish and modify each attribute as intended.

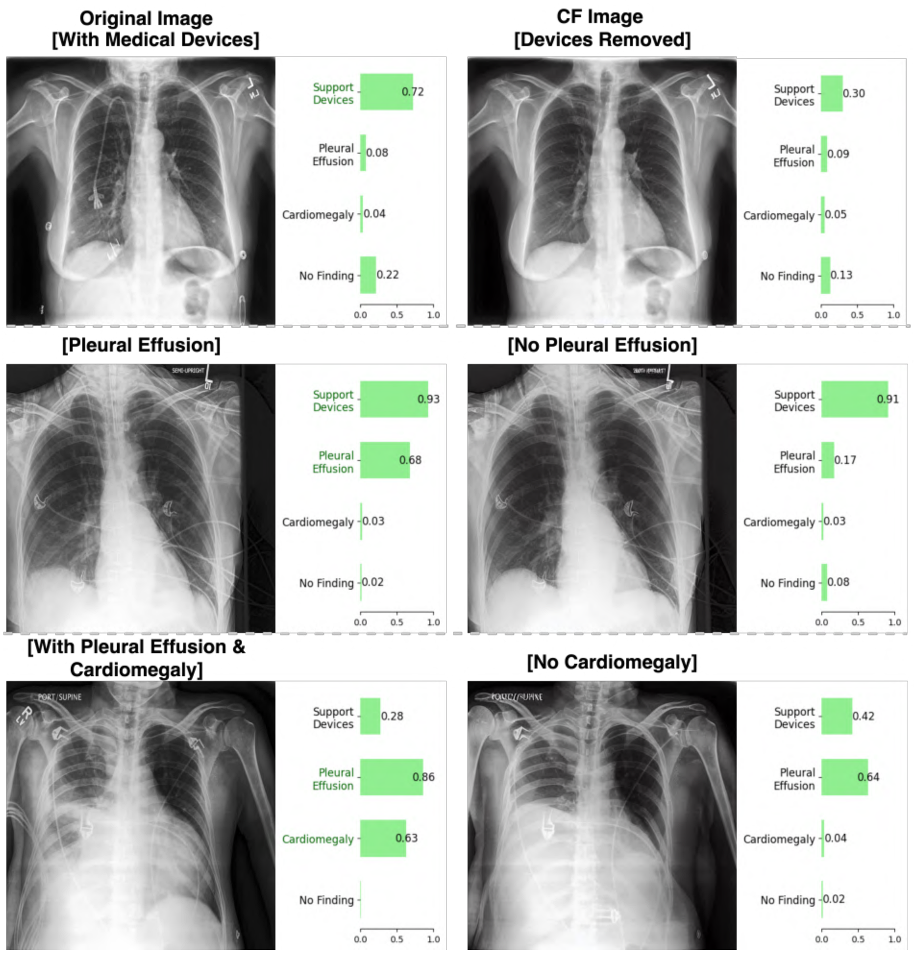


Figure 6: Classifier's performance on the original (left) and CF images (right). Note that the classifier is robust to changes made in the CF image. Text indicated in green shows the ground truth for the given image.

Appendix D. Performance of the robust classifier

To evaluate the utility of counterfactuals synthesized from PRISM for downstream tasks, we augment our dataset and retrain the original EfficientNet multi-head classifier (see Table 3). Notably, the original classifier, trained without augmented counterfactuals, continues to detect support devices even after their removal—likely due to the correlation between pleural effusion and medical devices in the dataset. By incorporating CF augmentation, the classifier learns the true features associated with the medical device, reducing its reliance on correlations with the disease, see Fig. 6.

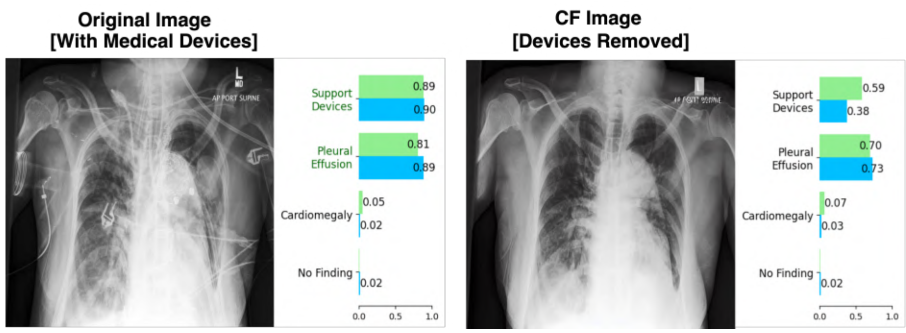


Figure 7: Original classifier and robust classifier performance on the original (left) and CF images (right) for the text prompt to remove medical devices. Note that the robust classifier is able to correctly identify the absence of medical devices in the CF image while the original classifier fails. Text indicated in green in the image shows the ground truth for the given image.

Appendix E. Validation of Image Modification Using State-of-the-Art VQA Models

To verify that the image modifications reflect expected anatomical changes, we use state-of-the-art VQA models to classify the images and analyze the disease-related features. We applied two state-of-the-art Vision Question Answering (VQA) models - Claude 3.5 Sonnet and LlaVA-Med. These models were chosen as they achieved high performance in disease diagnosis (Kurokawa et al., 2024). During inference, we input an image X and its corresponding edited image X' after modifying the text T: "an x-ray with severe pleural effusion" to T': "an x-ray with no pleural effusion". The resulting correspondence with the VQA models is shown in Table 4. The response from the VQA models indicate not only a successful shift in classification but also correct reasoning, aligning with the desired modifications.

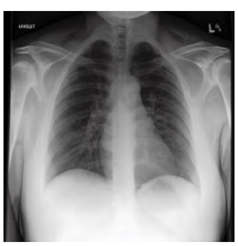
 Original Image	Question	Does this image have pleural effusion?
	Claude 3.5 Sonnet	`Yes, this chest X-ray shows evidence of pleural effusion. There appears to be blunting of the right costophrenic angle and increased opacity in the lower right hemithorax, which are characteristic radiographic signs of pleural effusion. The effusion appears to be unilateral (on the right side) and moderate in size based on this frontal (PA/AP) view.`
	LLaVA-Med	`Yes, the image shows bilateral pleural effusion, which means there is an accumulation of fluid in the pleural space on both sides of the lungs.`
 CF Image **[Ours]**	Question	Does this image have pleural effusion?
	Claude 3.5 Sonnet	`This chest X-ray appears to be normal without evidence of pleural effusion. The costophrenic angles (the corners where the ribs meet the diaphragm) are sharp and clear bilaterally, which is a key finding that suggests absence of pleural effusion. The lung fields appear symmetrical with normal transparency, and there is no blunting of the costophrenic angles or visible fluid levels that would indicate pleural effusion.`
	LLaVA-Med	`No, the chest X-ray image does not show any pleural effusion.`

Table 4: Validation of the original and edited image using state-of-the art VQA models.

Appendix F. Role of Confounders in CF Generation

An interesting correlation we observed when generating counterfactuals to explain the disease cardiomegaly (an enlarged heart) was the removal of the pacemaker. Note that a pacemaker can be a treatment for cardiomegaly. Indeed, when removing cardiomegaly, the CF showed that the pacemaker was also removed (Fig. 8), and other devices, such as EKG (electrocardiogram) stickers and tubes, are maintained (see final row of Fig. 6 and the bottom left example in Fig. 8). In the context where the original image contains both pleural effusion and cardiomegaly, the CF image with the task of removing cardiomegaly also removes the pacemaker. However, when generating a CF image to remove pleural effusion, no such effect occurs (Fig. 9). This suggests that the model associates the presence of a pacemaker specifically with cardiomegaly but not with pleural effusion.

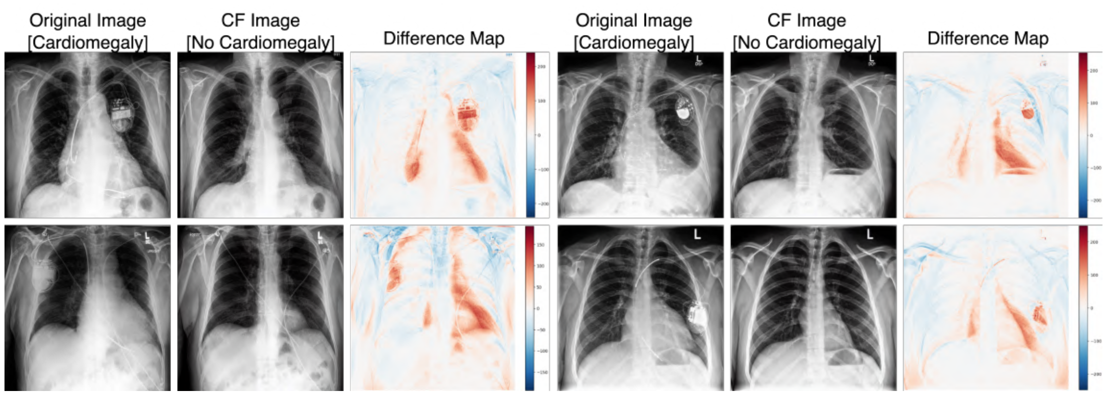


Figure 8: Samples where the removal of cardiomegaly, from the original image containing 'pacemaker'. Please note that our method removes the disease, cardiomegaly, and pacemaker.

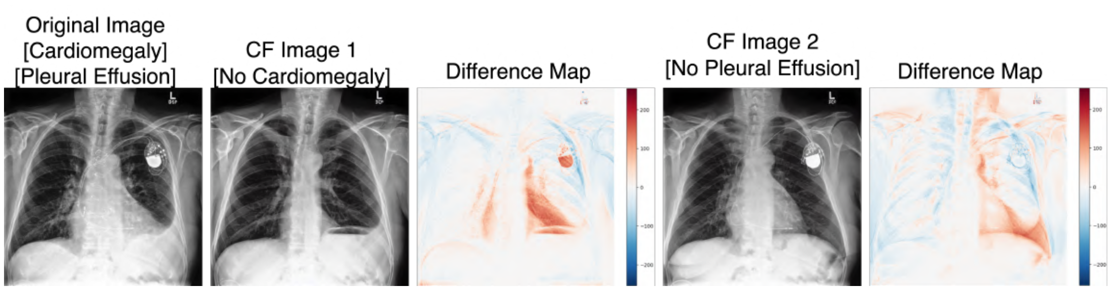


Figure 9: Comparing the change from original image with both cardiomegaly and pleural effusion to two different CFs. Note that when synthesizing the CF image with no pleural effusion the pacemaker is retained.

Appendix G. Validation: CF generation in Challenging Cases

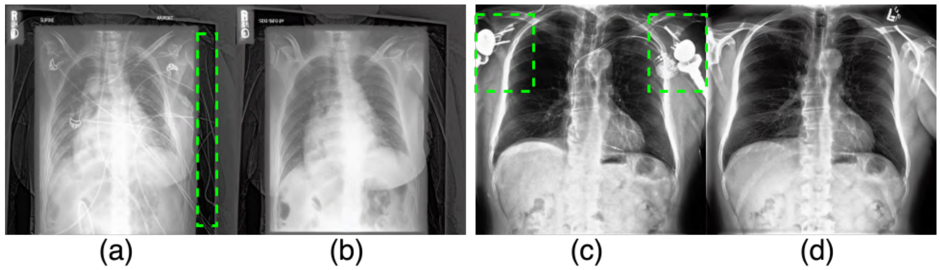


Figure 10: Samples showing challenging cases. (a,c): Original images with devices; (b,d): CF images without medical devices.

To demonstrate the robustness of PRISM, we examine cases that are particularly challenging to edit due to the placement of devices outside the field of view or devices in regions with bone structures. As shown in Fig. 10 (a), the device cables are located in low-light conditions near the arm. Fig. 10 (b) shows the edited image where the cables are accurately removed by our model without impacting the humerus. In Fig. 10 (c), the artificial shoulder joint creates high-intensity pixels. The corresponding edited image in Fig. 10 (d) shows the successful removal of the joints, replacing the affected pixels with feasible anatomical structures for the region. The structures in other areas are not altered. These examples demonstrate the robustness of the proposed method in challenging settings.

Appendix H. Limitations of PRISM

Although our method is capable of synthesizing high-resolution images (512×512), it faces difficulties in reproducing the small text written in the corner of radiographs (Fig 11) in both the inverted and CF images.

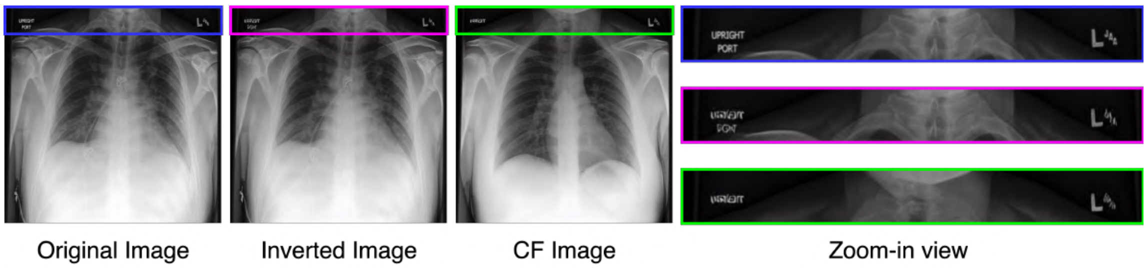


Figure 11: Text at the corner of the image remains unresolved in the inverted and edited images.

These are challenging settings in which the model struggles to maintain consistent edits. This variation is partly dependent on the complexity of the image. In cases where the original image is distorted, the CF image deviates from expected changes (see Fig. 12).

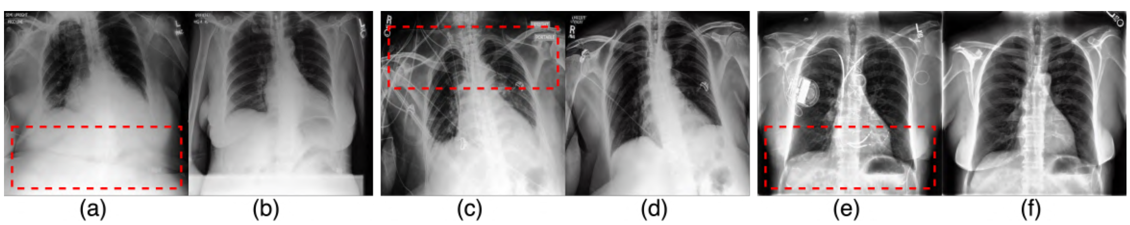


Figure 12: Examples of original (a, c, e) and CF image (b, d, f) pairs. The command was to remove the support device, and the edits were inconsistent with the expected outcome. Red boxes highlight areas where the changes are not as intended. (a-b): The radiograph shows a problem with the original image (at the bottom). The edited image incorrectly modifies this region instead of retaining it. (c-d): The red-boxed region contains multiple tubes. While removing the tubes, the model recreates the missing anatomical area improperly. (e-f): When removing the medical devices, the subject is depicted more strongly as female