Counterfactual Disease Removal and Generation in Chest X-Rays Using Diffusion Models

Ahmet Berke Gökmen
ETH Zürich Computer Vision Lab
Zürich, Switzerland
agoekmen@ee.ethz.ch

Ender Konukoğlu
ETH Zürich Computer Vision Lab
Zürich, Switzerland
ender.konukoglu@vision.ee.ethz.ch

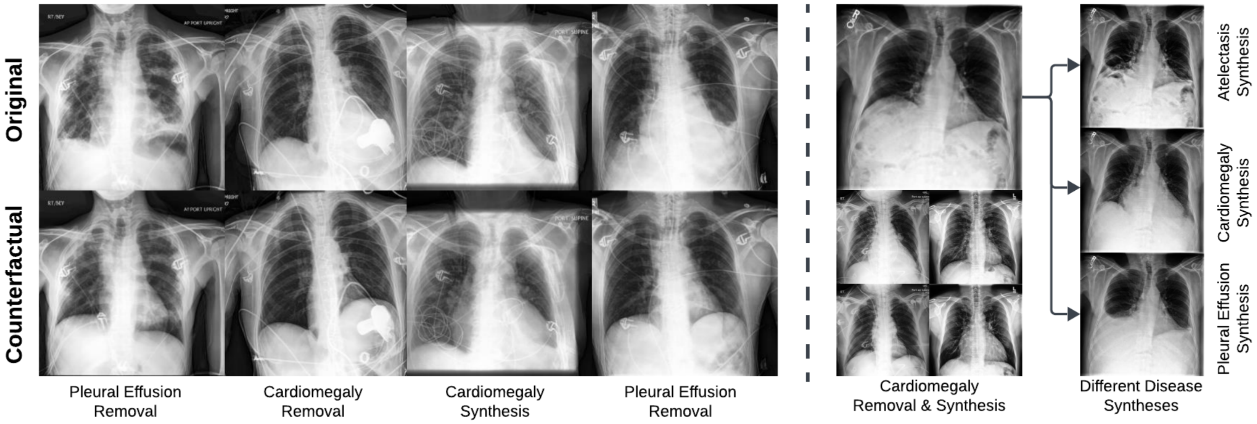


Figure 1. Our Method is Able to Generate Clinically Accurate Counterfactual Images While Preserving Patient Identity.

Abstract

Recent advancements in medical imaging have leveraged deep learning, particularly latent manipulation in autoencoders and StyleGAN-based models, to edit chest X-rays for disease modification. However, these methods often rely on fixed latent directions, limiting their flexibility and precision. Although diffusion models have recently been explored for medical imaging, they frequently alter critical patient-specific details, compromising identity preservation. We introduce a novel diffusion-based approach that enables disease generation and removal in chest X-rays while preserving patient identity, eliminating the need for specific disease masks. By fine-tuning attention layers, our model retains defining anatomical features, ensuring realistic yet flexible edits. A pretrained classifier guides the diffusion process, aligning modifications with clinical relevance. We evaluate our approach using Fréchet Inception Distance (FID) for image quality and Contrastive Language-Image Pretraining (CLIP) accuracy.

1. Introduction

Over the past decade, deep learning has revolutionized medical imaging by identifying complex visual patterns [7]. Generative models like generative adversarial networks (GANs) and autoencoders have shown significant potential in high-fidelity modification and generation of medical images [9, 21, 27]. These advancements enable counterfactual analysis—altering medical images to explore hypothetical scenarios—crucial for diagnostic insights, treatment planning, and synthetic data generation.

Chest X-ray manipulation is of special interest due to the modality's non-invasive nature and rapid diagnostic capabilities. However, interpretation often relies on subjective radiologist expertise, leading to interest in generative models that simulate hypothetical disease scenarios to enhance diagnostic processes. Prior work has mainly used latent space manipulation in GANs and autoencoders for these modifications [1, 5, 15, 17, 28]. While promising, these methods are limited in flexibility and depend on fixed latent directions for disease modification.

Diffusion models provide a distinct approach to genera-

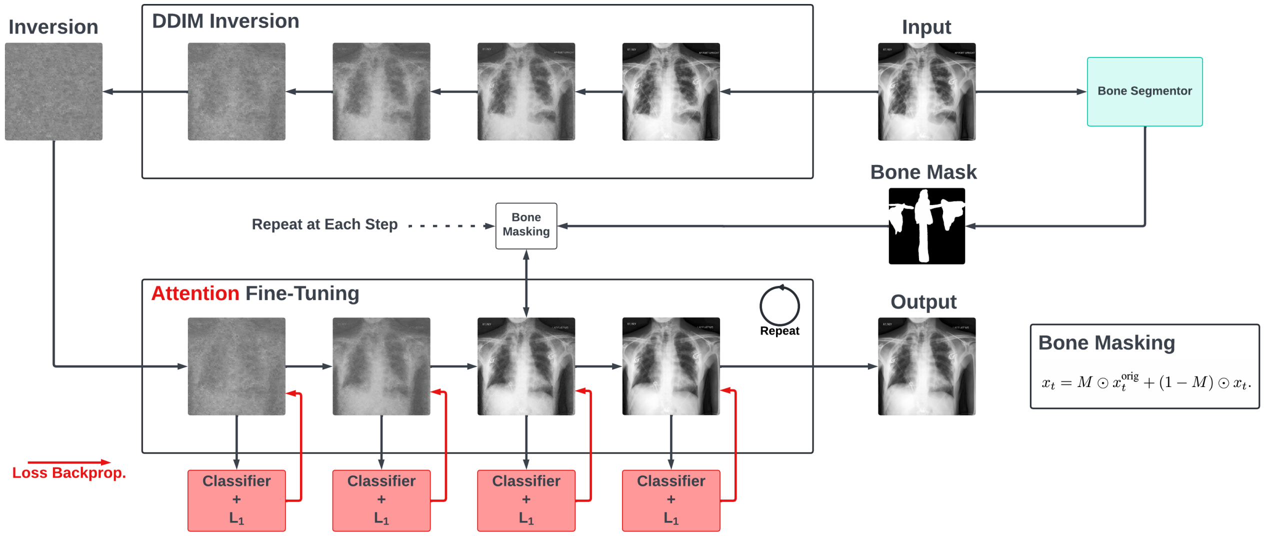


Figure 2. Overview of Our Method. We first apply DDIM inversion to get the latent. Then, we iteratively denoise the image while applying losses and masking the latent with the bone mask at each step.

tive image modeling, initially developed as a method for denoising data through a Markov process [29]. Unlike GANs, which rely on adversarial training, diffusion models add noise incrementally in a forward process and remove it iteratively, producing realistic images without GAN-related instabilities. While diffusion models have shown strong results in natural image generation, they remain underexplored in medical imaging, particularly for counterfactual disease manipulation in chest X-rays. One example, RadEdit [22], applies diffusion to stress-test biomedical vision models, identifying failure cases. RadEdit uses a generative text-to-image approach to modify chest X-rays by adding or removing abnormalities in masked regions based on text descriptions. These edits help evaluate models for disease classification or anatomy segmentation under varied conditions. However, RadEdit depends on predefined disease masks for editing, potentially limiting its flexibility in scenarios requiring unsupervised or diverse manipulation.

In this work, we introduce diffusion models for generating and removing diseases in chest X-ray images, addressing limitations of latent space and mask-based methods with a flexible, clinically relevant approach. Guided by a pretrained classifier, our model ensures that modifications are visually realistic and clinically aligned.

To validate our approach, we conducted visual experiments and assessed disease modification accuracy using the CheXzero model [31], designed specifically for medical imaging.

Our results demonstrate that diffusion models, combined with our method, outperform traditional latent and mask-based techniques in visual fidelity and disease specificity. While further evaluation is needed for scalability across datasets, this work represents a significant advancement in applying diffusion models to medical imaging and offers new possibilities for automated disease manipulation.

2. Related Works

Diffusion Models. Diffusion models [19] generate data by reversing a noise-adding process, making them highly effective for image synthesis. Key advancements include DDPM [11], which optimizes for clean data generation through score matching, and DDIM [30], which introduces a non-Markovian process for efficient, deterministic sampling. Latent Diffusion Models (LDM) [26] further improve scalability by operating in latent space, preserving data integrity with reduced computational demands.

Chest X-Ray Disease Classifiers. The CheXpert dataset [13] has enabled significant advances in chest X-ray disease classification, offering a large, labeled dataset for training robust machine learning models. Leveraging this resource, CNNs like DenseNet [12] have improved the accuracy and generalizability of automated classifiers, supporting the detection of conditions such as pneumonia and cardiomegaly, and enhancing clinical decision-making.

Medical CLIP Models. Recent advancements in medical imaging have introduced models like CheXZero [31], based on the CLIP framework [23]. CheXZero uses contrastive learning to align chest X-ray images with text descriptions, enabling zero-shot learning for enhanced interpretability and automated disease classification. Other models, such as CXR-CLIP [37] and BioMedCLIP [38], further expand this approach. CXR-CLIP improves diagnostic accuracy by enhancing contextual understanding specific to chest X-rays, while BioMedCLIP addresses a broader range of biomedical data, linking images with medical text. Collectively, these models demonstrate the potential of multimodal learning to integrate visual and textual information, advancing radiological disease diagnosis and classification.

Medical Image Generation. In the realm of medical image generation, the Cheff model [34] stands out for its innovative application of latent diffusion models (LDMs) in synthesizing chest X-ray images. Cheff leverages LDMs to perform diffusion within a compressed latent space, allowing for high-resolution image generation with reduced computational demands. During training, Cheff iteratively learns to refine noisy latent representations into realistic chest X-rays, guided by prior patterns in extensive medical imaging data.

Cheff was trained on a collection of large-scale chest X-ray datasets, including ChestX-ray8 [33], MIMIC-CXR [14], PadChest [3], BRAX [24], VinDr-CXR [18], and CheXpert [13]. These datasets collectively provide a diverse array of chest X-ray images with a wide range of labeled conditions, allowing the model to generalize well across various thoracic abnormalities. Through the use of LDMs, Cheff synthesizes highly realistic chest X-rays that capture both anatomical and pathological features, making it suitable for applications in clinical training and evaluation. Thus, we based our work on top of Cheff model.

For counterfactual generation, several methods have emerged to enhance interpretability and diagnostic support. GIFsplanation [5] uses an autoencoder-based approach to create visual explanations of medical images by utilizing a latent-shift approach, facilitating interpretability by simulating hypothetical scenarios. Style-Chexplain [1] combines a GAN with a classifier to generate contextually rich explanations for chest X-rays using specific latent directions to edit images.

DiffExplainer [8] employs a counterfactual generation approach to clarify decision-making in AI models by creating alternative images that provoke different classifications. This method is particularly useful for visualizing decision-influencing features in black-box models. Similarly, CoFE [15] introduces contrastive learning with counterfactual explanations to address bias in radiology report generation by comparing factual and counterfactual samples, thereby improving the reliability of automated reporting.

Atad et al. [2] propose a diffusion autoencoder framework for generating counterfactual explanations, demonstrating the utility of diffusion processes in crafting realistic counterfactuals without requiring extensive labeled data. Likewise, Xia et al. [35] address attribute amplification in counterfactual generation, aiming to refine the granularity and relevance of generated attributes in medical imaging contexts.

Score-based methods have also gained traction, such as the model by Wang et al. [32], which uses score-based diffusion for generating counterfactuals that support lesion localization and classification in medical images. Singla et al. [28] developed a counterfactual explainer for chest X-ray diagnosis, leveraging a GAN-based approach to generate smooth, progressive changes in query images that align with diagnostic shifts, making classifier decisions more transparent and clinically interpretable.

While these approaches are promising, the application of diffusion models in counterfactual generation remains limited. Further research is needed to harness the strengths of diffusion models in creating clinically meaningful counterfactuals that enhance diagnostic accuracy and decision transparency in medical imaging.

3. Method

3.1. Preliminary: Diffusion Models

Diffusion models [19, 20, 25, 29, 30] generate data by learning to reverse a noise-adding process. They consist of two stages: a forward process that incrementally adds Gaussian noise, and a reverse process that learns to denoise and reconstruct realistic samples.

In the forward process, Gaussian noise is added to data through a Markov chain:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where β_t defines the noise schedule. As t progresses, the data approaches a standard Gaussian distribution: $q(\mathbf{x}_T) \approx \mathcal{N}(0, \mathbf{I})$.

The reverse process, parameterized by θ, removes noise in steps to recover the data:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}). \tag{2}$$

The model is trained using a simplified objective:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, \epsilon\sim\mathcal{N}(0,\mathbf{I}), t}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2\right], \tag{3}$$

where ϵ_θ predicts the noise at each step.

Generation starts from Gaussian noise, with each sample refined iteratively:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t^2}}\epsilon_\theta(\mathbf{x}_t, t)) + \sigma_t\mathbf{z}, \tag{4}$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ and the process begins with $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$.

3.2. Preliminary: Chest X-Ray Classification

DenseNet121 [12], used by JF Healthcare [36] for the CheXpert challenge [13], is a convolutional neural network (CNN) well-suited for chest X-ray classification due to its efficient parameter use and deep feature representation. DenseNet121's architecture connects each layer to all previous layers, enhancing gradient flow and reducing the vanishing gradient problem. This structure supports robust learning for medical imaging tasks, with final predictions made through global average pooling and a fully connected layer with softmax for classification.

3.3. DDIM Inversion

DDIM inversion [11] enables controlled noise addition to an image's latent representation, gradually transforming it into pure noise. After encoding an image x into latent space using an autoencoder, we iteratively add noise at each timestep, eventually obtaining a fully noisy latent state.

For an input x at timestep t, the model predicts the noise component e_t:

$$e_t = \epsilon_\theta(x, t), \tag{5}$$

which is used to estimate the original denoised image $\hat{x}_0$:

$$\hat{x}_0 = \frac{x - \sqrt{1-\alpha_t} \cdot e_t}{\sqrt{\alpha_t}}. \tag{6}$$

To progress to the next timestep, we add a directional noise term dir_{x_t}:

$$\text{dir}_{x_t} = \sqrt{1-\alpha_{\text{next}}} \cdot e_t, \tag{7}$$

updating x_t as follows:

$$x_t = \sqrt{\alpha_{\text{next}}} \cdot \hat{x}_0 + \text{dir}_{x_t}. \tag{8}$$

By repeating this process across timesteps, we gradually transform the encoded latent representation into pure noise. This step prepares the noisy latent state for subsequent conditional generation, ensuring that the model can refine the latent structure based on specific target characteristics.

3.4. Classification Loss

To predict $\hat{x}_0$, we utilize the DDIM [11] process, which calculates an intermediate x_t and denoises it to produce $\hat{x}_0$. Given input x and timestep t, we initialize cumulative alpha values, α_t and α_{prev}, and calculate the noise prediction e_t as:

$$e_t = \epsilon_\theta(x, t) \tag{9}$$

Using e_t, $\hat{x}_0$ is predicted by:

$$\hat{x}_0 = \frac{x - \sqrt{1-\alpha_t} \cdot e_t}{\sqrt{\alpha_t}}, \tag{10}$$

where α_t controls the denoising at the current timestep. The prediction $\hat{x}_0$ is then passed to a classifier to adjust the confidence in a target class.

To achieve this, we apply a targeted classification loss. For **disease removal**, we set $y_{\text{target}} = 0$ and define the binary cross-entropy loss:

$$\mathcal{L}_{\text{remove}} = -\log(1 - p(y_{\text{target}}|\hat{x}_0)). \tag{11}$$

For **disease synthesis**, we set $y_{\text{target}} = 1$ with:

$$\mathcal{L}_{\text{synthesize}} = -\log p(y_{\text{target}}|\hat{x}_0). \tag{12}$$

This targeted loss formulation allows fine control over the presence or absence of the target class in $\hat{x}_0$, while preserving other class predictions.

3.5. Using Bone Masks to Preserve Identity

To maintain identity during image manipulation, we use bone masks that isolate key anatomical structures, such as the clavicles, scapulae, aorta, and spine, while excluding the lungs. These masks, generated by a bone segmentation model [4, 6], are initially extracted from the image domain and then downsampled to match the latent dimensions for application at the latent level.

At each denoising step, we create a progressively altered version of the original encoded image x^{orig} at timestep t:

$$x_t^{\text{orig}} = \sqrt{\alpha_t} \cdot x_{\text{orig}} + \sqrt{1-\alpha_t} \cdot \epsilon, \tag{13}$$

where α_t is the cumulative scaling factor for the current timestep, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ represents a perturbation component. This modified x_t^{orig} is then combined with the current image x_t using the downsampled bone mask M:

$$x_t = M \odot x_t^{\text{orig}} + (1 - M) \odot x_t. \tag{14}$$

This process is repeated at each timestep, confining changes to non-bone regions and preserving identity-defining bone structures throughout the refinement process in latent space.

3.6. Fine-Tuning of Attention Layers

To improve consistency and visual quality, we fine-tune the attention layers of the Cheff diffusion model, which is based on a UNet architecture with self-attention and residual blocks. By restricting updates to only the attention layers, we allow the model to focus on refining feature alignment and preserving details in each sample. This targeted fine-tuning applies precise adjustments where needed, enhancing visual coherence in the generated output without altering other parts of the network.

3.7. L_1 Regularization for Detail Preservation

To preserve details between the original and edited images, we apply an L_1 regularization term at each denoising step. After converting images back to the image domain, the loss minimizes the pixel-wise absolute difference between the original image x_{original} and the edited image x_{edited}:

$$\mathcal{L}_{L1} = \|x_{\text{edited}} - x_{\text{original}}\|_1, \tag{15}$$

This term, applied iteratively, ensures that modifications remain subtle, preserving structural and textural details throughout the refinement process.

4. Experimental Setup

4.1. Training

We have used CheXpert [13] dataset in line with the instructions from [36] to train the classification model. The dataset itself is split into train and validation sets. Therefore, we complete the training on the train set and use the validation set for testing. The counterfactual image generation is done on the validation set.

The fine-tuning operations on the Cheff model [34] are done using a single RTX 4090 GPU for 2 epochs where one epoch refers to a full denoising pass. The learning rate is $2e^{-5}$. For the optimizer AdamW [16] is utilized.

4.2. Baselines

We present qualitative and quantitative comparisons between our model and two competing methods: RadEdit [22] and GifExplanation [5].

RadEdit [22] is a diffusion-based generative method that uses user-defined editing prompts along with specific edit and non-edit masks to control which regions are modified. For fairness, we provided RadEdit with bone masks as non-edit masks to restrict modifications to non-bone areas just like in our model. In the original RadEdit model, only a small portion of the X-ray is open to edits by the mask, limiting changes outside this region; however, this approach can be restrictive for unsupervised editing tasks.

GifExplanation [5] is a latent shift-based method using an autoencoder combined with a classifier, where transformations occur within the latent space. We applied the autoencoder from [34] to generate outputs in a consistent framework, ensuring fair and standardized comparisons.

Additionally, since RadEdit is a prompt-based method rather than a class-based one, we used "No apparent {disease}" for removal prompts and "Present {disease}" for synthesis prompts to align with its editing paradigm.

4.3. Evaluation

CLIP Accuracy Score. Following [31], we measure probabilities for prompts "{Pathology}" and "No {Pathology}" to evaluate synthesis and removal tasks, respectively.

CLIP Embedding Shift Alignment Score (ESAS). This metric assesses alignment between the generated image and target transformation in embedding space. Let f_{orig}, f_{gen}, f_+, and f_- represent embeddings for the original image, generated image, positive prompt, and negative prompt, respectively [31].

The semantic direction in embedding space is given by:

$$f_{\text{dir}} = \begin{cases} f_- - f_+, & \text{for "removal"} \\ f_+ - f_-, & \text{for "addition"} \end{cases} \tag{16}$$

We obtain the target embedding f_{target} by shifting f_{orig} along this direction:

$$f_{\text{target}} = f_{\text{orig}} + f_{\text{dir}} \tag{17}$$

The Embedding Shift Alignment Score (ESAS) is the Mean Squared Error (MSE) between f_{gen} and f_{target}:

$$\text{ESAS} = \frac{1}{N}\sum_{i=1}^{N}\left(f_{\text{gen}}^{(i)} - f_{\text{target}}^{(i)}\right)^2 \tag{18}$$

ESAS provides a measure of identity preservation and semantic shift, with lower values indicating successful alignment.

Fréchet Inception Distance (FID). Fréchet Inception Distance (FID) [10] measures the quality of generated images by comparing feature distributions of real and synthetic images within a high-dimensional space. Using embeddings from a pretrained model, FID calculates the distance between the distributions of real and generated data. Lower FID scores indicate a closer match to real data. In our evaluations, we compute FID between edited images and corresponding ground truth images, targeting the presence or absence of specific diseases to assess modification effectiveness.

5. Experiments and Results

5.1. Results on Disease Removal

In this section, we show some the disease removal capabilities of our model.

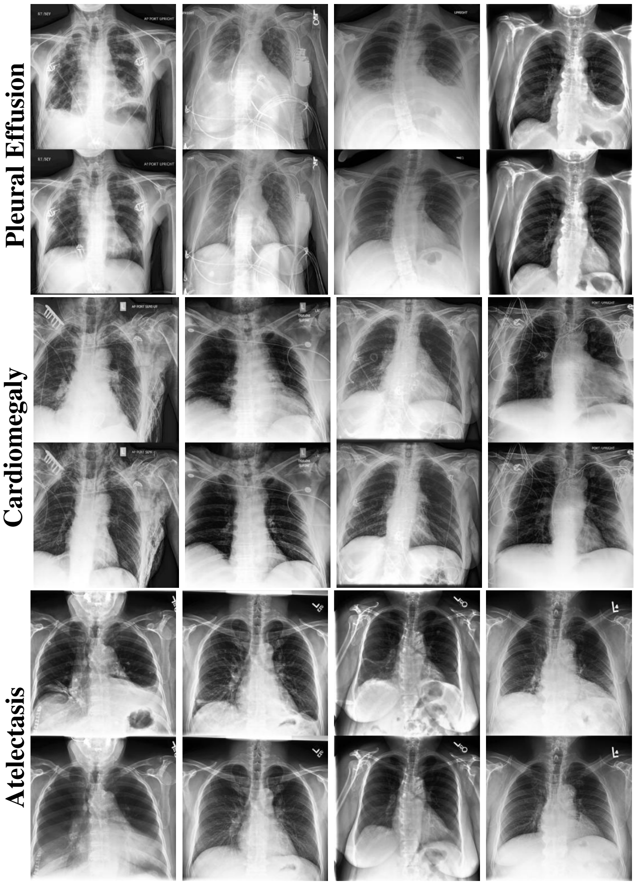
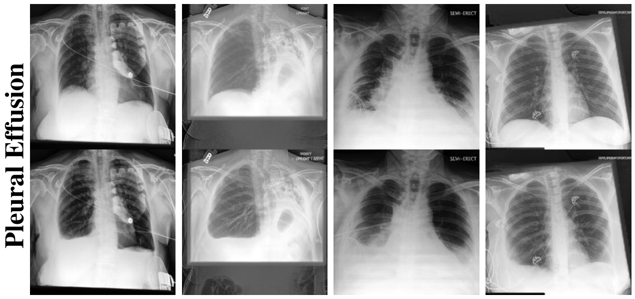


Figure 3. Qualitative Results of Disease Removal. First rows show the original image while second rows show the counterfactual one.

5.2. Results on Disease Synthesis

In this section, we show some the disease synthesis capabilities of our model.



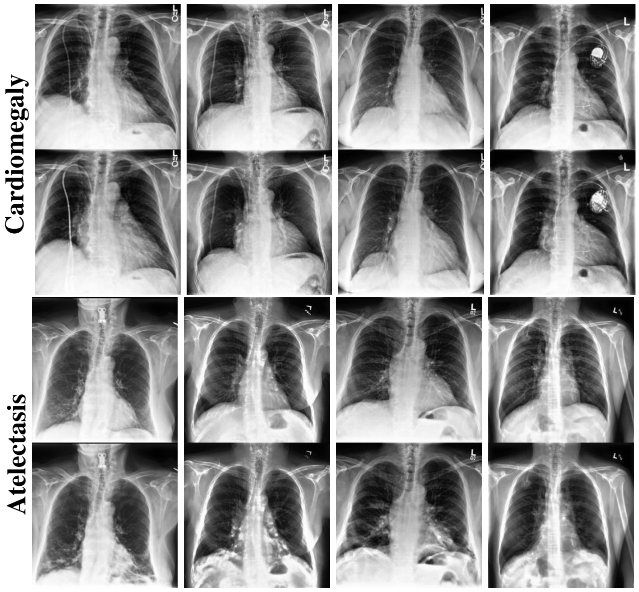


Figure 5. Qualitative Results of Disease Synthesis. First rows show the original image while second rows show the counterfactual one.

5.3. Comparison with Other Models

We present qualitative and quantitative results for our method alongside competing approaches.

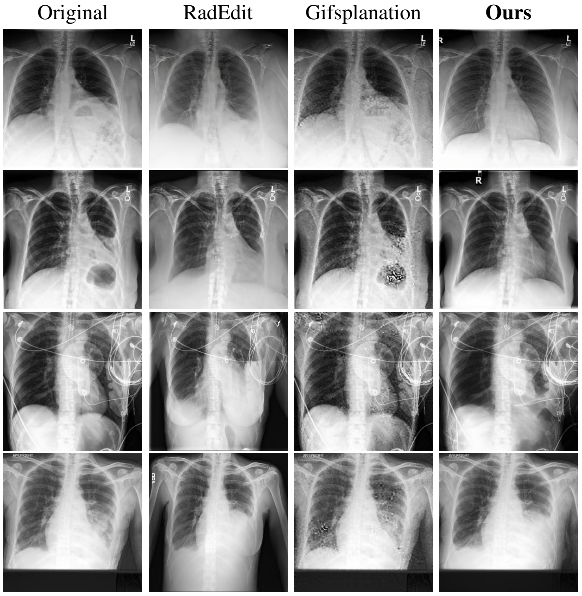


Figure 6. Qualitative Results of Ours and Competing Models on Pleural Effusion Removal (first two) and Synthesis (last two rows).

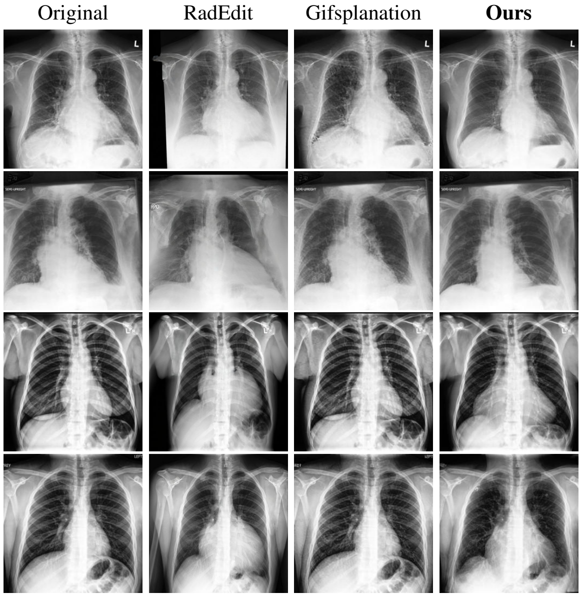


Figure 7. Qualitative Results of Ours and Competing Models on Cardiomegaly Removal (first two) and Synthesis (last two rows).

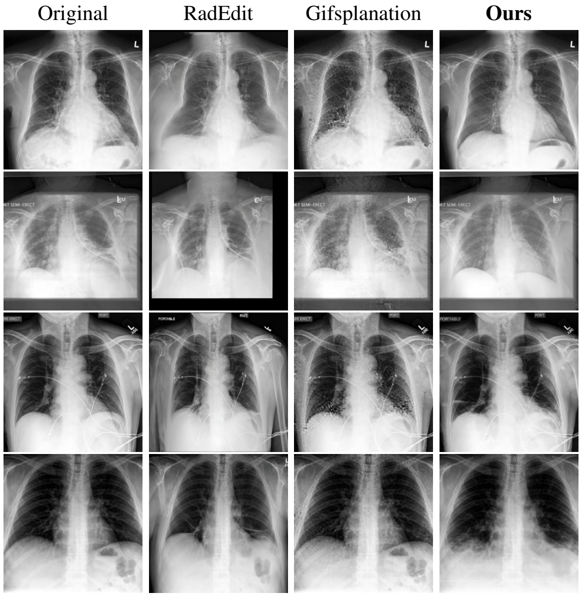


Figure 8. Qualitative Results of Ours and Competing Models on Atelectasis Removal (first two rows) and Synthesis (last two rows).

Method	CLIP Accuracy	ESAS	FID
RadEdit	0.69	0.145	69.42
Gifsplanation	0.37	0.159	84.23
Ours	0.85	0.075	49.91

Table 1. CLIP Accuracy, ESAS and FID for Different Models

5.4. Ablation Study

Using Bone Masks for Identity Preservation Here we show the effectiveness of incorporating bone masks into our model which helps preserve identity.

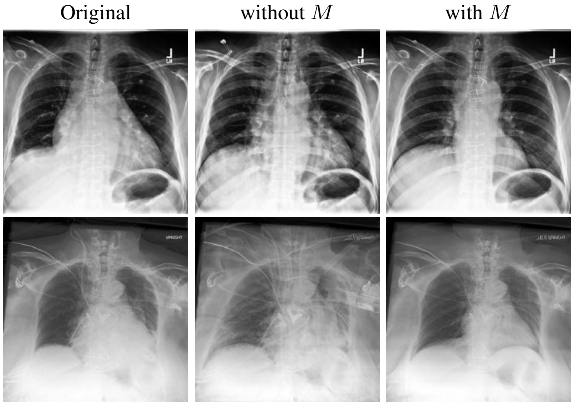


Figure 9. Qualitative Results of Ablation Study of Using Bone Masks on Cardiomegaly Removal.

Using L_1 Regularization Here we show the effectiveness of incorporating L_1 loss into our model which helps preserve details. L_1 loss is helpful at preserving details while allowing edits in the image according to our experiments.

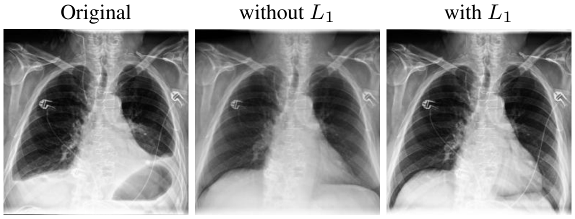


Figure 10. Qualitative Results of Ablation Study of L_1 Loss on Pleural Effusion Removal.

Fine-tuning Different Layers In this section, we show that fine-tuning only the attention layers lead to high quality counterfactual image generation while preserving the identity of the patient without the use of specific masks. Moreover, fine-tuning only the attention layers provides a good balance between CLIP Accuracy score and FID as compared to other layers.

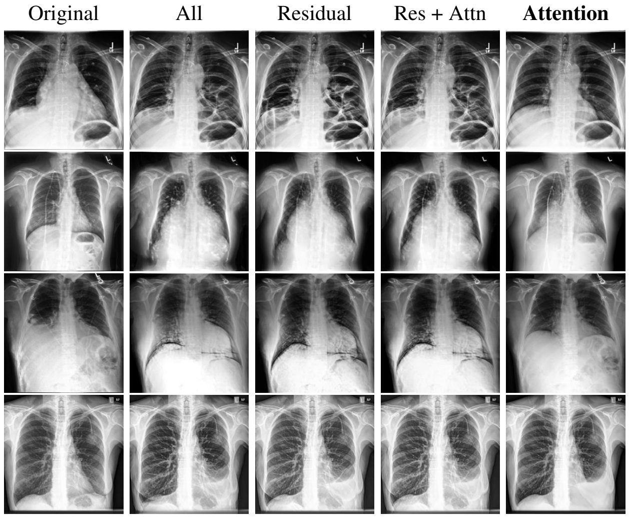


Figure 11. Qualitative Results of Fine-Tuning Different Layers on Cardiomegaly Removal (First Row), Cardiomegaly Synthesis (Second Row), Pleural Effusion Removal (Third Row) and Pleural Effusion Synthesis (Fourth Row). Best viewed zoomed-in.

Method	CLIP Accuracy	ESAS	FID
All	0.93	0.093	72.85
Residual	0.93	0.093	66.70
Res + Attn	0.93	0.094	68.70
Attention	0.85	0.075	49.91

Table 2. CLIP Accuracy, ESAS Score and FID for Different Methods. Scores are calculated using 500 edited images as the average of different tasks.

Effect of Extended Training We observe that running the model for more than 2–3 epochs typically offers no additional benefit, as the model stabilizes beyond this point. Additionally, as the classifier's prediction score decreases with further training, the edits begin to deviate, becoming less relevant to the intended modifications.

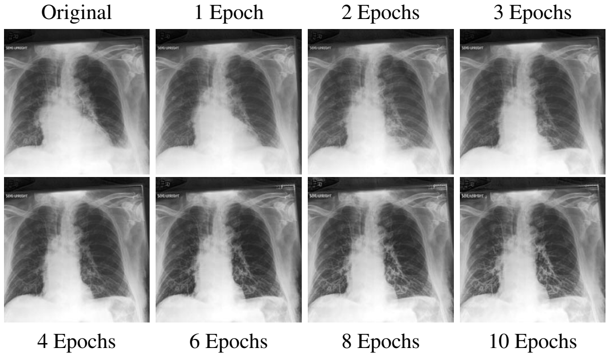


Figure 12. Qualitative Results of Ablation Study of Running the Model for 10 Epochs for Cardiomegaly Removal

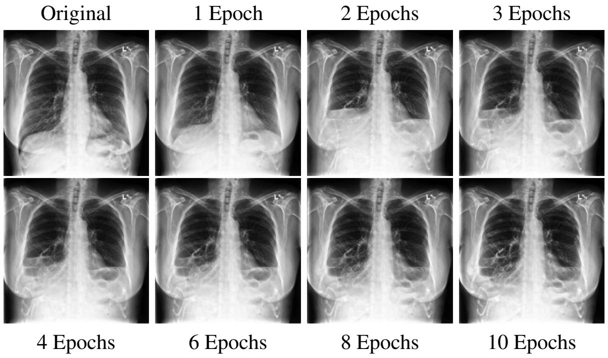


Figure 13. Qualitative Results of Ablation Study of Running the Model for 10 Epochs for Pleural Effusion Synthesis

6. Conclusions

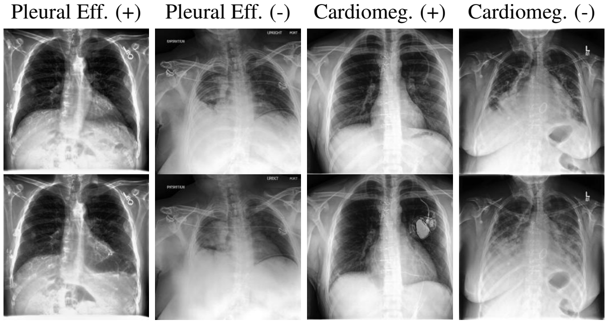


Figure 14. Qualitative Results of Failure Cases. First row shows the original images while the second row shows the failure cases. The (+) sign denotes synthesis while (-) denotes removal.

We proposed a diffusion-based approach for counterfactual disease manipulation in chest X-rays, capable of both disease removal and synthesis. Guided by a pretrained classifier, our method achieves realistic and clinically relevant edits by selectively modifying disease attributes while preserving patient-specific anatomical features. Our approach surpasses traditional latent and mask-based techniques in realism, identity preservation, and usability, making it a valuable tool for medical imaging.

Broader Impacts. This technique could advance medical imaging applications in diagnostic training, treatment planning, and education by enabling realistic disease manipulation in chest X-rays. While it can enhance clinical insights, careful management is needed to avoid ethical concerns or misinterpretation.

Limitations. Limitations remain, particularly if the classifier misidentifies disease, resulting in ineffective edits. These challenges point to a need for improved classifier accuracy and model guidance. Examples of failure cases are provided in Fig. 14.

References

[1] Matan Atad, Vitalii Dmytrenko, Yitong Li, Xinyue Zhang, Matthias Keicher, Jan Kirschke, Bene Wiestler, Ashkan Khakzar, and Nassir Navab. Chexplaining in style: Counterfactual explanations for chest x-rays using stylegan, 2022. 1, 3

[2] Matan Atad, David Schinz, Hendrik Moeller, Robert Graf, Benedikt Wiestler, Daniel Rueckert, Nassir Navab, Jan S. Kirschke, and Matthias Keicher. Counterfactual explanations for medical image classification and regression using diffusion autoencoder. *Machine Learning for Biomedical Imaging*, 2(iMIMIC 2023):2103–2125, 2024. 3

[3] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020. 3

[4] Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated x-ray prediction. In *Medical Imaging with Deep Learning*, 2020. 4, 12

[5] Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P. Lungren, and Akshay Chaudhari. Gifsplanation via latent shift: A simple autoencoder approach to counterfactual generation for chest x-rays, 2021. 1, 3, 5

[6] Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. In *Medical Imaging with Deep Learning*, 2022. 4, 12

[7] J. Duncan, M. Insana, and N. Ayache. Biomedical imaging and analysis in the age of big data and deep learning. *Proc. IEEE*, 108:3–10, 2020. 1

[8] Yingying Fang, Shuang Wu, Zihao Jin, Shiyi Wang, Caiwen Xu, Simon Walsh, and Guang Yang. Diffexplainer: Unveiling black box models via counterfactual generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 208–218, Cham, 2024. Springer Nature Switzerland. 3

[9] Lukas Fetty, Mikael Bylund, Peter Kuess, Gerd Heilemann, Tufve Nyholm, Dietmar Georg, and Tommy Löfstedt. Latent space manipulation for high-resolution medical image synthesis via the stylegan. *Zeitschrift fur medizinische Physik*, 2020. 1

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. 5

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

[12] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. 2, 4

[13] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019. 2, 3, 4, 5

[14] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 3

[15] Mingjie Li, Haokun Lin, Liang Qiu, Xiaodan Liang, Ling Chen, Abdulmotaleb Elsaddik, and Xiaojun Chang. Contrastive learning with counterfactual explanations for radiology report generation. In *Computer Vision – ECCV 2024*, pages 162–180, Cham, 2025. Springer Nature Switzerland. 1, 3

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 5

[17] 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, 2022. 1

[18] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022. 3

[19] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2, 3

[20] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[21] Muzaffer Ozbey, Salman UH Dar, Hasan Atakan Bedel, Onat Dalmaz, cSaban Ozturk, Alper Gungor, and Tolga cCukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42: 3524–3539, 2022. 1

[22] Fernando Pérez-García, Sam Bond-Taylor, Pedro P. Sanchez, Boris van Breugel, Daniel C. Castro, Harshita Sharma, Valentina Salvatelli, Maria T. A. Wetscherek, Hannah Richardson, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, Ozan Oktay, and Maximilian Ilse. Radedit: stress-testing biomedical vision models via diffusion image editing, 2024. 2, 5

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[24] Eduardo P Reis, Joselisa PQ De Paiva, Maria CB Da Silva, Guilherme AS Ribeiro, Victor F Paiva, Lucas Bulgarelli, Henrique MH Lee, Paulo V Santos, Vanessa M Brito, Lu-

cas TW Amaral, et al. Brax, brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):487, 2022. 3

[25] 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*, pages 10684–10695, 2022. 3

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2

[27] J. Schaefferkoetter, Jianhua Yan, Sangkyu Moon, Rosanna Chan, C. Ortega, U. Metser, A. Berlin, and P. Veit-Haibach. Deep learning for whole-body medical image generation. *European Journal of Nuclear Medicine and Molecular Imaging*, 48:3817 – 3826, 2021. 1

[28] Sumedha Singla, Motahhare Eslami, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. Explaining the black-box smoothly—a counterfactual approach. *Medical Image Analysis*, 84:102721, 2023. 1, 3

[29] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265, Lille, France, 2015. PMLR. 2, 3

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[31] Edward Tiu, Emily Talius, Priya Patel, et al. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 2022. 2, 5

[32] Ke Wang, Zicong Chen, Mingjia Zhu, Zhetao Li, Jian Weng, and Tianlong Gu. Score-based counterfactual generation for interpretable medical image classification and lesion localization. *IEEE Transactions on Medical Imaging*, 43(10): 3596–3607, 2024. 3

[33] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 3

[34] Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. Cascaded latent diffusion models for high-resolution chest x-ray synthesis. In *Advances in Knowledge Discovery and Data Mining: 27th Pacific-Asia Conference, PAKDD 2023*. Springer, 2023. 3, 5

[35] Tian Xia, Mélanie Roschewitz, Fabio De Sousa Ribeiro, Charles Jones, and Ben Glocker. Mitigating attribute amplification in counterfactual image generation, 2024. 3

[36] Wenwu Ye, Jin Yao, Hui Xue, and Yi Li. Weakly supervised lesion localization with probabilistic-cam pooling, 2020. 4, 5

[37] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K. Hong, Woonhyuk Baek, and Byungseok Roh. *CXR-CLIP: Toward Large Scale Chest X-ray Language-Image Pre-training*, page 101–111. Springer Nature Switzerland, 2023. 2

[38] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2024. 2

A. Appendix

A.1. Detailed Breakdown of Scores for Training Layer Ablation Study

Atelectasis Synthesis

Method	CLIP Accuracy	ESAS	FID
All	0.83	0.072	78.64
Residual	0.81	0.071	64.68
Res + Attn	0.82	0.078	68.17
Attention	0.70	0.049	44.63

Table 3. CLIP Accuracy, ESAS Score and FID for Different Methods. Scores are calculated using 500 edited images.

Cardiomegaly Synthesis

Method	CLIP Accuracy	ESAS	FID
All	0.96	0.105	79.30
Residual	0.94	0.104	72.26
Res + Attn	0.95	0.104	74.37
Attention	0.80	0.089	64.23

Table 4. CLIP Accuracy, ESAS Score and FID for Different Methods. Scores are calculated using 500 edited images.

Pleural Effusion Synthesis

Method	CLIP Accuracy	ESAS	FID
All	0.98	0.112	58.31
Residual	0.97	0.114	55.61
Res + Attn	0.95	0.113	57.67
Attention	0.91	0.089	49.62

Table 5. CLIP Accuracy, ESAS Score and FID for Different Methods. Scores are calculated using 500 edited images.

Atelectasis Removal

Method	CLIP Accuracy	ESAS	FID
All	0.93	0.073	81.82
Residual	0.94	0.070	78.38
Res + Attn	0.93	0.072	79.10
Attention	0.88	0.058	49.84

Table 6. CLIP Accuracy, ESAS Score and FID for Different Methods. Scores are calculated using 500 edited images.

Cardiomegaly Removal

Method	CLIP Accuracy	ESAS	FID
All	0.97	0.091	66.56
Residual	0.97	0.090	57.80
Res + Attn	0.98	0.092	63.01
Attention	0.89	0.080	40.23

Table 7. CLIP Accuracy, ESAS Score and FID for Different Methods. Scores are calculated using 500 edited images.

Pleural Effusion Removal

Method	CLIP Accuracy	ESAS	FID
All	0.96	0.109	72.48
Residual	0.96	0.109	71.52
Res + Attn	0.96	0.107	69.92
Attention	0.93	0.090	50.93

Table 8. CLIP Accuracy, ESAS Score and FID for Different Methods. Scores are calculated using 500 edited images.

A.2. Detailed Breakdown of Scores for Different Models

Atelectasis Synthesis

Method	CLIP Accuracy	ESAS	FID
RadEdit	0.65	0.112	58.36
Gifsplanation	0.17	0.135	78.44
Ours	0.70	0.049	44.63

Table 9. CLIP Accuracy, FID and Average Time for Different Models

Cardiomegaly Synthesis

Method	CLIP Accuracy	ESAS	FID
RadEdit	0.92	0.140	64.13
Gifsplanation	0.19	0.165	68.36
Ours	0.80	0.089	64.23

Table 10. CLIP Accuracy, FID and Average Time for Different Models

Pleural Effusion Synthesis

Method	CLIP Accuracy	ESAS	FID
RadEdit	0.98	0.169	80.41
Gifsplanation	0.18	0.182	121.23
Ours	0.91	0.089	49.62

Table 11. CLIP Accuracy, FID and Average Time for Different Models

Atelectasis Removal

Method	CLIP Accuracy	ESAS	FID
RadEdit	0.72	0.120	67.88
Gifsplanation	0.76	0.133	68.81
Ours	0.88	0.058	49.84

Table 12. CLIP Accuracy, FID and Average Time for Different Models

Cardiomegaly Removal

Method	CLIP Accuracy	ESAS	FID
RadEdit	0.12	0.172	76.79
Gifsplanation	0.40	0.164	52.02
Ours	0.89	0.080	40.23

Table 13. CLIP Accuracy, FID and Average Time for Different Models

Pleural Effusion Removal

Method	CLIP Accuracy	ESAS	FID
RadEdit	0.76	0.16	68.95
Gifsplanation	0.52	0.175	116.52
Ours	0.93	0.090	50.93

Table 14. CLIP Accuracy, FID and Average Time for Different Models

A.3. Bone Segmentation Model

The bone segmentation by [4, 6] produces 14 different segmentations for each of Left Clavicle, Right Clavicle, Left Scapula, Right Scapula, Left Lung, Right Lung, Left Hilus Pulmonis, Right Hilus Pulmonis, Heart, Aorta, Facies Diaphragmatica, Mediastinum, Weasand and Spine. Each segmentation can be seen in Fig. 15 in the order written.

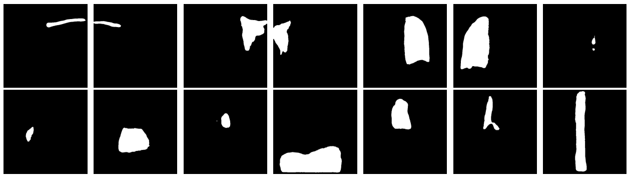


Figure 15. Qualitative Results of Different Masks Based On Segmentations.

Our final mask consists of individual masks of left clavicle, right clavicle, left scapula, right scapula, aorta, mediastinum, weasand and spine. Final mask can be seen in Fig. 16.

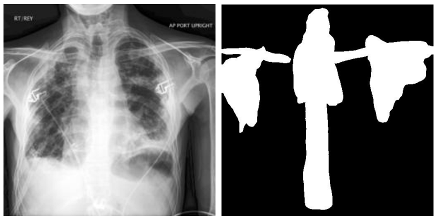


Figure 16. Qualitative Result of Final Used Mask.