Beyond Standard Sampling: Metric-Guided Iterative Inference for Radiologists-Aligned Medical Counterfactual Generation

Elham Ghelichkhan[1,2] ELHAM@SCI.UTAH.EDU
Tolga Tasdizen[1,3] TOLGA.TASDIZEN@UTAH.EDU
[1] *Scientific Computing and Imaging Institute, University of Utah, Utah, USA*
[2] *Kahlert School of Computing, University of Utah, Utah, USA*
[3] *Department of Electrical and Computer Engineering, University of Utah, Utah, USA*

Abstract

Generative counterfactuals offer a promising avenue for explainable AI in medical imaging, yet ensuring these synthesized images are both anatomically faithful and clinically effective remains a significant challenge. This work presents a domain-specific diffusion framework for generating "healthy" counterfactuals from chest X-rays with cardiomegaly, underpinned by a systematic metric-guided inference strategy. In contrast to methods relying on static sampling parameters, our approach iteratively explores the inference hyperparameter space to maximize our composite selection criterion, CF_Score, that integrates our novel Faithfulness-Effectiveness Trade-off (FET) metric.

We extend the evaluation of counterfactual utility beyond simple classification shifts by conducting the simultaneous validation against radiologist annotations and eye-tracking data. Using the REFLACX dataset, we demonstrate that difference maps derived from our counterfactuals exhibit strong spatial alignment with expert visual attention and annotation. Quantified by Normalized Cross-Correlation, Hit Rate, pixel-wise ROC-AUC, and AUC-IoU, our results confirm that metric-guided counterfactuals provide dense and clinically relevant localizations of pathology that closely mirror human diagnostic reasoning.

Keywords: Diffusion Models, Counterfactual, Chest X-ray, XAI, Evaluation

1. Introduction

Chest radiography is one of the most frequently performed imaging examinations worldwide for screening and diagnosing cardiothoracic disease. Recent advances in deep learning have enabled models that achieve radiologist-level performance on several chest X-ray (CXR) classification tasks, including detection of cardiomegaly. Despite this progress, concerns around transparency and trust remain: clinicians must be able to understand *why* a model predicts cardiomegaly on a given study, and whether that prediction is grounded in anatomically plausible image evidence. Most existing explainability approaches for CXR models rely on saliency-based methods such as class activation maps, gradient-based attribution, or perturbation-based importance scores (Selvaraju et al., 2017; Nazir et al., 2023). While these methods are simple to apply, they may highlight large non-specific regions, and can be difficult to interpret in terms of concrete image changes.

Generative diffusion models have recently emerged as powerful tools for high-fidelity image synthesis and editing (Ho et al., 2020; Rombach et al., 2021). In principle, they provide an attractive mechanism for counterfactual explanation: given a factual image

displaying cardiomegaly, a conditional diffusion model could be used to generate a "no-cardiomegaly" version of the same anatomy, and the difference between the two images could serve as an actionable explanation. However, the outcome of diffusion sampling is highly sensitive to inference hyperparameters, e.g., guidance strength, denoising strength, null-text optimization, and naive choices can lead to unrealistic or overly edited counterfactuals that alter irrelevant structures or introduce artifacts. Furthermore, there is a lack of rigorous, clinically grounded evaluation protocols for counterfactual explanations: it is unclear how to quantify whether a generated counterfactual simultaneously (i) removes the target pathology and (ii) remains faithful to the original image.

In this work, we address these gaps in the setting of cardiomegaly on frontal CXRs. Our key methodological idea is to explicitly quantify and balance *faithfulness* and *effectiveness*. Faithfulness measures how well the counterfactual preserves the patient's identity and non-target anatomy; we assess it using perceptual similarity (LPIPS), L2, and cosine similarity. Effectiveness measures whether the counterfactual actually removes evidence of cardiomegaly; we capture this via the change in logit output of a strong CXR classifier between the factual and counterfactual images. We combine these metrics into a single faithfulness–effectiveness trade-off score and use it to drive an iterative inference procedure that explores different choices of diffusion checkpoint, denoising strength, and guidance scale. For each factual image, we retain counterfactuals that satisfy stringent thresholds on all metrics, ensuring that only high-quality explanations are used downstream.

To evaluate whether diffusion-based counterfactuals provide clinically meaningful explanations, we study their spatial alignment with expert annotations and radiologist visual attention. We leverage the REFLACX (Bigolin Lanfredi et al., 2022, 2021) dataset which includes radiologists' eye-tracking data for a subset of MIMIC-CXR (Johnson et al., 2019d,c,a,b) studies and analyze the *difference maps* between factual and counterfactual images. Our experiments show that difference maps derived from our metric-guided counterfactuals are competitive with, and in several regimes surpass, eye-tracking as a spatial explanation signal for cardiomegaly: they robustly highlight the heart borders and surrounding regions implicated in radiologists' annotations, while remaining tightly localized within the thorax.

In summary, our contributions are twofold:

- We introduce a suite of complementary metrics that jointly quantify faithfulness and effectiveness of counterfactuals, and we propose a metric-guided inference scheme that searches over diffusion hyperparameters and null-text inversion (Mokady et al., 2023) settings to select high-quality counterfactuals on a per-image basis.

- We conduct, to our knowledge, the first systematic comparison of diffusion-based counterfactual difference maps with both radiologist annotations and eye-tracking heatmaps in CXRs, demonstrating that metric-guided counterfactuals provide spatial explanations that are well aligned with expert annotations and clinically relevant image regions.

2. Method

2.1. Dataset Refinement

We fine-tuned a Stable Diffusion Model (SDM) (Rombach et al., 2021) using frontal (PA/AP view) *No Finding* and *Cardiomegaly* images from the MIMIC-CXR-2.0.0 dataset (Johnson et al., 2019d,c,a,b). To ensure high data quality, we fine-tuned a Vision-Transformer (ViT) classifier (Dosovitskiy, 2020) pretrained on chest X-rays to filter out noisy, lateral, rotated, and out-of-domain samples (see Appendix A for data refinement details). For evaluation, we utilized the REFLACX dataset (Bigolin Lanfredi et al., 2022, 2021), a subset of MIMIC. Although the diffusion model encountered these images during self-supervised reconstruction training, it was never trained to generate counterfactuals, ensuring that our evaluation of counterfactual utility remains valid.

2.2. Training and Validation

Following (Kumar et al., 2025), we fine-tuned the text-conditional U-Net component of a Stable Diffusion Model (Rombach et al., 2021) pre-trained on general datasets, on our *Cardiomegaly/No Finding* MIMIC subset (Appendix A). In addition, we enabled Classifier-Free Guidance (Ho and Salimans, 2022) ($p_{uncond} = 0.1$) by conditioning on null tokens or the prompt *"this is a chest radiograph from AP/PA view WITHOUT [or WITH] CARDIOMEGALY"*. For inference, we selected the two checkpoints yielding the lowest L1 and L2 reconstruction losses, evaluated on 100 samples, randomly selected from MIMIC validation set, with applying Null-Text Inversion (Mokady et al., 2023).

2.3. Evaluation Metrics

This study aims to generate Normal counterfactuals for images labeled as *Cardiomegaly*. To identify optimal counterfactuals, we balance two competing objectives: *faithfulness* (preserving identity) and *effectiveness* (removing abnormality). Recognizing the inherent trade-off where strict faithfulness limits necessary edits and excessive effectiveness alters irrelevant features, our framework strictly enforces both criteria to ensure modifications remain target-specific.

Faithfulness. To quantify how well the counterfactual preserves the original anatomy, we measure the L2 loss in both pixel and embedding spaces and the cosine similarity between two images. We specifically utilize the Learned Perceptual Image Patch Similarity (LPIPS) metric (Zhang et al., 2018), which consists of L2 distance between feature maps of an image encoder for the factual and counterfactual images, offering a robust measure of perceptual similarity. Additionally, we compute the L2 distance in image space to capture raw visual fidelity. Finally, to strictly enforce faithfulness, we utilize the cosine similarity of CLIP (Radford et al., 2021) image embeddings alongside LPIPS and L2. Appendix B reviews popular evaluation metrics (Prabhu et al., 2023; Radford et al., 2021; Hore and Ziou, 2010; Zhang et al., 2018; Salimans et al., 2016; Heusel et al., 2017; Jain et al., 2025) employed in prior counterfactual studies and details the rationale for their exclusion from this work.

Effectiveness. We quantify the removal of pathological features using Classification Prediction Gain (CPG) (Nemirovsky et al., 2020). We fine-tuned a ViT binary classifier—pretrained

on five classes (TheSmartTechnologyLab, 2025) of CheXpert dataset (Irvin et al., 2019)—on a balanced MIMIC subset to calculate CPG based on the shift in raw *Cardiomegaly* logits for Factual (F) vs Counterfactual (CF) images:

$$\mathrm{CPG}(F, CF) = \mathrm{ViT}(F) - \mathrm{ViT}(CF), \tag{1}$$

A higher CPG value indicates a more successful removal of the pathology features. Note that this metric requires only a strong score-based classifier, rather than a perfectly calibrated one.

Table 1: *FET-score* in all faithfulness-effectiveness scenarios

	Effectiveness	**Faithfulness**	***FET-score***
(1)	✓($-\log\uparrow$)	✓($CPG\uparrow$)	*FET-score*↑↑
(2)	✓($-\log\uparrow$)	✓($CPG\downarrow$)	*FET-score*↑
(3)	✗($-\log\downarrow$)	✗($CPG\uparrow$)	*FET-score*↑
(4)	✗($-\log\downarrow$)	✓($CPG\downarrow$)	*FET-score*↓

Table 2: *Optimized τ in Eq. 3.*

Threshold	Value
τ_{LPIPS}	0.2
τ_{CLIP}	0.92
τ_{L2}	0.3
τ_{CPG}	0.1
τ_{FET}	4

Faithfulness-Effectiveness Trade-off score To jointly evaluate the balance between these objectives, we propose the Faithfulness-Effectiveness Trade-off (FET) score. This metric combines the effectiveness score (CPG) with latent distance metrics:

$$\textit{FET-score}(F, CF) = \mathrm{CPG}(F, CF) - \log(\mathrm{LL1}(F, CF) \cdot \mathrm{LL2}(F, CF)) \tag{2}$$

where LL1 and LL2 represent the L1 and L2 distances between the CLIP latent representations of F and CF. Since LL1 and LL2 are typically less than 1, the logarithmic term is positive; thus, higher faithfulness (smaller latent distance) results in a larger log value, rewarding the preservation of identity. Figure 1 illustrates sample images and their corresponding scores across different trade-off scenarios, as detailed in Table 1.

Proposed Criteria. As implied by the Eq. 2, a counterfactual that is highly faithful but ineffective, or highly effective but unfaithful, achieves a high FET-score. To ensure rejection of non-faithful or non-effective generated explanations, we combine all metrics into a final acceptance criterion:

$$\begin{aligned}\mathrm{CF\_Score} = \mathbb{1}(\mathrm{LPIPS} < \tau_{\mathrm{LPIPS}}) + \mathbb{1}(Sim_{CLIP} > \tau_{\mathrm{CLIP}}) + \mathbb{1}(\mathrm{L2} < \tau_{\mathrm{L2}}) \\ + \mathbb{1}(\mathrm{CPG} > \tau_{\mathrm{CPG}}) + \mathbb{1}(\mathrm{FET} > \tau_{\mathrm{FET}})\,,\end{aligned} \tag{3}$$

where $\mathbb{1}$ is the indicator function. The thresholds (τ) for each component should be optimized empirically. Table 2 shows optimized thresholds for our studies.

2.4. Inference Strategy

We generate counterfactuals using Null-Text Inversion (NTI) (Mokady et al., 2023) on our fine-tuned SDM. We optimize null embeddings ($\varnothing$) on the factual image (prompt: *"...WITH..."*) and perform Deep Denoising Implicit Model (DDIM) sampling with the

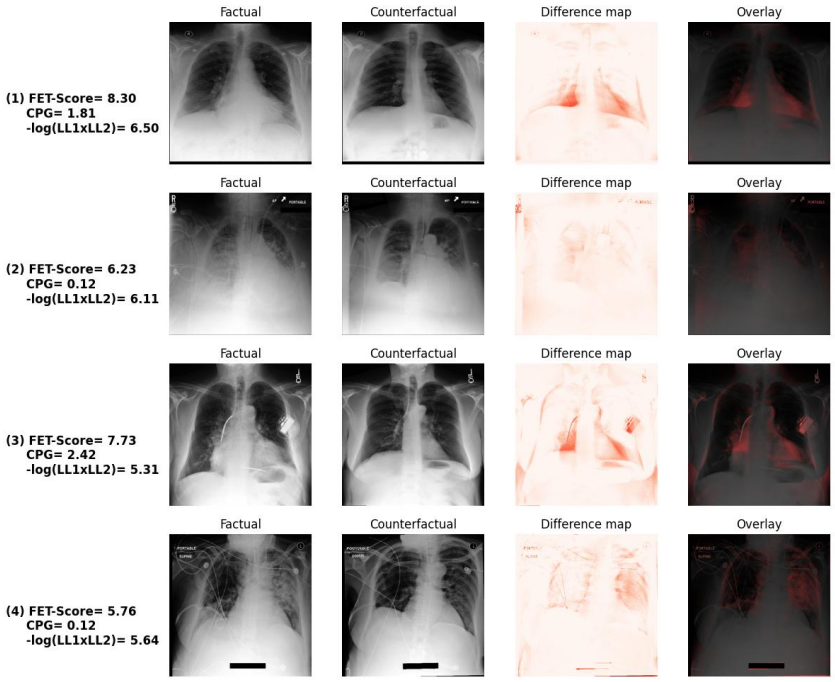

Figure 1: Different samples in Faithfulness-Effectiveness trade-off and their FET-score

target prompt (*"...WITHOUT..."*). Since single-pass inference often fails our strict criteria (see Appendix E), we employ a metric-guided search strategy (Algorithm 1) that iterates over denoising strength (δ), guidance strength (ω), model best checkpoints (θ), and null-text optimization hyperparameters ($\mathcal{H}_{null}$).

Recent work suggests counterfactual generation is most effective in a "balanced space," equidistant from the source manifold and generated distribution (Yan et al.). We approach this state using high denoising strengths ($\delta \in [0.8, 1.0]$) to facilitate large semantic changes, while leveraging null-text optimization to maintain structural fidelity (Mokady et al., 2023).However, high denoising and guidance ($\omega \in \{7, \ldots, 10\}$) are insufficient if structural constraints are too rigid. We identified the null-text optimization as a key factor in this rigidity: an overfitted null embedding memorizes high-frequency pixel details, restricting the model's ability to alter the cardiac silhouette. To enhance editability, our search includes "underfitting" null-text optimization settings ($\mathcal{H}_{null}$), which relax these constraints and allow the generative pipeline greater flexibility to modify the target anatomy while preserving the patient's identity.

Our search strategy increased the generation success rate from 33% (default hyperparameters) to 93%. As detailed in 3, the resulting counterfactuals exhibit high validity and consistency with human alignment and region of interest.

2.5. Evaluation

In this study, we define "segmentation masks" and "annotations" as the annotations of the *Enlarged Cardiac Silhouette* from the REFLACX dataset. This is an approximate segmentation, obtained by the radiologists placing an ellipse around the heart, when cardiomegaly is present. Additionally, we refer to the label-specific eye-tracking heatmaps for *Cardiomegaly* from (Bigolin Lanfredi et al., 2023) as "eye-tracking heatmaps".

Cardiomegaly is clinically characterized by an enlarged cardiac silhouette. Thus, a valid normal counterfactual must demonstrate heart size reduction. We therefore isolate the positive component of the difference map (Diff Map $= I_{\text{factual}} - I_{\text{counterfactual}} > 0$), capturing the relevant anatomical darkening along retracted borders while zeroing out non-diagnostic negative values. Furthermore, to eliminate confounding background noise from imperfect text reconstruction, we constrain our analysis—including difference maps, eye-tracking heatmaps, and annotations—to the chest bounding boxes provided in the REFLACX dataset. More details on importance of chest box is provided in Appendix D.

To assess the clinical validity of the generated counterfactuals, we compare the masked difference maps against two forms of expert ground truth: (1) radiologist segmentation masks of the abnormality and (2) eye-tracking heatmaps for label *Cardiomegaly*, representing radiologists' visual attention for diagnosing this abnormality. Ideally, a counterfactual difference map should highlight the reduction of the cardiac silhouette; we validate this using radiologist eye-tracking as a clinical benchmark: we hypothesize that a high-quality counterfactual should exhibit a spatial alignment with the heart borders, comparable to the alignment observed between expert visual attention and same boundaries.

We quantify this spatial alignment using Normalized Cross-Correlation (NCC), Hit Rate, pixel-wise ROC-AUC, and threshold-IoU. Additionally, we validate effectiveness of counterfactuals via Classification Prediction Gain (CPG) using two independent ViT classifiers distinct from the model used in our selection criteria.

3. Experiments and Results

We evaluated all generated counterfactuals (93% passed all criteria and 7% passed four out of five in Eq. 3) across three dimensions: (1) *Effectiveness*, measuring the successful reduction of pathological features as perceived by independent classifiers; (2) *Spatial Alignment*, quantifying the overlap with ground-truth radiologist annotations; and (3) *Explainability*, benchmarking the counterfactual difference maps against radiologist eye-tracking data.

3.1. Effectiveness

To validate that our counterfactuals reflect a "healthy" state, we analyzed the class probability shift for *Cardiomegaly* using two independent ViT classifiers distinct from the one used in our optimization loop. The first was pre-trained on the CheXpert dataset (TheSmartTechnologyLab, 2025), and we fine-tuned the second on five labels of the MIMIC dataset.

Figure 2 summarizes the Classification Prediction Gain (CPG), defined here as the reduction in class probability ($P_{\text{factual}} - P_{\text{counterfactual}}$). Both models demonstrate a consistent positive shift, confirming that the counterfactuals reduce the perceived pathology.

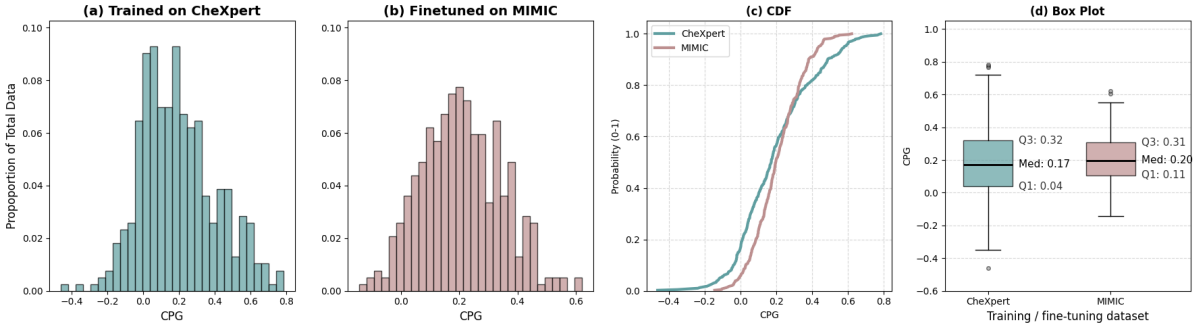

Figure 2: Classification Prediction Gain (CPG) analysis. (a) Histogram of CPG from a ViT-classifier trained on CheXpert. (b) Histogram of CPG from the same model fine-tuned on MIMIC. (c) Cumulative Distribution Function (CDF) of CPG for both models. (d) Box plot comparison.

The model fine-tuned on MIMIC exhibits a tighter distribution with higher confidence, reflecting its familiarity with the domain. In contrast, the CheXpert-trained model shows slightly greater variance; this is expected due to the domain gap between CheXpert and MIMIC datasets. In addition, the Cumulative Distribution Function (Figure 2(c)) reveals that even under domain shift, only a negligible fraction of samples have a negative CPG, demonstrating the robustness of our counterfactual generation.

3.2. Explainability and Spatial Alignment

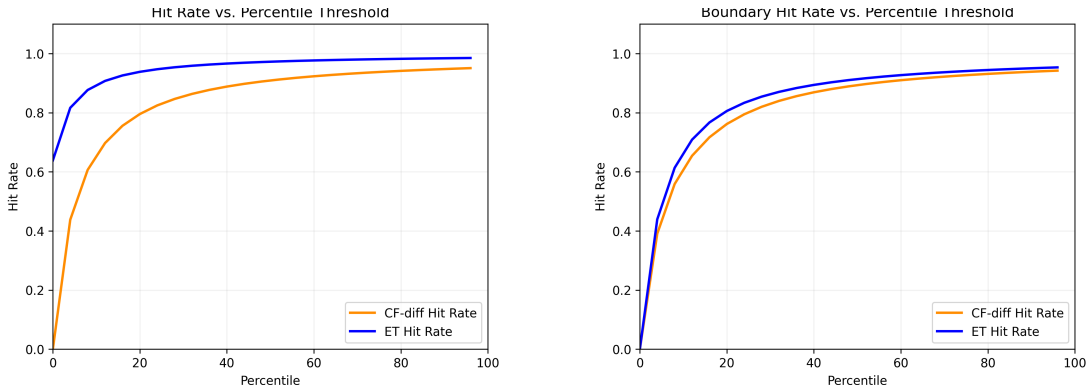

Figure 3: Hit Rate analysis. (a) Hit Rate of CF difference maps and eye-tracking heatmaps against *Cardiomegaly* segmentations. (b) Hit Rate against the boundary of the segmentation.

We benchmark the difference maps against radiologist eye-tracking (ET) heatmaps using Hit Rate, adapted from (Saporta et al., 2022; Arun et al., 2021), which measures the

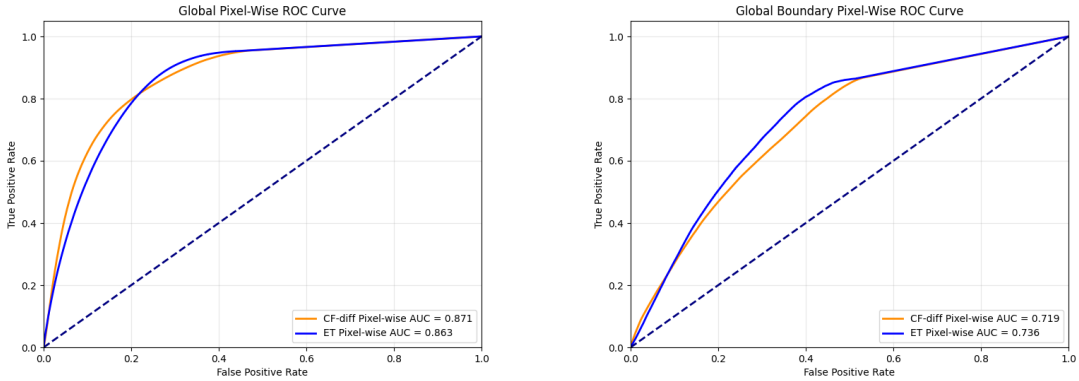

Figure 4: Pixel-wise ROC-AUC curves. (a) ROC-AUC against Cardiomegaly segmentations. (b) ROC-AUC against segmentation boundaries.

proportion of samples where the top-k percentile pixels of the heatmap fall within the ground-truth.

Figure 3(a) illustrates the Hit Rate against the full cardiomegaly segmentation. Radiologist eye-tracking (blue) shows high initial hit rate, with the most fixated pixels falling inside the annotation ∼65% of the time. Our counterfactual difference maps (orange) begin with zero hit rate but steadily narrow the performance gap relative to eye-tracking as the threshold relaxes. Figure 3(b) evaluates alignment with the *boundary* of the heart—the critical diagnostic feature. Both heatmaps exhibit similar trends, confirming that the counterfactuals, aligned with radiologists' eye-tracking heatmaps, localize the cardiac margins rather than irrelevant regions.

These findings are corroborated by the pixel-wise ROC-AUC curves (Figure 4). For the full segmentation (a), the counterfactual difference maps slightly outperform eye-tracking (AUC 0.871 vs. 0.863), indicating excellent separation of pathological pixels from the background. For the boundary task (b)—a significantly more challenging target defined by a single-pixel contour—performance degrades for both modalities, though eye-tracking retains a slight advantage (AUC 0.736 vs. 0.719), reflecting the precise visual attention of experts on edge definitions.

Table 3 presents the Normalized Cross-Correlation (NCC) scores. As noted in Section 2.5, text markers' reconstruction artifacts outside the chest field initially suppress the alignment scores (Appendix D). However, constraining the analysis to the chest bounding box significantly improves the NCC against annotations for both counterfactuals ($0.1786 \rightarrow 0.2842$) and eye-tracking heatmaps ($0.3983 \rightarrow 0.4424$). While eye-tracking retains a higher correlation with human annotations—expected since both are human-derived signals—the counterfactuals demonstrate meaningful anatomical alignment within the relevant region.

Crucially, this masking strategy reveals a strong structural concordance between the model's logic and human attention. When limiting the scope to the chest box, the direct alignment between eye-tracking heatmaps and counterfactual difference maps more than doubles, with the NCC rising from 0.1754 to 0.3696. This substantial increase confirms

Table 3: Normalized Cross-Correlation (NCC) between heatmaps (ET: Eye-Tracking, CF-Diff: Difference map) and Annotations (Ann). Top: Full heatmaps. Bottom: Masked to Chest Box.

Full Heatmaps	**ET**	**CF-Diff**	Ann
Ann	0.3983	0.1786	1.
CF-diff	0.1754	1.	-
ET	1.	-	-
In Chest Box			
Ann	0.4424	0.2842	1.
CF-diff	0.3696	1.	-
ET	1.	-	-

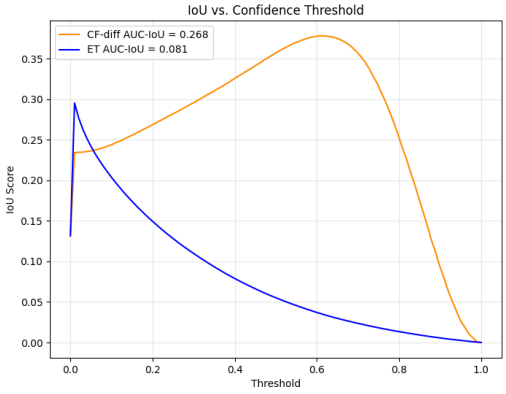

Figure 5: Threshold-IoU of heatmaps with *Cardiomegaly* annotations.

that within the chest, the counterfactual changes are consistent with radiologist visual attention; thereby validating the difference map as a clinically interpretable explanation.

Finally, we evaluated the Intersection over Union (IoU) across varying intensity thresholds (Figure 5). While eye-tracking achieves higher IoU at the strictest thresholds, counterfactual difference maps significantly outperform eye-tracking in overall shape coverage, yielding an AUC-IoU of 0.268 versus 0.081. This disparity reflects the fundamental difference between human diagnostic attention and generative reconstruction: Whereas radiologists often use a more efficient visual search, where fixating on just a portion of the heart's outline is sufficient to confirm enlargement, the diffusion model removes the entire pathological manifestation associated with the label. Consequently, the counterfactuals provide a dense, comprehensive map of the abnormality that aligns more fully with the dense segmentation masks.

3.3. Summary of Search Effectiveness

Direct benchmarking of generation yield for *Cardiomegaly* is challenging, as prior studies (Atad et al., 2022; Singla et al., 2023; Kumar et al., 2025) primarily rely on aggregate performance metrics or qualitative selection, rather than reporting per-sample success rates. However, our metric-guided search strategy proves highly robust: it produced valid (Eq. 3) counterfactuals for 361 of 387 images (93.3%), whereas naive inference with default hyperparameters yielded only 130 of 387 (33%) acceptable results, validating the efficacy of our proposed framework. For ablations studies on hyperparameter search effect, see Appendix E.

4. Discussion

This work introduces a domain-aware diffusion framework for generating healthy counterfactuals from cardiomegaly Chest X-rays. By employing a metric-guided inference strategy that satisfies our acceptance criteria (Eq. 3) and ensuring balanced Trade-off between effectiveness and faithfulness, we significantly outperformed naive sampling. To our knowledge,

this is the first study to validate generative difference maps against radiologist eye-tracking and annotations, demonstrating that our explanations exhibit strong spatial alignment with human diagnostic attention.

Clinical Applicability. Unlike opacities that require complete erasure, cardiomegaly necessitates a structural retraction of the heart border. Our framework is uniquely suited for such "boundary-shift" pathologies—including hyperinflation in COPD or disease progression (e.g., tumor growth)—where the difference map must highlight anatomical deformation rather than diffuse intensity changes.

Limitations. First, our iterative search strategy creates a computational bottleneck, requiring multiple inference passes per image. While our strict criteria demand substantial resources, the efficiency compromised for performance contributes to counterfactuals that align with radiologist annotations and attention. Future work could accelerate this process by narrowing the hyperparameter search space based on the specific failure modes encountered during generation. Second, imperfect text reconstruction introduces artifacts. We mitigate this by masking the output to the chest bounding box. Although chest bounding boxes are not available in most datasets, center-cropping can be an approximate substitute. Alternatively, more accurate chest boxes can be predicted by lightweight detection models, given the thorax's predictable large-scale structure. Finally, although zeroing out negative difference map values is not ideal for a pure counterfactual, it serves as an effective, automatic post-processing step to isolate relevant pathological features.

Future Work We aim to extend this framework to other chest abnormalities by developing pathology-specific metrics (e.g., distinguishing "removal" vs. "deformation"). Additionally, to address the strictness of single-pixel contour evaluation, we will employ dilated ground-truth borders to better account for the spatial uncertainty inherent in both human attention and generative editing, offering a more robust assessment of clinical validity.

5. Conclusion

The clinical adoption of AI in chest radiography relies on the transparency and trustworthiness of its predictions. While generative counterfactuals offer a powerful explainable mechanism for visualizing pathology, their utility has historically been limited by generation instability and a lack of rigorous, clinically grounded validation. In this work, we addressed these challenges by introducing a domain-aware diffusion framework driven by a novel **metric-guided inference strategy**. By dynamically optimizing the trade-off between anatomical faithfulness and pathological erasure via our proposed FET and CF_Score, our approach succeeds in synthesizing realistic healthy counterfactuals, significantly surpassing naive sampling methods.

Crucially, this study presents the first systematic validation of generative difference maps against radiologist eye-tracking data and their annotation. Our results demonstrate that metric-guided counterfactuals yield dense, clinically relevant explanations that exhibit strong spatial alignment with expert visual attention and segmentation. This confirms that generative models, when rigorously controlled, can mirror human diagnostic reasoning more effectively than traditional saliency methods. Future work will focus on optimizing the computational efficiency of the search strategy and extending this framework to diverse pathologies requiring different morphological changes.

References

Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021.

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. *arXiv preprint arXiv:2207.07553*, 2022.

R. Bigolin Lanfredi, M. Zhang, W. Auffermann, J. Chan, P. Duong, V. Srikumar, T. Drew, J.D. Schroeder, and T. Tasdizen. Reflacx: Reports and eye-tracking data for localization of abnormalities in chest x-rays (version 1.0.0)., 2021. URL https://doi.org/10.13026/e0dj-8498.

R. Bigolin Lanfredi, M. Zhang, W.F. Auffermann, J. Chan, P.A.T. Duong, V. Srikumar, T. Drew, J.D. Schroeder, and T. Tasdizen. Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Scientific data*, 9(1):350, 2022.

Ricardo Bigolin Lanfredi, Joyce D Schroeder, and Tolga Tasdizen. Localization supervision of chest x-ray classifiers using label-specific eye-tracking annotation. *Frontiers in radiology*, 3:1088068, 2023.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

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.

Suparshva Jain, Amit Sangroya, and Lovekesh Vig. Difclue: Generating counterfactual explanations with diffusion autoencoders and modal clustering. *arXiv preprint arXiv:2502.11509*, 2025.

A. Johnson, M. Lungren, Y. Peng, Z. Lu, R. Mark, S. Berkowitz, and S. Horng. Mimic-cxr-jpg: Chest radiographs with structured labels (version 2.0.0), 2019a. URL https://doi.org/10.13026/8360-t248.

A. Johnson, T. Pollard, S. Berkowitz, N. Greenbaum, M. Lungren, C. Deng, R. Mark, and S. Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042, 2019b. URL http://arxiv.org/abs/1901.07042.

A. Johnson, T. Pollard, S. Berkowitz, N. Greenbaum, M. Lungren, C. Deng, R. Mark, and S. Horng. Mimic-cxr database (version 2.0.0), 2019c. URL https://doi.org/10.13026/C2JT1Q.

A. Johnson, T. Pollard, R. Mark, S. Berkowitz, and S. Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. *Scientific Data*, 6(1):317, 2019d. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

Amar Kumar, Anita Kriz, Mohammad Havaei, and Tal Arbel. Prism: High-resolution & precise counterfactual medical image generation using language-guided stable diffusion. *arXiv preprint arXiv:2503.00196*, 2025.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023.

Sajid Nazir, Diane M Dickson, and Muhammad Usman Akram. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Computers in Biology and Medicine*, 156:106668, 2023.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, 2022.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

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.

TheSmartTechnologyLab. vit-chest-xray:. https://huggingface.co/codewithdark/vit-chest-xray, 2025.

Baohua Yan, Qingyuan Liu, Zhaobin Mo, Kangrui Ruan, and Xuan Di. Balanced latent space of diffusion models for counterfactual generation. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Appendix A. MIMIC Dataset Refinement

The limited size of our target subset initially hindered the Stable Diffusion Model's ability to accurately learn the data distribution. To rigorously implement the data filtration outlined in 2.1, we established a semi-automated pipeline to identify and exclude unsuitable images. Given the scale of MIMIC, manual review of the entire dataset was infeasible. Instead, we manually annotated a robust subset of over 80,000 images to serve as ground truth, categorizing them into the five target classes: *Frontal*, *Lateral*, *Rotated*, *Under-/Over-Sharpened*, and *Out-of-Domain*.

Leveraging these annotations, we fine-tuned a Vision Transformer (ViT) classifier —already pretrained on chest X-rays for abnormality detection (TheSmartTechnologyLab, 2025)—to distinguish between valid frontal views from the rest of categories. This domain-adapted classifier was then applied to the full MIMIC archive (comprising over 200,000 images) to predict image categories. To ensure high precision, we visually verified samples predicted as any class other than *Lateral*. Table 4 details the exclusion statistics for each category, and Figure 6 provides visual examples of the artifacts targeted by this pipeline.

Table 4: *Number of samples eliminated as non-frontal*

Class	Under-/Over-Sharpened	Lateral	Rotated	Out Of Domain
# eliminated samples	2560	641	236	641

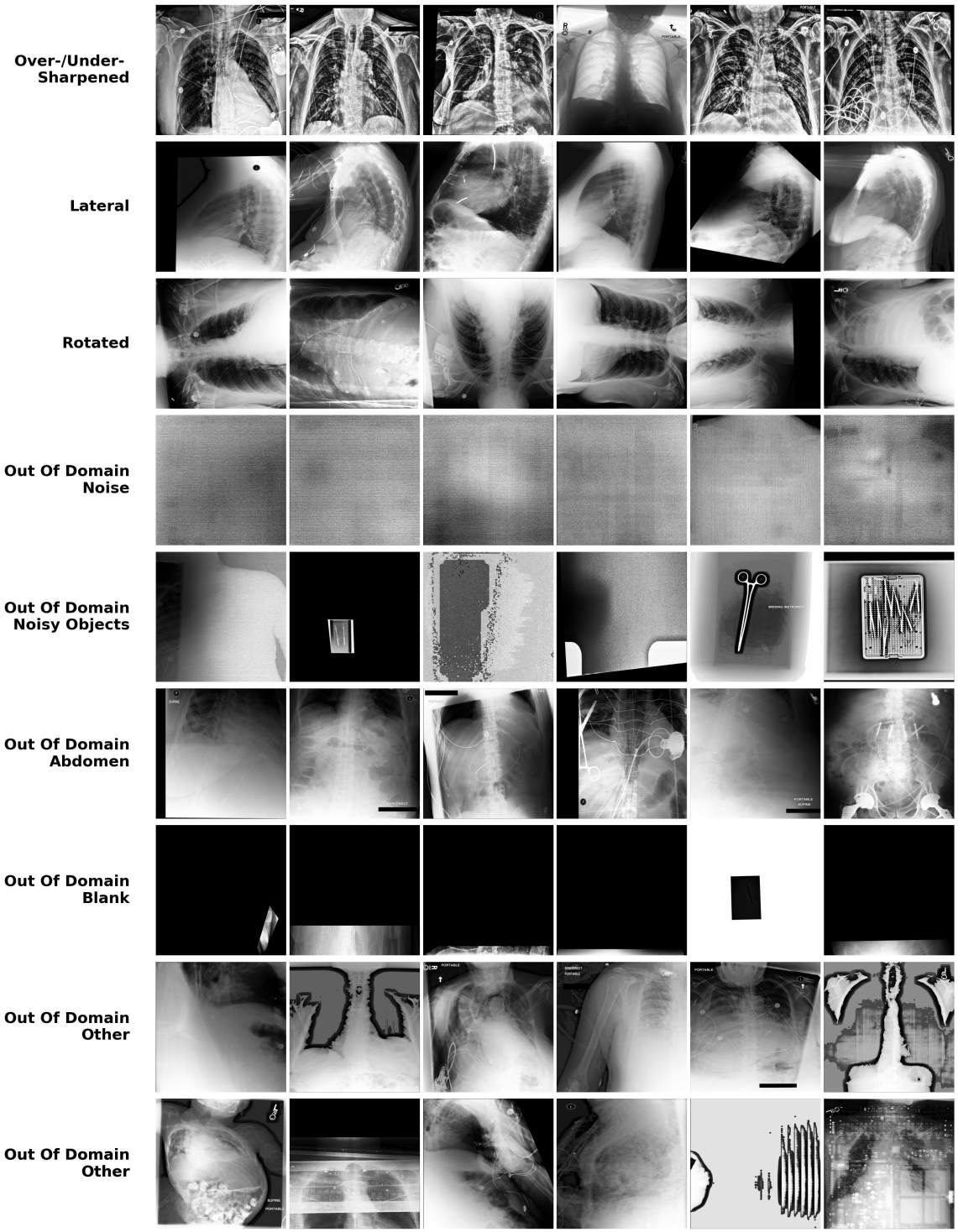

Figure 6: Representative samples of eliminated images.

Appendix B. Excluded Metrics in Criteria

- **Directional Similarity.** Some prior works (Prabhu et al., 2023; Kumar et al., 2025) employ CLIP Directional Similarity (Radford et al., 2021) to evaluate counterfactual quality. This metric compares the alignment between the change in image embeddings (ΔI) and the change in text embeddings (ΔT) according to:

$$\phi(x, x', c, c') = 1 - \frac{(E_I(x) - E_I(x')) \cdot (E_T(c) - E_T(c'))}{\|E_I(x) - E_I(x')\| \|E_T(c) - E_T(c')\|} \tag{4}$$

  where E_I, E_T, x, x', c, c' denote image encoder, text encoder, factual image, counterfactual image, textual prompt for *Cardiomegaly*, and textual prompt for *No Finding*, respectively. This metric relies on a meaningful directional shift in the text embedding space. However, in our specific use case, the factual and counterfactual prompts differ by only a single word ("with" vs. "without" cardiomegaly). Consequently, the difference in prompt embeddings is negligible, providing a weak reference for assessing directional change. In our experiments, this metric failed to distinguish between high-quality and poor-quality counterfactuals.

- **Structural**: The Structural Similarity Index (SSIM) (Hore and Ziou, 2010) evaluates image quality by quantifying degradation in luminance, contrast, and structure. However, traditional pixel-level metrics like SSIM and Peak Signal-to-Noise Ratio (PSNR) are known to correlate poorly with human perceptual judgments of semantic realism (Zhang et al., 2018), rendering them insufficient for evaluating complex generative changes.

- **Density-based**: Metrics such as Inception Score (IS) and Fréchet Inception Distance (FID) are standard benchmarks for evaluating the fidelity and diversity of generative models. Inception Score (IS), proposed by (Salimans et al., 2016), is defined as $\text{IS} = \exp(\mathbb{E}_{\mathbf{x}}[D_{KL}(p(y|\mathbf{x}) \parallel p(y))])$. A high IS indicates that the model generates images with high class-conditional confidence (low entropy for $p(y|\mathbf{x})$) while maintaining diversity across classes (high entropy for marginal $p(y)$). Fréchet Inception Distance (FID) measures the Wasserstein-2 distance between the feature distributions of real (μ_r, Σ_r) and generated (μ_g, Σ_g) images (Heusel et al., 2017):

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}\right)$$

  Lower FID values indicate that the generated distribution is closer to the real data manifold.

  While previous works have employed FID (Jain et al., 2025) or IS (Salimans et al., 2016) to assess general realism, these distribution-based metrics are unsuitable for our objective. They evaluate the global quality of a dataset, whereas counterfactual evaluation requires a one-to-one assessment of instance-level modifications.

Appendix C. Iterative Inference Algorithm

Algorithm 1: Iterative Search Algorithm for Valid Counterfactual Generation

Input: Image F, Source Prompt f_{pr}, Target Prompt cf_{pr}
Data: Finetuned weights $\Theta \in \{\text{Best}(L1), \text{Best}(L2)\}$
Output: CF, the optimal counterfactual image

$CF_{best} \leftarrow$ None
$max\_score \leftarrow 0$
// Define hyperparameter search spaces
$\mathcal{H}_{null} \leftarrow \{(10, 8), (10^{-5}, 5 \cdot 10^{-4})\}$
// Tuples of (iterations, max error) for null-text opt.
$\mathcal{S}_{denois} \leftarrow \{1.0, 0.9, 0.8\}$
$\Omega_{guide} \leftarrow \{7, 8, 9, 10\}$
for $(N_{null}, \epsilon) \in \mathcal{H}_{null}$ **do**
  for $\theta \in \Theta$ **do**
    for $\delta \in \mathcal{S}_{denois}$ **do**
      for $\omega \in \Omega_{guide}$ **do**
        // Map F to latent code z via DDIM noise scheduling and optimize the embedding of null-text $\varnothing$
        $z, \varnothing \leftarrow \text{NullInversion}(F, f_{pr}, \omega, \delta, N_{null}, \epsilon)$
        // Generate counterfactual CF using DDIM Sampling
        $CF \leftarrow \text{DDIMSample}(z, \text{cond} = cf_{pr}, \text{uncond} = \varnothing, \text{scale} = \omega, \text{start} = \delta)$
        // Evaluate criteria (0 to 5)
        $score \leftarrow \text{CF\_Score}(F, CF)$
        if $score = 5$ **then**
          return CF
        end
        else if $score > max\_score$ **then**
          $CF_{best} \leftarrow CF$
        end
      end
    end
  end
end
return CF_{best}

Appendix D. Limiting Difference Maps to Chest Bounding Boxes

A persistent challenge in diffusion-based image generation is the precise reconstruction of fine textual elements. In our experiments, while the generated images reconstruct text markers, the results are often imperfect. Since these markers and annotations are frequently the brightest components of a radiograph, even minor reconstruction errors can manifest as high-intensity artifacts in the difference map's margins—distractions that are irrelevant to the pathological assessment.

Given that the primary objective of this study is to generate counterfactuals for anatomical chest structures rather than metadata, we employ chest bounding boxes from the REFLACX dataset (Bigolin Lanfredi et al., 2022, 2021) to mask the difference maps. This step effectively eliminates artifacts arising from imperfect text reconstruction, allowing our evaluation to focus exclusively on the anatomical validity of the counterfactual.

Figure 7 presents three representative samples, displaying the factual image, the generated counterfactual, the raw difference map, and the difference map constrained to the chest bounding box. As illustrated, this masking strategy successfully removes peripheral noise caused by textual artifacts without compromising the relevant reconstruction of the chest anatomy.

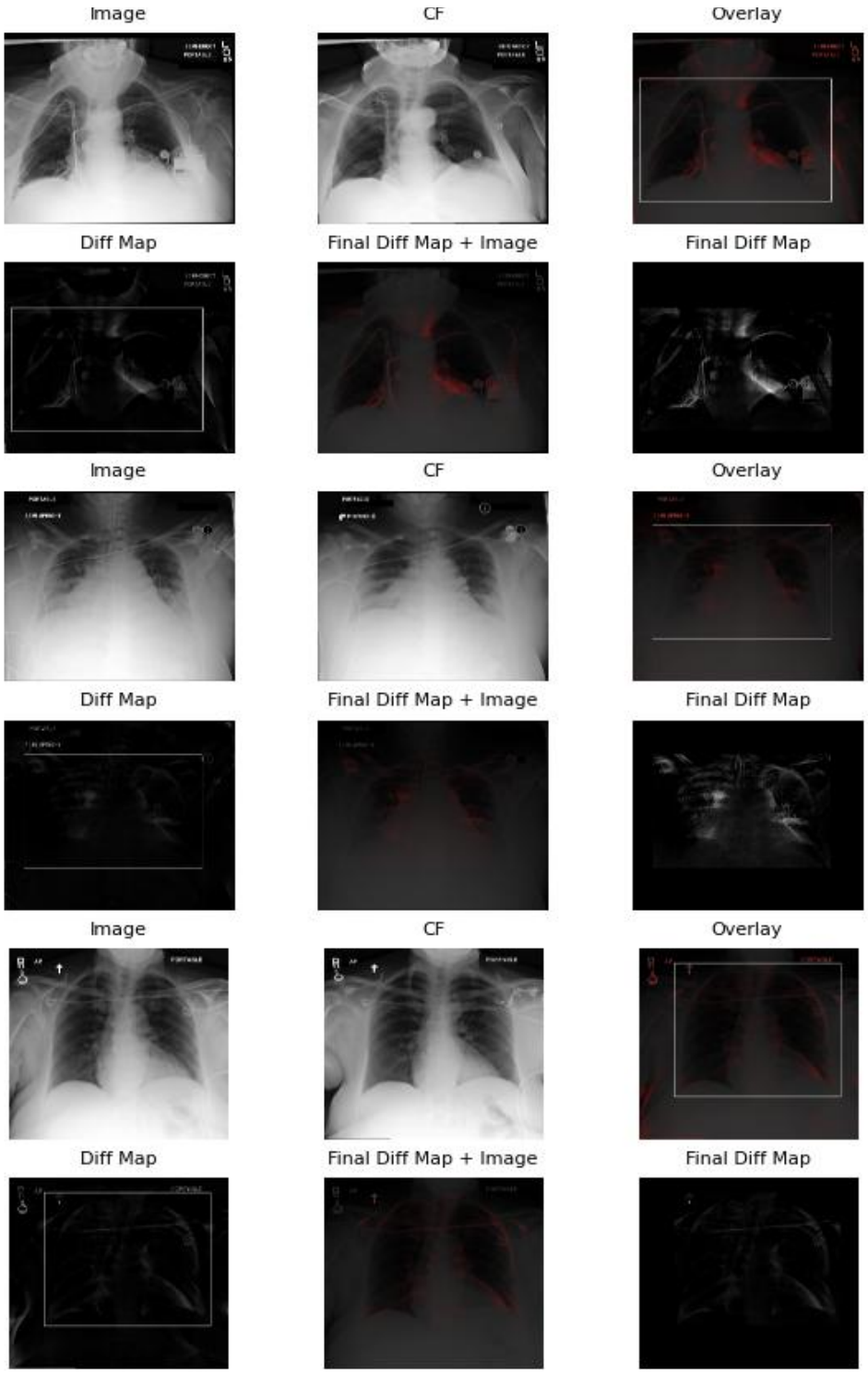

Figure 7: Using chest box to eliminate text

Appendix E. Ablation Study: Metric-Guided Search

We analyze the impact of our metric-guided inference strategy (Algorithm 1) by quantifying the contribution of each hyperparameter to the generation of valid counterfactuals. We define a "success" as a generated image that satisfies all five conditions of our strict composite criteria (Eq. 3). The following results report the proportion of successful samples contributed by specific hyperparameter configurations. Note that the total number of images (**#Im**) varies across experiments, as the search is hierarchical: difficult samples that fail under default settings are passed to subsequent stages of the search algorithm.

Denoising Strength (δ) and Guidance Scale (ω)

These two parameters are the primary tools for balancing the trade-off between faithfulness and effectiveness. Table 5 details the success rate distribution across different combinations of denoising strength (δ) and guidance scale (ω). We observe that higher denoising strengths (e.g., $\delta = 1.0$) generally yield the highest success rates (accounting for 40% of successful cases). This confirms our hypothesis that initiating generation from a near-isotropic noise state (balanced space) provides the necessary flexibility to remove the pathology effectively. Conversely, lower denoising strengths often result in higher faithfulness but fail to achieve the effectiveness threshold required to pass Eq. 3.

Table 5: *Distribution of successful counterfactuals across denoising strength (δ) and guidance scale (ω).*

δ/ω	7	8	9	10	$\sum_\omega$
0.8	0.25	0.04	0.03	0.02	0.34
0.9	0.22	0.02	0.01	0.1	0.25
1.0	0.37	0.2	0.01	0.0	0.40
$\sum_\delta$	0.84	0.08	0.05	0.03	**#Im**=366

Table 6: *Success rate using different model checkpoints (θ) and null-text embedding strategies (ϕ).*

Param	θ_{L2}	θ_{L1}	$\phi_{default}$	$\phi_{underfit}$
Success	0.83	0.80	0.90	0.50
#Im	387	82	387	18

Model Checkpoint (θ) and Null-Text Optimization (ϕ)

When standard inference parameters fail, our strategy iterates through alternative model states. We prioritized the checkpoint with the best validation L2 loss (θ_{L2}), followed by the best L1 loss (θ_{L1}). As shown in Table 6, θ_{L2} successfully handled the majority of samples (83%). However, for the subset of cases that failed with θ_{L2}, switching to θ_{L1} successfully rescued 80% of the remaining failures, highlighting the value of ensemble-like search strategies.

Furthermore, we found that rigid null-text optimization can sometimes hinder editability. While the default optimization ($\phi_{default}$) works for most images, employing "underfitted" null-text embeddings ($\phi_{underfit}$) proved crucial for challenging cases. Although $\phi_{underfit}$ was applied to a small subset of highly resistant images ($N = 18$), it achieved a 50% success rate where all other configurations had failed (Table 6).