Semi-Synthetic Localization Datasets for Radiological Findings on Chest X-Rays

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2026

Abstract

While large datasets for chest X-ray (CXR) finding classification are widely available, datasets for finding localization are scarce. Curating these localization datasets is costly and time-intensive, requiring manual annotation by medical experts, which often results in them being small and limited in scope. To overcome this, we introduce *SemiSynCXR*, a framework designed to automatically generate semi-synthetic localization datasets. *SemiSynCXR* operates by inpainting specific radiological findings into real healthy CXRs at anatomically plausible locations, which allows to output both the edited image and the ground-truth bounding box for each finding. *SemiSynCXR*-generated CXRs effectively augment existing localization datasets, mitigating data scarcity and improving generalization. Comprehensive quantitative and qualitative evaluations confirm that the generated findings are realistic and accurately localized, establishing *SemiSynCXR* as a practical solution for the generation of CXR finding localization datasets. Code will be released upon acceptance.
Keywords: Semi-synthetic CXRs, Inpainting, Diffusion models

1. Introduction

A chest X-ray (CXR) is a fast, effective, and rather inexpensive aid for diagnosing and monitoring thoracic conditions (Mayo Clinic, 2024). Due to their prevalence in clinical practice, CXRs are a popular image source for medical deep learning applications. Large datasets for radiological finding classification are common, often supported by automated labeling, whereas curating datasets for localization and segmentation requires costly, time-intensive manual annotation by medical experts. As a result, these datasets are not only scarce, but also tend to be small and limited in scope.

To date, only eight publicly available CXR datasets provide bounding boxes for more than one radiological finding (Wang et al., 2017; Boecking et al., 2022a; Nguyen et al., 2022; Pham et al., 2023; Bigolin Lanfredi et al., 2022; de Castro et al., 2025; Fan et al., 2024; Liu et al., 2020), and two provide segmentation masks (Liu et al., 2020; Saporta et al., 2022). Of these eight localization datasets, only two contain over 10 000 scans[1]. Moreover, these datasets suffer from class imbalance. Even in the largest dataset, CXR-AL14 (with about 165 000 images), some common thoracic findings have low representation. For instance, atelectasis is found in only 0.2% of its images.

To overcome data scarcity in medical imaging, data synthesis offers a compelling solution (Voetman et al., 2023; Khosravi et al., 2024; Ktena et al., 2024). Driven by advances in high-quality image generation models like latent diffusion models (Rombach et al., 2022), existing

1. For comparison, PASCAL VOC (20 classes, 22 591 images) is considered one of the smallest widely used benchmarks for object detection in computer vision.

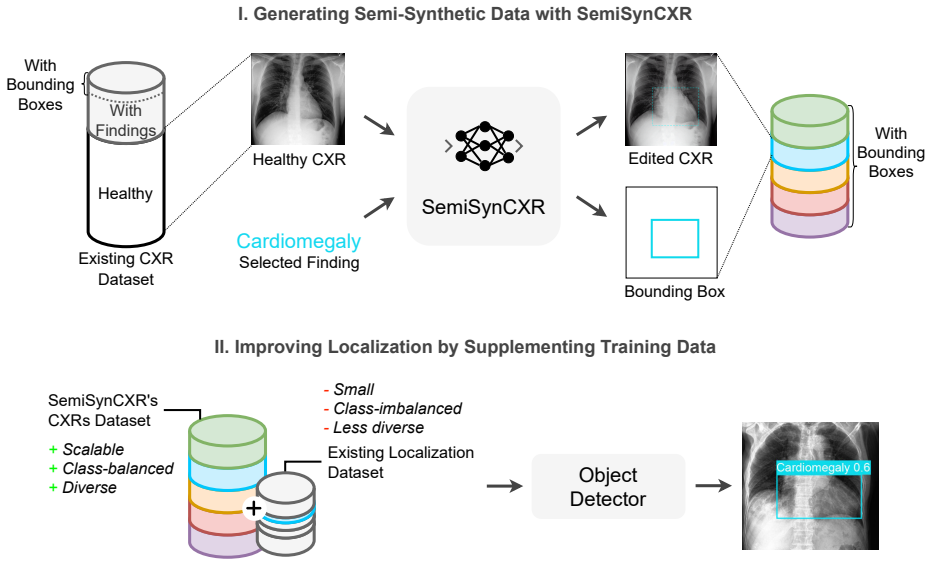


Figure 1: We propose *SemiSynCXR*, a framework that generates semi-synthetic CXRs by inpainting radiological findings into healthy images at plausible anatomical locations. *SemiSynCXR* automatically produces both the edited image and a precise bounding box, directly addressing data scarcity and class imbalance of existing localization datasets.

studies have demonstrated the potential of generating synthetic CXRs (Bluethgen et al., 2025; Weber et al., 2023; Han et al., 2024; Huang et al., 2024). However, they largely focus on creating unlabeled images, image-text pairs, or images with only global classification labels, leaving the generation of much-needed localization datasets still unaddressed.

Moreover, obtaining the finding's exact positional information is a significant challenge when generating localization datasets using fully-synthetic image generation. Since the generative model implicitly determines the finding placement, its precise location often remains unknown. We propose semi-synthetic image generation as an alternative, as illustrated in Figure 1. By using automated image editing to inpaint findings into healthy CXRs, we can explicitly define the finding's location using conditioning masks. Thereby, the process inherently guarantees ground-truth bounding box annotations for every generated image, directly overcoming a core limitation of fully-synthetic approaches.

Our contributions are summarized as follows:

- We introduce *SemiSynCXR*, a framework for automatically generating semi-synthetic CXRs with radiological findings. The framework's core strength is its ability to provide the generated image with intrinsically matching, precise bounding boxes at scale.
- For this, we develop an automated mask generation method for *SemiSynCXR* that places findings at anatomically plausible locations based on real-world spatial distributions. *SemiSynCXR* further leverages existing diffusion models for inpainting findings into healthy CXR images, which alleviates the need for training of new models.

- Using *SemiSynCXR*, we create a semi-synthetic dataset for CXR finding localization. Augmenting real training data with our generated samples significantly improves object detection performance and robustness, effectively mitigating data scarcity.
- Extensive quantitative and qualitative evaluations confirm that the generated findings are realistically placed and that the resulting CXRs resemble real images.

2. Related Work

2.1. Chest Radiological Findings Localization Datasets

The largest publicly available datasets for chest radiological findings are CXR-AL14 and VinDr-CXR. CXR-AL14 contains 165 988 posterior-anterior (PA) CXRs with bounding box annotations for 14 common abnormalities. These annotations were created through a "human-in-the-loop" process, where expert radiologists reviewed and corrected initial model annotations. VinDr-CXR is a smaller dataset of 18 000 PA CXRs, with manual local annotations for 22 critical findings. Despite their size, these datasets present significant limitations that could hinder model performance and robustness. Both originate from a very small number of institutions (one hospital in China for CXR-AL14, and two in Vietnam for VinDr-CXR), leading to low diversity in patient demographics, imaging equipment, and clinical workflows. This raises concerns about generalization to new clinical settings. Additionally, these datasets suffer from severe class imbalance, meaning some radiological findings are less frequently represented in the dataset, making their reliable detection challenging for models. More existing datasets are detailed in Appendix B.

2.2. Text- and Mask-Conditioned Diffusion Models for Editing

Using masks alongside text is an effective technique for guiding image editing, enabling precise, controlled changes without inadvertently altering adjacent areas (Huang et al., 2025). Hence, this approach is highly appealing for medical imaging applications, with several studies demonstrating its potential (Rouzrokh et al., 2023; Jin et al., 2024; Pérez-García et al., 2024; Hashmi et al., 2024; Chu et al., 2025). Some of these approaches require specific training or fine-tuning of the diffusion model, such as those by Rouzrokh et al. (2023) for brain tumor editing or Jin et al. (2024) for background alteration. Additionally, they often rely heavily on accurate, user-defined segmentation masks, which limits their scalability. Other approaches, conversely, eliminate the need for additional supervision by leveraging the iterative nature of the diffusion process itself. These frequently use multi-stage or multi-masking strategies, as seen in RadEdit (Pérez-García et al., 2024), XReal (Hashmi et al., 2024), and ChestX-rays_Mpe (Chu et al., 2025).

Reducing artifacts and preserving anatomical accuracy remain major challenges in CXR editing. Most existing methods either depend on user-defined masks, which inherently hinders large-scale applicability, or use broad anatomical regions, which limits the precision necessary to define ground-truth bounding boxes for specific findings. Additionally, comprehensive reporting on editing quality at the radiological finding level is often absent. Crucially, and to our knowledge, no prior studies have reported analyses on supplementing training data with (semi-)synthetic images specifically for chest radiological finding localization. This highlights an unaddressed gap that our study aims to fill.

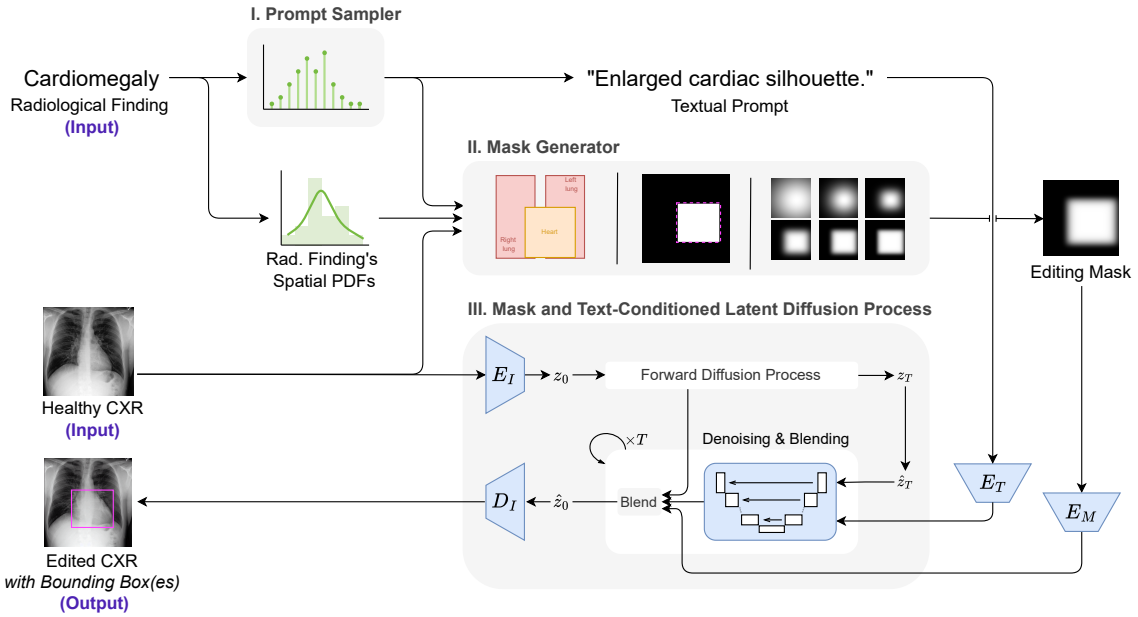


Figure 2: Overview of *SemiSynCXR*. Based on a real, healthy CXR and a specific finding, our method samples a textual prompt and generates a plausible spatial mask to guide the editing process. A latent diffusion model then uses the chest X-ray, prompt, and mask to inpaint the finding. The resulting semi-synthetic CXR and bounding boxes (derived from the conditioning mask) are used as targets for localization tasks.

3. Methodology

Our *SemiSynCXR* framework generates semi-synthetic CXRs by inpainting specific radiological findings into healthy images. The process, illustrated in Figure 2, begins with a real, healthy CXR and a target radiological finding. We then sample a textual prompt and generate a plausible spatial mask to guide the placement, drawing on real-world spatial distributions. Conditioned on these, a latent diffusion model (either RadEdit (Pérez-García et al., 2024) or RoentGen (Bluethgen et al., 2025)) then inpaints the finding into the healthy image. The output is a new semi-synthetic CXR and its ground-truth bounding boxes, directly derived from the conditioning mask. Seven findings are cureently supported: Atelectasis, Cardiomegaly, Consolidation, Edema, Lung Opacity, Pleural Effusion, and Pneumothorax.

3.1. Datasets

SemiSynCXR itself does not require any training data; however, we leverage the following datasets to source healthy images, create textual prompts, and guide mask generation:

- MIMIC-CXR-JPG (Johnson et al., 2019b, 2024; Goldberger et al., 2000) contains 377 110 CXRs derived from MIMIC-CXR (Johnson et al., 2019a). It provides healthy CXRs (Section 3.2), and radiology reports to create textual prompts (Section 3.3).

- MS-CXR (Boecking et al., 2022a,b; Goldberger et al., 2000), constructed from MIMIC-CXR, consists of 1 162 image-sentence pairs with bounding boxes. We use it as source for medical texts to create textual prompts (Section 3.3), and for estimating the expected spatial distribution of different findings in the lung (Section 3.4).
- CheXmask (Gaggion et al., 2024b,a; Goldberger et al., 2000) comprises 657 566 anatomical segmentation masks, generated by a HybridGNet (Gaggion et al., 2021), for CXR datasets including MIMIC-CXR-JPG and VinDr-CXR. It provides the lung and heart segmentation masks for the mask generation and CXR editing (Sections 3.4 and 3.5).
- Chest ImaGenome (Wu et al., 2021; Goldberger et al., 2000), an anatomy-centered scene graph dataset from MIMIC-CXR, includes 29 chest anatomical locations and a manually annotated subset for 500 unique patients (gold standard dataset). We use this subset to source anatomical reference locations for mask generation (Section 3.4).

3.2. Sourcing Real, Healthy Chest X-rays

Instead of fully synthetic generation (Bluethgen et al., 2025; Lee et al., 2023; Han et al., 2024; Huang et al., 2024; Hashmi et al., 2024; Chu et al., 2025), we use a semi-synthetic approach: inpainting findings into real, healthy CXRs. Thus, *SemiSynCXR* preserves authentic image characteristics in unaffected regions while enabling precise control over finding placement, which inherently guarantees knowing the ground-truth bounding boxes.

Healthy chest X-rays are sampled from MIMIC-CXR-JPG using the following criteria: (i) the image must be a `posterior-anterior (PA)` view with `erect` patient posture; (ii) it must not be included in the MS-CXR dataset; (iii) it must be labeled as `No Finding` and negative for `Support Devices`, according to the CheXpert (Irvin et al., 2019) annotations; and (iv) it must be classified as negative for all relevant radiological findings by the XVR DenseNet-121 model (Cohen et al., 2022). Based on these criteria, we identified 24 555 eligible CXRs for sampling.

3.3. Sampling Textual Prompts

To steer the diffusion model toward the desired radiological finding, we condition the model on a textual prompt. For each finding, we curate a set of phrases derived from medical texts (Appendix C). We then sample a prompt from the corresponding set during editing.

Phrases are sourced from MIMIC-CXR radiology reports (exclusively associated with a single CheXpert finding) and MS-CXR textual descriptions (also associated with a single finding). From MIMIC-CXR radiology reports, we extract the *Findings*, *Impression*, or last section following Johnson et al. (2022, 2018). Only phrases observed more than once are retained and simplified using the `gpt-oss-20b` language model (OpenAI, 2025), removing mentions of size and severity, which are instead controlled by the editing mask. After manual review of simplified phrases, we adjust the sampling probabilities to ensure balance sampling: phrases from MIMIC-CXR and MS-CXR are equally likely to be selected.

3.4. Generating the Editing Mask

To outline the location for inpainting, we sample a mask conditioned on the target finding, the sampled prompt (Section 3.3), and the anatomical structures of the sampled healthy

CXR (Section 3.2). More precisely, for lung-associated findings, we model the spatial distribution in relation to the lungs, while for cardiomegaly, we consider the cardiothoracic ratio (CTR – i.e., the ratio between the heart's and thorax's widths). Modeling the spatial distributions relative to these anatomical structures, rather than in pixel space, eliminates the need for image registration and exploits the fact that the heart and lungs are easily identifiable in CXRs (Gómez et al., 2020). Within *SemiSynCXR*, the lung and heart structures are identified using CheXmask-derived bounding boxes.

Spatial Distribution Estimation We estimate the (relative) center and size distributions of the findings' bounding boxes based on data provided in the MS-CXR dataset.

For each lung-related finding, we model the center and size either as multivariate (2D) log-normal distributions or as two independent univariate (1D) distributions, based on a Spearman's correlation test for independence. The 1D distributions are selected according to the residual sum of squares (RSS) goodness-of-fit criterion[2]. We additionally assume lateral symmetry for the left and right lungs and thus mirror the source data in the contralateral lung before computing the distributions. Although assumed to simplify the modeling process, note that lateral symmetry does not fully reflect reality: A two-sample Kolmogorov–Smirnov test revealed enough statistical evidence against symmetry for two (edema and pleural effusion) of the six considered lung-related findings.

For cardiomegaly, we estimate the distribution of the CTR. Cardiomegaly is conventionally defined as present if the CTR is greater than 0.5 on a PA CXR. Hence, anterior-posterior (AP) images are excluded during estimation as hearts appear enlarged in this projection.

Mask Sampling To generate the mask, we sample the relative center and size from the estimated distributions of the given finding and convert these into image coordinates using the lung/heart masks.

During sampling, we must consider several constraints to assure findings are plausibly placed. First, we constrain the sampled sizes so the final masks remain within the lung boundaries. Additionally, textual prompts may contain anatomy-specific references, such as "bibasilar atelectasis", which indicates the finding is located at the base of both lungs. To ensure alignment with the prompt, we estimate the bounding boxes of the lung's anatomical substructures (like the lung bases) relative to the full lung itself using the Chest ImaGenome's gold standard dataset. We then use these bounding boxes so the center of the sampled mask lies on the lung area indicated by the prompt. All constraints are enforced via probability distribution truncation, using inverse transform sampling for the 1D distributions and an efficient sampling method for truncated multivariate normal distributions (Brunzema and Kim, 2021; Botev, 2017) for the 2D distributions.

For *pleural effusion* and *cardiomegaly*, we developed specialized mask sampling methods. For pleural effusion, we use the full width of the lung (left or right) and sample the center y-coordinate and height of the mask, guaranteeing full coverage of the bottom of the lung's bounding box. For cardiomegaly, we sample from the CTR distribution, and compute the mask size from the sampled ratio using the lung and heart masks of the current X-ray. The mask's center is set as that of the heart's bounding box.

2. 1D considered distributions: normal, generalized extreme value, exponential, gamma, Pareto, log-gamma, lognormal, beta, Student's t, and uniform distribution.

Finally, the masks are blurred using generalized Gaussian filters. Blurring is found to help to reduce artifacts as inpainted findings blend more naturally into healthy CXRs.

3.5. Editing: Inpainting the Radiological Finding

We generate the final semi-synthesized CXR by editing the healthy CXR using a latent diffusion model conditioned on the sampled healthy image, sampled textual prompt, and generated editing mask. We employed either RoentGen (Bluethgen et al., 2025) or RadEdit (Pérez-García et al., 2024) as the latent diffusion model, depending on the radiological finding. Note that these models are used pre-trained, requiring no fine-tuning or training on bounding box data. While both models natively support conditioning on textual prompts, we extend RoentGen to additionally support mask conditioning.

In both models, we use the blending method (Avrahami et al., 2023) for mask conditioning, which leverages the iterative nature of the reverse diffusion process. We consider three different variations of blending: (i) blendling between the latents from the forward and reverse diffusion process before denoising (blending before; used with RoentGen only), (ii) blending between the latents from the forward and reverse diffusion process after denoising (blending after; used with RoentGen only), and restricting the classifier-free guidance (CFG) to the editing area (CFG masking; used with both RoentGen and RadEdit).

3.6. Finding the Optimal Configuration

We explored different design configurations of our framework to identify the best setting for each of the seven radiological findings under analysis. Specifically, we varied the diffusion model (RoentGen or RadEdit), mask blurring parameters, number of steps with mask conditioning, and hyperparameters of the diffusion inference process (Appendix E).

For each design configuration, we generated 35 semi-synthetic chest X-rays. We then used these samples to select the optimal *SemiSynCXR* configuration per finding based on four metrics: (i) Area Under the ROC Curve (AUROC), using a DenseNet-121 trained on XRV-all (Cohen et al., 2022); (ii) Fréchet Inception Distance (FID), obtained with InceptionV3 (layer 2048) (Szegedy et al., 2016); (iii) CLIP Score, derived from the XRayCLIP model (Chen et al., 2024); (iv) mean Average Precision (mAP@0.10:0.70), using an ensemble of YOLOv4 models (Wang et al., 2021; Vinokurov, 2021) trained on VinDr-CXR (Nguyen et al., 2022). We aggregated these metrics into a single selection score by computing the area of the polygon formed by their radar plot representation.

4. Experiments and Results

Using our *SemiSynCXR* framework, we generate a semi-synthetic localization dataset to supplement real training data for radiological finding localization in CXRs. By training object detectors on this combined data, we first assess the impact of our generated CXRs as supplementary training data (Section 4.1). We then evaluate the quality of our generated images (Section 4.2).

Table 1: Effect of *SemiSynCXR*-generated CXRs as supplementary training data for finding localization. YOLO11n and YOLOv8n detectors are trained on VinDr-CXR supplemented with varying amounts of our semi-synthetic images. We report mAP (IoU 0.1-0.7) and AP@30 on VinDr-CXR test set (in-distribution) and MS-CXR (out-of-distribution). Supplementing with our data increases mAP by up to 11% (VinDr-CXR) and 21% (MS-CXR), confirming that *SemiSynCXR* helps to effectively address data scarcity and improve model generalization.

Model	**Training Data**		**VinDr-CXR**		**MS-CXR**	
	Real	Synth	[mAP ↑]	[AP@30 ↑]	[mAP ↑]	[AP@30 ↑]
YOLO11n	15k	–	21.9	26.5	9.5	13.5
	15k	7k	22.9 (+ 5%)	28.7 (+ 8%)	10.3 (+ 8%)	14.5 (+ 7%)
	15k	17.5k	22.5 (+ 3%)	27.5 (+ 4%)	**10.8** (+13%)	**15.2** (+13%)
	15k	35k	**24.2** (+11%)	**29.8** (+12%)	9.6 (+ 0%)	13.8 (+ 2%)
YOLOv8n	15k	–	21.9	26.4	9.4	13.2
	15k	7k	22.9 (+ 5%)	27.8 (+ 5%)	9.6 (+ 3%)	14.1 (+ 7%)
	15k	17.5k	23.6 (+ 8%)	28.9 (+ 9%)	**11.3** (+21%)	**15.6** (+18%)
	15k	35k	**23.7** (+ 8%)	**29.1** (+10%)	10.7 (+14%)	14.6 (+11%)

4.1. SemiSynCXR for Supplementary Training Data Generation

To study how *SemiSynCXR*-generated images can be used to improve object detection training, we created a finding localization dataset of 35 000 samples and use it as an extension of the VinDr-CXR dataset (15 000 real images). Table 1 shows the results of training YOLO11n and YOLOv8n object detectors (Jocher and Qiu, 2024; Jocher et al., 2023) on VinDr-CXR alone or supplemented with subsets of our generated dataset. We report the mAP, averaged over IoU thresholds 0.1 to 0.7, and AP@30 on the VinDr-CXR test set (in-distribution) and MS-CXR dataset (out-of-distribution). Supplementing with our data improves the overall performance by up to 11% in mAP on VinDr-CXR and up to 21% in mAP on MS-CXR compared to using only on real samples. For VinDr-CXR, using all 35k generated samples leads to the best performance. For MS-CXR, however, the optimal performance is achieved when supplementing with 17.5k samples, suggesting adding too many semi-synthetic samples might sometimes introduce bias. Overall, these results confirm that *SemiSynCXR* helps to effectively address data scarcity and enhance the generalization capability of object detection models.

4.2. Generation Quality

We quantitatively assess the factual correctness of our generated images using classification models, following common practice, and localization models. Finding detection (whether desired finding is successfully inpainted into healthy image) is measure by the AUROC from a DenseNet-121 classifier trained on XRV-all (Cohen et al., 2022). Accurate placement (localization) is measured using the mAP from an ensemble of YOLOv4 models (Wang et al., 2021; Vinokurov, 2021) trained on VinDr-CXR (Nguyen et al., 2022), averaging over IoU thresholds 0.1 to 0.7. Edema is excluded from the mAP metric due to its underrepresenta-

Table 2: Factual correctness of *SemiSynCXR*-generated CXRs. We benchmark detectability (AUROC using a DenseNet121) and localization accuracy (mAP@0.1:0.7 using an ensemble of YOLOv4s trained on VinDr-CXR) against real and fully synthesized CXRs. Our approach yields findings detectable at levels comparable to, or superior to, fully synthesized CXRs. High AUROC scores for cardiomegaly and pleural effusion suggest these semi-synthetic CXRs closely resemble prototypical clinical cases. Strong mAP scores confirm *SemiSynCXR* successfully also produces realistic, well-localized findings.

Model	**Radiological Finding**							
	Atel.	Cmgl.	Cnls.	Edema	Opac.	P. Eff.	Pneum.	Avg.
Classification [AUROC ↑]								
XVR's benchmark (Cohen et al., 2022) (real)	0.88	0.88	0.91	0.92	0.86	0.92	0.81	0.88
RoentGen (Bluethgen et al., 2025) (synthetic)	0.76	0.82	0.69	0.85	0.74	0.90	0.61	0.76
CXRL (Han et al., 2024) (synthetic)	0.86	0.88	0.94	0.89	0.70	0.77	0.88	0.81
LLM-CXR (Lee et al., 2023) (synthetic)	0.81	0.78	0.82	0.81	0.83	0.82	0.75	0.80
Chest-Diffusion (Huang et al., 2024) (synthetic)	0.70	0.73	0.63	0.79	0.65	0.85	0.57	0.70
SemiSynCXR (ours)	0.72	0.98	0.73	0.82	0.67	0.97	0.58	0.78
Localization [mAP ↑]								
VinDr-CXR (Nguyen et al., 2022) (real)	6.97	76.04	19.32	–	5.05	63.47	29.68	30.57
SemiSynCXR (ours)	14.64	97.26	52.56	–	36.58	58.34	8.80	44.70

tion in VinDr-CXR, and bounding boxes for cardiomegaly and pleural effusion are rescaled to account for distributions shifts between our editing masks and VinDrCXR's annotations. We benchmark model performance on *SemiSynCXR* samples against their performance on real CXRs and on fully synthesized CXRs generated by other methods.

Results, presented in Table 2, show that our approach produces findings that are detectable by the classifier at levels comparable to, or even better than, fully synthesized CXRs, and strong mAP scores confirm successful localization. At the individual findings level, cardiomegaly and pleural effusion achieve particularly strong performance, suggesting that these findings resemble prototypical clinical cases. We attribute this success to their consistent anatomical placement (at the heart and lung bases, respectively). By contrast, pneumothorax shows comparatively weaker performance, likely because this finding often affects large portions of the lung beyond the localized inpainting region, making realistic generation challenging when constrained to narrow masks. Overall, the generated samples demonstrate high factual correctness for most findings, confirming their suitability as training data; however, performance differences compared to real CXRs indicate room for further improvement.

While high AUROC scores are necessary, they are not sufficient for guaranteeing image realism (e.g., they can result from a model over-exaggerating certain pathological features). This is why our quality evaluation also includes visual alignment with MIMIC-CXR using Fréchet Inception Distance (FID) and visual-text alignment with conditioning prompts using CLIPScore (Appendix A.1). Our approach achieves comparable performance to most state-of-the-art methods while uniquely providing ground-truth bounding boxes. Finally, a

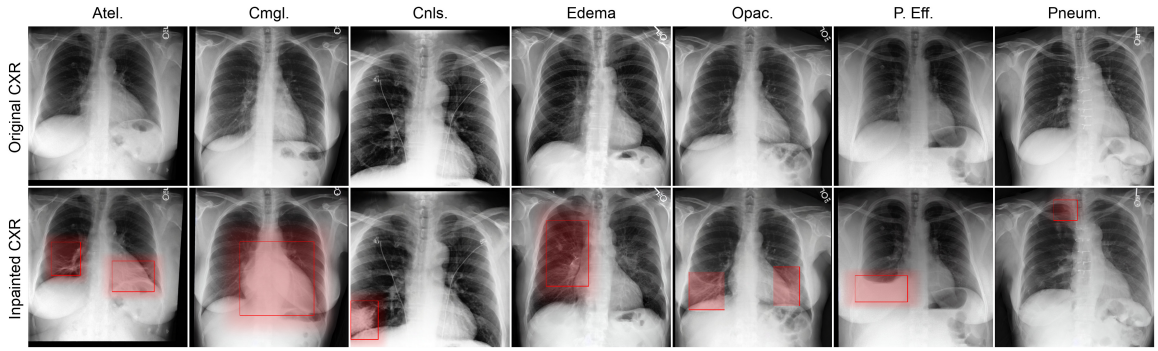


Figure 3: Examples of *SemiSynCXR*-generated CXRs. We show the real, healthy chest X-rays (top) and edited versions (bottom). The red outlines correspond to the conditioning masks alongside their non-blurred version, which serve as training targets (bounding boxes) for localization models.

qualitative study by three medical experts on 140 randomly selected CXRs (70 generated, 70 real) found that, on average, 36% of generated images were judged as real (compared to 71% of real images judged as real), and the intended finding was correctly recognized in 54% of generated cases (vs. 28% in real images) (Appendix A.2). Examples of the generated CXRs are shown in Figure 3.

5. Discussion and Conclusions

We introduce *SemiSynCXR*, a framework for automatically generating localization datasets for chest radiological findings. Our framework's core strength is its ability to provide the generated image with its intrinsically matching, precise bounding boxes at scale. Extensive evaluations confirm the quality and realism of the edited images at levels comparable to fully synthetic data, as well as their utility as a training data augmentation source. Our findings suggest that *SemiSynCXR* provides a practical and effective solution to address data scarcity in medical imaging.

Nonetheless, certain limitations remain. First, editing quality is constrained by both the capabilities of the underlying diffusion models and the effectiveness of our mask generation strategy. Although model performance could potentially be enhanced through fine-tuning, improving mask conditioning is more challenging. Specifically, our method may fail to fully model the changes when findings affect regions beyond their localized bounding boxes, as observed for pneumothorax. Promising future directions for addressing these limitations include iterative mask relaxation and anatomical region bounding boxes. Additionally, our current framework does not explicitly model finding size and severity, both of which could provide more nuanced control for image generation.

References

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended Latent Diffusion. *ACM Trans. Graph.*, 42(4), 2023.

Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F. Auffermann, Jessica Chan, Phuong-Anh T. Duong, Vivek Srikumar, Trafton Drew, Joyce D. Schroeder, and Tolga Tasdizen. REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Sci. Data*, 9(1), 2022.

Christian Bluethgen, Pierre Chambon, Jean-Benoit Delbrouck, Rogier Van Der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P. Langlotz, and Akshay S. Chaudhari. A vision–language foundation model for the generation of realistic chest X-ray images. *Nat. Biomed. Eng.*, 9(4):494–506, 2025.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. In *ECCV 2022*, pages 1–21. Springer Nature Switzerland, 2022a.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. MS-CXR: Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing (version 0.1). *PhysioNet*, 2022b.

Zdravko I. Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *J. R. Stat. Soc. Ser. B Methodol.*, 79(1):125–148, 2017.

Paul Brunzema and Jungtaek Kim. Efficient sampling from the truncated MVN. https://github.com/brunzema/truncated-mvn-sampler, 2021.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis Langlotz. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. Eprint arXiv:2401.12208, 2024.

Huan Chu, Xiaolong Qi, Huiling Wang, and Yi Liang. Multi-label pathology editing of chest X-rays with a Controlled Diffusion Model. *Med. Image Anal.*, 103:103584, 2025. ISSN 1361–8415.

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 *MIDL 2022*, volume 172, pages 231–249. PMLR, 2022. URL https://github.com/mlmed/torchxrayvision.

Daniel Coelho de Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L. Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, José María Salinas-Serrano, Javier Alvarez-Valle, Joaquín Galant-Herrero, and Antonio Pertusa. PadChest-GR: A Bilingual Chest X-Ray Dataset for Grounded Radiology Report Generation. *NEJM AI*, 2(7), 2025.

Weijie Fan, Yi Yang, Jing Qi, Qichuan Zhang, Cuiwei Liao, Li Wen, Shuang Wang, Guangxian Wang, Yu Xia, Qihua Wu, et al. A deep-learning-based framework for identifying and localizing multiple abnormalities and assessing cardiomegaly in chest X-ray. *Nat. Commun.*, 15(1), 2024.

Nicolás Gaggion, Lucas Mansilla, Diego H. Milone, and Enzo Ferrante. Hybrid Graph Convolutional Neural Networks for Landmark-Based Anatomical Segmentation. In *MICCAI 2021*, pages 600–610. Springer International Publishing, 2021.

Nicolás Gaggion, Candelaria Mosquera, Martina Aineseder, Lucas Mansilla, Diego H. Milone, and Enzo Ferrante. CheXmask Database: a large-scale dataset of anatomical segmentation masks for chest x-ray images (version 0.4). *PhysioNet*, 2024a.

Nicolás Gaggion, Candelaria Mosquera, Lucas Mansilla, Julia Mariel Saidman, Martina Aineseder, Diego H. Milone, and Enzo Ferrante. CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. *Sci. Data*, 11(1), 2024b.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Oscar Gómez, Pablo Mesejo, Oscar Ibáñez, Andrea Valsecchi, and Oscar Cordón. Deep architectures for high-resolution multi-organ chest X-ray image segmentation. *Neural Comput. & Applic.*, 32(20):15949–15963, 2020.

Woojung Han, Chanyoung Kim, Dayun Ju, Yumin Shim, and Seong Jae Hwang. Advancing Text-Driven Chest X-Ray Generation with Policy-Based Reinforcement Learning. In *MICCAI 2024*, pages 56–66. Springer Nature Switzerland, 2024.

Anees Ur Rehman Hashmi, Ibrahim Almakky, Mohammad Areeb Qazi, Santosh Sanjeev, Vijay Ram Papineni, Dwarikanath Mahapatra, and Mohammad Yaqub. XReal: Realistic Anatomy and Pathology-Aware X-ray Generation via Controllable Diffusion Model. Eprint arXiv:2403.09240, 2024.

Peng Huang, Xue Gao, Lihong Huang, Jing Jiao, Xiaokang Li, Yuanyuan Wang, and Yi Guo. Chest-Diffusion: A Light-Weight Text-to-Image Model for Report-to-CXR Generation. In *IEEE Int. Symp. Biomed. Imaging*, pages 1–5, 2024.

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion Model-Based Image Editing: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(6):4409–4437, 2025.

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. In *AAAI'19/IAAI'19/EAAI'19*. AAAI Press, 2019.

Qixuan Jin, Walter Gerych, and Marzyeh Ghassemi. MaskMedPaint: Masked Medical Image Inpainting with Diffusion Models for Mitigation of Spurious Correlations. Eprint arXiv:2411.10686, 2024.

Glenn Jocher and Jing Qiu. Ultralytics YOLO11. https://github.com/ultralytics/ultralytics, 2024.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics, 2023.

Alistair E. W. Johnson, David J. Stone, Leo A. Celi, and Tom J. Pollard. The MIMIC Code Repository: enabling reproducibility in critical care research. *J. Am. Med. Inform. Assoc.*, 25(1):32–39, 2018.

Alistair E. W. 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. *Sci. Data*, 6(1), 2019a.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-Ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. Eprint arXiv:1901.07042, 2019b.

Alistair E. W. Johnson, Tom J. Pollard, Jim Blundell, Brian Gow, erinhong, Nicolas Paris, shu98, JackieMe, Kien Dang, Angus Zhang, Eric Carlson, Qinyu Zhao, etheleon, Andrew Barros, a chahin, Lucas Bulgarelli, Hans0124SG, Peter Szolovits, Sicheng Hao, C. V. Cosgriff, Sravan R., Thomas M. Ward, alexmbennett2, Bibo Hao, cx1111, emilyva, Fernando Andreotti, Ishrar Hussain, James Wiggins, and Josh Gieringer. MIT-LCP/mimic-code: MIMIC Code v2.2.1 (v2.2.1). Zenodo. https://doi.org/10.5281/zenodo.6818823, 2022.

Alistair E. W. Johnson, Matthew P. Lungren, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.1.0). *PhysioNet*, 2024.

Bardia Khosravi, Frank Li, Theo Dapamede, Pouria Rouzrokh, Cooper U. Gamble, Hari M. Trivedi, Cody C. Wyles, Andrew B. Sellergren, Saptarshi Purkayastha, Bradley J. Erickson, and Judy W. Gichoya. Synthetically enhanced: unveiling synthetic data's potential in medical imaging research. *eBioMedicine*, 104, 2024.

Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, Alan Karthikesalingam, and Sven Gowal. Generative models improve fairness of medical classifiers under distribution shifts. *Nat. Med.*, 30(4):1166–1173, 2024.

Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. LLM-CXR: instruction-finetuned LLM for CXR image understanding and generation. Eprint arXiv:2305.11490, 2023.

Jingyu Liu, Jie Lian, and Yizhou Yu. ChestX-Det10: Chest X-ray Dataset on Detection of Thoracic Abnormalities. Eprint arXiv:2006.10550v3, 2020.

Mayo Clinic. Chest X-rays. https://www.mayoclinic.org/tests-procedures/chest-x-rays/about/pac-20393494, 2024.

Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Sci. Data*, 9(1), 2022.

OpenAI. gpt-oss-120b & gpt-oss-20b Model Card. Eprint arXiv:2508.10925, 2025.

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. In *ECCV 2024*, pages 358–376. Springer Nature Switzerland, 2024.

Hieu H. Pham, Ngoc H. Nguyen, Thanh T. Tran, Tuan N. M. Nguyen, and Ha Q. Nguyen. PediCXR: An open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children. *Sci. Data*, 10(1), 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR 2022*, pages 10684–10695, 2022.

Pouria Rouzrokh, Bardia Khosravi, Shahriar Faghani, Mana Moassefi, Sanaz Vahdati, and Bradley J. Erickson. Multitask Brain Tumor Inpainting with Diffusion Models: A Methodological Report. Eprint arXiv:2210.12113, 2023.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven Q. H. Truong, Chanh D. T. Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G. Blankenberg, Andrew Y. Ng,

Matthew P. Lungren, and Pranav Rajpurkar. Benchmarking saliency methods for chest X-ray interpretation. *Nat. Mach. Intell.*, 4(10):867–878, 2022.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *CVPR 2016*, pages 2818–2826, 2016.

Nikita Vinokurov. VinBigData Chest X-ray Abnormalities Detection. https://github.com/Beaver48/kaggle-vinbigdata, 2021.

Roy Voetman, Maya Aghaei, and Klaas Dijkstra. The Big Data Myth: Using Diffusion Models for Dataset Generation to Train Deep Detection Models. Eprint arXiv:2306.09762, 2023.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "scaled-yolov4: Scaling cross stage partial network". In *CVPR 2021*, pages 13029–13038, 2021.

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 *CVPR 2017*, pages 2097–2106, 2017.

Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. Cascaded Latent Diffusion Models for High-Resolution Chest X-ray Synthesis. In *PAKKD 2023*, pages 180–191. Springer Nature Switzerland, 2023.

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper S. Yao, Edward C. Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo A. Celi, Tanveer Syeda-Mahmood, and Mehdi Moradi. Chest ImaGenome Dataset (version 1.0.0). *PhysioNet*, 2021.

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. In *MICCAI 2023*, pages 101–111. Springer Nature Switzerland, 2023.

Appendix A. Supplemental Results

A.1. Visual and Visual-Text Alignment

We measure the visual alignment of our generated images with the MIMIC-CXR dataset using the Fréchet Inception Distance (FID) score, obtained with InceptionV3 (layer 2048) (Szegedy et al., 2016). To also assess the visual–text alignment, we employ the CLIPScore based on the CXR-CLIP model (You et al., 2023) and the XRayCLIP model (Chen et al., 2024)[3], which measures the similarity between each generated semi-synthetic image and its corresponding textual prompt.

The general and finding-level results are presented in Tables 3 and 4, respectively. Our approach achieves performance comparable to most existing methods, with only LLM-CXR yielding substantially better scores. However, a direct comparison of these scores across studies should be interpreted with caution due to: (i) the scores are computed on different subsets of finding classes, which alters the distribution of generated images, and (ii) the textual prompt distributions vary across studies, which can influence the CLIPScore. Overall, our method demonstrates image quality on par with state-of-the-art models, while uniquely providing ground-truth bounding boxes for the findings.

Table 3: Visual alignment (FID, InceptionV3) and visual–text (CLIPScore, CXR-CLIP) alignment of *SemiSynCXR*-generated CXRs. Our approach performs comparably to most existing methods, with only LLM-CXR achieving a notably better FID. These results, while broadly comparable, should be interpreted with caution due to differences in the findings considered and textual prompt distributions across studies. Overall, *SemiSynCXR* achieves competitive image quality while providing precise bounding boxes for radiological findings.

Model	$\text{FID}_{\text{InceptionV3}} \downarrow$	$\text{CLIPScore}_{\text{CXR-CLIP}} \uparrow$
RoentGen (Bluethgen et al., 2025) (synthetic)	64.60‡	0.29†‡
LLM-CXR (Lee et al., 2023) (synthetic)	**22.75**‡	0.20†‡
XReal (Hashmi et al., 2024) (synthetic)	55.12‡	–
CXRL (Han et al., 2024) (synthetic)	–	**0.34**‡
SemiSynCXR (ours)	63.99	0.30

†: Scores as reported in CXRL (Han et al., 2024).
‡: Limited comparability as these scores are averaged over a different set of finding classes.

A.2. Qualitative Assessment

We conducted a systematic study to qualitatively evaluate our generated CXRs. Three medical experts rated 140 randomly selected CXRs: 70 generated images (10 per finding) and 70 real images (10 per finding). For each CXR, raters identified (i) whether the image was real (realism) and (ii) which finding was present (finding recognition). Results, presented in Table 5, detail the False Positive Rate (FPR) for judging a generated image as

3. StanfordAIMIXrayCLIP__vit-l-14__laion2b-s32b-b82k

Table 4: Visual alignment and visual–text alignment per radiological finding, measured using FID (InceptionV3) and CLIPScore (CXR-CLIP, XRayCLIP), respectively.

	Atel.	Cmgl.	Cnls.	Edema	Opac.	P. Eff.	Pneum.	Avg.
Visual Alignment								
$\text{FID}_{\text{InceptionV3}} \downarrow$	61.62	73.22	63.24	67.09	58.57	61.34	62.84	63.99
Visual-Text Alignment								
$\text{CLIPScore}_{\text{CXR-CLIP}} \uparrow$	0.28	0.28	0.34	0.14	0.29	0.42	0.36	0.30
$\text{CLIPScore}_{\text{XrayCLIP}} \uparrow$	0.21	0.23	0.21	0.22	0.21	0.27	0.27	0.23

Table 5: Qualitative evaluation of our generated CXRs. Three medical experts assessed 140 scans (70 generated, 70 real with findings). On average, 36% of generated images were judged as real, and the intended finding was correctly identified in 64% of generated cases, suggesting that many inpainted findings are recognizable.

	Atel.	Cmgl.	Cnls.	Edema	Opac.	P. Eff.	Pneum.	Avg.
Generated Images								
As Real [FPR ↑]	0.10	**0.50**	0.30	0.40	0.37	0.47	0.40	0.36
With Finding [TPR ↑]	0.60	0.57	**0.83**	0.33	0.20	0.77	0.50	0.54
Real Images								
As Real [TPR ↑]	0.77	0.63	0.80	0.77	0.63	0.70	0.70	0.71
With Finding [TPR ↑]	0.13	0.43	0.20	0.17	0.13	0.53	0.37	0.28

real and the True Positive Rate (TPR) for correctly identifying the intended finding. For reference, we further report the TPRs in real images for realism and finding recognition.

On average, 36% of the 70 generated images are judged as real (vs. 71% on real images), and the intended finding is correctly identified in 54% of the generated cases (vs. 28% on real images). Pleural effusion and cardiomegaly demonstrate high rates for both realism and finding recognition. Atelectasis, consolidation, and pneumothorax also have high finding recognition rates, though there is room for improvement in their realism. Overall, these results suggest that many inpainted findings are recognizable, even so image generation artifacts may remain.

A.3. Ablation Studies

We perform an ablation study comparing editing pipelines (diffusion model with blending strategy – Section 3.5) under different mask blurring parameters. Specifically, we evaluated four pipelines: RoentGen with blending before, RoentGen with blending after, RoentGen with CFG masking, and RadEdit with CFG masking. The overall generation quality score, as defined in Section 3.6, is presented in Figure 4. RoentGen generally achieves better or comparable performance to RadEdit, except for pneumothorax and consolidation, where RadEdit is notably superior. The choice of blending strategy has only a minor impact. Additional mask conditioning results are presented in Figures 5 and 6.

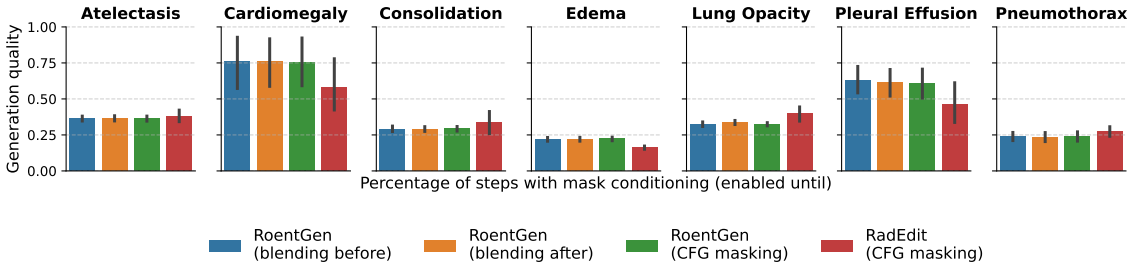


Figure 4: Ablation study on the different editing pipelines (RoentGen with blending before, RoentGen with blending after, RoentGen with CFG masking, and RadEdit with CFG masking). For each setting, we consider multiple mask blurring parameters shown as uncertainty intervals (Section 3.4) and compute the overall generation quality score following Section 3.6. We found that for most findings RoentGen generally performs better than or on par with RadEdit while the blending pipeline does not have a huge impact. However, note that for pneumothorax and consolidation RadEdit performs notably better.

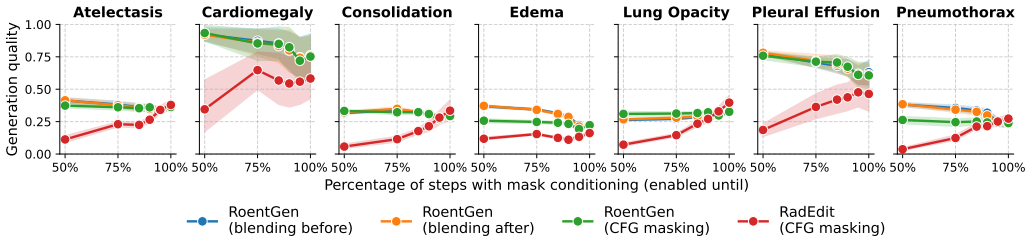


Figure 5: Ablation study on different blending pipelines and mask conditioning settings. We consider four different pipelines (RoentGen with blending before, RoentGen with blending after, RoentGen with CFG masking, and RadEdit with CFG masking) and vary the number of steps where mask conditioning is used. We enable mask conditioning for a specified percentage of steps (x-axis) before dropping it. For setting, we consider multiple mask blurring parameters (uncertainty intervals in the graph) and compute the overall generation quality score, following Sec. 3.6.

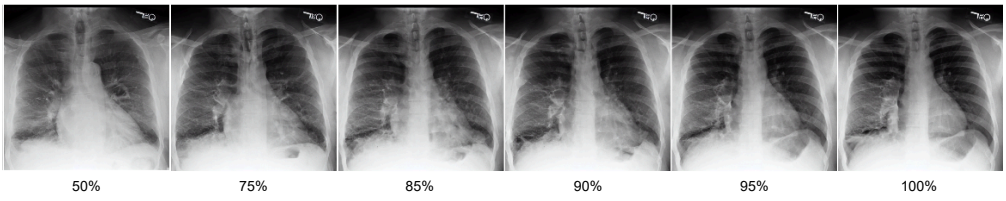
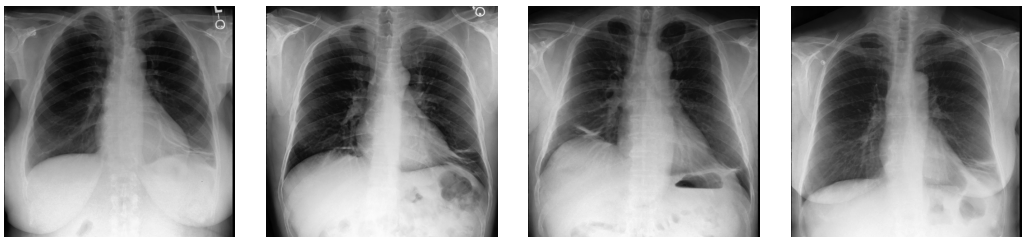
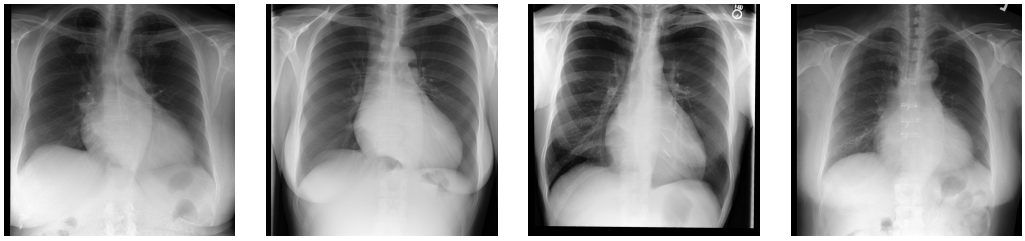


Figure 6: Example on the effect of using mask conditioning for a specified percentage of steps when inpainting edema with our approach.

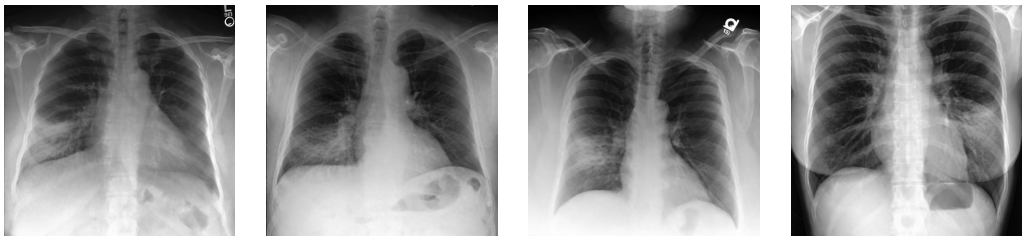
A.4. Examples of CXRs Generated by *SemiSynCXR*



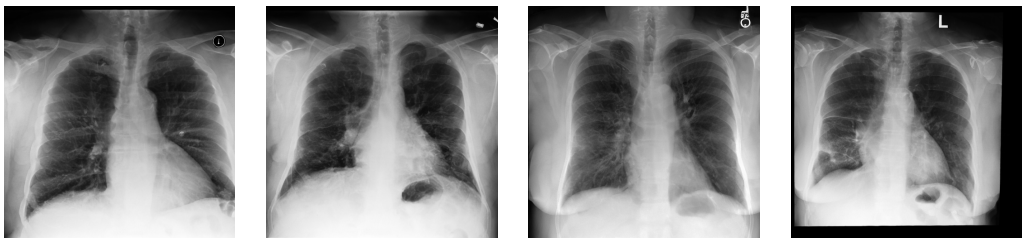
(a) Atelectasis



(b) Cardiomegaly

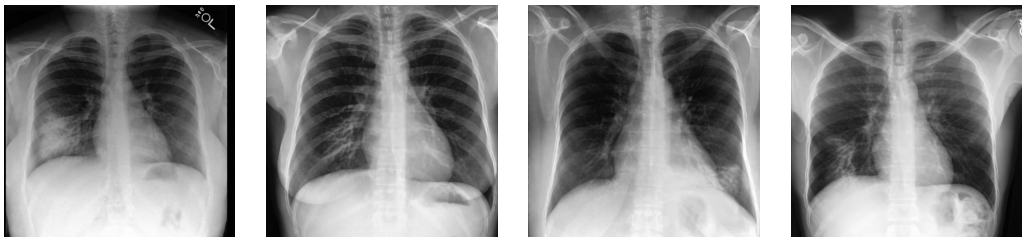


(c) Consolidation

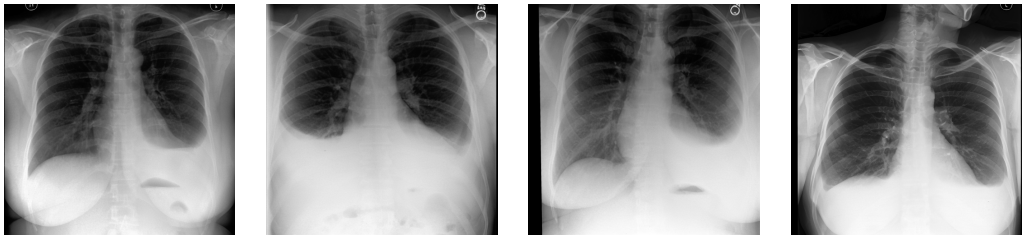


(d) Edema

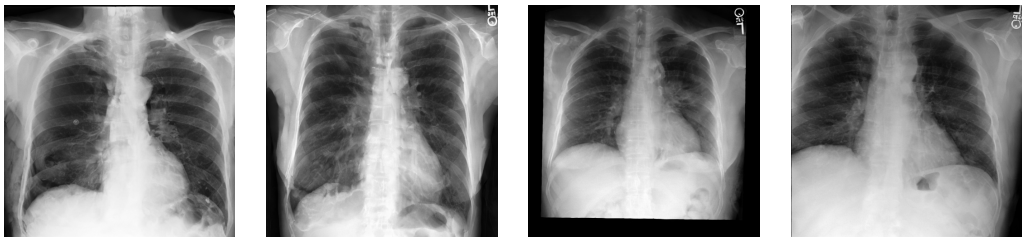
Table 6: Examples of images generated by our *SemiSynCXR* framework for atelectasis, cardiomegaly, consolidation, and edema.



(e) Lung Opacity



(f) Pleural Effusion



(g) Pneumothorax

Table 7: Examples of images generated by our *SemiSynCXR* framework for lung opacity, pleural effusion, and pneumothorax.

Appendix B. Existing Chest X-Rays Datasets for Radiological Finding Localization and Segmentation

We provide in Table 8 an overview of existing publicly available Chest X-ray (CXR) datasets for localization and segmentation of radiological findings. Datasets focusing on only one specific finding, such as those dedicated exclusively to pneumothorax or pulmonary nodules, have been excluded. Examples of the single-finding datasets omitted from this list include BIMCV-COVID19+, COVIDx CXR Dataset (COVID-Net) GRAZ+RSNA Pneumonia Detection Challenge, CANDID-PTX, PLCO Dataset (LIDC extension), SIIM-ACR Pneumothorax Segmentation.

Table 8: Overview of publicly available chest X-ray localization and segmentation datasets.

Dataset	Targets	Size (No. CXRs)	Origin (Country/Dataset)	Available	Notes
NIH ChestX-ray14	Disease labels (14), bounding boxes (8; subset)	112 120 frontal views; 880 with bounding boxes	United States	Kaggle, Google Cloud, and NIH download site	NLP-mined disease labels from radiology reports.
Vindr-CXR	Disease labels (6), bounding boxes (22; subset)	> 100 000; 18 000 with bounding boxes (15k train and 3k test)	Vietnam	PhysioNet	Training set labeled by three radiologists, test set labeled by consensus of 5 radiologists.
VinDr-PCXR	Disease labels (15), bounding boxes (36)	9 125 (7 728 train and 1 397 test)	Vietnam	PhysioNet	Labeled by experienced radiologists. Pediatric dataset.
MS-CXR	Bounding boxes (8) with descriptions	1 162	MIMIC-CXR	PhysioNet	Labels verified by board-certified radiologists.
PadChest-GR	Bounding boxes (24) with descriptions	4 555 frontal views	PadChest	BIMCV	Labeled by a team of 14 radiologists. Multilingual dataset (English and Spanish).
CXR-AL14	Bounding boxes (14)	165 988	China	Dong Zhang[a] or cxr-al14.top	Human-in-the-loop labeling.
REFLACX + LATTE-CXR	Eye-tracking with dictated reports, ellipses (localization), anatomical bounding boxes	3 032 frontal views	MIMIC-CXR	PhysioNet	Eye-tracking and dictated reports from five radiologists.
ChestX-Det Dataset	Bounding boxes (13), segmentation masks (13)	3 578	NIH ChestX-ray14	GitHub	Labeled by board-certified radiologists.
CheXlocalize	Segmentation masks (10), keypoints	902	MIMIC-CXR	Standford AIMI	Labeled by board-certified radiologists.

a. hszhangd@tmmu.edu.cn

Appendix C. Textual Prompts

We present below the curated set of radiological finding-related phrases used as textual prompts for conditioning the diffusion model during the editing process.

Rad. Finding	Textual Prompt	Probability
Atelectasis	Bibasilar atelectasis.	0.6406
	Left basilar atelectasis.	0.1647
	Basilar atelectasis.	0.0380
	Bibasilar subsegmental atelectasis.	0.0341
	Right basilar atelectasis.	0.0380
	Left lower lobe atelectasis.	0.0180
	Atelectasis in the lung bases.	0.0106
	Left basilar subsegmental atelectasis.	0.0053
	Streaky bibasilar atelectasis.	0.0042
	Subsegmental atelectasis.	0.0042
	Linear bibasilar atelectasis.	0.0063
	Atelectasis.	0.0158
	Left lower lobe collapse.	0.0032
	Right lower lobe atelectasis.	0.0021
	Right basilar subsegmental atelectasis.	0.0042
	Patchy bibasilar atelectasis.	0.0063
	Right upper lobe collapse.	0.0022
	Right middle lobe collapse.	0.0022
Cardiomegaly	Cardiomegaly.	0.7846
	Enlarged cardiac silhouette.	0.1940
	Enlargement of the cardiac silhouette.	0.0154
	Prominent cardiac silhouette.	0.0018
	Enlarged heart.	0.0042
Consolidation	Left lower lobe consolidation.	0.3064
	Right lower lobe consolidation.	0.2401
	Patchy consolidation in the mid left lung.	0.0704
	Patchy consolidation in the right lung.	0.0704
	Patchy consolidation in the right lower lobe.	0.1232
	Left consolidation.	0.0352
	Patchy bilateral pulmonary consolidations.	0.0352
	Bilateral consolidations.	0.0340
	Right middle lobe consolidation.	0.0511
	Right upper lobe consolidation.	0.0340
Edema	Pulmonary edema.	0.7310
	Interstitial pulmonary edema.	0.1333
	Interstitial edema.	0.1023
	Edema.	0.0175

	Peribronchial cuffing consistent with pulmonary edema.	0.0159
Lung Opacity	Right lower lobe infiltrate.	0.1635
	Right lower lobe opacity.	0.1499
	Left lower lobe opacity.	0.1908
	Bilateral lower lobe infiltrates.	0.0681
	Left lower lobe infiltrate.	0.0681
	Patchy bilateral pulmonary opacities.	0.0514
	Patchy left lower lobe opacity.	0.0681
	Bibasilar opacities.	0.0409
	Patchy ground-glass opacities at the right lung base.	0.0386
	Left basilar opacity.	0.0273
	Right basilar opacity.	0.0273
	Lower lung opacity.	0.0273
	Patchy ground-glass opacities in the left lower lung.	0.0257
	Patchy bibasilar opacities.	0.0530
Pleural Effusion	Bilateral pleural effusions.	0.3979
	Right pleural effusion.	0.2429
	Left pleural effusion.	0.2351
	Right effusion.	0.0388
	Bilateral effusions.	0.0310
	Right-sided pleural effusion.	0.0284
	Left-sided pleural effusion.	0.0207
	Left effusion.	0.0052
Pneumothorax	Right apical pneumothorax.	0.3472
	Left apical pneumothorax.	0.3208
	Right pneumothorax.	0.1774
	Left pneumothorax.	0.1245
	Pneumothorax.	0.0151
	Apical pneumothorax.	0.0075
	Bilateral pneumothoraces.	0.0075

Appendix D. Radiological Findings' Spatial Distributions

Table 10: Probability distributions for the spatial modeling of radiological finding bounding boxes. Parameters for the 1D best-fit distributions correspond to the location-scale form of the distribution in regard.

Radiological Finding	Bounding Box Element	Distribution	Parameters	Score
Atelectasis	Center (x-axis)	Beta	a : 194.8522, b : 78.1808×10^6, loc : -119.7766, scale : 66.6710×10^6	2.36×10^{-4}
	Center (y-axis)	Log-gamma	c : 0.7680, loc : 87.8522, scale : 6.9387	1.07×10^{-3}
	Width & height	Multivariate log-normal	$\boldsymbol{\mu}$: [4.3618, 3.4926], $\boldsymbol{\Sigma}$: [[0.0842, 0.0632], [0.0632, 0.2054]]	–
Cardiomegaly	CTR	Gamma	a : 40.4439, loc : 33.4765, scale : 0.6308	4.66×10^{-3}
Consolidation	Center (x-axis)	Log-normal	s : 0.1733, loc : -24.9657, scale : 69.8613	6.94×10^{-5}
	Center (y-axis)	Beta	a : 9.3284, b : 3.6820, loc : -32.9031, scale : 132.7595	1.20×10^{-4}
	Width & height	Multivariate log-normal	$\boldsymbol{\mu}$: [4.1543, 3.6383], $\boldsymbol{\Sigma}$: [[0.1449, 0.1393], [0.1393, 0.3113]]	–
Edema	Center	Multivariate log-normal	$\boldsymbol{\mu}$: [3.8485, 3.9856], $\boldsymbol{\Sigma}$: [[0.0968, -0.0336], [-0.0336, 0.0529]]	–
	Width & height	Multivariate log-normal	$\boldsymbol{\mu}$: [4.2697, 3.9856], $\boldsymbol{\Sigma}$: [[0.1678, 0.1776], [0.1776, 0.2681]]	–
Lung Opacity	Center (x-axis)	Beta	a : 8.3376, b : 4 440.2773, loc : 4.3958, scale : 2 2401.2265	2.24×10^{-4}
	Center (y-axis)	Log-gamma	c : 0.7223, loc : 77.5134, scale : 10.4228	1.89×10^{-4}
		log-normal	$\boldsymbol{\Sigma}$: [[0.0968, -0.0336], [-0.0336, 0.0529]]	
	Width & height	Multivariate log-normal	$\boldsymbol{\mu}$: [4.0257, 3.5337], $\boldsymbol{\Sigma}$: [[0.1754, 0.1410], [0.1410, 0.3298]]	–
Pleural Effusion	Center (y-axis)	GEV	c : 0.6393, loc : 88.9743, scale : 7.3628	6.34×10^{-4}
	Height	Beta	a : 1.2177, b : 3.5508, loc : 13.3016, scale : 22 401.2265	2.43×10^{-5}
Pneumothorax	Center	Multivariate log-normal	$\boldsymbol{\mu}$: [3.9222, 2.7920], $\boldsymbol{\Sigma}$: [[0.277, -0.3239], [-0.3239, 1.0157]]	–
	Width & height	Multivariate log-normal	$\boldsymbol{\mu}$: [4.1561, 3.2241], $\boldsymbol{\Sigma}$: [[0.1881, 0.0425], [0.0425, 0.4092]]	–

GEV: Generalized extreme value distribution
CTR: Cardiothoracic ratio.

Appendix E. Experimental Setup

As part of our experiments, we explored several design variations for generating semi-synthetic chest X-rays. These variations contemplated the following:

- Text prompts: Class labels directly (e.g., "Atelectasis") or sampled class-related phrases (see Section 3.3 and Appendix C).
- Mask blurring: No blur and six blurring alternatives using generalized Gaussian filters with standard deviation based on the size of the mask and a factor r, i.e.,

$$\boldsymbol{\sigma} = (\lfloor 0.5 \times \text{size(bounding box)} \rfloor)/r.$$

  Specifically, we tested with $\beta = 2.0$ and $r \in \{0.5, 1.0, 2.0\}$, and $r = 0.1$ and $\beta \in \{4.0, 6.0, 8.0\}$.
- Mask conditioning extend: 85%, 90%, 95%, and 100% of the total number of steps of the reverse diffusion process.
- Hyperparameters of the diffusion process:
  - Number of steps: 1.0, 1.33, and 2.0 times the default number of parameters.
  - Classifier-free guidance scale: RadEdit with default, 12.0, 15.0, and 19.0; RoentGen with default, 6.0, 7.5, and 10.0.
  - Strength: Default, 0.9, and 1.0.
  - Negative conditioning: $\emptyset$ (no negative prompt), and "No acute cardiopulmonary process."

  Default settings were used for all other underlying model hyperparameters.

Note on Implementation: *SemiSynCXR*'s editing component is built upon the Stable Diffusion Inpainting pipeline from HuggingFace (von Platen et al., 2022).