Supplementary Material

RadEdit: stress-testing biomedical vision models via diffusion image editing

A Medical terminology

With our editing approach being readily applicable to many (non-medical) applications, we tried our best to keep the paper as accessible as possible to a wider audience, using only a small number of medical terms. In the following section we describe the terms used in more detail.

Note, when interpreting a chest X-ray, it is important to remember that the left and right sides are switched. This is because we view the patient from their anatomical laterality point of view, as if we are facing them. So, what appears on the left in an image is actually the patient's right side, and vice versa.

A.1 Pathologies

Cardiomegaly This term refers to an enlarged heart, which is usually indicative of an underlying heart condition. The enlargement can include the entire heart, one side of the heart, or a specific area. On a chest X-ray, the heart may appear larger than normal.

Opacities In the context of a chest X-ray, opacity is a nonspecific descriptor for areas that appear whiter than normal lung. Normally, lungs look dark gray on an X-ray due to presence of air (note the black pure air surrounding the patient on x-ray for reference). If there are whiter areas, it means something is filling up that space inside the lungs, replacing the air.

Pulmonary Edema is caused by accumulation of fluid in the lungs. In the context of chest X-rays, pulmonary edema appears as increased opacity within and around the air space. In Fig. 12, we show a variety of pulmonary edema examples.

Consolidation In the context of chest X-rays, consolidation refers to a region of the lung where the air spaces are filled with fluid, cells, tissue, or other substances. This results in a white region on the X-ray. In Fig. 14, we show a variety of consolidation examples.

COVID-19 refers to pneumonia caused by SARS-CoV-2 virus which manifests most commonly as multifocal, bilateral opacities with predominance in the lower half of the lung.

Pneumothorax This condition occurs when air leaks into the pleural space (between the lung and chest wall), causing the lung to collapse. It can be a complete lung collapse or a collapse of only a portion of the lung. On a chest X-ray, a pneumothorax is seen as a dark region around the edge of the lung, lacking any white texture (except the ribs). The border of the collapsed lung can be seen as in Fig. 8a at the inferior contour of the mask. Often small pneumothorax can be hard to spot on a chest X-ray which contributed to computer vision models overly relying on chest drains for detection, see Sec. 5.3.

A.2 Support devices

Chest drain This is a tube inserted into the pleural space to remove unwanted air (pneumothorax) or fluid (pleural effusion). On an X-ray, you can see the tube in the form of two parallel thin white lines. Its position depends on what it is treating: for pneumothorax it is aimed towards the top; if it is draining fluid, it is towards the bottom.

Pacemaker This is a device placed under the skin near the collarbone. It helps control abnormal heart rhythms. It has two parts: a control unit (battery and electronics) and wires (white lines) that connect to the heart. In Fig. 13, we show a variety of pacemaker examples.

B Details for DDPM inversion and DiffEdit

In this section we provide some additional details on how the editing process is performed. Algorithm 2 describes the DDPM inversion process which is used by RadEdit to encode images to a sequence of vectors. Additionally, to explicitly see how RadEdit (Algorithm 1) differs from DiffEdit [11], we provide Algorithm 3, which describes the DiffEdit editing method using DDPM inversion.

Algorithm 2 DDPM inversion [30]

Require: image x_0, inversion prompt c_{inv}, diffusion model f_θ

▷ *Sample statistically independent* $\tilde{\epsilon}_t$

for $t \leftarrow 1$ to T **do**

  $\tilde{\epsilon}_t \sim \mathcal{N}(0, I)$

  $\hat{x}_t \leftarrow \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\tilde{\epsilon}_t$

▷ *Isolate* z_t *from series* $\hat{x}_{1:T}$

for $t \leftarrow T$ to 1 **do**

  $\epsilon_t \leftarrow f_\theta(\hat{x}_t, t, c_{\text{inv}})$

  $z_t \leftarrow (\hat{x}_{t-1} - \hat{\mu}_t(\hat{x}_t, \epsilon_t))/\sigma_t$

  ▷ *Avoid error accumulation*

  $\hat{x}_{t-1} \leftarrow \hat{\mu}_t(\hat{x}_t, \epsilon_t) + \sigma_t z_t$

return $(\hat{x}_{1:T}, z_{1:T})$

Algorithm 3 DiffEdit [11] w/ DDPM inversion

Require: image x_0, inversion prompt c_{inv}, edit prompt c, edit mask m_{edit}, CFG weight w, diffusion model f_θ

$(\hat{x}_{1:T}, z_{1:T}) \leftarrow$ DDPMINVERSION(x_0, c_{inv})

$x_T \leftarrow \hat{x}_T$

for $t \leftarrow T$ to 1 **do**

  $\epsilon_{\text{cond},t} \leftarrow f_\theta(x_t, t, c)$

  $\epsilon_{\text{uncond},t} \leftarrow f_\theta(x_t, t, c = \emptyset)$

  ▷ *classifier-free guidance (CFG)*

  $\epsilon_t \leftarrow \epsilon_{\text{uncond},t} + w(\epsilon_{\text{cond},t} - \epsilon_{\text{uncond},t})$

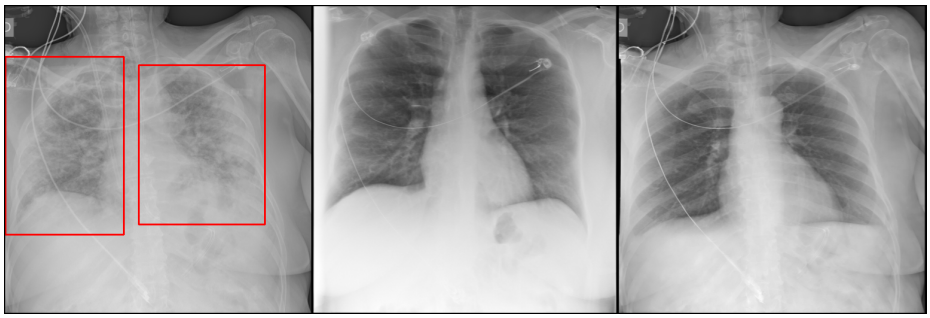
  $x_{t-1} \leftarrow \hat{\mu}_t(x_t, \epsilon_t) + \sigma_t z_t$

  $x_{t-1} \leftarrow m_{\text{edit}} \odot x_{t-1} + (1 - m_{\text{edit}}) \odot \hat{x}_{t-1}$

return edited version of x_0

C Details for the limitations of LANCE

During the development of RadEdit, we observed numerous artefacts when editing images from the BIMCV+ or CANDID-PTX datasets without using masks. In both instances, the pathology and the lateral markings or chest tubes were removed, leading to potential misinterpretations of the results if these edited images were used for stress-testing. Note, that instead of using a captioner and perturber as seen in the original implementation of LANCE, we manually select the prompts used for editing. In Fig. 7, we compare RadEdit with LANCE (which does not use masks) in editing images from the BIMCV+ dataset. This comparison follows the same experimental setup as in Sec. 5.2. RadEdit retains the laterality marker on the left of the image, whereas LANCE completely removes it. In both scenarios, we employ the prompt *'No acute cardiopulmonary process'*[3] to edit the image.



(a) Original Image **(b)** LANCE [53] **(c)** RadEdit (ours)

Fig. 7: Using LANCE (b) to remove COVID-19 features (rectangle in (a)),the laterality markers are missing. In addition, the field of view is changed. In contrast, RadEdit (c; ours) uses masks to preserve laterality markers, which also preserves anatomical structures in the process, and retains the original contrast.

Similarly, in Fig. 8, we attempt to remove only the pneumothorax from an image containing a pneumothorax and chest drain, using the prompt *'No acute cardiopulmonary process'*[3], while preserving the rest of the image, including the chest drain. For a more comprehensive description of the experimental setup, refer to Sec. 5.3. For LANCE (Fig. 8b), we note that not only is the region containing the pneumothorax altered, but the chest drain is also removed. This makes LANCE unsuitable for evaluations such as our manifestation shift evaluation (Sec. 5.3), which requires the preservation of support devices like chest drains. We argue that this artefact suggests that the diffusion model has learned correlations between pathologies and support devices, leading to the removal of support devices when prompted to remove a pathology.

In Fig. 9, we compare RadEdit with LANCE in editing images from the CANDID-PTX dataset using the prompt *'No pneumothorax'*. We observe that LANCE generates a variety of artefacts. While it retains most of the chest drain, LANCE fails to effectively remove the pneumothorax, instead altering its appearance to resemble a wire. Addition-

(a) Original Image **(b)** LANCE [53] **(c)** RadEdit (ours)

Fig. 8: Removing pneumothorax (red) from X-rays using LANCE (b) results in the spuriously correlated chest drain (blue) also being removed. RadEdit (c, ours) uses pneumothorax and chest drain masks to remove the pneumothorax while preserving the chest drain. LANCE results in decreased contrast and poorly defined anatomical structures, preserved by RadEdit.

ally, there are extensive bilateral artefacts, with modifications to the abdomen, face, and arms, altered gas pattern and heart, and the lung apices no longer being asymmetrical, raising questions about whether the X-rays are from the same patient.

One potential explanation for the artefacts seen in this section is found in recent literature on diffusion models for image-to-image translation. In Su et al. [71], the authors show that image-to-image translation can be performed with two independently trained diffusion models. They first obtain a latent representation $\hat{x}_t$ from a source image x_0 with the source diffusion model, and then decode the latent using the target model to construct the target image. We argue that since the diffusion model in Sec. 5.1 was not trained on data from BIMCV+ or CANDID-PTX, in those cases we perform image-to-image translation along with the image editing. I.e., editing images outside of the training distribution of the diffusion model leads to images that look more similar to images from within the training distribution. In the case of RadEdit, where we heavily rely on masks to control the editing, we only observe minor artefacts. However, in the case of LANCE, we observe major artefacts that make LANCE unsuitable for stress-testing of biomedical imaging models. To avoid artefacts, we tried different values for the LANCE hyperparameters, such as the guidance scale, without success.

D Details for the limitations of DiffEdit

In contrast to LANCE, DiffEdit employs a single mask m_{edit} for editing. As the editing is only applied within m_{edit}, DiffEdit avoids the artefacts described in the previous section. However, DiffEdit introduces new artefacts.

In general, DiffEdit consists of two steps. First, it predicts the edit mask m_{edit} using the difference between the original prompt and the editing prompt. Second, the editing, following the editing prompt, is applied inside the predicted mask m_{edit}, leaving the area outside of the mask unchanged. When applying DiffEdit to the experimental setups of Sec. 5.3 and Sec. 5.4 we find problems with both instances.

(a) Original Image **(b)** LANCE [53] **(c)** RadEdit (ours)

Fig. 9: Removing pneumothorax from X-rays using RadEdit (c; ours) results in a minimally modified X-ray, with the pneumothorax successfully removed and chest drain still present. In contrast, LANCE (b) fails to properly remove the pneumothorax while keeping most of the chest drain in place, instead modifying the appearance of the drain to look more like a wire; moreover, there are extensive artefacts bilaterally, with abdomen, face, and arms added, modified gas pattern and heart, as well as the lung apexes no longer being asymmetrical, making it unclear whether the X-rays are of the same patient. Blue: ground-truth annotation for chest drain; red: ground-truth annotation for pneumothorax.

Initially, we quantify how well the mask automatically predicted by DiffEdit aligns with the ground-truth annotation. We use the same setup as in Sec. 5.3: we take an image containing a pneumothorax and a chest drain (sourced from the CANDID-PTX dataset) and aim to remove only the pneumothorax. We create the editing prompt by splitting the original impressions into one part containing a description of the pneumothorax and another part containing a description of the chest drain. We then replace the part containing the description of the pneumothorax with *'No pneumothorax'*. Therefore, DiffEdit should predict a mask containing only the pneumothorax. We perform a grid search on the validation CANDID-PTX dataset over DiffEdit's hyperparameters, optimising for pneumothorax segmentation metrics, and then evaluate on the training set. In Fig. 6, we show that masks predicted by DiffEdit obtain poor quantitative metrics compared to the manually annotated masks, where parts of the pneumothorax are often missing, and the spuriously correlated chest drain is often included in the automatically predicted mask. As a result, masks predicted by DiffEdit are unsuitable for editing images that can be used for stress-testing.

Secondly, in contrast to RadEdit, which allows the area outside of the mask to change for consistency, DiffEdit restricts the changes to happen inside the mask. While this would generate valid edits for the experiment in Sec. 5.2, it can lead to artefacts in the case of the experiments in Sec. 5.3 and Sec. 5.4.

Following the setup from Sec. 5.4, our goal is to add consolidation to the left upper lung of a healthy patient. In Fig. 10, we compare the editing results of RadEdit and DiffEdit. While RadEdit leads to a realistic occlusion of the heart, DiffEdit fails to generate a realistic-looking edit. Instead, it creates a visible gap between the consolidation and the heart border, which makes the edited image unsuitable for stress-testing a lung segmentation model.

(a) Original Image **(b)** DiffEdit [11] **(c)** RadEdit (ours)

Fig. 10: Adding consolidation to the left lung using DiffEdit (b) results in a dark border along the original lung mask (red) since editing can only occur within the masked region. RadEdit (c; ours) allows the region outside of the mask to change to ensure consistency, resulting in more realistic edits. For both editing methods, we use ground-truth masks of the lung.

E Experimental details for Section 5.1: diffusion model

In this section, we provide additional details on how the diffusion model used for all experiments in Sec. 5 was trained. The VAE downsamples the input images by a factor of eight, meaning that the latent space has spatial dimensions 64×64. For the diffusion model, we use the linear $beta$ schedule and ϵ-prediction proposed by Ho et al. [28]. The U-Net architecture is as used by Rombach et al. [59], which we instantiate with base channels 128, channel multipliers (1, 2, 4, 6, 8), and self-attention at feature resolutions 32×32 and below, with each attention head being 32-dimensions. The BioViL-T text encoder [5] has a maximum token length of 128, so sentences within the impression are shuffled and then clipped to this length. An exponential moving average is used on model parameters, with a decay factor of 0.999. We drop the text conditioning with $p = 0.1$ during training to allow CFG when sampling [27]. Training was performed using 48 V100 GPUs for 300 epochs using automatic mixed precision. The AdamW [46] optimiser was used, with a fixed learning rate of 10^{-4}.

The preprocessing steps are:

1. Resize such that the short side of the image has size 512, using bilinear interpolation;
2. Centre-crop to 512×512 pixels;
3. Map minimum and maximum intensity values to $[-1, 1]$.

We use the following label categories for the CheXpert dataset:

1. Atelectasis
2. Cardiomegaly
3. Consolidation
4. Edema
5. Enlarged cardiomediastinum
6. Fracture
7. Lung lesion
8. Lung opacity
9. No finding
10. Pleural effusion
11. Pleural other
12. Pneumonia
13. Pneumothorax
14. Support devices

For ChestX-ray8, we use:

1. Atelectasis
2. Cardiomegaly
3. Consolidation
4. Edema
5. Effusion
6. Emphysema
7. Fibrosis
8. Hernia
9. Infiltration
10. Mass
11. No Finding
12. Nodule
13. Pleural thickening
14. Pneumonia
15. Pneumothorax

F Experimental details for Section 5.2: acquisition shift

The datasets used and their respective train / validation / test splits are as follows:

1. BIMCV+: 3008 / 344 / 384
2. BIMCV-: 1721 / 193 / never used for testing
3. MIMIC-CXR: 5000 / 500 / 500 (randomly sampled)
4. Synthetic: never used for training or validation / 2774 (after filtering)

All splits were made ensuring non-overlapping subject IDs.

The filtering of the synthetic test dataset was done using the prompts: *'Opacities'* and *'No acute cardiopulmonary process'*[3].

For training, we converted the original labels of the BIMCV datasets as follows: if an image has the label 'Negative for Pneumonia' or 'Atypical Appearance' we assign label 0; while if it has the label 'Typical Appearance' or 'Indeterminate Appearance' we assign label 1.

The classifier is trained using a ResNet50 architecture with batch size 32, 100 epochs and learning rate 10^{-5}. The model was evaluated at the point of best validation area under the receiver operating characteristic curve (AUROC).

The preprocessing steps are as in Appendix E, but image intensities are mapped to $[0, 1]$.

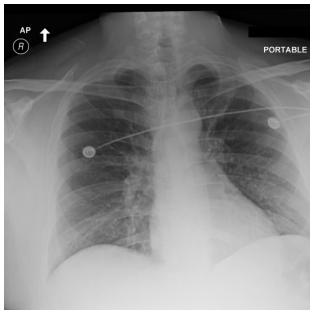
The following augmentations were used:

1. Random horizontal flip with probability 0.5
2. Random affine transformations with rotation $\theta \sim \mathcal{U}(-30, 30)$ degrees and shear $\phi \sim \mathcal{U}(-15, 15)$ degrees
3. Random colour jittering with brightness $j_b \sim \mathcal{U}(0.8, 1.2)$ and contrast $j_c \sim \mathcal{U}(0.8, 1.2)$
4. Random cropping with scale $s \sim \mathcal{U}(0.8, 1)$
5. Addition of Gaussian noise with mean $\mu = 0$ and standard deviation $\sigma = 0.05$

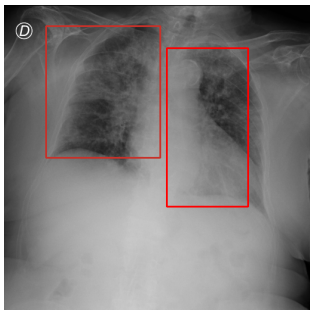
G Experimental details for Section 5.3: manifestation shift

The datasets used and their respective train / validation / test splits are as follows:

1. CANDID-PTX: 13 836 / 1539 / 1865
2. SIIM-ACR: 10 712 / 1625 / never used for testing



(a) Example image from MIMIC-CXR [34].



(b) Example image from BIMCV+ [76].

Fig. 11: Comparison of the visual appearance between the MIMIC-CXR and BIMCV+ datasets. As shown by [12] there are distinct differences in the laterality markings (top left corner) and field of views of the images. Bounding boxes in (b) indicate the presence of abnormalities caused by COVID-19.

3. Synthetic: never used for training or validation / 629 (after filtering)

All splits were made ensuring non-overlapping subject IDs.

The filtering of the synthetic test dataset was done using the prompts: *'Pneumothorax'* and *'No acute cardiopulmonary process'*[3].

After observing that the contours of the pneumothorax and chest drain masks often do not include the borders of the pneumothorax or chest drain we apply isotropic dilation with a radius of 5. Examples of such dilated masks can be seen in Fig. 9 (a).

For the weak predictor, the same model architecture, training hyperparameters and data augmentation are as described in Appendix F

In the case of the strong predictor model, a segmentation model is trained using the EfficientNet U-Net [72] architecture. We add a single classification layer to the lowest resolution of the U-Net. The segmentation model is trained to segment pneumothorax, and the classifier is used to detect the presence of pneumothorax.

The combined model is trained for 100 epochs with batch size 16, learning rate 5×10^{-4} , and a cosine scheduler with warm-up during the first 6% of steps. The model was evaluated at the point of best validation AUROC for the pneumothorax classifier.

Data preprocessing and augmentation were as described in Appendix F, with $s \sim \mathcal{U}(0.9, 1.1)$. Additionally, a random elastic transform with scale 0.15 (as implemented in Albumentations [7]) was used.

H Experimental details for Section 5.4: population shift

Prompts used are as follows:

- Pulmonary edema: *'Moderate pulmonary edema. The heart size is normal'*
- Pacemaker: *'Left pectoral pacemaker in place. The position of the leads is as expected. Otherwise unremarkable chest radiographic examination'*
- Consolidation: *'New [left/right] upper lobe consolidation'*

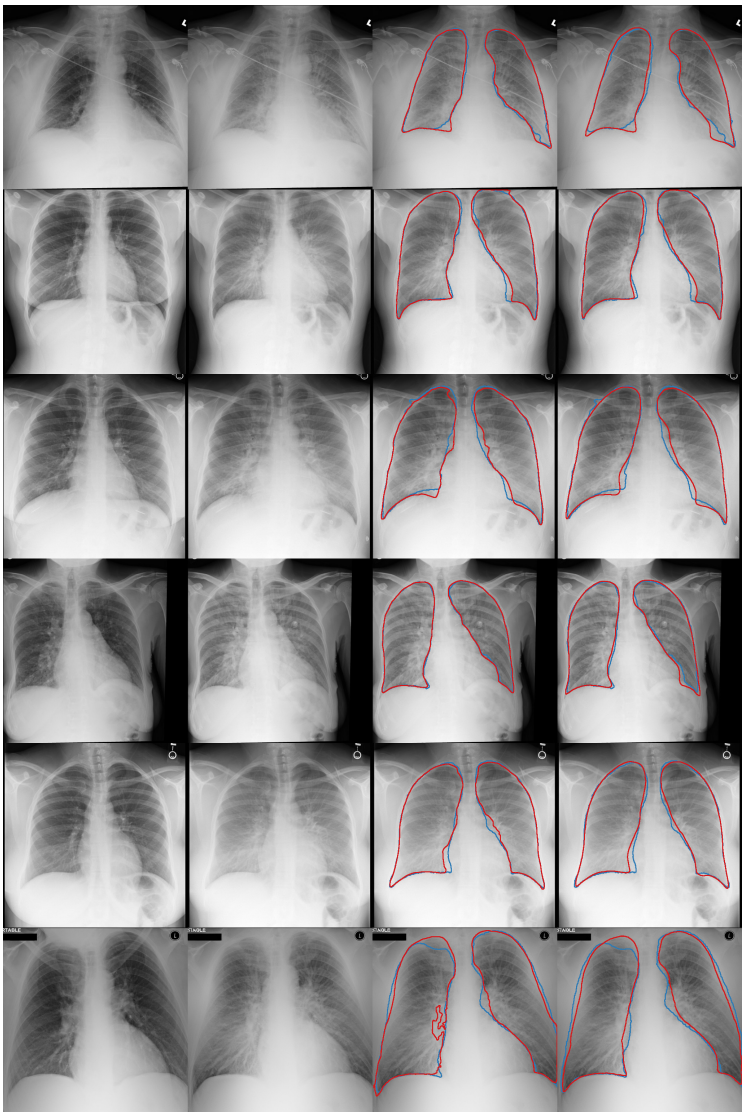
The datasets used and their respective train / validation / test splits are as follows:

1. MIMIC-Seg: 911 / 114 / 115
2. CheXmask: 169 206 / 36 580 / 36 407
3. Synthetic Edema: never used for training or validation / 787 (after filtering)
4. Synthetic Pacemaker: never used for training or validation / 744 (after filtering)
5. Synthetic Consolidation: never used for training or validation / 1577 (after filtering)

All splits were made ensuring non-overlapping subject IDs.

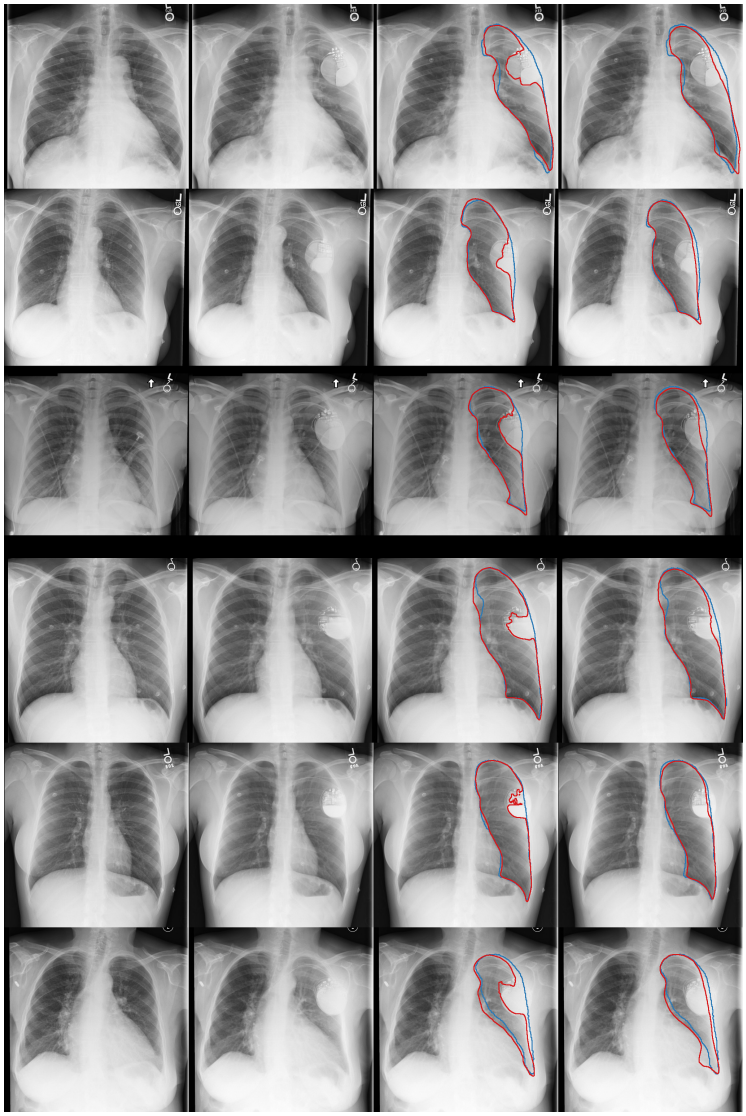
The same segmentation model architecture, training hyperparameters, and data augmentation/preprocessing steps are used as described in Appendix G.

In Figures 12 to 14 we show more examples of edits produced by RadEdit to stress test the segmentation models. RadEdit edits are high-quality, with both general anatomy maintained after the edit, as well as image markings.



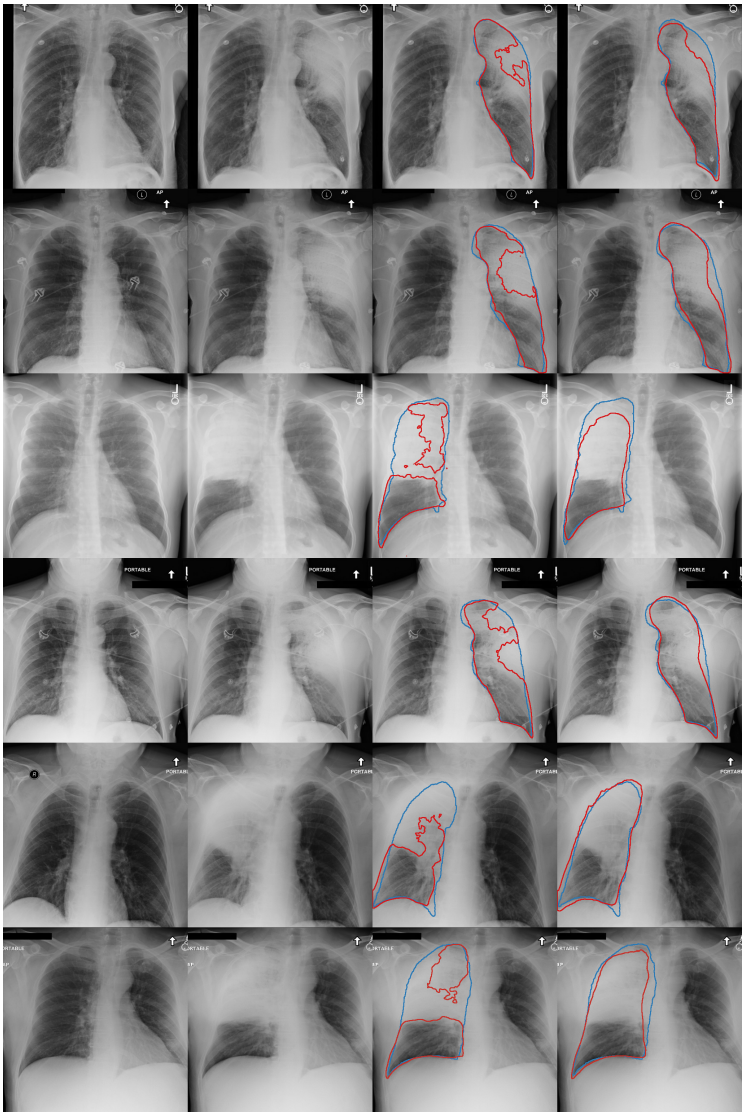
(a) Original (b) Edited (c) Weak Predictor (d) Strong Predictor

Fig. 12: Additional edits simulated by RadEdit for stress-testing two segmentation models. The 'weak predictor' (c) and the 'strong predictor' (d) are trained on MIMIC-Seg [10] and CheXmask [17] respectively, by adding pulmonary edema, via the prompt *'Moderate pulmonary edema. The heart size is normal.'* Blue: ground-truth mask: ; red: predicted. Similar to the example in Fig. 4, both segmentation models predict relatively accurate segmentation maps, indicating a high level of robustness to this pathology. Edits are visually high quality, with anatomy well maintained, and the edema clearly identifiable.



(a) Original (b) Edited (c) Weak Predictor (d) Strong Predictor

Fig. 13: Additional edits simulated by RadEdit for stress-testing two segmentation models. The 'weak predictor' (c) and the 'strong predictor' (d) are trained on MIMIC-Seg [10] and CheXmask [17] respectively, by adding pacemakers, which can be seen in the top left of images, via the prompt *'Left pectoral pacemaker in place. The position of the leads is as expected. Otherwise unremarkable chest radiographic examination.'* Blue: ground-truth mask: ; red: predicted. Similar to the example in Fig. 4, the segmentation model trained on MIMIC-Seg (which contains predominantly healthy patients) incorrectly segments around the pacemakers, while the model trained on CheXmask (which is larger and contains various abnormal cases), segments more accurately.



(a) Original (b) Edited (c) Weak Predictor (d) Strong Predictor

Fig. 14: Additional edits simulated by RadEdit for stress-testing two segmentation models. The 'weak predictor' (c) and the 'strong predictor' (d) are trained on MIMIC-Seg [10] and CheXmask [17] respectively, by adding upper-lobe consolidation, via the prompt *'New [left/right] upper lobe consolidation.'* Blue: ground-truth mask: ; red: predicted. Similar to the example in Fig. 4, both models are less able to segment the lungs accurately, however, segmentations by the model trained on MIMIC-Seg are notably worse, often excluding the consolidated region.