Appendix

In this supplemental material, we include additional backgrounds, visualization results, and analyses. The contents of the individual sections are:

- Appendix Sec. A: Background on domain generalization and causality.
- Appendix Sec. B: visualization results and qualitative analysis of the ablation study on the impact of fine-tuning the SD U-Net on source domain D_0 using DreamBooth [49] before the ControlNet [65] training.
- Appendix Sec. C: Additional visualization results of segmentation maps predicted by our method and competing baselines on the abdominal segmentation (AS), lumbar spine segmentation (LSS) and lung segmentation (LS) tasks.
- Appendix Sec. D: Additional visualization results of diffusion-generated augmentations for the AS, LSS and LS tasks with different style-intervention prompts.
- Appendix Sec. E: Dataset and preprocessing details.
- Appendix Sec. F: Implementation and training details of our method and baselines, including model architectures, hyperparameters, etc.

A. Background: Generalization and Causality

Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ represent input images and corresponding segmentation masks. We further assume that an expert (or oracle) is able to provide correct segmentation masks Y from observations X alone. In the context of *single-source* DG (SDG), we assume that we are given training pairs from a single domain D_0: $\{(X_i^0, Y_i^0)\}_{i=1}^{n_0}$, and a fixed set of target domains $D_1, \ldots, D_N$. Each dataset D_e contains independent and identically distributed (i.i.d.) samples from some probability distribution $P(X^e, Y^e)$. We aim to obtain an optimal predictor f to enable out-of-distribution (OOD) generalization. In particular, f is trained on D_0 such that f minimizes the worst case risk $R^e(f) := \mathbb{E}_{X^e,Y^e}\left[\ell(f(X^e), Y^e)\right]$ for any given target domain D_e:

$$R_{\text{OOD}}(f) = \max_{e\in\{1,\ldots,N\}} R^e(f). \tag{7}$$

Arjovsky et al. [2] demonstrate that an invariant predictor would obtain an optimal solution for Eq. (7):

Definition 1 ([2]). A data representation $\Phi : \mathcal{X} \to \mathcal{H}$ elicits an invariant predictor $w \circ \Phi$ across environments (domains) $\mathcal{E}$ if there is a classifier $w : \mathcal{H} \to \mathcal{Y}$ simultaneously optimal for all environments, that is,

$$w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w} \circ \Phi) \quad \text{for all } e \in \mathcal{E}.$$

Arjovsky et al. [2] propose an Invariant Risk Minimization (IRM) algorithm to obtain this invariant solution from *multiple* domains and show that this solution is optimal if and only if it uses only the direct causal parents of Y in the corresponding structural causal model (SCM) [42]. In this work, we demonstrate how to achieve this optimal solution given a *single* source domain.

We extend common assumptions [31, 40] on the data generative process from a causal perspective by allowing an additional causal relation from content to style as proposed in [59]:

$$\begin{aligned} C &:= g_C(U_c), \\ S &:= g_S(C, U_s), \\ X &:= g_X(C, S), \\ Y &:= g_Y(C), \end{aligned} \tag{8}$$

where C denotes the latent content variable, S denotes the latent style variable, $X \sim p_X$ is an observed input variable, Y is the observed segmentation mask, $(U_c, U_s) \sim p_{u_c} \times p_{u_s}$ are independent exogenous variables, and g_C, g_S, g_X, g_Y are deterministic functions. We assume that different domains are generated via an intervention on the style variable S.

Under the SCM described above (Eq. (8)), von Kügelgen et al. [59] theoretically prove that if the augmented pairs of views (X, X^+) in contrastive methods are generated under the principle of a "soft" intervention on S, then the InfoNCE [39] objective combined with an encoder function Φ (*e.g.*, neural network) *identifies* the invariant content C partition in the generative process described above (Eq. (8)):

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}_{X\sim p_X}\left[\log \frac{\exp(\text{sim}(\nu, \nu^+)/\tau)}{\sum_{X^-\in\mathcal{N}} \exp(\text{sim}(\nu, \nu^-)/\tau)}\right], \tag{9}$$

where $\nu = \Phi(X)$, $\text{sim}(\cdot,\cdot)$ is some similarity metric (*e.g.*, cosine similarity), τ is a temperature parameter, and X^- are negative pairs.

In real-world applications, direct intervention on the style variable is infeasible as it is unobserved. For instance, in medical imaging, this would require scanning the same patient using different imaging modalities multiple times under controlled conditions—a process that is often impractical for acquiring training data. However, Ilse et al. [19] introduce the concept of *intervention-augmentation equivariance*, formally demonstrating that data augmentation can serve as a surrogate tool for simulating interventions:

Definition 2 ([19]). The causal process g_X is *intervention-augmentation equivariant* if for every considered stochastic data augmentation transformation $\text{aug}(\cdot)$ on $X \in \mathcal{X}$ we have a corresponding noise intervention $do(\cdot)$ on S such that:

$$\text{aug}(X) = g_X(do(S = s), c). \tag{10}$$

B. Visualizations of Ablation Study

We conduct an ablation study on the LSS task (CT→MRI) to investigate the efficacy of the Style Swap technique [22] in the main paper. This technique involves two steps: first, fine-tuning the pretrained Stable Diffusion (SD) U-Net (U^B) on the source domain using DreamBooth [49] to obtain an instance-tuned SD U-Net U^{D_0}, and then train the ControlNet [65] with U^{D_0} so that it can focus on learning to inject image conditions into the generation process rather than domain information from D_0. Specifically, we compare this method to directly training the ControlNet with U^B, omitting the instance fine-tuning stage with DreamBooth.

To further illustrate this point, we visualize the images generated by U^B using two versions of ControlNet: one trained with U^{D_0} (fine-tuned on D_0 with DreamBooth) and another trained with U^B from the original SD model, using the same style-intervention prompt. As shown in Fig. 7, images generated using ControlNet trained directly with U^B on D_0 (top row) with the prompt "*sagittal lumbar spine MRI*" retain characteristics of CT modality rather than MRI. However, these images appear darker than typical CT images, which is more characteristic of MRI, indicating partial intervention on the style variable.

Conversely, images generated using the ControlNet trained with instance-fine-tuned U^{D_0} (bottom row) exhibit the appearance of MRI images rather than CT, demonstrating the effectiveness of style variable intervention through prompting when using the Style Swap technique.

C. Visualizations of Predicted Segmentation Masks

We present additional visualizations of segmentation results from our method and other domain generalization (DG) approaches for the AS, LSS and LS tasks in Fig. 8, Fig. 9 and Fig. 10, respectively. The first two columns display source and target domain images, illustrating the domain shift. Consistent with our main findings, these visualizations demonstrate that our method consistently produces superior segmentation masks in the unseen target domain. Notably, the generated masks exhibit not only higher accuracy but also enhanced spatial continuity of the foreground classes.

D. Visualizations of Diffusion-generated Augmentations

To demonstrate the effectiveness of our diffusion-based style intervention, we provide visualizations of images generated by our controlled diffusion model for the AS, LSS and LS tasks in Fig. 11, Fig. 12 and Fig. 13, respectively. These visualizations illustrate that our diffusion-based augmentation strategy successfully leverages the rich generative prior of the SD model, which has been trained on a large-scale medical imaging dataset comprising scans from diverse anatomies and imaging modalities. Our method performs comprehensive intervention on the style variable while preserving the content. These strong augmentations are crucial for extracting content features using contrastive learning.

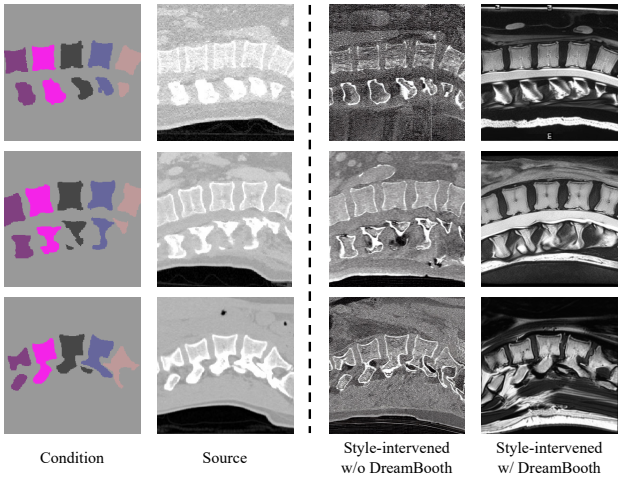

Figure 7. Visualization results on diffusion-based style intervention for LSS task (CT→MRI). The first and second column show the image conditions (*i.e.*, segmentation masks) and the corresponding source images, respectively. The third column shows the style-intervened images using the ControlNet trained direcly with pretrained SD U-Net (*i.e.*, no instance fine-tuning stage using DreamBooth). The forth column shows the style-intervened images using the ControlNet trained with instance-fine-tuned U-Net. Both of them are generated using the style-intervention prompt "*sagittal lumbar spine MRI*".

E. Dataset And Preprocessing Details

We evaluate our method on three cross-modality segmentation tasks: AS, LSS and LS tasks. For the AS task, we have 20/10 volumes for train/test CT data and 40/20 volumes for train/test MRI data. For the LSS task, we randomly split all datasets into approximately 90% train / 10% test, resulting in 117/15 train/test volumes for CT , 182/20 volumes for MRI and 350/50 images for X-Ray. For the LS task, we use the split proposed in the original paper [63] for CT data, which contains 36 volumes for training and 24 volumes for testing. For X-Ray scans, we first perform quality control (QT) on both images and labels. We filter out the cases where images are not X-Ray scans (*e.g.*, some scans are axial slices of CT scans) and/or labels contains no foreground or a rectangle foreground. After the QT, we randomly split the dataset into approximately 90% train / 10% test, resulting in 5409/601 train/test images. We adopt the preprocessing steps from [40] for all 3D volume data.

The common augmentations applied to all methods include affine transformations, elastic deformations, brightness and contrast adjustments, gamma corrections, and additive Gaussian noise.

F. Implementation Details

F.1. Loss Hyperparameters

The hyperparameters for weighting different losses across all methods are presented in Tab. 7. All the parameters are selected based on a grid search over the range indicated in the "*values*" column.

Table 7. Summary of hyperparameter optimization ranges for each method.

Method	Parameter	Values	Description
Ours	λ_{src}	$\{0.1, 1\}$	Source image seg. loss
	λ_{sty}	$\{0.1, 1\}$	Style-intervened image seg. loss
	$\lambda_{contrast}$	$\{1, 10\}$	InforNCE regularization
RandConv [62]	λ_{src}	$\{0.1, 0.5, 1\}$	Source image seg. loss
	λ_{aug}	$\{0.1, 0.5, 1\}$	Augmented image seg. loss
CSDG [40]	λ_{src}	$\{0.1, 0.5, 1\}$	Source image seg. loss
	λ_{aug}	$\{0.1, 0.5, 1\}$	Augmented image seg. loss
	λ_{kl}	$\{1, 10\}$	KL-DIV regularization
SLAug [56]	λ_{gla}	$\{0.1, 0.5, 1\}$	Global location-scale augmentation seg. loss
	λ_{sbf}	$\{0.1, 0.5, 1\}$	Saliency-balancing fused augmentation seg. loss

F.2. Baselines

MIXSTYLE: We implement MixStyle [70] using the authors' provided code[1]. To accommodate MixStyle's requirement for multiple source domains, we employ our diffusion-based augmentation to synthesize novel training domains with varied appearances across imaging modalities. We adopt the hyperparameters from the original implementation: mixing probability $p = 0.5$, Beta distribution parameter $\alpha = 0.1$, and "random" mixing method. Style mixing is applied twice after the first and second double-convolution blocks of the U-Net [48] encoder.

DUALNORM: We use the original implementation of Dual-Normalization [71][2]. For each training image, we generate two source-similar and three source-dissimilar augmentations using nonlinear Bézier transformations. We use the default control points to define the Bézier curves for both source-similar and source-dissimilar augmentations, with 1000 time steps in both cases.

CSDG: We adopt the model architecture from the original CSDG implementation [40][3]. For the global intensity non-linear augmentation (GIN) module, we use 4 convolutional layers and 2 intermediate layers, with "Frobenius" normalization after the final layer. The interventional pseudo-correlation augmentation (IPA) module uses blending parameters $\epsilon = 0.3$ and $\xi = 1e^{-6}$, with control point spacing of 32, downsample scale of 2, and interpolation order of 2. Control point parameters are initialized using a Gaussian distribution.

RANDCONV: We implement RandConv [62] based on the CSDG implementation [40], removing the IPA module to generate augmentations using only randomly-initialized convolutional layers. We adopt the same GIN module hyperparameters as CSDG.

SLAUG: We use the model architecture and augmentation from the original SLAug implementation [56][4]. For the non-linear Bézier transformation, we set the background threshold to 0.01 and use 100,000 time steps. We initialize 4 control points based on the input image's intensity values, and gradually add another two points by uniformly sample a random value between the first and last elements of the current point array. The probability of random inversion is set to 0.5. For saliency-based fusion, we use a 2D B-spline kernel with interpolation order 2 and grid size 3.

F.3. Diffusion-based Augmentation

We implement DreamBooth [49] and ControlNet [65] using the HuggingFace Diffusers library[5]. During the sampling process, we adopt the UniPC sampler to accelerate the generation with 20 steps [69].

[1] https://github.com/KaiyangZhou/mixstyle-release

[2] https://github.com/zzzqzhou/Dual-Normalization

[3] https://github.com/cheng-01037/Causality-Medical-Image-Domain-Generalization

[4] https://github.com/Kaiseem/SLAug?tab=readme-ov-file

[5] https://github.com/huggingface/diffusers

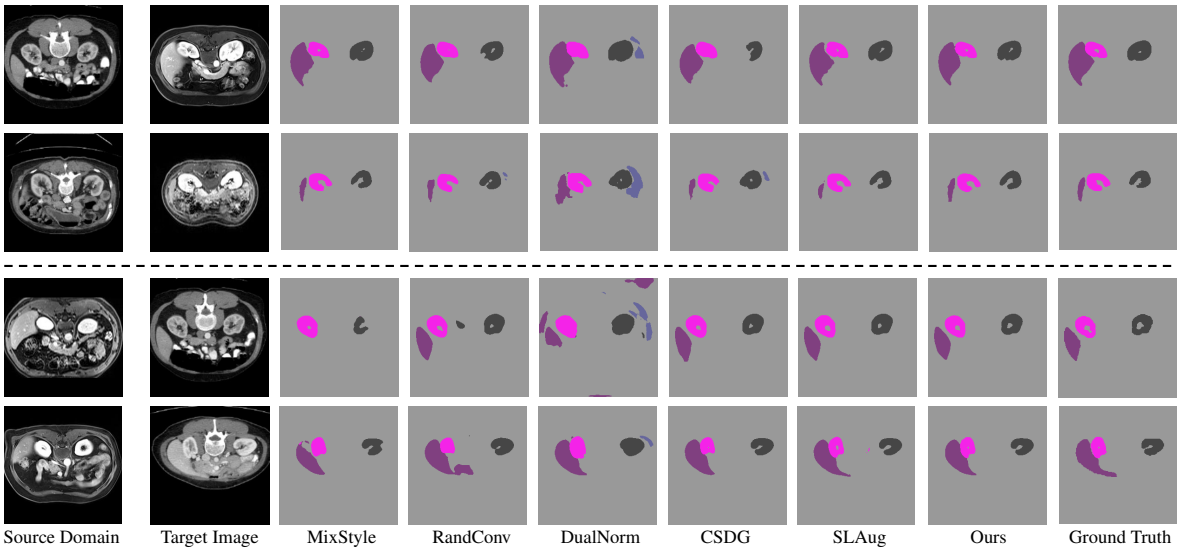

Figure 8. Additional visualization results on AS task of different methods. First two rows: "CT to MRI" task; Last two rows: "MRI to CT" task.

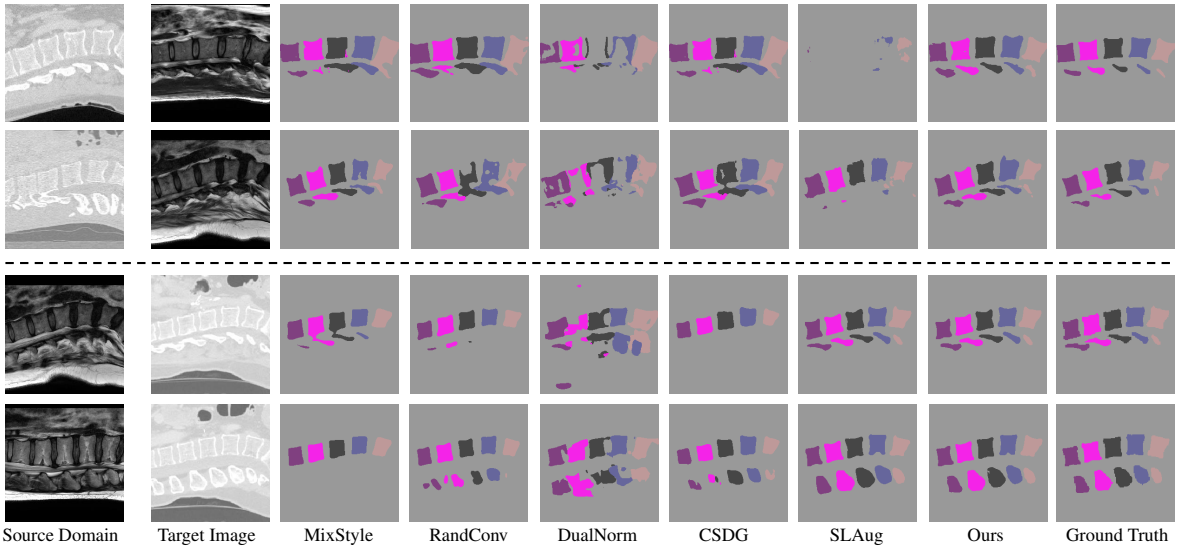

Figure 9. Additional visualization results on LSS task of different methods. First two rows: "CT to MRI" task; Last two rows: "MRI to CT" task.

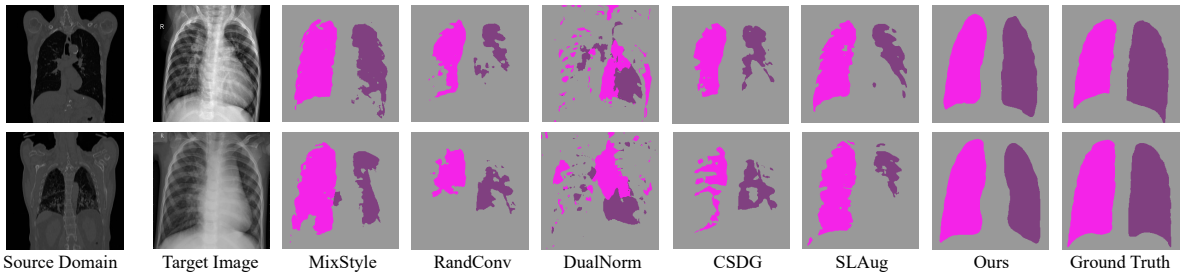

Figure 10. Additional visualization results on LS task of different methods. First two rows: "CT to X-Ray" task.

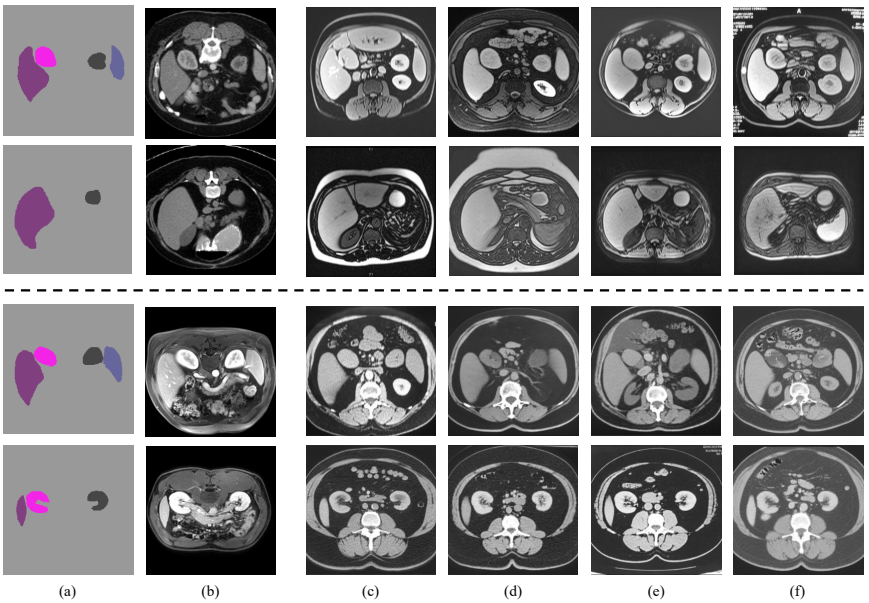

Figure 11. Visualization results of diffusion-based style intervention for the AS task. The top two rows show the "CT to MRI" task, while the bottom two rows display the "MRI to CT" task. The left columns present (a) the image conditions (*i.e.*, segmentation masks) and (b) the corresponding source images. The right columns (c-f) show the style-intervened images generated using specific style-intervention prompts. For the first row: "*axial liver left kidney right kidney spleen MRI*"; second row: "*axial liver spleen MRI*"; third row: "*axial liver left kidney right kidney spleen CT*"; and fourth row: "*axial liver left kidney right kidney CT*".

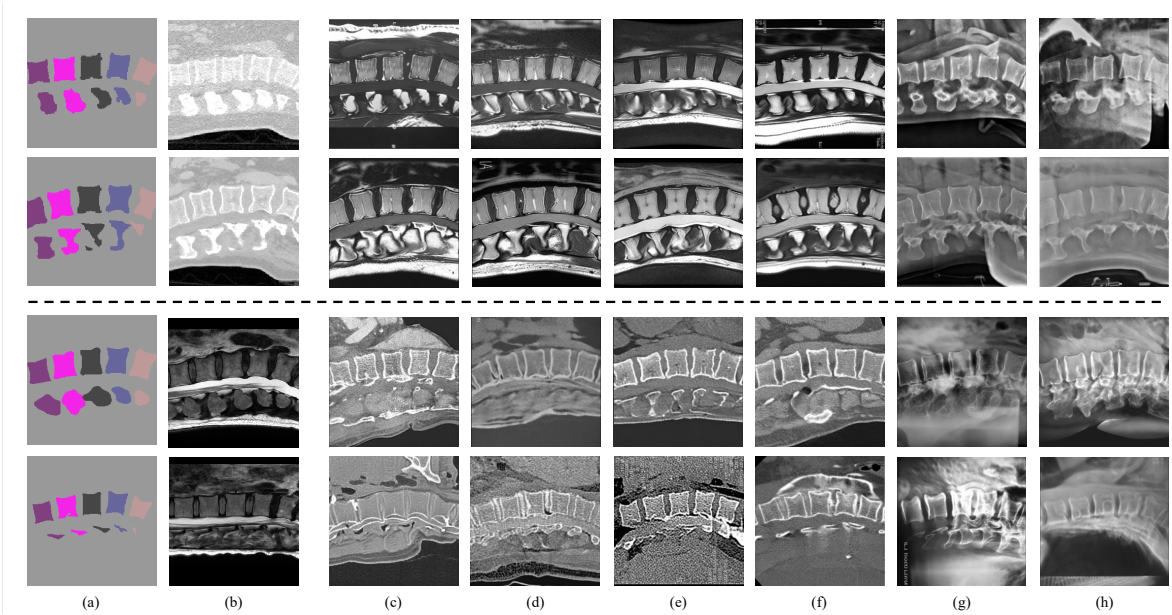

Figure 12. Visualization results of diffusion-based style intervention for the LSS task. The first two rows demonstrate the "CT to MRI" task, while the last two rows show the "MRI to CT" task. Columns on the left display (a) the image conditions (i.e., segmentation masks) and (b) the corresponding source images, respectively. For the first two rows (CT to MRI task), the right columns show style-intervened images generated using the following prompts: (c-d) "*sagittal lumbar spine T1-MRI*"; (e-f) "*sagittal lumbar spine T2-MRI*"; and (g-h) "*sagittal lumbar spine X-ray*". For the last two rows (MRI to CT task), the right columns present style-intervened images generated using these prompts: (c-f) "*sagittal lumbar spine CT*"; and (g-h) "*sagittal lumbar spine X-ray*".

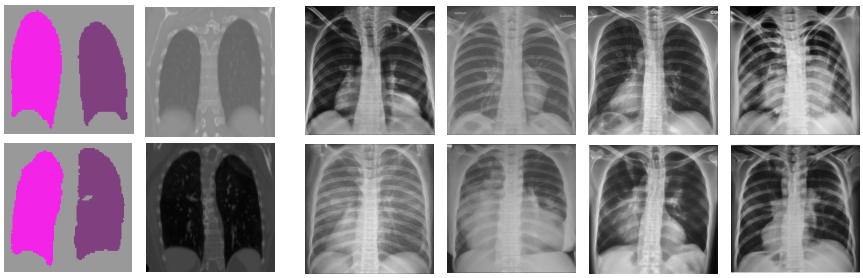

Figure 13. Visualization results of diffusion-based style intervention for the LS task. The top two rows show the ”CT to X-Ray” task. The left columns present (a) the image conditions (*i.e.*, segmentation masks) and (b) the corresponding source images. The right columns (c-f) show the style-intervened images generated using style-intervention prompt:”*chest CT left lung right lung*”.