SAFE-DIFF: WHERE ANATOMY MEETS FREQUENCY IN VIRTUAL CONTRAST ENHANCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Angiography is a contrast-enhanced imaging technique that improves the visualization of vascular structures, facilitating the detection and assessment of vascular diseases. However, many angiographic examinations rely on contrast agents that may pose risks to susceptible patients. Virtual angiography synthesis from non-contrast scans offers a promising alternative, yet existing medical image translation methods often struggle to preserve fine vascular anatomy, resulting in blurred boundaries, distorted vessels, or spurious structures. To address these challenges, we propose SAFE-Diff, a segmentation-conditioned and frequency-aware diffusion framework for paired cross-modality medical image synthesis. Specifically, SAFE-Diff combines cross-modal semantic alignment to provide global semantic guidance with anatomy-aware conditioning and frequency-aware feature modulation to preserve local anatomical structures during diffusion. We evaluate SAFE-Diff on 1,600 paired non-contrast CT and CTA volumes and a brain T1-to-T1ce synthesis, and the result shows that SAFE-Diff achieves the best overall performance across image fidelity metrics, including 25.1 dB PSNR, 0.77 SSIM, and 0.26 LPIPS on CT-to-CTA synthesis, and 29.7 dB PSNR, 0.87 SSIM, and 0.19 LPIPS on T1-to-T1ce synthesis. Beyond improvements in image fidelity, SAFE-Diff better preserves anatomically meaningful structures, achieving the highest performance in downstream anatomical evaluation. These results demonstrate that jointly modeling cross-modal semantics and anatomical guidance enables more faithful virtual contrast-enhanced image synthesis. Code will be released upon acceptance.

1 INTRODUCTION

Contrast-enhanced imaging is a cornerstone of modern radiology, enabling high-contrast visualization of vascular structures for the diagnosis of vascular diseases such as coronary artery disease, atherosclerosis, and aneurysms (Tsao et al., 2023; Wang et al., 2024b). However, acquiring contrast-enhanced images typically requires iodinated or gadolinium-based contrast agents and, for CTA and XRA, exposure to ionizing radiation, introducing risks of allergic reactions, nephrotoxicity, and cumulative radiation burden (Dorjsembe et al., 2024; Güngör et al., 2023). These limitations have motivated growing interest in *virtual contrast enhancement* (VCE), which aims to synthesize contrast-enhanced images directly from non-contrast scans, offering a promising strategy to reduce contrast-agent usage and radiation exposure while preserving diagnostic information.

Recent advances in generative models, particularly diffusion models, have substantially improved cross-modality medical image synthesis by learning realistic target-domain appearance from paired or unpaired data (Dar et al., 2019; Rombach et al., 2022; Croitoru et al., 2023; Özbey et al., 2023). However, visually realistic images do not necessarily imply anatomically faithful synthesis. Existing generative models primarily optimize target-domain distributions, which may introduce, remove, or spatially alter clinically relevant structures while still producing visually convincing results (Cohen et al., 2018). This limitation is particularly critical for VCE, where vessels, small branches, and enhancing boundaries occupy only a small fraction of the image volume. Consequently, the key challenge is to generate anatomically faithful contrast enhancement while preserving patient-specific structures across modalities.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

These limitations stem from two fundamental characteristics of virtual contrast enhancement. First, the substantial appearance discrepancy between non-contrast and contrast-enhanced images introduces a large cross-modal semantic gap, making target-domain enhancement difficult to infer directly from source intensities and limiting the effectiveness of pixel-level supervision alone. Second, diffusion models generate images through progressive global denoising without explicitly constraining local anatomical consistency throughout the reverse process. Furthermore, fine anatomical structures and image noise often exhibit similar high-frequency characteristics, making selective preservation of clinically meaningful details particularly challenging. Although recent studies have explored anatomical priors or frequency-aware modeling independently (Konz et al., 2024b; Li et al., 2024; Wang et al., 2025), denoisers struggle to adapt spectral responses to patient anatomy and distinguish fine structures from high-frequency noise. Coupling such anatomy-aware spectral modulation with paired cross-modal alignment remains underexplored.

To address these challenges, we propose **SAFE-Diff**, a segmentation-conditioned, frequency-aware diffusion framework for anatomy-preserving virtual contrast enhancement. Rather than relying solely on image-level supervision, SAFE-Diff progressively constrains the diffusion process through complementary semantic, structural, and spectral guidance. At the semantic level, dual modality-specific variational autoencoders together with a contrastive projection objective learn a shared latent representation that reduces the cross-modal semantic gap and provides robust conditioning for image generation. At the structural level, segmentation masks and thin anatomical outlines are injected into every denoising step, explicitly constraining image generation to preserve patient-specific anatomy while allowing realistic contrast enhancement. At the spectral level, we introduce a lightweight multi-band spectral attention (MBSA) module that adaptively reweights bottleneck features according to their frequency responses, emphasizing anatomically meaningful high-frequency structures while suppressing noise-like fluctuations. Together, these three components enable SAFE-Diff to jointly model cross-modal correspondence, anatomical consistency, and fine structural detail within a unified diffusion framework.

We conduct extensive experiments on large-scale CT-to-CTA and brain MRI datasets. SAFE-Diff consistently outperforms representative UNet-, GAN-, and diffusion-based baselines across both image-quality and anatomy-aware metrics, while comprehensive ablation studies further validate the contribution of each proposed component.

Our main contributions are summarized as follows:

- We formulate an overlooked anatomy–frequency coupling problem in structure-preserving diffusion. Anatomical guidance and frequency-aware modulation are typically designed as separate mechanisms, preventing the denoiser from adapting its spectral response to patient-specific anatomy. Consequently, fine anatomical details and noise-like high-frequency responses are not explicitly distinguished.
- We introduce an anatomy-conditioned Multi-Band Spectral Attention (MBSA) module that partitions 3D bottleneck spectra into learnable radial frequency bands, extracts multi-band energy descriptors, and fuses them with patient-specific boundary context to generate channel-wise recalibration weights. This makes frequency-aware feature modulation explicitly dependent on local anatomy rather than global spectral statistics alone.
- We further develop a hierarchical conditioning framework that combines paired cross-modal semantic alignment for global source–target correspondence with anatomy-conditioned spectral modulation for local structural preservation. Experiments on 3D cardiac CT-to-CTA and brain T1-to-T1ce synthesis demonstrate consistent improvements in image fidelity, with additional gains in downstream cardiac and vascular segmentation on the CT-to-CTA task.

2 Related Work

Medical image synthesis with diffusion. Diffusion models have become a strong paradigm for high-fidelity image generation because of their stable training and mode coverage (Fan et al., 2025; Dar et al., 2019; Özbey et al., 2023; Kazerouni et al., 2023; Gao et al., 2024b). In medical imaging, controllability is especially important, since spatial fidelity and anatomical consistency must be preserved. Recent methods therefore condition diffusion on structural priors such as organ masks,

contours, or frequency cues. For example, SegGuidedDiff (Konz et al., 2024a) conditions each reverse step on anatomical masks, while frequency-guided diffusion (Li et al., 2023b) separates low- and high-frequency pathways to retain fine structures. Surveys also highlight the robustness of diffusion models under distribution shifts and limited supervision (Kazerouni et al., 2023; Wang et al., 2024a).

Paired vs. unpaired translation. Supervision critically affects medical image translation. Unpaired methods such as CycleGAN-style models (Wang et al., 2023), StarGAN v2 (Choi et al., 2020), SynDiff (Özbey et al., 2023), and ContourDiff (Chen et al., 2024) rely on cycle consistency, adversarial constraints, or contour guidance to preserve content without spatially aligned pairs. They are valuable when paired data are unavailable, but solve a harder problem and may be less accurate when correspondence exists. Paired methods such as pix2pix (Isola et al., 2017), ResViT (Dalmaz et al., 2022), RegGAN (Kong et al., 2021), and BBDM (Li et al., 2023a) exploit voxel-level correspondence through adversarial, transformer-based, registration-aware, or Brownian-bridge formulations. Since CT/CTA scans are naturally paired within the same study, SAFE-Diff is designed as a paired method and is evaluated against both paired and unpaired baselines to avoid confounding gains with supervision advantages.

Contrast and modality transfer. Several studies have explored contrast generation beyond general modality translation. Classical CT contrast enhancement often relies on (cycle-)GANs (Choi et al., 2021; Han et al., 2024), but such methods can suffer from mode collapse or anatomical drift, especially in coronary imaging (Lyu et al., 2023). Diffusion-based methods have shown promise: DCE-Diff (Kishore Kumar et al., 2024) synthesizes DCE-MRI from non-contrast inputs, APT (Shin et al., 2025) preserves global and local structure via mutual-information fusion, conditional DDPMs generate synthetic dual-energy CT (Gao et al., 2024a), and outline-guided diffusion stabilizes vascular and organ boundaries (Fan et al., 2025). However, systematic paired CT-to-CTA synthesis that uses CT as a spatial condition remains underexplored. SAFE-Diff addresses this gap by jointly combining case-level cross-modal latent alignment, denoising-time segmentation/outline conditioning, and frequency-domain bottleneck modulation to reduce the CT–CTA modality gap while preserving fine anatomical structures.

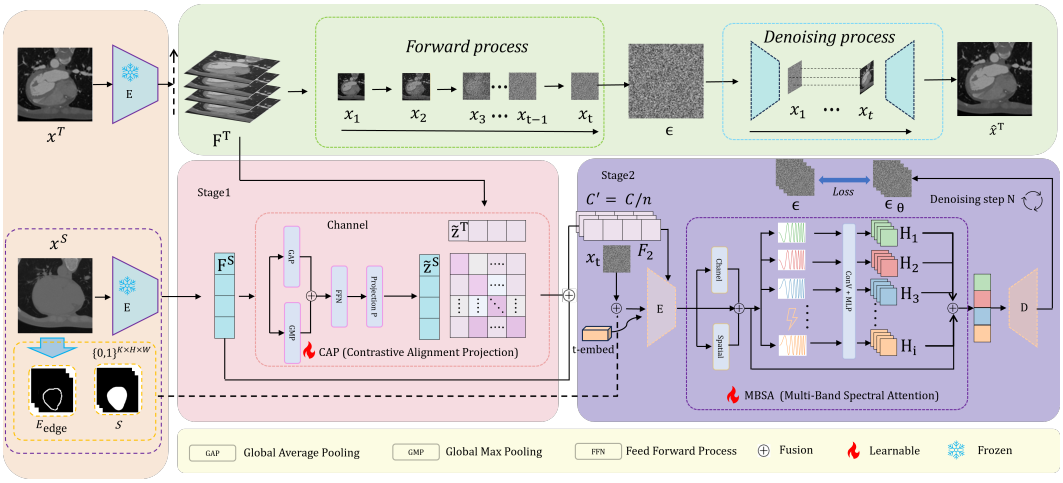


Figure 1: Overview of the proposed framework. Stage 1 learns a cross-modal latent space using a dual VAE encoder and a contrastive projection head, aligning source (CT/MR) and target (CTA/T1ce) features at the case level. Stage 2 trains a conditional diffusion model that takes the noisy target x_t concatenated with segmentation masks S and outline maps E_{edge}, and is further conditioned on the aligned latent features. A multi-band spectral attention (MBSA) module operates on UNet bottleneck features in the frequency domain to emphasize structurally meaningful high-frequency components.

3 METHOD

Fig. 1 illustrates SAFE-Diff, which aims to synthesize $\hat{\mathbf{x}}^T$ from a source volume $\mathbf{x}^S$ while preserving patient-specific anatomy. Paired contrast-enhanced targets $\mathbf{x}^T$ are used during training; inference requires only the source volume.

To address cross-modal semantic discrepancy and anatomical inconsistency, SAFE-Diff combines representation alignment with anatomically conditioned generation. Stage 1 aligns modality-specific VAE features in a shared embedding space through a contrastive projection head. Stage 2 performs diffusion in an independently learned target-domain latent space, using source semantics and structural priors as conditions. Multi-band spectral attention (MBSA) further modulates bottleneck features using spectral and anatomical context.

Let $\mathbf{x}^S \in \mathbb{R}^{C_S \times D \times H \times W}$ and $\mathbf{x}^T \in \mathbb{R}^{C_T \times D \times H \times W}$ denote the paired volumes. A pretrained nnU-Net (Isensee et al., 2021) predicts the source segmentation mask $\mathbf{S}$. Its class-wise outline is $\mathbf{E}_{edge} = \mathrm{Edge}(\mathbf{S})$, where Edge is a morphological boundary operator and $\mathbf{S}, \mathbf{E}_{edge} \in \{0,1\}^{K \times D \times H \times W}$. Here, K denotes the number of task-specific anatomical classes.

3.1 STAGE 1: CONTRASTIVE ALIGNMENT PROJECTION (CAP)

Cross-modal appearance differences can obscure anatomical correspondence even for paired volumes from the same patient. CAP therefore aligns their semantic embeddings while retaining modality-specific feature representations.

Modality-specific representations. We independently pretrain two VAEs, $(\mathcal{E}_S, \mathcal{D}_S)$ and $(\mathcal{E}_T, \mathcal{D}_T)$, to capture the characteristics of each modality. Their spatial features are $\mathbf{F}^m = \mathcal{E}(\mathbf{x}^m)$, $m \in \{S, T\}$, where $\mathcal{E}$ denotes the corresponding modality-specific encoder. After pretraining, both encoders and decoders remain frozen throughout subsequent training.

Cross-modal semantic alignment. Modality-specific reconstruction alone does not establish correspondence between paired source and target representations. We therefore apply global average pooling, a trainable projection head $\mathcal{P}$, and ℓ_2 normalization:

$$\mathbf{Z}^m = \mathrm{Norm}_2\left(\mathcal{P}\left(\mathrm{Pool}(\mathbf{F}^m)\right)\right), \qquad m \in \{S, T\}, \tag{1}$$

where $\mathbf{Z}^m$ is the case-level semantic embedding. For a minibatch of N paired volumes, same-patient embeddings form positive pairs, whereas cross-patient pairs serve as negatives. We optimize $\mathcal{P}$ using the symmetric InfoNCE loss (Oord et al., 2018):

$$\begin{aligned}\mathcal{L}_{\mathrm{NCE}} = -\frac{1}{2N}\sum_{i=1}^{N}\Bigg[&\log\frac{\exp(\mathrm{sim}(\mathbf{Z}_i^S, \mathbf{Z}_i^T)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(\mathbf{Z}_i^S, \mathbf{Z}_j^T)/\tau)} \\ &+ \log\frac{\exp(\mathrm{sim}(\mathbf{Z}_i^T, \mathbf{Z}_i^S)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(\mathbf{Z}_i^T, \mathbf{Z}_j^S)/\tau)}\Bigg],\end{aligned} \tag{2}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is the temperature parameter. After optimization, $\mathcal{P}$ is frozen, and both source spatial features and the case-level embedding $\mathbf{Z}^S$ condition Stage 2.

3.2 STAGE 2: ANATOMY- AND FREQUENCY-AWARE LATENT DIFFUSION

Semantic correspondence alone does not explicitly enforce local geometry during stochastic latent denoising (Konz et al., 2024b). Thin vessels and anatomical boundaries are particularly sensitive to spatial distortions. We therefore complement semantic conditioning with source-derived priors that provide region- and boundary-level anatomical guidance.

During reverse diffusion, cross-attention receives both source spatial features and the case-level semantic embedding $\mathbf{Z}^S$. The segmentation mask $\mathbf{S}$ identifies anatomical regions, while $\mathbf{E}_{edge}$ emphasizes their boundaries. Both priors are downsampled to the latent resolution and concatenated with the noisy target latent:

$$\mathbf{F}_2 = \mathrm{Concat}\left[\mathbf{z}_t^T, \mathbf{S}, \mathbf{E}_{edge}\right]. \tag{3}$$

Here, $\mathbf{z}_t^T$ denotes the target latent at diffusion step t; $\mathbf{S}$ and $\mathbf{E}_{edge}$ denote their downsampled versions within this concatenation.

3.3 MULTI-BAND SPECTRAL ATTENTION

Anatomical priors identify where structures should be preserved, but do not characterize the frequency content of denoising features. High-frequency responses may encode fine structures or noise-like fluctuations, motivating feature modulation that also incorporates anatomical context. MBSA therefore combines adaptive spectral descriptors with outline-derived context to recalibrate UNet bottleneck channels.

Learnable frequency bands. Let $\mathbf{H}_t \in \mathbb{R}^{B_s \times C \times D' \times H' \times W'}$ denote the bottleneck feature at diffusion step t. Here, B_s and C are the batch size and channel count. A channel-wise three-dimensional discrete cosine transform (DCT) (Ahmed et al., 1974) yields $\widehat{\mathbf{H}}_t = \mathcal{T}^{\mathrm{3D}}_{\mathrm{DCT}}(\mathbf{H}_t)$. The normalized radial frequency at coordinate (p, q, r) is

$$\rho_{p,q,r} = \frac{1}{\sqrt{3}} \sqrt{\left(\frac{p}{D'-1}\right)^2 + \left(\frac{q}{H'-1}\right)^2 + \left(\frac{r}{W'-1}\right)^2} \in [0, 1]. \tag{4}$$

To adapt the bands to feature responses while preventing collapsed intervals, we parameterize three positive band widths:

$$\Delta_k = \delta_{\min} + (1 - 3\delta_{\min}) \frac{\exp(a_k)}{\sum_{j=1}^{3} \exp(a_j)}, \qquad k \in \{1, 2, 3\}, \tag{5}$$

where a_k are learnable parameters and $0 < \delta_{\min} < 1/3$. The resulting boundaries are $\tau_1 = \Delta_1$ and $\tau_2 = \Delta_1 + \Delta_2$. For differentiable boundary learning, we define $s_k(\rho) = \sigma((\tau_k - \rho)/\eta)$, $k \in \{1, 2\}$, where σ is the sigmoid function and $\eta > 0$ controls smoothness. The low-, mid-, and high-frequency masks are

$$\begin{aligned} &\Omega_1(\rho) = s_1(\rho), \qquad \Omega_2(\rho) = s_2(\rho) - s_1(\rho), \\ &\Omega_3(\rho) = 1 - s_2(\rho), \end{aligned} \tag{6}$$

which satisfy $\sum_{k=1}^{3} \Omega_k(\rho) = 1$.

Spectral descriptors and structural gating. To compare bands with different bandwidths, we summarize each by its normalized spectral energy:

$$e_k(b) = \frac{\sum_{c,p,q,r} \Omega_k(\rho_{p,q,r}) \left| \widehat{\mathbf{H}}_{t,b,c,p,q,r} \right|^2}{C \sum_{p,q,r} \Omega_k(\rho_{p,q,r}) + \varepsilon}, \qquad k \in \{1, 2, 3\}, \tag{7}$$

where b indexes the samples and ε ensures numerical stability. Log compression reduces the dynamic range of the energies, yielding the descriptor $\mathbf{y}(b) = [\log(\varepsilon + e_1(b)), \log(\varepsilon + e_2(b)), \log(\varepsilon + e_3(b))] \in \mathbb{R}^3$.

Spectral energy alone does not identify the anatomical relevance of a feature response. We therefore project the downsampled outline map $\widetilde{\mathbf{E}}_{\mathrm{edge}}$ to the bottleneck channel dimension and apply global average pooling. This structural descriptor complements the spectral descriptor in computing channel-wise gates:

$$\begin{aligned} \mathbf{w} &= \sigma\left(\mathbf{y}\mathbf{W}_{\mathrm{spec}}^{\top} + \mathbf{b}_{\mathrm{spec}} + \mathrm{GAP}\left(\phi_{\mathrm{edge}}(\widetilde{\mathbf{E}}_{\mathrm{edge}})\right)\right), \\ \widetilde{\mathbf{H}}_t &= \mathbf{w} \odot \mathbf{H}_t. \end{aligned} \tag{8}$$

Here, $\mathbf{W}_{\mathrm{spec}} \in \mathbb{R}^{C \times 3}$, $\mathbf{b}_{\mathrm{spec}} \in \mathbb{R}^C$, and $\mathbf{w} \in (0, 1)^{B_s \times C}$. The gates are broadcast across spatial dimensions, and the recalibrated features $\widetilde{\mathbf{H}}_t$ are passed to the UNet decoder for noise prediction.

4 EXPERIMENTS

Datasets and Preprocessing We evaluate SAFE-Diff on cardiac CT angiography synthesis and brain MRI synthesis.

Cardiac CT. The cardiac cohort is an in-house dataset comprising 1,600 retrospectively collected paired non-contrast and contrast-enhanced CT volumes from unique patients acquired using Siemens CT scanners under routine clinical protocols. The dataset was fully anonymized and used under IRB approval with appropriate consent (consent waived where applicable). Patients are split into training, validation, and testing sets with a ratio of 7:1.5:1.5, resulting in 1,120, 240, and 240 cardiac CT cases, respectively. The validation set is used for model selection, and all results are reported on the held-out test set.

Brain MRI. We additionlly evaluate our method on the BraTS 2016/2017 datasetsplits (Menze et al., 2014; Bakas et al., 2017), following the original licensing and splits.

Preprocessing. All volumes are resampled to an isotropic resolution of 1.1 mm (trilinear for images, nearest-neighbour for labels). For the cardiac dataset, anatomical priors are generated using TotalSegmentator v2 (Wasserthal et al., 2023). We retrain the cardiac chambers and great vessels, remap them into K task-specific classes, and convert the resulting masks into one-hot representations after minor morphological post-processing. For BraTS, anatomical priors are predicted from T1 alone using a frozen nnU-Net trained exclusively on training-set subjects. Neither held-out T1ce images nor their annotations are used to construct these priors.

Evaluation Our primary task is angiography synthesis. We evaluate image quality using PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). To further assess anatomical fidelity, we evaluate downstream segmentation performance (Tab. 2) and robustness under degraded imaging conditions (Tab. 3).

Implementation Details. Stage 1 is trained on the VAE-GAN backbone of COLA (Jiang et al., 2023). The VAE is first pretrained for image reconstruction, after which the contrastive alignment module is optimized using AdamW with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. The trained encoder and projection head are frozen during Stage 2. Stage 2 adopts a 3D latent diffusion model based on COLA (Jiang et al., 2023). Volumes are resampled to $256 \times 144 \times 144$ and cropped into fixed-size 3D patches. The diffusion model is optimized using AdamW with a learning rate of 1×10^{-5} and a linear noise schedule with $T = 200$ diffusion steps

All experiments are implemented in PyTorch with MONAI and trained on NVIDIA H100 GPUs using mixed-precision distributed training. Unless otherwise specified, all methods use identical data splits, preprocessing, optimizer settings, and sampling protocols. Further details of data preprocessing are provided in Supplementary Sec. S2.

Table 1: Quantitative comparison with representative unpaired and paired baselines on cardiac CT→CTA and brain T1→T1ce synthesis. PSNR and SSIM are higher-is-better, while LPIPS is lower-is-better. Best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Supervision	Cardiac CT→CTA			Brain T1→T1ce		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
StarGAN v2 (Choi et al., 2020)	Unpaired	20.0	0.60	0.40	18.0	0.78	0.45
DC-CycleGAN (Wang et al., 2023)	Unpaired	22.0	0.63	0.35	22.4	0.73	0.29
SynDiff (Özbey et al., 2023)	Unpaired	23.5	0.65	0.33	24.6	0.71	0.27
ContourDiff (Chen et al., 2024)	Unpaired	21.7	0.55	0.42	19.3	0.55	0.35
cGAN (Dar et al., 2019)	Paired	21.2	0.62	0.39	27.4	**0.87**	0.22
pix2pix (Isola et al., 2017)	Paired	24.3	0.66	0.29	28.3	0.77	0.22
ResViT (Dalmaz et al., 2022)	Paired	22.8	0.66	0.34	<u>28.9</u>	<u>0.83</u>	<u>0.20</u>
BBDM (Li et al., 2023a)	Paired	20.2	0.49	0.48	21.3	0.39	0.46
RegGAN (Kong et al., 2021)	Paired	<u>24.6</u>	<u>0.72</u>	<u>0.27</u>	28.7	0.78	<u>0.20</u>
SAFE-Diff (Ours)	Paired	**25.1**	**0.77**	**0.26**	**29.7**	**0.87**	**0.19**

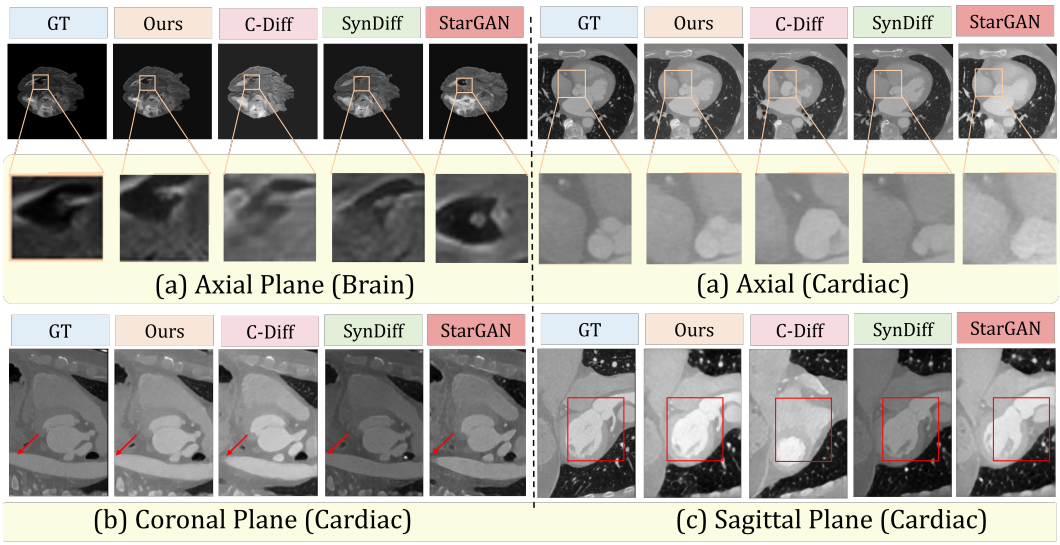


Figure 2: Synthesized contrast on BraTS brain MRI and cardiac CT. From left to right in each block: Ground Truth, Ours, ContourDiff (Chen et al., 2024), SynDiff (Özbey et al., 2023), StarGAN v2 (Choi et al., 2020).

4.1 Comparison with State of the Art

We compare SAFE-Diff with representative competing methods under both unpaired and paired supervision regimes. The unpaired baselines include GAN-based methods, StarGAN v2 (Choi et al., 2020) and DC-CycleGAN (Wang et al., 2023), as well as diffusion-based methods, SynDiff (Özbey et al., 2023) and ContourDiff (Chen et al., 2024). For paired translation, we include baselines spanning adversarial, transformer-based, registration-aware, and diffusion-based formulations: cGAN (Dar et al., 2019), pix2pix (Isola et al., 2017), ResViT (Dalmaz et al., 2022), RegGAN (Kong et al., 2021), and BBDM (Li et al., 2023a). All paired methods, including SAFE-Diff, are trained and evaluated using the same paired data, pre-processing pipeline, and splits, so that the comparison reflects modeling differences rather than differences in supervision or data access.

Quantitative results. Table 1summarizes the quantitative comparison. SAFE-Diff achieves the best performance across all evaluation metrics on both datasets. On cardiac CT→CTA synthesis, SAFE-Diff achieves 25.1 dB PSNR, 0.77 SSIM, and 0.26 LPIPS. On brain T1→T1ce synthesis, it reaches 29.7 dB, 0.87, and 0.19, respectively.

Compared with the strongest paired baselines, SAFE-Diff consistently improves both pixel-wise fidelity and perceptual quality. In particular, SAFE-Diff improves RegGAN by 0.5 dB PSNR, 0.055 SSIM, and 0.01 LPIPS on the cardiac dataset, and outperforms ResViT by 0.8 dB PSNR together with a 0.04 improvement in SSIM on the brain MRI dataset. These consistent gains demonstrate the effectiveness of combining semantic alignment, anatomical conditioning, and frequency-aware feature modulation.

Visual comparison. Fig. 2 provides representative qualitative results in the axial, coronal, and sagittal views. Unpaired methods tend to produce visually plausible contrast but suffer from local anatomical inconsistencies, including discontinuous vessels, blurred vascular boundaries, and occasional false enhancement. Although paired methods better preserve the global layout, they frequently over-smooth small branches or distort vessel caliber around complex anatomical structures. In contrast, SAFE-Diff more faithfully reproduces both the global enhancement pattern and local vascular geometry. In particular, it better preserves thin branches, sharper vessel boundaries, and more consistent vessel calibers, while reducing artifacts around high-curvature or densely packed vascular structures.

Anatomical Fidelity Evaluation. Image similarity metrics do not reflect whether synthesized CTA preserves anatomically meaningful structures. We therefore evaluate downstream segmentation by

Table 2: Downstream anatomical evaluation on 55 expert-labelled CTA cases. Predictions from TotalSegmentator and nnU-Net on real and synthesized CTA are evaluated against expert annotations. Great-vessel evaluation covers the aorta and pulmonary arteries (PA); coronary metrics cover the three annotated coronary arteries. Dice and clDice are reported as percentages, and ASSD in millimetres. Real CTA provides a reference for the segmentation pipeline. Blank entries indicate pending evaluation.

Input	Cardiac	Great vessels		Coronary arteries		
	mDice↑	Aorta Dice↑	PA Dice↑	mDice↑	clDice↑	ASSD↓
Real CTA (reference)	89.9	82.2	79.4	64.2	55.0	1.2
ContourDiff	47.2	65.2	53.5	8.5	2.5	4.5
cGAN	59.5	69.3	56.4	7.3	1.7	9.9
SynDiff	65.2	60.9	49.9	13.2	8.2	6.1
ResViT	63.1	66.8	58.6	9.6	4.9	9.0
RegGAN	58.3	62.1	59.2	15.3	11.0	4.1
SAFE-Diff (Ours)	**74.4**	**70.3**	**62.1**	**33.2**	**21.5**	**3.8**

applying pre-trained TotalSegmentator and nnU-Net to synthesized CTA images from 55 expert-annotated cases. As shown in Table 2, SAFE-Diff achieves the highest cardiac and vessel mDice among all synthesis methods (74.4% and 33.2%). Although a performance gap remains compared with real CTA, the results demonstrate that SAFE-Diff better preserves anatomically consistent structures.

Table 3: Structural-prior ablation and robustness on cardiac CT→CTA. (a) S denotes segmentation and E_{edge} denotes the outline; all variants use the projection head and diffusion backbone. (b) Mask Dice measures overlap between degraded and clean priors. ΔPSNR and ΔSSIM are relative to the clean-prior setting. PSNR is measured in dB; downstream cardiac and coronary mDice are reported as percentages.

(a) Structural-prior ablation

Variant	Cond.	PSNR ↑	SSIM ↑
No structural prior	–	24.2	0.74
Seg only	S	24.3	0.75
Edge only	E_{edge}	24.5	0.75
Seg + Edge (ours)	$S + E_{\text{edge}}$	**25.1**	**0.77**

(b) Robustness to prior degradation

Mask prior	Mask Dice↑	ΔPSNR	ΔSSIM	Cardiac mDice↑	Coronary mDice↑
Clean	1.00	0.0	0.00	74.4	33.2
Erosion	0.85	−0.9	−0.04	71.5	24.8
Shift	0.78	−0.8	−0.07	68.2	21.3
Erosion+Shift	0.72	−1.3	−0.08	65.8	17.5

Robustness to Segmentation Quality. To evaluate robustness to imperfect anatomical priors, we degrade the input masks with 3D erosion and random translations to simulate imperfect segmentors. As shown in Table 3(b), SAFE-Diff degrades gracefully as the mask quality decreases. Under the most severe corruption (erosion + shift), SSIM decreases by only 0.08, while downstream cardiac segmentation drops by 8.6 mDice points. This suggests that SAFE-Diff is robust to moderate errors in the segmentation priors and does not rely exclusively on precise anatomical masks.

4.2 Ablation Study

Components ablation. Table 4 summarizes the quantitative ablation results. Removing any of the three proposed components consistently degrades synthesis performance, demonstrating that each contributes complementary information to the final model. Among them, removing the structural priors or MBSA leads to the largest performance drop, highlighting the importance of anatomical guidance and frequency-aware feature modulation during image synthesis. Although removing the cross-modal semantic alignment module results in a relatively smaller decrease, it still consistently improves both PSNR and SSIM, indicating that semantic alignment provides effective global guidance for cross-modal translation.

Table 4: Ablation of key components on the cardiac CT→CTA task. "Proj." denotes the contrastive projection head, "Struct." denotes segmentation and outline priors, and "MBSA" denotes multi-band spectral attention. Image fidelity is evaluated using PSNR, SSIM, and LPIPS. Anatomical evaluation reports Dice scores for the aorta and pulmonary arteries (PA), and ASSD for the three annotated coronary arteries (Cor.). Dice is reported as a percentage and ASSD in millimetres. Blank entries indicate pending evaluation.

Variant	Components			Image fidelity			Anatomical fidelity		
	Proj.	Struct.	MBSA	PSNR(dB)↑	SSIM↑	LPIPS↓	Aorta Dice↑	PA Dice↑	Cor. ASSD↓
Baseline				21.0	0.58	0.45	50.1	41.7	8.3
w/o CAP		✓	✓	23.8	0.69	0.33	59.2	48.3	6.9
w/o Struct.	✓		✓	24.2	0.74	0.36	52.6	49.0	7.2
w/o MBSA	✓	✓		24.0	0.75	0.39	57.4	44.6	7.7
Full model	✓	✓	✓	**25.1**	**0.77**	**0.26**	**70.3**	**62.1**	**3.8**

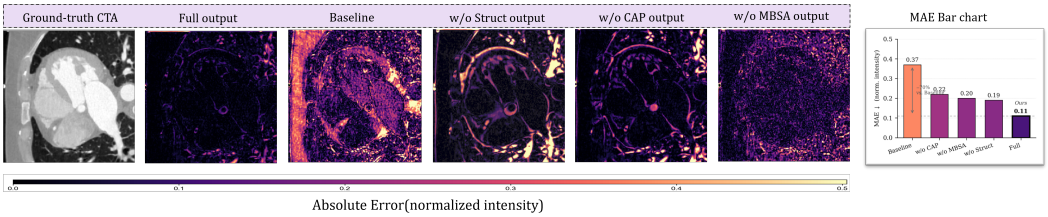


Figure 3: Voxel-wise heatmaps for SAFE-Diff without CAP, structural priors, or MBSA, compared with the full model on a held-out cardiac case. All maps share the same scale; darker colors indicate lower error.

Figure 3 further explains the quantitative results by localizing the errors introduced by removing each component. Compared with the full model, removing the structural priors leads to pronounced errors along the cardiac chambers and vascular boundaries, indicating reduced anatomical consistency. Without CAP, residual errors become more broadly distributed across the cardiac region, suggesting weaker global semantic guidance during cross-modal translation. Removing MBSA mainly increases errors around fine vessels and high-frequency contours, demonstrating its role in preserving fine anatomical details. In contrast, the full model produces the most spatially localized residual errors. Additional cross-modal embedding visualizations are provided in **Supplementary Doc.**.

Ablation of structural priors. Table 3(a) further analyzes the contributions of the region-level segmentation mask S and the boundary prior E_{edge}. Removing both priors reduces PSNR/SSIM from 25.1/0.77 to 24.2/0.74, indicating that semantic alignment alone is insufficient to preserve fine anatomical structures. Using S alone improves the result to 24.3/0.75, while using E_{edge} alone gives 24.5/0.75, suggesting that boundary cues are particularly useful for maintaining sharp vessel contours. The best performance is achieved when both priors are used together, confirming that region-level anatomy and boundary-level geometry are complementary for contrast-enhanced synthesis.

5 Conclusion

We presented SAFE-Diff, a diffusion framework for cross-modality medical image synthesis that integrates cross-modal semantic alignment, anatomical priors, and frequency-aware feature modulation. Experiments show that SAFE-Diff outperforms representative GAN-, transformer-, and diffusion-based methods in image quality and anatomical fidelity. Downstream segmentation, robustness, and qualitative analyses further support anatomical preservation, highlighting its potential for virtual contrast-enhanced imaging. However, SAFE-Diff assumes paired and reasonably aligned training data, which may limit its applicability to more challenging cross-domain scenarios.

REFERENCES

Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974.

S Bakas, H Akbari, A Sotiras, M Bilello, M Rozycki, JS Kirby, JB Freymann, K Farahani, and C Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific Data 4, 2017.

Yuwen Chen, Nicholas Konz, Hanxue Gu, Haoyu Dong, Yaqian Chen, Lin Li, Jisoo Lee, and Maciej A Mazurowski. Contourdiff: Unpaired image translation with contour-guided diffusion models. *arXiv preprint arXiv:2403.10786*, 2024.

Younghan Choi, Junhyung Kim, Seokmin Huh, Jeongjin Lee, Minseok Kim, Kyong Joon Lee, Yongsub Song, and Jong Chul Ye. Synthesis of contrast-enhanced chest ct from non-contrast ct using a 3d fully convolutional network. *Scientific Reports*, 11(1):23004, 2021.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, pp. 8188–8197, 2020.

Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, volume 11070 of *Lecture Notes in Computer Science*, pp. 529–536. Springer International Publishing, 2018. doi: 10.1007/978-3-030-00928-1_60.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE TPAMI*, 45(9):10850–10869, 2023.

Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: Residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022.

Salman UH Dar, Mahmut Yurt, Levent Karacan, Aykut Erdem, Erkut Erdem, and Tolga Cukur. Image synthesis in multi-contrast mri with conditional generative adversarial networks. *IEEE Transactions on Medical Imaging*, 38(10):2375–2388, 2019.

Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4084–4093, 2024.

J. Fan, Y. Wang, Z. Zhang, Q. Guo, W. Zhang, J. Cao, Y. Liu, C. Liu, Z. Wang, L. Shen, Y. Xu, and J. Li. Pre- to post-contrast medical image synthesis with outline-guided accelerated diffusion model. *Neural Networks*, 192:107851, 2025.

Yuan Gao, Richard L. J. Qiu, Huiqiao Xie, Chih-Wei Chang, Tonghe Wang, Beth Ghavidel, Justin Roper, Jun Zhou, and Xiaofeng Yang. Ct-based synthetic contrast-enhanced dual-energy ct generation using conditional denoising diffusion probabilistic model. *Physics in Medicine & Biology*, 69(16):165015, 2024a.

Zhuangzhi Gao, He Zhao, Zhongli Wu, Yuankai Wang, Gregory Yoke Hong Lip, Alena Shantsila, Eduard Shantsila, and Yalin Zheng. Coral-cvds: a consistent ordinal regression model for cardiovascular diseases grading. In *International Workshop on Ophthalmic Medical Image Analysis*, pp. 73–82. Springer, 2024b.

Alper Güngör, Salman UH Dar, Şaban Öztürk, Yilmaz Korkmaz, Hasan A Bedel, Gokberk Elmas, Muzaffer Ozbey, and Tolga Çukur. Adaptive diffusion priors for accelerated mri reconstruction. *Medical Image Analysis*, 88:102872, 2023.

Seunghyeok Han, Jiyeon Song, Woo Hyeon Kim, Sang Joon Park, et al. Dl-syncct: deep learning-based synthesis of contrast-enhanced abdominal ct images from non-contrast ct images. *Scientific Reports*, 14(1):16638, 2024.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pp. 1125–1134, 2017.

Lan Jiang, Ye Mao, Xiangfeng Wang, Xi Chen, and Chao Li. Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis. In *MICCAI*, pp. 398–408. Springer, 2023.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023.

M Kishore Kumar, Sriprabha Ramanarayanan, S Sadhana, Arunima Sarkar, Matcha Naga Gayathri, Keerthi Ram, and Mohanasankar Sivaprakasam. Dce-diff: Diffusion model for synthesis of early and late dynamic contrast-enhanced mr images from non-contrast multimodal inputs. In *CVPR Workshops*, pp. 5174–5183, 2024.

Lingke Kong, Chenyu Lian, Detian Huang, Zhenjiang Li, Yanle Hu, and Qichao Zhou. Breaking the dilemma of medical image-to-image translation. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:1964–1978, 2021.

Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A. Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *MICCAI*, pp. 88–98. Springer, 2024a.

Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A. Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15007 of *Lecture Notes in Computer Science*, pp. 88–98. Springer Nature Switzerland, 2024b. doi: 10.1007/978-3-031-72104-5_9.

Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *CVPR*, pp. 1952–1961, 2023a.

Yunxiang Li, Hua-Chieh Shao, Xiao Liang, Liyuan Chen, Ruiqi Li, Steve Jiang, Jing Wang, and You Zhang. Zero-shot medical image translation via frequency-guided diffusion models. *arXiv preprint arXiv:2304.02742*, 2023b.

Yunxiang Li, Hua-Chieh Shao, Xiao Liang, Liyuan Chen, Ruiqi Li, Steve Jiang, Jing Wang, and You Zhang. Zero-shot medical image translation via frequency-guided diffusion models. *IEEE Transactions on Medical Imaging*, 43(3):980–993, 2024. doi: 10.1109/TMI.2023.3325703.

Wanchun Lyu, Hao Li, Yue Han, Jun Jiang, Zhenghan Yang, et al. Deep learning unveils hidden angiography in noncontrast computed tomography scans. *Radiology*, 307(1):e222616, 2023.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34 (10):1993–2024, 2014.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Cukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12):3524–3539, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Yejee Shin, Yeeun Lee, Hanbyol Jang, Geonhui Son, Hyeongyu Kim, and Dosik Hwang. Anatomical consistency and adaptive prior-informed transformation for multi-contrast mr image synthesis via diffusion model. In *CVPR*, pp. 30918–30927, 2025.

Connie W Tsao, Aaron W Aday, Zaid I Almarzooq, Cheryl AM Anderson, Pankaj Arora, et al. Heart disease and stroke statistics—2023 update: a report from the american heart association. *Circulation*, 147(8):e93–e621, 2023.

Jiayuan Wang, QM Jonathan Wu, and Farhad Pourpanah. Dc-cyclegan: Bidirectional ct-to-mr synthesis from unpaired data. *Computerized Medical Imaging and Graphics*, 108:102249, 2023.

Zhenbin Wang, Lei Zhang, Lituan Wang, and Zhenwei Zhang. Soft masked mamba diffusion model for ct to mri conversion. *arXiv preprint arXiv:2406.15910*, 2024a.

Zhifeng Wang, Renjiao Yi, Xin Wen, Chenyang Zhu, and Kai Xu. Cardiovascular medical image and analysis based on 3d vision: A comprehensive survey. *Meta-Radiology*, 2(4):100102, 2024b.

Zhifeng Wang, Renjiao Yi, Xin Wen, Chenyang Zhu, and Kai Xu. Vastsd: Learning 3d vascular tree-state space diffusion model for angiography synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15693–15702, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

SUPPLEMENTARY MATERIAL

S1 IMPLEMENTATION DETAILS

S1.1 TRAINING CONFIGURATION

SAFE-Diff is implemented in PyTorch 2.2 with CUDA 12.1 and MONAI. All experiments are conducted using NVIDIA H100 GPUs with mixed-precision distributed training.

Training consists of two sequential stages. In Stage 1, the modality-specific VAE encoders and decoders are first pretrained for image reconstruction. After reconstruction pretraining, the VAE parameters are frozen, and the shared projection head is optimized using the symmetric contrastive objective described in the main paper. The projection head is trained using AdamW with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. After convergence, both the encoders and projection head are frozen and used to provide source-derived semantic conditions for Stage 2.

Stage 2 adopts a three-dimensional latent diffusion backbone based on COLA. The diffusion model is optimized using AdamW with a learning rate of 1×10^{-5}. We use a linear noise schedule with $T = 200$ diffusion steps. The anatomical segmentation mask and its corresponding boundary map are downsampled to the latent resolution and concatenated with the noisy target latent. The source latent features and projected global embedding are incorporated as semantic conditions, while MBSA is inserted at the U-Net bottleneck.

Unless otherwise specified, all competing methods and ablation variants use the same data partitions, preprocessing pipeline, optimization settings, and sampling protocol.

Table S1: Implementation configuration of SAFE-Diff.

Setting	Configuration
Framework	PyTorch 2.2, CUDA 12.1, MONAI
Hardware	NVIDIA H100 GPUs
Precision	Mixed precision
Stage 1 optimizer	AdamW
Stage 1 learning rate	1×10^{-4}
Stage 1 weight decay	1×10^{-4}
Stage 2 optimizer	AdamW
Stage 2 learning rate	1×10^{-5}
Diffusion steps	200
Noise schedule	Linear
Number of spectral bands	3
Cardiac spatial size	$256 \times 144 \times 144$

S2 DATA PREPROCESSING

S2.1 CARDIAC CT-TO-CTA DATASET

The cardiac cohort contains 1,600 paired non-contrast CT and CTA volumes from unique patients. All data are divided at the patient level into 1,120 training, 240 validation, and 240 test subjects, corresponding to a ratio of $7 : 1.5 : 1.5$. Each patient appears in exactly one partition. The validation partition is used for checkpoint and hyperparameter selection, whereas all reported cardiac results are calculated using the held-out test partition.

All cardiac volumes are resampled to an isotropic spacing of 1.1 mm. Trilinear interpolation is used for image volumes, whereas nearest-neighbor interpolation is used for discrete segmentation masks. The resampled volumes are processed at a spatial size of $256\times144\times144$ before fixed-size 3D patch extraction. All synthesis methods receive the same training, validation, and test partitions and the same preprocessed image volumes.

S2.2 BraTS T1-to-T1ce Dataset

For the brain synthesis experiment, we follow the official subject-level partitions of BraTS 2016/2017. The source T1 and target T1ce volumes are processed using the same spatial resampling strategy. No validation or test subject is used to optimize the modality-specific encoders, projection head, diffusion model, or segmentation-prior network.

S2.3 Generation of Anatomical Priors

For cardiac CT-to-CTA synthesis, the anatomical prior is generated from the source non-contrast CT using a fixed pipeline based on TotalSegmentator v2. The required cardiac chamber and great-vessel classes are retained and remapped into K task-specific classes. The resulting prediction is converted into a one-hot tensor

$$S \in \{0,1\}^{K\times D\times H\times W}. \tag{S1}$$

The corresponding class-wise boundary prior is generated deterministically:

$$E_{\text{edge}} = \text{Edge}(S). \tag{S2}$$

Neither the paired CTA nor its expert annotation is used to generate S or E_{edge} for validation or test subjects.

For BraTS, a dedicated nnU-Net is trained using only subjects from the BraTS training partition. Its input is the source T1 volume, and the corresponding training-set tumor annotation is used as supervision. After training, the nnU-Net is frozen. For validation or test subject i, the prior is obtained as

$$S_i = g_\phi\left(x_i^{\text{T1}}\right), \tag{S3}$$

where g_ϕ denotes the frozen nnU-Net. Neither the target T1ce image nor its annotation is used to construct the prior for a held-out subject.

Although the original BraTS training annotations were produced with access to multimodal images, they are used only as training labels for training-partition subjects. The provided annotations of validation and test subjects are not used as oracle priors. Therefore, the validation and test pipelines perform source-only prior prediction and do not have access to privileged target information.

Embedding space visualization. As shown in Fig. S1, we project ℓ_2-normalized case embeddings into 2D using t-SNE. Before alignment, source features (orange) form a manifold separated from the target cluster (blue), indicating a large modality gap. After CAP, source embeddings (green) overlap substantially with the target cluster, indicating improved cross-modal alignment and a stronger global prior for stable, anatomy-preserving denoising.

Figure S2 ,MAE decreases monotonically as components are added, from 0.37 for the unconditioned baseline to 0.11 for the full model. Among the leave-one-out variants, removing the cross-modal alignment (CAP) yields the largest error 0.22, in agreement with the broadly distributed residuals and indicating that a shared cross-modal representation provides the global foundation on which the subsequent structural and spectral conditioning operate.

S2.4 Separation of Prior Generation and Structural Evaluation

The cardiac structural evaluation is conducted on 20 expert-annotated CTA subjects drawn exclusively from the held-out test partition. These subjects and their annotations are not used to train the synthesis model, train or fine-tune the prior predictor, select a checkpoint, tune a hyperparameter, or train the evaluation models.

For every synthesis method, cardiac mDice is calculated using cardiac masks predicted by TotalSegmentator, whereas vessel mDice is calculated using the dedicated vessel nnU-Net. Both evaluation models are applied identically to real and synthesized CTA, and their predictions are compared with the expert annotations.

The prior generator operates on the source non-contrast CT before synthesis, whereas the evaluation models operate on the generated CTA after synthesis. Nevertheless, because the cardiac prior and cardiac evaluation may involve the same segmentation software family, the resulting scores are described as a *proxy structural evaluation*. They provide a common measurement protocol for comparing synthesis methods but should not be interpreted as independent clinical validation.

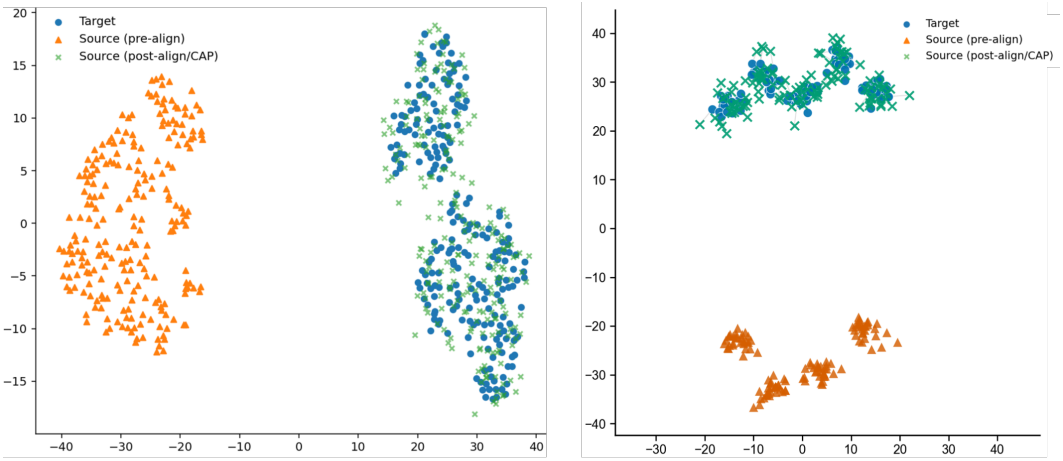


Figure S1: Cross-modal embedding alignment visualized using t-SNE. Target embeddings are shown in blue, source embeddings before alignment in orange, and source embeddings after CAP in green. The increased overlap after CAP indicates a reduced modality gap.

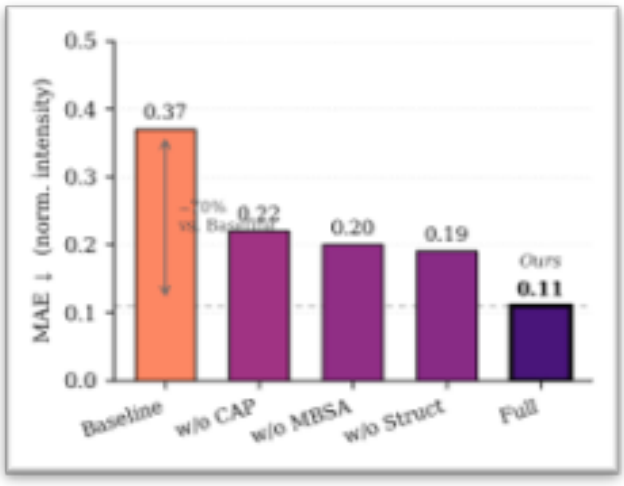


Figure S2: Component ablation in terms of image-level mean absolute error (MAE, ↓) on cardiac CT→CTA. Each bar is the volume-averaged absolute error of the corresponding voxel-wise error map and shares its Absolute-Error colour map, so that darker denotes lower error in both figures. Removing any component increases MAE, and the full model attains the lowest error 0.11, a 70% reduction over the unconditioned baseline (0.37).