GEOMETRY-AWARE SPECTRAL CONDITIONING FOR BIPLANAR X-RAY-TO-CT RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing a 3D computed tomography (CT) volume from two orthogonal X-ray projections is severely ill-posed because projection measurements collapse depth information and superimpose anatomical structures. We propose Frequency-Enhanced Dual-Encoding ControlNet (FEDE-ControlNet), a framework that adapts a frozen 3D latent diffusion model for biplanar X-ray-to-CT reconstruction through complementary geometric and spectral conditioning. A geometric pathway back-projects features extracted by a frozen X-ray encoder into a shared volumetric grid, establishing explicit detector-to-volume correspondence. In parallel, a dual-encoding pathway integrates spatial features with global and patch-local Fourier representations through frequency-domain cross-attention, then geometrically lifts the fused features into multi-scale volumetric conditions. A trainable control branch incorporates both pathways through zero-initialized residual connections while keeping the pretrained denoiser and CT autoencoder frozen. To complement latent-space denoising, multi-level volumetric feature supervision aligns decoded predictions with reference CT volumes using hierarchical representations from an independent frozen encoder. This auxiliary supervision is applied only during training and adds no inference-time computation. Experiments demonstrate that the proposed FEDE-ControlNet achieves superior reconstruction fidelity and structural similarity compared with the evaluated state-of-the-art methods.

1 INTRODUCTION

Reconstructing a 3D computed tomography (CT) volume from biplanar X-rays offers a potential means of recovering volumetric anatomical information from a limited radiographic acquisition. Unlike conventional CT, which collects projections over many angles, biplanar imaging provides only two views, typically posteroanterior (PA) and lateral (Lat) radiographs. This limited acquisition motivates the development of reconstruction methods that can complement two-dimensional radiographic assessment. However, each projection integrates attenuation along X-ray paths, superimposing structures at different depths. Two orthogonal views therefore leave substantial ambiguity: distinct 3D anatomies can produce similar measurements. The central challenge is to combine the available projection evidence with anatomical priors while maintaining fidelity to the individual target volume.

Learning-based methods have made progress in addressing this challenge. X2CT-GAN (Ying et al., 2019) demonstrates the feasibility of learning volumetric reconstructions from biplanar radiographs using adversarial training. Geometric correspondence was subsequently incorporated more explicitly: PerX2CT (Kyung et al., 2023) associates 3D query locations with features sampled from the input projections, making the relationship between detector evidence and target anatomy explicit. Such geometry-aware representations provide a useful inductive bias for reconstruction. Nevertheless, geometric lifting alone cannot determine where structures lie along a projection ray, and a plausible volumetric output may still differ from the anatomy that generated the observations.

Diffusion models provide learned priors for regularizing this underdetermined reconstruction problem. DiffuX2CT (Liu et al., 2024) and DX2CT (Jeong et al., 2025) explore diffusion-based reconstruction conditioned on radiographic information, while DVG-Diffusion (Xie et al., 2026) further investigates alignment between projection-derived features and the CT latent space. These studies

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.

motivate combining generative modeling with geometry-aware conditioning. In parallel, pretrained volumetric generators such as 3D MedDiffusion (Wang et al., 2025) offer an opportunity to reuse an existing anatomical prior through conditional adaptation. In this setting, an important design problem is how to translate sparse two-dimensional observations into informative conditions for a frozen three-dimensional generator, and how to supervise the resulting reconstruction beyond latent-space denoising.

To address these limitations, we propose Frequency-Enhanced Dual-Encoding ControlNet (FEDE-ControlNet), a geometry-aware framework that adapts a pretrained volumetric diffusion prior to biplanar X-ray-to-CT reconstruction. PA and lateral X-rays are first encoded by a frozen two-dimensional VAE and back-projected into a shared volumetric grid using calibrated acquisition geometry, establishing explicit correspondence between projection evidence and the target volume. Since back-projection alone may not sufficiently represent long-range anatomical dependencies and localized structural variations, we further construct a multi-scale spatial-frequency conditioning pathway. Global and patch-local Fourier processing capture complementary contextual and localized frequency information, which is fused with spatial features through frequency-domain cross-attention and geometrically lifted into scale-specific volumetric conditions.

A trainable control branch incorporates the back-projected volume at its input and the lifted spatial-frequency features at corresponding resolutions. Their guidance is injected into the frozen diffusion backbone through zero-initialized residual connections, enabling reconstruction-specific adaptation while preserving the pretrained denoiser and CT autoencoder. Because latent-space noise prediction does not directly constrain structural errors after decoding, we additionally introduce multi-level volumetric feature supervision. During training, decoded reconstructions and reference CTs are compared through a frozen volumetric feature model at multiple representation levels, providing complementary guidance for local details, intermediate morphology, and global volumetric structure. This auxiliary objective is evaluated periodically and introduces no additional inference cost. Experiments on LIDC-IDRI-derived biplanar projections demonstrate favorable quantitative performance and improved preservation of anatomical structures.

Our main contributions are threefold:

- We propose FEDE-ControlNet, which combines calibrated feature back-projection with residual adaptation of a frozen volumetric diffusion prior for biplanar X-ray-to-CT reconstruction.
- We develop a geometry-aligned, multi-scale spatial-frequency conditioning pathway that integrates global context and localized spectral information at corresponding control resolutions.
- We introduce multi-level supervision of decoded CT volumes, complementing latent-space denoising with hierarchical structural guidance without additional inference cost.

2 Related Work

2.1 Biplanar (Dual-view) CT Reconstruction

Reconstructing CT from two radiographs remains highly underdetermined despite their complementary anatomical information. Early learning-based methods addressed this ambiguity through learned anatomical priors. X2CT-GAN (Ying et al., 2019) combined biplanar feature fusion with adversarial learning, while Kasten et al. (2020) developed an end-to-end convolutional model for knee-bone reconstruction. Subsequent methods strengthened feature representation and volumetric prediction: TRCT-GAN (Wang et al., 2023) and XTransCT (Zhang et al., 2024) introduced transformer-based modeling to capture contextual dependencies and improve the mapping from projections to volumes.

More recently, diffusion models have provided expressive generative priors for this inverse problem. DiffuX2CT (Liu et al., 2024) develops conditional diffusion for biplanar reconstruction, while DX2CT (Jeong et al., 2025) uses position-aware projection features for mono- and biplanar inputs. DVG-Diffusion (Xie et al., 2026) combines an observed view with a synthesized complementary view through spatially aligned conditioning. In DCT-Net (Zhang et al., 2025), diffusion generates

bone-suppressed X-ray conditions for a dual-branch adversarial reconstruction network. These approaches highlight the importance of both anatomical priors and effective conditioning. Our work focuses on adapting a pretrained volumetric diffusion prior through calibrated biplanar features and complementary spatial-frequency guidance.

2.2 Geometry- and Frequency-Aware CT Reconstruction

Acquisition geometry provides a fundamental link between X-ray measurements and volumetric structure. Classical methods such as SART (Andersen & Kak, 1984) use the projection model to enforce measurement fidelity, while total-variation regularization (Sidky & Pan, 2008) constrains reconstruction from incomplete data. Learned representations further incorporate geometry into reconstruction. NAF (Zha et al., 2022) represents attenuation as a continuous field optimized against measured projections. DIF-Net (Lin et al., 2023) predicts volumetric intensities from projection features sampled at geometrically corresponding detector locations. For biplanar reconstruction, perspective projection-based methods (Kyung et al., 2023) similarly exploit the relationship between image coordinates and three-dimensional locations. These studies motivate explicit geometric alignment when transferring projection evidence into volumetric representations.

Frequency-aware methods provide complementary tools for recovering structural detail under sparse sampling. FreeSeed (Ma et al., 2023) uses frequency-aware processing and self-guidance to suppress artifacts in sparse-view CT images. DuFal (Van et al., 2026) combines spatial encoding with global and local Fourier processing for extremely sparse-view CBCT reconstruction. Building on this spatial-frequency encoding design, we construct a conditioning hierarchy that geometrically lifts projection features into volumetric representations at multiple resolutions. This connects spectral feature extraction to the spatial organization of the diffusion control branch.

Medical Generative Priors. Volumetric generative models learn anatomical distributions that can complement incomplete imaging measurements. Hierarchical Amortized GAN (Sun et al., 2022) supports high-resolution three-dimensional synthesis through hierarchical generation. Medical Diffusion (Khader et al., 2023) extends diffusion modeling to three-dimensional medical images, while latent and structured representations make volumetric generation more computationally practical. MAISI (Guo et al., 2025) combines volumetric compression with latent diffusion for scalable CT synthesis. 3D MedDiffusion (Wang et al., 2025) further develops controllable volumetric generation and demonstrates adaptation to downstream tasks, including sparse-view reconstruction.

3 Method

Let $y \in \mathbb{R}^{D\times H\times W}$ denote the target CT volume and $x_v \in \mathbb{R}^{H_d\times W_d}$ its projection at view $v \in \mathcal{V} = \{\mathrm{PA}, \mathrm{Lat}\}$. In the idealized physical model, the line-integral projection is

$$x_v(\mathbf{u}) = \mathcal{A}_v(y)(\mathbf{u}) = \int_{\mathcal{L}_v(\mathbf{u})} \mu_y(\mathbf{r})\,\mathrm{d}\ell, \tag{1}$$

where μ_y is the linear attenuation field associated with y, and $\mathcal{L}_v(\mathbf{u})$ is the ray through detector coordinate $\mathbf{u}$. In our experiments, the projector instead operates on clipped and normalized CT arrays. Its fixed geometry defines a consistent spatial mapping between detector coordinates and the CT grid; it does not calibrate the synthetic pixel values as physical attenuation measurements. Two projections leave substantial depth ambiguity, motivating a conditional model $p_\phi(y \mid x_{\mathrm{PA}}, x_{\mathrm{Lat}})$ that combines projection evidence with a learned anatomical prior.

As shown in Fig. 1, FEDE-ControlNet adapts a frozen volumetric latent diffusion model to biplanar X-ray-to-CT reconstruction through two conditioning pathways. In the geometric pathway, a frozen X-ray VAE encodes the PA and lateral projections, and calibrated back-projection lifts their features onto a shared volumetric grid to form the input-level condition. In parallel, Frequency-Enhanced Dual-Encoding (FEDE) combines spatial features with global and patch-local Fourier representations through frequency-domain cross-attention, then lifts the fused features into scale-specific volumetric conditions. A trainable control branch receives the geometric condition at its input and the FEDE conditions at corresponding resolutions, injecting their guidance into the frozen denoiser through zero-initialized residual connections. During training, a separate frozen volumetric feature model (VFM) compares decoded predictions with reference CTs at multiple feature levels.

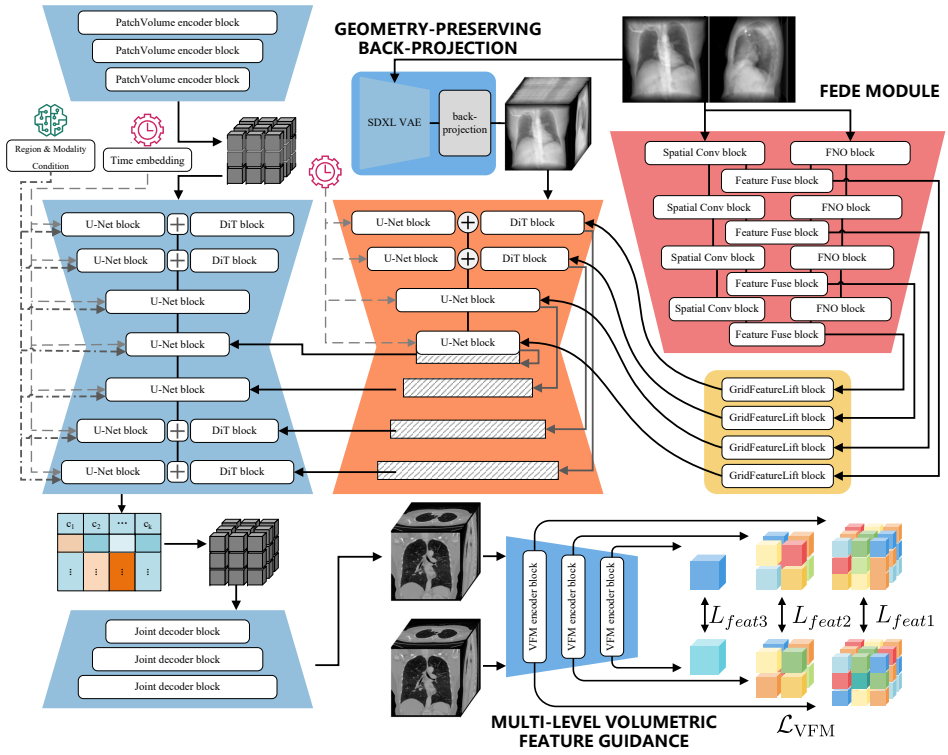


Figure 1: Overview of the proposed FEDE-ControlNet.

3.1 Geometry-Aligned Feature Back-Projection

The first pathway establishes a spatially organized condition before the generative prior resolves the remaining ambiguity. A frozen two-dimensional VAE encoder E_{X} maps each normalized projection to a compact feature map,

$$f_v = E_{\mathrm{X}}(x_v), \qquad f_v \in \mathbb{R}^{C_b \times H_b \times W_b}, \tag{2}$$

where C_b, H_b, and W_b denote the feature channels and spatial dimensions. Encoding before lifting provides multiple learned descriptors of the projection while reducing its spatial resolution.

We lift each feature channel onto a shared volumetric grid Ω_0 using a calibrated cone-beam back-projection operator $\mathcal{B}_v^b$ implemented with TIGRE (Biguri et al., 2025):

$$F_v^{\mathrm{bp}}[c, \mathbf{q}] = \mathcal{B}_v^b(f_v)[c, \mathbf{q}] = \sum_{\mathbf{u}} \omega_v^b(\mathbf{q}, \mathbf{u}) f_v[c, \mathbf{u}], \qquad \mathbf{q} \in \Omega_0. \tag{3}$$

The superscript b denotes sampling onto the target volumetric grid Ω_0 under the fixed geometry; ω_v^b contains the geometric and interpolation weights for the feature-map sampling grid. These weights account for the detector resolution and the coordinates of the target volume under the fixed computational geometry. The PA and Lat features are concatenated channel-wise:

$$F_{\mathrm{bp}} = \mathrm{Concat}\left(F_{\mathrm{PA}}^{\mathrm{bp}}, F_{\mathrm{Lat}}^{\mathrm{bp}}\right). \tag{4}$$

This condition preserves calibrated detector-to-grid correspondences and complementary view evidence. Since feature lifting does not determine where attenuation occurs along a ray, the remaining ambiguity is addressed by the learned conditioning hierarchy and volumetric prior.

3.2 Frequency-Enhanced Dual-Encoding

Geometric alignment establishes spatial correspondence between projection evidence and volumetric locations, while effective conditioning also requires features that capture global anatomical con-

text and localized structural detail. Inspired by DuFal (Van et al., 2026), we combine spatial encoding with global and patch-local Fourier processing, then geometrically lift the fused representations into a multi-scale conditioning hierarchy for volumetric diffusion. Each encoding branch shares its weights across the PA and Lat views. At each scale $s \in \{1, \dots, L\}$, the spatial and frequency branches produce features with matching spatial resolutions. The spatial hierarchy is defined as

$$S_v^s = \Phi_{\mathrm{spa}}^s(S_v^{s-1}), \qquad S_v^0 = x_v, \tag{5}$$

where Φ_{spa}^s includes spatial feature extraction and downsampling.

Global and local spectral encoding. Global spectral processing provides a broad receptive field, while patch-local processing preserves the spatial locality of fine-scale variations. Both branches use factorized Fourier operators. For an input $U \in \mathbb{R}^{C_{\mathrm{in}} \times h \times w}$, let $\widehat{U} = \mathcal{F}(U)$ denote its complex-valued two-dimensional Fourier representation. At a retained frequency coordinate $\mathbf{k} \in \mathcal{K}_s$, the global spectral operator is parameterized as

$$\widehat{V}[a, \mathbf{k}] = \sum_{c=1}^{C_{\mathrm{in}}} A_s[a, c] B_s[c, \mathbf{k}] \widehat{U}[c, \mathbf{k}], \tag{6}$$

where A_s mixes channels and B_s weights frequency modes within each input channel. Coefficients outside $\mathcal{K}_s$ are set to zero. This factorization uses $C_{\mathrm{out}}C_{\mathrm{in}} + C_{\mathrm{in}}|\mathcal{K}_s|$ complex coefficients, compared with $C_{\mathrm{out}}C_{\mathrm{in}}|\mathcal{K}_s|$ for an unfactorized spectral kernel. Inverse transformation produces the global feature response.

The local branch partitions the input into $N_p \times N_p$ non-overlapping patches and applies the same operator form independently to each patch, retaining all patch-level frequency modes. The processed patches are reassembled in their original spatial order. Let $\mathcal{G}_s$ and $\mathcal{P}_s$ denote the global and local branches, including their resolution alignment. Starting from $T_v^0 = x_v$, the frequency hierarchy is

$$T_v^s = \mathrm{GA}_s\big(\mathrm{GELU}\big[W_s(T_v^{s-1}) + \mathcal{G}_s(T_v^{s-1}) + \mathcal{P}_s(T_v^{s-1})\big]\big), \tag{7}$$

where W_s is the position-wise feature branch with the required downsampling, and GA_s denotes Galerkin attention (Cao, 2021). All three terms have the spatial resolution of S_v^s.

Spatial-frequency fusion. At each scale, cross-attention allows spatial features to selectively incorporate spectral information. Channel projections first align the two representations, followed by Fourier transformation:

$$\widehat{S}_v^s = \mathcal{F}(C_s^{\mathrm{spa}}(S_v^s)), \qquad \widehat{T}_v^s = \mathcal{F}\big(C_s^{\mathrm{fre}}(T_v^s)\big). \tag{8}$$

Writing their real and imaginary components as $\widehat{S}_v^s = S_v^{s,\mathrm{r}} + \mathrm{i}S_v^{s,\mathrm{i}}$ and $\widehat{T}_v^s = T_v^{s,\mathrm{r}} + \mathrm{i}T_v^{s,\mathrm{i}}$, we use spatial features as queries and frequency features as keys and values:

$$C_v^{s,\eta} = \mathrm{Attn}_{s,\eta}(S_v^{s,\eta}, T_v^{s,\eta}, T_v^{s,\eta}), \qquad \eta \in \{\mathrm{r}, \mathrm{i}\}. \tag{9}$$

The attention operator includes learned query, key, and value projections. Residual fusion and inverse transformation yield a real-valued projection feature:

$$M_v^s = \mathrm{Re}\big\{\mathcal{F}^{-1}\big[(S_v^{s,\mathrm{r}} + C_v^{s,\mathrm{r}}) + \mathrm{i}(S_v^{s,\mathrm{i}} + C_v^{s,\mathrm{i}})\big]\big\}. \tag{10}$$

Geometry-aligned volumetric conditions. The fused features are lifted onto the grid Ω_s associated with control resolution s. For each grid point $\mathbf{q}$, calibrated perspective projection Π_v^s gives its sampling coordinate in the corresponding feature map:

$$L_v^s(\mathbf{q}) = \mathrm{Bilinear}(M_v^s, \Pi_v^s(\mathbf{q})). \tag{11}$$

The two sampled features are concatenated and arranged into a dense volume,

$$F^{V,s} = \mathrm{Grid}_{\Omega_s}\Big(\{\mathrm{Concat}[L_{\mathrm{PA}}^s(\mathbf{q}), L_{\mathrm{Lat}}^s(\mathbf{q})]\}_{\mathbf{q}\in\Omega_s}\Big). \tag{12}$$

All grids share the same physical coordinate system. Thus, F_{bp} supplies an input-level geometric condition, while $\{F^{V,s}\}_{s=1}^L$ provides trainable spatial-frequency representations throughout the control hierarchy.

3.3 Residual Control of a Frozen Diffusion Model

We reuse the pretrained volumetric denoiser and CT autoencoder of 3D MedDiffusion (Wang et al., 2025), learning a reconstruction-specific control pathway following ControlNet (Zhang et al., 2023). Let E_{CT} and D_{CT} denote the frozen CT encoder and decoder, including the latent scaling and packing used by the pretrained model. Given $z_0 = E_{\mathrm{CT}}(y)$, training samples a noisy latent

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{13}$$

where $\bar{\alpha}_t$ is the cumulative noise-schedule coefficient.

The back-projected condition is aligned with z_t and concatenated at the control input:

$$h_t^0 = Z_{\mathrm{in}}(\mathrm{Concat}[z_t, P_b(F_{\mathrm{bp}})]), \tag{14}$$

where P_b performs the required channel and grid alignment, and Z_{in} is a zero-initialized input convolution. At control stage s, the lifted FEDE condition is added to the propagated control feature:

$$h_t^s = \Phi_{\mathrm{ctrl}}^s\big(D_s(h_t^{s-1}) + P_s(F^{V,s}), t\big), \qquad s = 1, \ldots, L. \tag{15}$$

Here, D_s aligns the preceding feature with the current stage, P_s adapts the condition channels, and Φ_{ctrl}^s denotes the corresponding control block, including its U-Net/DiT operations where applicable.

Zero-initialized output projections convert the control features into residuals,

$$r_t^s = Z_s(h_t^s), \qquad Z_s\big|_{\mathrm{init}} = 0, \tag{16}$$

which are added to the corresponding frozen-backbone activations. These output projections ensure that the controlled model initially reproduces the pretrained denoiser, while subsequent updates learn input-dependent corrections. The noise prediction is

$$\widehat{\epsilon} = \epsilon_\theta\big(z_t, t; \{r_t^s\}_{s=1}^L\big), \tag{17}$$

where θ is frozen. The trainable parameters ϕ comprise the FEDE encoder, control branch, and associated feature projections. Defining $d_1(A, B) = \|A - B\|_1 / \operatorname{numel}(A)$ as the elementwise mean absolute error, the denoising objective is

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{(y, x_{\mathrm{PA}}, x_{\mathrm{Lat}}), t, \epsilon}\left[d_1(\epsilon, \widehat{\epsilon})\right]. \tag{18}$$

3.4 Multi-Level Volumetric Feature Supervision

Informative conditioning improves the denoising input, but the latent noise objective does not directly measure structural discrepancies after CT decoding. We therefore supervise decoded predictions in an independent volumetric representation space. From the predicted noise, we estimate the clean latent and reconstruct the CT volume:

$$\widehat{z}_0 = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\,\widehat{\epsilon}}{\sqrt{\bar{\alpha}_t}}, \qquad \widehat{y} = D_{\mathrm{CT}}(\widehat{z}_0). \tag{19}$$

This prediction is obtained from the sampled training timestep, without running a complete reverse diffusion trajectory.

We use an independent frozen volumetric feature model (VFM) from MAISI (Guo et al., 2025). Let V_ψ^i denote its representation at layer $i \in \mathcal{I} = \{3, 6, 8\}$, and $\mathcal{N}_i$ the feature normalization applied identically to both volumes. The auxiliary objective is

$$\mathcal{L}_{\mathrm{VFM}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} d_1\big(\mathcal{N}_i(V_\psi^i(\widehat{y})), \mathcal{N}_i(V_\psi^i(y))\big). \tag{20}$$

These layers differ in spatial resolution and receptive field, providing complementary supervision for local detail and broader volumetric structure. Per-level averaging prevents larger feature tensors from dominating solely because they contain more elements. Although the CT decoder and VFM parameters remain frozen, gradients through the predicted-volume branch propagate to the trainable conditioning and control parameters.

To amortize full-volume decoding and feature extraction, we evaluate the auxiliary term once every k_u optimizer updates. For update index $u \geq 1$, the training objective is

$$\mathcal{L}(u) = \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{VFM}}\, \delta_{k_u}(u) \mathcal{L}_{\mathrm{VFM}}, \qquad \delta_{k_u}(u) = \mathbf{1}[u \bmod k_u = 0]. \tag{21}$$

Here, λ_{VFM} controls the auxiliary loss weight and k_u its evaluation frequency. This schedule reduces the average cost of volumetric supervision. At inference, the two projections supply the conditioning features, DDIM denoising produces the CT latent, and D_{CT} decodes the final volume. The VFM and reference CT are required only during training.

Table 1: Quantitative comparison of 3D CT reconstruction quality. Best results are highlighted in bold and second-best results are underlined.

Method	Slice PSNR ↑	Slice SSIM ↑	Slice LPIPS ↓	Vol. PSNR ↑	Vol. SSIM ↑
X2CT-GAN (Ying et al., 2019)	26.612	0.654	0.234	21.786	0.643
PerX2CT (Kyung et al., 2023)	24.359	0.657	0.234	<u>22.884</u>	0.650
C-MedDiffusion	<u>26.813</u>	<u>0.683</u>	0.220	22.135	<u>0.671</u>
DVG-Diffusion(Xie et al., 2026)	26.718	0.670	<u>0.218</u>	21.253	0.659
FEDE-ControlNet (ours)	**28.168**	**0.713**	**0.209**	**23.110**	**0.708**

Table 2: Factorial ablation on LIDC-IDRI dataset. Every variant includes BP conditioning. The BP baseline is C-MedDiffusion; VFM denotes multi-level volumetric feature supervision.

Method	Slice PSNR ↑	Slice SSIM ↑	Vol. PSNR ↑	Vol. SSIM ↑	Proj. PSNR ↑	Proj. SSIM ↑
C-MedDiffusion (Baseline)	26.8126	0.6831	22.1351	0.6710	28.2610	0.9240
C-MedDiffusion + Feature supervision	26.9464	0.6917	22.3573	0.6817	29.0290	0.9295
FEDE-ControlNet (w/o Feature supervision)	<u>27.6209</u>	<u>0.7114</u>	<u>23.0479</u>	<u>0.7056</u>	<u>29.8857</u>	<u>0.9402</u>
FEDE-ControlNet (Complete)	**28.1676**	**0.7129**	**23.1104**	**0.7078**	**30.7444**	**0.9419**

4 EXPERIMENTS

Our framework uses the frozen denoiser and Patch-Volume Autoencoder of 3D MedDiffusion (Wang et al., 2025). For 128^3 CT volumes, the autoencoder produces an eight-channel packed latent of spatial size 32^3. The pretrained 2D X-ray encoder and the auxiliary MAISI volumetric feature model (VFM) are also frozen; training updates the conditioning and control parameters. Our model and its 128^3 ablation variants are implemented in PyTorch and trained for 200 epochs on a single NVIDIA A800 GPU, using the same optimizer settings and a learning rate of 10^{-4}. A batch size of two with two gradient-accumulation steps gives an effective batch size of four. At inference, these variants use 1000-step DDIM sampling with $\eta = 0$. We use 916 LIDC-IDRI (Armato et al., 2011) CT volumes for reconstruction training and 102 held-out volumes for testing. We additionally evaluate transfer to 60 LCTSC (Yang et al., 2020) cases without dataset-specific adaptation. Full preprocessing and coordinate conventions are provided in the supplementary material.

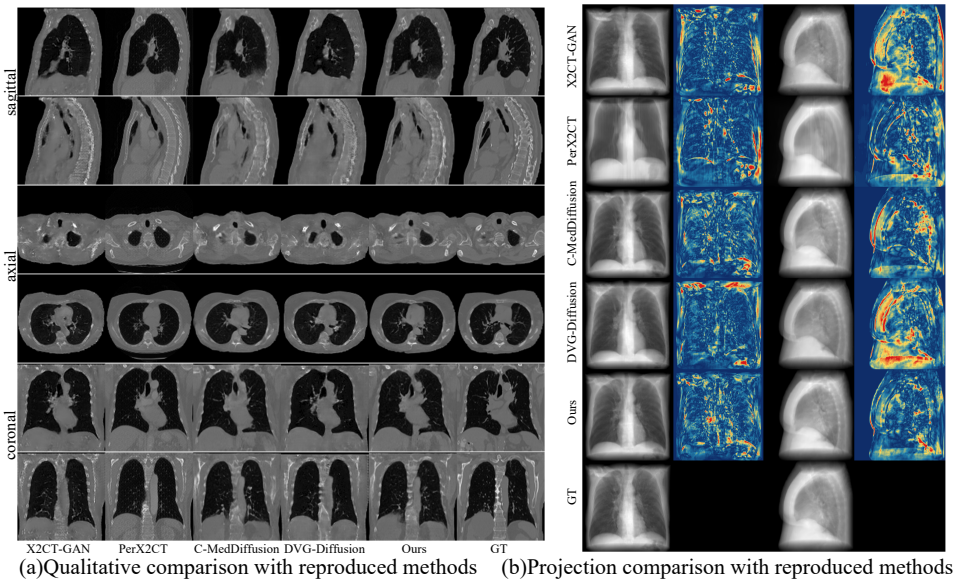
4.1 RECONSTRUCTION RESULTS ON LIDC-IDRI

We compare the proposed FEDE-ControlNet with X2CT-GAN (Ying et al., 2019), PerX2CT (Kyung et al., 2023), and DVG-Diffusion (Xie et al., 2026), using published checkpoints or locally retrained models. We also include C-MedDiffusion, a ControlNet baseline with the same frozen volumetric prior and back-projected (BP) conditioning, but without FEDE or VFM supervision.

CT reconstruction. FEDE-ControlNet achieves the best results among locally evaluated methods on all five CT metrics in Table 1. Relative to C-MedDiffusion, it increases Slice PSNR from 26.813 to 28.168 and Vol. PSNR from 22.135 to 23.110, while improving Slice SSIM from 0.683 to 0.713 and Vol. SSIM from 0.671 to 0.708. Slice LPIPS decreases from 0.220 to 0.209. The volume PSNR also exceeds PerX2CT, the strongest competing method on this metric, by 0.226. Together, these results show improvements in intensity fidelity, structural similarity, and perceptual agreement.

Projection consistency. Table 3 evaluates the consistency between the original X-ray inputs and the DRRs re-projected from the reconstructed CT volumes. FEDE-ControlNet also ranks first on all three projection-domain metrics in Table 3. Compared with C-MedDiffusion, projection PSNR increases by 2.4834 dB, SSIM increases by 0.0179, and LPIPS decreases from 0.1874 to 0.1659 (an 11.5% relative reduction). It exceeds PerX2CT, the next-best method in projection PSNR and SSIM, by 0.8792 dB and 0.0164, respectively. These results further demonstrate that the reconstructed volumes retain the spatial information contained in the original biplanar X-ray observations and better preserve the correspondence between the reconstructed 3D structures and the input projections.

Qualitative comparison. Figure 2 presents corresponding LIDC-IDRI slices and re-projected DRRs. In the displayed cases, FEDE-ControlNet more closely follows the reference lung contours and the arrangement of internal thoracic structures, while retaining clearer soft-tissue and skeletal boundaries. The sagittal and coronal views show improved agreement around the diaphragm and



(a)Qualitative comparison with reproduced methods (b)Projection comparison with reproduced methods

Figure 2: Qualitative comparison on LIDC-IDRI. (a) Corresponding sagittal, axial, and coronal slices reconstructed by different methods. (b) Re-projected PA and lateral DRRs with discrepancy maps relative to the GT-derived input projections.

Table 3: Projection consistency on LIDC-IDRI, measured between the input DRRs and projections of the reconstructed CT volumes.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
X2CT-GAN(Ying et al., 2019)	28.7591	0.9011	0.1981
PerX2CT(Kyung et al., 2023)	<u>29.8652</u>	<u>0.9255</u>	0.2074
C-MedDiffusion	28.2610	0.9240	<u>0.1874</u>
DVG-Diffusion(Xie et al., 2026)	27.1850	0.8957	0.1894
FEDE-ControlNet (ours)	**30.7444**	**0.9419**	**0.1659**

Table 4: Quantitative comparison of 3D CT reconstruction quality on LCTSC.

Method	Slice PSNR ↑	Slice SSIM ↑	Vol. PSNR ↑	Vol. SSIM ↑
X2CT-GAN(Ying et al., 2019)	13.5486	0.1625	12.6707	0.1349
PerX2CT(Kyung et al., 2023)	19.8627	<u>0.4014</u>	18.8171	0.3550
C-MedDiffusion	<u>20.0074</u>	0.3965	<u>19.2019</u>	<u>0.3801</u>
DVG-Diffusion(Xie et al., 2026)	19.1078	0.3851	18.1546	0.3694
FEDE-ControlNet (ours)	**20.3727**	**0.4165**	**19.5714**	**0.4035**

chest wall, and the axial views retain more recognizable pulmonary detail. The re-projected views provide a complementary comparison of the lung silhouette and central thoracic anatomy. Fine structures and localized projection discrepancies remain imperfect, consistent with the ambiguity of reconstruction from two views.

4.2 Evaluation on LCTSC

Table 4 presents the external evaluation on LCTSC without dataset-specific adaptation. Although all evaluated methods obtain lower scores than on LIDC-IDRI, FEDE-ControlNet achieves the best performance across all four reported metrics. Compared with C-MedDiffusion, the strongest competing method in Slice PSNR, Vol. PSNR, and Vol. SSIM, FEDE-ControlNet improves these metrics by 0.3653 dB, 0.3695 dB, and 0.0234, respectively. For Slice SSIM, it achieves 0.4165, exceeding the second-best result of 0.4014 obtained by PerX2CT by 0.0151. These consistent improvements indicate that the proposed method retains an advantage in both intensity fidelity and structural similarity when transferred to an unseen CT cohort.

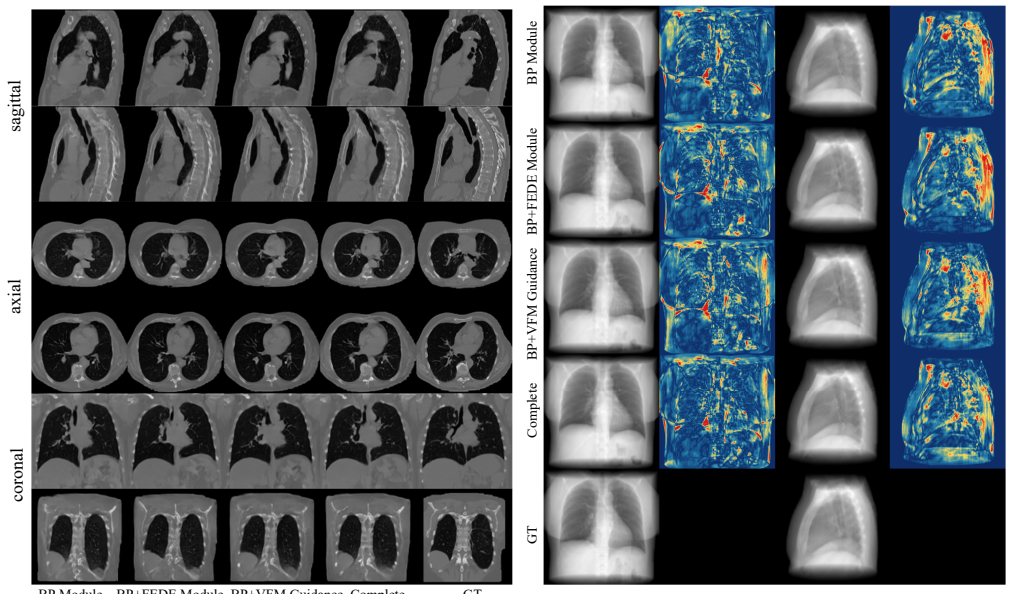


Figure 3: Qualitative ablation on LIDC-IDRI. (a) Corresponding CT slices from the BP baseline, BP + FEDE, BP + VFM, the complete model, and GT, from left to right. (b) Re-projected PA and lateral DRRs and discrepancy maps for the same variants, followed by the GT-derived projections.

4.3 Ablation Studies on LIDC-IDRI

Quantitative ablation. Table 2 shows that both components improve reconstruction quality, with FEDE providing larger gains when added independently to the BP baseline. Without VFM supervision, FEDE increases Slice PSNR, Vol. PSNR, and projection PSNR by 0.8083, 0.9128, and 1.6247 dB, respectively, together with corresponding SSIM gains of 0.0283, 0.0346, and 0.0162. VFM supervision alone yields smaller but consistent improvements across all six metrics, including PSNR gains of 0.1338, 0.2222, and 0.7680 dB in slice, volume, and projection space, respectively. Combining the two components further improves every reported metric over the FEDE-only variant: Slice PSNR and projection PSNR increase by 0.5467 and 0.8587 dB, while the additional gains in Vol. PSNR and Vol. SSIM are more modest at 0.0625 dB and 0.0022. The complete model achieves the best results across all six metrics, supporting the joint use of geometry-aligned spatial-frequency conditioning and feature supervision on decoded CT volumes.

Qualitative ablation. Figure 3 illustrates the same variants in CT and projection space. The BP baseline captures the coarse thoracic layout but leaves some internal structures and boundaries poorly resolved. FEDE produces more recognizable pulmonary detail, while the complete model more closely follows the reference contours and skeletal structures in the displayed examples. The projection panels show the corresponding changes in view consistency. These observations complement the aggregate metrics; visual sharpness alone is insufficient to establish anatomical accuracy.

5 Conclusion

We presented a geometry-aware FEDE-ControlNet framework for reconstructing 3D CT volumes from two orthogonal X-rays. The method adapts a frozen volumetric diffusion prior for biplanar X-ray-to-CT reconstruction through geometry-aligned back-projection and multi-scale spatial-frequency conditioning. Geometry-aware back-projection maps VAE-encoded X-ray features into a shared volumetric space. FEDE integrates spatial encoding with global and local Fourier processing through frequency-domain cross-attention, then geometrically lifts the fused features into multi-scale volumetric conditions. Experimental results demonstrate the proposed method improves reconstruction fidelity, structural similarity, and projection consistency.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and LaTeX formatting. The authors reviewed all AI-assisted revisions and verified the technical content, equations, citations, and reported experimental results. The authors take responsibility for the final content of this work.

REFERENCES

A. H. Andersen and A. Kak. Simultaneous algebraic reconstruction technique (sart): A superior implementation of the art algorithm. *Ultrasonic Imaging*, 6:81 – 94, 1984.

Samuel G. Armato, Geoffrey Mclennan, L. Bidaut, Michael McNitt-Gray, C. R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise Aberle, Claudia I. Henschke, Eric A. Hoffman, Ella A. Kazerooni, H. MacMahon, E. V. van Beek, D. Yankelevitz, A. Biancardi, P. H. Bland, Matthew S. Brown, R. Engelmann, Gary E. Laderach, Daniel Max, Richard C. Pais, D. Qing, Rachael Y. Roberts, Amanda R Smith, Adam Starkey, Poonam V. Batra, P. Caligiuri, Ali O. Farooqi, G. Gladish, C. Jude, R. Munden, I. Petkovska, L. Quint, L. Schwartz, Baskaran Sundaram, Lori E. Dodd, C. Fenimore, David Gur, Nicholas Petrick, J. Freymann, Justin S. Kirby, Brian Hughes, Alessi Vande Casteele, S. Gupte, M. Sallam, Michael D. Heath, Michael H. Kuhn, E. Dharaiya, R. Burns, David Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, Barbara Y. Croft, and Laurence P. Clarke. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38 2: 915–31, 2011.

Ander Biguri, Tomoyuki Sadakane, Reuben Lindroos, Yi Liu, Malena Sabaté Landman, Yi Du, Manasavee Lohvithee, Stefanie Kaser, Sepideh Hatamikia, Robert Bryll, Emilien Valat, Sarinrat Wonglee, Thomas Blumensath, and Carola-Bibiane Schönlieb. TIGRE v3: Efficient and easy to use iterative computed tomographic reconstruction toolbox for real datasets. *Engineering Research Express*, pp. 015011, 2025.

Shuhao Cao. Choose a transformer: Fourier or galerkin. *Advances in Neural Information Processing Systems*, 34:24924–24940, 2021.

Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin D. Simon, Mason J. Belue, Stephanie A. Harmon, B. Turkbey, and Daguang Xu. Maisi: Medical ai for synthetic imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4430–4441, 2025. doi: 10.1109/WACV61041.2025.00435.

Y. Jeong, Hye Bin Yoo, and Il Yong Chun. Dx2ct: Diffusion model for 3d ct reconstruction from bi or mono-planar 2d x-ray(s). In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888986.

Yoni Kasten, Daniel Doktofsky, and Ilya Kovler. End-to-end convolutional neural network for 3d reconstruction of knee bones from bi-planar x-ray images. In *Machine Learning for Medical Image Reconstruction (MLMIR 2020)*, pp. 123–133, 2020. doi: 10.1007/978-3-030-61598-7_12.

Firas Khader, G. Mueller-Franzes, S. T. Arasteh, T. Han, Christoph Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baessler, S. Foersch, J. Stegmaier, C. Kuhl, S. Nebelung, Jakob Nikolas Kather, and D. Truhn. Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 13(7303), 2023. doi: 10.1038/s41598-023-34341-2.

Daeun Kyung, Kyungmin Jo, J. Choo, Joonseok Lee, and E. Choi. Perspective projection-based 3d ct reconstruction from biplanar x-rays. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096296.

Yiqun Lin, Zhongjin Luo, Wei Zhao, and Xiaomeng Li. Learning deep intensity field for extremely sparse-view cbct reconstruction. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 13–23, 2023. doi: 10.1007/978-3-031-43999-5_2.

Xu-Hui Liu, Zhi Qiao, Runkun Liu, Hong Li, Juan Zhang, Xiantong Zhen, Zhen Qian, and Baochang Zhang. Diffux2ct: Diffusion learning to reconstruct ct images from biplanar x-rays. In *Computer Vision – ECCV 2024*, pp. 458–476, 2024. doi: 10.1007/978-3-031-72775-7_26.

Chenglong Ma, Zilong Li, Junping Zhang, Yi Zhang, and Hongming Shan. Freeseed: Frequency-band-aware and self-guided network for sparse-view ct reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 250–259, 2023.

E. Sidky and Xiao-Chuan Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in medicine and biology*, 53:4777 – 4807, 2008.

Li Sun, Junxiang Chen, Yanwu Xu, Mingming Gong, K. Yu, and K. Batmanghelich. Hierarchical amortized gan for 3d high resolution medical image synthesis. *IEEE journal of biomedical and health informatics*, 26:3966 – 3975, 2022.

Cuong Tran Van, Trong-Thang Pham, Ngoc Son Nguyen, D. Nguyen, and Ngan Le. Dufal: Dual-frequency-aware learning for high-fidelity extremely sparse-view cbct reconstruction. *Trans. Mach. Learn. Res.*, 2026, 2026.

Haoshen Wang, Zhentao Liu, Kaicong Sun, Xiaodong Wang, Dinggang Shen, and Zhiming Cui. 3d meddiffusion: A 3d medical latent diffusion model for controllable and high-quality medical image generation. *IEEE Transactions on Medical Imaging*, 44:4960–4972, 2025. doi: 10.1109/TMI.2025.3585372.

Yufeng Wang, Zhan-Li Sun, Zhigang Zeng, and Kin-Man Lam. Trct-gan: Ct reconstruction from biplane x-rays using transformer and generative adversarial networks. *Digit. Signal Process.*, 140: 104123, 2023.

Xing Xie, Jiawei Liu, Hui-Jie Fan, Zhi Han, Yandong Tang, and Liang-Qiong Qu. Dvg-diffusion: Dual-view-guided diffusion model for ct reconstruction from x-rays. *IEEE Transactions on Image Processing*, 35:1158–1173, 2026. doi: 10.1109/TIP.2026.3655171.

Jinzhong Yang, Harini Veeraraghavan, Wouter van Elmpt, Andre Dekker, Mark J. Gooding, and Gregory C. Sharp. Ct images with expert manual contours of thoracic cancer for benchmarking auto-segmentation accuracy. *Medical physics*, 47:3250 – 3255, 2020.

Xingde Ying, Heng Guo, Kai Ma, Jian Wu, Z. Weng, and Yefeng Zheng. X2ct-gan: Reconstructing ct from biplanar x-rays with generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10611–10620, 2019. doi: 10.1109/CVPR.2019.01087.

Ruyi Zha, Yanhao Zhang, and Hongdong Li. Naf: Neural attenuation fields for sparse-view cbct reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 442–452, 2022.

Chu-Long Zhang, Lin Liu, Jingjing Dai, Xuan Liu, Wenfeng He, Yinping Chan, Yao-Qin Xie, F. Chi, and Xiaokun Liang. Xtransct: ultra-fast volumetric ct reconstruction using two orthogonal x-ray projections for image-guided radiation therapy via a transformer network. *Physics in Medicine and Biology*, 69, 2024. doi: 10.1088/1361-6560/ad3320.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Zhiyu Zhang, Cong Shen, Ji-Jun Tang, and Zhi-Jun Liao. Dct-net: Dual-branch ct reconstruction from orthogonal x-rays with diffusion model and contrastive learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 150–160, 2025. doi: 10.1007/978-3-032-04947-6_15.

SUPPLEMENTARY MATERIAL

This supplement provides additional details on data construction, model implementation, and evaluation, together with a higher-resolution component ablation and additional qualitative comparisons.

A DATA CONSTRUCTION AND PREPROCESSING

Acquiring clinical radiographs and CT volumes with the exact geometric calibration and spatial correspondence required for supervised reconstruction is challenging. We therefore construct paired biplanar projections and CT volumes using a fixed virtual cone-beam geometry, following the synthetic-pair strategy used in prior X-ray-to-CT reconstruction work (Ying et al., 2019). Each projection pair has a known source volume and a consistent detector-to-volume mapping. LIDC-IDRI provides 1,018 thoracic CT cases, with 916 cases assigned to reconstruction training and 102 held out for testing. The split is established before paired files are prepared, and the same preprocessing and projection geometry are applied to both subsets. We additionally use all 60 LCTSC cases for external evaluation without dataset-specific adaptation.

A.1 LIDC-IDRI CT PREPROCESSING AND COORDINATE CONVENTIONS

For each LIDC-IDRI case, preprocessing starts from the thoracic CT array of size 256^3 stored in the raw HDF5 file. Stored values are clipped to $[0, 2500]$, divided by 2500, and trilinear resampled to 128^3. The clipping interval refers to the stored array values, rather than an independently specified HU window.

Let V_R denote the normalized 128^3 array in the raw-HDF5 orientation. Fixed axis permutations and flips map V_R into the TIGRE coordinate system for forward projection. A separate fixed mapping converts V_R into the diffusion coordinate system to obtain the target CT volume V_D. The autoencoder receives $2V_D - 1$; its packed latent is subsequently scaled using the pretrained autoencoder's codebook range. These orientation mappings keep the projection inputs, target CT, and latent representation anatomically aligned despite their different array conventions. The encoding and packing procedure is detailed in Section B.1.

Each paired HDF5 file stores the diffusion-coordinate target CT (128^3), the raw-coordinate CT used to construct the projections (128^3), the corresponding scaled latent ($8 \times 32 \times 32 \times 32$), and two normalized 512×512 projection images. The training loader reads these precomputed projections instead of evaluating the forward projector for every batch.

A.2 SYNTHETIC PROJECTION GEOMETRY AND DRR NORMALIZATION

A digitally reconstructed radiograph (DRR) is obtained by projecting a volume through a specified source–detector geometry. In the idealized physical model of Eq. (1), each detector value is a line integral of the attenuation field along the corresponding ray. In our implementation, the TIGRE forward operator (Biguri et al., 2025), implemented by `Ax`, acts on a clipped and normalized CT array:

$$p_v = \mathcal{A}_{g,\theta_v}(V_T), \qquad v \in \{\mathrm{PA}, \mathrm{Lat}\}, \tag{22}$$

where V_T is V_R reoriented to TIGRE coordinates, g denotes the fixed computational geometry, and θ_v specifies the view. These synthetic projections have a defined geometric relationship to the CT grid, but their pixel values are not calibrated physical radiographic intensities.

The cone-beam projector uses source-to-origin and source-to-detector distances of 5 and 7, respectively. The CT grid contains 128^3 voxels over a configured volume size of $(2, 2, 2)$, and the detector contains 512×512 pixels over a configured size of $(3.05, 3.05)$. Source and detector offsets are zero. Under the prescribed orientation convention, the angles $\pi/2$ and π are stored in this order as posteroanterior (PA) and lateral (Lat) views. We report these numerical settings in the implementation's geometry units without assigning an unverified physical length unit.

Each view is independently min–max normalized to obtain the input x_v used in the main paper:

$$x_v = \begin{cases} \dfrac{p_v - \min p_v}{\max p_v - \min p_v}, & \max p_v > \min p_v, \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

The paired files retain the complete detector images without cropping or resizing. Architecture-specific input transformations, where required by a comparison method, are applied subsequently. This projection-wise normalization is distinct from CT evaluation, which uses a fixed intensity range without per-case or per-slice min–max normalization.

A.3 LCTSC Preprocessing

The external cohort comprises all 60 LCTSC CT scans, including the 36 cases labeled Train and the 24 labeled Test in the original distribution. Both groups are used exclusively for external reconstruction evaluation. We use only the CT DICOM images and do not use the accompanying annotations.

Slices are assembled according to DICOM spatial metadata and converted to Hounsfield units (HU). A field centered at the physical center of each scan, measuring $320 \times 320 \times 256$ mm along the left–right, anterior–posterior, and superior–inferior directions, respectively, is linearly resampled to 128^3 voxels. The resulting sampling intervals are $2.5 \times 2.5 \times 2.0$ mm. Locations outside the acquired scan are filled with -1024 HU. The HDF5 array stores $\mathrm{HU} + 1024$; these stored values are subsequently clipped to $[0, 2500]$ and divided by 2500.

Projection pairs are generated using the same orientation convention, fixed cone-beam settings, and per-view normalization described in Section A.2. The TIGRE volume size remains $(2, 2, 2)$ and is not recalibrated to the physical sampling intervals of the LCTSC crop. Thus, this experiment evaluates transfer across CT cohorts under a shared computational projection geometry. Its results do not establish generalization to clinical radiographs or to different acquisition geometries.

B Architecture and Optimization Details

B.1 Frozen CT Autoencoder and Packed Latent Representation

We use the pretrained Patch-Volume Autoencoder of 3D MedDiffusion (Wang et al., 2025), implemented as a vector-quantized autoencoder (VQ-AE), with a spatial compression factor of 4 along each axis and eight latent channels. A 128^3 input volume is divided into eight non-overlapping 64^3 patches. Each patch is encoded independently into an eight-channel 16^3 latent. These patch latents are then assembled according to their original spatial locations, producing a packed tensor of shape $[B, 8, 32, 32, 32]$, where B is the batch size. Patch assembly therefore occurs along the spatial axes; the number of latent channels remains eight.

Consistent with the main paper, E_{CT} and D_{CT} include the packing and latent-scaling conventions of the pretrained model. The CT encoder, CT decoder, and pretrained volumetric denoiser remain frozen throughout reconstruction training.

B.2 Geometric and Spatial–Frequency Conditioning

The geometric pathway uses the pretrained SDXL VAE encoder as the frozen two-dimensional feature extractor E_{X}. Each projection is encoded into f_v, and each feature channel is lifted onto the common volumetric grid with the calibrated TIGRE back-projection operator $\mathcal{B}_v^b$ in Eq. (3). The PA and Lat feature volumes are concatenated to form F_{bp}. Back-projected outputs are reoriented into diffusion coordinates, and the alignment projection P_b matches the condition to the noisy CT latent at the control input.

The geometric back-projection operator has no learned parameters, whereas the X-ray encoder has pretrained parameters that are held fixed. The associated trainable alignment and control projections are part of the conditioning and control pathway specified in Section 3.3 of the main paper. Back-projection supplies a geometry-aligned condition rather than an exact inverse reconstruction from

two views. Its sampling weights account for the feature-map resolution under the fixed computational geometry.

The implemented Frequency-Enhanced Dual-Encoding (FEDE) module uses four resolution scales ($L = 4$), with five blocks distributed across those scales. This configuration reduces the number of scales from the initial five-scale design to control the parameter count and support the subsequent volumetric lifting. The number of blocks should therefore be distinguished from the number of distinct spatial resolutions.

As described in the main paper, each encoding branch shares its weights across PA and Lat views. The spatial hierarchy is complemented by global and patch-local Fourier processing; frequency-domain cross-attention fuses the corresponding spatial and spectral features. The fused features are lifted using calibrated perspective projection and bilinear sampling to obtain the scale-specific volumetric conditions $\{F^{V,s}\}_{s=1}^{L}$. These conditions enter the corresponding control stages, while F_{bp} enters at the control input. The FEDE encoder, control branch, and associated feature projections are trainable. Zero-initialized output projections inject their residuals into the frozen diffusion backbone.

B.3 Training and Inference Configuration

Consistent with the main paper, our 128^3 model and its 128^3 ablation variants are implemented in PyTorch and trained for 200 epochs on a single NVIDIA A800 GPU, with the same optimizer settings and a learning rate of 10^{-4}. A batch size of two and two gradient-accumulation steps yield an effective batch size of four. At inference, these variants use 1000-step DDIM sampling with $\eta = 0$, followed by decoding with D_{CT}. The external-evaluation implementations of C-MedDiffusion and FEDE-ControlNet use batch size one and a 1000-step diffusion schedule, with 1000 DDIM sampling steps and $\eta = 0$.

C Evaluation Protocol

C.1 Reference Volumes and Intensity Alignment

We evaluate reconstructed volumes against the original CT data, rather than autoencoder reconstructions or selected visualization slices. For 128^3 evaluation, the reference CT is clipped to the stored-value range $[0, 2500]$, divided by 2500, resampled to the target grid, and placed in the evaluation orientation. For LIDC-IDRI this includes trilinear resampling from the stored 256^3 array; LCTSC volumes are prepared as described in Section A.3.

For our diffusion model, the decoded prediction in the nominal $[-1, 1]$ range is mapped to $[0, 1]$ by $(\widehat{y} + 1)/2$, clipped to $[0, 1]$, and aligned with the reference. All comparison methods are evaluated in the same fixed $[0, 1]$ intensity domain after their required output conversion and orientation alignment. No per-case or per-slice min–max normalization is applied to CT volumes during evaluation.

C.2 Slice-Based and Volumetric CT Metrics

For a reference G and prediction P in $[0, 1]$, we compute

$$\mathrm{PSNR}(G, P) = \min\{100, -10 \log_{10}[\max(\mathrm{MSE}(G, P), 10^{-10})]\} \ \mathrm{dB}. \tag{24}$$

Slice PSNR is the arithmetic mean of the PSNR values over all axial, coronal, and sagittal slices: $3 \times 128 = 384$ slices per 128^3 case. Vol. PSNR is computed once from the MSE over all voxels in the complete volume. These quantities differ because averaging slice-level PSNR values is not equivalent to applying the logarithm to the volume-level MSE. Higher PSNR indicates lower squared intensity error.

We use `skimage.metrics.structural_similarity` with `data_range=1.0` and the remaining parameters at the installed library's defaults. Slice SSIM is the arithmetic mean over the same 384 slices per case. Vol. SSIM is obtained by applying this function directly to the two complete 128^3 arrays; it is not an average of two-dimensional SSIM scores. Higher values indicate greater similarity in local luminance, contrast, and structure.

Table 5: FEDE ablation at 256^3 on the LIDC-IDRI evaluation set. Both configurations include BP conditioning and omit VFM supervision.

Method	Slice PSNR ↑	Slice SSIM ↑	Vol. PSNR ↑	Vol. SSIM ↑	Proj. PSNR ↑	Proj. SSIM ↑
C-MedDiffusion (Baseline)	26.9370	0.6605	21.8915	0.6400	28.8500	0.9287
FEDE-ControlNet (w/o Feature supervision)	**27.2167**	**0.6646**	**22.1344**	**0.6459**	**29.5953**	**0.9408**

Learned Perceptual Image Patch Similarity (LPIPS) measures differences in pretrained visual features rather than only pointwise intensity error; lower values indicate greater perceptual similarity. We use a two-dimensional, VGG16-based LPIPS model with pretrained VGG weights and learned linear calibration weights. Each grayscale CT slice in $[0, 1]$ is replicated across three channels and mapped to $[-1, 1]$ before evaluation. The model compares channel-normalized features at five VGG stages, applies the learned linear weights, spatially averages each stage, and sums the stage scores. The reported *LPIPS* is the arithmetic mean of these scores over all 384 slices of each case; it is not a three-dimensional LPIPS computation.

We first obtain each metric for each case and then report the arithmetic mean across evaluated cases. In the 128^3 LIDC-IDRI evaluation, PSNR, Vol. PSNR, SSIM, Vol. SSIM, and LPIPS correspond to the implementation's `slice_psnr`, `volume_psnr`, `slice_ssim`, `volume_ssim`, and `slice_lpips`, respectively. The current cross-dataset LCTSC evaluation reports the first four CT metrics but does not compute LPIPS.

D Additional Experiments and Qualitative Results

D.1 Higher-Resolution FEDE Ablation

We additionally compare C-MedDiffusion with FEDE-ControlNet without VFM supervision on 256^3 LIDC-IDRI volumes. Both configurations retain BP conditioning. This comparison examines the benefit of adding FEDE at a higher reconstruction resolution; it does not evaluate the complete model with VFM supervision at 256^3.

As shown in Table 5, adding FEDE improves Slice PSNR by 0.2797 dB, Slice SSIM by 0.0041, Vol. PSNR by 0.2429 dB, and Vol. SSIM by 0.0059. Projection PSNR and SSIM increase by 0.7453 dB and 0.0121, respectively. These results support the usefulness of spatial–frequency conditioning at 256^3 in this comparison, with improvements in both CT reconstruction and projection consistency. The effects of adding VFM supervision at this resolution are not established by this two-configuration ablation.

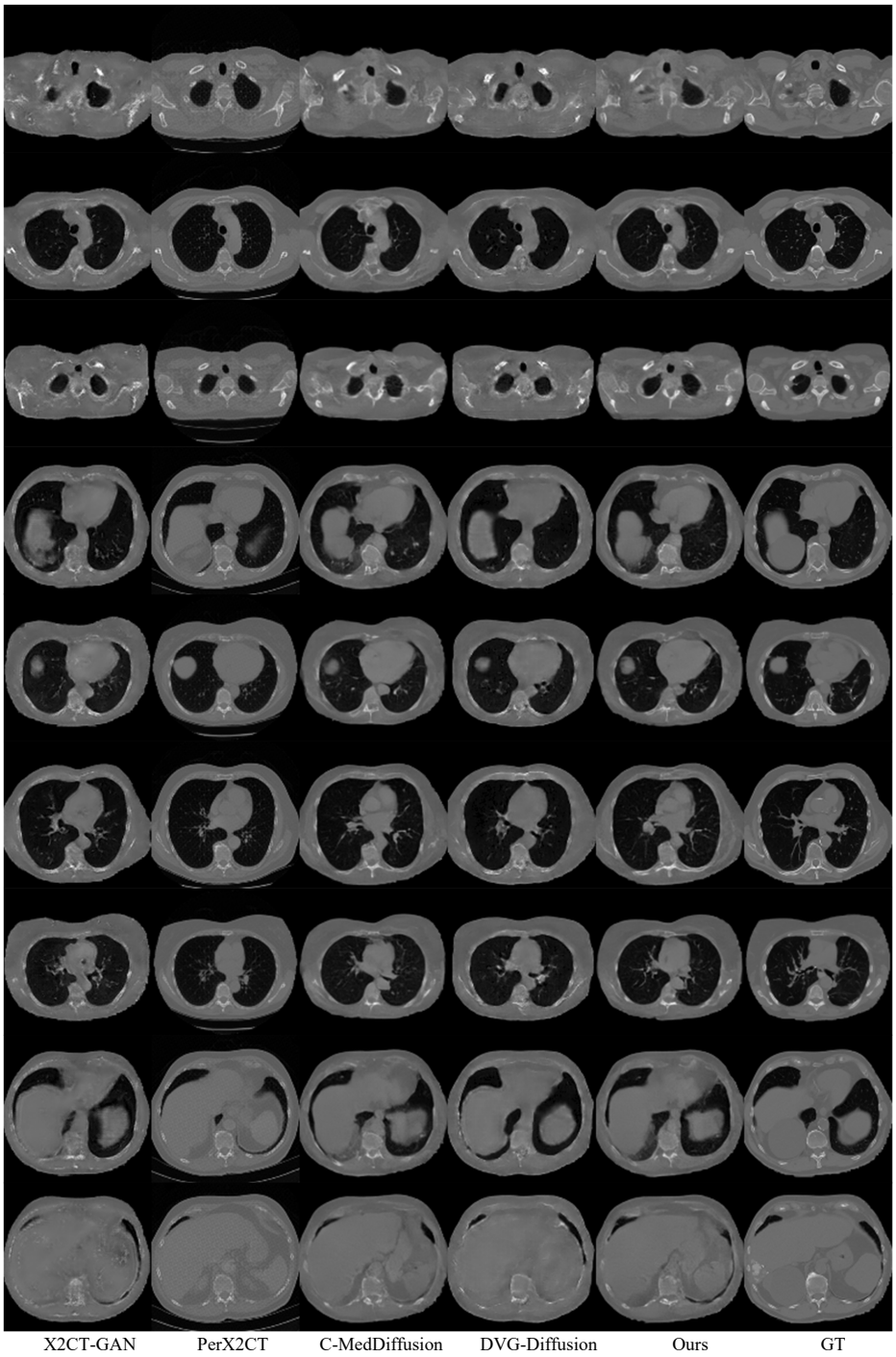


Figure 4: Additional axial CT comparisons on LIDC-IDRI. Each row contains anatomically corresponding slices.

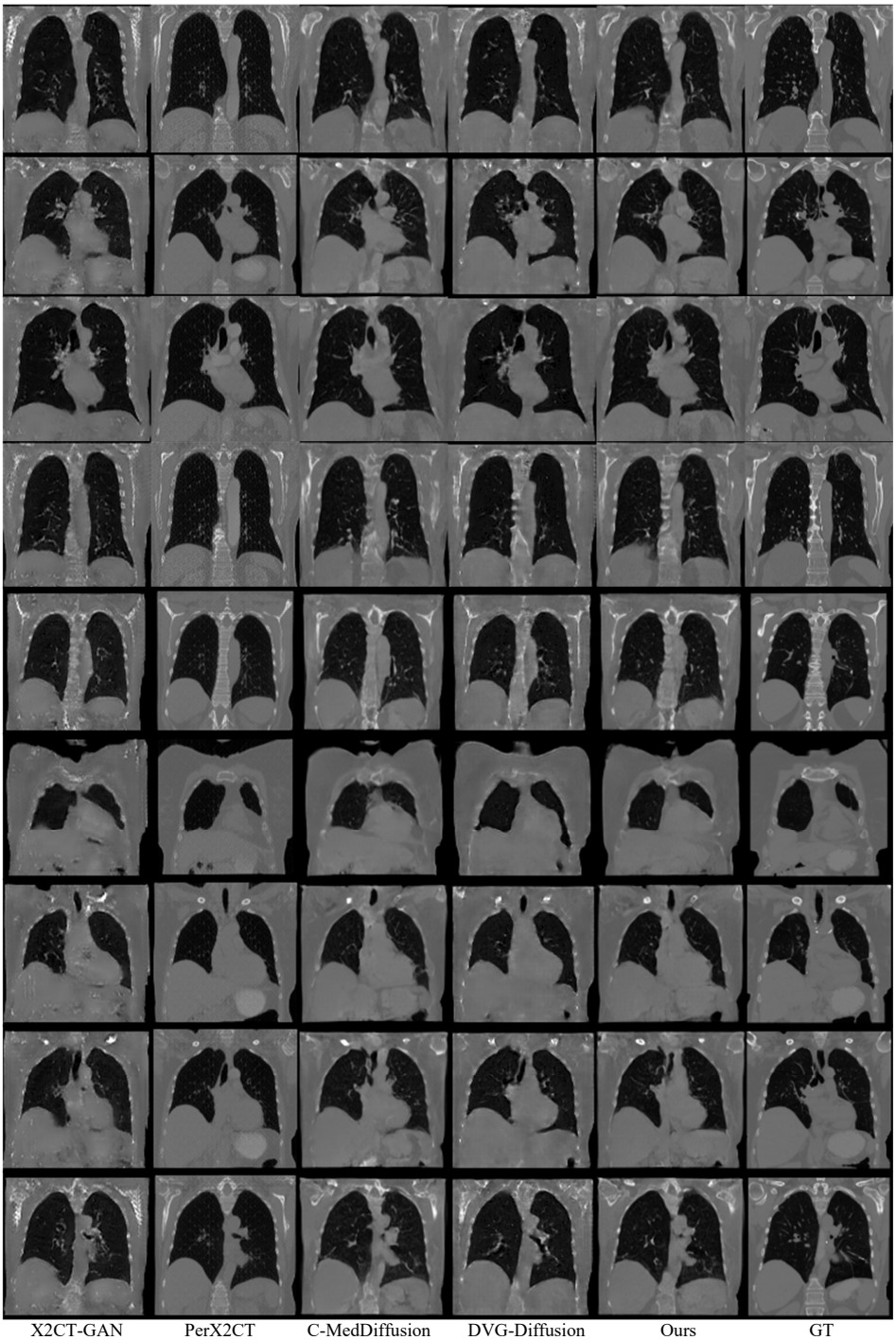


Figure 5: Additional coronal CT comparisons on LIDC-IDRI. Each row contains anatomically corresponding slices.

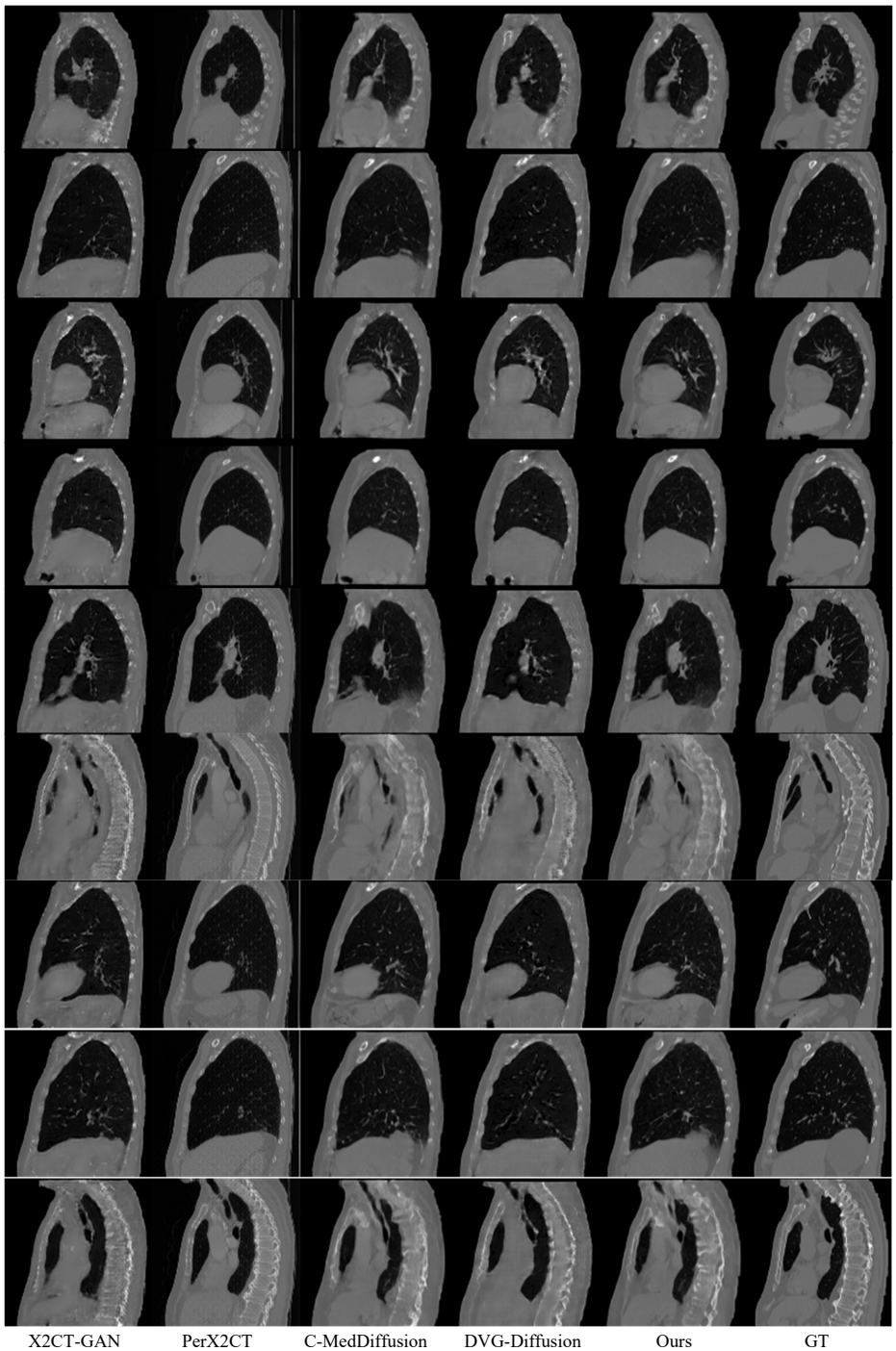


Figure 6: Additional sagittal CT comparisons on LIDC-IDRI. Each row shows anatomically corresponding slices.

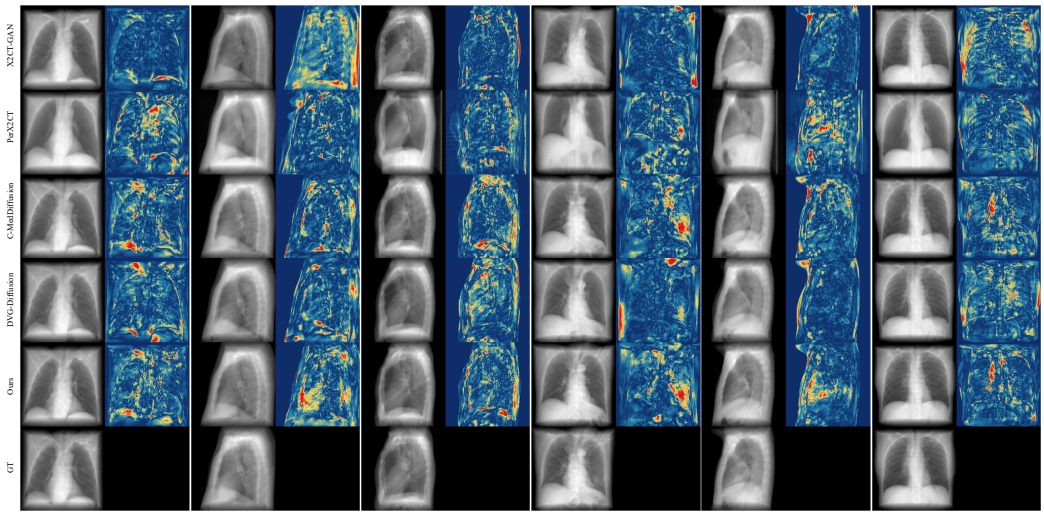


Figure 7: Projection-domain comparisons on LIDC-IDRI. Rows show X2CT-GAN, PerX2CT, C-MedDiffusion, DVG-Diffusion, FEDE-ControlNet (Ours), and GT. Columns contain representative PA and lateral DRRs. Color maps adjacent to reconstructed projections show discrepancies relative to the corresponding GT-derived views.

D.2 CT Reconstruction Comparisons

Figures 4–6 provide additional LIDC-IDRI examples in the axial, coronal, and sagittal planes. Each row compares anatomically corresponding slices reconstructed by X2CT-GAN, PerX2CT, C-MedDiffusion, DVG-Diffusion, and FEDE-ControlNet with the ground-truth (GT) CT. Across the displayed cases, our method generally preserves the extent of the lung fields and the relative positions of major thoracic structures, while retaining more recognizable internal detail.

The axial examples in Figure 4 show closer agreement with GT around the lung-field contours, mediastinum, and adjacent soft tissue in several cases. The coronal examples in Figure 5 illustrate the representation of lung boundaries, diaphragmatic contours, and central airway or vascular structures. The sagittal examples in Figure 6 show pulmonary branching patterns and the relative positions of the chest wall, diaphragm, and upper abdominal tissue. Fine structures remain imperfect. These selected examples complement the aggregate CT metrics; visual sharpness alone does not establish anatomical accuracy.

D.3 Projection-Domain Comparisons

Figure 7 compares DRRs re-projected from the reconstructed CT volumes with the corresponding GT-derived projections. Grayscale panels show representative PA and lateral views, and adjacent color maps visualize discrepancies from the corresponding GT projections. These views provide a complementary check on consistency with the observed biplanar anatomy. In several displayed examples, FEDE-ControlNet more closely follows the lung silhouette and central thoracic structures, with fewer prominent discrepancies. The trend agrees with the projection-domain metrics in Table 3 of the main paper, although localized projection errors remain.

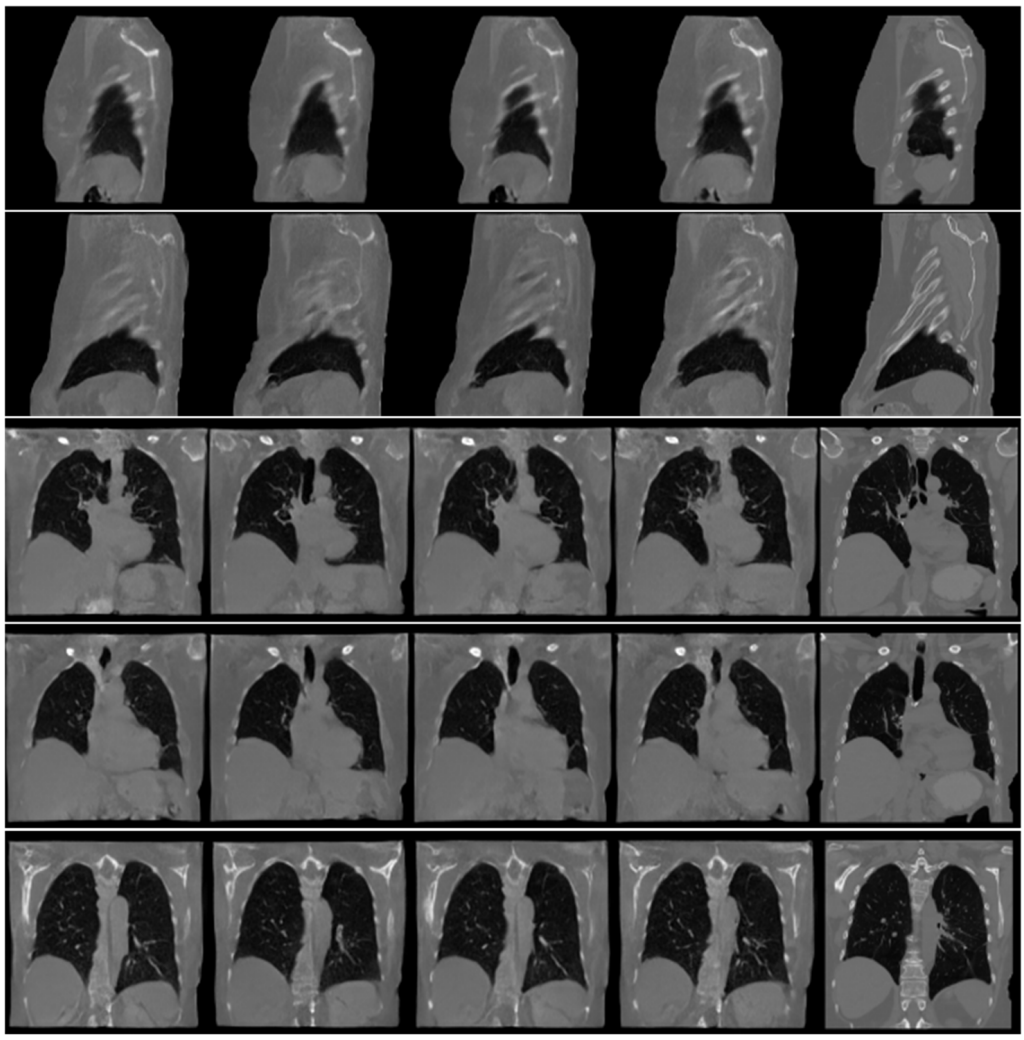


Figure 8: Additional CT-space ablation examples on LIDC-IDRI. Columns show the BP baseline, BP+FEDE, BP+VFM supervision, the complete model (Complete), and GT. The first two rows are sagittal slices and the remaining three rows are coronal slices.

D.4 CT-Space Component Ablations

Figure 8 extends the CT-space ablation to additional sagittal and coronal slices. All four learned variants retain BP conditioning and the frozen volumetric prior. In the figure labels, *BP Module* denotes the C-MedDiffusion baseline without FEDE or VFM supervision; *BP+FEDE Module* adds FEDE; *BP+VFM supervision* adds training-only VFM supervision to the baseline; and *Complete* denotes the complete FEDE-ControlNet. The BP baseline is therefore a learned reconstruction model, rather than the output of analytical back-projection alone.

The baseline captures the coarse thoracic layout, but some fine tissue and bone boundaries are less stable. Adding FEDE yields more recognizable local detail, while VFM supervision complements the denoising objective with structural constraints on decoded volumes. In the displayed cases, the complete model more consistently follows the reference lung contours, internal pulmonary structures, and skeletal detail. These observations complement the factorial ablation metrics in Table 2 of the main paper.

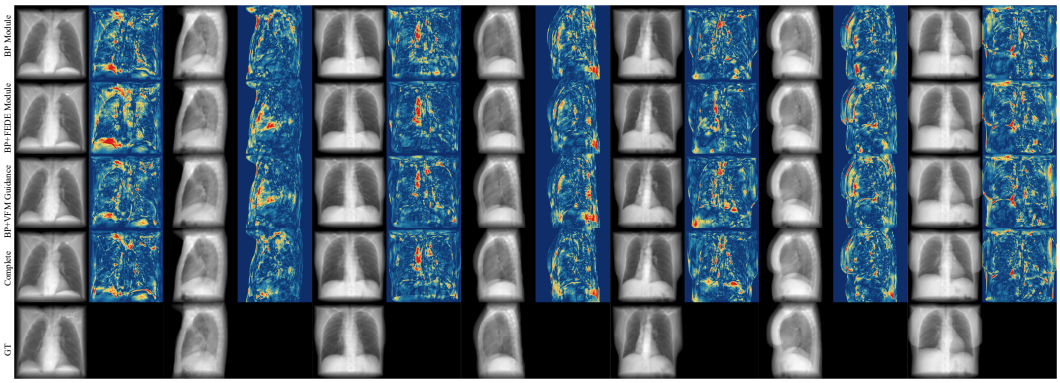


Figure 9: Projection-domain ablation examples on LIDC-IDRI. Rows show the BP baseline, BP+FEDE, BP+VFM supervision, the complete model (Complete), and GT. Columns show representative PA and lateral DRRs, with adjacent color maps indicating discrepancies from the corresponding GT-derived projections. Variant labels have the same meanings as in Figure 8.

D.5 Projection-Domain Component Ablations

Figure 9 shows re-projected DRRs and discrepancy maps for the same ablation configurations. The BP baseline captures the main projected anatomy, while FEDE and VFM supervision improve the fidelity of finer structures and boundaries in several examples. The complete model generally follows the corresponding GT views more closely, although localized discrepancies persist around high-contrast anatomical boundaries. Together with Figure 8, this comparison illustrates changes in both CT appearance and projection consistency, without implying that two-view agreement uniquely determines the underlying volume.