Extremely Sparse-View Computed Tomography from 2D Projections via Pose-Aware Diffusion Priors

Linrui Dai[1, 2] **Jinxiu Liang**[2*] **Chu Zhou**[2]
Mae Yamaguchi[1, 2] **Ryuya Takahashi**[1, 2] **Imari Sato**[1, 2]

[1]The University of Tokyo
[2]National Institute of Informatics

Abstract

Extremely sparse-view computed tomography (xSV-CT) is critical when scanning time, motion, or radiation tolerance limits each scan to only a few projections. In many such scenarios, the same acquisition limits also constrain dataset construction, making pre-reconstructed 3D volumes from dense scans unavailable. This exposes a gap in current sparse-view approaches: Optimization-based methods, which use the sparse measurements directly but lack data-driven priors to suppress artifacts, and learning-based methods, which use strong priors but usually require densely scanned training volumes acquired outside the xSV-CT regime. We study this train-sparse / infer-sparse setting by learning a pose-aware diffusion prior from one 2D projection per training sample, then guiding diffusion sampling with measured-view fidelity, shared-volume reprojection, and projection-physics consistency for tomographic reconstruction. On two public CT datasets across 1-, 4-, and 8-view settings, the proposed method achieves the best novel-view synthesis scores and best or competitive reconstruction quality compared with baselines that use dense-volume supervision.

1 Introduction

Computed tomography (CT) is essential for 3D diagnosis, but dense angular sampling is time-consuming and radiation-intensive. While sparse-view CT mitigates these problems by reconstructing from a sparser set of projections [1–4], several applications demand more extreme reduction, allowing only a few snapshots. This includes intra-operative C-arm navigation at 3-8 views [5], temporally dynamic anatomy imaging such as cardiac CT [6], and radiation-sensitive specimens that reach their absorption limit within one or two scans [7]. These scenarios motivate an extremely sparse-view CT (xSV-CT) regime: high-quality reconstruction from fewer than 10, or even a single, projection. Unlike prior formulations that aim to reduce radiation footprint at reconstruction [2–4], these applications impose acquisition constraints that extend beyond test-time reconstruction: The same acquisition limits can also make densely sampled projections or pre-reconstructed 3D training volumes difficult to collect for the target regime on a large scale for model training.

Existing SV-CT methods cover complementary sides of this acquisition problem. *Optimization-based methods* recover a 3D representation per sample using only its sparse input projections as supervision [3, 8–12]. This formulation automatically admits the xSV-CT regime. However, when pushed to extreme sparsity, their lack of anatomical priors to constrain the solution space results in severe reconstruction artifacts (Fig. 1(**a**)). *Learning-based methods* learn strong statistical priors from large-scale datasets of pre-reconstructed 3D volumes [2, 4, 13–16] and have demonstrated anatomically coherent reconstructions from 4-10 views. Their training signal usually comes from data acquired outside the target xSV-CT regime: each training volume or slice assumes sufficient

*Corresponding author (jinxiu_liang@nii.ac.jp).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

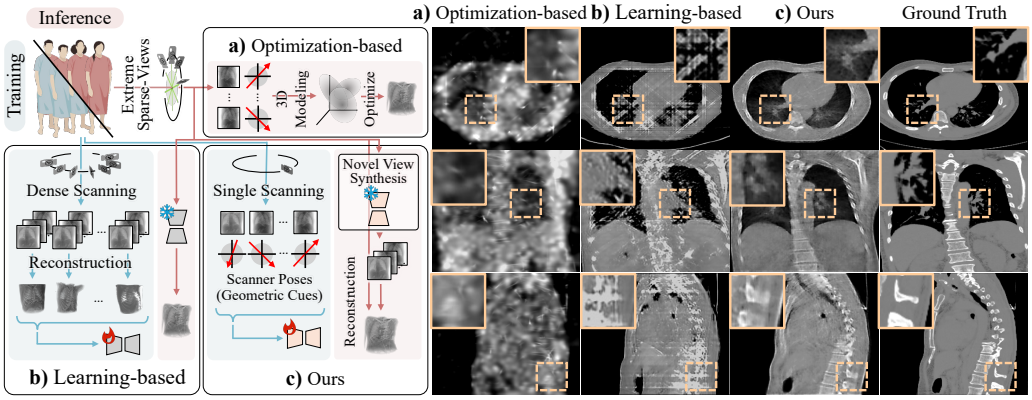

Figure 1: Tomographic reconstruction results from extremely sparse views at $0°$, $45°$, $90°$, $135°$. Current methods cover complementary sides of the extremely sparse view reconstruction problem: **a**) optimization-based methods [8] directly reconstruct from input views but produce severe artifacts when pushed to extreme sparsity at inference. **b**) Learning-based methods benefit from data-driven priors, but typically require reconstructed datasets from dense scans that are difficult to obtain for certain applications. **c**) To simultaneously respond to the sparsity limit constraining both the training and inference process, our method learns a pose-aware projection-domain diffusion prior *from a single 2D projection per training sample*, then uses physics-guided sampling to produce tomographically consistent novel projections for accurate reconstruction.

angular coverage for reliable reconstruction (Fig. 1(**b**)). For fragile, dynamic, or dose-sensitive targets, this training-data requirement becomes the main bottleneck.

Can we reduce X-ray scans at both training and inference? This setting prompts us to change the learning signal. A single training projection does not identify an individual 3D volume, but across many objects, projection–pose pairs define a pose-conditioned distribution of physically plausible measurements. If samples from this distribution are constrained at inference to agree with measured views and tomographic consistency, the learned projection prior can supply missing angular information without dense training scans. Motivated by this observation, we train a diffusion model on pairs of a projection and its acquisition angle, accessing *only one 2D projection per training sample*. At inference, we use this model to synthesize a dense set of candidate projections from sparse-view inputs for standard reconstruction (Fig. 1(**c**)).

Following this strategy, we separate the problem into a learning component and an inverse-problem component. (**1**) For training (Fig. 2(**a**)), we leverage the strong distribution modeling capability from diffusion models widely used in CT [17] and propose a *pose-aware projection-domain diffusion model* that injects scanner-angle geometry via sinusoidal pose encoding and multi-scale cross-attention at every network depth. This lets a 2D diffusion backbone model pose-conditioned projection statistics, including the reflection symmetry between complementary parallel-beam angles and the local sensitivity needed for adjacent views. (**2**) For inference (Fig. 2(**b**)), we introduce *consistency-aware guided generation* that steers the reverse diffusion process with two constraints: a *geometric coherency* loss that ties generated and measured projections to a shared reconstructed volume, and a *physics-informed consistency* loss that enforces consistency among generated views by projecting them onto a valid sinogram.

We evaluate our formulation on two public CT datasets with projection-volume pairs simulated under conventional CT noise models or from a verified forward model CatSim [18]. Reconstruction volumes are held out from our method and used only as metric references. Across 1-, 4-, and 8-view settings, our method achieves the best novel-view synthesis scores and competitive reconstruction quality against learning-based methods trained on full 3D reconstructions. This shows that, unlike uncorrected or feed-forward generated projections [19], a guided projection-domain sampler optimized with appropriate signals could also provide valuable information for sparse-view CT reconstruction.

Our main contributions are threefold: (**1**) We study xSV-CT under a train-sparse / infer-sparse setting with arbitrary acquisition poses, learning a pose-aware projection-domain diffusion prior from a single 2D projection per training sample and thereby removing the need for dense-volume supervision

in the target data regime; (**2**) We integrate sinusoidal pose encoding with multi-scale cross-attention injection, enabling a 2D diffusion backbone to model pose-conditioned projection distributions from per-sample projections; (**3**) We embed a discrete physical surrogate directly inside diffusion sampling, coupled with a shared-volume geometric coherency loss, so generated dense projections are refined toward a common tomographic explanation.

2 Related Work

Optimization-based methods reconstruct each test case from its own sparse input projections, with no learned data prior from a training database. Classical analytical methods such as filtered back-projection (FBP) accurately reconstruct from dense angular coverage, but severely degrade under sparse sampling due to the Fourier Slice Theorem: insufficient angular coverage leaves the inverse problem fundamentally ill-posed [20–22]. Iterative methods tackle sparse view tomographic reconstruction by minimizing energy functions with regularization [23–26], often combining network-based [27, 28] or plug-and-play denoisers to enforce smoothness [29, 30]. Neural-representation methods further adapt NeRF-based implicit representations [9, 10] or 3D Gaussian splatting [3, 8, 11, 12, 31, 32] to the CT geometry, achieving strong results at 20–50 views. At fewer than 10 views, geometry alone leaves too much angular information missing, and artifacts become severe [33]. This motivates a learned projection-domain prior that can be trained from 2D measurements while preserving the projection-domain inference setting.

Learning-based methods learn strong statistical priors over 3D anatomy from large-scale datasets of pre-reconstructed CT volumes [2, 4, 13–16, 34]. Volume-based regression methods learn a deterministic mapping from sparse projections to 3D volumes via deep unrolling [35, 36], GANs [14], or implicit neural fields [2, 3, 13], supervised by ground-truth volumes obtained from dense-angle scans. Volume-based diffusion methods train 2D prior models on anatomical slices extracted from CT volumes [15, 34, 37, 38] and apply posterior inference to reconstruct slices conditioned on sparse measurements. Subsequent work has then substantially reduced the inference cost of such priors for large volumes [39–41]. The training priors of these methods are learned from volumes reconstructed with dense angular coverage. For radiation-sensitive or rapidly changing targets, this creates a mismatch between the training-data regime and the acquisition regime of interest. Our approach learns a pose-aware projection-domain diffusion prior directly from 2D projections, removing dense-volume supervision during training while keeping measured-view fidelity and tomographic consistency at inference.

Few-view X-ray-to-CT generation methods have shown that networks can infer volumetric structure from one or a few radiographs at predefined poses, using deep regression [42], adversarial training [43, 44], or diffusion-based multi-modal generation [45]. Like the learning-based methods above, they form their training pairs from reconstructed volumes and are designed around a fixed acquisition protocol, so changing the number or poses of input views typically requires redesign or retraining. Closer to our setting, Lai *et al.* [46] learn solely in the projection domain with a self-supervised neural field, but require 90 projections per training sample and a fixed biplanar protocol at inference. In contrast, we train on a single projection per sample and accept measurements at arbitrary poses, which admits a train-sparse setting and broadens potential downstream applications.

3 Method

The tomographic reconstruction problem can be formulated as finding the optimal solution to inverse the forward transform operator (*e.g.*, Radon) $\mathcal{A}_{\boldsymbol{\theta}}$ [37], given the projections $\mathbf{x} = \{\mathbf{x}_i\}$ from a set of input projection angles $\boldsymbol{\theta} = \{\theta_i\}$:

$$\mathbf{x}_i = \mathcal{A}_{\theta_i}\mathbf{y} + \mathbf{n}_i, \tag{1}$$

where $\mathbf{y}$ is the 3D volume and $\mathbf{n}_i$ is the noise tensor induced by photon scattering and electronic noise [47]. To achieve high-quality xSV-CT from extremely sparse sets of projections, we propose a framework that generates consistent projections for reconstruction via a projection-domain generative prior trained from a single 2D projection per training sample (Fig. 2). The prior models angle-conditioned projection statistics rather than identifying a 3D volume from one training view. At inference, measured-view fidelity and tomographic consistency turn this prior into a constrained reconstruction procedure (Sec. 3.2).

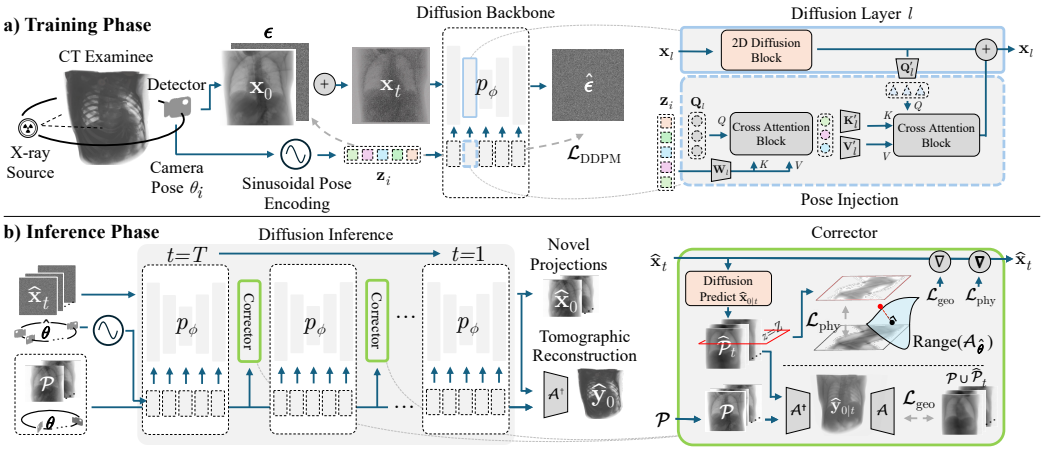

Figure 2: Overview of our framework. **a**) Training phase (Sec. 3.1): we learn a **pose-aware projection-domain diffusion prior** from 2D projections $\mathbf{x}_0$ paired with poses θ_i via network p_ϕ. Specifically, we prepare geometric cues θ_i via *sinusoidal pose encoding* and inject them with *multi-scale pose injection* (right). The model is trained with standard denoising loss $\mathcal{L}_{\text{DDPM}}$. **b**) Inference phase (Sec. 3.2): given M sparse input views $\mathcal{P}$ at poses $\boldsymbol{\theta}$, we generate N novel views $\hat{\mathbf{x}}_0$ at poses $\hat{\boldsymbol{\theta}}$ through **consistency-aware guided generation**. At each denoising timestep t, a *corrector module* (right) enforces two complementary constraints: *geometric coherency* loss $\mathcal{L}_{\text{geo}}$ pushes all generated and input views toward a shared 3D explanation via forward projection $\mathcal{A}$ and approximate inverse $\mathcal{A}^\dagger$; *physics-informed consistency* loss $\mathcal{L}_{\text{phy}}$ further enhances consistency among generated views $\hat{\mathcal{P}}_t$ by projecting them onto the consistent range of the Radon transform $\mathcal{A}_{\hat{\boldsymbol{\theta}}}$.

3.1 Pose-Aware Projection-Domain Diffusion Prior

We model the underlying data distribution of 2D X-ray projections using a denoising diffusion probabilistic model (DDPM) [48]. To generate projections that are consistent with different scanner angles at inference, we condition the model on the scanner pose at every layer of the network. Formally, during training the diffusion forward process progressively corrupts a clean projection $\mathbf{x}_0$ by adding Gaussian noise $\boldsymbol{\epsilon}$ over T timesteps via a variance schedule β_t ($0 \leq t \leq T$). At timestep t, the noisy projection $\mathbf{x}_t$ is generated as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{2}$$

where $\bar{\alpha}_t = \prod_{\tau=1}^{t} \alpha_\tau$ and $\alpha_t = 1 - \beta_t$. The diffusion model p_ϕ is then trained to reverse this degradation by predicting the added noise $\boldsymbol{\epsilon}$ from $\mathbf{x}_t$ via:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}} ||p_\phi(\mathbf{x}_t; \theta_i, t) - \boldsymbol{\epsilon}||_2^2, \tag{3}$$

where θ_i is the detector pose associated with the input projection $\mathbf{x}_0$. We promote the pose-awareness of our model p_ϕ through the following two components.

Sinusoidal Pose Encoding Due to X-ray transmissivity, projections from complementary angles can be closely related (*e.g.*, projections at 0° and 180° in Fig. 3 under parallel-beam geometry differ mainly by detector reflection), while adjacent angles can contain small but important changes. We encode each scalar angle θ_i (pose) into a K-dimensional feature vector $\mathbf{z}_i$ with a sinusoidal encoding module:

$$\mathbf{z}_i[k] = \sin((\lfloor k/2 \rfloor + 1)\,\theta_i + |k\pi/2|)\,. \tag{4}$$

The periodic basis gives the network a smooth angular coordinate system. It does not assume that complementary views are identical; instead, it exposes the symmetry and local angular variation in a form that cross-attention layers can learn from data. High-frequency components (k large) provide fine-grained angular sensitivity for detailed projection synthesis.

Multi-Scale Pose Injection Given encoded pose features $\mathbf{z}_i \in \mathbb{R}^K$ for θ_i, we inject the pose information throughout the diffusion backbone via layer-wise cross-attention mechanism [50]. At each layer l, we compute a layer-specific pose feature $\mathbf{z}_l$ through:

$$\mathbf{z}_l = \texttt{CrossAttn}\left(\mathbf{Q}_l, \mathbf{W}_l(\mathbf{z}_i), \mathbf{W}_l(\mathbf{z}_i)\right), \tag{5}$$

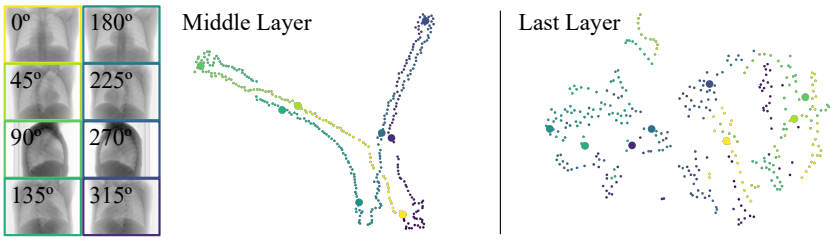

Figure 3: t-SNE visualizations [49] of learned features $\mathbf{z}_l$ at different layers. Larger dots in the feature space highlight features from the displayed sample projections. $\mathbf{z}_l$ at the middle layer (left) aligns features with the general structural distinctiveness of projections to resolve view ambiguity, while $\mathbf{z}_l$ at the last layer (right) retains high local sensitivity to individual angles, preserving a separable feature space for accurate pose-conditioned synthesis.

(See Fig.2(**a**), right), where $\mathbf{Q}_l$ is a learnable query and $\mathbf{W}_l$ is a learnable projection. The cross-attention operation is defined as $\texttt{CrossAttn}(Q, K, V) = \texttt{MLP}(\texttt{LayerNorm}(\texttt{Softmax}(\frac{QK^\top}{\sqrt{d}})V))$ with d being the size of the latent dimension of K [50]. This allows each layer to learn which components of the pose information $\mathbf{z}_i$ are the most relevant at its feature scale. As shown in Fig. 3, layers with larger receptive fields (middle layers) tend to collapse pose features based on the general structural distinctiveness of projections to resolve view ambiguity, while layers with smaller receptive fields (later layers) tend to preserve the fine-grained separability of adjacent views on the feature space for detailed synthesis. The pose feature $\mathbf{z}_l$ is then fused with the intermediate image feature map $\mathbf{x}_l$ via a second cross-attention module and a residual connection:

$$\mathbf{x}_l \leftarrow \texttt{CrossAttn}(\mathbf{Q}'_l(\mathbf{x}_l),\ \mathbf{K}'_l(\mathbf{z}_l),\ \mathbf{V}'_l(\mathbf{z}_l)) + \mathbf{x}_l, \tag{6}$$

where $\mathbf{Q}'_l, \mathbf{K}'_l, \mathbf{V}'_l$ are learnable projections mapping to a common latent dimension.

3.2 Consistency-Aware Guided Generation

After training the pose-aware projection-domain diffusion model p_ϕ, we generate a dense set of N novel projections $\hat{\mathbf{x}}_0$ at poses $\hat{\boldsymbol{\theta}}$ that are geometrically coherent with each other and consistent with the M sparse input measurements $\mathcal{P}$ at poses $\boldsymbol{\theta}$. Standard reverse sampling in DDPM produces clean $\hat{\mathbf{x}}_0$ from p_ϕ via repeated denoising from Gaussian noise [48]:

$$\hat{\mathbf{x}}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathbf{x}}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} p_\phi(\hat{\mathbf{x}}_t; \theta, t)\right) + \sqrt{\beta_t}\boldsymbol{\epsilon}. \tag{7}$$

For xSV-CT, batch-wise independently generated projections must be coupled to the measured views and to a common projection geometry. Inspired by classifier guidance [51], we guide the reverse diffusion process with geometry- and physics-based constraints.

At each denoising timestep t and novel pose $\hat{\theta} \in \hat{\boldsymbol{\theta}}$, we predict the clean projection $\hat{\mathbf{x}}_{0|t}$ through the standard denoising update [48]:

$$\hat{\mathbf{x}}_{0|t} = (\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t} p_\phi(\mathbf{x}_t; \hat{\theta}, t))/\sqrt{\bar{\alpha}_t} \in \hat{\mathcal{P}}_t, \tag{8}$$

Let $\hat{\mathcal{P}}_t$ denote the collection of $\hat{\mathbf{x}}_{0|t}$ evaluated at all novel projection angles $\hat{\boldsymbol{\theta}}$. We enforce consistency via two complementary constraints.

Geometric Coherency Loss To ensure geometric coherence, the generated dense projections $\hat{\mathcal{P}}_t$ at poses $\hat{\boldsymbol{\theta}}$ and the sparse measurements $\mathcal{P}$ at poses $\boldsymbol{\theta}$ must jointly correspond to a common 3D volume. Let $\mathcal{A}$ denote the forward Radon transform and $\mathcal{A}^\dagger$ its approximate inverse (FBP [52] for parallel- and fan-beam, FDK [23] for cone-beam). We reconstruct an intermediate volume $\hat{\mathbf{y}}_{0|t} = \mathcal{A}^\dagger_{\tilde{\boldsymbol{\theta}}}(\tilde{\mathcal{P}}_t)$ from all projections $\tilde{\mathcal{P}}_t = \mathcal{P} \cup \hat{\mathcal{P}}_t$ at combined poses $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} \cup \hat{\boldsymbol{\theta}}$. This intermediate volume can be decomposed into two components $\hat{\mathbf{y}}_{0|t} = \hat{\mathbf{y}}^C_{0|t} + \hat{\mathbf{y}}^R_{0|t}$, separately representing a reconstruction

from consistent parts in $\tilde{\mathcal{P}}_t$ and a residual volume [20]. To ensure $\hat{\mathbf{y}}^R_{0|t} \to \mathbf{0}$ as the denoising process progresses, our geometric coherency loss decomposes naturally into the following synergistic terms:

$$\mathcal{L}_{\text{geo}} = \underbrace{\|\mathcal{P} - \mathcal{A}_{\boldsymbol{\theta}}\hat{\mathbf{y}}_{0|t}\|_2^2}_{\text{measurement fidelity}} + \underbrace{\|\hat{\mathcal{P}}_t - \mathcal{A}_{\hat{\boldsymbol{\theta}}}\hat{\mathbf{y}}_{0|t}\|_2^2}_{\text{novel-view coherency}} + \mathcal{L}_{\text{TV}}(\hat{\mathbf{y}}_{0|t}). \tag{9}$$

The *measurement fidelity* term penalizes the discrepancy between the true sparse measurements $\mathcal{P}$ and their reprojection from $\hat{\mathbf{y}}_{0|t}$ at the original angles $\boldsymbol{\theta}$, ensuring the volume implied by the generated views is faithful to the observations. The *novel-view coherency* term enforces that the generated projections $\hat{\mathcal{P}}_t$ are themselves consistent with the same volume, pushing independently sampled denoising chains into alignment and preventing them from converging to mutually incompatible local anatomies. The total variation term $\mathcal{L}_{\text{TV}}$ further regularizes local intensity variations in $\hat{\mathbf{y}}_{0|t}$ [8, 28].

Physics-Informed Consistency Refinement $\mathcal{L}_{\text{geo}}$ enforces consistency *globally*: the generated views must reproject correctly from the reconstruction $\hat{\mathbf{y}}_{0|t}$. We pair this reconstruction-mediated constraint with a direct sinogram constraint derived from projection physics. In the continuous parallel-beam setting, valid Radon sinograms satisfy the Helgason–Ludwig Range Conditions [53]. We use a finite-angle, finite-order version of these conditions as a differentiable surrogate, which penalizes moment residuals outside the range of a coherent object. Consider the Radon transform operator $\mathcal{A}_{\hat{\boldsymbol{\theta}}}(r, \theta, z)$ on projection angles $\hat{\boldsymbol{\theta}}$ with r, θ, z representing a unit cylindrical coordinate system, its n-th moment curve $\mathcal{C}_n = \int_{-1}^{1} r^n \mathcal{A}_{\hat{\boldsymbol{\theta}}}(r, \theta, z)dr$ can be represented in a Fourier series:

$$\mathcal{C}_n = \alpha_0(z) + \sum_{i=1}^{\infty} \alpha_{2i-1}(z)\cos i\theta + \alpha_{2i}(z)\sin i\theta. \tag{10}$$

For fan-beam data, we apply this surrogate after standard rebinning to an approximate parallel-beam sinogram [54, 55]. This surrogate enters only $\mathcal{L}_{\text{phy}}$, which means $\mathcal{L}_{\text{geo}}$ relies solely on the forward operator $\mathcal{A}$ and its approximate inverse $\mathcal{A}^\dagger$, and therefore applies unchanged to cone-beam and non-coplanar geometries. The coefficient vector $\boldsymbol{\alpha}(z)$ for a slice at height z whose i-th component is $\alpha_i(z)$ satisfies the following Range Conditions for any order n: the i-th component of $\boldsymbol{\alpha}(z)$ is zero whenever $\lceil \frac{i}{2} \rceil + n$ is odd or $i > 2n$ [53, 56]. Consequently, $\boldsymbol{\alpha}(z)$ is sparse and possesses only $n+1$ degrees of freedom.

At each denoising step, we select a slice at height $z = Z$ and solve for this $(n+1)$-dimensional vector $\boldsymbol{\alpha}_0(Z)$ of non-zero coefficients in $\boldsymbol{\alpha}(z)$ via least squares. We construct a design matrix $\mathbf{A}_0 \in \mathbb{R}^{N\times(n+1)}$ based on the parity of n: Denoting Vandermonde-like matrices of trigonometric basis $\mathbf{C}_e, \mathbf{S}_e, \mathbf{C}_o, \mathbf{S}_o \in \mathbb{R}^{N\times\lceil \frac{n}{2} \rceil}$ whose i-th column vectors are $\cos 2i\hat{\boldsymbol{\theta}}, \sin 2i\hat{\boldsymbol{\theta}}, \cos(2i-1)\hat{\boldsymbol{\theta}}, \sin(2i-1)\hat{\boldsymbol{\theta}}$, $\mathbf{A}_0$ equals $[\mathbf{1}|\mathbf{C}_e|\mathbf{S}_e]$ when n is even and $[\mathbf{C}_o|\mathbf{S}_o]$ when n is odd. In practice, $(\mathbf{A}_0^\top \mathbf{A}_0)^{-1}$ is implemented with a numerically stable least-squares solver. The closed-form solution is:

$$\boldsymbol{\alpha}_0(Z) = (\mathbf{A}_0^\top \mathbf{A}_0)^{-1} \left(\mathbf{A}_0^\top \left(\int_{-1}^{1} r^n \hat{\mathcal{P}}_t(r, \theta, Z)dr \right) \right). \tag{11}$$

Our physics-informed consistency loss then refines the residual:

$$\mathcal{L}_{\text{phy}} = || \int_{-1}^{1} r^n \hat{\mathcal{P}}_t(r, \theta, Z)dr - \mathbf{A}_0\boldsymbol{\alpha}_0(Z)||_2^2. \tag{12}$$

This objective penalizes the distance from the generated views to the finite-order moment subspace. It complements $\mathcal{L}_{\text{geo}}$ by discouraging sinogram patterns that fit the diffusion prior locally but violate global projection physics.

Guided Denoising Update At each denoising step $t < T$, before performing reverse sampling with Eq. (7), we update each $\hat{\mathbf{x}}_t$ via gradient descent on the combined loss to steer its predicted clean projection $\hat{\mathbf{x}}_{0|t}$ toward geometric and physical consistency. The losses are evaluated with differentiable projection and reconstruction operators, so gradients flow through $\hat{\mathbf{x}}_{0|t}$ back to $\hat{\mathbf{x}}_t$. This guidance process is typically defined as [57]:

$$\hat{\mathbf{x}}_t \leftarrow \hat{\mathbf{x}}_t - \gamma \cdot \nabla_{\hat{\mathbf{x}}_t}\mathcal{L}, \tag{13}$$

where $\gamma > 0$ is a guidance scale (learning rate) and $\mathcal{L}$ is any loss function. In practice we use the Adam optimizer [58] with fixed learning rate γ_0 and moment decay rates β_1, β_2, which provides per-parameter adaptive gradient scaling for more stable convergence than vanilla gradient descent:

$$\hat{\mathbf{x}}_t \leftarrow \texttt{Adam}(\hat{\mathbf{x}}_t,\ \nabla_{\hat{\mathbf{x}}_t}(\lambda_1\mathcal{L}_{\text{geo}} + \lambda_2\mathcal{L}_{\text{phy}});\ \gamma_0, \beta_1, \beta_2)\,, \tag{14}$$

where λ_1, λ_2 are loss weights. This guided sampling process iteratively refines the ensemble of projections $\hat{\mathcal{P}}_t$, so that each final denoised output $\hat{\mathbf{x}}_0 \in \hat{\mathcal{P}}_0$ follows the learned projection statistics and remains geometrically coherent with all projections. The final tomographic reconstruction $\hat{\mathbf{y}}_0$ is obtained by applying any standard reconstructor to the combined sinogram $\mathcal{P} \cup \hat{\mathcal{P}}_0$.

The guided denoising update resembles DPS [59] and DDNM [60] in its construction of the consistency objectives through the Tweedie's estimate $\hat{\mathbf{x}}_{0|t}$, and its Range-Null space decomposition of $\hat{\mathbf{y}}_{0|t}$ to get consistent and residual volumes. Unlike Steerable Conditional Diffusion [61], which adapts the network to each measurement setting, one network serves every measurement configuration; we do not project directly onto the range space because $\mathcal{A}^\dagger$ introduces artifacts under sparse sampling. The key difference lies in the prior: DPS and DDNM assume a pretrained prior over the target (in our case, CT) domain, whereas ours is learned only from pose-conditioned projections and is paired with tomography-specific consistency objectives rather than generic data fidelity. We provide an algorithmic overview of this guided generation process in Appendix A.

4 Experiments

Dataset Curation We evaluate on xSV-CT regime with two public datasets separately on the pulmonary (CT-RATE) [62] and cranial (SinoCT) [63] cavities. We simulated 2D parallel-beam projections for CT-RATE using LEAP [64] package. For SinoCT, we use the provided fan-beam sinograms, which were simulated with GE's CatSim simulator [18] to model a polychromatic source and detector response. These public datasets were built from dense acquisitions, which lets us compute metrics against dense references; no component of our method accesses more than one view per object during training. More details can be found in Appendix B.

Controlled xSV-CT Simulation To evaluate against known ground truth while preserving the xSV-CT information pattern, we use these public CT datasets with three access rules: (i) at training, our model sees *one* 2D projection per case together with its scanner pose; (ii) at inference, only $M \in \{1, 4, 8\}$ projections plus their poses are exposed; (iii) the dense 3D volume is held out from every method component of ours and used only as a reference for metric computation. It is important to note that, unlike our method, concurrent learning-based methods require reconstructed slices or volumes at training and thus do not fully respect the xSV-CT regime.

Evaluation Metrics Following prior works [13, 61], we compute Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [65] between ground truths and reconstructions. We also use Learned Perceptual Image Patch Similarity (LPIPS) [66] to compare the perceptual authenticity of reconstructed volumes and projections. Reconstruction metrics are averaged over all axial slices and all images are rescaled to $[0, 1]$ before metric computation.

Implementation Details We implement our method with PyTorch [67] and train it on a 4 NVIDIA RTX 6000 Ada GPU cluster. Our generative training uses a constant learning rate of 4×10^{-5} with a batch size of 8 per GPU and a projection size of 256×256 for CT-RATE and 64×512 for SinoCT. Training uses AdamW [68] with momentum 0.99 and weight decay 10^{-5}. At inference, we select the degree of moment curve to be $n = 20$ in Eq. (11) with $\lambda_1 = 10^{-2}$ and $\lambda_2 = 1$. At N=180 on CT-RATE, inference takes 46–50 min for 4 to 32 input views on one GPU, comparable to recent accelerated slice-domain samplers [39, 40]. More details on dataset curation, method implementation and runtime analysis can be found in Appendix B.

4.1 Comparison Studies

We compare against 5 optimization-based methods, including classical analytical (FBP [52]), classical iterative (OS-SART [25]), and NeRF/Gaussian-volume methods (SAX-NeRF [10], R^2-Gaussian [11], 3DGR-CT [12]), as well as 3 learning-based methods, including a diffusion method trained on axial slices (DiffusionMBIR [37]) and two regression methods trained on dense 3D volumes (C^2RV [13], DSDF [14]). We evaluate the performance of each method on two tasks, **1**) 2D novel view synthesis and **2**) 3D tomographic reconstruction, with an extremely sparse set of $M \in \{1, 4, 8\}$ input views. To ensure fairness, we retrain each method according to their per-dataset or per-viewcount training schemes. More details can be found in Appendix B.

Novel View Synthesis As shown in Table 1, analytical reconstruction is strongly affected by the missing angular coverage in xSV-CT. Other optimization-based methods improve over FBP and

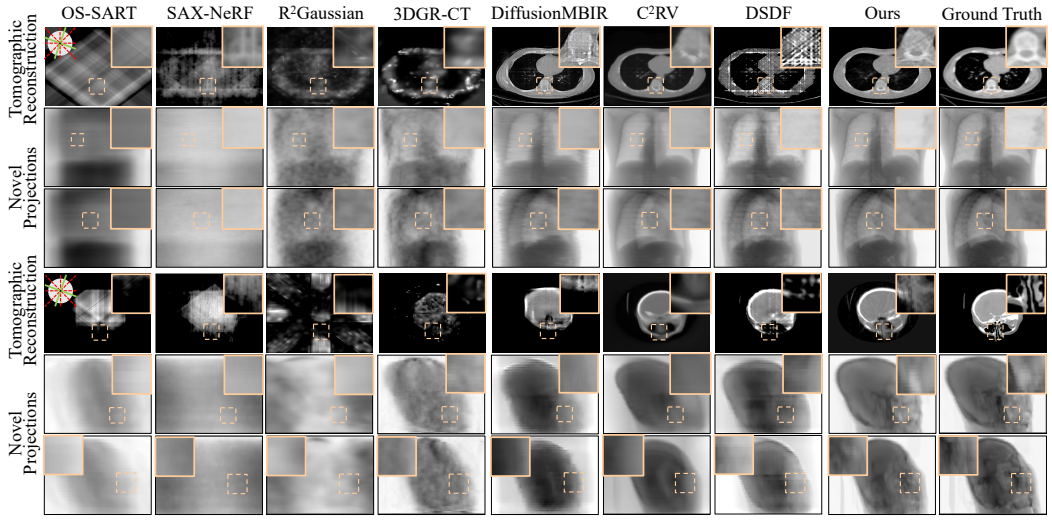

Figure 4: Qualitative comparison of tomographic reconstruction and novel view synthesis from extremely sparse (4) views on CT-RATE and SinoCT datasets. Input angles are marked in red dashed and novel angles in green solid lines in the inset polar plots.

Table 1: Quantitative comparison for novel view synthesis on a stratified subset of 50 test cases per dataset. Metrics are averaged across 180 synthesized novel projections covering the entire pose space. We highlight best, second best and third best results.

Metric		PSNR↑						SSIM↑						LPIPS↓					
Dataset		CT-RATE [62]			SinoCT [63]			CT-RATE			SinoCT			CT-RATE			SinoCT		
Inference Views		1	4	8	1	4	8	1	4	8	1	4	8	1	4	8	1	4	8
Optimization-based	FBP [52]	7.2	10.1	11.3	10.6	11.5	12.2	0.15	0.13	0.29	0.28	0.39	0.42	0.50	0.48	0.48	0.52	0.48	0.43
	OS-SART [25]	7.4	8.9	14.7	12.1	11.9	19.6	0.14	0.16	0.31	0.35	0.44	0.53	0.50	0.46	0.36	0.54	0.43	0.40
	SAX-NeRF [10]	15.0	19.7	20.5	11.4	12.3	15.1	0.32	0.56	0.61	0.28	0.42	0.49	0.43	0.42	0.39	0.52	0.43	0.40
	R^2-Gaussian [11]	14.1	15.4	25.8	11.2	13.6	16.8	0.46	0.52	0.59	0.21	0.31	0.44	0.45	0.45	0.40	0.58	0.51	0.46
	3DGR-CT [12]	15.0	24.6	26.8	11.5	12.6	16.1	0.48	0.51	0.61	0.28	0.44	0.49	0.45	0.40	0.37	0.49	0.46	0.39
Learning-based	DiffusionMBIR [37]	17.5	23.1	29.8	13.0	13.3	20.1	0.50	0.76	0.84	0.35	0.37	0.56	0.52	0.35	0.21	0.45	0.40	0.38
	C^2RV [13]	20.5	27.8	33.5	17.4	17.8	22.8	0.74	0.89	0.94	0.51	0.58	0.67	0.32	0.25	0.24	0.49	0.42	0.39
	DSDF [14]	21.1	23.6	27.3	18.1	18.1	20.6	0.75	0.81	0.90	0.56	0.59	0.69	0.39	0.32	0.21	0.48	0.45	0.40
	Ours	23.1	32.8	36.3	19.6	24.3	25.7	0.80	0.93	0.95	0.63	0.74	0.77	0.34	0.23	0.17	0.43	0.38	0.34

OS-SART but degrade as M decreases. Learning-based methods maintain reasonable quality at this sparsity level using priors trained from dense-view reconstructions. Our method achieves the best projection metrics while using one projection per training sample.

Tomographic Reconstruction As shown in Table 2, the performance hierarchy observed for novel view synthesis (Table 1) generally persists: learning-based methods outperform optimization-based approaches, and our method achieves comparable or superior results to learning-based methods that require densely scanned training data. Qualitative results are shown in Fig. 4 (4-view), with more inference-view counts and stress-test analysis in Appendix E. The gains are concentrated in the xSV-CT regime ($M \leq 8$) and narrow as M grows into the standard sparse-view setting (1–64 input views; see Appendix C). When more projections per sample are available at training time, performance improves smoothly (Appendix C); single-projection training is the most data-efficient end of this trade-off.

Inference Performance Scaling with Number of Input Views We evaluate how performance scales as the number of input views M grows from 1 to 64, which covers both the xSV-CT regime ($M \leq 8$) and the standard sparse-view setting. As shown in Fig. 5, our method performs best in novel view synthesis for $M \leq 8$ and remains competitive with learning-based methods as M increases, with the gap narrowing as more views become available. This contrasts with earlier findings that projection-domain diffusion priors underperform slice-domain priors because errors in independently generated projections accumulate during reconstruction [19]. Our ablation reproduces that failure when both guidance losses are removed (Table 3); coupling all generated views to a shared

Table 2: Quantitative comparison for tomographic reconstruction on the same 50-case stratified subset as Table 1.

	Metric	PSNR↑						SSIM↑						LPIPS↓					
	Dataset	CT-RATE [62]			SinoCT [63]			CT-RATE			SinoCT			CT-RATE			SinoCT		
	Inference Views	1	4	8	1	4	8	1	4	8	1	4	8	1	4	8	1	4	8
Optimization-based	FBP [52]	7.4	8.0	9.6	11.7	11.7	13.0	0.12	0.19	0.23	0.34	0.38	0.39	0.55	0.53	0.54	0.54	0.45	0.43
	OS-SART [25]	7.9	7.8	11.2	12.7	17.8	19.0	0.15	0.19	0.27	0.33	0.37	0.44	0.50	0.48	0.56	0.53	0.43	0.41
	SAX-NeRF [10]	6.3	11.4	20.2	9.4	15.0	17.9	0.15	0.37	0.54	0.29	0.40	0.46	0.66	0.52	0.40	0.49	0.42	0.36
	R^2-Gaussian [11]	9.8	15.9	24.8	9.3	12.0	19.1	0.21	0.40	0.56	0.23	0.28	0.42	0.55	0.45	0.29	0.56	0.47	0.41
	3DGR-CT [12]	11.7	19.6	26.2	9.1	15.4	20.3	0.21	0.38	0.53	0.33	0.34	0.53	0.51	0.41	0.31	0.48	0.45	0.40
Learning-based	DiffusionMBIR [37]	16.4	21.8	27.1	16.3	21.5	24.9	0.20	0.44	0.60	0.59	0.72	0.80	0.47	0.40	0.27	0.48	0.41	0.34
	C^2RV [13]	11.0	23.3	26.6	13.3	19.6	25.9	0.28	0.50	0.64	0.63	0.71	0.76	0.54	0.39	0.25	0.50	0.45	0.38
	DSDF [14]	9.3	22.1	27.0	10.4	17.8	23.9	0.23	0.43	0.64	0.40	0.71	0.77	0.43	0.36	0.30	0.48	0.36	0.35
	Ours	21.1	24.5	28.6	16.4	20.0	25.2	0.46	0.51	0.69	0.66	0.70	0.80	0.41	0.32	0.27	0.47	0.36	0.34

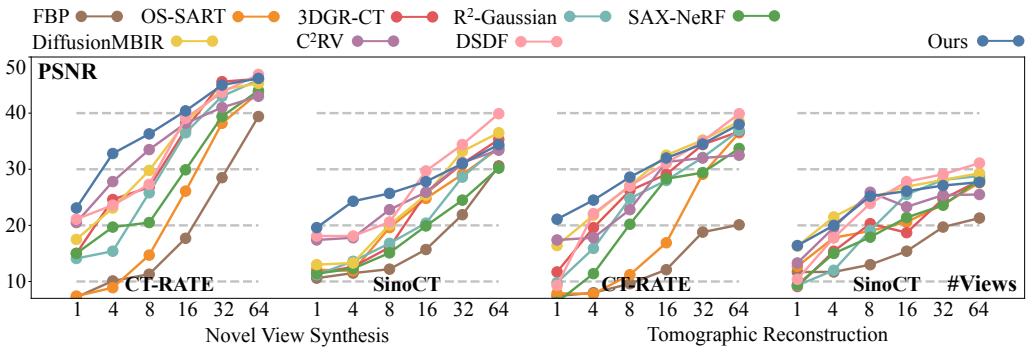

Figure 5: Performance scaling with the number of input views, metrics averaged over 30 cases.

volume and to the Radon range during sampling largely prevents it, making projection-domain priors a preferable choice when dense acquisition itself is the constraint. When dense volumes are available and M is not extremely small, we recommend a simpler slice- or volume-domain prior.

4.2 Ablation Studies

Table 3 ablates our key design choices. First, we evaluate our pose-aware projection-domain diffusion prior by comparing against (**1**) removing pose information entirely, and (**2**) using direct cross-attention on raw pose angles and the diffusion latents. Without pose information, the model generates uncalibrated projections even when guided by the consistency losses. Raw-angle cross-attention also underperforms because it lacks the periodic and reflection-related structure of projection angles. Second, we evaluate consistency losses, $\mathcal{L}_{\text{geo}}$ and $\mathcal{L}_{\text{phy}}$, during inference. With the pose-aware projection-domain prior fixed, either loss improves reconstruction, and using both losses gives the best results across both datasets. This supports the intended division of labor: the diffusion model proposes pose-conditioned projections, while the guidance losses align them with measured views and projection physics.

4.3 Robustness Analysis

We present the robustness of our model against key hyperparameters and faithfulness to sparse measurements in this section, with more analysis and failure cases provided in Appendices C and D.
Parameter Sensitivity Fig. 6 analyzes three key hyperparameters: number of novel projections N generated at inference, the order n of the moment curve used in Eq. (11), and the guidance scale γ_0 for inference losses. Generating more novel projections generally improves quality at the cost of longer inference, as a denser view set provides more stable solutions for Eq. (11) and improves the reconstruction quality of the FBP operator $\mathcal{A}^\dagger$. We select $N = 180$ as our default to balance reconstruction quality and computational cost. The moment order n works well across a wide range (20∼80), degrading only when n is too small to describe the sinogram moment structure or too large for stable optimization. The optimal γ_0 differs between datasets, likely because SinoCT uses

Table 3: Ablation study on key components evaluated with 8 input views on 50 test cases per dataset.

Task	Novel View Synthesis						Tomographic Reconstruction					
Dataset	CT-RATE			SinoCT			CT-RATE			SinoCT		
Metric	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
No pose	11.9	0.21	0.56	12.5	0.23	0.50	9.4	0.17	0.61	10.7	0.25	0.53
Naive CrossAttn	24.5	0.40	0.48	20.2	0.45	0.47	11.5	0.33	0.41	15.6	0.39	0.50
w/o $\mathcal{L}_{\text{phy}}, \mathcal{L}_{\text{geo}}$	15.6	0.32	0.55	14.6	0.32	0.49	9.2	0.19	0.59	10.0	0.18	0.53
w/o $\mathcal{L}_{\text{geo}}$	28.6	0.78	0.30	21.4	0.50	0.40	16.4	0.35	0.44	18.9	0.51	0.47
w/o $\mathcal{L}_{\text{phy}}$	30.1	0.88	0.26	22.1	0.56	0.39	22.0	0.44	0.39	19.1	0.56	0.47
Ours (full)	36.3	0.95	0.17	25.7	0.77	0.34	28.6	0.69	0.27	25.2	0.80	0.34

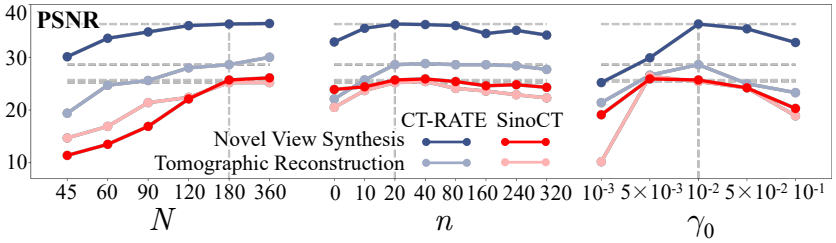

Figure 6: Parameter sensitivity analysis with 8 input views. Blue and red curves show PSNR on CT-RATE and SinoCT datasets respectively. Gray dashed lines indicate our default settings.

a higher axial resolution, which introduces more fine-grained variations across projections than CT-RATE, thus requiring a smaller guidance scale for refining the generated projections.

Measurement Anchoring Analysis Under 1- and 4-view inputs, generated angles whose nearest measured view is more than $\sim 30^\circ$ away draw more strongly on the learned prior. *(i) Cross-seed repeatability.* Running our method with multiple random seeds yields similar projections across runs: mean cross-seed ℓ_2 residuals are 5.48×10^{-3}, 1.31×10^{-3}, and 0.96×10^{-3} at 1-, 4-, and 8-view inputs, with standard deviations around 10^{-4} (Table 4). *(ii) Angular-distance dependence.* Residuals on generated projections grow with angular distance from the nearest input view (see Appendix C), matching the expected geometry: views closer to measurements are more directly constrained.

Views	*mean*$\times10^{-3}$	*std.*$\times10^{-4}$
1-view	5.48	4.64
4-view	1.31	1.74
8-view	0.96	0.91

Table 4: Cross-seed ℓ_2 residual on generated views.

5 Conclusion

We introduced a train-sparse / infer-sparse framework for xSV-CT, where acquisition limits constrain both reconstruction and training-data curation. We show that a projection-domain diffusion prior learned from single-view data can become useful when it is embedded inside a physics-guided sampling loop. Through this projection-domain prior, shared-volume reprojection, and finite-order Range Condition guidance, the method synthesizes tomographically coherent projections for standard reconstruction without dense-volume supervision in the target regime. Controlled experiments on two public CT datasets support this claim across 1-, 4-, and 8-view settings. We analyze the broader impacts in Appendix F. **Limitations:** We use public dense CT data to provide paired references for controlled evaluation; prospective sparse-only clinical or specimen studies are left for future work. Generated angles far from measured views draw more strongly on the learned prior, and our finite-order Range-Condition surrogate currently covers parallel-beam or rebinned fan-beam data. These points motivate calibrated trust maps and native non-parallel constraints such as Grangeat's Condition [69].

Acknowledgements

The research is made possible under the JST-Mirai program from Japanese Science and Technology Agency with grant number JPMJMI24G1.

References

[1] Jianglei Di, Juncheng Lin, Liyun Zhong, Kemao Qian, and Yuwen Qin. Review of sparse-view or limited-angle CT reconstruction based on deep learning. *Laser & Optoelectronics Progress*, 60(8):0811002, 2023. 1

[2] Yiqun Lin, Zhongjin Luo, Wei Zhao, and Xiaomeng Li. Learning deep intensity field for extremely sparse-view CBCT reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 13–23. Springer, 2023. 1, 3

[3] Yiqun Lin, Hualiang Wang, Jixiang Chen, and Xiaomeng Li. Learning 3D Gaussians for extremely sparse-view cone-beam CT reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 425–435. Springer, 2024. 1, 3

[4] Guofeng Zhang, Ruyi Zha, Hao He, Yixun Liang, Alan Yuille, Hongdong Li, and Yuanhao Cai. X-LRM: X-ray large reconstruction model for extremely sparse-view computed tomography recovery in one second. In *International Conference on 3D Vision (3DV)*, pages 176–185, 2026. 1, 3

[5] Nao Otomo, Haruki Funao, Kento Yamanouchi, Norihiro Isogai, and Ken Ishii. Computed tomography-based navigation system in current spine surgery: a narrative review. *Medicina*, 58(2):241, 2022. 1

[6] Harald Schöndube, Thomas Allmendinger, Karl Stierstorfer, Herbert Bruder, and Thomas Flohr. Temporal resolution and motion artifacts in single-source and dual-source cardiac CT. *Medical Physics*, 40(3):031112, 2013. 1

[7] Elspeth F Garman and Martin Weik. Radiation damage to biological samples: still a pertinent issue. *Journal of Synchrotron Radiation*, 28(5):1278–1283, 2021. 1

[8] Yuanhao Cai, Yixun Liang, Jiahao Wang, Angtian Wang, Yulun Zhang, Xiaokang Yang, Zongwei Zhou, and Alan Yuille. Radiative Gaussian splatting for efficient X-ray novel view synthesis. In *European Conference on Computer Vision*, pages 283–299. Springer, 2024. 1, 2, 3, 6

[9] 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*, pages 442–452. Springer, 2022. 3

[10] Yuanhao Cai, Jiahao Wang, Alan Yuille, Zongwei Zhou, and Angtian Wang. Structure-aware sparse-view X-ray 3D reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11174–11183, 2024. 3, 7, 8, 9

[11] Ruyi Zha, Tao Jun Lin, Yuanhao Cai, Jiwen Cao, Yanhao Zhang, and Hongdong Li. R^2-Gaussian: Rectifying radiative Gaussian splatting for tomographic reconstruction. *Advances in Neural Information Processing Systems*, 37:44907–44934, 2024. 3, 7, 8, 9

[12] Yingtai Li, Xueming Fu, Han Li, Shang Zhao, Ruiyang Jin, and S Kevin Zhou. 3DGR-CT: Sparse-view CT reconstruction with a 3D Gaussian representation. *Medical Image Analysis*, page 103585, 2025. 1, 3, 7, 8, 9, 21, 22

[13] Yiqun Lin, Jiewen Yang, Hualiang Wang, Xinpeng Ding, Wei Zhao, and Xiaomeng Li. C^2RV: Cross-regional and cross-view learning for sparse-view CBCT reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11205–11214, 2024. 1, 3, 7, 8, 9, 17

[14] Yongsheng Pan, Yiwen Ye, Yanning Zhang, Yong Xia, and Dinggang Shen. Draw sketch, draw flesh: Whole-body computed tomography from any x-ray views. *International Journal of Computer Vision*, 133(5):2505–2526, 2025. 3, 7, 8, 9, 17, 18, 21, 22

[15] Xing Xie, Jiawei Liu, Huijie Fan, Zhi Han, Yandong Tang, and Liangqiong Qu. DVG-Diffusion: Dual-view-guided diffusion model for CT reconstruction from X-rays. *IEEE Transactions on Image Processing*, 35:1158–1173, 2026. 3

[16] Xuhui Liu, Hong Li, Zhi Qiao, Yawen Huang, Xi Liu, Juan Zhang, Zhen Qian, Xiantong Zhen, and Baochang Zhang. D3T: Dual-domain diffusion transformer in triplanar latent space for 3D incomplete-view CT reconstruction. *International Journal of Computer Vision*, pages 1–24, 2025. 1, 3

[17] Linrui Dai, Rongzhao Zhang, Yongrui Yu, and Xiaofan Zhang. GuideGen: A text-guided framework for paired full-torso anatomy and CT volume generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 3470–3478, 2026. 2

[18] Bruno De Man, Samit Basu, Naveen Chandra, Bruce Dunham, Peter Edic, Maria Iatrou, Scott McOlash, Paavana Sainath, Charlie Shaughnessy, Brendon Tower, and Eugene Williams. CatSim: a new computer assisted tomography simulation environment. In Jiang Hsieh and Michael J. Flynn, editors, *Medical Imaging 2007: Physics of Medical Imaging*, volume 6510, page 65102G. International Society for Optics and Photonics, SPIE, 2007. 2, 7, 16

[19] Shiyan Su, Ruyi Zha, Danli Shi, Hongdong Li, and Xuelian Cheng. DiffNR: Diffusion-enhanced neural representation optimization for sparse-view 3D tomographic reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 9144–9152, 2026. 2, 8

[20] Mark A Anastasio, Xiaochuan Pan, and Eric Clarkson. Comments on the filtered backprojection algorithm, range conditions, and the pseudoinverse solution. *IEEE transactions on medical imaging*, 20(6):539–542, 2001. 3, 6

[21] Hiroyuki Kudo, Taizo Suzuki, and Essam A Rashed. Image reconstruction for sparse-view CT and interior CT—introduction to compressed sensing and differentiated backprojection. *Quantitative imaging in medicine and surgery*, 3(3):147, 2013.

[22] Avinash C Kak and Malcolm Slaney. *Principles of computerized tomographic imaging*. SIAM, 2001. 3

[23] Lee A Feldkamp, Lloyd C Davis, and James W Kress. Practical cone-beam algorithm. *Journal of the Optical Society of America A*, 1(6):612–619, 1984. 3, 5, 16

[24] John A Scales. Tomographic inversion via the conjugate gradient method. *Geophysics*, 52(2):179–185, 1987.

[25] Ge Wang and Ming Jiang. Ordered-subset simultaneous algebraic reconstruction techniques (OS-SART). *Journal of X-ray Science and Technology*, 12(3):169–177, 2004. 7, 8, 9

[26] Emil Y Sidky and Xiaochuan Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine & Biology*, 53(17):4777, 2008. 3

[27] Rushil Anirudh, Hyojin Kim, Jayaraman J Thiagarajan, K Aditya Mohan, Kyle Champley, and Timo Bremer. Lose the views: Limited angle CT reconstruction via implicit sinogram completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6343–6352, 2018. 3

[28] Guangming Zang, Ramzi Idoughi, Rui Li, Peter Wonka, and Wolfgang Heidrich. Intratomo: self-supervised learning-based tomography via sinogram synthesis and prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1960–1970, 2021. 3, 6

[29] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948. IEEE, 2013. 3

[30] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3929–3938, 2017. 3

[31] Feiran Wang, Jiachen Tao, Junyi Wu, Haoxuan Wang, Bin Duan, Kai Wang, Zongxin Yang, and Yan Yan. X-Field: A physically informed representation for 3D X-ray reconstruction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3

[32] Shaokai Wu, Yuxiang Lu, Yapan Guo, Wei Ji, Suizhi Huang, Fengyu Yang, Shalayiding Sirejiding, Qichen He, Jing Tong, Yanbiao Ji, et al. Discretized Gaussian representation for tomographic reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 25073–25082, 2025. 3

[33] Jiacheng Xie, Hua-Chieh Shao, Yunxiang Li, and You Zhang. Prior frequency guided diffusion model for limited angle (LA)-CBCT reconstruction. *Physics in Medicine & Biology*, 69(13):135008, 2024. 3

[34] Suhyeon Lee, Hyungjin Chung, Minyoung Park, Jonghyuk Park, Wi-Sun Ryu, and Jong Chul Ye. Improving 3D imaging with pre-trained perpendicular 2D diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10710–10720, 2023. 3

[35] Huan Li, Yibo Yang, Dongmin Chen, and Zhouchen Lin. Optimization algorithm inspired deep neural network structure design. In *Asian Conference on Machine Learning*, pages 614–629. PMLR, 2018. 3

[36] Haimiao Zhang, Bin Dong, and Baodong Liu. JSR-Net: A deep network for joint spatial-Radon domain CT reconstruction from incomplete data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3657–3661. IEEE, 2019. 3

[37] Hyungjin Chung, Dohoon Ryu, Michael T McCann, Marc L Klasky, and Jong Chul Ye. Solving 3D inverse problems using pre-trained 2D diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22542–22551, 2023. 3, 7, 8, 9, 17, 18

[38] Linchao He, Wenchao Du, Peixi Liao, Fenglei Fan, Hu Chen, Hongyu Yang, and Yi Zhang. Solving zero-shot sparse-view CT reconstruction with variational score solver. *IEEE Transactions on Medical Imaging*, 44(9):3586–3599, 2025. 3

[39] Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. In *International conference on learning representations*, volume 2024, pages 38922–38949, 2024. 3, 7, 17

[40] Bowen Song, Jason Hu, Zhaoxu Luo, Jeffrey A Fessler, and Liyue Shen. DiffusionBlend: Learning 3D image prior through position-aware diffusion score blending for 3D computed tomography reconstruction. *Advances in Neural Information Processing Systems*, 37:89584–89611, 2024. 7, 17

[41] Chenhe Du, Qing Wu, Xuanyu Tian, Jingyi Yu, Hongjiang Wei, and Yuyao Zhang. Improving 2D diffusion models for 3D medical imaging with inter-slice consistent stochasticity. *arXiv preprint arXiv:2602.04162*, 2026. 3

[42] Liyue Shen, Wei Zhao, and Lei Xing. Patient-specific reconstruction of volumetric computed tomography images from a single projection view via deep learning. *Nature biomedical engineering*, 3(11):880–888, 2019. 3

[43] Xingde Ying, Heng Guo, Kai Ma, Jian Wu, Zhengxin 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)*, pages 10611–10620. IEEE, 2019. 3

[44] Liang Zhao, Xiaoyuan Li, Sijia Hou, Xin Fan, Zhikui Chen, and Shuqiong Wu. Construction of simulated 3D CT from multiple-view X-ray images. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2951–2956. IEEE, 2024. 3

[45] Chenlu Zhan, Yu Lin, Gaoang Wang, Hongwei Wang, and Jian Wu. Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11502–11512. IEEE, 2024. 3

[46] Shuyang Lai, Xuanyu Tian, Qing Wu, Chenhe Du, Xiaojun Xu, Hongjiang Wei, Xiaojun Guan, and Yuyao Zhang. Reconstructing knee CT volumes from biplanar X-rays via self-supervised neural field. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. 3

[47] Lifeng Yu, Maria Shiung, Dayna Jondal, and Cynthia H McCollough. Development and validation of a practical lower-dose-simulation tool for optimizing computed tomography scan protocols. *Journal of computer assisted tomography*, 36(4):477–487, 2012. 3, 20

[48] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4, 5

[49] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 5

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 5

[51] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 5

[52] Lawrence A Shepp and Benjamin F Logan. The Fourier reconstruction of a head section. *IEEE Transactions on nuclear science*, 21(3):21–43, 1974. 5, 7, 8, 9, 16, 17

[53] Donald Ludwig. The Radon transform on euclidean space. *Communications on pure and applied mathematics*, 19(1):49–81, 1966. 6

[54] Philip Dreike and Douglas P Boyd. Convolution reconstruction of fan beam projections. *Computer graphics and image processing*, 5(4):459–469, 1976. 6

[55] Marc Kachelrieß, Stefan Schaller, and Willi A Kalender. Advanced single-slice rebinning in cone-beam spiral CT. *Medical Physics*, 27(4):754–772, 2000. 6

[56] Yixing Huang, Xiaolin Huang, Oliver Taubmann, Yan Xia, Viktor Haase, Joachim Hornegger, Guenter Lauritsch, and Andreas Maier. Restoration of missing data in limited angle tomography based on Helgason–Ludwig consistency conditions. *Biomedical Physics & Engineering Express*, 3(3):035015, 2017. 6

[57] Yingqing Guo, Hui Yuan, Yukang Yang, Minshuo Chen, and Mengdi Wang. Gradient guidance for diffusion models: An optimization perspective. *Advances in Neural Information Processing Systems*, 37:90736–90770, 2024. 6

[58] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[59] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. 7

[60] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022. 7

[61] Riccardo Barbano, Alexander Denker, Hyungjin Chung, Tae Hoon Roh, Simon Arridge, Peter Maass, Bangti Jin, and Jong Chul Ye. Steerable conditional diffusion for out-of-distribution adaptation in medical image reconstruction. *IEEE Transactions on Medical Imaging*, 2025. 7

[62] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*, 10(8):1610–1628, 2026. 7, 8, 9, 16, 17

[63] Sarah M Hooper, Jared A Dunnmon, Matthew P Lungren, Domenico Mastrodicasa, Daniel L Rubin, Christopher Ré, Adam Wang, and Bhavik N Patel. Impact of upstream medical image processing on downstream performance of a head CT triage neural network. *Radiology: Artificial Intelligence*, 3(4):e200229, 2021. 7, 8, 9, 16, 17

[64] Hyojin Kim and Kyle Champley. Differentiable forward projector for X-ray computed tomography. In *ICML 2023 Workshop on Differentiable Almost Everything: Differentiable Relaxations, Algorithms, Operators, and Simulators*, 2023. 7

[65] 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. 7

[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[67] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 7

[68] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7, 17

[69] Pierre Grangeat. Mathematical framework of cone beam 3D reconstruction via the first derivative of the Radon transform. In *Mathematical Methods in Tomography*, Lecture Notes in Mathematics, pages 66–97. Springer, 1991. 10, 23

[70] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10674–10685, 2022. 17

[71] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 17

[72] Yang Gao, Zhaoying Bian, Jing Huang, Yunwan Zhang, Shanzhou Niu, Qianjin Feng, Wufan Chen, Zhengrong Liang, and Jianhua Ma. Low-dose X-ray computed tomography image reconstruction with a combined low-mas and sparse-view protocol. *Optics express*, 22(12):15190–15210, 2014. 19

[73] Darius Rückert, Yuanhao Wang, Rui Li, Ramzi Idoughi, and Wolfgang Heidrich. NeAT: Neural adaptive tomography. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 20

[74] Stanislav Žabić, Qiu Wang, Thomas Morton, and Kevin M Brown. A low dose simulation tool for CT systems with energy integrating detectors. *Medical physics*, 40(3):031102, 2013. 20

[75] Ander Biguri, Manjit Dosanjh, Steven Hancock, and Manuchehr Soleimani. TIGRE: a MATLAB-GPU toolbox for CBCT image reconstruction. *Biomedical Physics & Engineering Express*, 2(5):055010, 2016. 20

[76] Yanbo Zhang and Hengyong Yu. Convolutional neural network based metal artifact reduction in X-ray computed tomography. *IEEE transactions on medical imaging*, 37(6):1370–1381, 2018. 22

[77] Heang K. Tuy. An inversion formula for cone-beam reconstruction. *SIAM Journal on Applied Mathematics*, 43(3):546–552, 1983. 23

This supplementary material provides: an algorithmic overview of our guided generation procedure, additional experimental details, extended analysis, failure analysis, potential societal impacts, and additional qualitative results. We also provide an accompanying **video** that visualizes 3D reconstructions across compared methods and ours for all cases presented in this appendix. The code for training and evaluation as well as model checkpoints will be available at https://github.com/OvO1111/xSV-CT.

A Algorithmic Overview

We provide an overview of our *Consistency-aware Guided Generation* process in Algorithm 1. Intuition and derivation of each component are described in Sec. 3.2 of the main text.

Algorithm 1 Consistency-Aware Guided Generation

Require: Sparse projections $\mathcal{P}$ at poses $\boldsymbol{\theta}$; novel poses $\hat{\boldsymbol{\theta}}$; trained diffusion model p_ϕ; Radon operator $\mathcal{A}$ and approximate inverse (via FBP [52] or FDK [23]) $\mathcal{A}^\dagger$; loss weights λ_1, λ_2; Adam parameters $(\gamma_0, \beta_1, \beta_2)$; Diffusion schedules $(\bar{\alpha}_t, \alpha_t, \beta_t)$

Ensure: Dense projections $\hat{\mathcal{P}}_0$ and reconstructed volume $\hat{\mathbf{y}}_0$.

Initialize noisy projections $\hat{\mathcal{P}}_T = \{\hat{\mathbf{x}}_T^i \sim \mathcal{N}(0, I)\}_{i=1}^N$
for $t = T, T-1, \ldots, 1$ **do**
  for each pose $\hat{\theta}_i \in \hat{\boldsymbol{\theta}}$ **do**
    Predict clean projection $\hat{\mathbf{x}}_{0|t}^i = (\hat{\mathbf{x}}_t^i - \sqrt{1-\bar{\alpha}_t}\, p_\phi(\hat{\mathbf{x}}_t^i; \hat{\theta}_i, t))/\sqrt{\bar{\alpha}_t}$
  end for
  Collect $\hat{\mathcal{P}}_t = \{\hat{\mathbf{x}}_{0|t}^i\}_{i=1}^N$, $\tilde{\mathcal{P}}_t = \mathcal{P} \cup \hat{\mathcal{P}}_t$, $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} \cup \hat{\boldsymbol{\theta}}$
  Reconstruct intermediate volume $\hat{\mathbf{y}}_{0|t} = \mathcal{A}_{\tilde{\boldsymbol{\theta}}}^\dagger(\tilde{\mathcal{P}}_t)$
  Compute $\mathcal{L}_{\text{geo}} = \|\mathcal{P} - \mathcal{A}_{\boldsymbol{\theta}}\hat{\mathbf{y}}_{0|t}\|_2^2 + \|\hat{\mathcal{P}}_t - \mathcal{A}_{\hat{\boldsymbol{\theta}}}\hat{\mathbf{y}}_{0|t}\|_2^2 + \mathcal{L}_{TV}(\hat{\mathbf{y}}_{0|t})$
  Construct design matrix $\mathbf{A}_0$ at height $z = Z$, rebin $\hat{\mathcal{P}}_t$ for fan-beam geometries and solve Range Condition coefficients $\boldsymbol{\alpha}_0(Z) = (\mathbf{A}_0^\top \mathbf{A}_0)^{-1}\mathbf{A}_0^\top \int_{-1}^1 r^n \hat{\mathcal{P}}_t(r, \theta, Z)dr$
  Compute $\mathcal{L}_{\text{phy}} = || \int_{-1}^1 r^n \hat{\mathcal{P}}_t(r, \theta, Z)dr - \mathbf{A}_0\boldsymbol{\alpha}_0(Z)||_2^2$
  for each projection $\hat{\mathbf{x}}_t^i$ **do**
    Guided update $\hat{\mathbf{x}}_t^i \leftarrow \texttt{Adam}\Big(\hat{\mathbf{x}}_t^i,\ \nabla_{\hat{\mathbf{x}}_t^i}\left(\lambda_1\mathcal{L}_{\text{geo}} + \lambda_2\mathcal{L}_{\text{phy}}\right); \gamma_0, \beta_1, \beta_2\Big)$
    Reverse diffusion step $\hat{\mathbf{x}}_{t-1}^i = \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathbf{x}}_t^i - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} p_\phi(\hat{\mathbf{x}}_t^i; \hat{\theta}_i, t)\right) + \sqrt{\beta_t}\boldsymbol{\epsilon}$
  end for
end for
Generated consistent projections $\hat{\mathcal{P}}_0 = \{\hat{\mathbf{x}}_0^i\}_{i=1}^N$
Final reconstruction $\hat{\mathbf{y}}_0 = \mathcal{A}_{\hat{\boldsymbol{\theta}}}^\dagger(\mathcal{P} \cup \hat{\mathcal{P}}_0)$

B Additional Experimental Details

Dataset Configuration. We list the projection parameters and dataset statistics in Table 5. To construct the training and evaluation sets, we generated 3D volume projections from the CT-RATE [62] dataset using a parallel-beam geometry. For the SinoCT [63] dataset, we used the existing fan-beam projections simulated with CatSim [18] (source-isocenter distance of 541 mm, a source-detector distance of 949 mm) provided by the original paper and resized them to a consistent shape. We selected 6,000 cases from CT-RATE and 4,000 from SinoCT based on the number of slices and image intensity range, discarding CTs from the same patient to prevent data leakage (in CT-RATE). All projection angles are defined in an object-centered frame: $\theta = 0^\circ$ corresponds to the frontal (anteroposterior, AP) view, with the X-ray source in front of the object, and an increasing θ implies a counterclockwise rotation when viewed from above. These cases are split 7:1 for training and validation. Evaluations are done on a separate left-out test set of 50 cases. Optimization-based methods are directly optimized on the input projections at inference for testing. In comparison, our method operates in the projection domain and requires only a single projection per training sample,

Table 5: Projection geometry and dataset splits (training, validation and testing) for CT-RATE and SinoCT. Our method uses only a single 2D projection per training sample, while learning-based baselines require the full 3D volumes reconstructed from dense angular sampling.

Dataset	Projection Geometry				No. Training	No. Validation
	Detector Shape	Volume Shape	Detector Pixel Size	Voxel Size	Cases	Cases
CT-RATE [62]	256×256	$256 \times 256 \times 256$	2×2	$2 \times 2 \times 2$	4200	600
SinoCT [63]	64×512	$64 \times 512 \times 512$	4×1	$4 \times 1 \times 1$	2800	400

while learning-based methods require original ground-truth 3D volumes (CT-RATE) or volumes reconstructed from all available projections via FBP algorithm [52] (SinoCT). This fundamental difference in training data requirements motivates our radiation reduction argument at the method level.

Implementation Details. We implemented our diffusion model for smooth reproduction based on Stable Diffusion [70], which uses a U-Net [71] backbone interleaved with cross-attention layers. We set the model channels of our U-Net to be [64, 128, 256, 512, 1024] and apply pose injection at all encoder and decoder layers except the first and second U-Net layer for efficient computation, with a pose context dimension of 64. The final model has 140M parameters.

Training protocol. We use AdamW [68] for diffusion model training with a constant learning rate of 4×10^{-5}, momentum 0.99, weight decay 10^{-5}, and batch size 8 per GPU on 4 NVIDIA RTX 6000 Ada GPUs. This diffusion model was trained for 300 epochs (which takes approximately 5 days) on our GPU cluster.

Inference hyperparameters. For inference, we use moment order n=20 for the Range Condition (Eq. (11) in the main text), loss weights λ_1=10^{-2} and λ_2=1, and generate N=180 novel projections. Our method does *not* require retraining when the number of input views M changes at inference, unlike several learning-based methods (*e.g.*, DSDF). The same trained model handles 1-, 4-, and 8-view inputs (as well as other view counts in Fig. 7 and Fig. 5 hereafter), with the consistency losses adapting automatically to the available measurements.

Reconstruction cost. Our model is trained only once per dataset and can be applied at inference with different numbers of input views without retraining. The inference time is dominated by the number of synthesized novel projections N, following an approximately linear trend of $t \approx 10 + 0.2N$ minutes. We synthesize 48 projections at a time on one GPU; four GPUs bring the runtime to about 35% at unchanged quality. For reference, DDS [39] and DiffusionBlend++ [40] report 48 and 32 minutes for 8-view reconstruction [40], which is comparable with 47 minutes for our method at $N = 180$ and 8 views. On CT-RATE with 256×256 projections, the peak VRAM usage is 4.1 GB during training and 0.9 GB during inference per batch.

Table 6: Training paradigms of learning-based baselines and ours.

Method	Supervision	Retraining
DiffusionMBIR [37]	Reconstructed axial slices	Per dataset (view-count agnostic)
C^2RV [13]	Reconstructed 3D volumes	Per dataset and per view count
DSDF [14]	Reconstructed 3D volumes	Per dataset and per view count (both stages)
Ours	X-ray projections	Per dataset (view-count agnostic)

Comparison methods. Our method is trained on a single projection at a randomly sampled pose per sample and accepts inputs at arbitrary poses at inference. To evaluate all methods under this setting, we use the official implementation of each comparison method and adapt it to our training setup where needed. Optimization-based methods need no adaptation, since they are optimized directly on the input projections of each test case. DiffusionMBIR needs none either: its prior is trained on axial slices, and the acquisition geometry enters only through the forward operator at inference, so we follow its original training setup. We retrain C^2RV and DSDF separately for each view count M, drawing the M input projections at randomly sampled poses in every training iteration. All learning-based methods use the same splits and preprocessing as our method, with geometry and resolution matched to ours, and are trained until the loss stops decreasing for 10 epochs. Table 6 summarizes their supervision and retraining scope.

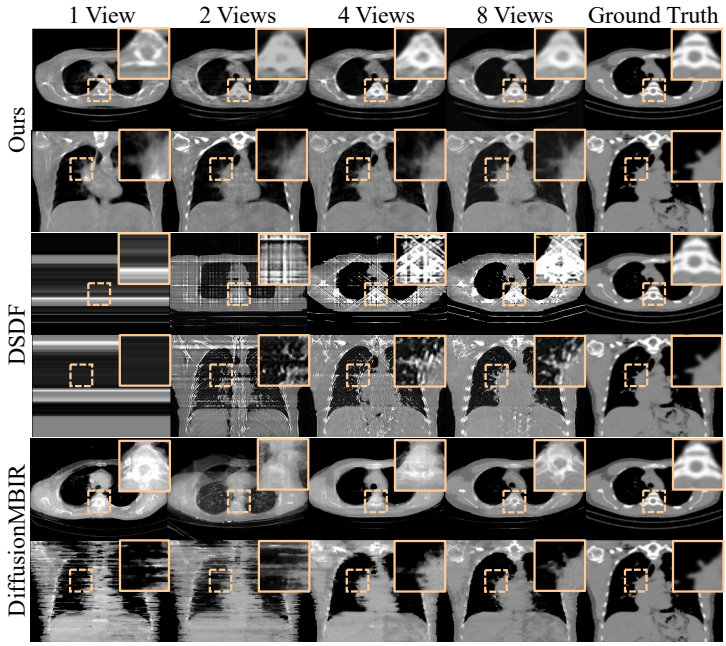

Figure 7: Tomographic reconstruction across different xSV-CT settings comparing our method against two best-performing methods [14, 37]. One axial and one coronal slice is shown for each case. Our method produces more plausible reconstructions with fewer artifacts across all sparsity levels.

C Extended Analysis

Table 7: Effect of training-time data availability: PSNR for novel view synthesis / tomographic reconstruction on CT-RATE dataset with more views per sample seen at training. With a single projection per training sample, our method achieves competitive results. Additional training views provide gains, particularly for single-view inference, while multi-view inference is less sensitive to the number of views per training sample.

#Train Views/Sample	#Test Views		
	1	4	8
1 (Ours)	23.1 / 21.1	32.8 / 24.5	36.3 / 28.6
4	23.0 / 21.1	33.2 / 24.8	36.9 / 28.6
32	26.0 / 21.3	33.6 / 24.8	38.1 / 29.3

Generation Faithfulness and Consistency. A critical question for any generative approach in medical imaging is the extent to which generated content is constrained by measurements versus injected by the learned prior. Under extreme sparsity (1–4 views), the problem is severely ill-posed and much of the reconstruction is necessarily prior-driven. In the main text, we evaluate the measurement-dependence of our generation quantitatively in Table 3, where removing $\mathcal{L}_{\text{geo}}$ causes a large performance drop. The measurement fidelity term in $\mathcal{L}_{\text{geo}}$ (Eq. (9)) constrains generated views to be compatible with the actual X-ray measurements, not just anatomically plausible. We emphasize that measurement fidelity at a few input angles does not guarantee anatomical correctness at all generated angles, since structures far from the measured views remain substantially prior-driven. To evaluate the consistency, we run our method 4 times with different random seeds under the xSV-CT settings ($M \in \{1, 4, 8\}$) and report the ℓ_2 residual of (i) generated novel projections at different angles (Fig. 8, top), and (ii) reconstructed CT slices (Fig. 8, bottom). The model behaves consistently across seeds: generated projections at angles closer to the sparse input views retain more structural detail than those with larger angular separation, and this effect diminishes as the number of available projections M at inference increases. The low cross-seed variance suggests that the guided generation produces repeatable outputs rather than arbitrary samples from the prior. This consistent angular pattern also makes it possible to estimate a per-view confidence map without ground truth,

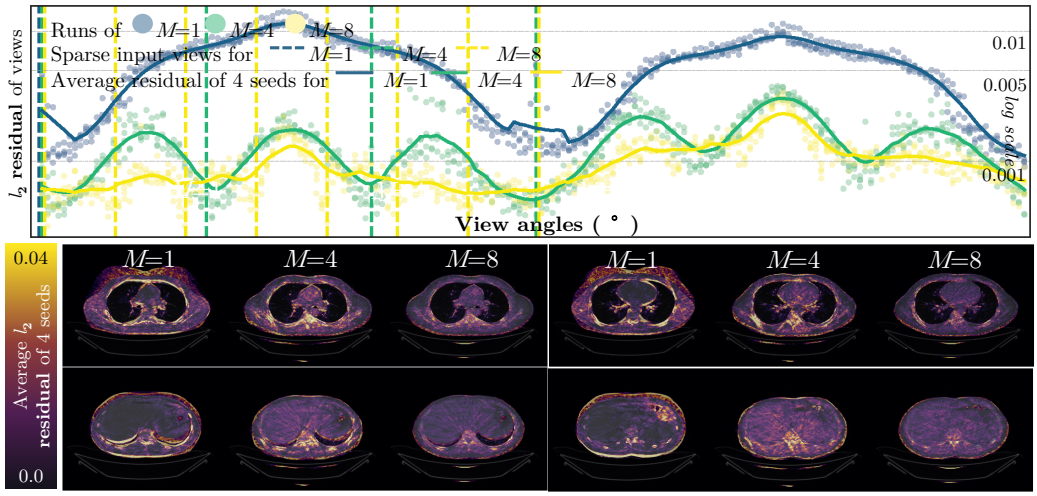

Figure 8: Consistency analysis via ℓ_2 residuals under the xSV-CT settings ($M \in \{1, 4, 8\}$) across 4 random seeds. **Top:** residuals of novel view synthesis at different angles, showing reduced error near input angles and consistent behavior across seeds. **Bottom:** average residuals of tomographic reconstruction overlayed on ground-truth axial slices, showing that reconstruction error decreases as more views become available. The low cross-seed variance suggests that the guided generation produces similar outputs across different initializations.

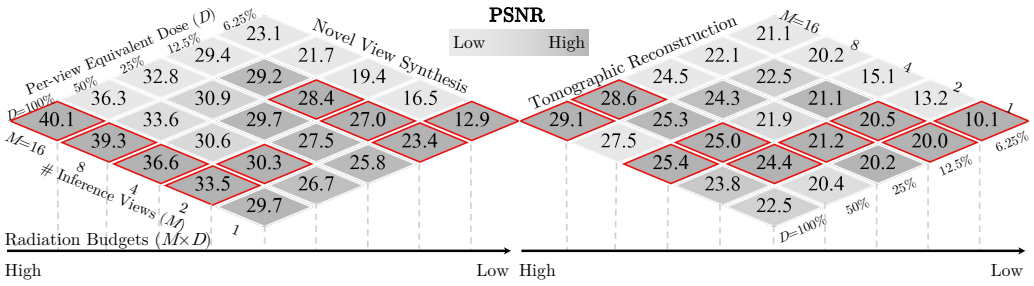

Figure 9: Performance under fixed radiation budgets: each vertically aligned group of cells corresponds to a constant radiation budget (inference views M × per-view dose D). Red outlines indicate the best-performing combination of M and D under each budget. Numbers denote PSNR values for either Novel View Synthesis (left) or Tomographic Reconstruction (right) on CT-RATE. The optimal operating point consistently involves moderate sparsity paired with moderate per-view dose, suggesting that our framework could provide flexibility in balancing view count and per-view exposure under radiation constraints.

from the variance of generated projections across random seeds. Even a simple estimate based on the 4 seeds above correlates with the per-view ℓ_2 error (Pearson $r = 0.60$ for $M = 4$), suggesting that cross-seed variance can flag generated views that rely more heavily on the learned prior to a certain extent.

Radiation Budget Trade-off. In clinical CT, the total radiation budget $B = M \times D$ (number of views × per-view dose) is the primary safety constraint. Given a fixed budget, should one acquire few high-dose views or many low-dose views? We evaluate this trade-off under simulated fixed radiation budgets in Fig. 9 by sweeping over inference views M and per-view dose D under fixed radiation budgets. The results show that the model performs best under a moderate combination of view sparsity and measurement noise for both novel view synthesis and tomographic reconstruction tasks. This observation highlights the importance of developing xSV-CT frameworks that provide flexibility in balancing projection sparsity and noise levels when operating under a fixed radiation budget in clinical settings [72].

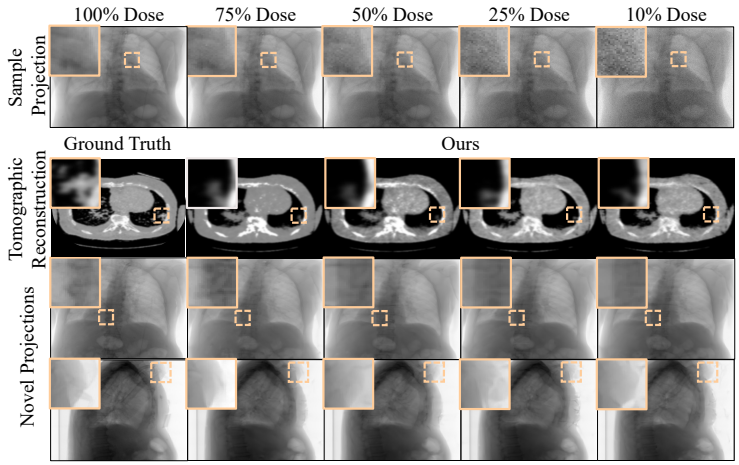

Figure 10: Qualitative evaluation under noisy input projections. Each row shows reconstruction from 8 noisy projections (one example shown in the first row) under simulated 75%, 50%, 25% and 10% of original X-ray dosage to guide the inference process.

Table 8: Robustness to noisy input projections (left) and angular uncertainty (right). Metrics (**P**SNR/**S**SIM/**L**PIPS) are averaged on 50 CT-RATE cases with 8 input views.

Equiv. Dosage	Novel View Syn.			Reconstruction		
	P↑	**S**↑	**L**↓	**P**↑	**S**↑	**L**↓
100%	36.3	0.95	0.17	28.6	0.69	0.27
75%	34.3	0.90	0.20	26.1	0.66	0.28
50%	33.6	0.89	0.23	25.3	0.66	0.30
25%	30.6	0.85	0.24	25.0	0.65	0.31
10%	31.2	0.85	0.24	24.3	0.61	0.30

$\Delta\theta$	Novel View Syn.			Reconstruction		
	P↑	**S**↑	**L**↓	**P**↑	**S**↑	**L**↓
$\pm 0^\circ$	36.3	0.95	0.17	28.6	0.69	0.27
$\pm 2.5^\circ$	34.4	0.93	0.21	27.3	0.68	0.33
$\pm 5^\circ$	34.5	0.94	0.20	24.9	0.66	0.35
$\pm 7.5^\circ$	32.8	0.88	0.27	21.6	0.57	0.39
$\pm 10^\circ$	29.1	0.82	0.35	20.6	0.54	0.39

Training-time Scalability. We further evaluate the scalability of our model when trained with increasing numbers of projections per sample. As shown in Tab. 7, performance improves gradually as more views are available during training, indicating that our model can effectively use additional training projections and scale with increased data availability.

Model Robustness. Inspired by the model self-calibration analysis in NeAT [73], we evaluate the robustness of our method under two practical perturbation settings: (**1**) Robustness to noisy projections: In real-world acquisition, even with nominally accurate geometric calibration, projection quality can be degraded by changes in the imaging environment, reduced photon counts, and electronic readout noise. This issue is amplified in xSV-CT, where the limited number of projections provides fewer measurements over which noise can be averaged out. Following the low-dose X-ray noise simulation protocols in [47, 74], we use TIGRE [75] to inject mixed Poisson and Gaussian noise into the input views, modeling photon-counting noise and detector/electronic noise under reduced-dose acquisition. As shown in Table 8 (left) and Fig. 10, our method degrades gracefully under noisy inputs: the global anatomical structure remains stable, while most degradation is confined to fine-scale details. This behavior suggests that the diffusion prior provides a useful structural constraint when individual sparse projections become unreliable. (**2**) Robustness to uncertainty of camera poses: We further evaluate robustness to geometric miscalibration, where the assumed acquisition angles deviate from the true scanning geometry. Specifically, we perturb each input angle $\theta \in \boldsymbol{\theta}$ by a random offset sampled within a prescribed range $\Delta\theta$. Table 8 (right) shows that our method outperforms competing approaches when the angular error is within $\pm 2.5^\circ$ and remains competitive up to $\pm 7.5^\circ$, indicating tolerance to moderate pose-calibration errors. Fig. 11 provides visual comparisons under both noisy projections and angular perturbations. Compared with selected baselines, our method better preserves coherent anatomical structures and fine details under both types of acquisition uncertainty.

D Failure Analysis

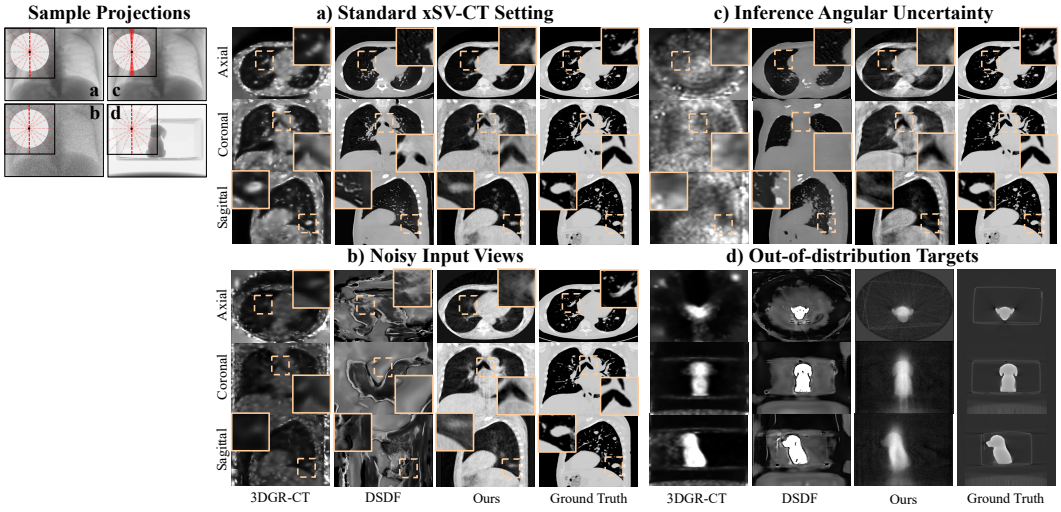

Figure 11: Systematic analysis of failure modes across four conditions, comparing our method with the best-performing optimization-based (3DGR-CT [12]) and learning-based methods (DSDF [14]) using 8 input views (0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°). The model of ours and the learning-based method are trained on CT-RATE dataset. Sample input projections as well as input angles are shown on the left. We provide tomographic reconstruction results under 4 different settings: **a**) Standard xSV-CT setting with 8 input views: our method preserves fine-grained anatomical details. **b**) xSV-CT with a simulated 5% dose of the 8 input views: our diffusion prior maintains structural integrity while baselines introduce severe artifacts. **c**) xSV-CT with the 8 input view angles randomly perturbed within $\pm 10°$: our method stays anchored to the available projections, while baselines develop geometric artifacts (3DGR-CT) or hallucinate structure (DSDF). **d**) xSV-CT on an out-of-distribution (OOD) target with respect to the training datasets: our method exhibits graceful blurring along the superior-inferior axis, yet still produces more faithful reconstructions than both baselines (note that for optimization-based methods there are no OOD targets in this sense).

Understanding failure modes is essential for responsible deployment. While our method demonstrates robustness to noisy projections and angular uncertainties (Table 8), we characterize the conditions under which it exhibits graceful degradations. For example, as shown in Table 8, the performance metrics drop when the input X-ray projections have excessive noise due to low dosage outputs (*e.g.*, less than a simulated 10% of that in the normal-dose CT dataset it is trained on) or if the angular input uncertainty exceeds a certain threshold (*e.g.*, more than a $\pm 7.5°$ of error between the model's angular input and the true camera angular pose). We visualize these failure modes in Fig. 11(**b**,**c**), with reconstructions from ours and the best-performing optimization-based (3DGR-CT [12]) and compared learning-based (DSDF [14]) methods. For context, Fig. 11(**a**) shows results under the standard xSV-CT setting, where our method excels at preserving fine-grained details.

Input X-ray dosage. When the input X-ray dosage drops below $\sim$10% of the normal-dose level, the signal-to-noise ratio of the input projections deteriorates to the point where $\mathcal{L}_{\text{geo}}$'s measurement fidelity term anchors the generation to noisy observations. Nonetheless, the diffusion prior maintains the global structural integrity of the reconstruction. Our method degrades more gracefully than both baseline methods, which introduce severe artifacts.

Angular uncertainty. When the angular input error exceeds $\pm 7.5°$, the mismatch between the assumed and true acquisition angles causes the geometric coherency loss to enforce consistency with an incorrect geometry. This manifests as local geometric distortions. Compared methods either develop geometric artifacts [12] or hallucinate structure [14], whereas our method stays anchored to the available projections and recovers the anatomy more faithfully.

Out-of-distribution Targets. Like the learning-based methods, our approach's performance is impaired when applied to out-of-distribution (OOD) targets, *i.e.*, data that lies outside the training distribution. As visualized in Fig. 11(**d**), OOD inputs introduce noticeable blurring effects along the

superior-inferior axis of the target and slight distortions at anatomical boundaries with our method. This is expected: the learned prior cannot synthesize anatomy it has never observed. However, even under these OOD conditions, our model still produces more visually faithful reconstructions compared to both the optimization-based [12] and learning-based [14] compared methods, thanks to the physics-based consistency constraints that remain valid regardless of anatomy.

The identified failure modes suggest concrete directions for future work: (1) noise-aware training or adaptive guidance weighting for low-dose robustness, (2) joint angle-projection optimization for angular uncertainty, and (3) foundation model pre-training or few-shot adaptation for OOD generalization.

Robustness to off-model physics To stress-test structured violations of the Radon model, we simulate metal-induced beam hardening following [76]. We insert several ellipsoidal metal implants (about 4000 voxels each, with intensities of 3000 HU under standard FBP reconstruction) into 10 CT-RATE cases and compute PSNR against the clean, metal-free volume (Table 9). FBP from 360 corrupted views propagates beam-hardening streaks throughout the volume, whereas our 8-view reconstruction retains fewer artifacts (26.4 vs. 20.7 dB). Outside the metal region, our method reaches 27.7 dB, only 0.9 dB below reconstruction from clean projections, while FBP remains at 21.2 dB. The implants remain correctly localized, so this gain reflects suppression of streak artifacts, not of the implant. Other off-model regimes behave differently. Severe sinogram truncation remains a limitation because both training and inference assume complete measurements, a limitation shared by most sparse-view CT methods. The effect of X-ray scatter is beyond the scope of this work without an accurate physical forward model.

Table 9: Reconstruction under simulated metal-induced beam hardening. PSNR is computed against the clean, metal-free volume; values in parentheses give the change relative to reconstructing from clean projections with our method at $M = 8$.

Variant	PSNR↑
FBP, M=360, corrupted projections	20.7
FBP, M=360, corrupted projections, non-metal region	21.2
Ours, M=8, corrupted projections	26.4 (-2.2)
Ours, M=8, corrupted projections, non-metal region	27.7 (-0.9)

E Additional Qualitative Results

We present more qualitative analysis for tomographic reconstruction and novel view synthesis across CT-RATE and SinoCT datasets under three challenging xSV-CT settings (1, 4, and 8 views). Two representative cases from each dataset are detailed in Table 10, which provides an index to the corresponding visualizations listed in this appendix. These comparisons show that our method achieves superior quantitative performance (Tables 1 and 2) and produces more visually plausible results. Please zoom in to compare fine anatomical details across methods. For further qualitative 3D validation, we refer to the accompanying **video**, which visualizes these cases across compared methods and ours.

Table 10: Index of qualitative results. Each entry links to the corresponding figure showing novel view synthesis (NVS) or tomographic reconstruction (TR) for a specific case and sparsity level.

Case No.		Case 1			Case 2		
No. Inference Views		1-view	4-view	8-view	1-view	4-view	8-view
CT-RATE	NVS	Fig. 12	Fig. 13	Fig. 14	Fig. 15	Fig. 16	Fig. 17
	TR	Fig. 18	Fig. 19	Fig. 20	Fig. 21	Fig. 22	Fig. 23
SinoCT	NVS	Fig. 24	Fig. 25	Fig. 26	Fig. 27	Fig. 28	Fig. 29
	TR	Fig. 30	Fig. 31	Fig. 32	Fig. 33	Fig. 34	Fig. 35

F Broader Impacts

Beyond the conventional motivation of sparse-view reconstruction, *i.e.*, reducing radiation exposure and associated health risks for both patients and operators, xSV-CT has the potential to broaden the range of scenarios in which tomographic reconstruction is practical. In intraoperative navigation, for example, current workflows often rely directly on planar X-ray projections for guidance; an xSV-CT regime could provide volumetric information from only a few acquisitions. For dynamic imaging scenarios, such as cardiac imaging, or high-throughput industrial inspection, reducing the number of required projections can shorten acquisition time, thereby improving temporal resolution or productivity. For radiation-sensitive biological samples, cultural-heritage artifacts, or archaeological specimens, the reduced imaging dose may also enable tomographic analysis of objects that would otherwise be difficult or unsafe to scan repeatedly.

However, these potential benefits should also be interpreted with caution. Like other sparse-view CT methods, xSV-CT operates under severely limited measurements and can therefore introduce reconstruction bias, hallucinated structures, or degradation of fine details. Such risks are particularly important in safety-critical clinical or industrial settings, where reconstructed volumes should not be treated as definitive evidence without appropriate validation, uncertainty assessment, and expert oversight. In addition, the current computational cost makes our method more suitable as an asynchronous reconstruction module: a full reconstruction may still require tens of minutes. Compared with analytical reconstruction methods that can produce results within minutes or less, this latency may limit applicability in real-time workflows. On geometry, the finite-order Range-Condition surrogate enters only $\mathcal{L}_{\mathrm{phy}}$, and $\mathcal{L}_{\mathrm{geo}}$ applies unchanged to cone-beam and non-coplanar geometries. Replacing the surrogate with Grangeat's condition [69] would remove the rebinning approximation for cone-beam data; full 3D reconstruction then also requires Tuy's sufficiency condition [77] through non-planar acquisition orbits. Reducing inference cost and quantifying reconstruction uncertainty remain important directions for future deployment.

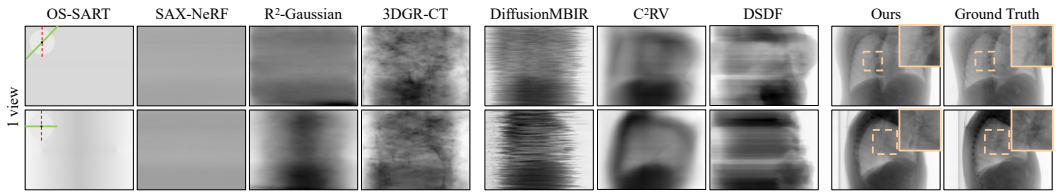

Figure 12: Novel View Synthesis results (45°, 90°) from 1 view (0°) on Case 1 from the CT-RATE dataset, the input (red) and novel (green) angles are also indicated in the inset polar graph. We provide zoomed-in details for results with plausible anatomy compared to their corresponding ground truths.

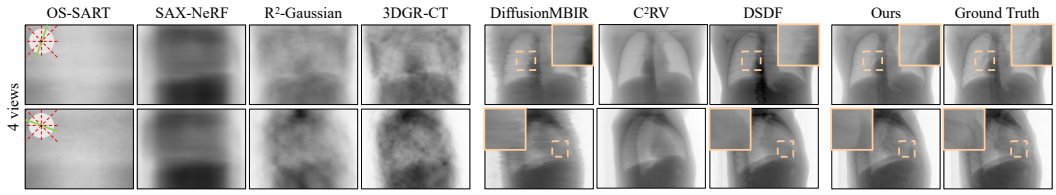

Figure 13: Novel View Synthesis results (22.5°, 112.5°) from 4 views (0°, 45°, 90°, 135°) on Case 1 from the CT-RATE dataset, the input (red) and novel (green) angles are also indicated in the inset polar graph. We provide zoomed-in details for results with plausible anatomy compared to their corresponding ground truths.

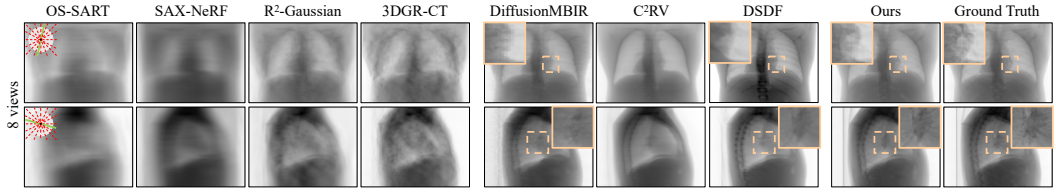

Figure 14: Novel View Synthesis results (11.25°, 101.25°) from 8 views (0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°) on Case 1 from the CT-RATE dataset, the input (red) and novel (green) angles are also indicated in the inset polar graph. We provide zoomed-in details for results with plausible anatomy compared to their corresponding ground truths.

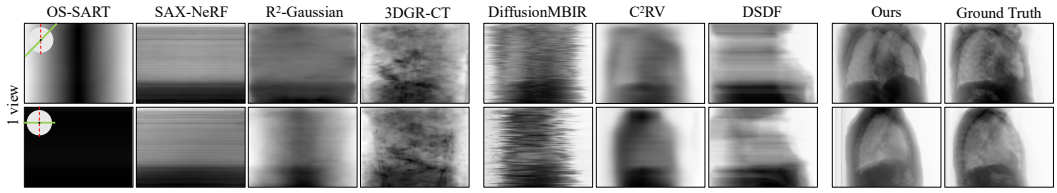

Figure 15: Novel View Synthesis results (45°, 90°) from 1 view (0°) on Case 2 from the CT-RATE dataset, the input (red) and novel (green) angles are also indicated in the inset polar graph.

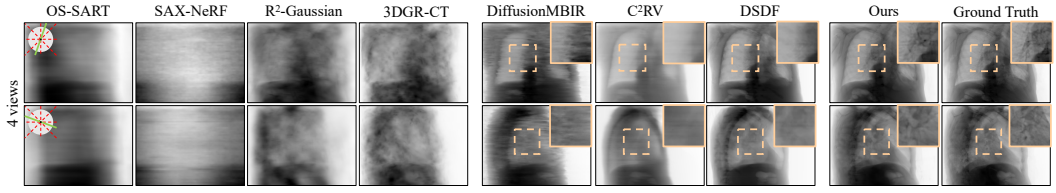

Figure 16: Novel View Synthesis results (22.5°, 112.5°) from 4 views (0°, 45°, 90°, 135°) on Case 2 from the CT-RATE dataset, the input (red) and novel (green) angles are also indicated in the inset polar graph. We provide zoomed-in details for results with plausible anatomy compared to their corresponding ground truths.

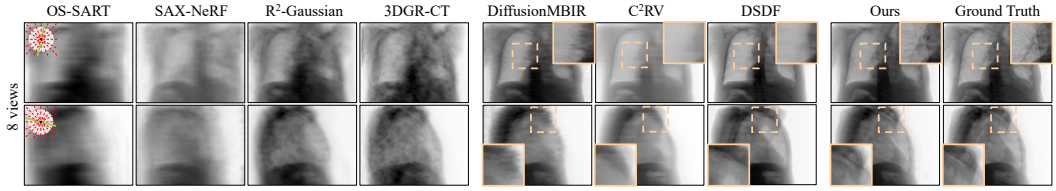

Figure 17: Novel View Synthesis results (11.25°, 101.25°) from 8 views (0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°) on Case 2 from the CT-RATE dataset, the input (red) and novel (green) angles are also indicated in the inset polar graph. We zoom in on methods with plausible details with the ground truths.

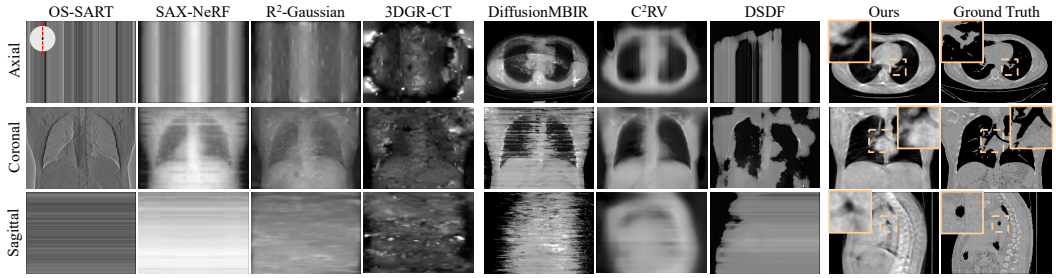

Figure 18: Tomographic Reconstruction results from 1 view (0°) on Case 1 from the CT-RATE dataset, the input angle is also indicated in the inset polar graph. We zoom in on methods with plausible details with the ground truths.

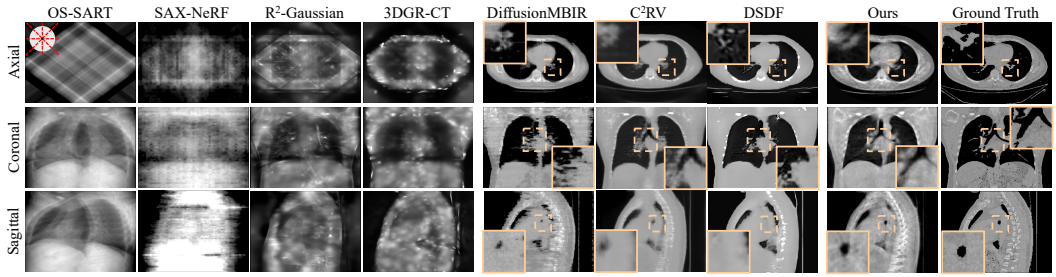

Figure 19: Tomographic Reconstruction results from 4 views (0°, 45°, 90°, 135°) on Case 1 from the CT-RATE dataset, the input angles is also indicated in the inset polar graph. We zoom in on methods with plausible details with the ground truths.

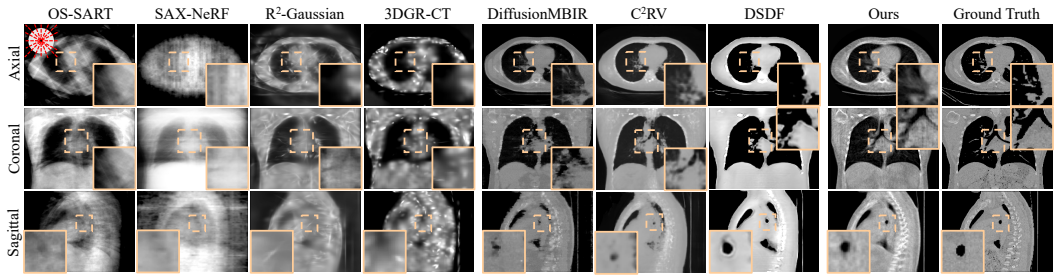

Figure 20: Tomographic Reconstruction results from 8 views (0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°) on Case 1 from the CT-RATE dataset, the input angles is also indicated in the inset polar graph. We zoom in on methods with plausible details with the ground truths.

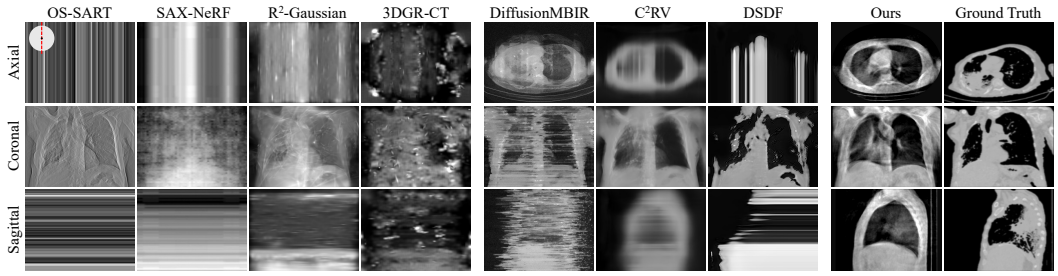

Figure 21: Tomographic Reconstruction results from 1 view (0°) on Case 2 from the CT-RATE dataset, the input angle is also indicated in the inset polar graph. We zoom in on methods with plausible details with the ground truths.

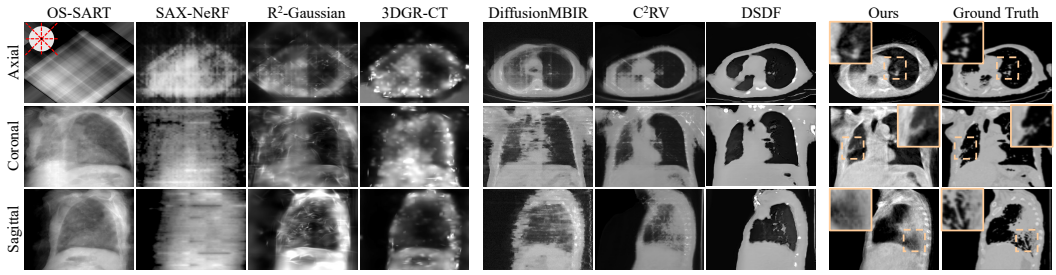

Figure 22: Tomographic Reconstruction results from 4 views (0°, 45°, 90°, 135°) on Case 2 from the CT-RATE dataset, the input angles is also indicated in the inset polar graph. We zoom in on methods with plausible details with the ground truths.

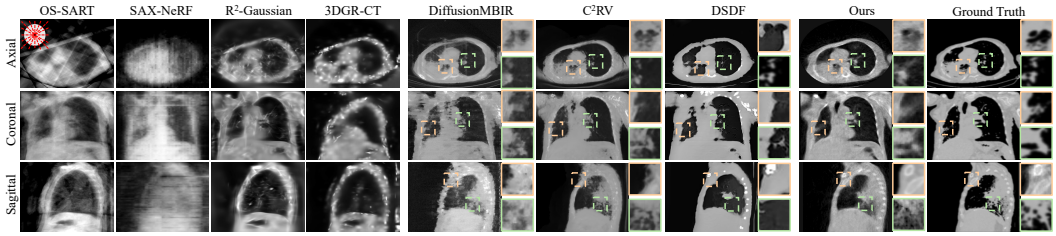

Figure 23: Tomographic Reconstruction results from 8 views (0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°) on Case 2 from the CT-RATE dataset, the input angles is also indicated in the inset polar graph. We zoom in on methods with plausible details with the ground truths.

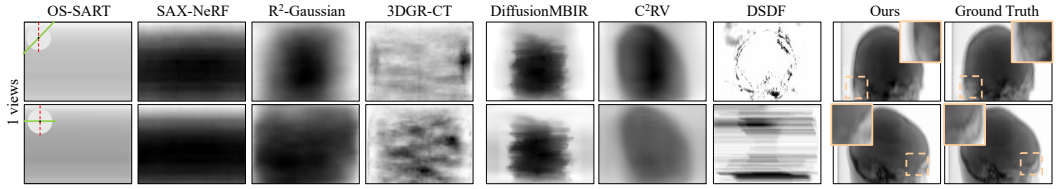

Figure 24: Novel View Synthesis results (45°, 90°) from 1 view (0°) on Case 1 from the SinoCT dataset, the input (red) and novel (green) angles are also indicated in the inset polar graph. We provide zoomed-in details for results with plausible anatomy compared to their corresponding ground truths.

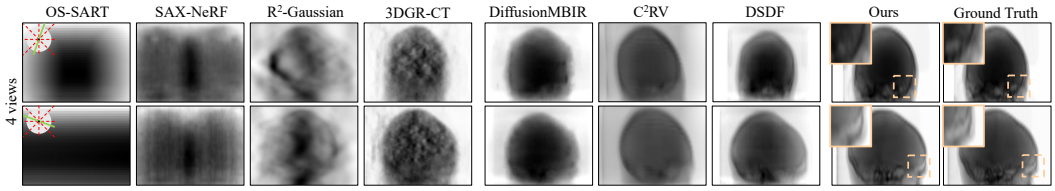

Figure 25: Novel View Synthesis results (22.5°, 112.5°) from 4 views (0°, 45°, 90°, 135°) on Case 1 from the SinoCT dataset, the input (red) and novel (green) angles are also indicated in the inset polar graph. We provide zoomed-in details for results with plausible anatomy compared to their corresponding ground truths.

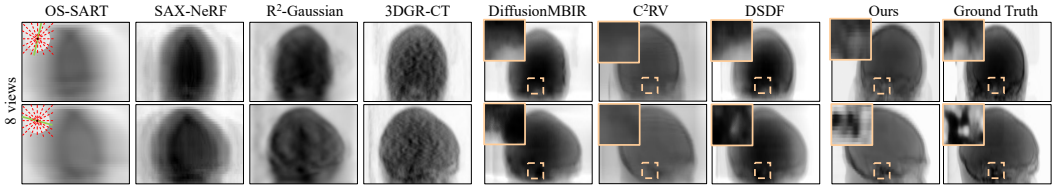

Figure 26: Novel View Synthesis results (11.25°, 101.25°) from 8 views (0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°) on Case 1 from the SinoCT dataset, the input (red) and novel (green) angles are also indicated in the inset polar graph. We provide zoomed-in details for results with plausible anatomy compared to their corresponding ground truths.

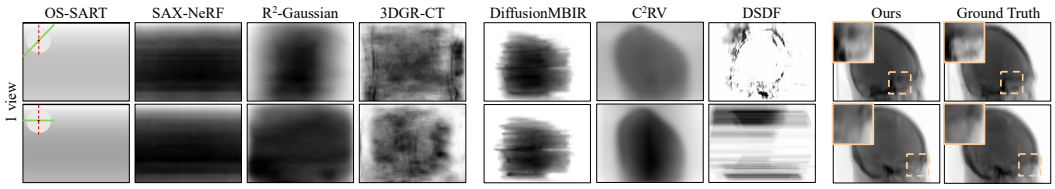

Figure 27: Novel View Synthesis results (45°, 90°) from 1 view (0°) on Case 2 from the SinoCT dataset, the input (red) and novel (green) angles are also indicated in the inset polar graph. We provide zoomed-in details for results with plausible anatomy compared to their corresponding ground truths.

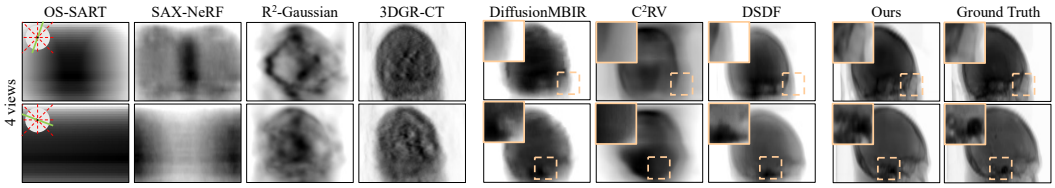

Figure 28: Novel View Synthesis results (22.5°, 112.5°) from 4 views (0°, 45°, 90°, 135°) on Case 2 from the SinoCT dataset, the input (red) and novel (green) angles are also indicated in the inset polar graph. We provide zoomed-in details for results with plausible anatomy compared to their corresponding ground truths.

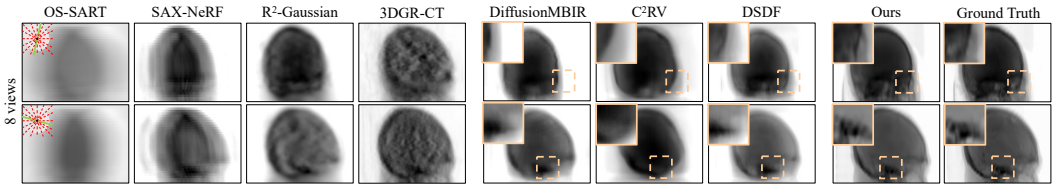

Figure 29: Novel View Synthesis results (11.25°, 101.25°) from 8 views (0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°) on Case 2 from the SinoCT dataset, the input (red) and novel (green) angles are also indicated in the inset polar graph. We provide zoomed-in details for results with plausible anatomy compared to their corresponding ground truths.

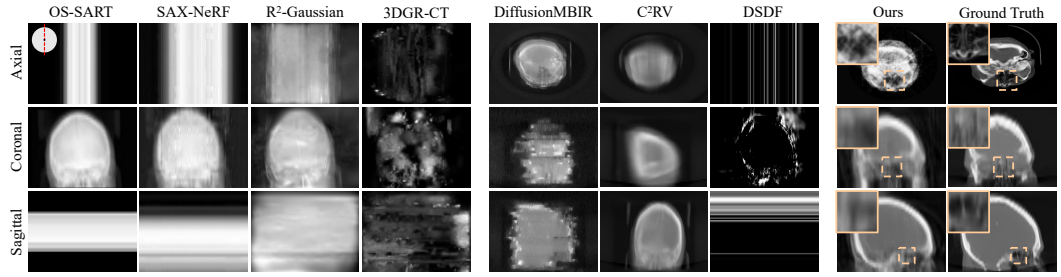

Figure 30: Tomographic Reconstruction results from 1 view (0°) on Case 1 from the SinoCT dataset, the input angle is also indicated in the inset polar graph. We zoom in on methods with plausible details with the ground truths.

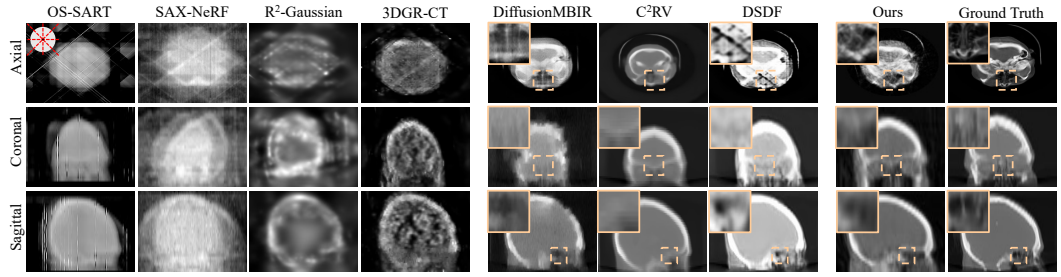

Figure 31: Tomographic Reconstruction results from 4 views (0°, 45°, 90°, 135°) on Case 1 from the SinoCT dataset, the input angles is also indicated in the inset polar graph. We zoom in on methods with plausible details with the ground truths.

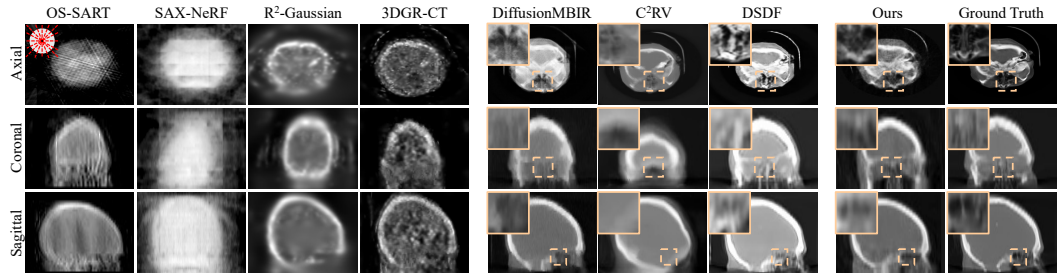

Figure 32: Tomographic Reconstruction results from 8 views (0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°) on Case 1 from the SinoCT dataset, the input angles is also indicated in the inset polar graph. We zoom in on methods with plausible details with the ground truths.

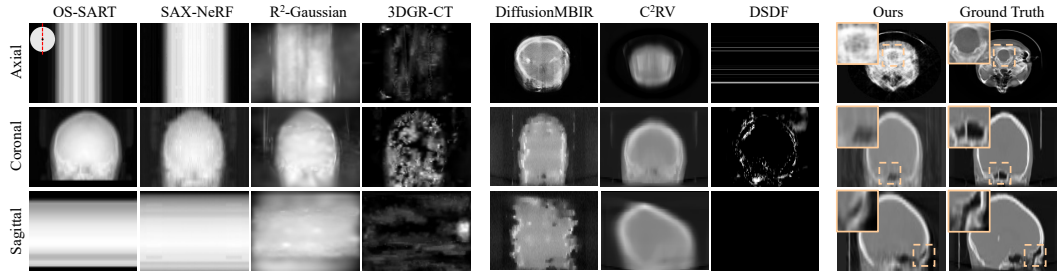

Figure 33: Tomographic Reconstruction results from 1 view (0°) on Case 2 from the SinoCT dataset, the input angle is also indicated in the inset polar graph. We zoom in on methods with plausible details with the ground truths.

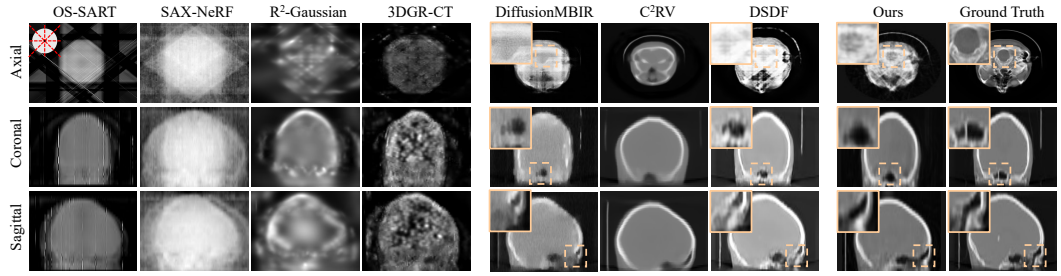

Figure 34: Tomographic Reconstruction results from 4 views (0°, 45°, 90°, 135°) on Case 2 from the SinoCT dataset, the input angles is also indicated in the inset polar graph. We zoom in on methods with plausible details with the ground truths.

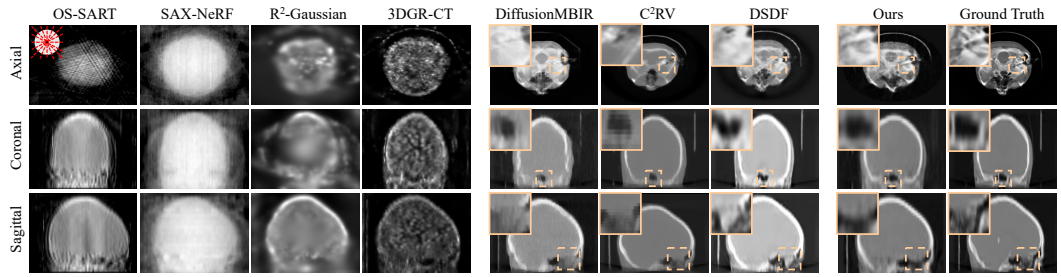

Figure 35: Tomographic Reconstruction results from 8 views (0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°) on Case 2 from the SinoCT dataset, the input angles is also indicated in the inset polar graph. We zoom in on methods with plausible details with the ground truths.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contribution of our paper is to enable the xSV-CT regime that requires extreme projection sparsity at both dataset curation and downstream application, which has been discussed in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of prior works has been discussed in abstract, introduction, and related works in more details.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper provides no novel theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The code, dataset links to public datasets used in this paper as well as data splits, and training checkpoints will be available in the GitHub link at the start of the appendix. Also, we provided extensive details of our dataset and implementation inside our appendices which helps reproduction.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have disclosed our core algorithm with a pseudo code in the appendix, the full disclosure of code, dataset links, as well as model checkpoints will be available in the GitHub link at the start of the appendix.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We include inference details are discussed in full in the Experiments section, exact splits along with other materials will be available in the GitHub link at the start of the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Standard deviation of ℓ_2 error of our method across 4 runs on novel view synthesis are reported in the Experiments section as well as more are reported in the appendices.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The memory footprints and execution time can be found in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have read and followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the method's broader impacts in the appendices.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our model is strictly trained on licensed public dataset sources, thus should not pose such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have read and complied with their dataset or code licenses, adequate citations have also been added to the main text.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will provide README files with the codebase in the GitHub link at the start of the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper provides no such experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper includes no such experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used strictly for paper polishing, and the core concept and code are originally constructed.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.