

Conditional Diffusion Posterior Alignment for Sparse-View CT Reconstruction

Luis Barba*

Johannes Kirschner†

Benjamín Béjar‡

Abstract

Computed Tomography (CT) is a widely used imaging modality in medical and industrial applications. To limit radiation exposure and measurement time, there is a growing interest in sparse-view CT, where the number of projection views is significantly reduced. Deep neural networks have shown great promise in improving reconstruction quality in sparse-view CT, especially generative diffusion models. However, these methods struggle to scale to large 3D volumes due to several reasons: (i) the high memory and computational requirements of 3D models, (ii) the lack of large 3D training datasets, and (iii) the inconsistencies across slices when using 2D models independently on each slice. We overcome these limitations and scale diffusion-based sparse-view CT reconstruction to large 3D volumes by combining conditional diffusion with explicit data consistency. We propose Conditional Diffusion Posterior Alignment (CDPA) to enable scalable 3D sparse-view CT reconstruction. A 2D U-Net diffusion model is conditioned on an initial 3D reconstruction to improve inter-slice consistency, combined with data-consistency alignment to match measured projections. Experiments on synthetic and real Cone Beam CT (CBCT) data show state-of-the-art performance, with ablations that confirm the synergistic effects of the proposed pipeline. Finally, we show that the same principles also strengthen fast denoising U-Nets, yielding near-diffusion quality at a fraction of the computational cost.

Funding: This project acknowledges funding support from the SDSC CHIP project under grant no. C22-11L.

1 Introduction.

Computed Tomography (CT) is a non-invasive imaging technique that reconstructs cross-sectional images of an object from its X-ray projections. It has become an essential tool in medical diagnostics, industrial inspection, and scientific research. However, acquiring high-quality CT images often requires a large number of projection views, leading to increased radiation dose and, in some settings, to longer scan times. *Sparse-view* CT aims to reduce the number of projections acquired while maintaining image quality, but it remains a challenging inverse problem due to the ill-posed nature of the reconstruction task.

When the number of projection angles is drastically reduced, to lower radiation dose or scan time, classical analytical algorithms such as *filtered back-projection* (FBP) and its Cone Beam CT (CBCT) counterpart, the *Feldkamp-Davis-Kress* (FDK) algorithm, produce severe streak artifacts. In the particular case of CBCT where we focus our experimental results, the reconstructions cannot be decomposed into independent slices as in parallel-beam geometry CT. Therefore, CBCT poses an intrinsically 3D problem, where the aforementioned artifacts

*Swiss Data Science Center (SDSC) in Paul Scherrer Institute (PSI), Villigen, Switzerland. E-mail: luis.barba-flores@psi.ch.

†Swiss Data Science Center (SDSC) and ETH Zurich, Switzerland. E-mail: johannes.kirschner@ethz.ch.

‡Swiss Data Science Center (SDSC) and Paul Scherrer Institute (PSI), Villigen, Switzerland. E-mail: benjamin.bejar@psi.ch.

span across different slices, and where 2D approaches tend to be insufficient. Early iterative reconstruction schemes with hand-made regularizers (e.g., total-variation) improve these artifacts but require long run-times, are sensitive to hyper-parameter tuning, and still fall short under high-sparsity scenarios.

Deep learning has transformed sparse-view CT by introducing data-driven priors that learn image statistics directly from examples. Supervised convolutional neural networks (CNNs), in particular U-Net-based models such as FBPCNet [14], have emerged as the standard reference method when used as post-processing on FBP/FDK reconstructions, providing good artifact suppression. We refer to these models as *FBP/FDK-denoisers*. More recent approaches have improved reconstruction quality through geometry-aware encoders [19] or sliding 3D sub-volumes [34], setting new state-of-the-art benchmarks for CBCT. In parallel, unsupervised methods such as Deep Image Prior [28] and score-based diffusion models [2, 4, 26] have emerged as powerful priors achieving state-of-the-art results. However, these methods are computationally expensive, making it prohibitive to scale them to large 3D-volumes. This leaves us with slice-by-slice approaches using 2D models that out-of-the-box suffer from inconsistencies across slices in 3D reconstructions.

Two strategies can restore 3D consistency to such a slice-by-slice 2D diffusion prior. The first imposes an explicit, hand-crafted regularizer coupling adjacent slices directly inside the sampling process, e.g. a total-variation penalty solved via ADMM, as in DiffusionMBIR [5]. The second, which we adopt, inherits 3D-consistent behavior by conditioning the diffusion model on an initial 3D reconstruction. We show in Sec. 6.4 that although the two strategies reach comparable PSNR and SSIM, the explicit-regularization route is not a free alternative: it over-smooths the axial axis relative to the anatomy’s own structure, reduces the sampler’s posterior variance, undermining the benefit of posterior averaging via the $\mu(\cdot)$ operator, and produces substantially worse perceptual quality. Our conditioning strategy matches its fidelity without incurring any of these three costs.

1.1 Contributions.

Motivated by these limitations, we propose the use of conditioning on slices of the FBP/FDK volume reconstructions to improve 3D consistency. We present two complementary 2D U-Net-based approaches for sparse-view CT: (i) Conditional Diffusion models with Posterior Alignment (CDPA) and (ii) refined FBP/FDK-denoising networks, both with slice-positional encoding and data-consistency alignment. Although our method applies to all CT geometry settings, we evaluate them in CBCT datasets that deal with an intrinsically 3D reconstruction problem, where sparse-view artifacts span different slices, and which is a more prominent setting in the medical community. Our key contributions are as follows.

- We demonstrate that conditioning on FBP/FDK reconstructions significantly improves the quality and 3D slice consistency of diffusion model output, addressing a known limitation of unconditional slice-by-slice approaches.
- We introduce CDPA, a conditional ReSample-style diffusion posterior-alignment framework for large-volume sparse-view CT. We show that it systematically outperforms unconditional diffusion on all benchmarked baselines.
- We show that a simple Monte Carlo averaging strategy over diffusion posterior samples yields substantial PSNR and, surprisingly, also SSIM improvements, suggesting near-unimodal posteriors in practice.
- We show that an alternative, explicit-regularization strategy for enforcing 3D consistency (DiffusionMBIR’s ADMM+TV guidance, applied to our own pretrained checkpoints with

no retraining) reaches comparable PSNR/SSIM to CDPA, but at the cost of axial over-smoothing, a collapsed posterior, and substantially worse perceptual quality – costs that our conditioning avoids entirely.

- We revisit the FBPCnvNet paradigm [14] with modern U-Net architectures incorporating cross-attention and slice-position conditioning for better artifact removal. Combined with inference-time data-consistency fine-tuning, this approach nearly matches diffusion performance at a fraction of the computational cost and surpasses all previous approaches.
- We provide extensive experiments across synthetic and real-world CBCT datasets, demonstrating state-of-the-art PSNR and SSIM compared to recent methods.

The code supporting the results in this paper can be found in https://github.com/SwissDataScienceCenter/cbct_cdpa.

2 Related Work.

Classical and iterative reconstruction. The filtered back-projection (FBP) algorithm is the classical analytical solution in parallel-beam CT, exploiting the Fourier slice theorem for exact inversion under dense angular sampling. Its cone-beam counterpart, the Feldkamp-Davis-Kress (FDK) algorithm, serves the same role in CBCT. Under aggressive sparse-view or low-dose regimes, however, both FBP and FDK violate their sampling assumptions and produce globally oriented coherent streak artifacts that obscure fine structure. Algebraic reconstruction techniques (ART) and variational formulations (e.g. quadratic data fidelity plus sparsity or total-variation penalties) alleviate streaking by explicitly regularizing the ill-posed inverse. Yet they entail (i) many forward/backprojection sweeps, (ii) sensitive balancing between data fidelity and regularization strength, and (iii) difficulty adapting the prior to heterogeneous textures (e.g. simultaneously preserving sharp edges and smooth backgrounds). As sparsity increases, hand-crafted priors either over-smooth or leave residual structured artifacts, motivating learned data-driven priors.

Supervised CNN post-processing. A widely adopted pattern is *FBP+CNN*: compute a fast baseline reconstruction (FBP or FDK) and feed it to a convolutional network that maps streak-corrupted images to clean images. Jin *et al.* [14] demonstrated that a moderately sized U-Net (FBPCnvNet) can remove severe streaks from extremely sparse sinograms (e.g., tens of views) in near real-time per slice while outperforming total-variation iterative solvers. We refer to models following this paradigm as *FBP/FDK-denoisers*. Subsequent variants refined network depth, multi-scale skip connections, loss functions, and training curricula, cementing U-Nets as a strong supervised baseline.

Recent state-of-the-art methods for CBCT have incorporated geometric awareness and 3D processing. GAAL [19] uses a geometry-aware encoder-decoder framework that encodes multi-view 2D features from various 2D X-ray projections with a 2D CNN encoder, back-projects the multi-view 2D features into 3D space to formulate a comprehensive volumetric feature map, followed by a 3D CNN decoder to recover the 3D CBCT image. S-STAR Net [34] tackles sparse-view reconstruction with sliding 3D sub-volumes, a difference-enhancement loss, volume-attention residuals, and Fourier-transform convolutions, achieving state-of-the-art PSNR/SSIM on walnut and CQ500 datasets under 30-view acquisition. Other supervised 3D baselines include PTNet3D [35], which learns a direct 3D mapping for sparse-view reconstruction, and Pix2Pix3D [13], which adapts conditional image-to-image translation (pix2pix) to 3D volumes. All these 3D-algorithms are, however, limited to low-resolution volumes due to memory constraints, and do not scale to the high-resolution CBCT volumes that we target in this work.

Unsupervised and zero-shot methods. In scenarios where large paired sinogram–image datasets are unavailable, *zero-shot* methods exploit architectural or generative priors without external training data. Deep Image Prior (DIP) [28] reconstructs a single scan by optimizing the weights of a randomly initialized CNN to fit the measurements, exploiting architectural bias alone as an implicit prior. The network, fed with random noise as input, is iteratively updated to minimize data fidelity; the resulting image often converges to a plausible solution before overfitting measurement noise. Although effective, DIP incurs substantial per-scan optimization overhead. Warm-start strategies [3] pre-train the network on synthetic samples and then fine-tune per instance, reducing iteration counts and stabilizing optimization. An “educated warm start” approach uses FBP reconstructions as network input instead of pure noise, further accelerating convergence. Still, DIP-style methods remain computationally heavy and do not scale well to large 3D volumes, making them less suited for medical settings and high-resolution scientific imaging.

Diffusion models for inverse problems. Score-based diffusion models [12] have emerged as powerful generative priors, learning time-indexed score functions of noisy data distributions and enabling high-fidelity sampling decoupled from specific forward operators. For inverse problems, a general strategy iteratively denoises from pure noise while nudging samples toward measurement consistency.

Diffusion Posterior Sampling (DPS) [4] formalized this conditioning as adding an approximate likelihood gradient to the learned score, improving stability under noise and allowing flexible noise models without retraining the diffusion prior. DDRM (Denoising Diffusion Restoration Models) [15] leverages a spectral factorization to incorporate measurement information analytically for certain linear operators, though computing or storing such decompositions is impractical for large CT system matrices.

Latent diffusion with hard data consistency (e.g., ReSample [26]) extends these ideas to work with latent image representations. It introduces an optimization-based latent refinement that enforces exact forward consistency before a resampling step restores adherence to the learned latent manifold. Recent work [2] has adapted these ideas to non-latent diffusion for CT, further improving reconstruction quality. Deterministic samplers such as DDIM (Denoising Diffusion Implicit Models) [27] reduce wall-clock time by using non-Markovian trajectories with fewer function evaluations, which is critical given the computational burden of repeated forward/backprojection inside each reverse diffusion step.

Closely related to our work, DOLCE [18] introduced conditional diffusion models for limited-angle CT reconstruction, integrating data-consistency updates with the sampling steps of a diffusion model conditioned on the limited-angle measurements. DOLCE enforces measurement consistency via proximal gradient steps at each reverse diffusion iteration. Our CDPA method is conceptually related, but differs in targeting large-volume sparse-view CBCT and in using a ReSample-style posterior-alignment approach [26] rather than proximal gradients, combined with FDK-conditioning and slice-position conditioning for 3D consistency.

Despite their success, when applied slice-by-slice with 2D models to 3D volumes, diffusion methods can produce inter-slice inconsistencies because each slice is processed independently without enforcing 3D coherence. While 3D diffusion models could address these limitations in theory, in practice their memory and computational demands are prohibitive at high resolutions, not to mention the increased amount of training data required. Two lines of work address this with 2D priors only: Lee et al. [16] attempted to address this issue by modeling the 3D data distribution as a product of 2D distributions sliced in different axes and using them alternately during reconstruction, and DiffusionMBIR [5] augments a 2D diffusion prior with a one-dimensional total-variation penalty along the remaining axis inside an ADMM data-consistency step. We compare directly against DiffusionMBIR’s guidance mechanism in Sec. 6.4, applying it to our own pretrained DPA/CDPA checkpoints with no retraining. Despite reaching

comparable but lower PSNR/SSIM to CDPA there, we find it introduces three concrete costs – axial over-smoothing, a collapsed posterior, and substantially worse perceptual quality – that FDK-conditioning avoids.

Summary and positioning. Supervised CNNs exhibit good performance but fall short of state-of-the-art performance. Zero-shot methods (DIP) remove dependency on large datasets but incur high per-instance cost and do not exploit data priors, thus being surpassed by models using learned priors. Diffusion models unify strong generative fidelity with flexible conditioning, and recent advances (DPS, latent consistency, accelerated samplers) make them increasingly practical for CT. However, scaling diffusion methods to large 3D volumes remains challenging due to memory, compute, and training data limitations, leading to slice-by-slice approaches that suffer from inter-slice inconsistencies.

Our work centers on exploiting diffusion posterior structure and data-consistency refinement under sparse-view regimes. We introduce Conditional Diffusion Posterior Alignment to scale 2D-diffusion approaches to 3D volumes using data-consistency alignment, as well as slice-position and FDK/FBP conditioning. We also show that these techniques can be applied to FBP/FDK-denoisers elevating them to near-diffusion performance without major architectural changes. Our 2D-based methods scale to much higher resolution than intrinsic 3D models, due to the independent processing of slices.

3 Method.

In this section, we describe the methods used in our experiments. We focus on two main approaches: diffusion-based models and FBP/FDK-denoising models, both based on the U-Net architecture.

Problem setup and notation. We denote by $\mathbf{x}_0 \in \mathbb{R}^N$ the (unknown) target image and by $\mathcal{A} : \mathbb{R}^N \rightarrow \mathbb{R}^M$ the (discrete) forward operator (parallel beam Radon transform or Cone Beam Projection). We index projection views by an angle $\psi \in \Psi$, where the finite acquired set is $\boldsymbol{\psi} = \{\psi_1, \dots, \psi_K\}$. We decompose $\mathcal{A}$ into per-angle operators $\mathcal{A}_\psi$ so that the full set of projections is

$$\mathbf{y} = \{\mathbf{y}_\psi\}_{\psi \in \Psi}, \quad \mathbf{y}_\psi = \mathcal{A}_\psi \mathbf{x}_0 + \epsilon_\psi,$$

with ϵ_ψ measurement noise (assumed i.i.d. Gaussian). Thus $\mathbf{y}_\psi$ denotes the detector readings for view ψ .

3.1 Score-based diffusion priors and posterior sampling.

The main idea of our method is to train a diffusion model to learn the target image distribution, and then use data-consistency guidance to align the diffusion sample with the measurements as done in [2]. However, in addition to this guidance step, we make the diffusion model conditional by passing the sparse FDK/FBP reconstruction as input to approximate the conditional score function. Our approach combines these two conditional sampling approaches for superior performance. We describe this method in detail below.

We adopt score-based diffusion models as learned image priors. Given a clean sample $\mathbf{x}_0$, we construct a noisy $\mathbf{x}_t$ at noise level (or time) $t \in [0, 1]$ via $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$ with known scalar schedules (α_t, σ_t) (e.g., derived from a variance schedule in DDPM (Denoising Diffusion Probabilistic Models) [12]). The marginal $p_t(\mathbf{x}_t)$ denotes the probability density of the random variable $\mathbf{x}_t$ at time t when $\mathbf{x}_0$ is drawn from the empirical data distribution. A *score network* $s_\theta(\mathbf{x}_t, t)$ is trained to approximate $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ (the *score* of the noisy marginal) by minimizing a denoising-score matching objective. We use the common ϵ -parameterization that predicts the added noise $\epsilon_\theta(\mathbf{x}_t, t)$; the two are related by a scaling factor. From a network output we form

a denoised (Tweedie) estimate $\hat{\mathbf{x}}_0(\mathbf{x}_t)$ of $\mathbf{x}_0$ which we will later use inside a data-consistency loss function. When conditioning on measurements $\mathbf{y}$ our goal is to sample from the posterior distribution $p(\mathbf{x}_0 | \mathbf{y})$. At inference, we use a fast sampler based on DDIM [27].

Diffusion posterior alignment (DPA). For reconstruction we target the posterior over clean images given measurements $p(\mathbf{x}_0 | \mathbf{y})$. Using the diffusion prior, we can express the posterior at time t as $p_t(\mathbf{x}_t | \mathbf{y})$ and its score as:

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \mathbf{y}) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p_t(\mathbf{y} | \mathbf{x}_t). \quad (1)$$

Diffusion Posterior Sampling (DPS) [4] approximates the term $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{y} | \mathbf{x}_t)$ by propagating a data-consistency gradient from a current denoised estimate $\hat{\mathbf{x}}_0(\mathbf{x}_t)$ (e.g., via Tweedie’s formula) back to $\mathbf{x}_t$. We define the per-step data-consistency loss

$$L_{\text{DC}}(\mathbf{x}_t) = \sum_{\psi \in \Psi} \|\mathcal{A}_\psi \hat{\mathbf{x}}_0(\mathbf{x}_t) - \mathbf{y}_\psi\|_2^2. \quad (2)$$

DPS [4] uses $\nabla_{\mathbf{x}_t} L_{\text{DC}}$ (through $\hat{\mathbf{x}}_0$) as a likelihood score surrogate (scaled by an empirically tuned guidance weight). Each reverse diffusion update is then nudged toward the measurement manifold while the prior score preserves realism. This however requires to backpropagate through the diffusion model, which is computationally expensive, and is not compatible with the DDIM sampler we want to use. Instead, we use a variant introduced by [26] that takes the gradient $\nabla_{\hat{\mathbf{x}}_0(\mathbf{x}_t)} L_{\text{DC}}$ with respect to the Tweedie’s formula estimation $\hat{\mathbf{x}}_0(\mathbf{x}_t)$. That is, we stay in image space without backpropagating the gradient to the noisy manifold on $\mathbf{x}_t$. After performing several gradient steps using $\nabla_{\hat{\mathbf{x}}_0(\mathbf{x}_t)} L_{\text{DC}}$, ensuring data consistency, we map back the resulting image to the noisy manifold at time t . To this end, we use the ReSample algorithm introduced by [26], but avoiding the use of latent space diffusion models. Our approach is the same as in [2], but we extend it to CBCT geometries in this paper. We refer to this unconditional diffusion posterior sampling with data consistency alignment method as *Diffusion Posterior Alignment* (DPA).

Conditional diffusion posterior alignment (CDPA). In the conditional formulation, we use our neural network to approximate directly the conditional score function by training it to predict the score of the conditional distribution with $\mathbf{y}$ as input in the form of its FBP/FDK reconstruction, i.e.,

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \mathbf{y}) \approx \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \text{FDK}(\mathbf{y})). \quad (3)$$

We end thus with two ways to approximate the posterior score: the conditional score network and the data-consistency gradient. Our final posterior score approximation is given by the combination of both (1) and (3). We achieve this by using the conditional diffusion model to obtain the estimate $\hat{\mathbf{x}}_0$ through Tweedie’s formula, and then applying data-consistency gradient descent updates to enforce data consistency on this estimate before mapping back to the noisy manifold at time t using the aforementioned ReSample step. Throughout this work, we refer to this method as *Conditional Diffusion Posterior Alignment* (CDPA).

Technically, one could use only the conditional score network without the data-consistency alignment step. However, we found that not using the data-consistency alignment step results in much inferior reconstruction quality and reconstructions that are almost never data-consistent.

Slice conditioning. We noticed that due to the cone-beam geometry of CBCT, artifacts generated by FDK in the sparse-view setting differ across regions of the reconstructed volume. This led us to design a model that is position-aware and is capable of handling artifacts differently in different parts of the volume. To this end, we add an additional conditioning to the U-Net behind the diffusion model by using the slice-index of the reconstructed volume. More

specifically, we used the cross-attention layers of the U-Net architecture to attend to a positional encoding of this slice index in the bottleneck layers of the U-Net model.

3.2 FBP/FDK-denoising models.

We revisit the ideas of [14, 24] but using larger modern U-Net architectures with cross-attention layers. The network takes as input an initial FBP/FDK reconstruction $\hat{\mathbf{x}}_{\text{FBP}}$ obtained from the projections $\mathbf{y}$ and produces as output a denoised image $\hat{\mathbf{x}}_0$. The network is trained to minimize the difference between $\hat{\mathbf{x}}_0$ and the corresponding ground-truth image $\mathbf{x}_0$ in the ℓ_2 sense. The training dataset consists of pairs of sinograms and ground-truth images, coming from the same distributions as the images used in the testing phase. In addition to the FBP/FDK reconstruction, we also condition the network on the slice index within the volume, following the same positional encoding strategy as in transformer models. This spatial-aware conditioning helps the network to better adapt to the varying artifact patterns that depend on the slice position which are common in CBCT reconstructions.

We notice that $\hat{\mathbf{x}}_0$ is almost never data consistent. That is, the output $\hat{\mathbf{x}}_0$ does not satisfy the data-consistency condition $\mathcal{A}\hat{\mathbf{x}}_0 = \mathbf{y}$. This is a common issue with FBP/FDK-denoising models, as they focus on denoising the FBP reconstruction without explicitly enforcing consistency with the measured data. Rather than adding an additional data-consistency term to the loss function during training, which did not produce satisfactory results, we propose a simple fine-tuning step at inference time to enforce data-consistency. We project $\hat{\mathbf{x}}_0$ onto the data-consistency manifold by minimizing the data-consistency loss L_{DC} from equation (2) for a few iterations using a small learning rate with the Adam optimizer. This step refines the reconstruction by enforcing consistency with the measured data $\mathbf{y}$.

4 Datasets.

To evaluate the effectiveness of our methods, we conducted CBCT experiments using both simulated datasets (dental and spine [6]) and a real-world dataset (walnut [7]). Our low-resolution experiments (256^3 voxel volumes) follow the same setup and evaluation metrics as in [19] for fair comparison with recent state-of-the-art methods. These low-resolution datasets are available on Hugging Face at https://huggingface.co/datasets/Zhentao-Liu/TMI2024_SVCT_dataset. Our high-resolution experiments on the full resolution walnut dataset [7] (501^3 voxel volumes) further demonstrate the scalability and effectiveness of our methods in more challenging scenarios. We summarize the datasets, preprocessing steps, and evaluation metrics below.

4.1 CBCT Datasets.

Walnut (real CBCT). We adopt 42 real walnut scans, split 37/5 [7]. This dataset consists of three scans at different heights (high/middle/low), each with 1200 views at 0.3° steps, with recorded flat-field (I_1) and dark-field (I_0). A reference *ground-truth* CBCT is reconstructed by combining the three height scans using an iterative procedure. We work with two resolution settings for this dataset. The first, at low resolution, allows us to compare our method with baselines that struggle to scale to full resolution. Following exactly [19], raw projections are downsampled from 768×972 with an effective detector pixel size of 0.1496 mm to a third of the size, resulting in projections of size 256×324 and reference volumes with 256^3 voxels. For model inputs we use uniformly spaced $N = 20$ views from the *middle* scan.

We also work with full resolution projections resulting in reference volumes with 501^3 voxels. In this setting, we restrict our models to work with the more challenging *high* scan (1200 views at 0.3° steps). This produces more artifacts to further highlight the efficiency of our methods. Inputs use $N \in \{20, 40, 60, \dots, 160, 180\}$ uniformly spaced views to test the effect of varying sparsity. To compare directly with S-STAR Net [34], we also train and test with $N = 30$ views.

However, since their algorithm works only with a crop due to computational resources, we crop our 501^3 reconstructed volumes to a central crop with resolution $360 \times 360 \times 150$ to exactly match their setting.

Dental (simulated). For this dataset, we follow the exact dataset from [19]. They used 130 dental CBCT volumes from patients resampled to 256^3 voxels at $(0.3133 \text{ mm})^3$, split 100/10/20 for train/val / test. For each case, we generate synthetic X-ray projections with a circular cone-beam trajectory using the ASTRA toolbox [21, 29, 30]: $N = 20$ uniformly spaced views over $[0^\circ, 360^\circ)$, detector size 256×256 with pixel spacing 0.4386 mm, source-object distance $L_{sb} = 500$ mm, and object-detector distance $L_{bd} = 200$ mm [19]. Both the spine and dental datasets are third-party, de-identified public releases. CTSpine1K states that all volumes are de-identified in accordance with the IRB policies of its contributing sites [6]; the dental CBCT volumes are redistributed under the GAAL/TMI2024 release [19] under its original license. The present study performed no new human-subject data acquisition and required no additional ethical approval.

Spine (CTSpine1K-derived, simulated). For this dataset, we follow the exact dataset from [19]. They used 130 CT volumes from CTSpine1K [6], resample/crop/pad each to 256^3 with $(2 \text{ mm})^3$ voxels, split 100/10/20, and mirror the dental view protocol ($N = 20$ uniform views) to generate synthetic projections. These projections use a 256×256 detector with 3 mm pixels, $L_{sb} = 1000$ mm and $L_{bd} = 500$ mm. This protocol is noiseless, while the walnut experiments use real, noisy measurements.

Projection correction for real data (walnut). Raw detector counts I are converted to line integrals via Beer-Lambert with per-pixel flat/dark correction:

$$P = -\ln\left(\frac{I - I_0}{I_1 - I_0}\right),$$

where I_0 and I_1 are the dark-field and flat-field measurements, respectively.

5 Experiments.

For each of the datasets described in Section 4, we train both 2D diffusion-based models and FBP/FDK-denoising models using the methods described above and the corresponding training splits. The low-resolution datasets are restricted to $N = 20$ views, while for the high-resolution walnut dataset we experiment with varying numbers of views $N \in \{20, 40, 60, \dots, 160, 180\}$. For this high-resolution setting, we train a single model for all view counts by randomly selecting the number of views at each training iteration. This yields a single model that can reconstruct images from different levels of sparsity without training a separate model for each case. Our model then includes an additional conditioning input on the number of views, allowing it to differentiate between sparsity levels. For the specific case of $N = 30$ views, we fine-tuned the model with samples at this specific sparsity to enable a fair comparison with S-STAR Net [34]. We also use this setting to compare our method with explicit-regularization (DiffusionMBIR); see below.

To train our diffusion model, we use the standard MSE loss on the noise prediction as in [12] for both the unconditional and conditional models. For the FBP/FDK-denoising models, we use the ℓ_2 loss between the denoised output and the ground-truth image. We use the Adam optimizer for all our experiments. More details about the training hyperparameters and architectures are provided in the supplementary material A.

Table 1: Quantitative comparison of reconstruction methods across three CBCT datasets ($n = 20$ test volumes for Dental and Spine, $n = 5$ for Walnut). Values are reported as mean $\pm$ standard deviation. DPA stands for Diffusion Posterior Alignment, CDPA for Conditional Diffusion Posterior Alignment, and FT for Fine-Tuning. The operator μ indicates the mean of 10 samples from the posterior distribution. Runs with * indicate results taken from GAAL’s paper [19]; DIF-Net and GAAL (unstarred) were re-run by us using the authors’ released checkpoints/code under our evaluation protocol, all other unstarred rows are our own methods. The best value per column is in **bold**.

Method	Dental (20 views)		Spine (20 views)		Walnut (20 views)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
FDK	22.54 $\pm$ 0.57	0.430 $\pm$ 0.021	22.77 $\pm$ 1.43	0.381 $\pm$ 0.040	16.07 $\pm$ 0.22	0.194 $\pm$ 0.004
GD _{zero}	27.68 $\pm$ 0.95	0.785 $\pm$ 0.015	27.77 $\pm$ 2.18	0.868 $\pm$ 0.039	23.93 $\pm$ 0.39	0.590 $\pm$ 0.008
GD _{FDK}	27.97 $\pm$ 0.46	0.745 $\pm$ 0.018	28.34 $\pm$ 1.55	0.821 $\pm$ 0.051	23.36 $\pm$ 0.24	0.582 $\pm$ 0.006
NAF [32]*	28.77 $\pm$ 0.86	0.793 $\pm$ 0.020	30.80 $\pm$ 1.93	0.912 $\pm$ 0.024	25.36 $\pm$ 0.22	0.680 $\pm$ 0.006
SCOPE3D [11]*	29.39 $\pm$ 0.87	0.807 $\pm$ 0.019	30.91 $\pm$ 2.00	0.908 $\pm$ 0.024	25.92 $\pm$ 0.19	0.714 $\pm$ 0.009
SNAF [9]*	30.93 $\pm$ 0.51	0.844 $\pm$ 0.015	31.69 $\pm$ 1.39	0.902 $\pm$ 0.035	27.43 $\pm$ 0.29	0.748 $\pm$ 0.006
PatRecon [25]*	19.95 $\pm$ 0.85	0.569 $\pm$ 0.038	18.57 $\pm$ 1.75	0.641 $\pm$ 0.047	17.30 $\pm$ 0.55	0.699 $\pm$ 0.033
PixelNeRF [31]*	26.85 $\pm$ 0.57	0.775 $\pm$ 0.014	28.68 $\pm$ 1.64	0.890 $\pm$ 0.033	24.79 $\pm$ 0.15	0.732 $\pm$ 0.004
DIF-Net [17]	30.48 $\pm$ 0.80	0.870 $\pm$ 0.015	32.28 $\pm$ 1.75	0.901 $\pm$ 0.034	25.99 $\pm$ 0.14	0.707 $\pm$ 0.004
GAAL [19]	31.44 $\pm$ 1.00	0.891 $\pm$ 0.016	33.14 $\pm$ 1.88	0.938 $\pm$ 0.020	26.84 $\pm$ 0.37	0.889$\pm$0.006
DPA	30.82 $\pm$ 0.80	0.831 $\pm$ 0.021	32.85 $\pm$ 1.66	0.892 $\pm$ 0.028	25.34 $\pm$ 0.26	0.648 $\pm$ 0.011
CDPA	33.32 $\pm$ 0.69	0.890 $\pm$ 0.013	35.84 $\pm$ 1.74	0.939 $\pm$ 0.020	29.58 $\pm$ 0.34	0.795 $\pm$ 0.012
FDK-denoiser	33.07 $\pm$ 0.90	0.901 $\pm$ 0.013	36.62 $\pm$ 1.69	0.950 $\pm$ 0.018	28.18 $\pm$ 0.98	0.888 $\pm$ 0.005
μ (DPA)	33.61 $\pm$ 0.77	0.906 $\pm$ 0.012	35.11 $\pm$ 1.68	0.942 $\pm$ 0.014	26.21 $\pm$ 0.27	0.666 $\pm$ 0.006
μ (CDPA)	34.72$\pm$0.70	0.921$\pm$0.010	37.08$\pm$1.83	0.952$\pm$0.016	30.50$\pm$0.83	0.813 $\pm$ 0.014
FDK-denoiser + FT	33.30 $\pm$ 0.72	0.882 $\pm$ 0.012	36.90 $\pm$ 1.63	0.944 $\pm$ 0.024	28.90 $\pm$ 0.54	0.740 $\pm$ 0.014

5.1 Baseline Methods.

We compare our proposed methods against several state-of-the-art baselines for sparse-view CT reconstruction. For the 256³ resolution CBCT experiments, we compare against all baselines reported in GAAL [19].

FDK. The Feldkamp-Davis-Kress (FDK) algorithm serves as the analytical baseline for CBCT reconstruction. Under extreme sparse-view conditions, FDK produces severe streak artifacts that motivate data-driven approaches.

Iterative methods. We include gradient descent-based iterative reconstruction (GD) using data-consistency terms but lacking learned priors. We benchmark two variants of this approach, one initialized with zeros, GD_{zero}, and a second initialized with the FDK reconstruction, GD_{FDK}. Gradients are computed with respect to (2). To this end, we developed a PyTorch wrapper in the ASTRA toolbox [21, 29, 30] that allows us to compute projections and backpropagate gradients to image space. We use the Adam Stochastic Gradient Descent optimizer with mini-batches of 30 projections to improve performance and speed. Our fine-tuning steps and alignment diffusion guidance are implemented by calling this GD module.

Supervised CNN baselines. We compare against several recent approaches reported in GAAL [19]: PatRecon [25] and SCOPE3D (a 3D convolutional approach) [11].

Neural rendering methods. We include NAF (neural attenuation fields) [32], SNAF (sparse-view neural attenuation fields) [9] and PixelNeRF [31] which leverage neural rendering techniques for 3D reconstruction. The results of these methods are taken from [19].

Recent state-of-the-art. We compare against two recent state-of-the-art methods: GAAL [19], which employs a geometry-aware encoder-decoder framework with multi-view 2D features and 3D volumetric reconstruction, and DIF-Net [17], a deep image fusion network for sparse-view CT.

Explicit-regularization guidance (DiffusionMBIR). We also compare against Diffusion-MBIR [5], which augments a *pretrained, unconditional* 2D diffusion prior with an ADMM data-consistency step that enforces an explicit 1D total-variation penalty coupling adjacent slices along the axial axis, solved via alternating CG least-squares updates and soft-thresholding at every reverse-diffusion timestep. We implement this guidance mechanism as a drop-in replacement for our own alignment guidance (Sec. 3), which lets us apply it directly to our existing pretrained DPA and CDPA checkpoints without retraining: DPA+ADMM-TV reproduces DiffusionMBIR, while CDPA+ADMM-TV additionally retains our FDK-conditioning, letting us test whether the two consistency mechanisms are complementary. Because the ADMM penalty weight and iteration budget require per-dataset tuning, we restrict this comparison to the walnut dataset at $N = 30$ views, full 501^3 resolution – the real-measurement, high-resolution regime central to this paper’s scaling claim (Sec. 6.4).

All baseline results marked with * in Table 1 are taken directly from the GAAL paper [19] to ensure fair comparison under identical evaluation protocols.

For the 501^3 resolution CBCT experiments, we compare our methods against FDK and gradient descent-based iterative reconstruction (GD) baselines only, as most other methods struggle to scale to this higher resolution. For the case of 30 views, we compare our results with S-STAR Net [34] and the baselines presented there, which is the only method, together with DiffusionMBIR, that we are aware of that scales close to this resolution, though S-STAR Net only operates on a center crop of the full-resolution walnut volume. Since no code was provided for S-STAR Net, we compare against the results reported in their paper. Besides S-STAR-Net, we include their results on PTNet3D [35] and Pix2Pix3D [13] in our comparison.

5.2 Evaluation Metrics.

We evaluate the reconstruction quality in sparse-view tomography using three groups of metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) for pixel-wise fidelity, the Learned Perceptual Image Patch Similarity (LPIPS) for perceptual quality, and an adjacent-slice total-variation ratio that directly measures inter-slice consistency. PSNR measures the ratio between the maximum possible pixel value squared and the Mean Squared Error (MSE) between the reconstructed and reference images, providing a quantitative measure of reconstruction quality. SSIM, on the other hand, assesses the visual impact of structural information in the image, making it a more perceptually relevant metric. Because generative reconstruction methods can score well on PSNR/SSIM while still hallucinating implausible texture, or score worse on them while looking diagnostically more natural, we complement both with LPIPS, and we further check for a specific structural failure mode: inter-slice inconsistency along the axis our 2D model processes, with a dedicated TV-ratio metric.

For the low-resolution CBCT results, including the dental, spine, and walnut datasets, we follow the exact evaluation protocol used in [19] to ensure fair comparison with recent state-of-the-art methods. This includes a predefined clamping range of values for PSNR and SSIM computation. For the dental, spine, and walnut datasets, respectively, we use the following ranges: dental: $[0.0, 0.090]$, spine: $[0.0, 0.051]$, walnut: $[0.0, 0.084]$ (unit-less attenuation coefficients). For SSIM we use an average of the 2D SSIM computed slice-by-slice along the three orthogonal axes (axial, coronal, sagittal) to better capture 3D structural similarity. PSNR is computed in 3D directly.

Perceptual quality (LPIPS). PSNR and SSIM are computed directly on attenuation values and are not designed to penalize plausible-but-incorrect texture, a failure mode specific to generative reconstruction methods. We therefore complement them with the Learned Perceptual Image Patch Similarity (LPIPS) [33], computed slice-by-slice as the distance between deep-network feature activations of the reconstructed and reference images and averaged over the volume (lower is better). Because posterior averaging (e.g. $\mu(\text{CDPA})$) is by construction the MSE-optimal, pixel-wise estimator, LPIPS lets us quantify the realism/accuracy trade-off it introduces relative to a single posterior sample, rather than relying on a PSNR/SSIM gap alone. We report LPIPS for the high-resolution walnut experiments, alongside the explicit-regularization comparison (Sec. 6.4, Table 3).

Inter-slice consistency (TV ratio). Our 2D diffusion model processes slices independently along the axial axis and has no explicit mechanism to enforce consistency between neighbouring axial slices, unlike the coronal and sagittal directions, which it never processes directly; volume-averaged PSNR, SSIM and LPIPS cannot expose this asymmetry. We therefore measure inter-slice consistency directly with an adjacent-slice total-variation (TV) ratio. For ground-truth volume $\mathbf{x}_0$ and reconstruction $\hat{\mathbf{x}}_0$, and for each axis $a \in \{\text{axial}, \text{coronal}, \text{sagittal}\}$, let $\text{TV}_a(V)$ be the mean absolute difference between adjacent slices of V along axis a . We define

$$\text{tv_ratio}_a = \frac{\text{TV}_a(\hat{\mathbf{x}}_0)}{\text{TV}_a(\mathbf{x}_0)}, \quad (4)$$

so that $\text{tv_ratio}_a \approx 1$ means the reconstruction varies from slice to slice along axis a by the same amount as the ground truth, while $\text{tv_ratio}_a > 1$ indicates spurious jitter beyond what the anatomy itself produces. The two off-axis ratios serve as a control for a method’s overall sharpness or blur (a uniformly blurry method scores below 1 on *all three* axes), which lets us isolate the axis-specific pathology as

$$\text{tv_excess} = \text{tv_ratio}_{\text{axial}} - \frac{1}{2}(\text{tv_ratio}_{\text{coronal}} + \text{tv_ratio}_{\text{sagittal}}). \quad (5)$$

$\text{tv_excess} > 0$ specifically flags the axial-direction jitter characteristic of slice-by-slice 2D processing, robust to a method’s general smoothness level. We report $\text{tv_ratio}_{\text{axial}}$ and tv_excess throughout Sec. 6.

For our high-resolution CBCT results on the walnut dataset, we compute PSNR and SSIM with the same clamping range as in the low-resolution setting using the full volume. However, when comparing with S-STAR Net [34], we compute PSNR and SSIM on the central crop of size $360 \times 360 \times 150$ to match their evaluation protocol.

6 Results.

6.1 Low-resolution CBCT results.

Table 1 presents a comprehensive quantitative comparison of our proposed methods against several state-of-the-art baselines across three CBCT datasets with volume size 256^3 : Dental, Spine, and Walnut. The results are reported in terms of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), with values presented as mean $\pm$ standard deviation.

Our baseline diffusion model (DPA) shows decent performance, but struggles to surpass some of the recent supervised CNN-based methods such as GAAL [19], as already noticed by the authors of the GAAL paper. However, when we introduce conditioning on the FDK reconstruction (CDPA), we observe a significant performance boost across all datasets, achieving state-of-the-art results and improving 3D consistency. Furthermore, when considering the mean of multiple posterior samples with $\mu(\text{CDPA})$, we achieve the best PSNR scores across all datasets, strongly suggesting that the posterior distribution is unimodal and that averaging

samples effectively reduces noise and artifacts. This mean posterior estimation also helps the baseline DPA method achieve competitive results, although it is still lagging behind CDPA.

Interestingly, our FDK-denoising U-Net model with slice conditioning also achieves competitive results, particularly in SSIM, highlighting the potential of simpler architectures when combined with data-consistency enforcement. It matches or surpasses all previous methods in both PSNR and SSIM, and we can notice that the use of fine-tuning (FDK-denoiser + FT) further improves PSNR results almost to the point of CDPA, this fine-tuning however, might decrease SSIM results, which suggests that it might be overfitting to the MSE loss and losing some of the perceptual quality of the reconstruction.

$\mu(\text{CDPA})$ is, by construction, an approximation of the posterior mean, i.e. the Bayes-optimal estimator, which explains why it wins on PSNR essentially by definition. Standard bias-variance reasoning predicts this should come at a cost in perceptual/structural quality: averaging several plausible textures is pixel-wise closer to the ground truth, but smoother than any single sample, which SSIM’s local structure term can penalize. Surprisingly, this expected trade-off largely fails to materialize: on Dental, $\mu(\text{CDPA})$ is also the best method on SSIM (0.921 vs. 0.901 for FDK-denoiser), and on Spine it is also the best method on SSIM (0.952 vs. 0.950 for FDK-denoiser), i.e., $\mu(\text{CDPA})$ improves on both fidelity and structural similarity simultaneously, rather than trading one for the other. On Walnut at 20 views, the most severely underdetermined setting in the paper (Sec. 4), $\mu(\text{CDPA})$ clearly trails on SSIM (0.813 vs. 0.888 for FDK-denoiser and 0.889 for GAAL), though there is still a clear increase in both PSNR and SSIM when averaging over posterior samples.

To further illustrate the advantage of our slice conditioning and the fine-tuning step, we conducted an ablation study discussed in Section 6.7 (although at higher resolution). The results confirm that both components significantly contribute to the overall performance, with slice conditioning providing substantial gains and fine-tuning further enhancing the results to nearly match our diffusion-based methods. Without these components, the U-Net architecture alone falls quite short of state-of-the-art performance.

6.2 Low-resolution qualitative results.

Figures 1, 2 and 3 present qualitative reconstruction results for the Walnut, Spine and Dental datasets, respectively, at a resolution of 256^3 voxels using 20 views. These figures show the visual quality of our proposed methods compared to the FDK baseline, GAAL [19] and gradient descent, minimizing data-consistency loss. We can see that both conditional diffusion and FDK-denoising methods significantly reduce artifacts and enhance structural details compared to FDK and gradient descent, leading to clearer and more accurate reconstructions. The improvements of our methods are particularly evident in regions with complex anatomical structures, where we effectively recover fine details that are often lost by other methods. We observe that hallucinations still exist in this high-sparsity regime because the information is simply not present in the available measurements. However, $\mu(\text{CDPA})$ ameliorates this effect by taking the average over a few samples, artifacts hallucinated by a single reconstruction are voted out, leaving a sort of “majority voting” solution to the inverse problem. Noticeably, this average reconstruction is sharp, opposite to the usual average reconstruction found in models such as variational autoencoders (VAEs). This suggests a clear unimodal posterior distribution where our mean reconstruction lies closer to the ground truth.

For the walnut data based on real measurements, we can see that the coronal and sagittal views (top and bottom rows in Figure 1) still exhibit some inter-slice inconsistencies, compared with the axial view where the diffusion process is carried out. We can also see that, overall, GAAL produces results with substantially less detail and sharpness. However, producing less detailed reconstructions also means that GAAL can avoid some inter-slice inconsistencies.

Compared to the FDK-denoiser, we can see how diffusion models always produce detailed realistic structures, even if they are not always correct, while FDK-denoising models sometimes

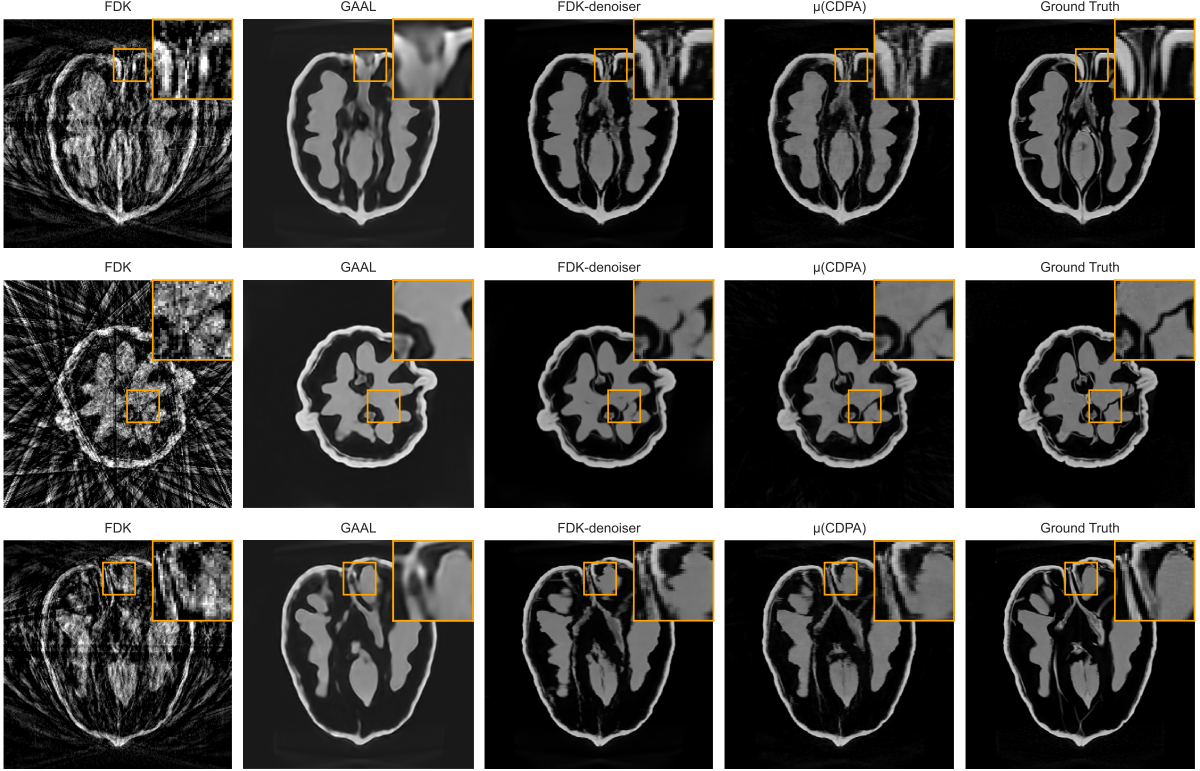


Figure 1: Slices (coronal, axial and sagittal from top to bottom) of the walnut reconstruction results with resolution 256^3 from the test dataset using 20 views uniformly spaced from the 1200 available views from the middle scan.

produce more unnatural structures, as can be seen in the walnut reconstructions in Figure 1.

6.3 High-resolution CBCT results vs number of views.

Figure 6 presents the reconstruction quality (PSNR and SSIM) of our methods on the high-resolution walnut dataset with 501^3 voxels using varying numbers of views from the high scan. For each sparsity level, we selected a near-uniform subset of views from the 1200 available views in the high scan. We compare our methods against the ground-truth reconstruction using 3600 views from all three scans (high/middle/low). With 20 views, our models already achieve significantly better PSNR and SSIM than the FDK baseline using all 1200 views from the high scan, and with 40 views they outperform the gradient descent reconstruction using all 1200 views from the high scan, highlighting the effectiveness of our methods in extremely sparse-view scenarios. CDPA with mean posterior averaging, $\mu(\text{CDPA})$, provides the best performance across all view counts, with a consistent improvement of approximately 1 dB of PSNR over standalone CDPA, demonstrating its ability to effectively leverage both learned priors and data consistency. Similarly, FDK-denoising with fine-tuning (+ FT) achieves competitive results, closely matching CDPA, especially at higher view counts.

6.4 High-resolution results at $N = 30$ views and inter-slice consistency.

For the specific case of 30 views, we fine-tuned our model with samples at this specific sparsity. We acknowledge that this fine-tuning step provides a significant advantage over the model trained for varying numbers of views, as shown in Figure 6, where we compare metrics for the whole generated volume.

We compare our results for $N = 30$ with S-STAR Net [34] in Table 2, showing that our

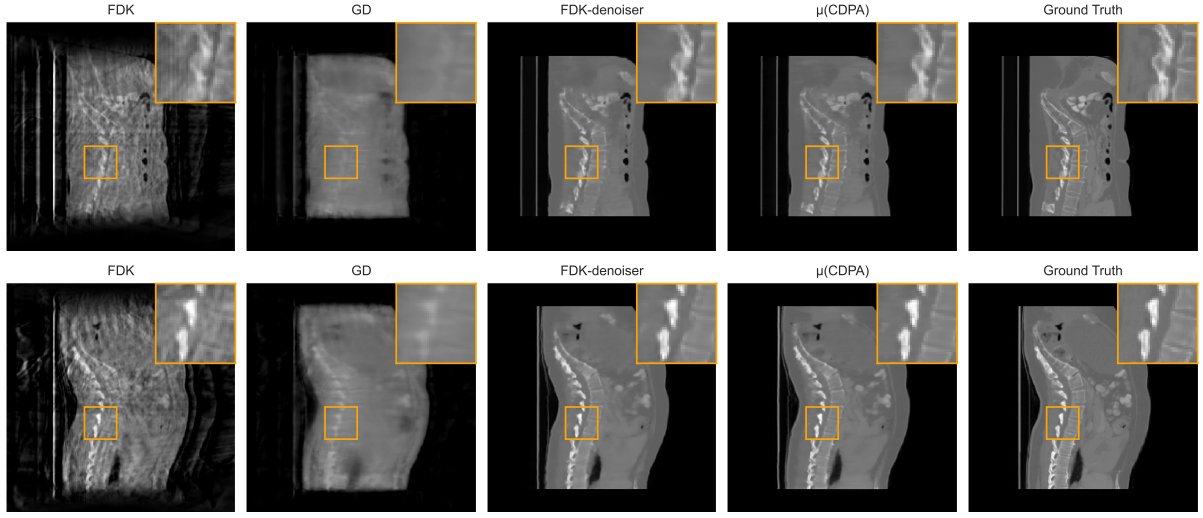


Figure 2: Coronal slices of the spine reconstruction results with resolution 256^3 from the test dataset using 20 views uniformly spaced from 0 to 360 degrees.

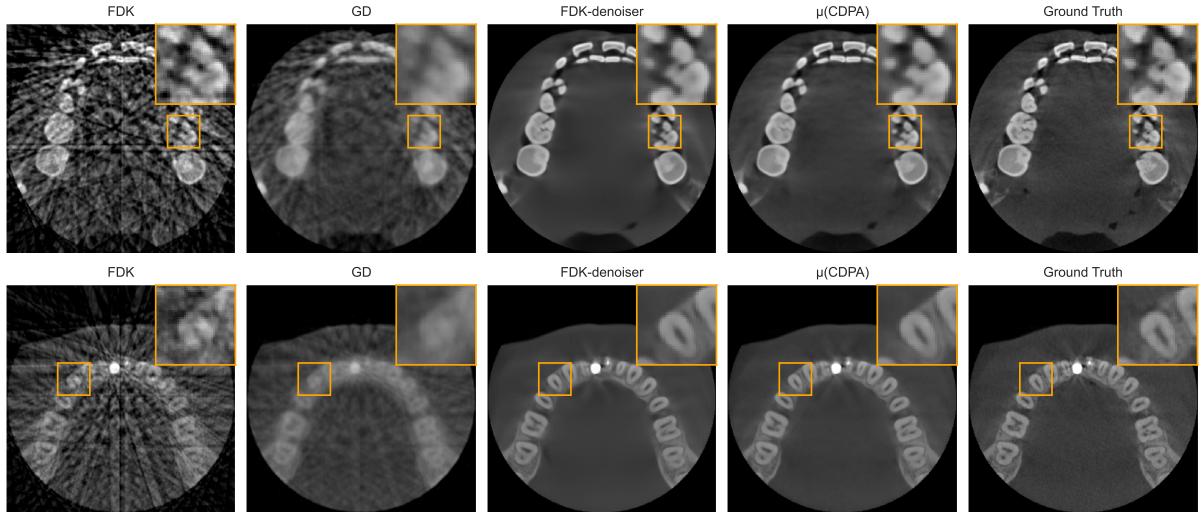


Figure 3: Axial slices of the dental reconstruction results with resolution 256^3 from the test dataset using 20 views uniformly spaced from 0 to 360 degrees.

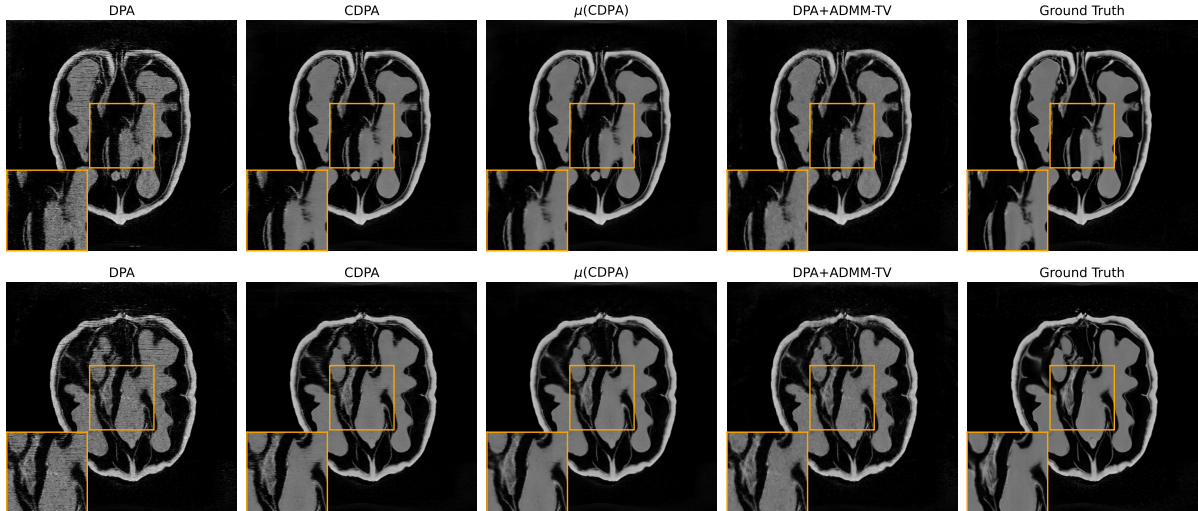


Figure 4: Slices (coronal and sagittal from top to bottom) of the walnut reconstruction results with resolution 501^3 from the test dataset using 30 views uniformly spaced from the 1200 available views from the middle scan.

method slightly exceeds S-STAR Net and outperforms other methods in both SSIM and PSNR, demonstrating that our 2D approach scales to the performance of state-of-the-art 3D approaches. Recall that these numbers are not directly comparable with other tables as they are computed only within a central crop; thus, while the SSIM values remain similar, the PSNR values differ from those of the entire volume, as almost no background is considered.

Table 2: Walnut high-resolution results for the center $360 \times 360 \times 150$ crops conditioned on 30 views. Results with a star are taken from [34].

Algorithm	SSIM $\uparrow$	PSNR $\uparrow$
U-former*	0.87 ± 0.01	24.94 ± 1.37
PTNet3D*	0.82 ± 0.01	22.09 ± 0.36
Pix2Pix3D*	0.85 ± 0.01	23.74 ± 0.56
S-STAR Net*	0.90 ± 0.01	27.17 ± 0.95
$\mu(\text{CDPA})$ (ours)	0.91 ± 0.01	27.35 ± 1.47

Table 2 follows S-STAR Net’s protocol of evaluating only a central $360 \times 360 \times 150$ crop, since that baseline cannot process the full volume. This means our defensible advantage there is not the small margin in the table itself, but that our 2D approach reaches this quality on the *entire* 501^3 volume with the same $N = 30$ -view fine-tuned model, which we report in Table 3 using the same PSNR, SSIM, LPIPS [33] and TV-ratio metrics defined in Sec. 5.2.

This also lets us test the inter-slice-consistency claim on real CBCT measurements at a view count where every method has enough information to produce a reasonable reconstruction. The conditioning benefit is clear, DPA’s axial TV excess lands at +1.00 and CDPA’s at +0.55, and $\mu(\text{DPA})$ ’s excess of +0.50 is cut to $\mu(\text{CDPA})$ ’s +0.14. PSNR and SSIM improve substantially for every method between 20 and 30 views, and the CDPA-over-DPA margin widens rather than narrows ($\mu(\text{CDPA})$ reaches 34.57 dB/0.906 SSIM vs. $\mu(\text{DPA})$ ’s 29.56 dB/0.747), consistent with conditioning extracting more benefit from FDK reconstructions that themselves improve with more views.

FDK-Conditioning vs. Explicit Regularization. An alternative to FDK-conditioning for enforcing inter-slice consistency is to impose an explicit hand-crafted regularizer directly on the

Table 3: Walnut, full 501³ volume, $N = 30$ views (fine-tuned model), $n = 5$ test volumes. tv_excess as defined in Eq. 5. DPA+ADMM-TV and CDPA+ADMM-TV apply DiffusionMBIR’s explicit ADMM+1D-TV guidance (Sec. 6.4) to our pretrained checkpoints; all rows are averaged over the same $n = 5$ test volumes.

Method	PSNR	SSIM	$\text{tv_ratio}_{\text{axial}}$	tv_excess	LPIPS
FDK	16.27±0.40	0.124±0.002	11.556	−1.125	0.809±0.004
DPA	27.63±0.59	0.727±0.021	3.265	+1.001	0.337±0.024
CDPA	32.57±0.55	0.879±0.008	1.509	+0.551	0.107±0.009
DPA+ADMM-TV	32.04±0.57	0.767±0.012	1.486	−0.269	0.227±0.012
CDPA+ADMM-TV	32.60±0.66	0.772±0.012	1.490	−0.273	0.228±0.012
FDK-denoiser	32.04±0.72	0.891±0.007	0.655	+0.022	0.166±0.009
$\mu(\text{DPA})$	29.56±0.53	0.747±0.019	2.634	+0.495	0.304±0.020
$\mu(\text{DPA+ADMM-TV})$	32.36±0.60	0.772±0.012	1.465	−0.252	0.220±0.012
$\mu(\text{CDPA+ADMM-TV})$	32.87±0.67	0.777±0.011	1.471	−0.255	0.220±0.011
$\mu(\text{CDPA})$	34.57±0.68	0.906±0.007	1.001	+0.144	0.092±0.006
FDK-denoiser + FT	31.11±0.58	0.763±0.020	2.160	+0.232	0.278±0.019

diffusion sampling process. DiffusionMBIR [5] augments a *pretrained, unconditional* 2D diffusion prior with an ADMM data-consistency step that solves $\min_{\mathbf{x}} \frac{1}{2} \|\mathcal{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \|D_z \mathbf{x}\|_1$ at every reverse-diffusion timestep, where D_z is the first-difference operator along the axial axis – a 1D total-variation penalty coupling adjacent slices, solved via alternating conjugate-gradient (CG) least-squares updates and soft-thresholding. We implement this guidance mechanism as a drop-in replacement for our own ReSample-style gradient-descent guidance (Sec. 3), which lets us apply it to our *existing, pretrained* DPA and CDPA checkpoints with no retraining – DPA+ADMM-TV is exactly DiffusionMBIR; CDPA+ADMM-TV additionally retains our FDK-conditioning, testing whether the two consistency mechanisms are complementary or redundant. Parameter λ and the ADMM/CG iteration budget have no principled default and must be tuned per acquisition, an expensive per-dataset hyperparameter search CDPA does not require.

Fidelity is close, but that is not the whole story. The DPA+ADMM-TV and CDPA+ADMM-TV rows of Table 3 show that CDPA remains the strongest single-sample method on both PSNR and, especially, SSIM, even against this properly-tuned explicit-regularization baseline: SSIM trails CDPA by 0.107–0.112, while PSNR is within 0.5 dB. On PSNR/SSIM alone, the two strategies for restoring 3D consistency look close to interchangeable. However, they are not: the following three paragraphs identify three independent, easily measurable axes on which explicit regularization loses, none of which a pixel-fidelity metric alone would reveal.

Cost 1: axial over-smoothing. Every other row of Table 3 has a *positive* tv_excess – the axial-specific jitter the metric was designed to detect (Sec. 6.4). Both ADMM-TV variants instead show a *negative* tv_excess (−0.269 and −0.273): the ADMM penalty does not merely remove the jitter, it over-corrects, smoothing the axial direction *more* than the anatomy’s own in-plane structure would predict. This is the opposite pathology from the one motivating CDPA in the first place, and it is invisible to PSNR/SSIM – exactly why a metric that measures inter-slice consistency directly, rather than inferring it from a fidelity score, is needed to catch it.

Cost 2: a collapsed posterior. Our results show that averaging posterior samples improves both fidelity and structural similarity for DPA/CDPA, which we attributed to a near-unimodal posterior: in Table 3, $\mu(\text{DPA})$ and $\mu(\text{CDPA})$ gain +1.93 dB and +2.00 dB PSNR respectively over their single-sample counterparts, with a substantial SSIM gain for CDPA (0.879 → 0.906). The $\mu(\text{DPA+ADMM-TV})$ and $\mu(\text{CDPA+ADMM-TV})$ rows test whether this benefit survives

once the explicit, deterministic ADMM constraint is added to the same sampling loop ($n = 10$ posterior samples, same $n = 5$ test volumes and hyperparameters as the single-sample ADMM-TV rows). For both variants the gain collapses to well under a sixth of DPA/CDPA’s own gain – $+0.32$ and $+0.27$ dB, with a correspondingly modest SSIM gain ($0.767 \rightarrow 0.772$ and $0.772 \rightarrow 0.777$) – regardless of conditioning. This is a genuine posterior-diversity effect rather than a metric artifact: because the CG-solved ADMM update is a deterministic function of the fixed measurement $\mathbf{y}$, it pulls every draw toward nearly the same fixed point regardless of the diffusion model’s own stochastic noise, leaving little independent sample-to-sample variation for Monte Carlo averaging to exploit. This collapse is not a conditioning artifact – it is essentially identical with and without FDK-conditioning – so it is a property of the explicit regularizer itself, not of its interaction with conditioning.

Cost 3: substantially worse perceptual quality. Table 3’s LPIPS column tells a third, independent story. Both ADMM-TV variants score 0.227 – 0.228 – more than double CDPA’s 0.107 and more than double $\mu(\text{CDPA})$ ’s 0.092 – despite reaching comparable PSNR and SSIM to CDPA. Enforcing 3D consistency through an explicit spatial penalty evidently costs perceptual realism that pixel-wise fidelity metrics cannot see, consistent with the over-smoothing already identified via `tv_excess` (Cost 1) and independent of the posterior-collapse argument (Cost 2), which concerns sample diversity rather than any single sample’s quality. Posterior averaging, which meaningfully reduces CDPA’s LPIPS ($0.107 \rightarrow 0.092$), barely moves it for either ADMM-TV variant ($0.227 \rightarrow 0.220$ and $0.228 \rightarrow 0.220$) – a third symptom of Cost 2’s collapsed diversity: with little independent variation left to average over, there is little perceptual benefit left to gain from averaging either.

6.5 High-resolution qualitative results.

Figure 5 presents qualitative reconstruction results for the high-resolution walnut dataset with 501^3 voxels using between 20 and 180 views from the high scan. The results show a progressive improvement in the quality of the reconstruction as the number of views increases. With only 20 views, the reconstructions exhibit still significant hallucinations, and there is not enough inter-slice consistency, although the quality of the reconstruction remains high. However, as the number of views increases, our method gets closer and closer to the ground truth, with almost no discernible difference at 100 views compared to the ground truth that uses 3600 views from all three scans (high, middle and low scans). Already with 40 views (or 30 in the fine-tuned model), inter-slice inconsistencies mostly disappear, which is supported by the largest jump in SSIM and PSNR in Figure 6 after going above 20 views.

6.6 Computational Efficiency.

In terms of speed, Figure 7 compares the running time of our different methods on the CBCT datasets with size 256^3 . Conditional diffusion requires fewer data-consistency steps than its unconditional counterpart due to the information already present in the FDK reconstruction. While these numbers may vary with the size of the GPU and how many slices are produced in one single batch, there is no denying that diffusion-based methods are significantly more computationally intensive than FDK-denoising approaches. Notably, FDK-denoising achieves performance comparable to diffusion-based models but with up to $18\times$ faster runtime, making it a practical choice for real-world applications. However, neither method has been optimized for speed, and deployment ready software is not the objective of this paper.

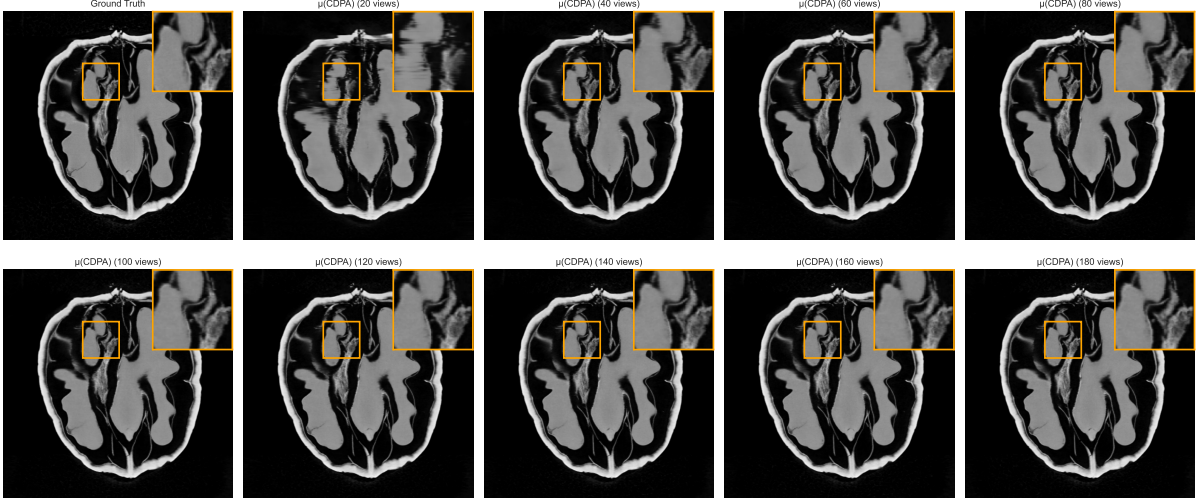


Figure 5: Axial slices of a walnut reconstruction using $\mu(\text{CDPA})$ at resolution 501^3 for the first walnut in the test dataset using progressively more views. Each view count is spaced uniformly from the 1200 available views from the high scan. The 3D consistency improves progressively as well; with only 20 views one can see inter-slice inconsistencies, but shortly after they become imperceptible.

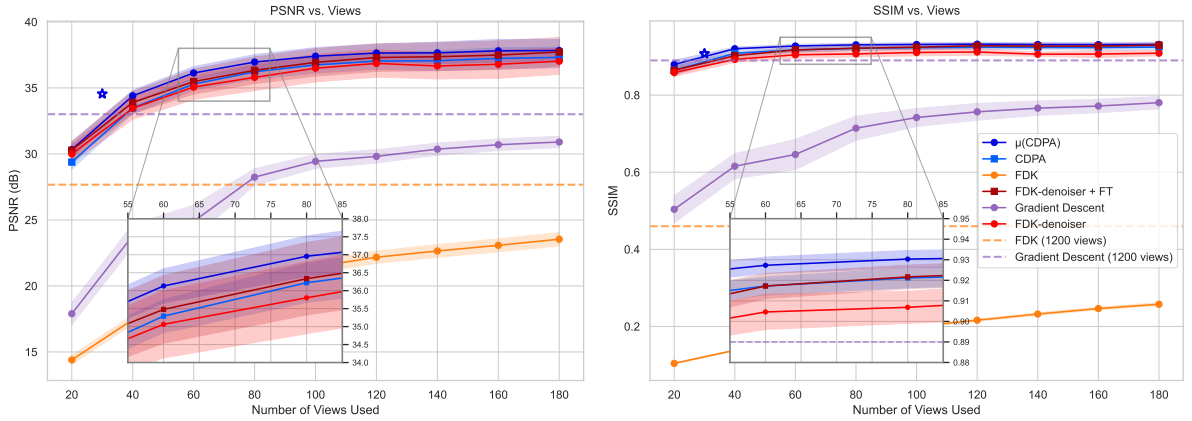


Figure 6: PSNR and SSIM vs number of views for the high-resolution walnut dataset at resolution 501^3 using the high-scan setting. The model was jointly trained for all sparsities. Dotted lines represent the baseline reconstruction using FDK and GD with all 1200 views from this high scan. All metrics are computed by comparing with the ground-truth reconstruction using 3600 views from all three high/middle/low scans. The star represents the result of the fine-tuned model for 30 views, showing the improvement that fine-tuning on a specific sparsity provides.

6.7 Ablation Studies.

Figure 7 (right) presents an ablation study of slice conditioning and data-consistency fine-tuning (FT) on the Walnut dataset at full resolution of 501^3 voxels. The results confirm that both components significantly contribute to the overall performance, with slice conditioning providing substantial gains and fine-tuning further enhancing the results to nearly match our diffusion-based methods. Interestingly, FT has a detrimental effect when working with only 20 views and without slice conditioning, but becomes beneficial when combined with slice conditioning at this high sparsity. This suggests that the fine-tuning step becomes beneficial when the original denoising network is already producing reasonable reconstructions, but is susceptible to overfitting in the ultra-sparse regime. Moreover, fine-tuning becomes increasingly beneficial in all settings as the number of views increases, matching the intuition that data-consistency becomes the dominant source of information as more measurements are available.

Without these components, the U-Net architecture alone falls quite short of state-of-the-art performance. The improvement carries over to different sparsities, though it is more dramatic at higher sparsities (fewer views). These two components are enough to make FDK-denoising competitive with diffusion-based models, and redefine state-of-the-art, while being significantly faster.

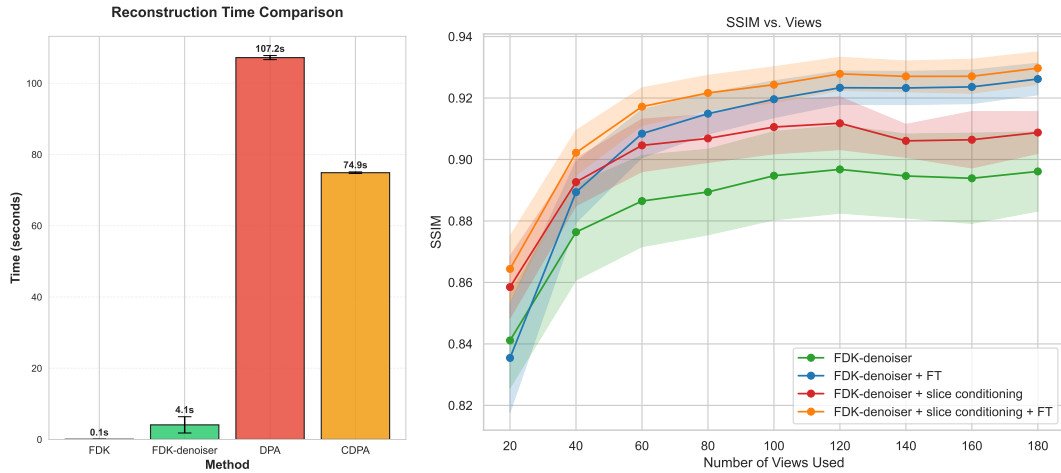


Figure 7: **Left:** Running time comparison of different methods on the CBCT datasets with size 256^3 . Conditional diffusion requires fewer data-consistency steps than its unconditional counterpart due to the information already provided by the conditioning FDK reconstruction. FDK-denoising with fine-tuning is significantly faster than diffusion-based methods while achieving comparable performance. **Right:** Ablation study of slice conditioning and data-consistency fine-tuning (FT) on the Walnut dataset at full resolution of 501^3 voxels. A modern U-Net architecture on its own is competitive but is not state-of-the-art. The addition of slice conditioning significantly improves performance, and data-consistency fine-tuning (FT) further boosts the results to nearly match our proposed diffusion-based methods.

7 Uncertainty Quantification.

Uncertainty quantification is an important aspect in medical imaging applications, where understanding the confidence of the reconstruction can inform clinical decision-making and guide further imaging or interventions. We explore the use of $\mu(\text{CDPA})$ to quantify uncertainty in the reconstructions, by computing multiple samples from the posterior distribution. More specifically, we sample several times from $p(\mathbf{x}|\mathbf{y})$ using $\mu(\text{CDPA})$ and consider the mean of these samples, as well as their pixel-wise standard deviation (STD). To assess the reliability of

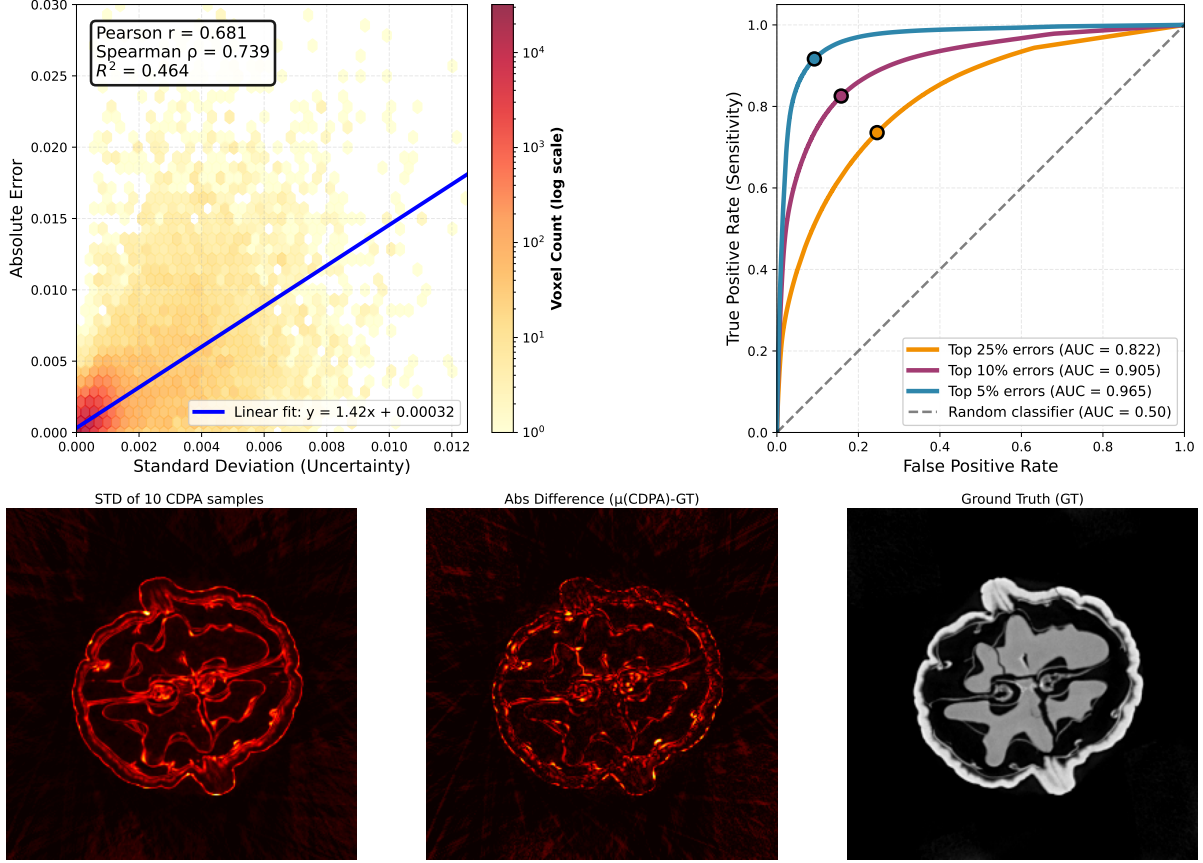


Figure 8: **(A) Top left:** Correlation between standard deviation (STD) of diffusion samples for the 256^3 walnut dataset and reconstruction error using 20 projections. The hexbin density plot shows strong positive correlation (Pearson $r = 0.680$, Spearman $\rho = 0.735$, $R^2 = 0.462$), with linear fit $\text{Error} \approx 1.42 \times \text{STD} + 0.0003$. STD explains 46.2% of error variance, demonstrating its utility as an uncertainty indicator. **(B) Top right:** Receiver Operating Characteristic (ROC) curves for identifying high-error regions using STD thresholds. The method achieves excellent discrimination with AUC values of 0.820, 0.902, and 0.961 for detecting the top 25%, 10%, and 5% highest-error voxels, respectively. Optimal operating points (circles) demonstrate practical trade-offs between sensitivity and specificity. At the optimal threshold for top 5% errors, the method achieves 91.3% sensitivity with only 9.2% false positive rate. **(C) Bottom:** Spatial comparison of a representative axial slice showing ground truth, uncertainty map (STD), and absolute error. High-uncertainty regions (bright areas in STD map) spatially correspond to high-error regions, validating STD as a reliable spatial predictor of reconstruction quality.

our uncertainty estimates, we analyzed the relationship between the STD of diffusion samples and reconstruction error against ground truth. Figure 8 presents a comprehensive uncertainty quantification analysis with a qualitative comparison of a representative axial slice, a correlation analysis between STD and absolute error, and a Receiver Operating Characteristic (ROC) analysis for identifying high-error regions using STD thresholds.

Correlation Analysis. The STD exhibits a positive correlation with the absolute reconstruction error (Pearson $r = 0.680$, Spearman $\rho = 0.735$, both $p < 10^{-10}$), explaining 46.2% of error variance ($R^2 = 0.462$) (Fig. 8A). The linear relationship ($\text{Error} \approx 1.42 \times \text{STD} + 0.0003$) indicates that regions with higher sampling variability consistently exhibit larger reconstruction errors. The higher Spearman correlation suggests a monotonic relationship that slightly deviates from perfect linearity, typical of heterogeneous structures.

Discriminative Power for Quality Control. To evaluate the practical utility for identifying problematic reconstructions, we performed ROC analysis treating STD (normalized) as a predictor probability score for high-error regions (Fig. 8B). The method achieves strong discrimination across different error thresholds: AUC = 0.820 for top 25% errors, AUC = 0.902 for top 10% errors, and AUC = 0.961 for top 5% errors. Notably, the highest AUC for the most severe errors indicates that STD is particularly reliable for identifying critical reconstruction failures.

Discussion. Our analysis shows promise of using the Diffusion posterior for uncertainty quantification. The results represented in Fig. 8 demonstrate clearly that the standard deviation of the Diffusion posterior is a strong predictor of the reconstruction error and, as such, can aid in the interpretation of the reconstruction. However, reliable uncertainty quantification for tomographic reconstruction remains a challenging topic. Directly relying on Diffusion posterior sampling does not alleviate the possibility of misspecification and hallucinations, and Diffusion posterior alignment does not sample the exact Bayesian posterior, c.f. [20]. Moreover, the Diffusion posterior samples do not come with formal asymptotic guarantees, such as, for example, Bootstrap based methods [23] or the Laplace approximation [1], which we therefore expect to perform better in the dense, noiseless data limit. Prior work has further investigated Bayesian models, which typically rely on approximations such as Markov-Chain Monte Carlo (MCMC), e.g. [8, 22]. Finally, recent work explores the use of sequential likelihood mixing to obtain formal confidence guarantees for neural network and diffusion-based reconstructions [10], however these are defined for a level set of the likelihood function, and not by the STD map alone. We leave further exploration of this topic for future work.

8 Conclusions.

In this work, we have presented two state-of-the-art reconstruction methods for sparse-view CBCT reconstruction: Conditional Diffusion Posterior Alignment (CDPA) and a new perspective on the FBP/FDK-denoiser models that ensures data consistency at inference time using fine-tuning.

CDPA leverages the power of diffusion models while incorporating conditioning on the FDK reconstruction to enhance performance and 3D consistency, and the guidance provided by the data-consistency through the diffusion steps. That is, we are effectively combining two techniques to sample from the posterior distribution and show that this combination is synergistic and leads to significant performance improvements. In addition, we noticed that averaging multiple posterior samples further boosts performance, suggesting that the posterior distribution is unimodal and that averaging helps reduce noise and artifacts. This is only the case if the

conditioning is strong enough, as is the case in several inverse problems. In addition, computing several samples from the posterior distribution allows us to quantify uncertainty in the reconstructions, which is an important aspect in medical imaging applications.

FDK-conditioning is, however, only one of two ways to restore 3D consistency to a slice-by-slice 2D diffusion prior that we study; the other imposes an explicit, hand-crafted regularizer directly on the sampling process, as DiffusionMBIR does with an ADMM+total-variation guidance step. We compared the two directly, applying the explicit-regularization mechanism to our own pretrained checkpoints with no retraining (Sec. 6.4), and found that the raw fidelity metrics alone are close to a tie: PSNR is comparable and SSIM trails CDPA by only 0.11 SSIM. Three further, independently measurable properties break that tie decisively in favor of FDK-conditioning: explicit regularization over-smooths the axial axis rather than merely removing its inconsistency (a negative rather than positive `tv_excess`), it reduces the sampler’s posterior variance to the point that averaging via the $\mu(\cdot)$ operator loses most of its value, and it produces reconstructions that a perceptual metric (LPIPS) scores as markedly worse. None of these three costs is visible in PSNR or SSIM alone, which is precisely why we measured them directly rather than relying on aggregate fidelity to adjudicate between the two strategies. We conclude that, whenever paired training data is available to learn the conditioning in the first place, it is preferable to hand-crafted spatial regularization even at matched fidelity.

The FDK-denoising approach, on the other hand, combines a U-Net architecture with slice conditioning and a data-consistency fine-tuning step to achieve competitive results with significantly reduced computational cost compared to diffusion-based methods. An ablation of these two components confirms their importance in achieving state-of-the-art performance, with slice conditioning providing substantial gains and fine-tuning further enhancing the results to nearly match our diffusion-based methods. Moreover, the computational efficiency of the FDK-denoising approach makes it a practical choice for real-world applications, especially in clinical settings where time constraints are critical.

Our methods scale to high-resolution CBCT reconstruction, where other approaches have failed to do so. The approach can be further scaled up by using latent diffusion techniques, which would allow to work with even higher resolution volumes, up to Gigavoxels resolution, though data-consistency updates require significant changes to work in latent space [26]. However, the limiting factor becomes the data-consistency gradient-descent step that at the moment consumes around 100GB of memory for $\approx 1500^3$ resolution volumes.

While 3D-based diffusion models should further improve performance, especially in terms of 3D consistency, they come at a significant computational cost that is at the moment prohibitive for high-resolution volumes. Moreover, a limited number of training volumes, as it is the case for the walnut dataset, makes it almost impossible to train a 3D-based diffusion model without overfitting, while 2D-based models can be trained with a much smaller number of volumes by leveraging the large number of slices available in each volume.

We believe the use of high-sparsity reconstructions (e.g., 20 views) might be only for benchmarking purposes, real clinical applications would likely use more views to ensure high-fidelity reconstructions, but our results show that as low as 180 views (15% of the original 1200 views from the high scan and only 5% from the 3600 total views used for the ground truth) are sufficient to achieve near-perfect reconstructions, significantly reducing radiation dose while maintaining high image quality.

Limitations. The method we presented is supervised and sits at the frontier of what can be achieved with high quality training data. We acknowledge, however, that in many applications such training data is not readily available. Moreover, generative reconstruction can hallucinate structure; the uncertainty map mitigates but does not eliminate this, and no reader study or task-based validation has been performed, so these methods are not yet validated for diagnostic use.

Acknowledgments.

The authors would like to thank AmirEhsan Khorashadizadeh, Raphael Winkler and all participants of the CHIP project of SDSC for useful discussions.

References.

- [1] Javier Antoran, Riccardo Barbano, Johannes Leuschner, José Miguel Hernández-Lobato, and Bangti Jin. Uncertainty estimation for computed tomography with a linearised deep image prior. *Transactions on Machine Learning Research*, 2023.
- [2] Luis Barba, Johannes Kirschner, Tomas Aidukas, Manuel Guizar-Sicairos, and Benjamín Béjar. Diffusion active learning: Towards data-driven experimental design in computed tomography. *arXiv preprint arXiv:2504.03491*, 2025.
- [3] Riccardo Barbano, Johannes Leuschner, Maximilian Schmidt, Alexander Denker, Andreas Hauptmann, Peter Maass, and Bangti Jin. An educated warm start for deep image prior-based micro ct reconstruction. *IEEE Trans. Computational Imaging*, 9:134–148, 2023.
- [4] Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *Proc. ICLR*, 2023.
- [5] 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 (CVPR)*, pages 22542–22551, 2023.
- [6] Yang Deng, Ce Wang, Yuan Hui, Qian Li, Jun Li, Shiwei Luo, Mengke Sun, Quan Quan, Shuxin Yang, You Hao, et al. Ctspine1k: A large-scale dataset for spinal vertebrae segmentation in computed tomography. *arXiv preprint arXiv:2105.14711*, 2021.
- [7] Henri Der Sarkissian, Felix Lucka, Maureen Van Eijnatten, Giulia Colacicco, Sophia Bethany Coban, and Kees Joost Batenburg. A cone-beam x-ray computed tomography data collection designed for machine learning. *Scientific data*, 6(1):215, 2019.
- [8] Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient bayesian computation by proximal markov chain monte carlo: when langevin meets moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.
- [9] Yu Fang, Lanzhuju Mei, Changjian Li, Yuan Liu, Wenping Wang, Zhiming Cui, and Dinggang Shen. Snaf: Sparse-view cbct reconstruction with neural attenuation fields. *arXiv preprint arXiv:2211.17048*, 2022.
- [10] Matteo Gätzner and Johannes Kirschner. Principled confidence estimation for deep computed tomography. *arXiv preprint arXiv:2602.05812*, 2026.
- [11] Yuhan Gu, Qing Wu, Zhechen Yuan, Xiangyu Zhang, Wenyan Su, Yuyao Zhang, and Xin Lou. An fpga accelerator for 3d cone-beam sparse-view computed tomography reconstruction. In *2024 IEEE 6th International Conference on AI Circuits and Systems (AICAS)*, pages 577–581. IEEE, 2024.
- [12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

- [13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [14] Kyong Hwan Jin, Michael T. McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Trans. Image Processing*, 26(9):4509–4522, 2017.
- [15] Bahjat Kawar, Gregory Vaksman, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Proc. NeurIPS*, 2022.
- [16] 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.
- [17] 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.
- [18] Jiaming Liu, Rushil Anirudh, Jayaraman J. Thiagarajan, Stewart He, K. Aditya Mohan, Ulugbek S. Kamilov, and Hyojin Kim. DOLCE: A model-based probabilistic diffusion framework for limited-angle CT reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6343–6353, 2023.
- [19] Zhentao Liu, Yu Fang, Changjian Li, Han Wu, Yuan Liu, Dinggang Shen, and Zhiming Cui. Geometry-aware attenuation learning for sparse-view CBCT reconstruction. 44(2):1083–1097.
- [20] Liam Moroy, Guillaume Bourmaud, Frédéric Champagnat, and Jean-François Giovannelli. On the posterior gap in plug & play diffusion methods for sparse-view computed tomography. *IEEE Journal of Selected Topics in Signal Processing*, 2026.
- [21] Willem Jan Palenstijn, Kees Joost Batenburg, and Jan Sijbers. Performance improvements for iterative electron tomography reconstruction using graphics processing units (gpus). *Journal of structural biology*, 176(2):250–253, 2011.
- [22] Marcelo Pereyra, Luis Vargas Mieses, and Konstantinos C Zygalakis. Accelerating proximal markov chain monte carlo by using an explicit stabilized method. *SIAM Journal on Imaging Sciences*, 13(2):905–935, 2020.
- [23] Marcelo Pereyra and Julián Tachella. Equivariant bootstrapping for uncertainty quantification in imaging inverse problems. In *International Conference on Artificial Intelligence and Statistics*, pages 4141–4149. PMLR, 2024.
- [24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.
- [25] 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.
- [26] Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. In *International Conference on Learning Representations*, 2024.

- [27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.
- [28] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proc. CVPR*, pages 9446–9454, 2018.
- [29] Wim Van Aarle, Willem Jan Palenstijn, Jeroen Cant, Eline Janssens, Folkert Bleichrodt, Andrei Dabravolski, Jan De Beenhouwer, K Joost Batenburg, and Jan Sijbers. Fast and flexible x-ray tomography using the astra toolbox. *Optics express*, 24(22):25129–25147, 2016.
- [30] Wim Van Aarle, Willem Jan Palenstijn, Jan De Beenhouwer, Thomas Altantzis, Sara Bals, K Joost Batenburg, and Jan Sijbers. The astra toolbox: A platform for advanced algorithm development in electron tomography. *Ultramicroscopy*, 157:35–47, 2015.
- [31] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021.
- [32] 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.
- [33] 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.
- [34] Shiang Zhang, Yibo Hu, Ziheng Deng, Yujie Wang, Jun Zhao, and Jianqi Sun. Sliding volume-based streak artifact reduction network (s-STAR net) for ultra-sparse-view computed tomography. *BMC Medical Imaging*, 25(1):364, 2025.
- [35] Xuzhe Zhang, Xinzi He, Jia Guo, Nabil Ettehad, Natalie Aw, David Semanek, Jonathan Posner, Andrew Laine, and Yun Wang. Ptnet3d: A 3d high-resolution longitudinal infant brain mri synthesizer based on transformers. *IEEE transactions on medical imaging*, 41(10):2925–2940, 2022.

A Supplementary Material.

A.1 Full Inter-Slice-Consistency Tables (256^3).

Tables 4 and 5 give the full per-method, per-dataset breakdown behind the compacted discussion in Sec. 6.4, covering all nine methods on all three 256^3 datasets rather than only the DPA/CDPA/ μ (CDPA) comparison highlighted there.

A.2 Network Architectures.

We utilize the networks and the optimizer from the diffusers library from Hugging Face <https://huggingface.co/docs/diffusers/en/index>. The diffusion model is based on a U-Net architecture with cross-attention layers, similar to the one used in Stable Diffusion. The network takes as input a noisy image and the time step, and outputs the predicted noise. For the conditional diffusion model (CDPA), we concatenate the FDK reconstruction to the noisy image along the channel dimension, allowing the network to leverage the information from the FDK reconstruction during denoising. The model has cross-attention layers that allow us to attend to the diffusion step, the slice index, and in the case of high-resolution walnut experiments, the number of views.

Table 4: Adjacent-slice total-variation ratio along the axial axis, $\text{tv_ratio}_{\text{axial}}$ (Eq. 4), and its axial excess over the in-plane average, tv_excess (Eq. 5), at $256^3/20$ views. Values near 0 for tv_excess indicate no axis-specific inconsistency; larger positive values indicate spurious inter-slice jitter specific to the axial direction. Same n per dataset as Table 1.

Method	Dental		Spine		Walnut	
	$\text{tv_ratio}_{\text{axial}}$	tv_excess	$\text{tv_ratio}_{\text{axial}}$	tv_excess	$\text{tv_ratio}_{\text{axial}}$	tv_excess
FDK	1.776	-0.482	2.838	-0.456	5.479	-0.781
GD_{zero}	0.804	+0.058	1.132	+0.192	2.357	+0.273
GD_{FDK}	1.045	-0.090	1.995	+0.377	2.715	+0.275
DPA	1.952	+1.061	2.367	+1.369	2.560	+0.490
CDPA	1.232	+0.439	1.429	+0.628	1.510	+0.355
FDK-denoiser	0.645	-0.003	0.750	-0.037	0.751	-0.027
$\mu(\text{DPA})$	0.990	+0.268	1.206	+0.352	2.321	+0.380
$\mu(\text{CDPA})$	0.788	+0.112	0.855	+0.106	1.209	+0.113
FDK-denoiser + FT	0.823	+0.106	0.757	-0.025	1.372	+0.086

Table 5: Per-axis SSIM: axial value and axial-vs-off-axis gap ($\text{SSIM}_{\text{axial}} - \frac{1}{2}(\text{SSIM}_{\text{coronal}} + \text{SSIM}_{\text{sagittal}})$), at $256^3/20$ views. A gap near 0 indicates the axial plane, in which our 2D model operates, is scored no differently than the two off-axis planes; a positive gap is the SSIM-side signature of inter-slice inconsistency.

Method	Dental		Spine		Walnut	
	$\text{SSIM}_{\text{axial}}$	gap	$\text{SSIM}_{\text{axial}}$	gap	$\text{SSIM}_{\text{axial}}$	gap
FDK	0.4145	-0.0234	0.3594	-0.0323	0.1976	+0.0052
GD_{zero}	0.7748	-0.0158	0.8693	+0.0020	0.5847	-0.0083
GD_{FDK}	0.7339	-0.0172	0.8243	+0.0054	0.5775	-0.0069
DPA	0.8580	+0.0406	0.9082	+0.0237	0.6499	+0.0026
CDPA	0.8942	+0.0063	0.9434	+0.0059	0.7987	+0.0064
FDK-denoiser	0.9008	-0.0001	0.9510	+0.0016	0.8861	+0.0000
$\mu(\text{DPA})$	0.9093	+0.0044	0.9451	+0.0047	0.6669	+0.0009
$\mu(\text{CDPA})$	0.9202	-0.0018	0.9527	+0.0012	0.8135	+0.0009
FDK-denoiser + FT	0.8841	+0.0027	0.9434	-0.0001	0.7409	+0.0011

For the FDK-denoising U-Net model, we use the same U-Net architecture as in the diffusion model, but without the time conditioning. The network takes as input the FDK reconstruction and outputs the denoised image. We also concatenate the slice index to the input along the channel dimension to provide spatial awareness to the network, and use cross-attention on the slice index for conditioning.

We use simple data augmentations during training, including random flips and rotations, to improve the generalization of the models.

We train with the AdamW optimizer with a learning rate of 10^{-4} for all models and 1000 diffusion steps for training and 50 for inference. For the guidance, we used 5 epochs with a batch size of $\min\{30, \text{num views}\}$ and a guidance learning rate of 5×10^{-4} . We used an A100 to train our models, and the training time for the diffusion models was around 24 hours, while for the FDK-denoising U-Net model it was around 6 hours. The inference time is reported in Figure 7 using the same A100 GPU.