

Towards reconstructing experimental sparse-view X-ray CT data with diffusion models

Nelas J. Thomsen^{*†}, Xinyuan Wang[†], Felix Lucka[†], Ezgi Demircan-Tureyen[†]

^{*}Martin-Luther-University Halle-Wittenberg, Institute of Physics, Halle, Germany

[†]Centrum Wiskunde & Informatica, Computational Imaging Group, Amsterdam, The Netherlands

Abstract—Diffusion-based image generators are promising priors for ill-posed inverse problems like sparse-view X-ray Computed Tomography (CT). As most studies consider synthetic data, it is not clear whether training data mismatch (“domain shift”) or forward model mismatch complicate their successful application to experimental data. We measured CT data from a physical phantom resembling the synthetic Shepp-Logan phantom and trained diffusion priors on synthetic image data sets with different degrees of domain shift towards it. Then, we employed the priors in a Decomposed Diffusion Sampling scheme on sparse-view CT data sets with increasing difficulty leading to the experimental data. Our results reveal that domain shift plays a nuanced role: while severe mismatch causes model collapse and hallucinations, diverse priors match or exceed well-matched but narrow priors. Forward model mismatch pulls the image samples away from the prior manifold, which causes artifacts but can be mitigated with annealed likelihood weight schedules that also increase computational efficiency. Overall, we demonstrate that performance gains do not immediately translate from synthetic to experimental data, and future development must validate against real-world benchmarks.

Index Terms—X-ray Computed Tomography, computational imaging, diffusion priors, image generation.

I. INTRODUCTION

Motivated by the need to reduce radiation and scan times, sparse-view X-ray CT presents an ill-posed inverse problem historically addressed by model-based [1], [2], and learned [3]–[5] methods. More recently, diffusion-based image generators are increasingly used as plug-and-play solvers [6]–[10]. These methods hinge on balancing two competing forces: the prior (what the diffusion model says a typical image should look like) and the likelihood (what the CT measurement data requires the image to be). The interplay between these two components is critical, as excessive reliance on the prior can lead to loss of detail or hallucinations, while overemphasizing data fidelity may amplify artifacts or fail to resolve ambiguities arising from the ill-posedness of the inverse problem.

The balance becomes critical for experimental data, where the gap between model and reality widens. The likelihood is challenged by the mismatch between the simplified forward model and the physics of real-world acquisition. While simulated measurements follow exact mathematical projectors, experimental scans include geometric inaccuracies, beam hardening, scatter, and complex noise textures that the standard likelihood term often cannot capture. As a result, the solver

may either amplify artifacts by fitting the wrong data model or rely too heavily on the prior to mask the mismatch.

The challenge in the prior, on the other hand, stems from the distribution shift between the training data and the experimental target. Standard diffusion models are often trained on idealized synthetic datasets that lack the geometric and textural variation of real objects. Priors trained on such narrow distributions can become brittle on experimental data, failing to reconstruct features that fall outside the training set. Conversely, priors specialized to experimental geometries may generalize poorly to unseen variations. Robust experimental reconstruction therefore requires not only handling forward model mismatch but also designing priors that bridge the gap between clean synthetic and imperfect experimental domains.

In this work, we examine how the performance of diffusion-based solvers for sparse-view CT translates to real experimental settings. To enable a controlled comparison, we designed a physical phantom that resembles the standard Shepp-Logan phantom, isolating specific sources of error. Our contributions are threefold: First, we analyze the impact of training distribution and demonstrate that while rigid geometric mismatch leads to drastic failure, a diverse prior may offer superior robustness against geometric challenges inherent to real-world objects. Second, we investigate forward model mismatch, showing that it manifests as not only severe streak artifacts, but also as structural misalignments, a specific form of hallucination where the solver distorts features to fit physical inconsistencies. Finally, we address practical inference strategies, demonstrating that an annealed likelihood weight schedule can reconcile the diffusion prior with imperfect experimental measurements even with significantly fewer sampling steps.

II. RELATED WORK

A. Preliminaries on Diffusion Models

Diffusion models [11] generate images by learning to invert a predefined noising process. In denoising diffusion probabilistic models (DDPMs) [11], this inversion is formulated as a Markovian reverse-time process, where each denoising step is modeled as a Gaussian distribution parameterized by a neural network ϵ_θ . The network is trained to predict the noise realization at a given state, which effectively estimates the mean of the reverse conditional distribution. Deterministic sampling schemes like DDIM [12] leverage this to define a non-Markovian trajectory, allowing for the direct estimation of the clean image $\hat{\mathbf{x}}_t$ from a noisy observation $\mathbf{x}_t$ at any timestep.

This estimate, often referred to as the Tweedie denoised estimate [13], is given by $\hat{\mathbf{x}}_t := \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_{\theta^*}^{(t)}(\mathbf{x}_t) \right)$, where $\bar{\alpha}_t$ is the noise schedule parameter and ϵ_{θ} is the predicted noise.

B. Diffusion-based Priors for Tomography

Assuming the tomographic measurement model $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{n}$, the goal is to reconstruct the unknown image $\mathbf{x}_0 \in \mathbb{R}^n$ from projection data $\mathbf{y} \in \mathbb{R}^m$, where $\mathbf{n} \in \mathbb{R}^m$ models measurement noise. Here $\mathbf{A} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ denotes the known tomography operator such as a discretized Radon transform mapping an image to line-integral measurements. By Bayes' Rule, the posterior score (i.e., the gradient of the log-posterior) decomposes as $\nabla_{\mathbf{x}_t} \log p(\mathbf{x} | \mathbf{y}) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}) + \nabla_{\mathbf{x}_t} \log p(\mathbf{y} | \mathbf{x})$ with the score $\nabla_{\mathbf{x}_t} \log p(\mathbf{x})$ being approximated by a pretrained model $\epsilon_{\theta}(\mathbf{x}_t, t)$. Thus, posterior sampling can be seen as combining a learned ‘‘prior direction’’ $\epsilon_{\theta}(\mathbf{x}_t, t)$ with a physics-based ‘‘data direction’’ $\mathbf{A}^\top(\mathbf{y} - \mathbf{A}\mathbf{x})$, assuming $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_y^2 \mathbf{I})$. Sampling approaches vary in the way they combine these two directions. Plug-and-play (PnP) approaches operate by injecting a likelihood-guidance term into each reverse diffusion step [6]–[10]. Given the data consistency loss $\ell(\mathbf{x}) = \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2$, the widely used diffusion posterior sampler (DPS) [8] updates the iterate at each reverse diffusion step according to:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} (\hat{\mathbf{x}}_t - \gamma_t \nabla_{\mathbf{x}_t} \ell(\hat{\mathbf{x}}_t)) + \tilde{\mathbf{w}}_t(\eta) \quad (1)$$

where γ_t denotes the data-consistency step size and $\tilde{\mathbf{w}}_t(\eta)$ is the overall noise term controlled by $\eta \in [0, 1]$, interpolating between fully deterministic and fully stochastic ancestral sampling. Decomposed Diffusion Sampler (DDS) [10] replaces the computationally expensive Jacobian term in Eq. (1) with a gradient-based surrogate, converting a step on the noisy generative manifold into a projected step onto the clean image manifold $\mathcal{M}$, i.e., $\hat{\mathbf{x}}_t - \gamma_t \nabla_{\mathbf{x}_t} \ell(\hat{\mathbf{x}}_t) = \mathcal{P}_{\mathcal{M}}(\hat{\mathbf{x}}_t - \zeta_t \nabla_{\hat{\mathbf{x}}_t} \ell(\hat{\mathbf{x}}_t))$. This projection is then implemented via an M -step conjugate gradient (CG), i.e., $\text{CG}(\mathbf{A}^\top \mathbf{A}, \mathbf{A}^\top \mathbf{y}, \hat{\mathbf{x}}_t, M)$.

III. METHODOLOGY

To realize an experimental scan of an object resembling the canonical Shepp-Logan (SL) phantom, we designed a physical phantom by laser-cutting ellipses into a 6mm-thick transparent polymethyl methacrylate (PMMA) plate. Two ellipses were left air-filled, while the others were filled with deep-pour transparent epoxy resin (see supplementary Fig. S1 for a photograph). Scans were acquired on a custom-built laboratory X-ray CT scanner in the FleX-ray Lab at CWI in Amsterdam. We collected 901 equally spaced projections over a full rotation in cone-beam geometry, with a source-to-object distance of 234.92 mm and a source-to-detector distance of 400 mm. The X-ray tube operated at 70 kVp and 600 μA with a 30 ms exposure. Projections were recorded with an effective detector pixel size of 0.2992 mm and 478 detector channels per view. Only the central detector row was used, yielding a fan-beam-equivalent sinogram of size 901 $\times$ 478 (views $\times$ channels). Dark-field and flat-field references were acquired for correction.

TABLE I: Parameters Used for Tailoring Training Sets

Dataset	Center	σ_A	$\sigma_{a,b}$	σ_{x_0, y_0}	σ_ϕ	$p_{\text{add, drop}}$
$\mathcal{X}_{\text{exp}}$	E_{exp}	0.03	0.02	0.01	10 $^\circ$	0.01
$\mathcal{X}_{\text{std}}$	E_{std}	0.03	0.02	0.01	10 $^\circ$	0.01
$\mathcal{X}_{\text{mix}}$	E_{mix}^a	0.03	0.03	0.02	45 $^\circ$	0.03

$$^a E_{\text{mix}} = \pi E_{\text{exp}} + (1 - \pi) E_{\text{std}}$$

A. Training Data

Let a phantom be defined as a set of ellipses $E = \{e_i\}_{i=1}^m$, where each ellipse $e_i := (A_i, a_i, b_i, x_{0,i}, y_{0,i}, \phi_i)$ encodes intensity, semi-axes, center coordinates, and orientation. For each generated phantom, we construct a sampled set E by perturbing a mean set $E_\mu = \{e_{i,\mu}\}_{i=1}^m$. Specifically, for each ellipse i and each parameter $p \in \{A, a, b, x_0, y_0, \phi\}$, we sample $p_i = p_{i,\mu} + s_p \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma_p^2)$ where s_p is the corresponding scale, i.e.,

$$s_p = \begin{cases} |p_{i,\mu}|, & p \in \{A, a, b\}, \\ \text{image_width, image_height}, & p = x_0, p = y_0, \\ 1, & p = \phi. \end{cases}$$

and σ_p is listed in Table I. Additionally, each ellipse is randomly dropped, or a new random ellipse is added, with probability $p_{\text{add, drop}}$, also listed in the table. This way, we have created three training sets, each of which has 10K images:

- *Standard Shepp-Logan dataset* ($\mathcal{X}_{\text{std}}$): For each phantom, we set $E_\mu := E_{\text{std}}$, where E_{std} is the set of ellipses from the standard Shepp-Logan test image, and sample E by perturbing all ellipses as above.
- *Experimental Shepp-Logan dataset* ($\mathcal{X}_{\text{exp}}$): For each phantom, we set $E_\mu := E_{\text{exp}}$, where E_{exp} is the set of ellipses from our laser-cut experimental SL, and sample E by perturbing all ellipses as above.
- *Mixed Shepp-Logan dataset* ($\mathcal{X}_{\text{mix}}$): For each phantom, we first sample $Z \sim \text{Bernoulli}(\pi)$. If $Z = 1$, we set $E_\mu := E_{\text{exp}}$; otherwise $E_\mu := E_{\text{std}}$. We then sample E by perturbing all ellipses as above. Here $\pi \in [0, 1]$ controls the fraction of experimental- versus SL-centered samples in the mixed dataset.

We refer the reader to supplementary Fig. S2 for visual examples of the generated phantoms from these datasets.

B. Generative Priors

Using these datasets, we train three corresponding priors. The *Standard SL prior* ($\mathbf{f}_{\text{std}}$) models the narrow distribution of ideal canonical phantoms. The *Experimental SL prior* ($\mathbf{f}_{\text{exp}}$) is explicitly tailored to the geometric design of our physical laser-cut phantom. Finally, the *Mixed SL prior* ($\mathbf{f}_{\text{mix}}$) provides a broader distribution by encompassing both standard and experimental variations.

C. Reconstruction Pipeline

Following Section II-B, we use DDS for reverse diffusion sampling from random noise with a pre-trained model (guided-diffusion codebase¹), guided by measurement-based CG steps.

¹<https://github.com/openai/guided-diffusion>

To match the real setup, the forward operator $\mathbf{A}$ models a fan-beam geometry with equally spaced sparse projections over $[0^\circ, 180^\circ]$. Unless stated otherwise, we reconstruct 128×128 images and use a cosine noise schedule [14] to preserve structure across timesteps at low resolution. In Sections IV-A and IV-B, we fix the number of model evaluations (NFE) to 1000, $M = 5$, $\eta = 0.85$, $\sigma_y = 10^{-7}$, and $\gamma_t = 1; \forall t$.

D. Test domains

We evaluate our models across four progressively challenging domains. The *Standard simulation* ($\mathbf{y}_{\text{sim}(\text{std})}$) is synthesized directly from the canonical SL image. The *CAD simulation* ($\mathbf{y}_{\text{sim}(\text{cad})}$) is generated from the laser-cut design parameters E_{exp} , mimicking standard piecewise geometry but with reduced contrast. The *Recon. simulation* ($\mathbf{y}_{\text{sim}(\text{recon})}$) introduces realistic textures by forward-projecting from a full-view reconstruction of the physical phantom. Finally, the *Experimental measurement* ($\mathbf{y}_{\text{exp}}$) is acquired directly from the physical phantom setup.

IV. EXPERIMENTAL RESULTS

A. Impact of Training Distribution

Fig. 1 quantifies reconstruction performance across a representative simulation-to-experimental spectrum. In the idealized $\mathbf{y}_{\text{sim}(\text{std})}$ domain (left), the domain-specific $\mathbf{f}_{\text{std}}$ achieves the highest PSNR, while $\mathbf{f}_{\text{mix}}$ shows a moderate generalization penalty and $\mathbf{f}_{\text{exp}}$ drops to baseline due to geometric mismatch between the standard SL and experimental laser-cut phantoms. In the second panel, where the test domain ($\mathbf{y}_{\text{sim}(\text{cad})}$) introduces a geometric shift toward the experimental phantom, $\mathbf{f}_{\text{mix}}$ unexpectedly exceeds the domain-specific $\mathbf{f}_{\text{exp}}$ by about 5 dB for more than five projections, possibly because its exposure to high-contrast standard phantoms provides stronger edge cues that transfer to boundary preservation despite the contrast difference. In the final two panels, the reference shifts from idealized phantoms to full-view reconstructions of experimental data, linking evaluation more directly to achievable experimental quality. These panels differ only in the forward model: $\mathbf{y}_{\text{sim}(\text{recon})}$ uses simulated projections from the reconstructed volume, whereas $\mathbf{y}_{\text{exp}}$ uses real measurements. In both cases, the PSNR gap between $\mathbf{f}_{\text{mix}}$ and the domain-specific $\mathbf{f}_{\text{exp}}$ largely closes across projection angles. However, the fully experimental case ($\mathbf{y}_{\text{exp}}$) shows consistently lower absolute PSNR than ($\mathbf{y}_{\text{sim}(\text{recon})}$), isolating the additional impact of non-ideal forward-model effects.

Fig. 2 shows reconstructions from 12 sparse projections across four domains. Baseline reconstructions with Conjugate Gradient Least Squares (CGLS) confirm that the measurements alone are insufficient. While diffusion priors generally recover high-fidelity structures, large distribution shifts (such as applying $\mathbf{f}_{\text{std}}$ to the laser-cut phantom or $\mathbf{f}_{\text{exp}}$ to the standard SL) lead to catastrophic failure. In contrast, the mixed prior $\mathbf{f}_{\text{mix}}$ generalizes robustly, reconstructing all domains with performance comparable to specialized priors on their native distributions. The narrower priors, $\mathbf{f}_{\text{std}}$ and $\mathbf{f}_{\text{exp}}$, can still leave residual streaks even on in-distribution

samples, suggesting that the solver struggles to balance prior adherence with data consistency. The broader support of $\mathbf{f}_{\text{mix}}$ appears to suppress these streaks, indicating that a looser prior may better accommodate the likelihood term and help the diffusion process find natural solutions without overfitting high-frequency inconsistencies.

B. Impact of Forward Model Mismatch

As noted above, the performance gap between $\mathbf{y}_{\text{sim}(\text{recon})}$ and $\mathbf{y}_{\text{exp}}$ is attributable to forward model mismatch. Fig. 3 illustrates this through line profiles from row 25 of the laser-cut phantom, which contains three fine holes representing its highest-frequency details. In the $\mathbf{y}_{\text{sim}(\text{cad})}$ domain, especially with $\mathbf{f}_{\text{mix}}$, the profiles closely follow the ground truth. In contrast, experimental data introduce a systematic offset. In the fully experimental domain, $\mathbf{y}_{\text{exp}}$, the mismatch causes the likelihood term to dominate the sampling process and pull the trajectory away from the clean prior manifold. This increases the risk of hallucinations, appearing here as a spatial shift of the reconstructed profile dips relative to the ground truth. The misalignment suggests that, to satisfy data consistency under a mismatched physics model, the solver not only propagates artifacts but also distorts features.

Since discretization errors typically decrease at finer grids, we tested whether higher reconstruction resolution could reduce the performance gap. Fig. 4 shows results at 128, 256, and 512 resolution. Although the gap between experimental ($\mathbf{y}_{\text{exp}}$) and simulated ($\mathbf{y}_{\text{sim}(\text{recon})}$) reconstructions narrows as resolution increases, a systematic deficit remains even at 512×512 (See supplementary Fig. S3 for reconstruction examples and corresponding error maps.) This persistent gap confirms that physical model mismatch, such as spectral inconsistency and geometric calibration error, remains a major source of experimental CT error.

C. Impact of Inference Design

In DDS, the trade-off between the generative prior and measured data is controlled by the number of CG iterations and the likelihood weight γ , introduced in [10] to modulate the measurement constraint. This balance is enforced in the proximal update by minimizing

$$L(x) = \frac{\gamma}{2} \ell(\mathbf{x}) + \frac{1}{2} \|\mathbf{x} - \hat{\mathbf{x}}_t\|_2^2, \quad (2)$$

which yields

$$\text{CG}(\gamma \mathbf{A}^\top \mathbf{A} + \mathbf{I}, \hat{\mathbf{x}}_t + \gamma \mathbf{A}^\top \mathbf{y}, \hat{\mathbf{x}}_t, M).$$

Here, γ scales the likelihood relative to the proximity term imposed by the diffusion prior.

In experimental settings, where forward model mismatch and measurement noise are unavoidable, a static γ creates a difficult sim2real trade-off. Table II shows that settings that work well in simulation can fail catastrophically on real data (see Const $\gamma = 5$, NFE= 1000). Because simulated data exactly matches the assumed forward operator, a high constant measurement weight yields high fidelity (26.82 dB

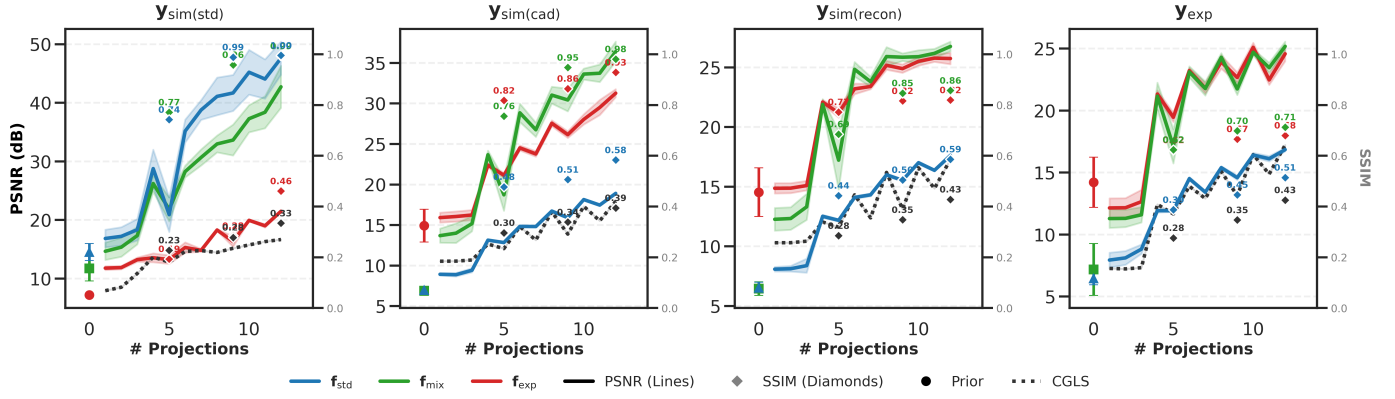


Fig. 1: PSNR (dB) as a function of the number of projections across four different test domains for the reconstructions obtained using CGLS and DDS with three different diffusion priors. Solid lines show PSNR, while diamond markers indicate sampled SSIM values at selected projection counts (5, 9, 12). Shaded areas denote ± 1 standard deviation across 10 random seeds.

TABLE II: Quantitative comparison (PSNR / SSIM) of constant versus linear decay likelihood weight scheduling across different NFEs and view counts at a 256×256 resolution, for both $y_{\text{sim(recon)}}$ and y_{exp} . Subscripts indicate standard deviations across nine runs. Bold values indicate the best mean results within each column for a given NFE group. Const: Constant schedule; Linear: Linear Decay schedule (γ_m denotes γ_{max}). See Supplementary Figs. S4 and S5 for visual samples from the diffusion trajectory, and Supplementary Table T1 for an additional row with further reduced sampling steps (NFE=10).

NFE	Schedule	12 views		24 views	
		$y_{\text{sim(recon)}}$	y_{exp}	$y_{\text{sim(recon)}}$	y_{exp}
1000	Const ($\gamma = 0.5$)	24.35 _{0.29} / 0.831 _{0.001}	24.07 _{0.21} / 0.774 _{0.002}	25.94 _{0.30} / 0.851 _{0.001}	24.91 _{0.45} / 0.767 _{0.002}
	Const ($\gamma = 5$)	25.53 _{0.14} / 0.826 _{0.001}	22.34 _{0.72} / 0.739 _{0.002}	26.82 _{0.29} / 0.838 _{0.001}	14.07 _{0.36} / 0.728 _{0.002}
	Linear ($\gamma_m = 5$)	23.60 _{0.13} / 0.829 _{0.001}	23.52 _{0.14} / 0.827 _{0.001}	24.14 _{0.09} / 0.853 _{0.000}	24.09 _{0.10} / 0.847 _{0.001}
	Linear ($\gamma_m = 50$)	24.60 _{0.13} / 0.852 _{0.001}	24.84 _{0.18} / 0.803 _{0.001}	25.51 _{0.34} / 0.871 _{0.001}	25.23 _{0.34} / 0.821 _{0.000}
	Linear ($\gamma_m = 150$)	25.03 _{0.17} / 0.849 _{0.001}	25.25 _{0.15} / 0.781 _{0.001}	26.29 _{0.24} / 0.867 _{0.001}	25.69 _{0.18} / 0.781 _{0.001}
100	Const ($\gamma = 0.5$)	23.29 _{0.59} / 0.826 _{0.001}	22.86 _{0.38} / 0.791 _{0.002}	24.12 _{0.63} / 0.837 _{0.001}	23.68 _{0.46} / 0.785 _{0.002}
	Const ($\gamma = 5$)	24.62 _{0.44} / 0.805 _{0.002}	24.05 _{0.41} / 0.728 _{0.001}	25.58 _{0.37} / 0.805 _{0.002}	24.43 _{0.50} / 0.694 _{0.003}
	Linear ($\gamma_m = 5$)	23.95 _{0.19} / 0.839 _{0.002}	23.85 _{0.22} / 0.833 _{0.002}	24.31 _{0.14} / 0.854 _{0.001}	24.45 _{0.14} / 0.851 _{0.001}
	Linear ($\gamma_m = 50$)	24.49 _{0.28} / 0.846 _{0.002}	24.37 _{0.27} / 0.799 _{0.002}	25.07 _{0.24} / 0.854 _{0.001}	24.54 _{0.41} / 0.803 _{0.002}
	Linear ($\gamma_m = 150$)	24.56 _{0.33} / 0.836 _{0.002}	24.44 _{0.29} / 0.775 _{0.001}	25.34 _{0.31} / 0.841 _{0.001}	24.56 _{0.51} / 0.769 _{0.002}

at 24 views). On real data, however, unmodeled physical effects and the approximate posterior used by diffusion-based solvers cause the same setting to accumulate errors over 1000 iterations, dropping performance to 14.07 dB due to severe global intensity drift. A scheduling strategy helps bridge this gap. Using a high γ in the early, high-noise stages locks the trajectory onto the correct global geometry while mismatch errors remain masked by noise. Annealing γ in later stages, specifically $\gamma_t = \gamma_{\text{max}} \frac{t}{T}$ for reverse timestep $t \leq T$, then relaxes the measurement constraint, preventing the model from fitting experimental artifacts and allowing the prior to suppress residual noise and recover fine detail. For real data (y_{exp}), linear decay schedules consistently give the best results across all NFEs and view counts. For example, at NFE= 1000 and 24 views, the linear decay schedule ($\gamma_{\text{max}} = 150$) reaches 25.69 dB, fully recovering from the failure of the constant schedule.

Moving further towards the requirements of experimental CT, computational feasibility becomes a major bottleneck. A key strength of DDS-like sampling schemes is their potential to maintain high reconstruction quality even with drastically

reduced sampling steps. When the sampling budget is reduced to NFE=100, the scheduled approach not only remains stable but achieves the highest SSIM across all real-world configurations (0.851 at 24 views). Table II also reveals a trade-off regarding how steep this linear decay should be. A high initial weight ($\gamma_m = 150$) favors pixel-intensity accuracy (PSNR), likely because it enforces data consistency strongly in the early generative steps. Conversely, a lower weight ($\gamma_m = 5$) yields the highest structural similarity (SSIM). A lower maximum weight means the generative prior has more relative influence throughout the entire process, allowing the model to enforce cleaner, structurally coherent geometric edges, even if the absolute pixel intensities drift. We must note a critical “sweet spot” regarding these metrics. As highlighted by the perception-distortion tradeoff [15], visual inspection suggests that while PSNR can mask the retention of streak artifacts, SSIM might hide hallucinations by favoring plausible, piecewise-constant structures [16], [17]. Consequently, despite the slightly lower numerical peaks, the intermediate schedule ($\gamma_{\text{max}} = 50$) likely offers the most well-balanced

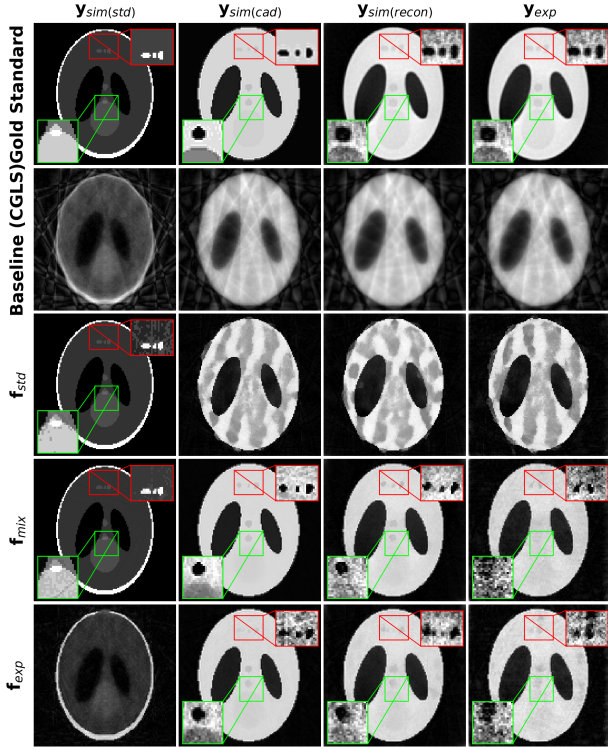


Fig. 2: Example reconstructions (averaged over nine samples) from 12 sparse projections across four test domains (rows). From top to bottom: gold standard, baseline without a prior, and reconstructions using the standard SL (f_{std}), mixed (f_{mix}), and experimental SL (f_{exp}) diffusion priors.

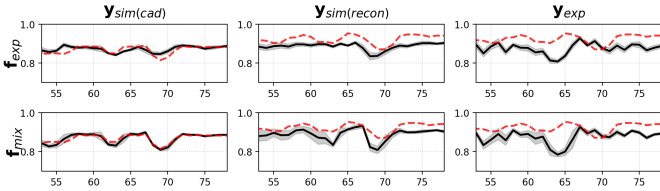


Fig. 3: Line profiles from Fig. 2 through the three small holes at row 25 of the laser-cut phantom. Results involving f_{std} and/or $y_{sim(cad)}$ are omitted. The gold standard is shown as a red dashed line, and shaded regions indicate ± 1 standard deviation over 10 random seeds.

reconstructions in our preliminary experiments.

V. CONCLUSION

We have shown that although diffusion-based solvers achieve excellent reconstruction quality on idealized synthetic data, their performance degrades when moving to experimental settings. This transition introduces multiple layers of complexity, from geometric variation in the object to physical inaccuracies in the measurement process. In our controlled study, these effects fall into two main categories: distributional shift in the prior and physical mismatch in the forward model. On the prior side, the best results were obtained for the high-

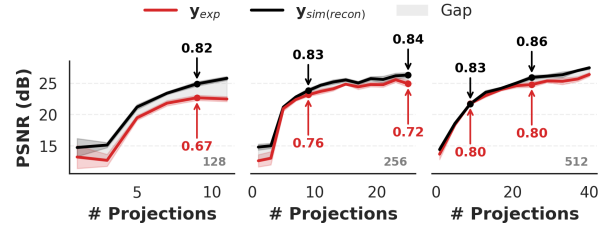


Fig. 4: PSNR (dB) vs number of projections for three resolutions (128, 256, 512) using f_{exp} on experimental data y_{exp} and simulated data $y_{sim(recon)}$, revealing the performance gap caused by forward model mismatch. Arrows indicate SSIM values at selected projection counts (9, 25). Shaded bands denote ± 1 standard deviation.

contrast standard SL phantom using a narrowly tailored prior. Regarding the prior, the best results were obtained on the high-contrast standard SL phantom with a narrowly tailored prior. As difficulty increased with the low-contrast design phantom, reconstruction quality generally declined. Yet a broader and more diverse prior surpassed the specialized model on this phantom, suggesting that exposure to varied structures can help even in a specific target domain. An additional shift appeared in full-view reconstructions of experimental data, which no longer exhibited piecewise-constant regions. The largest drop, however, occurred with real measurements, where forward-model mismatch introduced a trade-off: constant likelihood weighting either amplified artifacts or induced hallucinations. To address this, while also improving computational feasibility, we found that an annealed linear likelihood schedule outperformed constant weighting by emphasizing data consistency in early high-noise stages and prior guidance later. Preliminary results further suggest that this scheduled inference, also with fewer sampling steps, can match or exceed exhaustive sampling, offering a practical route toward robust experimental use.

Our study uses a canonical 2D phantom as a controlled, idealized test case. We view this as a best-case scenario, since the simple geometry allows a closely matched prior. Within this setting, we propose initial strategies to reduce the sim2real gap. In real applications, however, complex texture-rich volumes and severe noise are likely to widen this gap. Extending to 3D adds further challenges, including dependence on scan geometry, increased forward-model mismatch, and heuristic extensions of 2D diffusion models. Fully closing this gap will require further work, particularly on improved likelihood-weight scheduling and evaluation metrics that better reflect reconstruction fidelity.

CODE AND DATA AVAILABILITY

Code for all experiments is available at <https://github.com/ezgidetu/DDS-experimentalCT>. The experimental dataset is openly available on Zenodo at <https://zenodo.org> [18].

REFERENCES

- [1] E. Y. Sidky and X. Pan, "Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization," *Physics in Medicine & Biology*, vol. 53, no. 17, pp. 4777–4807, 2008.
- [2] G.-H. Chen, J. Tang, and S. Leng, "Prior image constrained compressed sensing (pics): a method to accurately reconstruct dynamic ct images from highly undersampled projection data sets," *Medical physics*, vol. 35, no. 2, pp. 660–663, 2008.
- [3] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE transactions on image processing*, vol. 26, no. 9, pp. 4509–4522, 2017.
- [4] H. Chen, Y. Zhang, Y. Chen, J. Zhang, W. Zhang, H. Sun, Y. Lv, P. Liao, J. Zhou, and G. Wang, "Learn: Learned experts' assessment-based reconstruction network for sparse-data ct," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1333–1347, 2018.
- [5] J. Adler and O. Öktem, "Learned primal-dual reconstruction," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1322–1332, 2018.
- [6] Y. Song, L. Shen, L. Xing, and S. Ermon, "Solving inverse problems in medical imaging with score-based generative models," in *International Conference on Learning Representations*, 2022.
- [7] H. Chung, B. Sim, D. Ryu, and J. C. Ye, "Improving diffusion models for inverse problems using manifold constraints," in *36th Conference on Neural Information Processing Systems, NeurIPS 2022*. Neural information processing systems foundation, 2022.
- [8] H. Chung, J. Kim, M. T. McCann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," in *The Eleventh International Conference on Learning Representations*, 2023.
- [9] B. Zhang, W. Chu, J. Berner, C. Meng, A. Anandkumar, and Y. Song, "Improving diffusion inverse problem solving with decoupled noise annealing," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 895–20 905.
- [10] H. Chung, S. Lee, and J. C. Ye, "Decomposed diffusion sampler for accelerating large-scale inverse problems," in *12th International Conference on Learning Representations, ICLR 2024*, 2024.
- [11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [12] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.
- [13] B. Efron, "Tweedie's formula and selection bias," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1602–1614, 2011.
- [14] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171.
- [15] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6228–6237.
- [16] S. Bhadra, V. A. Kelkar, F. J. Brooks, and M. A. Anastasio, "On hallucinations in tomographic image reconstruction," *IEEE transactions on medical imaging*, vol. 40, no. 11, pp. 3249–3260, 2021.
- [17] P. Kc, R. Zeng, N. Soni, and A. Badano, "sfrc for assessing hallucinations in medical image restoration," *arXiv preprint arXiv:2603.04673*, 2026.
- [18] X. Wang, N. J. Thomsen, E. Demircan-Tureyen, and F. Lucka, "Experimental 2d fan-beam x-ray ct dataset of a laser-cut shepp-logan-like phantom," 2026. [Online]. Available: <https://doi.org/10.5281/zenodo.19933728>