ERROR-CONDITIONED GENERATIVE SOLVERS FOR ILL-POSED INVERSE PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised conditional diffusion models solve inverse problems in a single sampling run with the measurement as conditioning, but never check whether the reconstruction reproduces the measurement under the forward operator. Plug-and-play posterior samplers enforce the measurement by optimizing a data-fidelity objective, which requires differentiating the operator and often the denoiser at every step, needs per-problem tuning, and is an unreliable proxy for reconstruction error when the operator has a large null space. We introduce *error-conditioned generative solvers*: at every sampling step the denoiser proposes a clean field, the forward operator is applied to it, and the resulting residual, together with a physics residual when governing equations are known, is fed back to the same denoiser as an input. Trained with the standard denoising loss, the network learns how to act on its own residual, with no backpropagation through the operator, denoiser, or sampling chain. We evaluate on six severely underdetermined problems: a new benchmark of three compressible flows reconstructed from an initial state and radiographs, together with compressed-sensing MRI, sparse-view CT and inverse scattering. Against the same conditional model at an equal number of network evaluations, residual feedback reduces reconstruction error by 6–40%. Our method attains the lowest reconstruction error of any posterior sampler we evaluate, while running 9–13$\times$ faster than plug-and-play samplers. Its samples preserve shock structure that a lower-error regression model smears, and 85% of its residual-driven correction lies in the operator's null space, where the measurement is blind.

1 INTRODUCTION

Inverse problems arise in medical imaging (Lustig et al., 2007), microscopy (Choi et al., 2007), geophysics (Virieux & Operto, 2009), and experimental fluid dynamics (Bell et al., 2025), where a full field must be recovered from indirect, noisy measurements. We study the severely ill-posed regime, where the measurement constrains a few percent of the directions in the field and the posterior is broad and often multimodal, so its mean is typically not a plausible field, motivating the use of generative inverse solvers.

Two families of diffusion solvers are widely used, and each has a limitation. *Supervised conditional models* (Adler & Öktem, 2018; Saharia et al., 2022; Bendel et al., 2023; Mammadov et al., 2024) are trained on paired data with the measurement supplied as conditioning and reconstruct in a single sampling run with no test-time optimization, but the measurement enters only as an input: the network never sees its own output through the forward operator, and in our experiments the conditional baseline attains up to three times our observation error (Figure 2). *Plug-and-play posterior samplers* (Chung et al., 2023; Zhang et al., 2025; Wu et al., 2024) instead steer an unconditional prior with the gradient of a data-fidelity term. One prior serves any operator, but every step differentiates the operator and often the denoiser, and how well the enforcement works depends on the problem: guidance is competitive on CS-MRI and inverse scattering, but on our flow benchmark, which is far more underdetermined, the most competitive sampler, DPS, has roughly 30–90% higher NRMSE than our method, and the other two considerably more.

We close the measurement loop inside a supervised sampler by conditioning it on its own measurement errors. We build on the recent error-conditioned neural solver framework (ENS; Jiang et al., 2026), which supplies a PDE residual to a network as an *input* and lets it learn an iterative correction. ENS refines a single deterministic estimate; we place the error-conditioning mechanism inside the

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

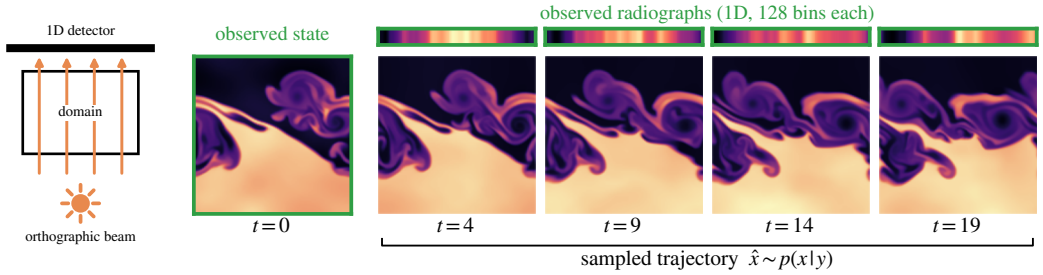

Figure 1: **Flow observation setting.** Green marks the observed data: the noisy state at $t = 0$ and a one-dimensional orthographic radiograph of 128 bins at $t = 4, 9, 14, 19$ (left: measurement geometry). The model samples the full 20-frame trajectory; the body forcing is never observed.

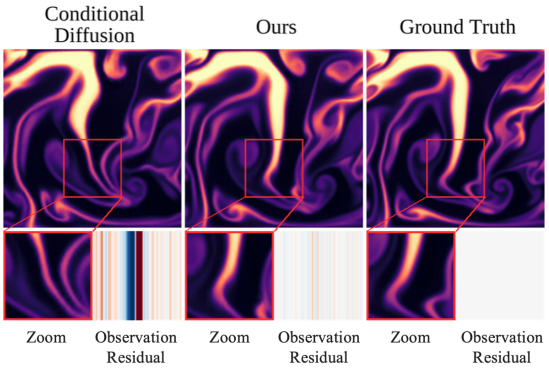

Figure 2: **Rayleigh–Taylor task reconstructions.** Density at the last frame of a 20-frame rollout, recovered from a noised initial state and four radiographs, for conditional diffusion and our method against the ground truth. Below each field is a detail region, and beside it the projection error over the 128 detector bins horizontally. Our method returns a field sample whose radiographs agree with the measured ones far more closely than the one sampled via conditional diffusion.

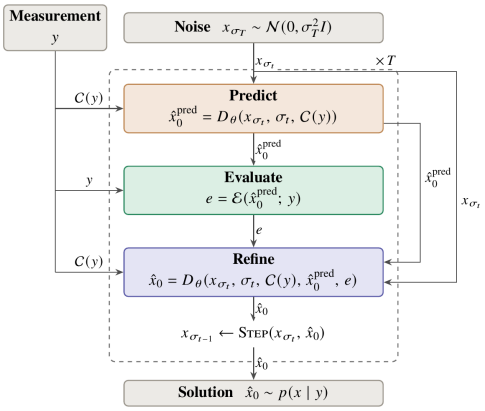

Figure 3: **Error-conditioned Generative Solver.** Schematic of the inference sampling process. The predict pass denoises the current iterate, the error operator $\mathcal{E}$ evaluates the residual e of that estimate, and the refine pass denoises again with the estimate and its residual as extra conditioning. Both passes share one network D_θ.

diffusion trajectory. At every sampling step the denoiser proposes a clean field, the forward operator is applied to that proposal, and its residual is supplied to the same denoiser, whose second evaluation produces the estimate the denoising sampler consumes (Figure 3). For a linear operator the residual channel is the backprojected mismatch $A^\top(A\hat{x} - y)$, the same quantity a data-fidelity gradient would use; the difference is that the network, conditioned on the noise level and its own estimate, decides how far to trust that signal and how to propagate it into the null space of A, the directions the measurement cannot see. The method therefore learns its correction from the residual rather than descending on it, in the lineage of recurrent inference machines (Putzky & Welling, 2017; Monga et al., 2021), placed inside a posterior sampler.

Our primary testbed is a new benchmark of three compressible-flow problems, which combine operator learning with an inverse problem; we also evaluate on compressed-sensing MRI, sparse-view CT and linear inverse scattering to show that the method is not specific to flows. A flow trajectory is reconstructed from a noisy initial state and four one-dimensional radiographs under an unobserved time-varying forcing, so that neither the initial condition nor the measurements determine the solution, a setting modeled on radiographic hydrodynamics experiments (Scarpetti et al., 2007; Kuranz et al., 2018) where only a handful of images can be recorded.

Our contributions are threefold. **(i) Error conditioning inside a sampler:** a shared-weight predict–evaluate–refine step in which the denoiser reads the residual of its own proposal, using only applications of A and $A^\top$ and no gradients through the operator, denoiser or sampler (Sec. 4). **(ii)**

Gains over baselines: at an equal number of network evaluations, error feedback lowers NRMSE by 6–40% on all six tasks; the method attains the lowest NRMSE of any posterior sampler on every task (a tie on inverse scattering) and is 9–13× faster than plug-and-play samplers, its samples keep shock structure, and most of its correction lies in the operator's null space (Sec. 6). **(iii) A combined forward–inverse flow benchmark** of a shock-forming compressible Euler flow and Kelvin–Helmholtz and Rayleigh–Taylor instabilities, with hidden forcing and sparse radiographs, modeled on radiographic hydrodynamics experiments (Sec. 5).

2 Background

2.1 Underdetermined Inverse Problems

We recover an unknown field $x \in \mathbb{R}^n$ from noisy measurements

$$y = Ax + \eta, \tag{1}$$

where $A : \mathbb{R}^n \to \mathbb{R}^m$, with $m \ll n$, is a known linear forward operator that is inexpensive to evaluate, and η is noise. With a prior $p(x)$ over plausible fields, inference targets the posterior $p(x \mid y) \propto p(y \mid x)\, p(x)$. The prior may marginalize unobserved causes of variability, $p(x) = \int p(x \mid z)\, p(z)\, dz$; in our flow problems z includes a time-varying forcing field that is not observed.

Because $m < n$, A has a nontrivial null space $\mathcal{N}(A)$, so $A(x + v) = A(x)$ and the measurement residual $r_{\text{obs}}(x) = Ax - y$ is invariant along that subspace:

$$r_{\text{obs}}(x + v) = r_{\text{obs}}(x), \qquad v \in \mathcal{N}(A), \qquad \text{hence} \qquad A\hat{x} = Ax \;\not\Rightarrow\; \hat{x} = x. \tag{2}$$

The likelihood is therefore constant on each coset $x + \mathcal{N}(A)$, and along those directions the posterior equals the prior. On the three static imaging tasks m/n lies between 0.031 and 0.063, so more than 93% of the directions in the field are of this kind; on the flow tasks the first frame is observed directly and the four radiographs supply 512 numbers for the roughly 1.2 to 1.6 million unknowns of the remaining nineteen frames.

Some problems supply further structure, such as governing PDEs, which we write as $g_k(x) \approx 0$ for $k = 1, \ldots, K$ and collect with the observation into the available error signals

$$e(x) = \big[\, Ax - y,\; g_1(x), \ldots, g_K(x) \,\big]. \tag{3}$$

Physical constraints narrow the set of plausible fields but do not uniquely determine one field, and numerical residuals are inexact: our finite-difference PDE residual does not vanish even on the ground truth (Figure 4), so a small residual $\|e(\hat{x})\|$ does not imply a small error $\|\hat{x} - x\|$. The measurement residual constrains only the row-space component of an estimate, so every method, ours included, relies on the prior for the null-space component, and methods differ in how they use the residual. We treat residuals as evidence of how a candidate violates the constraints, not as objectives whose minimizer is the solution.

2.2 Diffusion Sampling and Guidance

A diffusion model represents $p(x)$ by learning to reverse the noising process $x_\sigma = x + \sigma\epsilon$, $\epsilon \sim \mathcal{N}(0, I)$ (Song et al., 2021). In the EDM parameterization (Karras et al., 2022) a denoiser D_θ predicts the clean field, and Tweedie's formula (Efron, 2011) relates it to the score,

$$\nabla_{x_\sigma} \log p_\sigma(x_\sigma) = \frac{D_\theta(x_\sigma, \sigma) - x_\sigma}{\sigma^2}, \qquad D_\theta(x_\sigma, \sigma) \approx \mathbb{E}[x \mid x_\sigma]. \tag{4}$$

We write $\hat{x}_0$ for the denoiser output. Because $\hat{x}_0$ lies in the clean-field space, the error signals of Eq. 3 can be evaluated on it at every sampling step, and guidance-based samplers such as DPS (Chung et al., 2023) add the gradient of $\frac{1}{2}\|A\hat{x}_0 - y\|^2$ with respect to x_σ to the reverse update, which requires backpropagating through D_θ. At high noise levels, however, $\hat{x}_0$ is a conditional mean over many clean fields and lies far from the data manifold, so residuals evaluated on it need not define reliable descent directions (Sabour et al., 2025). A supervised conditional model $D_\theta(x_\sigma, \sigma, \mathcal{C}(y))$, trained on pairs with an encoding $\mathcal{C}(y)$ of the measurement (Saharia et al., 2022), avoids guidance altogether but never evaluates A. Error-conditioned neural solvers (Jiang et al., 2026) take a third route for PDE surrogates: a corrector refines an estimate as

$$\hat{u}^{(k+1)} = \hat{u}^{(k)} + \beta\, C_\phi\Big(\hat{u}^{(k)}, e^{(k)}\Big), \tag{5}$$

receiving the residual $e^{(k)}$ as an *input* rather than minimizing it, so the network learns how the structure of its error should change the estimate. That refinement is deterministic and its output is a point estimate; our method moves that correction inside the diffusion trajectory, so that each residual is evaluated at a known noise level and the output remains a posterior sample.

3 Related Work

Sections 1 and 2 position our method against plug-and-play samplers, supervised conditional models, and error-conditioned neural solvers (Jiang et al., 2026); Appendix E surveys these families and physics-informed solvers in full. Here we compare with the work closest to our mechanism: methods that give a network information derived from its own residual.

Recurrent inference machines (Putzky & Welling, 2017) and unrolled networks (Monga et al., 2021; Adler & Öktem, 2018; Aggarwal et al., 2019; Hammernik et al., 2018) supply the backprojected residual, the same signal our refine pass reads, but they are point estimators trained by backpropagating through the unrolled chain; our correction is applied inside a noise-conditioned posterior sampler and is trained one step at a time with a denoising loss.

Concurrently, ReCo-Diff (Choi et al., 2026) applies a related two-pass scheme to sparse-view CT, but it is deterministic and tied to one geometry, with the view count as its diffusion schedule; we condition on the residual against a fixed measurement through a general underdetermined operator, add physics residuals, and sample the posterior across imaging and flow tasks. Payeur et al. (2026) place a recurrent inference machine inside reverse diffusion, which requires differentiating the simulator; PRISMA (Sawhney et al., 2025) modulates a diffusion neural operator with a PDE residual on the state grid rather than through a forward operator; MACS (Tanevardi et al., 2025) uses the residual only as a scalar control of stochasticity; and self-conditioning (Chen et al., 2023) evaluates the network twice without any residual.

4 Method

Our method is a supervised conditional diffusion model whose denoiser is evaluated twice per sampling step (Figure 3): once to propose a clean field, and once more, given that proposal's residual, to produce a refined estimate the sampler consumes. Let $\mathcal{C}(y)$ denote a task-specific encoding of the measurement as image-space channels aligned with x, such as a zero-filled adjoint or a filtered backprojection; the denoiser $D_\theta(x_\sigma, \sigma, \cdot)$ follows the EDM parameterization (Karras et al., 2022).

Error operator. We collect the available constraint violations of a candidate $\hat{x}$ in an error operator

$$e = \mathcal{E}(\hat{x}; y) = \Big[\, e_{\text{obs}}(\hat{x}; y) \,\Big\|\, e_1(\hat{x}) \,\Big\|\, \cdots \,\Big\|\, e_K(\hat{x}) \,\Big], \tag{6}$$

where $\|$ is concatenation along channels. The observation component e_{obs} maps the measurement mismatch $A\hat{x} - y$ into the image space expected by the denoiser, which for our linear operators is its adjoint, $e_{\text{obs}} = A^\top(A\hat{x} - y)$, up to normalization; when the measurement has several blocks, each block's residual is mapped back separately and the results are concatenated. The components e_k are auxiliary residuals when they exist; on the flow tasks $\mathcal{E}$ additionally returns a strong-form PDE residual. Section 5 lists the components of $\mathcal{E}$ for each task, and App. B.3 defines the PDE residual.

Predict–evaluate–refine. A conventional conditional denoiser passes $D_\theta(x_\sigma, \sigma, \mathcal{C}(y))$ to the sampler. At every step we instead compute

$$\hat{x}_0^{\text{pred}} = D_\theta(x_\sigma, \sigma, \mathcal{C}(y);\ b_{\text{refine}} = \text{FALSE}), \tag{7}$$

$$e = \mathcal{E}\Big(\hat{x}_0^{\text{pred}}; y\Big), \tag{8}$$

$$\hat{x}_0 = D_\theta\Big(x_\sigma, \sigma, \mathcal{C}(y), \hat{x}_0^{\text{pred}}, e;\ b_{\text{refine}} = \text{TRUE}\Big), \tag{9}$$

and pass $\hat{x}_0$ to the sampler. The two evaluations share all weights and are distinguished by the flag b_{refine}; in refine mode the network receives the proposal and its residual as additional input channels. The method therefore adds no parameters beyond the extra input channels. Unlike a predictor–corrector Langevin step (Song et al., 2021) or the second evaluation of a Heun step, the second evaluation replaces the clean-field estimate passed to the sampler with one computed from the residual of the first.

Table 1: **Benchmark tasks.** All fields are 128^2; n unknowns, m real measurements. Each task is evaluated on 100 held-out items, one per independent source (trajectory, knee volume, CT patient, field). *On the flows almost all of m is the observed first frame, so the entry is radiograph bins per unknown in the remaining 19 frames.

	Compressible Euler	Kelvin–Helmholtz	Rayleigh–Taylor	CS-MRI	Sparse-view CT	Inverse scattering
Channels × frames	4×20	5×20	5×20	2×1	1×1	1×1
n	1,310,720	1,638,400	1,638,400	32,768	16,384	16,384
m	66,048	82,432	82,432	2,048	512	640
m/n	4.1×10^{-4}*	3.3×10^{-4}*	3.3×10^{-4}*	0.063	0.031	0.039
Training set	50k traj.	50k traj.	10k traj.	19.8k slices	100k slices	9.8k fields
Source	generated (Clawpack)			fastMRI	CT-RATE	InverseBench

A learned correction, not a gradient step. For a data-fidelity objective $\ell(\hat{x}) = \frac{1}{2}\|A\hat{x} - y\|_2^2$ the gradient is $A^\top(A\hat{x} - y)$, which is exactly the observation channel e_{obs}. The refine pass therefore receives the same information that a guidance step would use. The two differ in how that information is used. A guidance step moves x_σ along the fixed direction $J_\theta^\top A^\top(A\hat{x}_0 - y)$, where J_θ is the denoiser Jacobian, with a hand-tuned step size. The refine pass instead learns the correction $\Delta_\theta = \hat{x}_0 - \hat{x}_0^{\text{pred}}$ as a function of x_σ, σ, the measurement, the proposal and its residual, trained so that the corrected estimate is accurate. It can therefore scale the correction by how reliable the proposal is at the current noise level, discount measurement components it has learned to be unreliable, and use the prior to decide how an observed inconsistency should change the unobserved components of the field. Learning the update from the gradient signal is the role RIMs (Putzky & Welling, 2017) play in point estimation; here it is applied once per noise level inside a posterior sampler and trained with a per-step denoising loss. On sparse-view CT, most of the correction attributable to the residual lies in the null space of A, where a data-consistency step has no component (Appendix A.6).

Computational Cost. Each step costs two backbone passes and one application of A and $A^\top$, with no gradients through either; at $T = 100$ steps this matches the 200 backbone passes of the conditional baseline (Algorithm 2, Appendix B.2).

4.1 Training and Inference

We train one model per task on pairs (x, y). For each noised example we evaluate the predict pass, compute its residual with the same $\mathcal{E}$ used at inference, evaluate the network in refine mode, and supervise both outputs:

$$\mathcal{L} = \mathbb{E}_{x,y,\sigma,\epsilon}\Big[w(\sigma)\,\|\hat{x}_0 - x\|_2^2 + w(\sigma)\,\|\hat{x}_0^{\text{pred}} - x\|_2^2\Big], \qquad w(\sigma) = \frac{1}{\sigma^2 + \sigma_{\text{data}}^2}, \tag{10}$$

with $\sigma_{\text{data}} = 0.5$, the EDM preconditioning, and σ drawn from a cosine-interpolated density; the weighting is the soft-min-SNR schedule of Crowson et al. (2024), a smoothed variant of Min-SNR (Hang et al., 2023), expressed in clean-field space. The proposal is detached before the error operator is applied, so no gradient reaches the predict pass through $\hat{x}_0^{\text{pred}}$ or e: the two evaluations are coupled only by the shared weights, and each is supervised by its own loss term. Supervising the proposal keeps it a genuine reconstruction whose residual has a stable meaning; without it, the first pass could drift into an intermediate representation interpretable only by the second. No term in the objective minimizes $\|\mathcal{E}(\hat{x}; y)\|$: the residual influences the output only through the conditioning, and the reconstruction loss is what makes the network use the residual. At *inference*, we sample with a second-order DPM-Solver (Lu et al., 2025) with restart sampling (Xu et al., 2023), and temporal score rescaling (Xu et al., 2026), all training-free and independent of $\mathcal{E}$ (ablated in Appendix A.7).

5 Benchmarks

Flow benchmark. Operator learning maps a complete initial condition to the trajectory it determines; an inverse problem recovers a static field from an indirect measurement. We introduce a

benchmark that combines the two, modeled on radiography of driven hydrodynamic experiments. In these experiments the initial state is well characterized, and the evolution is recorded through a handful of line-integrated images, as in sequential radiography at DARHT (Scarpetti et al., 2007) or planar Rayleigh–Taylor experiments at the National Ignition Facility (Kuranz et al., 2018). Our three tasks mimic this setting (Figure 1, Table 1). Each is a two-dimensional compressible flow on a 128^2 grid over 20 frames, observed through a noisy full state at the first frame and 128-bin orthographic radiographs $\int \rho_{\mathrm{r}}\, dx$ of the radiographed density at frames 4, 9, 14 and 19. A smooth, time-varying body force, drawn per sample and given to no method, acts throughout. It stands in for unmodeled physics and for the mismatch between a simulation and the experiment it describes, which a method applied to real measurements has to tolerate. With the force unobserved, neither the initial state nor the radiographs determine the trajectory. **Compressible Euler** solves the forced Euler equations from a two-dimensional Riemann initial condition that forms interacting shocks; **Kelvin–Helmholtz** and **Rayleigh–Taylor** solve compressible Navier–Stokes with a species tracer for a sheared and a gravitationally unstable interface, and radiograph the heavy-fluid density. $\mathcal{C}(y)$ stacks the observed state, backprojected radiographs and masks; $\mathcal{E}$ returns the observed-state and projection residuals and a finite-difference residual of the governing equations (Appendix B.3).

Imaging tasks. **CS-MRI** recovers a complex single-coil fastMRI knee image (Zbontar et al., 2018) from 8 of 128 phase-encode lines (acceleration 16, mask redrawn per sample). **Sparse-view CT** recovers a CT-RATE chest slice (Hamamci et al., 2026) from four parallel-beam views in the LoDoPaB-CT geometry (Leuschner et al., 2021). **Linear inverse scattering** uses the InverseBench operator (Zheng et al., 2025) unchanged, with 8 transmitters and 40 receivers, fewer than published. $\mathcal{C}(y)$ is, respectively, the zero-filled adjoint with the mask, a filtered and an unfiltered backprojection, and the adjoint image; e_{obs} is always the adjoint of the measurement residual. Noise models, operator details, splits and example inputs (Figure A6) are in Appendix C.

6 Experiments and Analysis

Baselines. *Amortized* baselines are trained, like our method, on pairs of ground-truth fields and their measurements: FNO (Li et al., 2021); *conditional diffusion* (Saharia et al., 2022), our backbone without residual feedback, run at the same 200 backbone passes per sample as ours; and *regression*, our backbone trained with squared error, i.e. a posterior-mean estimator. *Plug-and-play samplers*, DPS (Chung et al., 2023), DAPS (Zhang et al., 2025) and PnP-DM (Wu et al., 2024), share one unconditional prior per task and run at their published budgets with per-task tuned step sizes. *Physics-informed* baselines apply to the flows: PINO (Li et al., 2024), PINO with test-time optimization (TTOP), and a per-instance PINN (Raissi et al., 2019). All diffusion models and the regression model share the transformer backbone of Li et al. (2025) (Appendices D.1 and D.2).

Metrics. On 100 held-out items per task (Table 1) we report NRMSE, PSNR, SSIM, and observation conformance (Obs) against the noise-free measurement (Appendix D.3). These are pointwise and favor the posterior mean, so Section 6.2 adds measures of sharpness and uncertainty.

6.1 Main Results

Among posterior samplers, ours attains the lowest NRMSE and highest SSIM on all six tasks and the highest PSNR on five (on inverse scattering DPS's is higher, 20.10 against 19.95), and it outperforms FNO and the physics-informed solvers (Table 2, Table 3). Against conditional diffusion, the same model without feedback at the same number of backbone passes, NRMSE falls by 6–40% and Obs by 17–70%. A regression model estimates the posterior mean and should beat any single sample on pointwise metrics; the like-for-like comparison is with our own posterior mean, and the mean of 20 of our samples is more accurate than regression on all three flows (‡ rows). Section 6.2 shows what regression loses in the field.

The effectiveness of minimizing the residual depends on the problem. On CS-MRI, DPS is competitive and has the lowest Obs (0.030 against our 0.034). On the flows the strongest baselines are those that never evaluate the operator, FNO on NRMSE and conditional diffusion on Obs, and DPS's Obs on Kelvin–Helmholtz is more than twice conditional diffusion's. On inverse scattering the two come apart within one task: DAPS and PnP-DM reach the lowest Obs (0.050, 0.048) but the two highest NRMSE of any sampler (0.437, 0.504). When the operator has a large null space, the data-fidelity objective is a weak proxy for reconstruction error, the residual–reconstruction gap of Jiang et al.

Table 2: **Physics problems.** Bold/underline: best/second best over the 100-item split. Rows marked ‡ compare the two posterior-mean estimates, regression (a deterministic squared-error model on our backbone) and the mean of 20 of our samples, on 5 items; they are not comparable to the rows above, and bold marks the better of the two.

Method	Compressible Euler				Kelvin–Helmholtz				Rayleigh–Taylor			
	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓
FNO	<u>0.154</u>	<u>28.75</u>	0.871	0.092	<u>0.141</u>	<u>26.93</u>	0.845	0.069	<u>0.200</u>	26.45	0.802	0.080
PINN	0.295	22.95	0.716	0.287	0.308	20.73	0.766	0.140	0.375	23.76	0.749	0.151
PINO	0.219	25.37	0.795	0.271	0.166	25.25	0.816	0.163	0.305	22.08	0.705	0.374
PINO+TTOP	0.258	24.12	0.765	0.245	0.510	15.81	0.663	0.194	0.350	19.12	0.709	0.102
Cond. diffusion	0.174	27.95	<u>0.888</u>	<u>0.081</u>	0.187	24.25	0.862	<u>0.023</u>	0.253	24.78	0.830	<u>0.025</u>
DPS	0.250	24.63	0.803	0.234	0.165	25.28	<u>0.874</u>	0.054	0.204	<u>26.54</u>	<u>0.858</u>	0.038
DAPS	0.399	20.90	0.701	0.247	0.489	16.49	0.623	0.066	0.538	18.97	0.619	0.068
PnP-DM	0.487	18.58	0.627	0.250	0.365	18.12	0.638	0.068	0.524	17.21	0.599	0.061
Ours	**0.133**	**30.56**	**0.921**	**0.044**	**0.129**	**27.73**	**0.909**	**0.007**	**0.151**	**29.47**	**0.895**	**0.009**
Regression‡	0.117	31.53	0.927	0.055	0.099	30.50	0.924	0.023	0.127	31.69	0.908	0.027
Ours, posterior mean‡	**0.098**	**33.36**	**0.949**	**0.030**	**0.091**	**31.14**	**0.945**	**0.005**	**0.112**	**32.85**	**0.927**	**0.009**

Table 3: **Imaging problems.** Bold/underline: best/second best. Time: wall clock per sample on one NVIDIA L40S GPU, measured on inverse scattering (full inference compute details in Table A1).

Method	CS-MRI				Sparse-view CT				Inverse scattering				
	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓	Time (s)
Cond. diffusion	0.328	26.03	0.591	0.096	<u>0.272</u>	<u>19.32</u>	<u>0.530</u>	0.065	0.355	19.30	<u>0.788</u>	0.127	4.01
DPS	<u>0.278</u>	<u>27.81</u>	<u>0.632</u>	**0.030**	0.299	18.47	0.493	<u>0.062</u>	<u>0.336</u>	**20.10**	0.576	0.084	43.4
DAPS	0.352	25.59	0.553	0.032	0.316	17.94	0.475	0.070	0.437	17.22	0.460	<u>0.050</u>	38.7
PnP-DM	0.420	24.07	0.507	<u>0.031</u>	0.344	17.27	0.382	0.078	0.504	16.33	0.691	**0.048**	55.2
Ours	**0.257**	**28.27**	**0.699**	0.034	**0.235**	**20.63**	**0.617**	**0.054**	**0.334**	<u>19.95</u>	**0.801**	0.095	4.11

(2026). Supplying the residual as an input gives the lowest NRMSE of any posterior sampler on all six tasks and the lowest Obs on four, at 4.1 s per sample against 39–55 s for plug-and-play samplers, which take 1000–2651 network passes and differentiate the operator up to 2×10^4 times (Table 3; Appendix A.1).

6.2 Shock Structure and Sharpness

Pointwise metrics cannot distinguish a sharp field from a smoothed average of plausible ones, so on compressible Euler we measure sharpness directly: shock thickness along 229,116 shock normals located on the ground truth, and the fine-scale energy in each frame's power spectrum (Appendix A.3).

Figure 4 and Figure 5 show two opposite failures. Deterministic methods thicken shocks and lose fine-scale energy (FNO by 1.29 cells at 29% of the true band energy, regression by 0.41 at 64%), while plug-and-play samplers are also too thick, and their samples have 1.2–2.0× the true band energy, as spurious high-frequency content. Single draws of ours match the true thickness to within 0.04 cells on average, as conditional diffusion does, at 98.5% of the band energy, and have the lowest per-shock thickness error (0.21 cells). Regression has the lowest density NRMSE (0.0680 against 0.0689), yet its thickness bias grows to +1.13 cells by the last frame, while ours stays below +0.07. When a point estimate is wanted, the mean of our samples is more accurate than regression (Table 2, ‡ rows), though smoother than any individual draw.

Sample spread. Each of our draws has a sharp shockfront, and the draws differ in the position of that front (Figure 5), so the uncertainty appears as a distribution over front positions rather than as a smoothed front. The variation between draws concentrates on fronts and interfaces and is near zero

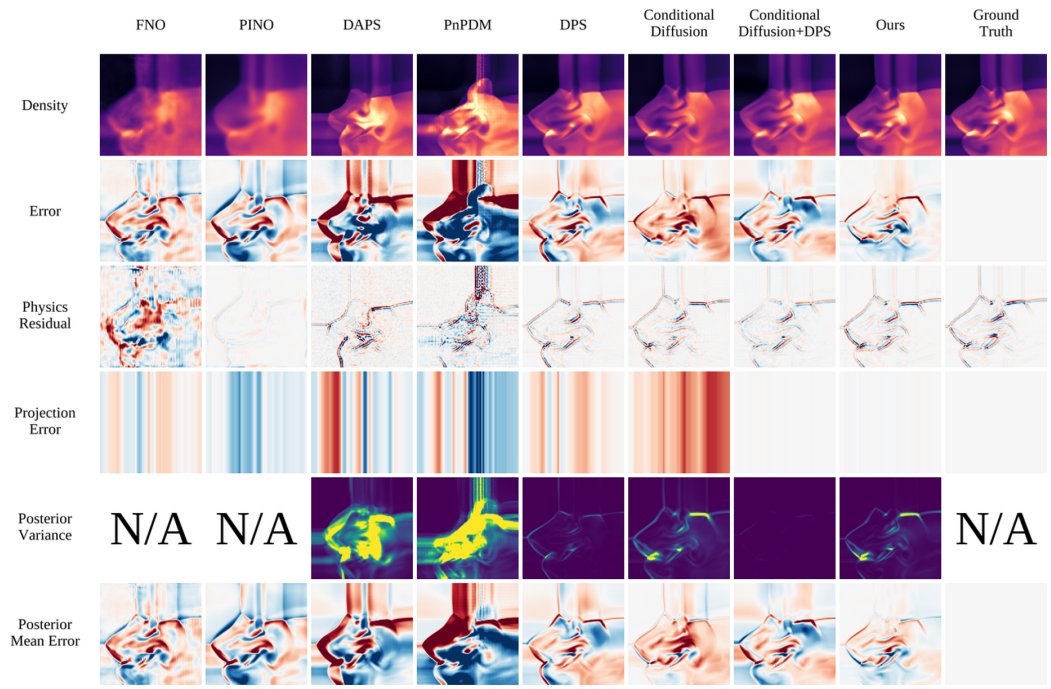

Figure 4: **Qualitative comparison on compressible Euler**, final frame of a held-out trajectory. The physics residual is the strong-form Euler residual, nonzero even on the ground truth since it is a coarse finite-difference approximation; the projection error spans 128 detector bins (horizontal); the last two rows are computed from 20 samples. The conditional diffusion + DPS column is analyzed in Appendix A.2.

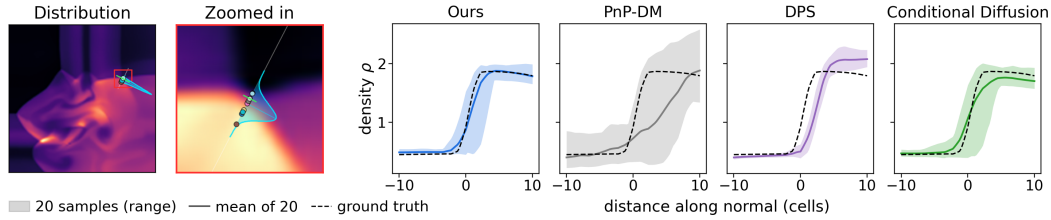

Figure 5: **Shock profile analysis.** One shock normal (frame 18). Left: shock locations along the normal in 20 of our draws (dots), a fitted Gaussian, and the ground truth (green tick). Right: density along the normal for our method and three baselines.

elsewhere (Figure 4), whereas DPS's samples barely differ and DAPS and PnP-DM vary most over the regions they reconstruct wrongly.

6.3 Ablation and Subspace Analysis

On compressible Euler every residual type contributes: removing the projection, observation or physics residual raises NRMSE by 11%, 4% and 0.6% (Table A4). A two-pass control with the residual input zeroed reaches 0.1674, only marginally better than single-pass conditional diffusion (0.1741) and far from ours (0.1400; all under the plain sampler), so the gain comes from the residual rather than from the second denoising evaluation.

To see how the residual is used, we split each correction $\Delta = \hat{x}_0 - \hat{x}_0^{\text{pred}}$ on four-view CT, where A has rank 510 of $16{,}384$, into the part the measurement constrains, $\text{row}(A)$, and the part it cannot see, $\text{null}(A)$. Any data-consistency step, a gradient $A^\top(A\hat{x}_0 - y)$ or a projection $A^+(y - A\hat{x}_0)$, lies entirely in $\text{row}(A)$, and DPS guidance is 70% row-space. The residual-driven part of our correction, Δ_{attr}, the change when the residual is swapped for another sample's, is 15.5% row-space (Figure 6): most of what the refine pass does with the residual happens where the measurement is blind. This null-space part points against the proposal's error, toward the ground truth (cosine $+0.18$

Table 4: **Shock structure on compressible Euler**, one draw per method over the 100-item evaluation split. δ: shock thickness in cells; $E_{16\text{–}50}/E_{\mathrm{gt}}$: high-band spectral energy relative to the ground truth; NRMSE_ρ: density only, so not comparable to Table 2. Bold/underline: best/second best, meaning closest to 0 (bias), to 1 (ratio), or lowest. Definitions in Appendix A.3.

	Thickness		Sharpness	
Method	δ bias	δ err.	$E_{16\text{–}50}/E_{\mathrm{gt}}$	NRMSE_ρ
FNO	$+1.29$	1.34	0.291	0.1006
Regression	$+0.41$	0.49	0.641	**0.0680**
Cond. diffusion	$+\mathbf{0.04}$	<u>0.29</u>	<u>0.951</u>	0.1002
DPS	$+0.40$	0.84	1.205	0.1790
DAPS	<u>$+0.15$</u>	0.77	1.260	0.2307
PnP-DM	$+0.21$	0.86	2.027	0.3159
Ours	$+\mathbf{0.04}$	**0.21**	**0.985**	<u>0.0689</u>

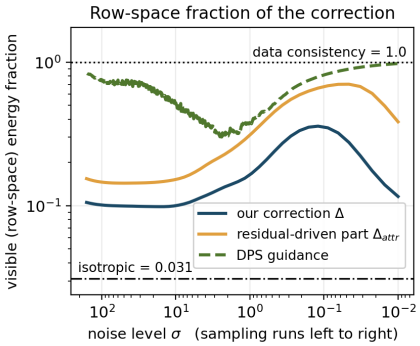

Figure 6: **Correction decomposition analysis on the CT task.** Fraction of each correction's energy in $\mathrm{row}(A)$, the part visible to the measurement. The dash-dotted line, $\mathrm{rank}(A)/n$, is the expected fraction for an isotropically random direction. Details in Appendix A.6.

with the negated error, mostly at high noise levels), and the error that remains is almost entirely invisible to the measurement (Appendix A.6). The refine pass therefore does not behave simply as a learned data-fidelity step. It receives a residual confined to $\mathrm{row}(A)$ and changes the field mainly in $\mathrm{null}(A)$, where the proposal's remaining error lies and where only the learned prior can determine the content.

7 Conclusion and Limitations

We introduced error-conditioned generative solvers, diffusion inverse solvers that evaluate the forward operator on intermediate predictions at each sampling step and read the resulting residual as an input rather than minimizing it. Across six severely underdetermined problems in compressible flow, medical imaging and inverse scattering, they attain the lowest reconstruction error among posterior samplers on all six and the best observation conformance on four. Our comparisons suggest that driving the constraint residual down is not the most effective route to an accurate reconstruction: plug-and-play samplers that do so are competitive on some problems but not others, and each requires a per-problem search over its measurement weight. Supplying the residual as an input is accurate on every task with a single inference configuration, and its samples are physically faithful: on compressible Euler its single draws match the true shock thickness to within 0.04 cells on average, with the lowest per-shock error, where deterministic baselines smear the front and plug-and-play samplers add spurious details.

Unlike plug-and-play samplers, our method is amortized: each problem needs paired training data and its own model, and the forward operator must be fixed before training. In exchange, inference needs no gradients of the operator or the denoiser and is an order of magnitude faster. Our evaluation is limited to two-dimensional problems with linear operators, the flow problems are reconstructed over a fixed time window, and our sample-spread diagnostics are not a calibration study. Extending the method to nonlinear and three-dimensional problems, to trajectories of arbitrary length, and to downstream inference from its posterior samples are promising directions for future work.

AI use statement

In this work we used generative AI tools for implementation: models assisted in writing, refactoring and debugging the training, sampling and evaluation code for our method, and the data generation pipeline for the flow benchmark of Section 5. We also used them for research hypothesis refinement and methodology design, as a sounding board during ideation and while deciding which ablations and analyses to run; the research questions, the design of the method and the choice of experiments are the authors' own. We additionally used generative AI for literature identification, to search

for and summarize related work and to suggest method families to position against, and for paper drafting and editing, to draft and revise passages of the manuscript and to format references.

We reviewed all AI-assisted work. All AI-assisted code was verified by the authors and AI-assisted text was checked against the results, and edited by the authors. We take responsibility for the final content of this work, including the text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 4 and Algorithms 1 and 2 specify the method. Appendix B gives the sampler configuration shared by all six tasks and the PDE residual. Appendix C describes how the three flow datasets were generated and the sources, operators and splits of the imaging tasks. Appendix D describes the architecture and training procedure, the configuration and tuning of the baselines, and the evaluation metrics. We commit to releasing the code for the method and the datasets upon acceptance.

REFERENCES

Jonas Adler and Ozan Öktem. Deep Bayesian inversion, 2018. URL https://arxiv.org/abs/1811.05910.

Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1322–1332, 2018. doi: 10.1109/TMI.2018.2799231. URL https://doi.org/10.1109/TMI.2018.2799231.

Hemant K. Aggarwal, Merry P. Mani, and Mathews Jacob. MoDL: Model-based deep learning architecture for inverse problems. *IEEE Transactions on Medical Imaging*, 38(2):394–405, 2019. doi: 10.1109/TMI.2018.2865356. URL https://doi.org/10.1109/tmi.2018.2865356.

Evan Bell, Daniel A. Serino, Ben S. Southworth, Trevor Wilcox, and Marc L. Klasky. Learning robust parameter inference and density reconstruction in flyer plate impact experiments. *Scientific Reports*, 15:40846, 2025. doi: 10.1038/s41598-025-24570-y. URL https://doi.org/10.1038/s41598-025-24570-y.

Matthew Bendel, Rizwan Ahmad, and Philip Schniter. A regularized conditional GAN for posterior sampling in image recovery problems. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68673–68684, 2023. doi: 10.52202/075280-3005. URL https://doi.org/10.52202/075280-3005.

Gabriel Cardoso, Yazid Janati El Idrissi, Sylvain Le Corff, and Éric Moulines. Monte Carlo guided denoising diffusion models for Bayesian linear inverse problems. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/bfb3c1db0deda449ff4027f77dd48704-Paper-Conference.pdf.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3itjR9QxFw.

W. Choi, C. Fang-Yen, K. Badizadegan, S. Oh, N. Lue, R. R. Dasari, and M. S. Feld. Tomographic phase microscopy. *Nature Methods*, 4(9):717–719, 2007. doi: 10.1038/nmeth1078. URL https://doi.org/10.1038/nmeth1078.

Yong Eun Choi, Hyoung Suk Park, Kiwan Jeon, Hyun-Cheol Park, and Sung Ho Kang. ReCo-Diff: Residual-conditioned deterministic sampling for cold diffusion in sparse-view CT, 2026. URL https://arxiv.org/abs/2603.02691.

Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Clawpack Development Team. Clawpack software, 2026. URL http://www.clawpack.org. Version 5.14.0.

Katherine Crowson, Stefan Andreas Baumann, Alex Birch, Tanishq Mathew Abraham, Daniel Z. Kaplan, and Enrico Shippole. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In *Proceedings of the 41st International Conference on Machine Learning, ICML 2024*, 2024. URL https://arxiv.org/abs/2401.11605.

Hongkun Dou, Zike Chen, Zeyu Li, Hongjue Li, Lijun Yang, and Yue Deng. Constrained particle seeking: Solving diffusion inverse problems with just forward passes. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 20870–20878, 2026. doi: 10.1609/aaai.v40i25.39227. URL https://doi.org/10.1609/aaai.v40i25.39227.

Zehao Dou and Yang Song. Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tplXNcHZs1.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, December 2011. doi: 10.1198/jasa.2011.tm11181. URL http://dx.doi.org/10.1198/jasa.2011.tm11181.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Ahmet Kaplan, Zhiyong Lu, Malgorzata Polacin, Bernhard Kainz, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*, 10(8):1610–1628, February 2026. doi: 10.1038/s41551-025-01599-y. URL http://dx.doi.org/10.1038/s41551-025-01599-y.

Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P. Recht, Daniel K. Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated MRI data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018. doi: 10.1002/mrm.26977. URL http://dx.doi.org/10.1002/mrm.26977.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via Min-SNR weighting strategy. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7407–7417, 2023. doi: 10.1109/ICCV51070.2023.00684. URL https://doi.org/10.1109/ICCV51070.2023.00684.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for PDEs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 72525–72624, 2024. doi: 10.52202/079017-2311. URL https://doi.org/10.52202/079017-2311.

Jiahe Huang, Guandao Yang, Zichen Wang, and Jeong Joon Park. DiffusionPDE: Generative PDE-solving under partial observation. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*, volume 37, pp. 130291–130323, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb3878c1dcbfff9ee95d5d033e5f5942-Abstract-Conference.html.

Haina Jiang, Liam Wang, Peng-Chen Chen, Min Seop Kwak, Seungryong Kim, Brian Bell, and Jeong Joon Park. Error-conditioned neural solvers, 2026. URL https://arxiv.org/abs/2606.27354.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9,*

2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems 35*, NeurIPS 2022, pp. 23593–23606. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022. doi: 10.52202/068431-1714. URL http://dx.doi.org/10.52202/068431-1714.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023. URL http://jmlr.org/papers/v24/21-1524.html.

C. C. Kuranz, H.-S. Park, C. M. Huntington, A. R. Miles, B. A. Remington, T. Plewa, M. R. Trantham, H. F. Robey, D. Shvarts, A. Shimony, K. Raman, S. MacLaren, W. C. Wan, F. W. Doss, J. Kline, K. A. Flippo, G. Malamud, T. A. Handy, S. Prisbrey, C. M. Krauland, S. R. Klein, E. C. Harding, R. Wallace, M. J. Grosskopf, D. C. Marion, D. Kalantar, E. Giraldez, and R. P. Drake. How high energy fluxes may affect Rayleigh–Taylor instability growth in young supernova remnants. *Nature Communications*, 9:1564, 2018. doi: 10.1038/s41467-018-03548-7. URL https://doi.org/10.1038/s41467-018-03548-7.

Johannes Leuschner, Maximilian Schmidt, Daniel Otero Baguer, and Peter Maass. LoDoPaB-CT, a benchmark dataset for low-dose computed tomography reconstruction. *Scientific Data*, 8(1), April 2021. doi: 10.1038/s41597-021-00893-z. URL http://dx.doi.org/10.1038/s41597-021-00893-z.

Edward Li, Zichen Wang, Jiahe Huang, and Jeong Joon Park. VideoPDE: Unified generative PDE solving via video inpainting diffusion models, 2025. URL https://arxiv.org/abs/2506.13754.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM / IMS Journal of Data Science*, 1(3):1–27, 2024. doi: 10.1145/3648506. URL https://doi.org/10.1145/3648506.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025. doi: 10.1007/S11633-025-1562-4. URL https://doi.org/10.1007/s11633-025-1562-4.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021. doi: 10.1038/s42256-021-00302-5. URL https://doi.org/10.1038/s42256-021-00302-5.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11451–11461, 2022. doi: 10.1109/CVPR52688.2022.01117. URL https://doi.org/10.1109/CVPR52688.2022.01117.

Michael Lustig, David Donoho, and John Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58:1182–1195, 2007. doi: 10.1002/mrm.21391. URL https://doi.org/10.1002/mrm.21391.

Abbas Mammadov, Hyungjin Chung, and Jong Chul Ye. Amortized posterior sampling with diffusion prior distillation, 2024. URL https://arxiv.org/abs/2407.17907.

Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/7711026951f3dc3afb2be427bf98bc73-Paper-Conference.pdf.

Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021. doi: 10.1109/MSP.2020.3016905. URL https://doi.org/10.1109/msp.2020.3016905.

Guillaume Payeur, Laurence Perreault-Levasseur, Gabriel Missael Barco, and Yashar Hezaveh. Strong gravitational lensing posterior sampling in pixel-space using diffusion models and recurrent inference machines, 2026. URL https://arxiv.org/abs/2607.19459.

Patrick Putzky and Max Welling. Recurrent inference machines for solving inverse problems, 2017. URL https://arxiv.org/abs/1706.04008.

M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045. URL https://doi.org/10.1016/j.jcp.2018.10.045.

Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (RED). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017. doi: 10.1137/16M1102884. URL https://doi.org/10.1137/16m1102884.

François Rozet and Gilles Louppe. Score-based data assimilation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 40521–40541, 2023. doi: 10.52202/075280-1763. URL https://doi.org/10.52202/075280-1763.

Amirmojtaba Sabour, Michael S. Albergo, Carles Domingo-Enrich, Nicholas M. Boffi, Sanja Fidler, Karsten Kreis, and Eric Vanden-Eijnden. Test-time scaling of diffusions with flow maps, 2025. URL https://arxiv.org/abs/2511.22688.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In Munkhtsetseg Nandigjav, Niloy J. Mitra, and Aaron Hertzmann (eds.), *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Vancouver, BC, Canada, August 7 - 11, 2022*, pp. 15:1–15:10. ACM, 2022. doi: 10.1145/3528233.3530757. URL https://doi.org/10.1145/3528233.3530757.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2023. doi: 10.1109/tpami.2022.3204461. URL http://dx.doi.org/10.1109/TPAMI.2022.3204461.

Medha Sawhney, Abhilash Neog, Mridul Khurana, and Anuj Karpatne. PRISMA: Improving the accuracy-latency frontier of diffusion-based PDE solvers using physics-informed spectral attention, 2025. URL https://arxiv.org/abs/2512.01370.

R. D. Scarpetti, S. Nath, J. Barraza, C. A. Ekdahl, E. Jacquez, B. T. McCuistian, K. Nielsen, M. Schulze, J. Seitz, G. J. Caporaso, Y.-J. Chen, G. Logan, and F. Bieniosek. Status of the DARHT 2nd axis accelerator at the Los Alamos National Laboratory. In *Proceedings of the 2007 IEEE Particle Accelerator Conference (PAC07)*, pp. 831–835, Albuquerque, New Mexico, USA, 2007. doi: 10.1109/PAC.2007.4441094. URL https://proceedings.jacow.org/p07/PAPERS/TUXC01.PDF.

Jo Schlemper, Jose Caballero, Joseph V. Hajnal, Anthony N. Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE Transactions on Medical Imaging*, 37(2):491–503, 2018. doi: 10.1109/TMI.2017.2760978. URL https://doi.org/10.1109/tmi.2017.2760978.

Dule Shu, Zijie Li, and Amir Barati Farimani. A physics-informed diffusion model for high-fidelity flow field reconstruction. *Journal of Computational Physics*, 478:111972, 2023. doi: 10.1016/j.jcp.2023.111972. URL https://doi.org/10.1016/j.jcp.2023.111972.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9_gsMA8MRKQ.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2011.13456.

Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C. Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Michael Johnson. End-to-end variational networks for accelerated MRI reconstruction. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 12262 of *Lecture Notes in Computer Science*, pp. 64–73, 2020. doi: 10.1007/978-3-030-59713-9_7. URL https://doi.org/10.1007/978-3-030-59713-9_7.

Amirreza Tanevardi, Pooria Abbas Rad Moghadam, Seyed Mohammad Eshtehardian, Sajjad Amini, and Babak Khalaj. MACS: Measurement-aware consistency sampling for inverse problems, 2025. URL https://arxiv.org/abs/2510.02208.

Utkarsh Utkarsh, Pengfei Cai, Alan Edelman, Rafael Gomez-Bombarelli, and Christopher Vincent Rackauckas. Physics-constrained flow matching: Sampling generative models with hard constraints. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.04171.

Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 945–948. IEEE, 2013. doi: 10.1109/GlobalSIP.2013.6737048. URL https://doi.org/10.1109/globalsip.2013.6737048.

Jean Virieux and Stéphane Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, 2009. doi: 10.1190/1.3238367. URL https://doi.org/10.1190/1.3238367.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=mRieQgMtNTQ.

Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine L. Bouman. Principled probabilistic imaging using diffusion models as plug-and-play priors. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/d65c4ce22241138c1784ff753d4c746c-Abstract-Conference.html.

Yanbo Xu, Yu Wu, Sungjae Park, Zhizhuo Zhou, and Shubham Tulsiani. Temporal score rescaling for temperature sampling in diffusion and flow models. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2510.01184.

Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi S. Jaakkola. Restart sampling for improving generative processes. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/f2543511e5f4d4764857f9ad833a977d-Abstract-Conference.html.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An open dataset and benchmarks for accelerated MRI, 2018. URL https://arxiv.org/abs/1811.08839.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 20895–20905. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01946. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Improving_Diffusion_Inverse_Problem_Solving_with_Decoupled_Noise_Annealing_CVPR_2025_paper.html.

Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy Feng, Caifeng Zou, Yu Sun, Nikola Borislavov Kovachki, Zachary E. Ross, Katherine L. Bouman, and Yisong Yue. InverseBench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=U3PBITXNG6.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1219–1229. IEEE, June 2023. doi: 10.1109/cvprw59228.2023.00129. URL http://dx.doi.org/10.1109/CVPRW59228.2023.00129.

APPENDIX

A ADDITIONAL RESULTS

A.1 INFERENCE COST

Table A1 reports inference cost, measured per reconstruction on a single NVIDIA L40S GPU. Our method produces a reconstruction in $4.11\,\mathrm{s}$, against 38.7 to $55.2\,\mathrm{s}$ for the plug-and-play methods, a factor of 9 to 13, while attaining the lowest NRMSE and the highest SSIM on this task. The structural difference appears in the two gradient columns: DPS differentiates the backbone at each of its 1000 steps, all three plug-and-play methods differentiate the forward operator between 10^3 and 2×10^4 times, and our method does neither, requiring only applications of the operator and its adjoint. Our steps cost twice what conditional diffusion's do, $0.041\,\mathrm{s}$ against $0.020\,\mathrm{s}$, because each step evaluates the denoiser twice. Against the plug-and-play methods the per-step costs are close: ours is between DPS's $0.043\,\mathrm{s}$ and DAPS's $0.039\,\mathrm{s}$, and is double PnP-DM's $0.021\,\mathrm{s}$. What separates us is the number of network evaluations. We take 100 steps, 200 backbone passes, where their published configurations take 1000 to 2651 forward passes, and our method attains the lowest NRMSE at that budget.

Table A1: **Inference cost, measured on inverse scattering.** Values are for one reconstruction on a single NVIDIA L40S GPU. NFE is reported in each method's own unit, which is sampling steps for conditional diffusion and our method (each of our steps runs the backbone twice), and network forward passes for the plug-and-play rows. DAPS and PnP-DM counts are measured per run rather than derived from their configurations, because PnP-DM's inner scheduler truncates at the current annealing noise level.

Method	**NFE**	**Backbone fwd**	**Backbone bwd**	**Operator evals**	**s/NFE**	**s/sample**
Cond. diffusion	200	200	0	0	0.020	4.01
DPS	1000	1000	1000	1000	0.043	43.4
DAPS	1000	1000	0	20000	0.039	38.7
PnP-DM	2651	2651	0	4000	0.021	55.2
Ours	100	200	0	100	0.041	4.11

A.2 CONDITIONAL MODEL WITH GUIDANCE

The two families can be combined: a supervised conditional model serves as the prior for a guidance-based sampler, so that the measurement enters both as conditioning and through a data-fidelity gradient. Table A2 reports that combination on the three flow tasks, beside each of its two components and our method. We pair the conditional diffusion model of Table 2 with DPS guidance at the 1000-step schedule of the DPS rows; Figure 4 shows the same combination qualitatively on compressible Euler.

Table A2: **Conditional diffusion under DPS guidance.** Bold indicates the best performance and underline the second best. The conditional diffusion, DPS and ours rows are those of Table 2, repeated here for comparison.

Method	**Compressible Euler**				**Kelvin–Helmholtz**				**Rayleigh–Taylor**			
	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓
Cond. diffusion	<u>0.174</u>	<u>27.95</u>	<u>0.888</u>	<u>0.081</u>	0.187	24.25	0.862	<u>0.023</u>	0.253	24.78	0.830	<u>0.025</u>
DPS	0.250	24.63	0.803	0.234	0.165	25.28	0.874	0.054	0.204	26.54	0.858	0.038
Cond. diffusion + DPS	0.233	25.20	0.808	0.245	<u>0.152</u>	<u>25.92</u>	<u>0.877</u>	0.063	<u>0.179</u>	<u>28.43</u>	<u>0.862</u>	0.063
Ours	**0.133**	**30.56**	**0.921**	**0.044**	**0.129**	**27.73**	**0.909**	**0.007**	**0.151**	**29.47**	**0.895**	**0.009**

Guidance improves the conditional model's reconstruction on Kelvin–Helmholtz and Rayleigh–Taylor, by 19% and 29% NRMSE, and degrades it on compressible Euler by 34%. On all three

tasks the combination is less accurate than our method, at 12 to 16 times the wall clock, and it has the cost structure of the guidance family: a backward pass through the denoiser and an operator evaluation at each of 1000 steps.

On observation conformance the ordering reverses. Adding the gradient raises Obs above the conditional model's on all three tasks, from 0.081 to 0.245 on compressible Euler, from 0.023 to 0.063 on Kelvin–Helmholtz, and from 0.025 to 0.063 on Rayleigh–Taylor. Scoring the two parts of the measurement separately accounts for this increase. The observed initial state contributes $65{,}536$ of the $66{,}048$ measured numbers on compressible Euler and $81{,}920$ of $82{,}432$ on the other two tasks, so Obs is dominated by the initial state. On the four radiographs alone the combination attains the lowest projection residual of any method we run, 0.0071, 0.0046 and 0.0040 against our 0.0072, 0.0118 and 0.0203. On the initial state alone it has the largest residual of the four, 0.2465, 0.0628 and 0.0633 against our 0.0447, 0.0069 and 0.0093.

Guidance from a conditional prior therefore lowers the residual on the radiographs and raises it on the initial state, which dominates Obs.

A.3 Shock Structure and Sharpness

Pointwise accuracy metrics cannot separate a sharp field from a smoothed average of plausible fields. On compressible Euler the governing equations produce shocks, and a shock has a thickness that a reconstruction either reproduces or does not, so we measure it directly. Everything below is on density in physical units, over frames 1 to 19 of the 100-item evaluation split. Frame 0 is the observed initial condition and is excluded.

Shock thickness. We locate shocks on the ground truth only and evaluate every method along the same lines, so no method benefits from failing to detect a front. We take points where $|\nabla\rho|$ of the ground truth exceeds 0.15, keep the crest of each front by non-maximum suppression along the ground-truth normal $\hat{n} = \nabla\rho/|\nabla\rho|$, and sample ρ and p bilinearly along $\hat{n}$ over ± 10 cells at a spacing of 0.25 cells. A ground-truth profile counts as a shock if its pressure ratio exceeds 1.5, the entropy p/ρ^{γ} increases across it, and its plateaus are flat, which excludes the contact discontinuities and rarefactions that carry most of the high-gradient fronts in this data. That leaves 229,116 shock lines over the evaluation split, one of which is drawn in Figure 5. On each line we read the upstream and downstream densities ρ_1, ρ_2 as the median over the outer three cells on each side, and define the thickness $\delta = |\rho_2 - \rho_1| / \max_s |d\rho/ds|$, with s the distance along the line. The thickness figures move by 0.1 cells or less across gradient thresholds of 0.1, 0.15 and 0.25.

Band energy. For each frame we compute the radially averaged power spectrum of density on the 128^2 grid: the frame mean is removed, a separable Hann taper is applied with its power loss compensated, and the power $|\hat{\rho}(k)|^2$ is averaged over rings of integer wavenumber k, in cycles per domain. The band energy is the power summed over $16 \le k \le 50$, and we report its ratio to the ground truth's band energy, averaged over frames and items. A ratio below one is lost fine-scale structure, and a ratio above one is spurious high-frequency content. The upper edge is the Nyquist wavenumber of the 100^2 grid on which the data was solved; above it the stored fields contain only interpolation artifact from the resize to 128^2. The lower edge is a fixed choice that excludes the large-scale structure every method reconstructs roughly correctly, which would otherwise dominate the ratio.

A.4 Deterministic Squared-Error Baseline

A model trained to minimize squared error against the ground truth converges to the posterior mean. Where the posterior is broad, that mean is a smoothed field rather than a plausible sample, and the standard distortion metrics favor it over a sample. Figure A1 shows the effect on the compressible Euler task.

The three deterministic reconstructions lose the fine structure of the field. The vortex roll-ups along the interface, magnified in the detail row, are reproduced by conditional diffusion and by our method and smoothed away by FNO, PINO and the regression model. The physics residual row shows the same failure in another form. A sharp reconstruction leaves a structured residual that follows the interface, as the ground truth's own residual does, while the smoothed reconstructions leave a

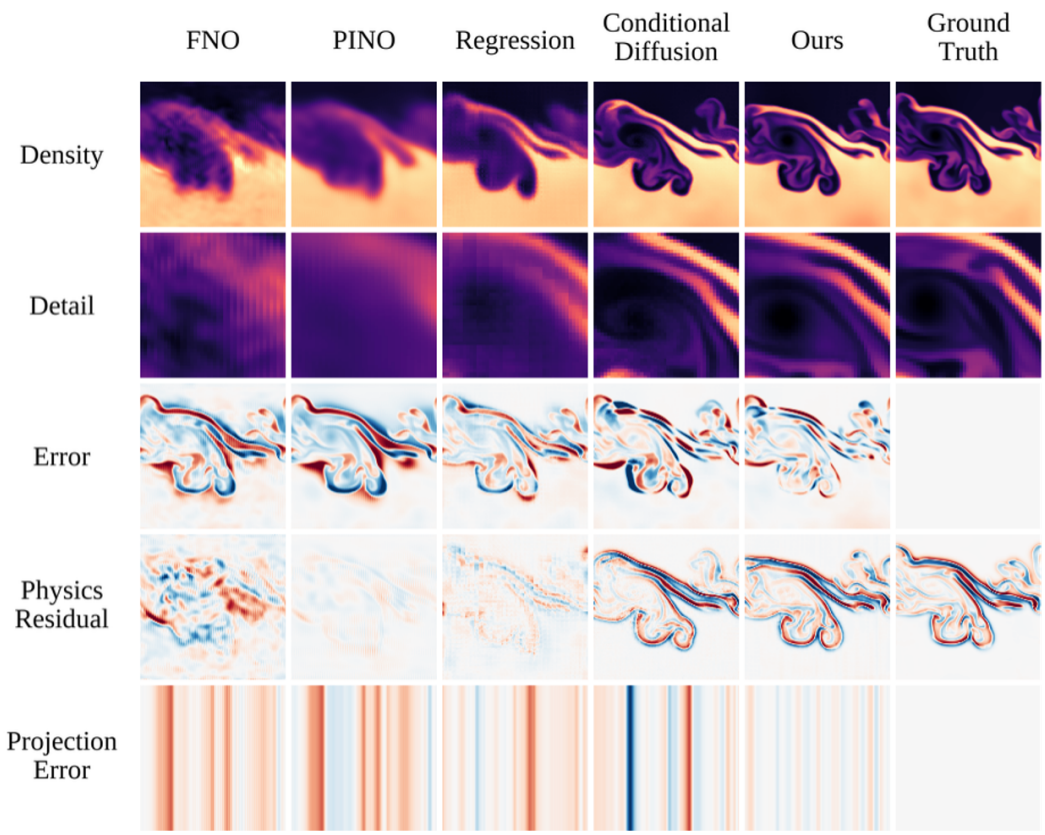

Figure A1: **Squared-error training smooths the field.** One evaluation item from the compressible Euler task. FNO, PINO and a regression model trained on our backbone, our data and our conditioning are deterministic; conditional diffusion and our method are samplers. The detail row magnifies one vortex roll-up. The physics residual row is the strong-form Euler residual of each reconstruction, and the ground truth column gives the residual of the true field, which is not zero.

Table A3: **Comparison against a regression baseline**, evaluated on sparse-view CT over 5 items. The regression model is deterministic and has no sample spread, so its single value stands against both of ours. *Single sample* is one ordinary sample. Bold indicates the best performance and underline the second best.

Method	NRMSE↓	PSNR↑	SSIM↑	Obs↓
Regression	<u>0.2490</u>	<u>20.45</u>	0.6130	<u>0.0545</u>
Ours, single sample	0.2496	20.40	<u>0.6241</u>	0.0562
Ours, mean of 20	**0.1945**	**22.67**	**0.6744**	**0.0357**

residual that is weak and diffuse, because a field with small derivatives cannot violate the equations by much.

None of this appears in the accuracy columns. Over the 100 evaluation items of this task, the regression model attains the best NRMSE, PSNR and SSIM of any method we run, 0.114, 31.53 and 0.926, against our 0.133, 30.56 and 0.921 (Table 2 lists regression over five items, where its NRMSE is 0.117). Those metrics measure squared error, which the regression model is trained to minimize, and on a broad posterior the conditional mean is the minimizer.

Table A3 gives the quantitative counterpart on sparse-view CT. A single sample from our method matches the regression point estimate on NRMSE, 0.2496 against 0.2490, since a sample does not minimize squared error. Averaging twenty of our samples estimates the posterior mean, and that es-

Table A4: **Residual ablation on the compressible Euler task.** Bold indicates the best performance and underline the second best. Each row is a separate training run. Every row samples with the plain sampler, without the restart schedule and temporal score rescaling of Section 4.1, so that the residual is the only difference between rows. The no-refine-pass row is the same model and configuration as the conditional diffusion row of Table 2.

Variant	NRMSE↓	PSNR↑	SSIM↑	Obs↓
No refine pass (= cond. diffusion)	0.1741	27.95	0.8878	0.0813
Without any residual	0.1674	28.38	0.8907	0.0819
Without observation residual	0.1457	29.73	0.9100	0.0590
Without projection residual	0.1549	29.14	0.9030	0.0601
Without physics residual	<u>0.1409</u>	<u>30.03</u>	<u>0.9128</u>	<u>0.0523</u>
All residuals (ours)	**0.1400**	**30.10**	**0.9145**	**0.0504**

timate is 22% more accurate than the regression model's, 0.1945 against 0.2490. On these five items, our method therefore provides a more accurate estimate of the posterior mean than the regression model at the cost of additional samples, while its individual samples remain plausible fields.

A.5 Residual Ablation

This ablation measures the contribution of each residual type. Table A4 varies which of them the refine pass reads, with a separate training run per row, all sampled without the restart schedule and temporal score rescaling so that the residual is the only difference between rows. Against no feedback, our method improves NRMSE by 20%. Every residual type contributes to that margin, and the two measurement residuals matter most: removing the projection residual costs 11% NRMSE, the observation residual 4%, and the physics residual 0.6%. Appendix A.7 ablates the sampler components.

A.6 Subspace Decomposition of the Correction

The refine pass reads a residual that carries information only about the directions the measurement constrains, and produces a correction over the whole field. Accuracy numbers cannot tell whether that correction is a learned data-consistency step, such as a gradient step on the data-fidelity term $\frac{1}{2}\|A\hat{x}_0 - y\|_2^2$, which lies in those same directions, or a correction that extends into the directions the measurement cannot see. The decomposition below separates the two on a trained checkpoint, with no retraining.

Setup. At the four-view CT geometry of Section 5 the forward operator is linear with rank 510 against $16{,}384$ unknowns, and the field space splits orthogonally as $\mathbb{R}^n = \mathrm{row}(A) \oplus \mathrm{null}(A)$. We build A as a dense matrix from the forward map and take $P = A^{+}A$ as the projector onto $\mathrm{row}(A)$; truncating at rank 450 instead moves every number below by less than 0.01. The visible fraction f_{row} of a vector is the fraction of its energy in $\mathrm{row}(A)$. At each sampling step we record the correction $\Delta = \hat{x}_0 - \hat{x}_0^{\text{pred}}$ and a control Δ_{shuf}, computed at the same state with the residual channels replaced by another sample's. Their difference $\Delta_{\text{attr}} = \Delta - \Delta_{\text{shuf}}$ is the part of the correction driven by the content of the residual rather than by a second pass conditioned on the first. We write $\varepsilon = \hat{x}_0^{\text{pred}} - x$ for the proposal's error. All numbers are over 100 evaluation samples, energy-weighted along the trajectory, under the plain sampler; every ordering is the same under the configuration of Section 4.1.

Reference levels. A likelihood gradient $A^{\mathsf{T}}(A\hat{x}_0 - y)$ and a projection step $A^{+}(y - A\hat{x}_0)$ both lie in $\mathrm{row}(A)$, so any data-consistency correction has a visible fraction of exactly 1.0, and that is the reference we compare against. The expected visible fraction of a direction drawn isotropically at random, $\mathrm{rank}(A)/n = 510/16{,}384 = 0.031$, is not a useful reference: four line-integral views see smooth structure far better than a random direction, and image-shaped differences with no model involved, such as a slice minus its blur or one patient's slice minus another's, have visible fractions between 0.09 and 0.50.

Results. Swapping the residual for another sample's changes the correction by more than its own magnitude, $\|\Delta_{\text{attr}}\|/\|\Delta\| = 1.20$, so the refine pass responds to the content of the residual rather

than to the presence of an input. The residual itself lies in $\mathrm{row}(A)$, at a visible fraction of 0.999, but the correction it drives does not: $f_{\text{row}}(\Delta_{\text{attr}}) = 0.155$, so 85% of its energy lies in $\mathrm{null}(A)$, where the likelihood is constant. That null-space component has a cosine of $+0.18$ with the negated proposal error $-\varepsilon$, so it moves unmeasured content toward the ground truth; the alignment is concentrated at high noise levels and falls to zero as $\sigma \to 0$. DPS's guidance direction, by comparison, is 70% row-space, short of 1.0 only because the gradient is carried back through the denoiser Jacobian. Finally, the proposal's error has a visible fraction of 0.029, just below that of an isotropic random direction: almost everything the proposal gets wrong is invisible to the measurement, out of reach of any step confined to $\mathrm{row}(A)$.

Limitations. This is an attribution measurement and not a proof of mechanism: Δ_{attr} does not isolate a causal pathway inside the network. It is also one task, and the binary split is the right model only for a linear operator.

A.7 Sampler Ablation

Two training-free modifications are applied to the sampler of Section 4.1: a restart schedule and temporal score rescaling. Table A5 varies both on a single Kelvin–Helmholtz checkpoint, so no row here requires retraining. Restarts carry most of the effect and temporal score rescaling adds little on its own, but the two together are best on four of the five columns.

Table A5: **Sampler ablation on Kelvin–Helmholtz.** Bold indicates the best performance and underline the second best. All four rows are the same checkpoint under different sampler configurations. $E_{16\text{–}64}/E_{\text{gt}}$ is the high-band spectral energy ratio of Appendix A.3, computed on density over the evaluation split; its target is 1.0. The band extends to 64 cycles per domain, the Nyquist wavenumber of the 128^2 grid on which this task is solved.

Variant	NRMSE↓	PSNR↑	SSIM↑	Obs↓	$E_{16\text{–}64}/E_{\text{gt}}$
Neither	0.1397	26.96	0.8957	0.0094	0.958
Restart only	<u>0.1295</u>	<u>27.68</u>	<u>0.9059</u>	0.0093	<u>0.974</u>
TSR only	0.1394	26.99	0.8984	**0.0069**	0.967
Restart and TSR	**0.1289**	**27.73**	**0.9090**	<u>0.0070</u>	**0.982**

A.8 Robustness to Measurement Degradation

Two properties of the measurement limit how well a field can be reconstructed: the number of readings taken, and the noise on each reading. A harder acquisition usually degrades both together, so the two cannot be separated within a single benchmark setting. We vary them independently on sparse-view CT, holding everything else fixed, and report for each method its accuracy at every setting together with the change in its margin over our method along the axis. A margin that stays flat along an axis indicates that two methods degrade at the same rate along it; a margin that widens indicates that they do not.

Setup. The field is a 128^2 attenuation slice measured by a parallel-beam Radon transform with 128 detector bins, so there are $16{,}384$ unknowns throughout. Measurement noise is the Gaussian approximation to log-transformed Poisson counts, with per-ray variance $\exp(p)/I_0$ for a line integral p, so the noise on each ray varies with the attenuation along that ray. We train on $100{,}000$ slices and evaluate on 100 held-out slices. The conditioning is two image channels at every setting, an FBP reconstruction and an unfiltered backprojection, so the input shape is constant along both axes and a single checkpoint serves an entire range. We train one model of ours and one conditional diffusion model per experiment for $300{,}000$ steps across that experiment's full range, without the sampler modifications of Section 4.1.

Baselines. DPS requires no training, since its prior is unconditional and one checkpoint remains valid at any geometry or dose. Conditional diffusion is amortized and receives conditioning identical to ours, so the difference between conditional diffusion and our method isolates the refine pass, while the difference between conditional diffusion and DPS isolates amortization. DPS's guidance step size determines how far the sampler trusts the measurement, and the default value is poorly suited to this task, so we tune it per setting.

Table A6: **Variable-view sparse-view CT at fixed dose.** Bold indicates the best performance and underline the second best, ranked within each view count. DPS runs at a step size tuned separately for each view count.

	DPS				Cond. diffusion				Ours			
Views	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓
4	<u>0.2891</u>	<u>18.75</u>	<u>0.5089</u>	<u>0.0659</u>	0.2958	18.57	0.4929	0.0744	**0.2772**	**19.16**	**0.5257**	**0.0597**
16	<u>0.2043</u>	<u>21.86</u>	<u>0.6251</u>	<u>0.0497</u>	0.2084	21.68	0.6146	0.0571	**0.1918**	**22.43**	**0.6393**	**0.0487**
64	0.1683	23.61	0.6795	<u>0.0381</u>	<u>0.1628</u>	<u>23.90</u>	<u>0.6841</u>	0.0405	**0.1557**	**24.33**	**0.6967**	**0.0369**
256	0.1418	25.15	0.7290	<u>0.0283</u>	<u>0.1370</u>	<u>25.49</u>	<u>0.7334</u>	0.0311	**0.1286**	**26.09**	**0.7485**	**0.0272**
1024	0.1265	26.24	0.7634	0.0300	<u>0.1196</u>	<u>26.74</u>	<u>0.7761</u>	<u>0.0247</u>	**0.1062**	**27.82**	**0.8011**	**0.0194**

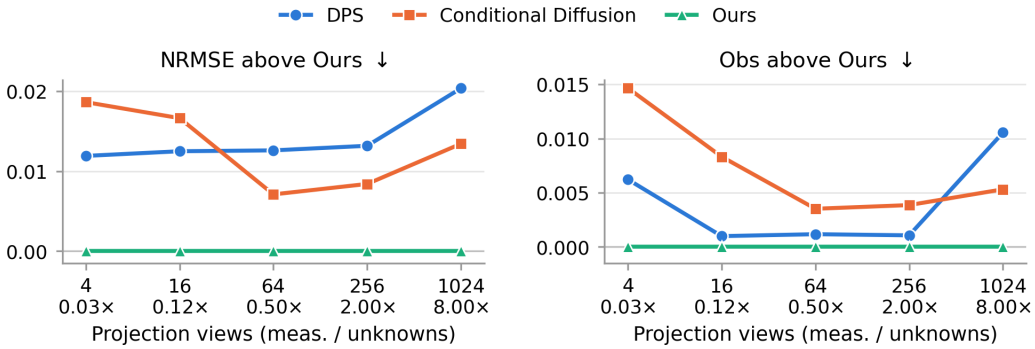

Figure A2: **Varying the number of projections.** Each baseline's error minus ours, in the metric's own units, so our method is the zero line and the vertical axis gives the margin directly. The margin over DPS is close to constant across the under-determined range and widens only at $8\times$ oversampling. The margin over conditional diffusion is U-shaped, largest at both ends of the range and smallest near the determined boundary.

A.8.1 Measurement quantity

We vary the projection count over powers of four, $V \in \{4, 16, 64, 256, 1024\}$, at a fixed photon count of $I_0 = 4096$. At 128 detector bins the problem is exactly determined at 128 views, so the range runs from $32\times$ under-determined to $8\times$ over-determined, at measurement-to-unknown ratios of 0.03, 0.12, 0.50, 2.00 and 8.00. Table A6 reports accuracy at every view count.

The margin over DPS is a near-constant offset that widens only under heavy oversampling. In NRMSE it lies between 0.0119 and 0.0132 from 4 views to 256, moving by 11% across a $64\times$ change in the amount of measurement, and then rises by 55% to 0.0204 at 1024 views. The flat portion already includes 256 views, which is twice over-determined, so the rise is specific to the $8\times$ setting.

The margin over conditional diffusion is U-shaped. In NRMSE it runs 0.0186, 0.0166, 0.0071, 0.0084 and 0.0134, with its minimum at 64 views, the setting closest to exactly determined, and observation conformance follows the same shape. The refine pass therefore contributes most where measurement is scarce and where it is abundant, and least near the determined boundary. At the two sparsest settings conditional diffusion is less accurate than DPS, by 2.3% at 4 views and 2.0% at 16, although it is more accurate than DPS at every dose level of the second experiment.

We attribute the two arms of that curve to different causes, both related to what the conditioning image carries. At 4 views the FBP reconstruction and the backprojection are dominated by streak artifacts, so a model conditioned on them likely has little usable information and no way to check its own output against the measurement; evaluating $A(\hat{x}_0^{\text{pred}})$ supplies what those images lack, and the refine pass is worth 6 to 8% there. Near the determined boundary the conditioning image is already close to the field, so a residual has little left to add, and the margin narrows to 4.4%. At 1024 views

Table A7: **Variable-dose sparse-view CT at fixed geometry.** Photon counts of $I_0 = 4096$, 512 and 64 give sinogram signal-to-noise ratios of 29.1, 20.0 and 11.0 dB. Bold indicates the best performance and underline the second best, ranked within each dose. DPS runs at a step size tuned separately for each dose.

	DPS				Cond. diffusion				Ours			
I_0	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓	NRMSE ↓	PSNR ↑	SSIM ↑	Obs ↓
4096	0.1683	23.61	0.6795	0.0381	0.1649	23.78	0.6811	0.0409	**0.1578**	**24.20**	**0.6932**	**0.0373**
512	0.2162	21.33	0.6072	0.0614	0.2042	21.83	0.6203	0.0607	**0.2005**	**22.00**	**0.6286**	**0.0582**
64	0.2910	18.67	0.5035	0.1036	0.2558	19.81	0.5546	0.0933	**0.2540**	**19.87**	**0.5603**	**0.0910**

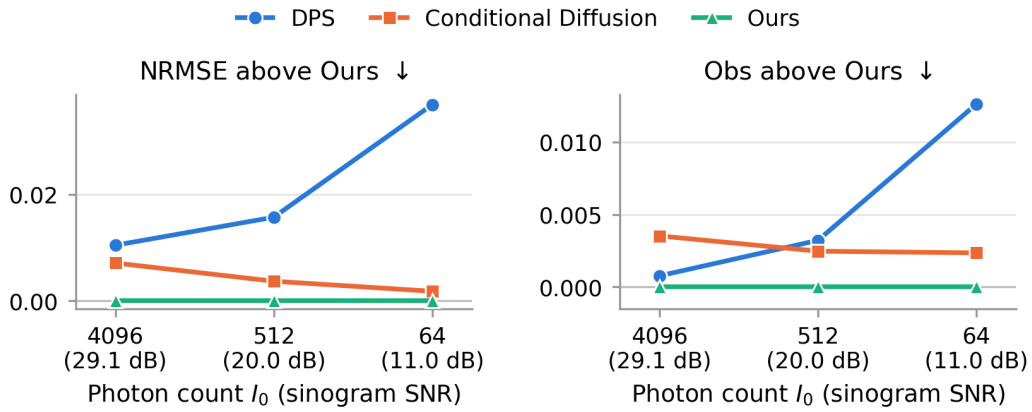

Figure A3: **Varying the photon count.** Plotted as in Figure A2. The two baselines move in opposite directions as the measurement is corrupted: DPS falls further behind our method, while the conditional diffusion margin narrows.

the conditioning image is close to the field as well, but the residual is now measured in 131,072 numbers rather than 512 and can localize the remaining error precisely, and the refine pass is worth 11%. On this account, the refine pass is most useful where the conditioning is weakest and where the residual is most informative, and least useful in the middle, where neither holds.

DPS is the only method whose observation conformance degrades as it is given more measurements. Its values run 0.0659, 0.0497, 0.0381, 0.0283 and 0.0300, improving monotonically to 256 views and then worsening with four times more data, while both amortized methods improve monotonically over the whole range.

A.8.2 Measurement noise

We hold the geometry at 64 parallel-beam views over 180°, giving 8,192 measurements for 16,384 unknowns, and vary the photon count I_0 over 4096, 512 and 64, drawn per sample. Table A7 reports accuracy at every dose.

The two baselines move in opposite directions (Figure A3). As the measurement is corrupted, DPS falls further behind our method on both metrics while the conditional diffusion margin narrows. The NRMSE margin over DPS grows by a factor of 3.5, from 0.0105 to 0.0369, and the observation conformance margin by a factor of 15, from 0.0008 to 0.0126. Over the same range the conditional diffusion NRMSE margin shrinks from 0.0071 to 0.0018, which is within 0.7% of our error.

On observation conformance the two baselines cross. At $I_0 = 4096$, the lowest noise level, DPS is the better of them, 0.0008 above us against conditional diffusion's 0.0035, as expected of a method whose guidance term optimizes that quantity directly. By $I_0 = 512$ the order has reversed, and at $I_0 = 64$ DPS's observation margin over our method is five times conditional diffusion's, 0.0126

against 0.0023. DPS, whose guidance term targets measurement fidelity, has the worst observation conformance of the three once the photon count drops to 512.

A.8.3 Joint analysis

Amortization accounts for robustness to noise, and the refine pass accounts for robustness to scarcity. The refine pass's contribution falls to 0.7% under heavy noise but remains between 4.4% and 11.2% at every projection count, and conditional diffusion is less accurate than DPS at the two sparsest settings while remaining more accurate at every dose level.

Our margin over DPS is flat to within 11% across a $64\times$ change in projection count and grows by a factor of 3.5 across a $64\times$ change in dose. Our method runs at 2.0 to 5.5 s per sample against DPS's 43 to 136 s, with no backward passes through the denoiser and no operator gradients.

A.9 Additional Qualitative Results

Figure A4 shows one reconstruction per method on Kelvin–Helmholtz and Rayleigh–Taylor (compressible Euler is in Figure 4), with the reconstruction error, the PDE residual and the projection error below each field. Figure A5 shows the same for the three imaging problems, with the reconstruction error and the observation residual.

B Method Details

B.1 Training and Sampling Pseudocode

Algorithms 1 and 2 give the training step and the sampler of our method as implemented. The predict and refine passes are evaluations of the same network D_θ, whose conditioning input is the channel stack $[\,\mathcal{C}(y) \,\|\, \hat{x}_0^{\text{pred}} \,\|\, e \,\|\, b_{\text{refine}}\,]$, with the flag b_{refine} broadcast as a constant channel. In predict mode the proposal and residual slots are filled with zeros and the flag is 0; in refine mode they hold the proposal, its residual and a flag of 1. The input width is therefore the same in both modes, and the refine pass adds no parameters beyond these input channels.

Algorithm 1 Training an error-conditioned generative solver

Require: paired training fields x, forward operator A and noise model, conditioning map $\mathcal{C}$, error operator $\mathcal{E}$, denoiser D_θ, training noise density $p(\sigma)$, σ_{data}, EMA decay β (half-life of 5×10^5 training examples)

repeat
  $x \sim p_{\text{data}}, \quad y \leftarrow Ax + \eta$ ▷ fresh measurement noise every step
  $\sigma \sim p(\sigma), \quad \epsilon \sim \mathcal{N}(0, I), \quad x_\sigma \leftarrow x + \sigma\epsilon$
  $\hat{x}_0^{\text{pred}} \leftarrow D_\theta(x_\sigma, \sigma, \mathcal{C}(y);\ b_{\text{refine}} = \text{FALSE})$ ▷ predict
  $e \leftarrow \mathcal{E}(\text{sg}(\hat{x}_0^{\text{pred}}); y)$ ▷ evaluate, no gradient
  $\hat{x}_0 \leftarrow D_\theta(x_\sigma, \sigma, \mathcal{C}(y), \text{sg}(\hat{x}_0^{\text{pred}}), e;\ b_{\text{refine}} = \text{TRUE})$ ▷ refine
  $\mathcal{L} \leftarrow w(\sigma)\big(\|\hat{x}_0 - x\|_2^2 + \|\hat{x}_0^{\text{pred}} - x\|_2^2\big), \quad w(\sigma) = 1/(\sigma^2 + \sigma_{\text{data}}^2)$ ▷ equation 10
  $\theta \leftarrow \text{AdamW}\big(\theta, \text{ClipNorm}(\nabla_\theta \mathcal{L}, 1)\big)$
  $\bar{\theta} \leftarrow \beta\bar{\theta} + (1-\beta)\,\theta$ ▷ EMA weights, used for all evaluation
until 300,000 steps
return $D_{\bar{\theta}}$

B.2 Inference Configurations

Every sample of our method is drawn with one sampler configuration, shared by all six tasks. We run the predict–evaluate–refine step of Section 4 inside the second-order DPM-Solver++ (Lu et al., 2025) for 100 steps (Algorithm 2), so a sample costs 200 backbone passes, and add a restart schedule and temporal score rescaling on top. Both modifications are training-free.

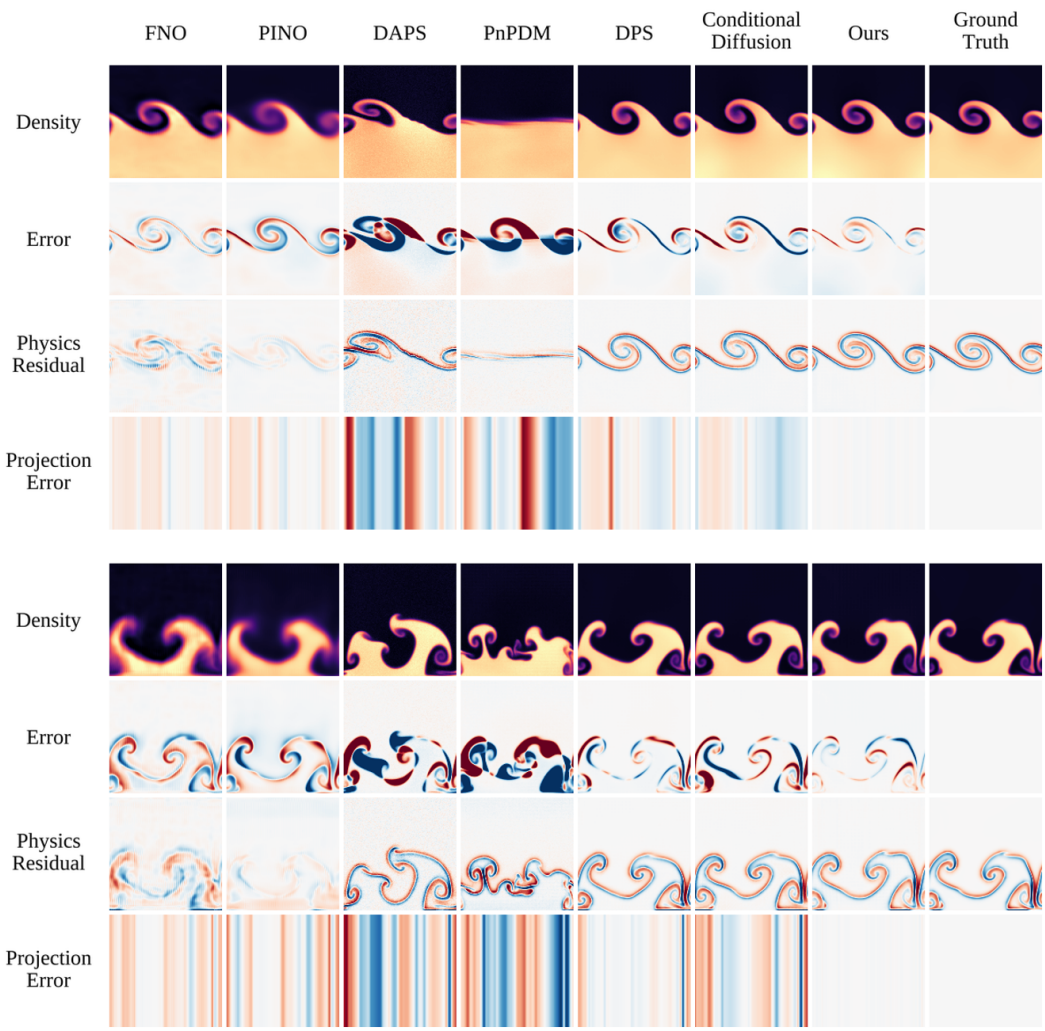

Figure A4: **Qualitative results on the two instability tasks.** One held-out item per task, Kelvin–Helmholtz (top) and Rayleigh–Taylor (bottom), with every method on the same item. Each shows the reconstructed density, its error against the ground truth, the strong-form physics residual, and the projection error against the measured radiographs. Compressible Euler is shown in Figure 4.

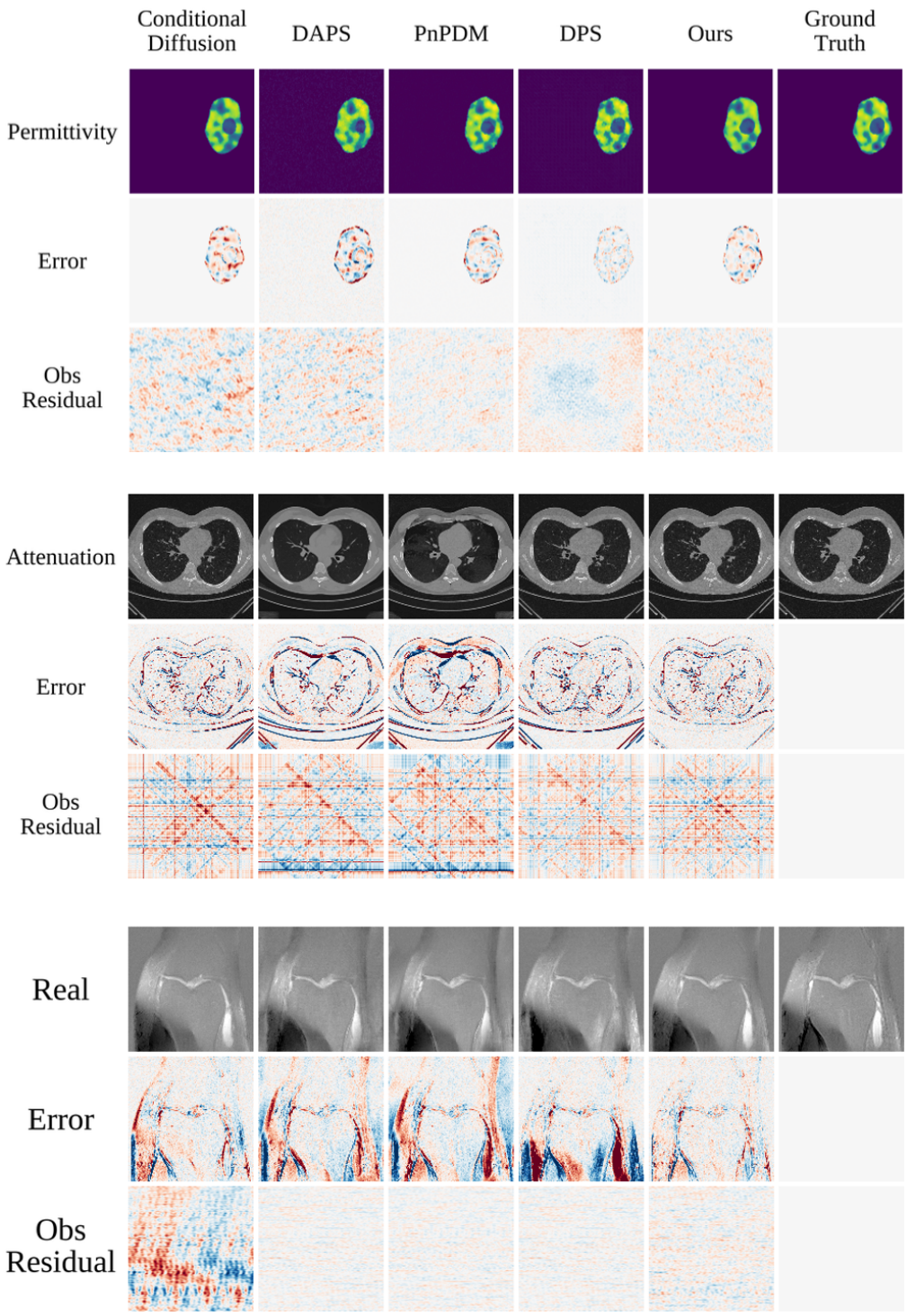

Figure A5: **Qualitative results on the three imaging tasks.** Tasks are inverse scattering (permittivity contrast), sparse-view CT (attenuation) and CS-MRI (real part of the knee image); each shows the reconstruction, its error against the ground truth, and the observation residual.

Algorithm 2 Sampling with an error-conditioned generative solver

Require: measurement y, conditioning map $\mathcal{C}$, error operator $\mathcal{E}$, denoiser D_θ, noise levels $\sigma_T > \cdots > \sigma_0 = 0$

$x_{\sigma_T} \sim \mathcal{N}(0, \sigma_T^2 I), \quad \hat{x}_0^{\text{prev}} \leftarrow \varnothing$
for $t = T, \ldots, 1$ **do**
  $\hat{x}_0^{\text{pred}} \leftarrow D_\theta(x_{\sigma_t}, \sigma_t, \mathcal{C}(y);\ b_{\text{refine}} = \text{FALSE})$ ▷ predict
  $e \leftarrow \mathcal{E}(\hat{x}_0^{\text{pred}}; y)$ ▷ evaluate
  $\hat{x}_0 \leftarrow D_\theta(x_{\sigma_t}, \sigma_t, \mathcal{C}(y), \hat{x}_0^{\text{pred}}, e;\ b_{\text{refine}} = \text{TRUE})$ ▷ refine
  if $\sigma_{t-1} = 0$ **then**
    $x_{\sigma_{t-1}} \leftarrow \hat{x}_0$
  else ▷ DPM-Solver++(2M) step
    $h \leftarrow \log(\sigma_t / \sigma_{t-1})$
    $x_{\sigma_{t-1}} \leftarrow (\sigma_{t-1}/\sigma_t)\, x_{\sigma_t} + (1 - e^{-h})\, \hat{x}_0$
    if $\hat{x}_0^{\text{prev}} \neq \varnothing$ **then**
      $x_{\sigma_{t-1}} \leftarrow x_{\sigma_{t-1}} + \frac{h}{h_{\text{prev}}}\big(1 - \frac{1-e^{-h}}{h}\big)\big(\hat{x}_0 - \hat{x}_0^{\text{prev}}\big)$
    end if
    $h_{\text{prev}} \leftarrow h$
  end if
  $\hat{x}_0^{\text{prev}} \leftarrow \hat{x}_0$
end for
return x_{σ_0}

Restart schedule. Instead of descending the ramp monotonically, the sampler periodically returns to a higher noise level by a forward-noising jump, $x \leftarrow x + \sqrt{\sigma_{\text{hi}}^2 - \sigma_{\text{lo}}^2}\,\epsilon$. This is similar to Restart Sampling (Xu et al., 2023), and resampling schemes of the same shape appear in inpainting (Lugmayr et al., 2022). Our schedule restarts one band, $\sigma \in [0.2, 3.0]$, eight times, and the node spacing is solved so that the schedule costs the same 100 steps as the monotone ramp: a jump is a noise injection and costs no network evaluation, so the restart and non-restart configurations use the same number of network evaluations.

Temporal score rescaling. We apply the rescaling of Xu et al. (2026), which multiplies the predicted noise by a time-dependent ratio and is equivalent, in clean-field space, to $x_\sigma + \rho(\sigma)\,(\hat{x}_0 - x_\sigma)$ with

$$\rho(\sigma) = \frac{\text{SNR}(\sigma)\, s^2 + 1}{\text{SNR}(\sigma)\, s^2/k + 1}, \qquad \text{SNR}(\sigma) = \sigma_{\text{data}}^2/\sigma^2. \tag{11}$$

We use $k = 1.5$ and $s = 1$. Values above one sharpen the prediction, $k = 1$ recovers the unmodified sampler, and $\rho \to k$ as $\sigma \to 0$ whatever s is. Appendix A.7 ablates the restart schedule and this rescaling.

B.3 PDE Residual

On the three flow tasks the error operator $\mathcal{E}$ of Section 4 returns a strong-form residual of the governing equations alongside the two measurement residuals.

With $\boldsymbol{q}$ the conservative state and $\boldsymbol{F}, \boldsymbol{G}$ its fluxes along x and y, the residual is the strong form of the governing equations, evaluated by central differences in space and a midpoint rule in time,

$$e_{\text{pde}}[n] = \frac{\boldsymbol{q}[n+1] - \boldsymbol{q}[n]}{\Delta t} + \tfrac{1}{2}\Big(\partial_x \boldsymbol{F}[n+1] + \partial_x \boldsymbol{F}[n]\Big) + \tfrac{1}{2}\Big(\partial_y \boldsymbol{G}[n+1] + \partial_y \boldsymbol{G}[n]\Big) - \boldsymbol{S}, \tag{12}$$

with $\partial_x \boldsymbol{F} \approx (\boldsymbol{F}_{i+1} - \boldsymbol{F}_{i-1})/2\Delta x$ and likewise in y, and $\boldsymbol{S}$ the source term of the task, which is zero for compressible Euler and carries the viscous, conduction and diffusion terms for the two instability tasks. The residual of the transition out of frame n is placed at frame n, and the last frame of the window is left at zero.

This residual is fast and approximate by design. It is one finite-difference pass rather than a solver rollout, which is what makes it cheap enough to evaluate at every sampling step. As a consequence, it does not vanish on the true field: a finite-difference stencil on a finite grid leaves a nonzero residual wherever the flow varies over a few cells, so the ground truth itself scores nonzero (Figure 4).

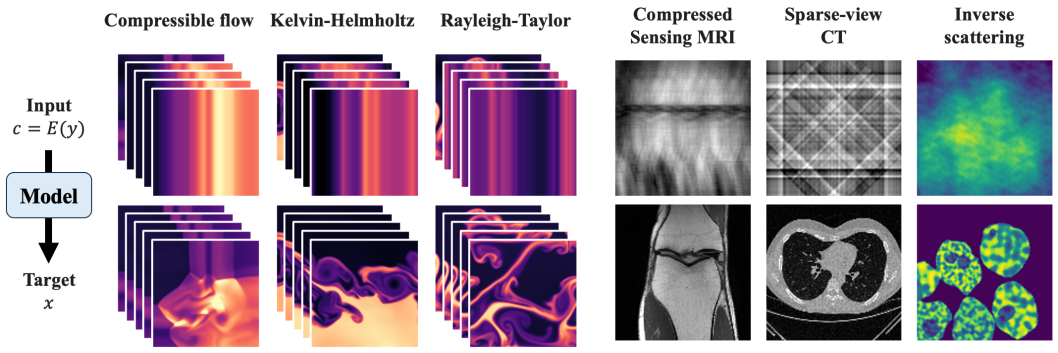

Figure A6: **The six tasks.** Every task gives the model a conditioning input $c = \mathcal{C}(y)$ formed from the measurement (top) and asks for the field x (bottom). The three flow tasks are trajectories, so their inputs and targets are stacks of frames, while the three imaging tasks are single fields.

C Benchmark Details

Section 5 defines each forward operator, its noise model and its data source. We continue that description here with how the data was generated and split.

Flow data. The three flow datasets are generated for this work with Clawpack (Clawpack Development Team, 2026). Each trajectory is simulated for 100 frames on the unit square and stored on the 128^2 grid of Section 5, and we use random 20-frame windows as the domain for the task. Every sample has two independent random inputs: a perturbation of the initial condition, which is visible in the first frame, and a body force, which is not.

Compressible Euler. The initial condition is the four-quadrant two-dimensional Riemann problem of the Clawpack `quadrants` example, with $\gamma = 1.4$ and the four quadrants meeting at $(0.8, 0.8)$. The primitive states (ρ, u, v, p) are $(1.5, 0, 0, 1.5)$ in the upper right quadrant, $(0.5323, 1.206, 0, 0.3)$ in the upper left, $(0.1380, 1.206, 1.206, 0.029)$ in the lower left and $(0.5323, 0, 1.206, 0.3)$ in the lower right. In each sample, a fractal noise field of amplitude 0.15 is added to the density, and the velocity and pressure are left unchanged. The noise is two octaves of Perlin noise, the coarser with two lattice periods across the domain. Trajectories run to $t = 0.8$ on a 100^2 grid with extrapolation boundaries and are resized to 128^2.

Kelvin–Helmholtz. Two streams with a velocity jump of 1 meet at a single interface at $y = 0.5$, with density 1 below and 4 above and a uniform pressure of 2.5, in a domain periodic in x with free-slip walls in y. Density, streamwise velocity and tracer share one $\tanh$ profile of width 0.025 across the interface, and the tracer is 1 in the lower fluid. The velocity is expressed in the frame of the density-weighted mean velocity, in which the billows are stationary. The instability is seeded with a transverse velocity $v = 0.01\,[\sin(4\pi x) + 0.5\sin(2\pi x)]\,\exp\!\big(-((y - 0.5)/0.15)^2\big)$: two primary wavelengths and their subharmonic at a fixed phase, so that the two billows pair. The Gaussian envelope is negligible at the walls. In each sample, a random combination of the four longest periodic modes $\sin(2\pi k x + \phi_k)$, $k = 1, \ldots, 4$, is added to the bracketed profile, with Gaussian amplitudes scaled by $1/k$, uniform phases ϕ_k, and a peak of 0.15 after normalization. The flow is compressible Navier–Stokes with $\gamma = 5/3$, $\mathrm{Re} = 12{,}800$ and $\mathrm{Pr} = 2/3$, where Re is based on the lower density, the velocity jump and the primary wavelength $\lambda = 0.5$. Trajectories run for 25 linear e-folding times ($t \approx 4.97$) on a 128^2 grid.

Rayleigh–Taylor. A heavy fluid of density 3 lies above a light fluid of density 1 under gravity $g = 1$ along $-y$, at rest and in discrete hydrostatic balance with a pressure of 10 at the interface. The domain is periodic in x with free-slip walls in y. The interface lies at $y = 0.75 + 0.01\cos(4\pi x)$, two wavelengths with an amplitude of two percent of the wavelength, smoothed by a $\tanh$ profile of width 0.01. In each sample, the same four-mode random perturbation as in Kelvin–Helmholtz, in cosines, is added to the interface displacement, with a peak of 0.15 times the seeded amplitude. The equations and γ are those of Kelvin–Helmholtz, with $\mathrm{Re} = \sqrt{g\lambda}\,\lambda/\nu = 6{,}400$ for $\lambda = 0.5$ and $\mathrm{Pr} = 2/3$. Trajectories run for 10 linear e-folding times ($t \approx 3.99$) on a 256^2 grid and are area-averaged to 128^2.

Body force. The body force is an acceleration field $\boldsymbol{f}(\boldsymbol{x}, t)$ applied to the momentum and energy equations as a source term with Godunov splitting. Its two components are independent draws of three-dimensional fractal Perlin noise over (t, x, y), with two octaves at persistence $1/2$. The coarser octave has four lattice periods over the 100-frame trajectory and two across the domain in each spatial direction, and the finer octave doubles both. The noise is drawn on a $128 \times 32 \times 32$ lattice, resized in space to the simulation grid with a Lanczos filter and interpolated linearly in time, so the force is smooth in both space and time. The force is drawn independently of the initial-condition perturbation and scaled by an amplitude of 4.0 on compressible Euler, 0.23 on Kelvin–Helmholtz and 0.45 on Rayleigh–Taylor.

Splits. Every evaluation set contains one item per independent source, so 100 items are 100 independent samples. CS-MRI is split by volume, and we take the middle slice of each of 100 held-out knee volumes. Sparse-view CT is split by patient, with 20 slices per training patient and one slice from each of 100 held-out patients. Inverse scattering uses the held-out split published with InverseBench (Zheng et al., 2025). The flow tasks are split by trajectory, with evaluation trajectories generated from seeds disjoint from the training set.

D Experimental Setup

D.1 Model and Training Details

Every diffusion model in the paper uses the same backbone, the video transformer denoiser of Li et al. (2025) under the EDM preconditioner (Karras et al., 2022). It is a two-level hierarchy over space-time patches of size $2 \times 4 \times 4$: a high-resolution level of 2 blocks at width 384 with neighborhood attention over a $2 \times 7 \times 7$ window, and a low-resolution level of 11 blocks at width 768 with global attention, with a 1-block mapping network at width 768 for the noise-level and conditioning embedding. Only the input stem differs between methods and tasks, since the conditioning and residual channels of Section 5 change its width; the refine pass adds no parameters beyond those input channels. The three plug-and-play baselines share one unconditional model of the same architecture per task.

Training uses AdamW at learning rate 5×10^{-4}, $\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$ and weight decay 10^{-2}, with the mapping network at one third of the learning rate, a 1000-step linear warm-up and a constant rate thereafter. Noise levels are drawn from a cosine-interpolated density with $\sigma_{\mathrm{data}} = 0.5$, and the loss is weighted by the soft-min-SNR schedule of Section 4.1. Every run trains for 300,000 steps. Batch size is 16 on the single-frame imaging tasks and 2 on the 20-frame flow tasks.

Each model is trained on a single L40S GPU. A run takes roughly 30 hours on the single-frame imaging tasks and 60 to 80 hours on the 20-frame flow tasks.

D.2 Baseline Configurations

All three plug-and-play baselines share a single unconditional prior per task, trained under the EDM recipe on the same data as our method. DPS runs 1000 ancestral steps. DAPS runs 200 annealing steps, each with 5 diffusion steps and 100 Langevin steps. PnP-DM runs 100 annealing steps with a truncated partial scheduler and 40 Langevin steps; we measure its network forward count per run rather than deriving it from the configuration, because its inner scheduler truncates at the current annealing noise level.

Each plug-and-play method has hyperparameters setting the strength of its measurement term: the guidance step size for DPS, and the Langevin step size and temperature for DAPS and PnP-DM. Published values are set for the operators and measurement scales of the tasks they were introduced on, so we retune them per task. For DAPS and PnP-DM, the step size lr and temperature τ enter the sampler only through the effective likelihood step $\mathrm{lr}/2\tau^2$. For all three methods we select the hyperparameters by grid search, running each configuration on five items from the evaluation split and keeping the configuration with the lowest NRMSE.

Operator and physics-informed baselines. FNO and PINO receive the same conditioning channels as the diffusion models, together with a normalized space–time coordinate grid, and predict the full 20-frame window in one forward pass. Both are three-dimensional FNOs over time and space, with

$(8, 16, 16)$ Fourier modes, width 64 and four layers, trained with the optimizer, schedule, batch size and step count of Appendix D.1 and evaluated with EMA weights. FNO is trained with squared error alone. PINO adds the finite-difference PDE residual of Appendix B.3, the same residual our error operator returns, to the loss with weight 0.1. PINO+TTOP starts each evaluation item from a copy of the trained PINO weights and optimizes all of them for 300 Adam steps at learning rate 10^{-4}, on the PDE residual with weight 0.1 and on squared-error fits to the observed initial state and to the four radiographs, each with weight 10. The instance optimization of Li et al. (2024) uses the PDE residual alone; we add the measurement terms so that TTOP is given the same measurement as every other method. The PINN is a tanh MLP with six hidden layers of width 256 on (x, y, t), trained from scratch on each evaluation item for $15{,}000$ Adam steps at learning rate 10^{-3}. Its loss combines the strong-form residual, evaluated by automatic differentiation at $8{,}192$ random collocation points per step with weight 1, with the same initial-state and radiograph terms at weight 10.

D.3 Metric Definitions

Every metric below is computed over the 100-item evaluation split of each task.

NRMSE. We report the relative ℓ_2 error per channel, averaged over the C channels,

$$\mathrm{NRMSE} = \frac{1}{C} \sum_{c=1}^{C} \frac{\|\hat{x}_c - x_c\|_2}{\|x_c\|_2}. \tag{13}$$

PSNR and SSIM. PSNR is computed per channel, pooled over all frames and pixels of an item, and then averaged over channels, $\mathrm{PSNR} = \frac{1}{C} \sum_c 10 \log_{10} \left(R_c^2 / \mathrm{MSE}_c \right)$. SSIM is computed per channel and per frame, each frame treated as a single-channel image under the standard 11×11 Gaussian window ($\sigma = 1.5$, $k_1 = 0.01$, $k_2 = 0.03$), then averaged over frames and over channels.

Both require a stated dynamic range R_c, so we fix R_c per task and per channel as the $[0.5, 99.5]$ percentile span of the ground truth over that task's evaluation split.

Observation conformance. We measure consistency with the measurement by $\mathrm{Obs} = \|A(\hat{x}) - A(x)\|_2 / \|A(x)\|_2$, taken over the whole measurement. It is scored against the noise-free measurement $A(x)$ of the ground truth.

E Extended Related Work

Plug-and-play diffusion priors. Plug-and-play methods (Venkatakrishnan et al., 2013; Romano et al., 2017) pair an unconditional diffusion prior with a likelihood term applied at inference. Zheng et al. (2025) group them into guidance-based methods that add a likelihood gradient to the reverse update (Chung et al., 2023; Song et al., 2023; Kawar et al., 2022; Wang et al., 2023), variable-splitting methods that alternate denoising with a separate data-fidelity step (Zhang et al., 2025; Wu et al., 2024; Zhu et al., 2023), variational methods (Mardani et al., 2024), and sequential Monte Carlo methods (Dou & Song, 2024; Cardoso et al., 2024); the same construction is used to combine sparse observations into state trajectories (Rozet & Louppe, 2023). The strength of this family is that one prior serves any forward operator without retraining. The cost is that the measurement is treated as an objective, which requires differentiating the operator, and often the denoiser, at every step, and requires tuning how strongly it is enforced for each new problem. Gradient-free variants avoid the operator gradient by searching over an ensemble of candidate particles scored by forward evaluations alone (Dou et al., 2026), at the cost of many more operator calls.

Supervised reconstruction. A second family trains a reconstruction model on paired data for a fixed forward operator, so that no optimization is solved at test time. Adler & Öktem (2018) introduced the generative form with a conditional GAN, and Palette (Saharia et al., 2022) established the conditional diffusion form now used throughout image-to-image reconstruction (Saharia et al., 2023); later work distills a diffusion prior into an amortized sampler (Mammadov et al., 2024) or regularizes a conditional GAN to preserve sample spread (Bendel et al., 2023). Our method belongs to this family, and what separates it from the rest is that none of these models apply the forward operator at inference: the measurement enters only as conditioning, and the network is never shown how its own output looks through the operator.

Operator learning and physics-informed solving. Closely related architectures appear in PDE surrogate modeling, where the learned map produces a solution field that is expected to satisfy the governing equations rather than a measurement. Neural operators learn a solution map between function spaces (Li et al., 2021; Lu et al., 2021; Kovachki et al., 2023), and foundation models extend this to multi-PDE pretraining (Herde et al., 2024); both are physics-blind at inference. PINNs (Raissi et al., 2019) impose the governing equations on a single instance by optimizing a network per problem, and PINO (Li et al., 2024) combines an operator-level physics loss during training with per-instance fine-tuning at test time. Methods that bring the residual into generation often use it as an optimization signal, either as a gradient that steers the reverse update (Huang et al., 2024) or as a Gauss–Newton projection onto the constraint manifold at each integration step (Utkarsh et al., 2025). Others supply it to the network instead: Shu et al. (2023) encode the gradient of the PDE residual as classifier-free conditioning for a flow reconstruction model, PRISMA (Sawhney et al., 2025) modulates the spectral features of a diffusion neural operator by the PDE residual, and Jiang et al. (2026) condition a corrector on that residual; Section 2.2 states the latter formulation. All three read a residual of the governing equations evaluated on the state itself, rather than a measurement residual formed through a forward operator. We draw several of our baselines from this literature.

Learned iterative reconstruction. Before diffusion priors, the standard way to give a network access to the measurement during reconstruction was to unroll an optimization algorithm and learn its steps (Monga et al., 2021). Recurrent inference machines (Putzky & Welling, 2017) supply the gradient of the log-likelihood, which for a linear operator is the backprojected measurement residual, to a weight-shared recurrent network at every iteration, and learned primal-dual (Adler & Öktem, 2018), MoDL (Aggarwal et al., 2019) and the deep cascade line (Schlemper et al., 2018; Hammernik et al., 2018; Sriram et al., 2020) interleave learned blocks with data-consistency steps. These are all point estimators: none is noise-conditioned and none samples, and they are trained by backpropagating through the unrolled chain rather than by a per-step denoising loss.