Scaling Test-Time Compute of Diffusion-Based Posterior Samplers for Physical Inverse Problems

Eric Hou
Independent Researcher
eric.x.hou@gmail.com

Haoxuan Chen
Stanford University
haoxuanc@stanford.edu

Abstract

For physical inverse problems, pretrained diffusion models commonly provide priors for posterior samplers whose inference cost is often fixed. Motivated by the recent success of scaling test-time compute in generative AI, particularly for large language models (LLMs) and diffusion models, we investigate whether the same recipe can be applied to diffusion-based solvers of inverse problems in physical sciences. Our methodology here is based on the Approximation-Free Diffusion Posterior Sampler (AFDPS) developed in existing work [4], where multiple trajectories are sampled in a parallel way under the sequential Monte Carlo (SMC) framework. We tested our method on three problems from InverseBench [43]: linear inverse scattering, full waveform inversion (FWI), and the Navier-Stokes equation. The computational cost at test time is operator-dependent: scattering uses cached linear algebra, FWI uses forward and adjoint wave solves, and Navier-Stokes uses differentiation through a fixed-step simulator. Numerical experiments further show that our method can improve the reconstruction quality by increasing the number of particles during inference, especially for ill-posed problems.

1 Introduction

Scaling compute at inference rather than at training has become a central theme in modern machine learning and generative artificial intelligence (GenAI). For large language models (LLMs), multiple related methods have been developed: chain-of-thought [35], self-consistency [33], tree search [39], and repeated sampling [2] all spend extra compute at test time [28, 38, 22], which all lead to better empirical performance than scaling the model parameters during training. This reflects a broader shift from training-time scaling laws [14, 11] to *inference-time scaling*, where additional compute is allocated to search, selection, verification, or sampling during inference. For a comprehensive review, the readers may refer to [41, 1]. The same shift is underway for pretrained diffusion models, where many inference-time tasks reduce to sampling from a *tilted* version of the pretrained distribution: annealing to low temperature, classifier-free guidance, reward-tilted sampling and posterior sampling under measurement constraints.

Existing inference-time methods can be classified into two categories. The first one is *length scaling*, which lengthens or iteratively refines a single trajectory, while the second one is *width scaling*, which explores multiple candidate trajectories in parallel. Width scaling transfers most readily beyond language, since it requires only a way to score candidates rather than a notion of a reasoning step. Sampling-based width scaling typically targets the tilted law through the sequential Monte Carlo (SMC) framework [6, 8], in which an interacting ensemble of particles is propagated under proposed dynamics and reweighted by an incremental potential prescribed by the path [37, 3, 27, 26, 4, 17, 23]. For physical inverse problems, however, it remains largely unexplored whether width scaling delivers comparable gains, and at what cost: the tilt is a physical likelihood rather than a learned reward, and each nonlinear evaluation may require solving a PDE, while a fixed linear operator can be cached. This raises the following question:

Submitted to the AI for Science workshop (NeurIPS 2026).

Can one develop SMC-based posterior samplers scaling test-time compute for physical inverse problems, using a fixed pretrained diffusion model that encodes the prior distribution?

In this paper, we tackle the question above by adopting the existing approximation-free ensemble posterior sampler [4], which evolves J weighted particles under the inherited AFDPS Feynman–Kac formulation instead of approximating a guided score along one trajectory, for multiple inverse problems from physical sciences. Below we start with describing prior work on solving physical inverse problems: Given measurements y of an unknown field x through a forward operator $\mathcal{A}$, plug-and-play posterior samplers reuse a pretrained diffusion prior to target $p(x \mid y) \propto p_0(x)\, p(y \mid x)$ without retraining [5, 34, 16, 44, 20, 29], and have been adopted across many physical inverse problems like optical microscopy, seismology, and fluid dynamics [43]. Here the tilt is the physical likelihood $\exp(-\mathcal{L}_y(x))$, which is specified by the measurement model rather than learned. One difficulty of width scaling therefore disappears: whereas the achievable gain elsewhere is bounded by the quality of a learned reward or verifier [18, 10, 25], here the scorer is explicit and differentiable under the stated forward model. Its reliability remains conditional on model and solver fidelity. What replaces reward-model uncertainty is operator cost: evaluating $\nabla\mathcal{L}_y$ for a nonlinear PDE requires a forward solve and an adjoint or reverse-mode gradient, so scoring rather than generation can dominate the budget, reversing the ordering in the LLM setting. Nearly all deployed samplers sidestep this by evolving a single reverse trajectory at a cost fixed by the algorithm, leaving no mechanism to convert spare compute into accuracy.

In addition, we remark that the particle count J here gives the sampler an inference-time compute axis that single-trajectory methods lack, and the price of moving along it is set by the algebraic structure of $\mathcal{A}$ rather than by the sampler. An FWI likelihood gradient requires a forward and an adjoint wave solve per source, and a Navier–Stokes gradient requires reverse-mode differentiation through 500 fixed solver steps, whereas for a linear operator with a cached SVD the gradient and curvature are closed-form and an ensemble is nearly free (Section 2). For instance, when we increase the number of particles from $J = 8$ to $J = 16$ for the FWI problem, the relative L_2 error drops from 0.197 to 0.155.

2 Background and method

2.1 Bayesian inverse problems

Let $x \in \mathbb{R}^d$ denote the unknown physical field and let $y \in \mathbb{R}^m$ be an observation generated by

$$y = \mathcal{A}(x) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_y^2 I), \tag{1}$$

where $\mathcal{A}$ is a differentiable forward operator and σ_y is the physical observation-noise scale. A prior $p_0(x)$ and the Gaussian observation model define the posterior

$$p(x \mid y) \propto p_0(x) \exp[-\mathcal{L}_y(x)], \qquad \mathcal{L}_y(x) = \frac{\|\mathcal{A}(x) - y\|_2^2}{2\sigma_y^2}. \tag{2}$$

The physical problems considered below are ill posed because $\mathcal{A}$ is non-injective, nonlinear, or observed only through a coarse terminal map. The prior supplies information that the measurement alone cannot identify [31].

2.2 Diffusion models for prior distributions

We represent p_0 with a pretrained diffusion model. A continuous forward process drives the target data distribution p_0 to the Gaussian distribution via the following stochastic differential equation (SDE):

$$dx_s = f(s)x_s\, ds + g(s)\, dW_s. \tag{3}$$

The corresponding reverse-time process, which uses the score $\nabla_x \log p_s(x)$ to transport samples back toward the data distribution, can be written as follows

$$dx_s = \left[f(s)x_s - g(s)^2 \nabla_x \log p_s(x_s)\right] ds + g(s)\, d\bar{W}_s, \tag{4}$$

where time is integrated from high noise to low noise. We use the *Elucidating the Design Space of Diffusion Models* (EDM) parameterization of Karras et al. [15]. With noise level σ, the denoiser D_θ provides the score estimate

$$\nabla_x \log p_\sigma(x) \approx \frac{D_\theta(x,\sigma) - x}{\sigma^2}. \tag{5}$$

All methods in our controlled comparisons use the same pretrained prior. The experiments therefore change inference-time computation without retraining the diffusion model.

2.3 AFDPS and SMC-based inference-time scaling

AFDPS augments the reverse diffusion process with likelihood-driven transport and sequential Monte Carlo reweighting [4]. At reverse step k, particle j has latent state x_k^j, denoised estimate $z_k^j = D_\theta(x_k^j, \sigma_k)$, and normalized weight w_k^j. The operator adapter evaluates the likelihood at the denoised point,

$$q_k^j = \nabla_z \mathcal{L}_y(z_k^j), \qquad c_k^j = \Delta_z \mathcal{L}_y(z_k^j), \tag{6}$$

and supplies these quantities to the inherited particle update. In schematic form, the population is propagated and reweighted by

$$x_{k-1}^j \sim K_k(\cdot \mid x_k^j, q_k^j), \qquad \widetilde{w}_{k-1}^j = w_k^j \exp\left[-\Delta G_k^j\right], \qquad w_{k-1}^j = \frac{\widetilde{w}_{k-1}^j}{\sum_{\ell=1}^J \widetilde{w}_{k-1}^\ell}, \tag{7}$$

where K_k is the AFDPS proposal kernel and ΔG_k^j is the incremental potential derived in Chen et al. [4]. The generic algorithm permits effective-sample-size resampling; all reported runs disable it. Initial particles are drawn from the high-noise endpoint of the pretrained prior.

The particle count J is the test-time width. Increasing J changes the number of likelihood evaluations and the fidelity of the weighted empirical posterior while leaving the checkpoint and diffusion schedule fixed. The continuous-time method and its many-particle guarantees are inherited from Chen et al. [4]. For the physical inverse problems tested here, we study how physical operator structure changes the implementation and cost of that width axis. Each adapter below must specify an evaluation point, a likelihood gradient, a trace or Laplacian term, numerical safeguards, and an output reduction.

3 Operator-adaptive likelihoods and costs

3.1 Linear inverse scattering

Let u_{in} be an incident field in a homogeneous background with wavenumber k_b, and let $f(\mathbf{r})$ denote the unknown scattering potential. The total field $u = u_{\mathrm{in}} + u_{\mathrm{sc}}$ satisfies the inhomogeneous Helmholtz equation

$$\left(\nabla^2 + k_b^2\right) u(\mathbf{r}) = -f(\mathbf{r}) u(\mathbf{r}), \qquad \left(\nabla^2 + k_b^2\right) u_{\mathrm{in}}(\mathbf{r}) = 0. \tag{8}$$

The background Green's function g converts the induced source fu into the scattered field observed on the receiver set Γ,

$$u_{\mathrm{sc}}(\mathbf{r}) = \int_\Omega g(\mathbf{r} - \mathbf{r}') f(\mathbf{r}') u(\mathbf{r}')\, \mathrm{d}\mathbf{r}', \qquad \mathbf{r} \in \Gamma. \tag{9}$$

The first-Born approximation replaces the unknown total field in the integrand by u_{in}. Discretization and real–imaginary stacking then give $y = Ax + \varepsilon$. For the cached singular-value decomposition $A = U \operatorname{diag}(s) V^\top$, define $\rho = U^\top(Ax - y)$. At the adapter's effective likelihood scale r_k,

$$\nabla \mathcal{L}_{r_k}(x) = V\left[\frac{s \odot \rho}{r_k^2}\right], \qquad \Delta \mathcal{L}_{r_k}(x) = \frac{\sum_j s_j^2}{r_k^2}. \tag{10}$$

Both quantities are exact for this discretized quadratic likelihood once the SVD is cached. The setting of r_k and the normalized observation scale is specified with the scattering experiment.

The production scattering sampler is a specialized AFDPS variant. It uses full anisotropic PiGDM guidance [29], retains an isotropic Feynman–Kac weighting term for particle selection, adds the

likelihood-value term, and integrates the linear guidance substep exactly in the singular-vector basis. These choices should be understood as operator-specific adaptations rather than the generic Euler update in Section 2. The anisotropic guidance used by this path has the singular-mode form

$$g_{\mathrm{PiGDM}}(x) = V\left[\frac{s \odot U^\top (Ax - y)}{\widetilde{\sigma}_y^2 + \sigma_k^2 s^2}\right], \tag{11}$$

where $\widetilde{\sigma}_y$ is the observation-noise scale after the affine normalization used by the scattering implementation. For the linear likelihood substep, writing $z = V^\top x$ and $b = U^\top y$, each nonzero singular mode is advanced analytically,

$$z_j^+ = e^{-\varphi_{j,k}} z_j + \left(1 - e^{-\varphi_{j,k}}\right) \frac{b_j}{s_j}, \qquad \varphi_{j,k} = \log \frac{\widetilde{\sigma}_y^2 + \sigma_k^2 s_j^2}{\widetilde{\sigma}_y^2 + \sigma_{k-1}^2 s_j^2}, \tag{12}$$

for a decreasing step $\sigma_k > \sigma_{k-1}$. Modes with $s_j = 0$ remain prior-controlled. This split avoids a stiffness restriction from the near-noiseless likelihood and makes the additional operator work per particle a pair of cached matrix–vector products. The current implementation forms a Feynman–Kac-weighted mean of $J = 512$ particles. A controlled diagnostic finds effective sample size one at the final step, making this output numerically one particle; the paired reduction experiment therefore also evaluates a uniform mean over all particles.

3.2 Full-waveform inversion

For compressional velocity $v(\mathbf{r})$, the acoustic wavefield satisfies

$$\frac{1}{v^2}\frac{\partial^2 u}{\partial t^2} - \nabla^2 u = q, \qquad y = Pu, \tag{13}$$

where q is the source and P samples receivers. Devito computes the forward wavefield and the adjoint-state gradient of the receiver residual [32, 19]. Each likelihood-gradient evaluation requires one forward and one adjoint wave solve per source. Before every solve, normalized particles are mapped to velocity and clamped to $[1.5, 4.5]$ km/s. Failed likelihood-gradient solves contribute zero gradient; non-finite particle states are excluded from terminal selection.

The configured curvature is approximated via the Hutchinson estimator applied to the forward-difference scheme [12],

$$\widehat{h}(x) = \frac{1}{M}\sum_{j=1}^{M} \xi_j^\top \frac{g(x + \epsilon\xi_j) - g(x)}{\epsilon}, \qquad \xi_j \sim \text{Rademacher}. \tag{14}$$

It estimates the full likelihood-Hessian trace; it is not the optional Gauss–Newton path. Each particle receives an independent probe. The clean width comparison uses $\tau = 0.5$ and $\gamma = 8$; the $\sigma_y = 0.5$ noisy run uses $\texttt{sn} = 0.6$ and $\gamma = 10$. All reported FWI runs use 200 EDM steps, no resampling, and maximum-weight output.

3.3 Navier–Stokes initial-condition recovery

We infer initial vorticity ω_0 from a subsampled terminal field. On the 2π torus,

$$\partial_t \omega + u \cdot \nabla \omega = \nu \Delta \omega + f, \qquad \nabla \cdot u = 0, \qquad y = P\omega(T) + \varepsilon. \tag{15}$$

The production path uses the pretrained 'ns-5m' diffusion checkpoint. A fixed-step pseudo-spectral solve with $\Delta t = 0.002$ and $T = 1$ creates a 500-step differentiable computation graph. Reverse-mode PyTorch autodiff yields the exact discrete adjoint of that implementation. A separately coded continuous adjoint is retained as a cross-check and is not used for the reported table.

The curvature is then approximated by combining the central finite difference scheme with the Hutchinson estimator:

$$\widehat{h}(\omega_0) = \frac{1}{M}\sum_{j=1}^{M} \xi_j^\top \frac{g(\omega_0 + \epsilon\xi_j) - g(\omega_0 - \epsilon\xi_j)}{2\epsilon}, \tag{16}$$

again with independent probes across particles. Runs use $J = 32$, 400 EDM steps, no resampling, and maximum-weight output. Guidance strengths and likelihood floors were tuned separately across acquisition settings; the floor is part of the inference likelihood and may exceed the physical observation noise.

Table 1: Configured physics-model-verification work per particle and diffusion step. Score-network evaluations are additional. $N_s = 16$ for FWI and $M = 1$ in both nonlinear experiments.

Operator	Gradient	Curvature	Forward/adjoint pairs
Scattering	cached SVD	exact constant	0
FWI	Devito adjoint state	forward FD, M probes	$(1 + M)N_s$
Navier–Stokes	discrete autodiff adjoint	central FD, M probes	$1 + 2M$

3.4 Per-particle verification cost

Table 1 reports configured operator evaluations rather than measured runtime. One gradient evaluation means a forward solve and its adjoint or reverse-mode backward pass. FWI's forward-difference estimator uses the base gradient plus M perturbed gradients. Navier–Stokes central differences use the base gradient plus $2M$ perturbed gradients.

4 Experiments

Evaluation protocol. We use the physical operators, pretrained diffusion checkpoints, test sets, and evaluators released with InverseBench [43]. We report relative L_2 error, PSNR, and SSIM, with arrows in each table giving the preferred direction. FWI and Navier–Stokes aggregates use ten cases. The scattering table uses 100 test images. Standard deviations are sample standard deviations when case-level values are available. Appendix D records the algorithm configurations and local qualifications.

4.1 Full-waveform inversion

We recover ten 128×128 CurveFaultB velocity maps [7] using the pretrained 'fwi-5m' EDM prior. The acquisition uses 16 sources at 1270 m depth, 129 receivers at 10 m depth, a 5 Hz Ricker wavelet, 1 ms temporal sampling, and a one-second record. The width comparison uses noise-free observations. Its $\tau = 0.5$ is an inference likelihood temperature rather than a physical noise level. Both AFDPS widths use 200 reverse steps, the same checkpoint, and the same ten cases. The comparison includes DAPS [40] and LGD [30] among the plug-and-play baselines.

Increasing the ensemble from $J = 8$ to $J = 16$ lowers mean relative L_2 from 0.197 to 0.155, raises PSNR from 14.10 to 15.70, and raises SSIM from 0.487 to 0.563 in Table 2. The $J = 16$ row has the best relative L_2 and SSIM among the prior-only rows. Published DiffPIR retains the best prior-only PSNR, 16.11. LBFGS uses an initialization formed by blurring the ground-truth velocity and is retained as an informative-initialization reference.

This two-width experiment establishes a local gain under one fixed FWI protocol. It does not determine a scaling exponent or a saturation point. Appendix C reports a selected noisy reconstruction at physical noise $\sigma_y = 0.5$. On that reconstruction, AFDPS $J = 16$ records $(0.1224, 18.98, 0.5283)$ in (relative L_2, PSNR, SSIM), outperforming DiffPIR single, DiffPIR r16, and LGD on all three metrics. This is a selected-case illustration; over the seven cases complete for every method, LGD has the best mean on all three metrics.

4.2 Linear inverse scattering

We recover 128×128 HL60 permittivity maps [36] from first-Born scattered fields using the HL60 EDM prior. Each observation uses 20 illuminations and physical complex-noise standard deviation 10^{-4}. The test set contains 100 images, and receiver count R controls acquisition density. AFDPS uses $J = 512$ and 200 steps.

A controlled $R = 360$ rerun resolves the output-reduction question. The current Feynman–Kac weights collapse to effective sample size one, so their weighted mean is numerically one particle. Uniformly averaging all particles improves mean PSNR from 32.78 to 34.50 dB and mean SSIM from 0.9786 to 0.9871, with gains on all 100 paired cases.

Committed DDNM, DPS, and RED-diff configurations were then rerun on the same 100 targets and byte-identical cached observations. Table 3 reports a selected reconstruction from this comparison,

Table 2: FWI on ten CurveFaultB cases, mean (std). AFDPS and both retuned DiffPIR rows use sample standard deviations from the earlier width-comparison run; Appendix C reports a separate noisy run. Only the earlier run family is used in this table. Published rows preserve their source convention. † denotes an informative initialization derived from a blurred ground-truth velocity. Retuning A uses $\sigma_n = 0.5, \lambda = 25$; retuning B uses $\sigma_n = 0.75, \lambda = 10$; both use $\xi = 0.05$ and 200 steps. Best prior-only value is bold. We show the baseline subset used in this study; the full InverseBench table also reports PnP-DM and data misfit.

Method	rel. $L_2 \downarrow$	PSNR↑	SSIM↑
LBFGS†	0.070 (0.023)	23.40 (2.75)	0.704 (0.077)
DAPS†	0.201 (0.103)	14.91 (4.18)	0.321 (0.067)
DPS	0.250 (0.154)	14.11 (6.82)	0.491 (0.161)
LGD	0.244 (0.024)	12.29 (0.89)	0.341 (0.047)
RED-diff	0.319 (0.102)	10.37 (2.65)	0.280 (0.108)
DiffPIR (published)	0.204 (0.129)	**16.11** (6.96)	0.554 (0.191)
DiffPIR (retuned A)	0.188 (0.134)	15.57 (6.26)	0.559 (0.160)
DiffPIR (retuned B)	0.190 (0.136)	15.43 (6.22)	0.559 (0.159)
AFDPS ($J = 8$)	0.197 (0.112)	14.10 (4.32)	0.487 (0.142)
AFDPS ($J = 16$)	**0.155 (0.066)**	15.70 (3.45)	**0.563 (0.110)**

Table 3: Selected inverse-scattering reconstruction at $R = 360$ and $\sigma_y = 10^{-4}$. All methods use the same target and byte-identical observation. Higher is better; bold marks the best value. DPS and RED-diff are committed-configuration reruns and fall below their previously reported aggregates.

Method	PSNR↑	SSIM↑
AFDPS (uniform mean)	**42.81**	**0.9980**
DDNM rerun	39.66	0.9147
DPS rerun	30.36	0.8982
RED-diff rerun	11.03	0.4073

with the full reconstruction and error plate in Appendix B. Over the full test set, AFDPS attains the higher mean SSIM (0.987 against 0.927 for DDNM), while DDNM retains the higher mean PSNR (37.3 against 34.5 dB).

On this reconstruction, AFDPS exceeds the strongest comparator, DDNM, by 3.16 dB PSNR and 0.0833 SSIM. These fixed-width results do not establish particle-width scaling because only $J = 512$ is evaluated.

4.3 Navier–Stokes initial-condition recovery

We recover ten 128×128 initial-vorticity fields at Reynolds number 200 from spatially subsampled terminal observations using the pretrained 'ns-5m' EDM prior. The physical solver advances one time unit with forcing $f = -4\cos(4x_2)$. AFDPS uses $J = 32,400$ steps, and the maximum-weight particle. Table 4 shows the retained settings where AFDPS has the lowest mean relative L_2 error; Table 7 gives the complete sweep. We compare against EnKG [42] and EKI [24].

Against the fixed-step reruns, AFDPS has lower mean error than EnKG and EKI for $\sigma_y \geq 2$, while EnKG leads through $\sigma_y = 1$ in the full sweep. These EnKG and EKI values are reruns rather than the published InverseBench aggregates. At the overlapping $\mathrm{ds} = 8, \sigma_y = 2$ setting, published EnKG records 0.773, slightly below AFDPS at 0.780. Different timestepping and the absence of a shared observation cache prevent a controlled same-simulator claim; Appendix D states the required aligned rerun.

5 Related work

Plug-and-play diffusion priors. Posterior samplers combine a pretrained generative prior with a problem-specific likelihood. DPS approximates likelihood guidance along one trajectory [5]; DDNM exploits the null space of a linear operator [34]; DiffPIR and RED-diff use splitting or variational

Table 4: Navier–Stokes settings where AFDPS has the lowest mean relative L_2 error over ten fields. Lower is better. Physical noise is σ_y. EnKG and EKI are fixed-step reruns rather than published InverseBench aggregates. The complete sweep, including the low-noise settings where EnKG leads, is in Table 7.

ds	σ_y	AFDPS	EnKG	EKI
8	2	**0.780**	0.813	1.140
8	3	**0.843**	1.007	1.306
8	4	**0.909**	1.182	1.411
8	5	**1.040**	1.313	1.512

objectives [44, 20]. These constructions avoid retraining the prior for every observation, though their approximations and stability properties differ. InverseBench compares the families on physical operators and documents failures when noisy iterates enter constrained PDE solvers [43]. Our denoised-point evaluation on the nonlinear paths and FWI clamping address that implementation failure mode. The cached-SVD scattering path instead exploits linear structure unavailable to the nonlinear operators.

Particle diffusion and inference-time width. Population methods propagate interacting candidates and attach Feynman–Kac weights to conditional generation [3, 27, 26]. Under the assumptions of its continuous-time analysis, AFDPS bounds score-model error and establishes a many-particle limit for its weighted-particle evolution [4]. The source paper also reports particle-count convergence on a synthetic posterior and a mild image-domain width ablation. The source also identifies geoscience and fluid dynamics as natural extensions; we instantiate that direction and measure how operator structure changes the price of particle width. Each added particle can require an adjoint PDE solve. Drift-control methods instead modify the proposal to reduce weight degeneracy [23]. That direction is complementary to our experiment, which holds the inherited sampler fixed while varying J.

Those inherited results do not certify the finite EDM discretization, denoised-point likelihoods, stochastic traces, or domain-dependent output reductions used here. We study the operator layer needed to evaluate its likelihood terms. This differs from drawing independent terminal samples and selecting the lowest-residual result. Physics-model information enters every reverse step and also changes the particle weights.

Inference-time scaling work asks how additional sequential or parallel computation changes performance [28, 38]. Parallel scaling is useful only when candidates can be ranked by a meaningful signal. Learned reward models and process verifiers can fail outside their training distribution or favor artifacts of the scoring rule [18, 25]. A physical residual avoids a learned reward model, yet it inherits the assumptions of the simulator, discretization, and observation law. It is therefore a model-based verifier rather than ground truth. This conditional status is central to the workshop setting and motivates reporting the solver and acquisition details alongside reconstruction metrics.

Physical inference-time scaling and learned inverse operators. SCaSML spends inference-time compute to correct a learned forward PDE surrogate with stochastic simulation [9]. Neural Inverse Operators instead learn problem-specific maps from observations to unknown coefficients [21]. Our setting keeps the pretrained prior fixed, solves each posterior at inference time, and scales the number of weighted particles.

Scope of evidence. The contribution is narrower than an empirical scaling law. We provide a controlled two-width comparison on FWI, a fixed-width linear specialization, and a higher-noise Navier–Stokes diagnostic whose aggregate baseline path has a known timestepping mismatch. Together they show how verification cost and candidate benefit depend on the forward operator. They do not establish an optimal particle budget, calibrated posterior coverage, or universal advantage over single-trajectory and ensemble-Kalman methods.

6 Discussion and conclusion

In this paper, we study how to scale test-time compute for physical inverse problems by adopting an existing approximation-free ensemble posterior sampler, with ensemble size as the main inference-

time compute variable. The cost of using more particles is set by the forward operator. For a linear map with a cached SVD, the likelihood gradient and Feynman–Kac curvature are closed-form, and the linear guidance substep is integrated exactly. FWI instead requires $(1 + M)N_s$ forward-adjoint pairs per particle and diffusion step, while Navier–Stokes requires $1 + 2M$ forward and reverse-mode pairs under its central-difference estimator. On FWI, doubling the ensemble gives the largest improvement we measure and takes the sampler to the best relative L_2 and SSIM among prior-only methods. The fixed-width scattering result does not establish saturation. On Navier–Stokes, EnKG leads in the clean regime and the recorded ordering reverses at higher noise, although different timestepping prevents a same-budget claim.

Interpretation across operators. The cost table and the reconstruction tables answer separate questions. Table 1 states what one additional particle requires under each implementation. Only the FWI width pair measures what that extra particle buys while the remaining protocol is fixed. The scattering experiment shows that a large ensemble can be coupled cheaply to a cached linear operator, without a smaller-J row from which to infer saturation. The Navier–Stokes sweep changes observation noise at fixed J, so it is a robustness result rather than a second width curve. Taken together, the experiments support a narrow conclusion: particle width is a usable inference-time control, and its measured benefit here is domain-specific.

Verification perspective. The physical likelihood is strongest as a verifier when the discretized forward model is trusted over the candidate region. Scattering comes closest to that case: its linear residual, gradient, and curvature are exact after normalization and SVD caching. The nonlinear residuals depend on numerical solvers and on safeguards that keep candidates inside a stable regime. FWI clamping and Navier–Stokes likelihood floors are therefore part of the verifier definition, not incidental engineering. A low residual may still reward a field that exploits model error or an inaccurate noise law. We guard against that reading by reporting reconstruction metrics against retained ground truth and by keeping unfavorable regimes visible. The resulting evidence concerns verification under the stated simulators, rather than unrestricted physical validity.

Limitations. The strongest empirical statement rests on two particle counts for one operator. This is enough to show an improvement under a fixed FWI protocol, while leaving the shape and eventual saturation of the width curve unknown. Table 1 counts configured solver calls rather than wall-clock time. Hardware utilization, score-network batching, failed PDE solves, and memory pressure can change the practical exchange rate between particles and accuracy. The benchmark record is also uneven. The scattering comparison is a controlled rerun on cached observations with per-case records, while FWI and Navier–Stokes retain case-level workbook values. The Navier–Stokes high-noise ordering uses baseline aggregates generated with different timestepping. We preserve these results because they answer a useful robustness question, withhold a same-budget claim, and disclose what a controlled rerun must repair.

Future measurements. A denser sweep in J with wall-clock and operator-call measurements would locate where the FWI curve saturates and permit compute-optimal allocation [28, 13]. Scattering reduction is no longer an open measurement: the current Feynman–Kac weights collapse to effective sample size one, uniform averaging improves both point metrics, and best-particle selection is inferior to the mean. The nonlinear paths still return the maximum-weight particle, and neither output rule measures calibration. Since guidance temperature trades candidate diversity against fidelity [26], future experiments should retain particle-level predictions and report coverage or task-specific posterior functionals in addition to point-estimate metrics. Finally, an aligned Navier–Stokes rerun should cache identical observations and use the same discrete simulator for every method before the high-noise crossover is treated as a controlled comparative result.

References

[1] Vidhisha Balachandran, Jingya Chen, Lingjiao Chen, Shivam Garg, Neel Joshi, Yash Lara, John Langford, Besmira Nushi, Vibhav Vineet, Yue Wu, et al. Inference-time scaling for complex tasks: Where we stand and what lies ahead. *arXiv preprint arXiv:2504.00294*, 2025.

[2] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

[3] Gabriel Cardoso, Sylvain Le Corff, Eric Moulines, et al. Monte carlo guided denoising diffusion models for bayesian linear inverse problems. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[4] Haoxuan Chen, Yinuo Ren, Martin Renqiang Min, Lexing Ying, and Zachary Izzo. Solving inverse problems via diffusion-based priors: An approximation-free ensemble sampling approach. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=qN8ASsfjKs.

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

[6] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436, 2006.

[7] Chengyuan Deng, Shihang Feng, Hanchen Wang, Xitong Zhang, Peng Jin, Yinan Feng, Qili Zeng, Yinpeng Chen, and Youzuo Lin. Openfwi: Large-scale multi-structural benchmark datasets for full waveform inversion. *Advances in Neural Information Processing Systems*, 35:6007–6020, 2022.

[8] Arnaud Doucet, Adam M Johansen, et al. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

[9] Zexi Fan, Yan Sun, Shihao Yang, and Yiping Lu. Physics-informed inference time scaling for solving high-dimensional partial differential equations. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

[10] Lin Gui, Cristina Gârbacea, and Victor Veitch. Bonbon alignment for large language models and the sweetness of best-of-n sampling. *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 2851–2885, 2024.

[11] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *Advances in Neural Information Processing Systems*, 35:30016–30030, 2022.

[12] Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

[13] Yuichi Inoue, Kou Misaki, Yuki Imajuku, So Kuroki, Taishi Nakamura, and Takuya Akiba. Wider or deeper? scaling llm inference-time compute with adaptive branching tree search. *arXiv preprint arXiv:2503.04412*, 2025.

[14] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[15] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:26565–26577, 2022.

[16] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:23593–23606, 2022.

[17] Cheuk Kit Lee, Paul Jeha, Jes Frellsen, Pietro Lio, Michael S Albergo, and Francisco Vargas. Debiasing guidance for discrete diffusion with sequential monte carlo. *arXiv preprint arXiv:2502.06079*, 2025.

[18] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[19] M. Louboutin, M. Lange, F. Luporini, N. Kukreja, P. A. Witte, F. J. Herrmann, P. Velesko, and G. J. Gorman. Devito (v3.1.0): an embedded domain-specific language for finite differences and geophysical exploration. *Geoscientific Model Development*, 12(3):1165–1187, 2019. doi: 10.5194/gmd-12-1165-2019. URL https://www.geosci-model-dev.net/12/1165/2019/.

[20] Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[21] Roberto Molinaro, Yunan Yang, Björn Engquist, and Siddhartha Mishra. Neural inverse operators for solving pde inverse problems. *arXiv preprint arXiv:2301.11167*, 2023.

[22] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 20286–20332, 2025.

[23] Yinuo Ren, Wenhao Gao, Lexing Ying, Grant M Rotskoff, and Jiequn Han. Driftlite: Lightweight drift control for inference-time scaling of diffusion models. *arXiv preprint arXiv:2509.21655*, 2025.

[24] Claudia Schillings and Andrew M Stuart. Convergence analysis of ensemble kalman inversion: the linear, noisy case. *Applicable Analysis*, 97(1):107–123, 2018.

[25] Amrith Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. Scaling test-time compute without verification or rl is suboptimal. In *International Conference on Machine Learning (ICML)*, pages 54058–54094. PMLR, 2025.

[26] Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen Mckeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *International Conference on Machine Learning (ICML)*, pages 55810–55827. PMLR, 2025.

[27] Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alan Aspuru-Guzik, Arnaud Doucet, Rob Brekelmans, Alexander Tong, and Kirill Neklyudov. Feynman-kac correctors in diffusion: Annealing, guidance, and product of experts. In *International Conference on Machine Learning (ICML)*, pages 55906–55949. PMLR, 2025.

[28] Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[29] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023.

[30] Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning (ICML)*, pages 32483–32498. PMLR, 2023.

[31] Andrew M Stuart. Inverse problems: a bayesian perspective. *Acta numerica*, 19:451–559, 2010.

[32] Jean Virieux and Stéphane Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, 2009.

[33] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[34] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:24824–24837, 2022.

[36] David Wiesner, David Svoboda, Martin Maška, and Michal Kozubek. Cytopacq: a web-interface for simulating multi-dimensional cell imaging. *Bioinformatics*, 35(21):4531–4533, 2019.

[37] Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:31372–31403, 2023.

[38] Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for llm problem-solving. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[39] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[40] Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing, 2024. URL https://arxiv.org/abs/2407.01521.

[41] Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, et al. A survey on test-time scaling in large language models: What, how, where, and how well? *arXiv preprint arXiv:2503.24235*, 2025.

[42] Hongkai Zheng, Wenda Chu, Austin Wang, Nikola Kovachki, Ricardo Baptista, and Yisong Yue. Ensemble kalman diffusion guidance: A derivative-free method for inverse problems. *arXiv preprint arXiv:2409.20175*, 2024.

[43] Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy T Feng, Caifeng Zou, Yu Sun, Nikola Kovachki, Zachary E Ross, et al. Inversebench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=U3PBITXNG6.

[44] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1219–1229, 2023.

A Qualitative reconstructions

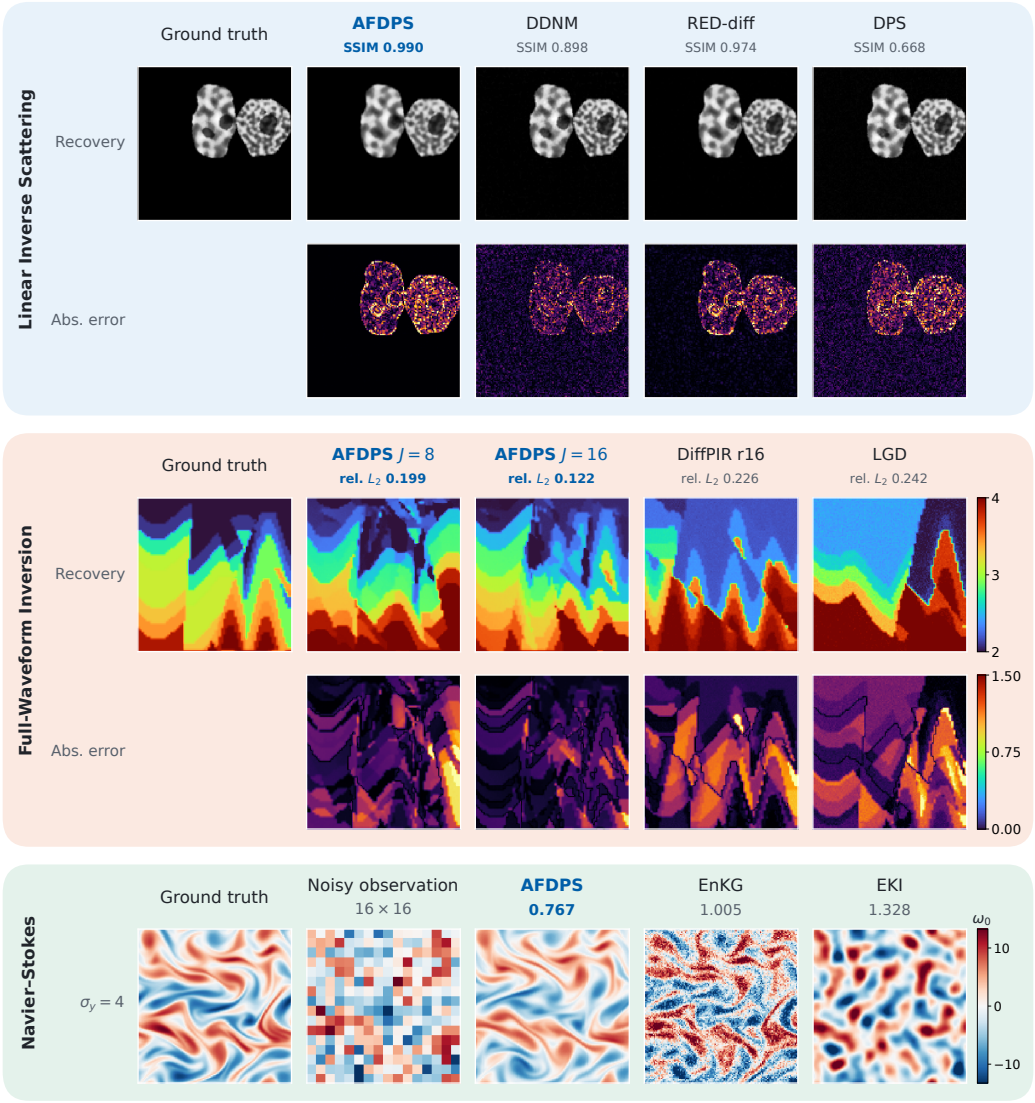


Figure 1: **Cross-domain qualitative reconstructions. Top left:** Inverse scattering shows the ground truth, four reconstructions, and pointwise absolute-error maps; headers report SSIM, and error colors are not comparable across columns. The receiver count is not retained in this qualitative export. **Top right:** a selected noisy CurveFaultB reconstruction at $\sigma_y = 0.5$, the selected reconstruction reported in Table 5, compares AFDPS at $J \in \{8, 16\}$ with DiffPIR r16 and LGD; headers report relative L_2, and the recovery and error rows use shared scales. **Bottom:** the Navier–Stokes $\sigma_y = 4$ showcase displays the ground truth, noisy observation, and recoveries from AFDPS, EnKG, and EKI for one retained field, with relative L_2 beneath each recovery. These values are field-level results and therefore differ from the ten-field means in Table 4. This setting provides the clearest high-noise separation among the recorded Navier–Stokes outputs; EnKG leads in the clean regime retained in Table 7. Unverified posterior-standard-deviation panels remain excluded.

The qualitative panels emphasize settings in which the physical structure and the effect of additional inference-time compute are both visible. Quantitative comparisons and compute counts are reported

beside the affected figures. The raw observation tensor was retained in the Navier–Stokes export. The IS and FWI exports retain recoveries and pointwise error maps but omit the raw measurement. For those domains we state the observation tensor structure rather than fabricating a raster from unavailable run artifacts.

Observation availability. The retained scattering and FWI exports omit their raw measurement tensors, so Figure 1 does not fabricate observation panels. Scattering error maps retain source-specific display normalization, and the displayed FWI panels are the selected reconstruction reported in Table 5.

B Selected inverse-scattering reconstruction

Figure 2 shows the complete reconstruction and error plate for the selected reconstruction of Table 3.

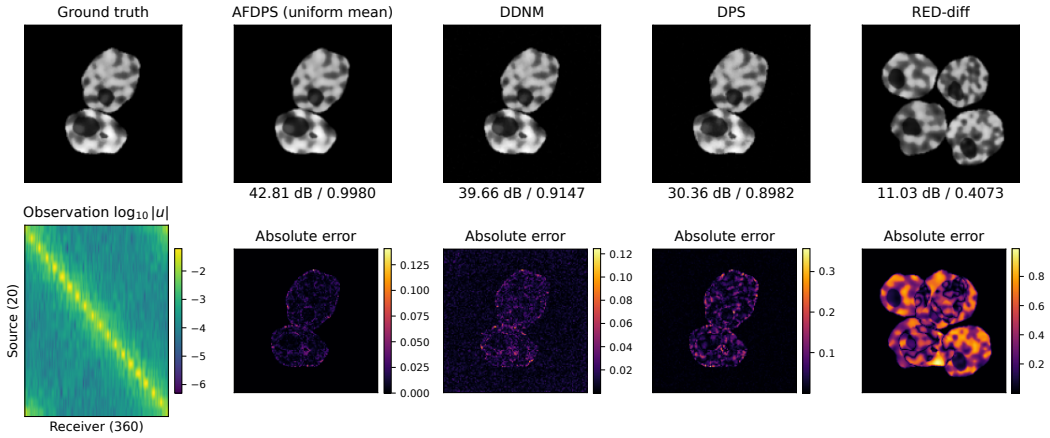


Figure 2: The selected inverse-scattering reconstruction of Table 3 at $R = 360$. **Top:** ground truth and reconstructions on a shared $[0, 1]$ gray scale, with PSNR and SSIM beneath each method. **Bottom:** the scattered-field observation magnitude, and pointwise absolute-error maps; each error panel carries its own color scale, so error colors are not comparable across columns. The AFDPS image is a same-configuration recovery rerun that reproduces the reported metrics to displayed precision on the byte-identical observation; the comparator panels are the original rerun tensors. RED-diff hallucinates two additional cells, while DDNM and DPS leave boundary and interior errors that AFDPS avoids.

C Selected noisy FWI reconstruction

From ten noisy CurveFaultB reconstructions generated by Eq. (1), we report a selected reconstruction at additive Gaussian noise $\sigma_y = 0.5$ in native receiver-trace units. It is selected because the unflagged AFDPS $J = 16$ reconstruction has the strongest retained separation from every displayed comparator. The runs use the 'fwi-5m' prior, 200 reverse steps, and seed 0. AFDPS uses the recorded setting $\texttt{sn} = 0.6$ and $\gamma = 10$. DiffPIR uses $\sigma_n = 0.5$, $\lambda = 2500$, and $\xi = 0.05$. Its r16 variant draws 16 trajectories and returns the reconstruction with the lowest data misfit. AFDPS $J = 16$ requires 32 PDE-gradient evaluations per step; DiffPIR r16 uses 16, DiffPIR single uses one, and LGD uses three Monte Carlo gradients. The retained $J = 8$ output for this case has a caught zero-gradient solve and is therefore excluded from the quantitative table.

Table 5: Selected noisy FWI reconstruction at $\sigma_y = 0.5$. Lower relative L_2 and higher PSNR and SSIM are better. Bold marks the best value. This is a selected single-case illustration rather than an aggregate comparison. AFDPS $J = 16$ uses 32 PDE-gradient evaluations per reverse step; DiffPIR r16 uses 16.

Method	rel. $L_2 \downarrow$	PSNR↑	SSIM↑	PDE gradients/step
AFDPS ($J = 16$)	**0.1224**	**18.98**	**0.5283**	32
DiffPIR (single)	0.2174	13.97	0.3651	1
DiffPIR (r16)	0.2263	13.63	0.3459	16
LGD	0.2418	13.05	0.3738	3

On this selected reconstruction, AFDPS $J = 16$ has 43.7% lower relative L_2, 5.01 dB higher PSNR, and 0.1545 higher SSIM than the strongest comparator values. This is an outcome-conditioned single-case result. Across the seven cases with completed runs for every method, LGD has the best mean relative L_2, PSNR, and SSIM; the selected reconstruction therefore does not establish aggregate AFDPS dominance. Figure 1 shows the retained width illustration for this reconstruction, including the flagged $J = 8$ output.

D Detailed experimental settings

Table 6: Documented AFDPS settings. The controlled scattering rerun uses a uniform mean; the exact reduction and receiver-dependent scale of the original evaluation are not retained.

Setting	Scattering	FWI	Navier–Stokes
Prior	HL60 EDM checkpoint	CurveFaultB EDM checkpoint	'ns-5m' EDM checkpoint
Particles J	512	8 or 16	32
EDM steps	200	200	400
Likelihood point	noisy particle	Tweedie estimate	Tweedie estimate
Guidance	full anisotropic PiGDM	$\gamma = 8$ clean; $\gamma = 10$ noisy	fixed, tuned by setting
Likelihood scale	physical 10^{-4} plus annealing	temperature $\tau = 0.5$	physical noise plus tuned floor
Trace	exact SVD constant	forward FD, $M = 1$, $\epsilon = 10^{-3}$	central FD, $M = 1$, $\epsilon = 10^{-3}$
Output	original reduction not retained; uniform mean in paired rerun	maximum-weight particle	maximum-weight particle
Resampling	disabled	disabled	disabled

Scattering geometry. InverseBench uses an 18 cm square domain, 0.84 cm wavelength, 20 transmitters, a 1.6 m acquisition radius, and background relative permittivity one. The HL60 test set contains 100 images; the reported comparison uses $R = 360$.

FWI geometry and stabilization. The 1.27 km square domain uses a 128×128 grid, sixteen sources at 1270 m depth, 129 receivers at 10 m depth, and a 5 Hz Ricker source sampled at 1 ms for one second. AFDPS evaluates the likelihood at the denoised field, clamps velocity to $[1.5, 4.5]$ km/s, isolates failed solves, and uses zero-field initialization. For the earlier $J = 8$ and $J = 16$ width run, DiffPIR respectively uses $(\sigma_n, \lambda, \xi) = (0.5, 25, 0.05)$ and $(0.75, 10, 0.05)$. The separate $\sigma_y = 0.5$ run uses $\texttt{sn} = 0.6$, $\gamma = 10$ for AFDPS and $(0.5, 2500, 0.05)$ for DiffPIR; r16 returns the lowest-data-misfit trajectory.

Table 7: Complete Navier–Stokes fixed-step rerun sweep over ten fields. Lower mean relative L_2 is better. EnKG leads through $\sigma_y = 1$; AFDPS leads at $\sigma_y \geq 2$. The abbreviated main-text table retains only the latter rows.

ds	σ_y	AFDPS	EnKG	EKI
2	0	0.591	**0.124**	0.621
4	0	0.634	**0.133**	0.638
8	0	0.655	**0.313**	0.877
8	1	0.730	**0.562**	0.979
8	2	**0.780**	0.813	1.140
8	3	**0.843**	1.007	1.306
8	4	**0.909**	1.182	1.411
8	5	**1.040**	1.313	1.512

Navier–Stokes solver. The state uses a 128×128 periodic grid at Reynolds number 200 with strided terminal-state measurements. AFDPS fixes $\Delta t = 0.002$ for a differentiable graph; the original baseline uses adaptive timestepping. Observation caching and aligned solvers were added later for qualitative figures. The displayed table uses the pretrained diffusion prior rather than the repository's optional analytic Gaussian-random-field score.