DECOUPLED DIFFUSION SOLVER FOR INVERSE PROBLEMS ON FUNCTION SPACES

Thomas Y.L. Lin[1,*] **Jiachen Yao**[2,*] **Lufang Chiang**[3] **Julius Berner**[4] **Anima Anandkumar**[2]

[1]University of Washington, Seattle, USA
[2]California Institute of Technology, Pasadena, USA
[3]National Taiwan University, Taipei, Taiwan
[4]NVIDIA Corporation, USA
[*]Equal contribution.

ABSTRACT

We propose a data-efficient, physics-aware generative framework in function space for inverse PDE problems. Existing plug-and-play diffusion posterior samplers represent physics implicitly through joint coefficient-solution modeling, requiring substantial paired supervision. In contrast, our Decoupled Diffusion Inverse Solver (DDIS) employs a *decoupled* design: an unconditional diffusion learns the coefficient prior, while a neural operator explicitly models the forward PDE for guidance. This decoupling enables superior data efficiency and effective physics-informed learning, while naturally supporting *Decoupled* Annealing Posterior Sampling (DAPS) to avoid over-smoothing in Diffusion Posterior Sampling (DPS). Theoretically, we prove that DDIS avoids the guidance attenuation failure of joint models when training data is scarce. Empirically, DDIS achieves state-of-the-art performance under sparse observation, improving l_2 error by 11% and spectral error by 54% on average; when data is limited to 1%, DDIS maintains accuracy with 40% advantage in l_2 error compared to joint models.

1 INTRODUCTION

Inverse problems are commonplace across science and engineering fields. Typically, such problems are ill-posed, non-unique, and nonlinear, requiring computationally expensive solvers and handcrafted knowledge (Tarantola, 2005). Inverse problems governed by partial differential equations (PDEs) aim to infer unknown coefficient fields $a \in \mathcal{A}$ (e.g., coefficients, sources) from partial or noisy observations u_{obs}. Formally, with forward operator $L : \mathcal{A} \to \mathcal{U}$, mask operator M, and noise ϵ, the observation model is

$$u = L(a), \quad u_{\text{obs}} = M \odot u + \epsilon. \tag{1.1}$$

Applications such as weather forecasting (Manshausen et al., 2024) and geophysical imaging (Herrmann & Li, 2012), operate under *sparse* sensor coverage, where observations are available only on a small fraction of the spatial domain. Separately, acquiring paired training data (a, u) requires repeatedly solving the underlying PDE (Chen et al., 2024), leading to an imbalanced regime with abundant coefficients but scarce paired coefficient-solution samples.

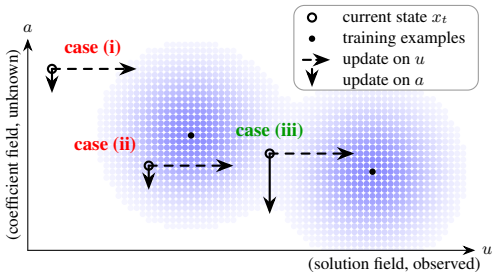

Figure 1: **Gradient guidance by joint-embedding models vanishes under scarce paired data.** Blue blobs visualize regions supported by the learned joint model around individual training examples (black dots), and the circle marker denotes diffusion state x_t. We consider three cases: when x_t is far from all blobs (i), near a single training sample (ii), or close to multiple training samples (iii). In all cases, the update on u (dashed arrows) remains valid; however, the update on a (solid arrows) vanishes in (i) and (ii), and is nonzero only in (iii). In high dimensions, case (iii) is rare, rendering coefficient-space guidance ineffective.

Bayesian generative modeling has emerged as a principled paradigm for solving PDE inverse problems by sampling from the posterior $p(a \mid u_{\text{obs}})$. Guidance-based methods such as Diffusion Posterior Sampling (DPS (Chung et al., 2022)) evaluate the observation likelihood to steer prior

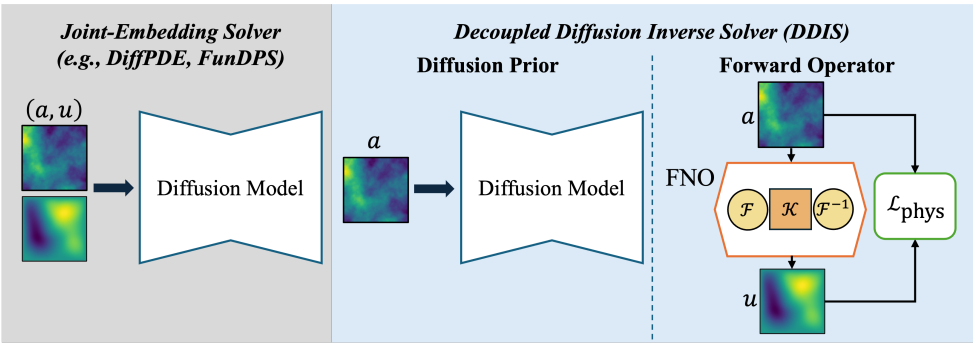

Figure 2: Training pipeline comparison. Left (Gray): Joint-embedding methods rely on *paired* data (a, u) to learn the joint distribution $p(a, u)$. Right (Cyan): DDIS decouples the architecture: the diffusion prior learns $p(a)$ utilizing abundant *unpaired* coefficients, while the neural operator takes paired data (a, u) to directly learn the forward physics map.

samples toward the posterior during the reverse diffusion process. Recent frameworks, including DiffusionPDE (Huang et al., 2024) and FunDPS (Yao et al., 2025), implement this by learning a diffusion prior over the *joint* distribution of coefficients and solutions $p(a, u)$ and apply guidance by masking the observed components. As a result, joint-embedding models must recover the underlying physics purely from statistical cross-field correlations. Conceptually, this reduces inverse PDE solving to inpainting problems in the joint space, raising a fundamental question:

Can joint-embedding models effectively support cross-field guidance?

Our analysis reveals that the answer is no, especially under data scarcity. We identify two hurdles that limit the effectiveness of joint embeddings. First, when data is scarce, joint models suffer from *guidance attenuation* (Sec. 4.1); we characterize this through geometric criteria (Fig. 1): non-vanishing guidance requires at least two training data points in the local neighborhood of current diffusion state, which the curse of dimensionality renders nearly impossible to satisfy. Second, for advanced samplers like DAPS (Zhang et al., 2025), sparse observations cause correlations between observed points and their spatial neighbors to collapse to zero in joint models (Sec. 4.2). These motivate a decoupled design that avoids joint embeddings and enforces physics consistency through explicit representations.

We therefore propose **Decoupled Diffusion Inverse Solver (DDIS)**, a modular framework that separates prior modeling from physics-induced likelihood evaluation.

Decoupling Physics during Training. Our key insight is that priors and physics serve fundamentally different roles: the prior is defined over the coefficient space, while the likelihood is defined by the PDE solution operator mapping a to u. Modeling the physics via a forward operator is more data-efficient than learning it implicitly through joint embeddings. Motivated by this decoupling, DDIS (i) learns the coefficient prior using a diffusion model in function space, and (ii) uses a neural operator $L_\phi(a)$ to represent the forward physics and evaluate the likelihood (Fig. 2). This enables data-efficient learning by exploiting abundant coefficient prior samples for prior modeling and limited paired data for operator training.

Decoupling Sampling during Inference. Posterior sampling requires combining the learned prior with the physics-induced likelihood. Unbiased posterior samplers such as *Decoupled* Annealing Posterior Sampling (DAPS (Zhang et al., 2025)) require dense guidance signals, which joint-embedding models fail to provide under sparse sensor coverage. In contrast, the neural operator L_ϕ in DDIS propagates sparse observations in the solution space into dense guidance over the coefficient space. (See Fig. 3.) This enables the effective use of DAPS during inference, avoiding the over-smoothing artifacts observed in prior DPS-based models.

Advantages. DDIS's key advantages include:

- **Data-efficient Learning.** DDIS leverages operator learning to bypass the limitations of joint embedding, thereby dominating the Pareto front of the accuracy-data tradeoff and sustaining performance under extreme data scarcity.

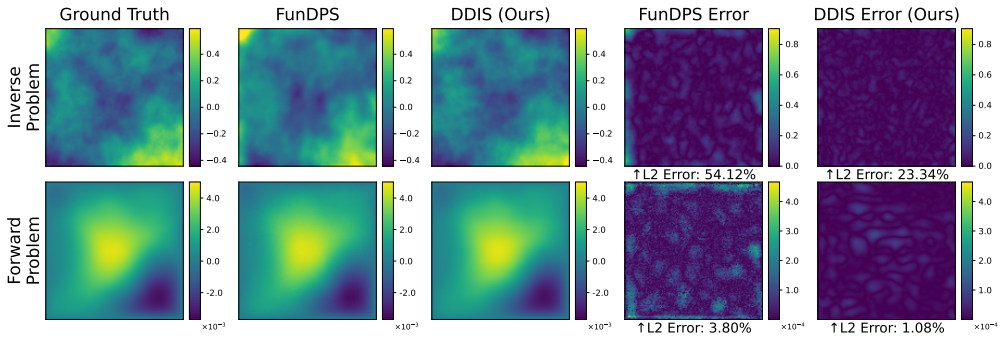

Figure 3: Comparison of FunDPS and our DDIS on inverse Poisson problem reconstruction under 1% paired data scarcity. FunDPS suffers from over-smoothing due to Jensen's gap while DDIS achieves sharp and dense guidance with improved accuracy.

- **Physics Integration.** DDIS uses a neural operator surrogate to explicitly represent the forward PDE mapping and enforce physics consistency by design. Incorporating physics loss during surrogate training is significantly more robust than previous attempts. Our physics-informed variant performs on par with the 100%-data variant with only 1% of paired data.
- **Theoretical Justification.** We provide theoretical justification for the decoupled design of DDIS by analyzing failure modes of joint-embedding diffusion models. In particular, we show that under precise geometric conditions, (i) joint-embeddings suffer guidance attenuation, (ii) DAPS fails due to sparse-guidance collapse, and (iii) DDIS avoids these limitations.
- **Empirical Performance.** DDIS achieves state-of-the-art accuracy and runtime efficiency on three challenging inverse PDE problems under sparse supervision, improving ℓ_2 error by 11% and spectral error by 54% on average; the advantage increases to 40% under data scarcity.

Organization. Sec. 2 introduces preliminaries. Sec. 3 presents the proposed Decoupled Diffusion Inverse Solver (DDIS). Sec. 4 analyzes joint-embedding failures and advantages of decoupling. Sec. 5 reports benchmark experimental results. Related work is deferred to App. C. Design rationales, ablations, and supplementary experiments appear in App. A and B.

2 Posterior Sampling for Inverse Problems

We employ diffusion models (Ho et al., 2020; Song et al., 2020a;b) to learn a prior distribution $p(a)$ over the coefficient space. The diffusion model defines forward and reverse stochastic processes that enable sampling from $p(a)$ by denoising from Gaussian noise. We refer readers to App. D for a detailed exposition of diffusion models and the diffusion prior.

Given $a = a_0$, the observation model (1.1) specifies a likelihood $p(u_{\text{obs}} \mid a_0)$. Combining with the diffusion prior, our goal is to sample from the posterior distribution

$$p(a_0 \mid u_{\text{obs}}) \;\propto\; p(a_0)\, p(u_{\text{obs}} \mid a_0), \tag{2.1}$$

where $p(a_0)$ is the prior and $p(u_{\text{obs}} \mid a_0)$ the likelihood. Posterior sampling requires modifying the prior score $\nabla_{a_t} \log p(a_t)$ in the reverse process (D.2) to the posterior score $\nabla_{a_t} \log p(a_t \mid u_{\text{obs}})$. Yet, direct posterior score computation is intractable due to the dependence on the unknown $p(a_0 \mid a_t)$. Thus, practical posterior samplers rely on approximations or auxiliary latent transitions (Kawar et al., 2021; 2022; Dou & Song, 2024; Wu et al., 2024). Before introducing our approach, we briefly review two representative methods: DPS (Chung et al., 2022) and DAPS (Zhang et al., 2025).

Diffusion Posterior Sampling (DPS) DPS decomposes the posterior score $\nabla_{a_t} \log p(a_t \mid u_{\text{obs}})$ into the sum of the unconditional score and a likelihood gradient. Since the likelihood $p(u_{\text{obs}} \mid a_t)$ requires marginalizing over the unknown conditional $p(a_0 \mid a_t)$, DPS approximates it by a mean estimate $\mathbb{E}[a_0 \mid a_t]$. The approximation introduces Jensen gap and results in over-smoothed reconstructions, as shown in Rmk. D.1 and Fig. 3. A full derivation of DPS is provided in App. D.2.

Decoupled Annealing Posterior Sampling (DAPS) Posterior sampling is naturally formulated as a correction on the noisy latent variable a_t, yet the likelihood term $p(u_{\text{obs}} \mid a_0)$ is defined on the clean variable a_0. DPS forces the correction to act on a_t, facing the intractability issue of $p(a_0 \mid a_t)$. DAPS instead separates these roles for updating a_t to a_{t-1}: it first estimates the clean variable

Table 1: Comparison between DDIS and joint-embedding diffusion methods.

	DDIS	**Joint-Embedding Models**
Training Stage		
Architecture	$G_\theta : \mathcal{Z} \to \mathcal{A}$ and $L_\phi : \mathcal{A} \to \mathcal{U}$	$H_\psi : \mathcal{Z} \to \mathcal{A} \times \mathcal{U}$
Objective	model $p(a)$ and learn $L(\cdot)$	model $p(a, u)$
Data Utilization	prior: $n_u + n_p$; operator: n_p	n_p only
Complexity[‡]	prior: d_P; operator: d_L	$d_J \geq \max(d_P, d_L)$
Generalization Bound[‡]	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{d_L}{n_p}} + \sqrt{\frac{d_P}{n_p + n_u}}\right)$	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{d_J}{n_p}}\right)$
Physics	explicit via $L_\phi(a)$	statistical correlations
Posterior Sampling Stage		
Target	$p(a \mid u_{\text{obs}})$	$p(a \mid u_{\text{obs}})$ via marginalizing out u
Sampler	DAPS (Zhang et al., 2025)	DPS (Chung et al., 2022)
Likelihood	$\log p(u_{\text{obs}} \mid \widehat{a}_{0\|u})$	$\log p(u_{\text{obs}} \mid \mathbb{E}[a_0 \mid a_{t+1}])$
Bias	asymptotically unbiased[†]	Jensen-gap bias
Spectral Feature	preserves high-frequency details	degraded at high frequencies
Inference Cost	low, diffusion forward + FNO backward	high, diffusion forward and backward

[†] Asymptotic with infinite Langevin steps and paired data; DAPS also uses Gaussian approximation in practice.
[‡] Detailed derivation in App. F.

a_0, then applies the likelihood-based correction to this clean estimate, and finally re-noises the corrected estimate to a_{t-1}. DAPS avoids the intractability issue in DPS and improves reconstruction quality in many inverse problems (Zheng et al., 2025). However, under sparse observations and joint-embeddings, the likelihood gradients act only locally and offer ineffective update (Rmk. D.2). We provide the details of its sparse-guidance failure in App. D.3.

3 Decoupled Diffusion Inverse Solver (DDIS)

DDIS separates the roles of priors and physics-induced likelihoods in a Bayesian inverse formulation. Instead of forcing a diffusion model to represent coefficient-solution correlations directly from paired data, DDIS learns these components independently in their native domains and integrates them only during sampling. DDIS consists of: (i) diffusion prior learning in coefficient space, (ii) neural operator learning of the forward physics, and (iii) DAPS-based physics-aware posterior sampling (at inference). Fig. 2 compares (i)+(ii) to previous joint-learning methods, and Fig. 4 visualizes (iii). Moreover, each component offers immediate benefits: (i) scalable prior learning using abundant prior-only data, (ii) explicit physics consistency, and (iii) dense guidance in posterior sampling, even under sparse observations. We elaborate on each component below.

3.1 Diffusion Prior in Coefficient Space

We train a diffusion prior $p(a)$ over PDE coefficients using a score-based diffusion model $\boldsymbol{s}_\theta(a_t, t)$ based on noisy a_t samples. Training follows standard noise-prediction loss without paired supervision:

$$\mathcal{L}_{\text{prior}} = \mathbb{E}_{a,\epsilon,t}\left[\|\boldsymbol{s}_\theta(a_t, t) - \epsilon\|_2^2\right]. \tag{3.1}$$

3.2 Neural Operator as a Physics Surrogate

Given scarce paired (a, u) samples, we learn a surrogate of the forward map $L : \mathcal{A} \to \mathcal{U}$ using a neural operator L_ϕ. The model is trained via supervised regression with an optional physics regularizer

$$\mathcal{L}_{\text{operator}} = \mathbb{E}_{(a,u)}\left[\|L_\phi(a) - u\|_2^2 + \overbrace{\lambda\|\text{Res}(L_\phi(a), a)\|_2^2}^{\text{PINO regularization}}\right], \tag{3.2}$$

where $\text{Res}(\cdot)$ evaluates PDE residual, weighted by λ.

The forward PDE problem $L : \mathcal{A} \to \mathcal{U}$ is well-posed, making it a natural target for neural operator learning (Kovachki et al., 2023). Unlike joint-embedding diffusion models that must reconstruct the full coefficient-solution correlations, the neural operator is trained only to represent the forward

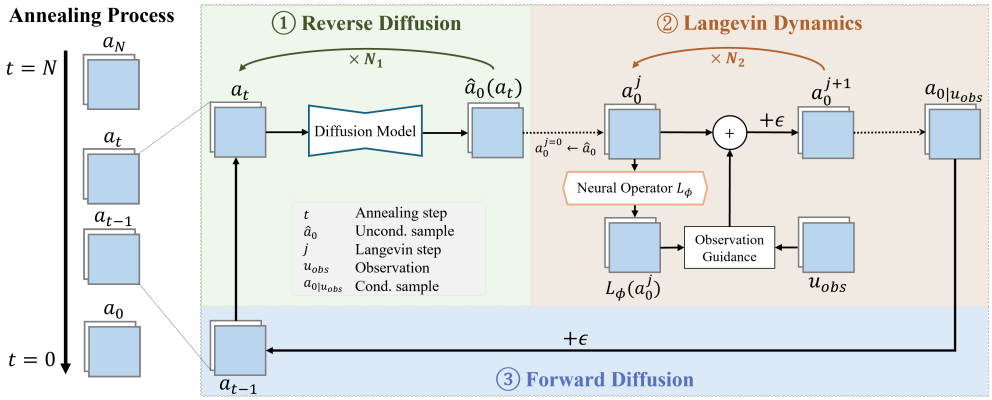

Figure 4: DDIS sampling process. Each annealing step alternates between ① **reverse diffusion**, which estimates an unconditional sample $\widehat{a}_0(a_t)$ by diffusion model; ② **Langevin dynamics** guided by the neural operator L_ϕ to enforce physics consistency; and ③ **forward diffusion**, which re-injects noise for the next annealing level. The process iteratively refines the posterior sample $a_0 \mid u_{\text{obs}}$.

physics. This reduces reliance on paired data and shifts learning complexity into the coefficient prior. Operating directly in function space, neural operators are resolution-invariant and can exploit low- or multi-resolution paired data while supporting high-resolution inference. Moreover, the optional PDE-residual regularization term in (3.2) further reduces the need for paired supervision, potentially enabling purely physics-based training. Collectively, these characteristics enable DDIS to maximize data utilization and achieve data efficiency in multiple regimes.

3.3 Physics-Aware Posterior Sampling via DAPS

Given sparse observations $u_{\text{obs}} = M \odot L(a) + \epsilon$, DDIS performs posterior inference using DAPS (App. D.3). DAPS operates on the observation-conditioned time marginals $p_t(a_t \mid u_{\text{obs}})$ along the reverse diffusion process. As the noise level t decreases, these marginals anneal toward the desired posterior $p(a_0 \mid u_{\text{obs}})$ at $t = 0$.

At each diffusion timestep, DAPS combines the diffusion prior with likelihood-based guidance to approximate sampling from $p_t(a_t \mid u_{\text{obs}})$. In DDIS, this guidance is instantiated through the neural operator surrogate L_ϕ and the denoised estimate $\widehat{a}_0(a_t)$ induced by the diffusion prior. Concretely, given the current latent a_t, DDIS applies Langevin MCMC loop (D.7):

$$\begin{aligned} a_0^{(j+1)} =& a_0^{(j)} - \eta \nabla_{a_0^{(j)}} \frac{\|a_0^{(j)} - \widehat{a}_0(a_t)\|^2}{r_t^2} \\ & - \eta \nabla_{a_0^{(j)}} \frac{\|M \odot L(a_0^{(j)}) - u_{\text{obs}}\|^2}{2\beta_y^2} + \sqrt{2\eta}\epsilon_j, \end{aligned} \tag{3.3}$$

where $\epsilon_j \sim \mathcal{N}(0, I)$, $\eta > 0$ is step size, r_t and β_y are prior and likelihood scale. The first term enforces consistency with the diffusion prior, while the second injects physics-aware guidance via the neural operator. After N Langevin steps, DAPS samples the next latent $a_{t-1} \sim p_{t-1}(a_{t-1} \mid u_{\text{obs}})$ by re-noising a_0^N according to (D.9). Iterating this procedure yields a trajectory towards the desired posterior. Alg. 1 shows the complete sampling pipeline.

The neural operator plays a critical role in enabling effective DAPS-based posterior sampling under sparse observations. While prior joint-embedding models suffer from sparse-guidance failure (Rmk. D.2), the global receptive field of L_ϕ propagates observation errors across the entire coefficient space, ensuring dense guidance. As a result, DDIS produces sharp, effective, and physics-aware samples without retraining the diffusion prior for new observation patterns. App. A further discusses design rationale and choices.

4 Theoretical Analysis

This section serves two purposes: (i) to formalize the guidance attenuation phenomenon that motivates decoupled designs, and (ii) to provide a comparison between joint and decoupled formulations, explaining the empirical results in Sec. 5. We begin by analyzing guidance behavior under data

scarcity for joint-embedding and decoupled models (Sec. 4.1). Particularly, the guidance in joint models provably attenuates (Sec. 4.1.1), whereas our DDIS preserves robust guidance (Sec. 4.1.2). We further analyze a naive application of DAPS in joint models and prove its sparse-guidance failure under sparse observations (Sec. 4.2). We also show that the decoupled design offers a tighter generalization bound than joint-embedding in App. F.

4.1 Guidance Behaviors under Data Scarcity

Assume observations are generated by a physical operator with additive homoscedastic noise:

$$u_{\text{obs}} = L^*(a) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_{\text{obs}}^2 I), \tag{4.1}$$

where $L^* : \mathcal{A} \to \mathcal{U}$ is the true forward operator and σ_{obs}^2 is a fixed sensor noise variance. Thus, the ideal inference-time guidance is given by the likelihood score

$$\nabla_a \log p(u_{\text{obs}}|a) = \frac{1}{\sigma_{\text{obs}}^2} \big(\nabla_a L^*(a)\big)^\top \big(u_{\text{obs}} - L^*(a)\big), \tag{4.2}$$

which defines a non-zero gradient on the function space $\mathcal{A}$, with fixed noise scale $1/\sigma_{\text{obs}}^2$.

4.1.1 Guidance Attenuation in Joint-Embedding Models

We first summarize why guidance necessarily attenuates in joint-embedding diffusion models under data scarcity; full derivations are deferred to App. E.

Joint-embedding methods learn a diffusion prior over the joint variable $x = (a, u)$ and impose observations with

$$u_{\text{obs}} = M x_0 + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \widetilde{\sigma}_{\text{obs}}^2 I).$$

For Diffusion Posterior Sampling (DPS), the likelihood gradient with respect to the noisy state x_t is approximated using a denoised estimate $\widehat{x}_0(x_t, t) = \mathbb{E}[x_0 \mid x_t]$,

$$\nabla_{x_t} \log p(u_{\text{obs}} \mid x_t) \approx \frac{1}{\widetilde{\sigma}_{\text{obs}}^2} J_{\widehat{x}_0}(x_t, t)^\top M^\top \big(u_{\text{obs}} - M\widehat{x}_0(x_t, t)\big). \tag{4.3}$$

We define the *scale-free guidance* (Def. E.1):

$$g(x_t, t) \;:=\; J_{\widehat{x}_0}(x_t, t)^\top M^\top r(x_t, t), \quad r = u_{\text{obs}} - M\widehat{x}_0. \tag{4.4}$$

Using the Tweedie estimator for score-based diffusion, $\widehat{x}_0(x_t, t) = \alpha_t^{-1}(x_t + \sigma_t^2 s_\theta(x_t, t))$, the guidance admits the block decomposition (Lem. E.2):

$$g(x_t, t) \propto \begin{pmatrix} \sigma_t^2\, \partial_{a_t} s_{\theta,u}(x_t, t)^\top \\ I + \sigma_t^2\, \partial_{u_t} s_{\theta,u}(x_t, t)^\top \end{pmatrix} r(x_t, t). \tag{4.5}$$

The coefficient update g_a depends on the cross-partial $\partial_{a_t} s_{\theta,u}$; thus, observations can influence a only through learned a-u coupling in the joint score model.

Under data scarcity, the learned score is well approximated by an empirical Gaussian mixture (Lem. E.3). Let $\varphi_n(x)$ denote the n-th component density and $w_n(x) = \varphi_n(x)/\sum_j \varphi_j(x)$ its responsibility. Since a single isotropic Gaussian cannot couple a and u (Rmk. E.1), all cross-component guidance arises solely through the responsibility gradients $\partial_a w_n$ (Lems. E.5 and E.6). We therefore analyze when these gradients vanish or remain nontrivial.

Theorem 4.1 (Local dominance $\Rightarrow$ vanishing responsibility gradients (informal Thm. E.1))**.** Fix (x, t). If there exists $k \in [N]$ such that $\varphi\big(x - \alpha_t x_0^{(k)}\big) \gg \varphi\big(x - \alpha_t x_0^{(j)}\big),\ \forall j \neq k$, then all responsibility gradients vanish: $\partial_a w_n(x, t) \approx 0,\ \forall n \in [N]$.

This yields a sufficient condition for guidance attenuation.

Corollary 4.1.1 (Locality implies attenuation (informal Corollary E.1.1))**.** If x_t lies close to a single mixture center $\alpha_t x_0^{(k)}$, the a-component of the scale-free guidance attenuates by $g_a(x_t, t) \approx 0$.

The first result shows that, under the locality condition, responsibility gradients sufficiently vanish. We now state the complementary: nontrivial responsibility gradients can occur only when x lies in an overlap region of the mixture (informal Thm. E.2).

Theorem 4.2 (Non-vanishing responsibility gradients require overlap)**.** Fix (x, t). If $\left\|\partial_a w_n(x,t)\right\| > 0$ for some $n \in [N]$, then there exist $p \neq q$ such that $\left|\|x - \alpha_t x_0^{(p)}\|_2^2 - \|x - \alpha_t x_0^{(q)}\|_2^2\right| \lesssim \sigma^2(t)$.

We then have a necessary condition for non-zero guidance.

Corollary 4.2.1 (Non-vanishing guidance requires overlap (informal Corollary E.2.1))**.** If the a-component of the scale-free guidance is nonzero $\|g_a(x_t, t)\| > 0$, then x_t must lie in an overlap region of the mixture.

Thms. 4.1 and 4.2 offer a clear geometric interpretation of guidance behavior in joint-embedding diffusion models during posterior sampling. Guidance attenuates when x lies far from all mixture centers and also when x is dominated by a single component (Corollary 4.1.1). Non-vanishing guidance is possible only when x lies in an overlap region where at least two mixture components have comparable responsibility (Corollary 4.2.1). We provide a visualization for the two criteria in Fig. 1. Under data scarcity, such overlap regions may be rare or entirely absent, causing coefficient-space guidance to collapse throughout the sampling process. This failure mode is intrinsic to joint-embedding formulations and cannot be remedied by alternative noise schedules or sampling heuristics. These observations motivate a decoupled formulation in which observation consistency is enforced through an explicit forward operator rather than learned joint correlations.

4.1.2 Guidance Robustness in DDIS

Unlike joint-embedding formulations, DDIS enforces observation consistency through a neural operator rather than learned joint correlations. As a result, likelihood-based guidance in DDIS is mediated directly by the Jacobian of a deterministic operator and does not rely on data-dependent coupling between coefficients and solutions.

DDIS approximates the physical operator by a neural operator $L_\phi \approx L^*$ and performs posterior sampling via DAPS, where likelihood corrections are applied to clean-variable estimates $a_0^{(j)}$. For a noise scale $\widetilde{\sigma}_{\text{obs}}$, the resulting guidance approximates (4.2) by

$$\frac{1}{\widetilde{\sigma}_{\text{obs}}^2}\left(\nabla_a L_\phi(a_0^{(j)})\right)^\top \left(u_{\text{obs}} - L_\phi(a_0^{(j)})\right), \tag{4.6}$$

which directly approximates the ideal likelihood score (4.2).

Here, the magnitude of this guidance is governed by the explicit Jacobian $\nabla_a L_\phi(a)$. Since L_ϕ is trained via regression independently of any joint data density, data scarcity affects approximation accuracy but does not induce guidance attenuation, unlike joint-embedding models.

Proposition 4.1 (Structural robustness of DDIS guidance)**.** In contrast to joint-embedding models, DDIS admits no data-dependent mechanism that forces guidance attenuation as the number of paired data decreases.

4.2 Joint Embeddings and DAPS: Failure Modes with Sparse Guidance

In this section, we analyze a structural failure mode of applying DAPS (Zhang et al., 2025) with joint embeddings. The Langevin dynamics inside DAPS create spatially discontinuous gradient updates under sparse observations. These updates alter the covariance structure of generated samples. Thus, the sampling processes produce outputs that are out-of-distribution relative to the diffusion model's learned manifold, resulting in poor performance. Extending DAPS to function spaces does not resolve this issue. We present an informal derivation here; the rigorous one is in App. G.

The DAPS inner Langevin chain (Eq. (D.7)) targets a density proportional to $p(x \mid x_t)\, p(u_{\text{obs}} \mid x)$, where $x = (a, u)$ is the joint variable. We extend DAPS's Gaussian assumptions to function spaces as $p(x \mid x_t) \approx \mathcal{N}(\widehat{x}_0(x_t), C)$, where $\widehat{x}_0$ is an estimator of x_0 given x_t and C is a covariance function. Under sparse observations, it suffices to analyze a single point observation $x_i = c$ (well-separated constraints contribute approximately additively), which we model as $p(u_{\text{obs}} \mid x) \approx \mathcal{N}(c, \sigma_s^2)$. Taking the continuous-time limit yields the preconditioned Langevin SDE:

$$dx_t = -\Sigma\, \nabla U(x_t)\, dt + \sqrt{2}\Sigma^{1/2} dW_t, \tag{4.7}$$

where Σ is the noise covariance and the potential $U(x)$ combines the GRF prior and the pointwise constraint:

$$U(x) = \frac{1}{2}(x - x_0)^T C^{-1}(x - x_0) + \frac{1}{2\sigma_s^2}(e_i^T x - c)^2. \tag{4.8}$$

Table 2: Standard supervision results for the inverse PDE problems under 3% sparse observations. Relative ℓ_2 error (%), spectral error E_s, and runtime per sample are reported. Rows are grouped by comparable time budgets. Fig. 5 visualizes the table.

Method	Poisson		Helmholtz		N-S		Time (s)
	ℓ_2	E_s	ℓ_2	E_s	ℓ_2	E_s	
DiffusionPDE	74.68	.566	46.10	.315	32.78	2.099	17.75
FunDPS	19.96	.192	17.16	.140	8.99	.382	14.58
DDIS	**15.78**	**.074**	**15.08**	**.044**	**8.93**	**.165**	16.75
DiffusionPDE	35.53	.281	23.86	.164	11.01	.221	35.63
FunDPS	17.14	.135	16.05	.157	8.23	.423	28.75
DDIS	**14.36**	**.086**	**14.24**	**.055**	**8.02**	**.178**	32.67
DiffusionPDE	16.65	**.085**	17.73	**.096**	9.24	.219	142.62
FunDPS	14.73	.136	14.14	.198	7.98	.479	113.83
DDIS	**12.32**	.087	**12.20**	.097	**7.81**	**.188**	127.42
ECI-sampling*	94.63	/	92.83	/	42.36	/	0.20
ECI-sampling*	93.47	/	93.23	/	41.68	/	0.38
OFM*	71.87	/	49.60	/	37.57	/	470.44
OFM*	47.04	/	42.07	/	20.98	/	4366.34

Table 3: Relative ℓ_2 error (%) under scarce paired-data supervision (100%, 5%, and 1% data) for inverse PDE problems. Bold numbers highlight better performance under same data scarcity. Two models in the same group share comparable time budgets as in Tab. 2.

Data Scarcity	Poisson		Helmholtz		N-S		T (s)
	DDIS	Fun.‡	DDIS	Fun.‡	DDIS	Fun.‡	
100%	**16.36**	20.47	**15.19**	17.16	**8.22**	8.48	16
5%	**17.28**	23.24	**15.69**	20.97	**9.21**	11.56	
1%	**18.70**	35.81	**16.40**	41.69	**12.05**	13.65	
1%+Phys	**16.56**	35.81	**16.05**	41.69	/ †	/ †	
100%	**15.34**	17.30	**14.03**	16.05	8.00	**7.67**	32
5%	**15.76**	22.66	**15.10**	19.49	**9.12**	11.15	
1%	**17.79**	35.79	**15.90**	41.44	**12.28**	13.50	
1%+Phys	**15.63**	35.81	**15.21**	41.69	/ †	/ †	

† Physics loss is not applicable for Navier-Stokes due to insufficient information; see App. H.1.
‡ Fun. stands for FunDPS.
* See App. A.5 for analysis of the flow models.

We denote by Σ_∞ the stationary covariance of x_t.

Theorem 4.3 (Sparse constraint induces correlation shrinkage (informal Thm. G.1))**.** Under the dynamics induced by Eqs. (4.7) and (4.8), all covariance terms involving the constrained index i are uniformly scaled:

$$(\Sigma_\infty)_{ik} = (\Sigma_\infty)_{ki} = \frac{\sigma_s^2}{\sigma_s^2 + C_{ii}} C_{ik}, \quad \forall k \tag{4.9}$$

Corollary 4.3.1 (Strong sparse guidance collapses correlations)**.** If $\sigma_s^2 \ll C_{ii}$, then

$$(\Sigma_\infty)_{ik} = (\Sigma_\infty)_{ki} \approx 0 \quad \forall k.$$

In practice, sparse-guidance methods often set σ_s^2 to a small value (e.g., 10^{-3} (Zhang et al., 2025)), so the constrained point rapidly becomes (nearly) uncorrelated from the rest.

This covariance collapse has a geometric interpretation. Physical coefficient fields are typically continuous and correlated across space, but under sparse constraints, the constrained location becomes nearly independent of its neighbors, yielding discontinuities that are atypical under the diffusion prior. As a result, the sample is pushed off the data manifold, which degrades reconstruction quality.

5 Experiments

Task. We consider inverse Helmholtz, Poisson, and Navier-Stokes problems with sparse observations. For each instance, we observe 500 randomly sampled solution points from $u(x) \in \mathbb{R}^{128^2}$ ($\sim$3% of the domain) and reconstruct the underlying coefficient field $a(x)$. Performance is measured by relative ℓ_2 error and spectral energy error E_s between reconstructed and ground-truth coefficient.

To reflect practical scenarios of available data, we train the neural operator under three supervision regimes: (1) *Standard supervision*, where the operator is trained on full-resolution 128^2 paired data; (2) *Scarce paired-data supervision*, where the operator is trained using only 5% or 1% of the full dataset; and (3) *Low-/multi-resolution supervision*, where the operator is trained either on full 64^2 low-res data, or on a mixed dataset consisting of 64^2 data and 10% of 128^2 high-res samples.

Benchmark. We compare DDIS against prior diffusion-based solvers and variants, including DiffusionPDE (Huang et al., 2024), previous state-of-the-art FunDPS (Yao et al., 2025), recent flow-based ECI-sampling (Cheng et al., 2025) and OFM (Shi et al., 2025). For ablations, we test FunDAPS, a baseline where the posterior sampler in FunDPS is replaced by DAPS, and DecoupledDPS, which applies DPS sampling within our modular DDIS framework. This setup allows us to isolate the effects of (i) the sampling scheme (DPS vs. DAPS) and (ii) the architectural choice of decoupling prior learning from operator-informed sampling.

Standard supervision. We compare our DDIS with the benchmarks DiffusionPDE (Huang et al., 2024) and FunDPS (Yao et al., 2025). Besides relative ℓ_2 error, we report a spectral error E_s to assess model's ability to reconstruct features across different spatial frequencies, measured by wave number k. Since the power spectrum scales logarithmically as k decreases, directly averaging errors would be heavily biased toward low-frequency modes. We thus take the geometric mean of the spectral error over all k. DDIS achieves state-of-the-art performance on the inverse Poisson, Helmholtz, and Navier-Stokes problems across various budgets (Tab. 2).

As the number of sampling steps increases, DDIS continues to improve, whereas the baseline models saturate. As shown in Fig. 5, DDIS lies near the accuracy-runtime Pareto frontier for both Poisson and Helmholtz. Across all tasks, including Navier-Stokes where ℓ_2 differences are small, DDIS achieves substantially lower spectral error than FunDPS by a factor up to 3.2. We visualize the spectral error in App. B.2, showing that FunDPS fails to capture high-frequency features, whereas DDIS preserves fidelity across frequencies.

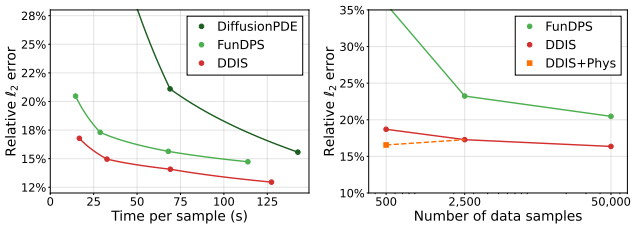

Figure 5: Pareto frontiers for the inverse Poisson problem. (Left) Relative ℓ_2 error versus inference time. (Right) Relative ℓ_2 error versus training set size. DDIS dominates baselines across both computational and data efficiency dimensions, with the physics-informed variant (DDIS+Phys) further enhances sample efficiency.

Scarce paired-data supervision. To assess the data-efficiency of DDIS, we consider data-scarcity scenarios with 100%, 5%, and 1% of the original paired-data supervision and FunDPS as the benchmark model, keeping the observation sparsity at around 3%. As paired data become scarce, FunDPS degrades sharply, while DDIS maintains stable reconstruction accuracy even at 1% paired data, indicating superior data efficiency (Tab. 3). We further observe that DecoupledDPS improves over FunDPS (Tab. 6), but DDIS achieves larger gains, highlighting the additional benefit of DAPS enabled by decoupling. This confirms our claim: a deterministic representation of physics is more appropriate and data-efficient than joint statistical modeling.

Low-/Multi-resolution supervision To examine the resolution-invariant properties of DDIS, we evaluate performance when the neural operator is trained on low-resolution (64^2) data or on a mixed dataset combining 64^2 data with 10% of the 128^2 samples.

As shown in Tab. 5, DDIS achieves the best accuracy when trained on full 128^2 data, while training on low- or mixed-resolution data results in only modest degradation. This demonstrates DDIS's robustness to resolution enabled by the resolution-invariant neural operator.

6 Conclusion and Discussion

We proposed Decoupled Diffusion Inverse Solver (DDIS), a physics-aware generative framework for inverse PDE problems under sparse observations and imbalanced data regimes. Unlike prior joint-embedding approaches, DDIS separates prior modeling in coefficient space from physics-based likelihood evaluation via a neural operator, aligning with Bayesian inverse formulations (Sec. 3).

Our central finding is that joint-embedding models fail to provide effective cross-field guidance under data scarcity or sparse sensor layouts (Sec. 4). Specifically, we characterize geometric conditions under which joint models suffer guidance attenuation (Sec. 4.1) and identify a failure mode of joint models with DAPS-based sampling under sparse observation (Sec. 4.2). By decoupling, DDIS provides reliable guidance via the neural operator (Sec. 3.3).

Empirically (Sec. 5), DDIS achieves state-of-the-art performance. Across budgets, DDIS creates new accuracy-runtime Pareto frontiers (Fig. 5) and attains lower spectral error (Tab. 2). Under scarce paired-data supervision, DDIS remains stable down to 1% paired data, whereas joint-embedding methods degrade (Tab. 3). DDIS is also robust to low- and mixed-resolution supervision (Tab. 4).

ACKNOWLEDGMENTS

Anima Anandkumar is supported in part by Bren endowed chair, ONR (MURI grant N00014-23-1-2654), and the AI2050 senior fellow program at Schmidt Sciences. Jiachen Yao is supported in part by the Naren and Vinita Gupta Fellowship.

REFERENCES

Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P Bruinsma, Tom R Andersson, Michael Herzog, Nicholas D Lane, Matthew Chantry, J Scott Hosking, et al. End-to-end data-driven weather prediction. *Nature*, 641(8065):1172–1179, 2025.

Marc Amorós-Trepat, Luis Medrano-Navarro, Qiang Liu, Luca Guastoni, and Nils Thuerey. Guiding diffusion models to reconstruct flow fields from sparse data. *Physics of Fluids*, 38(1), January 2026. ISSN 1089-7666. doi: 10.1063/5.0304492. URL http://dx.doi.org/10.1063/5.0304492.

Marcin Andrychowicz, Lasse Espeholt, Di Li, Samier Merchant, Alexander Merose, Fred Zyda, Shreya Agrawal, and Nal Kalchbrenner. Deep learning for day forecasts from sparse observations. *arXiv preprint arXiv:2306.06079*, 2023.

Ricardo Baptista, Agnimitra Dasgupta, Nikola B Kovachki, Assad Oberai, and Andrew M Stuart. Memorization and regularization in generative diffusion models. *arXiv preprint arXiv:2501.15785*, 2025.

James V Beck, Ben Blackwell, and Charles R St Clair. *Inverse heat conduction: Ill-posed problems*. James Beck, 1985.

Heli Ben-Hamu, Omri Puny, Itai Gat, Brian Karrer, Uriel Singer, and Yaron Lipman. D-flow: Differentiating through flows for controlled generation, 2024. URL https://arxiv.org/abs/2402.14017.

Julius Berner, Miguel Liu-Schiaffini, Jean Kossaifi, Valentin Duruisseaux, Boris Bonev, Kamyar Azizzadenesheli, and Anima Anandkumar. Principled approaches for extending neural architectures to function spaces for operator learning. *arXiv preprint arXiv:2506.10973*, 2025.

Gabriel Cardoso, Yazid Janati El Idrissi, Sylvain Le Corff, and Eric Moulines. Monte carlo guided diffusion for bayesian linear inverse problems. *arXiv preprint arXiv:2308.07983*, 2023.

Wuyang Chen, Jialin Song, Pu Ren, Shashank Subramanian, Dmitriy Morozov, and Michael W Mahoney. Data-efficient operator learning via unsupervised pretraining and in-context learning. *Advances in Neural Information Processing Systems*, 37:6213–6245, 2024.

Chaoran Cheng, Boran Han, Danielle C. Maddix, Abdul Fatir Ansari, Andrew Stuart, Michael W. Mahoney, and Bernie Wang. Gradient-free generation for hard-constrained systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=teE4pl9ftK.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Richard Courant, Kurt Friedrichs, and Hans Lewy. On the partial difference equations of mathematical physics. *IBM journal of Research and Development*, 11(2):215–234, 1967.

Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024.

Ranjan Das. A simulated annealing-based inverse computational fluid dynamics model for unknown parameter estimation in fluid flow problem. *International Journal of Computational Fluid Dynamics*, 26(9-10):499–513, 2012.

Daniel T Davis, Zhengxiao Chen, Jenq-Neng Hwang, Leung Tsang, and Eni Njoku. Solving inverse problems by bayesian iterative inversion of a forward model with applications to parameter mapping using smmr remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 33(5): 1182–1193, 1995.

J.R. Dormand and P.J. Prince. A family of embedded runge-kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, 1980. ISSN 0377-0427.

Zehao Dou and Yang Song. Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

Valentin Duruisseaux, Jean Kossaifi, and Anima Anandkumar. Fourier neural operators explained: A practical perspective, 2025. URL https://arxiv.org/abs/2512.01421.

Tiffany Fan, Kailai Xu, Jay Pathak, and Eric Darve. Solving inverse problems in steady-state navier-stokes equations using deep neural networks. *arXiv preprint arXiv:2008.13074*, 2020.

A Gholami and HR Siahkoohi. Regularization of linear and non-linear geophysical ill-posed problems with joint sparsity constraints. *Geophysical Journal International*, 180(2):871–882, 2010.

James Gregson, Ivo Ihrke, Nils Thuerey, and Wolfgang Heidrich. From capture to simulation: connecting forward and inverse problems in fluids. *ACM Transactions on Graphics (ToG)*, 33(4): 1–11, 2014.

Charles W Groetsch and CW Groetsch. *Inverse problems in the mathematical sciences*, volume 52. Springer, 1993.

Felix J Herrmann and Xiang Li. Efficient least-squares imaging with sparsity promotion and compressive sensing. *Geophysical prospecting*, 60(4-Simultaneous Source Methods for Seismic Data):696–712, 2012.

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

Jiahe Huang, Guandao Yang, Zichen Wang, and Jeong Joon Park. Diffusionpde: Generative pde-solving under partial observation. *arXiv preprint arXiv:2406.17763*, 2024.

Christian Jacobsen, Yilin Zhuang, and Karthik Duraisamy. Cocogen: Physically consistent and conditioned score-based generative models for forward and inverse problems. *SIAM Journal on Scientific Computing*, 47(2):C399–C425, 2025.

Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems*, 34:21757–21769, 2021.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35:23593–23606, 2022.

Gavin Kerrigan, Giosue Migliorini, and Padhraic Smyth. Functional flow matching, 2023. URL https://arxiv.org/abs/2305.17209.

Jeongsol Kim, Bryan Sangwoo Kim, and Jong Chul Ye. Flowdps: Flow-driven posterior sampling for inverse problems, 2025. URL https://arxiv.org/abs/2503.08136.

VS Kostsov. General approach to the formulation and solution of the multi-parameter inverse problems of atmospheric remote sensing with measurements and constraints of different types. *International journal of remote sensing*, 36(11):2963–2994, 2015.

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.

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference*, pp. 1–11, 2023.

Shingyu Leung, Jianliang Qian, and Jiangtao Hu. A level-set adjoint-state method for transmission traveltime tomography in irregular domains. *SIAM Journal on Scientific Computing*, 43(3):A2352–A2380, 2021.

Zeyu Li, Hongkun Dou, Shen Fang, Wang Han, Yue Deng, and Lijun Yang. Physics-aligned field reconstruction with diffusion bridge. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=D042vFwJAM.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for partial differential equations' operator learning. *arXiv preprint arXiv:2205.13671*, 2022.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier neural operator with learned deformations for pdes on general geometries. *Journal of Machine Learning Research*, 24(388):1–26, 2023.

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/JMS Journal of Data Science*, 1(3):1–27, 2024.

Lu Lu, Pengzhan Jin, and George Em Karniadakis. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193*, 2019.

Peter Manshausen, Yair Cohen, Peter Harrington, Jaideep Pathak, Mike Pritchard, Piyush Garg, Morteza Mardani, Karthik Kashinath, Simon Byrne, and Noah Brenowitz. Generative data assimilation of sparse weather station observations at kilometer scales. *arXiv preprint arXiv:2406.16947*, 2024.

Morteza Mardani, Enhao Gong, Joseph Y Cheng, Shreyas Vasanawala, Greg Zaharchuk, Marcus Alley, Neil Thakur, Song Han, William Dally, John M Pauly, et al. Deep generative adversarial networks for compressed sensing automates mri. *arXiv preprint arXiv:1706.00051*, 2017.

Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*, 2023.

Rebecca E Morss, Kerry A Emanuel, and Chris Snyder. Idealized adaptive observation strategies for improving numerical weather prediction. *Journal of the Atmospheric Sciences*, 58(2):210–232, 2001.

Jennifer L Mueller and Samuli Siltanen. *Linear and nonlinear inverse problems with practical applications*. SIAM, 2012.

Andrew D Parsekian, Kamini Singha, Burke J Minsley, W Steven Holbrook, and Lee Slater. Multiscale geophysical imaging of the critical zone. *Reviews of Geophysics*, 53(1):1–26, 2015.

Maitreya Patel, Song Wen, Dimitris N. Metaxas, and Yezhou Yang. Flowchef: Steering of rectified flow models for controlled generations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15308–15318, October 2025.

Grigorios A Pavliotis. Stochastic processes and applications. *Texts in applied mathematics*, 60, 2014.

R-E Plessix. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, 167(2):495–503, 2006.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Medha Sawhney, Abhilash Neog, Mridul Khurana, and Anuj Karpatne. Beyond loss guidance: Using pde residuals as spectral attention in diffusion neural operators. *arXiv preprint arXiv:2512.01370*, 2025.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950. ISSN 00034851. URL http://www.jstor.org/stable/2236561.

Yaozhong Shi, Zachary E Ross, Domniki Asimaki, and Kamyar Azizzadenesheli. Stochastic process learning via operator flow matching. *arXiv preprint arXiv:2501.04126*, 2025.

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.

Emil Y Sidky, Iris Lorente, Jovan G Brankov, and Xiaochuan Pan. Do cnns solve the ct inverse problem? *IEEE Transactions on Biomedical Engineering*, 68(6):1799–1810, 2020.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

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*, pp. 32483–32498. PMLR, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021.

Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

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

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Yanfei Wang. Regularization for inverse models in remote sensing. *Progress in physical geography*, 36(1):38–59, 2012.

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

Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16293–16303, 2022.

Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine Bouman. Principled probabilistic imaging using diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 37:118389–118427, 2024.

Xingyu Xu and Yuejie Chi. Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. *Advances in Neural Information Processing Systems*, 37:36148–36184, 2024.

Jiachen Yao, Abbas Mammadov, Julius Berner, Gavin Kerrigan, Jong Chul Ye, Kamyar Azizzadenesheli, and Anima Anandkumar. Guided diffusion sampling on function spaces with applications to pdes. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.17004.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20895–20905, 2025.

Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy Feng, Caifeng Zou, Yu Sun, Nikola Borislavov Kovachki, Zachary E Ross, Katherine Bouman, and Yisong Yue. Inversebench: Benchmarking plug-and-play diffusion models for scientific inverse problems. In *The Thirteenth International Conference on Learning Representations*, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1219–1229, 2023.

Appendix

A Design Rationale and Justification **16**
- A.1 Why not train conditional diffusion? . . . 16
- A.2 Why not use posterior sampling directly with numerical solver? . . . 16
- A.3 Why use DAPS-based posterior sampling? . . . 16
- A.4 Why not use joint-embedding with better posterior sampling? . . . 17
- A.5 Current Challenges in Flow-Based Posterior Sampling . . . 17

B Ablation Studies and Supplementary Results **18**
- B.1 Pareto Fronts: Accuracy-Speed Trade-offs . . . 18
- B.2 Power Spectrum Comparison . . . 19
- B.3 Mixed-resolution Training . . . 20
- B.4 Mixed-resolution Sampling . . . 20

C Related Works **20**

D Preliminaries **22**
- D.1 Diffusion Model . . . 22
- D.2 Diffusion Posterior Sampling (DPS) . . . 22
  - D.2.1 Extension: DecoupledDPS . . . 22
- D.3 Decoupled Annealing Posterior Sampling (DAPS) . . . 23
  - D.3.1 Asymptotic Guarantee for DAPS . . . 24

E Detailed Derivation of Guidance Attenuation in Joint-Embedding Models **25**

F Sample-Complexity Analysis **32**
- F.1 Hypothesis Class of Joint and Decoupled Models . . . 32
- F.2 Generalization Bound . . . 33

G Proofs of the Failure Modes of Joint Embeddings + DAPS **33**

H Experimental Setup **35**
- H.1 Dataset . . . 35
- H.2 Details on Decoupled Diffusion Inverse Solver . . . 35
- H.3 Details on Flow-based Models . . . 35
  - H.3.1 Prior Learning and Hyperparameters . . . 35
  - H.3.2 Inference Setup . . . 37
  - H.3.3 Quantitative Results and Analysis . . . 37
- H.4 Details on Other Baseline Methods . . . 37

I Qualitative Results **39**
- I.1 Inverse Poisson problem . . . 39
- I.2 Inverse Helmholtz problem . . . 40
- I.3 Inverse Navier-Stokes problem . . . 41
- I.4 Operator Flow Matching . . . 42

A Design Rationale and Justification

In this section, we analyze the design choices behind DDIS by addressing questions regarding alternative approaches in App. C. We provide theoretical justifications and empirical evidence to explain why DDIS adopts a decoupled generative framework with neural operator surrogates over other common strategies.

A.1 Why not train conditional diffusion?

A straightforward approach to inverse problems is to train a conditional diffusion model $p_\theta(a|u_{obs})$ directly. While effective for fixed tasks, this "supervised" strategy suffers from severe limitations in scientific contexts:

- **Combinatorial Generalization Failure:** Conditional models are tied to the specific observation mask M and forward operator L seen during training. Changing the sensor layout (mask) or the physics (operator) requires retraining the entire model from scratch.
- **Data Inefficiency:** Learning the conditional distribution requires massive paired datasets (a, u_{obs}) covering all potential measurement configurations. In contrast, DDIS learns the prior $p(a)$ unconditionally from unpaired data, allowing it to generalize to any observation mask and operator at inference time without retraining.

A.2 Why not use posterior sampling directly with numerical solver?

In theory, direct posterior sampling strategies could handle arbitrary, sparse observations by masking the exact numerical PDE solver during likelihood evaluation, i.e., computing $\log p(u_{obs} \mid M \odot L_{\text{num}}(a))$. However, this approach is impractical due to:

- **Backpropagation Computational Instability:** Backpropagating gradients through iterative solvers is known to suffer from computational instability and vanishing gradients (Zheng et al., 2025) even under dense observation.
- **Violation of Solver Stability Conditions:** The noisy intermediate samples generated during diffusion frequently violate strict stability conditions (e.g., CFL condition (Courant et al., 1967)), causing the solver to diverge or provide unreliable gradients (Zheng et al., 2025).
- **Ill-Conditioning under Sparsity:** These instabilities are amplified under sparse observations, where the gradient calculation becomes ill-conditioned (Plessix, 2006; Leung et al., 2021) and the solver fails to propagate observations back into a meaningful coefficient update.

In contrast, DDIS adopts learned neural operators in place of numerical solvers, avoiding the above failure modes and offering the following advantages.

- **Differentiability and Stability:** Neural operator provides a smooth, differentiable surrogate. It ensures stable gradient flow for guidance and enables rapid likelihood evaluation ($1000\times$ faster) during the iterative sampling process.
- **Dense Guidance:** Neural operator processes information globally using a spectral basis. This naturally "smears" the sparse pointwise errors across the spatial domain, converting sparse observations into the dense, global guidance required for stable Langevin updates.
- **Resolution Invariance:** As shown in Tab. 4, the FNO allows DDIS to be trained on multi-resolution data and sampling at resolutions different from the observation grid.

A.3 Why use DAPS-based posterior sampling?

We employ DAPS (Zhang et al., 2025) as our sampling backend after ablating other unsupervised posterior sampling strategies. Our selection is motivated by the specific limitations of alternative methods in non-linear PDE inverse problems:

- **Inapplicability of Decomposition Methods:** Decomposition-based methods such as DDRM (Kawar et al., 2022) and DDNM (Wang et al., 2022) rely on the linearity of the forward operator, so they are inapplicable to the non-linear PDE.
- **Bias in Guidance-Based Methods:** Guidance-based methods like DPS (Chung et al., 2022) and LGD (Song et al., 2023) rely on approximations of the intractable likelihood score. In practical implementations (e.g., using Tweedie's formula or limited Monte Carlo samples), these approximations introduce Jensen gap (Rmk. D.1) that results in higher error rates (Tab. 6) and over-smoothed reconstructions lacking high-frequency physical details (Fig. 3).

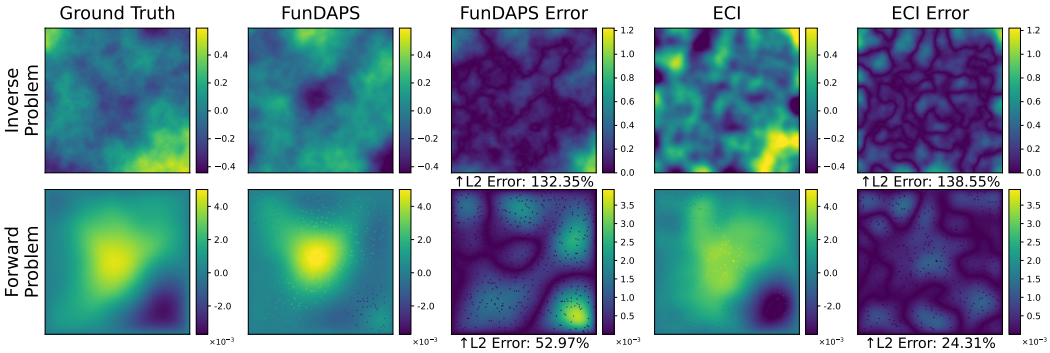

Figure 6: **Failure of FunDAPS and ECI-sampling under sparse observations. Top:** The joint-embedding model fails to recover the coefficient field despite using DAPS or ECI-sampling. **Bottom:** Forward error reveals overfitting to specific sensor locations (visible as dots), with no generalization to unobserved regions.

- **Inefficiency of Asymptotically Exact Samplers:** Asymptotically exact samplers like FPS (Dou & Song, 2024) and MCGDiff (Cardoso et al., 2023) avoid this approximation bias by using Sequential Monte Carlo (SMC) to properly weigh samples. However, the heavy computational costs make them prohibitive for high-dimensional PDE problems.

DAPS effectively functions as an efficient asymptotically exact sampler. Unlike SMC-based methods that require maintaining and resampling a large population to represent the posterior distribution, DAPS operates on a single sample trajectory and achieves the rigorous sampling quality of exact samplers, making it the our most Pareto-optimal choice.

A.4 Why not use joint-embedding with better posterior sampling?

A natural ablation is to integrate DAPS into joint-embedding framework like FunDPS—a baseline we term *FunDAPS*. However, our analysis (Sec. 4.2) shows that this combination fails under sparse observations. The failure comes from the joint-embedding design because:

- **Correlation does not imply Causation:** Joint-embedding models learn $p(a, u)$ via statistical correlations rather than an explicit causal mechanism. They lack a representation of physics operator to transport information between the solution space $\mathcal{U}$ and the coefficient space $\mathcal{A}$.
- **Sparse Guidance Failure:** As shown in Sec. 4.2 and Fig. 6, applying Langevin dynamics to a joint embedding (e.g., FunDAPS) under sparse constraints causes the stationary covariance to collapse. Without a differentiable operator to smooth these signals, the sampling trajectory is pushed off the data manifold.

A.5 Current Challenges in Flow-Based Posterior Sampling

A natural extension is to integrate flow-based inverse solvers, such as ECI-sampling (Cheng et al., 2025) and OFM regression (Shi et al., 2025), into a joint solution–coefficient embedding framework. However, our empirical analysis (Sec. 5; experimental details are provided in App. H.3) shows that these approaches exhibit clear limitations under sparse observations, either in terms of reconstruction performance or computational cost. These failures stem from structural issues in how joint embeddings are handled during inference, for the following reasons:

- **Heuristic Cross-Channel Coupling:** As shown in Tab. 2 and Fig. 6, ECI-sampling's heuristic replacement of u with observations—lacking both $\nabla_a \|M \odot u - u_{\text{obs}}\|^2$ to guide the a-channel and a manifold projection step—causes the inference trajectory to drift from the joint manifold, even destabilizing the u-channel's own manifold consistency. As also noted in PCFM (Utkarsh et al., 2025), ECI shows limited robustness in scenarios with discontinuities, making it less effective for our sparse and masked observations that require resolving high-frequency details from partial data.
- **High Computational Cost of Exact Langevin Inference.** As shown in Tab. 2 and Tab. 12, OFM regression requires a large number of Langevin iterations to obtain satisfactory posterior samples. Using accurate ODE solvers further necessitates backpropagation through the entire trajectory, resulting in substantial computational and memory overhead (App. H.3.3) and making it impractical.

Algorithm 1 DDIS-DAPS Sampler

Require: Sparse observation u_{obs}, mask M; diffusion model $\widehat{a}_0(\cdot)$; neural operator L_ϕ; noise schedule $\{\sigma(t_i)\}$; prior scale $\{r_{t_i}\}$; likelihood scale β_y; Langevin step size η and step count N_{L}.

$a_N \sim \mathcal{N}(0, I)$ {*Initialize from prior*}
for $i = N$ **to** 1 **do**
  $a_0^{(0)} \leftarrow \widehat{a}_0(a_i)$ {*Reverse diffusion (denoising)*}
  for $j = 0$ **to** $N_{\text{L}} - 1$ **do**
    $\epsilon_j \sim \mathcal{N}(0, I)$
    $g_{\text{prior}} \leftarrow -\nabla_{a_0^{(j)}} \dfrac{\|a_0^{(j)} - a_0^{(0)}\|_2^2}{r_{t_i}^2}$
    $g_{\text{like}} \leftarrow -\nabla_{a_0^{(j)}} \dfrac{\|M \odot L_\phi(a_0^{(j)}) - u_{\text{obs}}\|_2^2}{2\beta_y^2}$
    $a_0^{(j+1)} \leftarrow a_0^{(j)} + \eta\big(g_{\text{prior}} + g_{\text{like}}\big) + \sqrt{2\eta}\,\epsilon_j$ {*Langevin MCMC update*}
  end for
  $\xi_i \sim \mathcal{N}(0, I)$
  $a_{i-1} \leftarrow a_0^{(N_{\text{L}})} + \sigma(t_{i-1})\,\xi_i$ {*Re-noising with Gaussian approximation*}
end for
return a_0

B Ablation Studies and Supplementary Results

B.1 Pareto Fronts: Accuracy-Speed Trade-offs

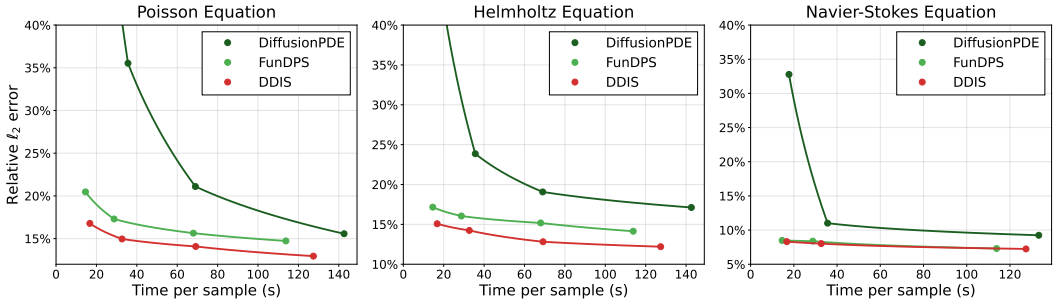

Figure 7: **DDIS consistently dominates the Pareto frontier across three PDEs.** Pareto fronts of relative ℓ_2 error versus wall-clock time (see Table 2 for settings). DDIS (red envelopes) achieves superior accuracy-speed trade-offs, demonstrating Pareto optimality over all joint-embedding baselines (green envelopes). Even on Navier-Stokes, where the advantage on ℓ_2 error over FunDPS is modest, DDIS achieves 60% lower spectral error, resulting in significantly higher quality.

B.2 POWER SPECTRUM COMPARISON

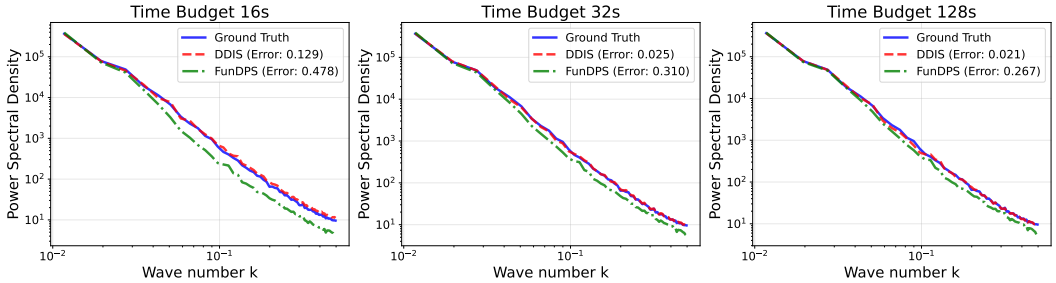

Figure 8: **Power spectrum comparison for the inverse Poisson problem.** A representative batch comparison between FunDPS and DDIS is shown. The predicted power spectral density is plotted against the ground truth as a function of wave number k.

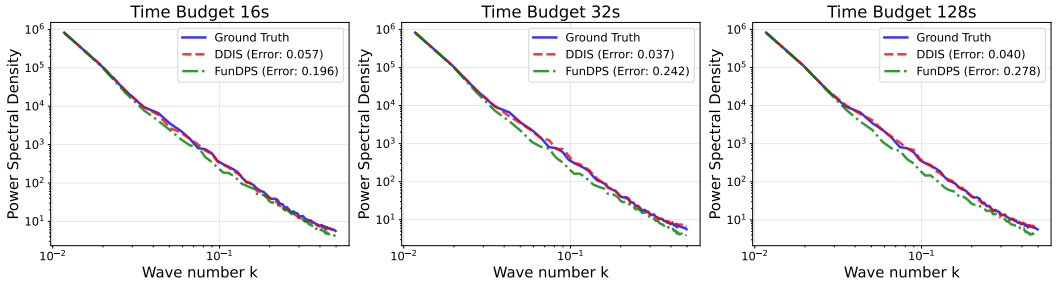

Figure 9: **Power spectrum comparison for the inverse Helmholtz problem.** DDIS accurately captures the fine-scale structures.

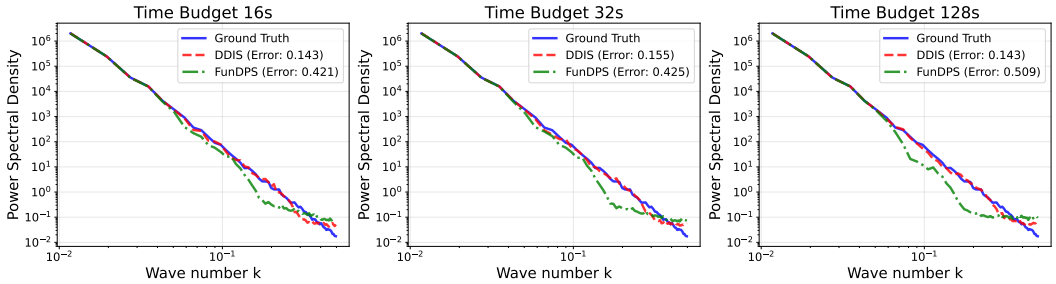

Figure 10: **Power spectrum comparison for the inverse Navier-Stokes problem.** This case is particularly challenging due to the high nonlinearity of the Navier-Stokes equation, reflected in FunDPS's gap at high wave numbers. Yet DDIS accurately captures them.

B.3 MIXED-RESOLUTION TRAINING

Table 4: DDIS performance under low- and mixed-resolution supervision. Relative ℓ_2 error (%) is reported. Mixed-resolution training uses full 64^2 data with an additional 10% of 128^2 data, maintaining accuracy while significantly reducing training costs.

Training Resolution	Poisson	Helmholtz	Steps (Diff/Lang)
128	17.19	16.50	500/20
64	18.28	16.47	
Mixed	17.75	16.39	
128	16.28	15.37	1000/20
64	17.27	15.67	
Mixed	16.76	15.58	
128	15.93	14.94	1000/50
64	17.64	15.10	
Mixed	16.61	14.85	

B.4 MIXED-RESOLUTION SAMPLING

Table 5: Mixed-resolution sampling results under 3% sparse observations. *DDIS (Full)* performs posterior sampling entirely at 128^2 resolution, whereas *DDIS (Mixed)* uses a coarse-to-fine pipeline (first half of the annealing at 64^2, then 128^2). Relative ℓ_2 error (%), spectral error E_s, and runtime per sample are reported. DDIS (Mixed) achieves comparable accuracy with 30% lower runtime.

Method	Poisson		Helmholtz		N-S		Time (s)
	ℓ_2	E_s	ℓ_2	E_s	ℓ_2	E_s	
DDIS (Full)	**16.36**	.066	15.19	.050	**8.22**	**.163**	25.92
DDIS (Mixed)	16.78	**.059**	**15.08**	**.044**	8.31	.185	16.75
DDIS (Full)	15.34	**.057**	**14.03**	**.053**	**8.00**	.174	51.83
DDIS (Mixed)	**14.96**	.080	14.24	.055	8.04	**.162**	32.67
DDIS (Full)	12.99	.098	12.28	**.079**	7.31	**.188**	207.83
DDIS (Mixed)	**12.95**	**.071**	**12.20**	.097	**7.24**	.200	127.42

C RELATED WORKS

Diffusion Models for Inverse Problems. Diffusion models have become a dominant paradigm for solving inverse problems, broadly categorized into two strategies. The first and more trivial strategy, *conditional diffusion*, trains models directly on the conditional distribution $p(x|y)$ or modifies the reverse process with task-specific architectures (Saharia et al., 2022; Whang et al., 2022). While effective for fixed observation patterns, this supervised approach lacks flexibility, requiring retraining for every new forward operator or measurement configuration. The alternative strategy, *unsupervised posterior sampling*, leverages a pre-trained unconditional prior $p(x)$ and enforces measurement consistency at inference time. For linear inverse problems, decomposition-based methods exploit the linear structure of the operator: DDRM (Kawar et al., 2022) utilizes Singular Value Decomposition (SVD) to efficiently perform diffusion in the spectral space, whereas DDNM (Wang et al., 2022) employs Range-Null Space Decomposition to avoid heavy computation in high-dimensional tasks (Daras et al., 2024). For general non-linear problems, guidance-based methods approximate the intractable likelihood score. Notably, DPS (Chung et al., 2022) employs Tweedie's formula to estimate the clean data mean, while LGD (Song et al., 2023) incorporates loss-based gradient guidance. Beyond simple guidance, optimization-centric approaches like DiffPIR (Zhu et al., 2023), PnP-DM (Wu et al., 2024), and DPnP (Xu & Chi, 2024) utilize variable splitting techniques to alternate between proximal data updates and denoising (Zheng et al., 2025). Recently, researchers have proposed asymptotically exact samplers using Sequential Monte Carlo, such as FPS (Dou & Song, 2024) and MCGDiff (Cardoso et al., 2023), as well as variational frameworks like RED-diff

(Mardani et al., 2023). Within this landscape, DAPS (Zhang et al., 2025) distinguishes itself by decoupling the denoising and likelihood updates via Langevin dynamics to mitigate approximation bias. However, as highlighted in (Zheng et al., 2025), even these advanced posterior sampling methods remain fragile: they violate the stability conditions of PDE solvers and yield spatially discontinuous updates when observations are sparse.

Inverse Problems with Arbitrary, Sparse Observations. Practical scientific inverse problems often rely on sparse, irregular measurements, emerging to be a new structural challenge. Inverse solvers in the previous paragraph actually treat inverse problems as natural image problems, and thus mostly fail under sparse observation. Conditional generative approaches, such as PIDM (Shu et al., 2023), CoCoGen (Jacobsen et al., 2025), PalSB (Li et al., 2025), and PRISMA (Sawhney et al., 2025), fail to handle this variability due to a combinatorial explosion of possible sensor layouts. As noted in (Daras et al., 2024), they lack the flexibility to generalize to varying masks without exhaustive retraining. Alternatively, direct posterior sampling strategies could theoretically handle arbitrary masks by masking the numerical PDE solver during likelihood evaluation. However, backpropagating gradients through iterative solvers is already known to suffer from computational instability and vanishing gradients even in dense settings (Zheng et al., 2025). This instability is amplified under sparse observations, where the gradient calculation becomes ill-conditioned (Plessix, 2006; Leung et al., 2021). While certain gradient-free replacement techniques (Amorós-Trepat et al., 2026) can satisfy sparse observations without backpropagation, they are difficult to generalize to complex physics-constrained settings. Furthermore, these methods focus on solution field reconstruction and are not designed to resolve the underlying coefficient space.

To avoid these solver-based issues, recent works like DiffusionPDE (Huang et al., 2024) and FunDPS (Yao et al., 2025) adopt joint-embedding methods. By training a diffusion model on the concatenated distribution of coefficients and solutions $p(a, u)$, these methods effectively recast the complex inverse problem as a data-driven inpainting task. While this bypasses unstable solvers, it relies on a questionable premise: modeling physics as a statistical correlation rather than a causal mechanism. Intuitively, we argue that enforcing strict physical laws into a joint distribution places a heavy representational burden on the model, forcing it to "rediscover" the forward map purely from data. This suggests joint-embedding methods may be inherently data-inefficient compared to approaches that explicitly distinguish the prior and the forward operator.

Flow Matching for PDE Inverse Problems. Beyond diffusion models, flow-based models have emerged as a promising paradigm for inverse problems due to the computational efficiency demonstrated in other domains like computer vision. Several recent works have explored flow-based models for image inverse problems. For example, D-Flow (Ben-Hamu et al., 2024) differentiates the cost function through continuous normalizing flows, while FlowChef (Patel et al., 2025) derives Jacobian-based guidance by treating the clean sample as an intermediate variable. FlowDPS (Kim et al., 2025) further leverages Tweedie's formula to decompose denoised and noisy estimates, applying gradients to appropriate components for posterior sampling. For scientific inverse problems, function-space modeling plays a central role. Functional Flow Matching (FFM) (Kerrigan et al., 2023) extends flow matching to infinite-dimensional settings, providing a principled formulation grounded in measure theory. More recently, ECI-sampling (Cheng et al., 2025) and PCFM (Utkarsh et al., 2025) propose training-free algorithms for enforcing physical constraints. From a more general perspective, Operator Flow Matching (OFM) (Shi et al., 2025) introduces a stochastic process learning framework for universal functional regression, enabling posterior sampling via SGLD-based updates. However, existing flow-based methods have largely focused on inverse problems defined on a single state variable, and have not yet been developed for joint coefficient-solution inverse problems. As a result, the structured dependencies between coefficients and solutions remain largely unexploited in current flow-based formulations.

Neural Operators as Physics Representation. Neural operators are architectures designed to learn mappings between infinite-dimensional function spaces rather than fixed vector spaces (Kovachki et al., 2023; Berner et al., 2025; Duruisseaux et al., 2025). Unlike standard CNNs or MLPs constrained to specific discretizations, neural operators are resolution-invariant, allowing models trained on coarse discretizations to work on finer grids (Li et al., 2020). Although architectures like DeepONet (Lu et al., 2019) offer flexibility for complex geometries, the Fourier Neural Operator (FNO) (Li et al., 2020) achieves state-of-the-art efficiency on uniform grids, particularly for massive-scale climate modeling

(Kurth et al., 2023) and chaotic fluid dynamics (Li et al., 2020). To address specific physical tasks, recent works have extended these backbones: PINO (Li et al., 2024) incorporates PDE constraints into the loss function to enable unsupervised or semi-supervised learning; Geo-FNO (Li et al., 2023) adapts spectral methods to arbitrary meshes via geometric deformations; and OFormer (Li et al., 2022) integrates Transformer attention to capture long-range dependencies. These architectures have become robust surrogates for traditional solvers, especially when repeated evaluations of complex physics are needed.

D Preliminaries

D.1 Diffusion Model

The minimal form of diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020a;b) learns to sample from a prior distribution. In the inverse problems we consider, the prior is over the coefficient $p(a)$. Formally, a diffusion model defines a forward stochastic process $\{a_t\}_{t=0}^{T}$ with the initial one $a_0 \coloneqq a$ and is governed by:

$$da_t = f(a_t, t)dt + g(t)dw, \tag{D.1}$$

where dw is a Wiener process and f, g are drift and diffusion coefficients. Let $p(a_t)$ be the marginal distribution of a_t. As $T \to \infty$, the distribution $p(a_T)$ converges to an unstructured noise distribution determined by (D.1), irrespective of $p(a)$. Then, by sampling a_T from noise distribution, e.g., Gaussian noise $\mathcal{N}(0, I)$, and reversing the process:

$$da_t = \Big[f(a_t, t) - g^2(t)\nabla_{a_t} \log p(a_t)\Big]dt + g(t)dw, \tag{D.2}$$

one ultimately obtains samples a_0 from $p(a)$. Here, the score term $\nabla_{a_t} \log p(a_t)$ is learned by a neural network $s_\theta(a_t, t)$ during forward process (D.1) via score-matching (Vincent, 2011). We define the diffusion prior as the distribution $p(a_0)$ obtained by evolving the reverse-time dynamics (D.2) from a noisy sample at $t = T$ down to $t = 0$.

D.2 Diffusion Posterior Sampling (DPS)

DPS modifies the reverse process (D.2) so that we obtain samples from the posterior $p(a \mid u_{\text{obs}})$:

$$da_t = \Big[f(a_t, t) - g^2(t)\big(\nabla_{a_t} \log p(a_t) + \nabla_{a_t} \log p(u_{\text{obs}} \mid a_t)\big)\Big]dt + g(t)dw_t. \tag{D.3}$$

Here, $\nabla_{a_t} \log p(a_t)$ is obtained via score-matching (Vincent, 2011) and $\nabla_{a_t} \log p(u_{\text{obs}} \mid a_t)$ is the log-likelihood gradient that guides the reverse process toward samples consistent with u_{obs}.

However, it is impossible to compute the likelihood by $p(u_{\text{obs}} \mid a_t) = \int p(u_{\text{obs}} \mid a_0)p(a_0 \mid a_t)da_0$ due to the intractability of $p(a_0 \mid a_t)$. Therefore, DPS takes the approximation:

$$p(u_{\text{obs}} \mid a_t) = \mathbb{E}_{a_0 \sim p(\cdot \mid a_t)}\left[p(u_{\text{obs}} \mid a_0)\right] \approx p\big(u_{\text{obs}} \mid a_0 = \widehat{a}_0(a_t)\big), \quad \widehat{a}_0(a_t) \coloneqq \mathbb{E}[a_0 \mid a_t]. \tag{D.4}$$

In practice, DPS implements the process (D.3) iteratively, effectively constructing a Markov chain $p(a_t \mid u_{\text{obs}}, a_{t+1})$. Each update is therefore incremental, limiting the ability to make large corrective moves toward the posterior. This motivates the Decoupled Annealing Posterior Sampling (DAPS) approach, which avoids the Markov chain formulation and enables direct posterior alignment.

Remark D.1 (Jensen-gap Approximation of DPS.)**.** The approximation (D.4) incurs a *Jensen's gap*:

$$\mathbb{E}[f(X)] \neq f(\mathbb{E}[X]),$$

where $f(a_0) \coloneqq \log p(u_{\text{obs}} \mid a_0)$. Thus, this approximation is exact only when $p(u_{\text{obs}} \mid a_0)$ is affine in a_0, and otherwise introduces bias that depends on the distribution of $a_0 \mid a_t$.

D.2.1 Extension: DecoupledDPS

As an ablation, we consider combining the DDIS insight with the standard DPS update in (D.3), yielding a *DecoupledDPS* variant. Unlike joint-embedding DPS (Huang et al., 2024; Yao et al., 2025), which defines the likelihood directly on the noisy latent a_t, this variant preserves the DDIS separation between the diffusion prior and the physics-induced likelihood.

Table 6: Ablation of decoupled DPS. Relative ℓ_2 error (%) for FunDPS, DecoupledDPS, and DDIS under 100% and 1% paired data. DecoupledDPS improves over joint embeddings but remains inferior to DDIS. All methods use $T = 16$s.

Method	Data	Poisson	Helmholtz	Navier–Stokes
FunDPS	100%	20.47	17.16	8.48
	1%	35.81	41.69	13.65
DecoupledDPS	100%	19.23	19.24	8.37
	1%	24.25	21.68	13.10
DDIS	100%	**16.36**	**15.19**	**8.22**
	1%	**18.70**	**16.40**	**12.05**

Concretely, the diffusion prior is defined over coefficients a as in Sec. 3.1, while the likelihood is specified on the clean variable through the neural operator surrogate:

$$p(u_{\text{obs}} \mid a_0) \;\propto\; \exp\left(-\frac{\|M \odot L_\phi(a_0) - u_{\text{obs}}\|_2^2}{2\beta_y^2}\right). \tag{D.5}$$

Following the DPS approximation in (D.4), the likelihood gradient in (D.3) is instantiated by substituting $a_0 \approx \widehat{a}_0(a_t)$:

$$\nabla_{a_t} \log p(u_{\text{obs}} \mid a_t) \;\approx\; \nabla_{a_t} \log p(u_{\text{obs}} \mid a_0 = \widehat{a}_0(a_t)), \tag{D.6}$$

where $\widehat{a}_0(a_t)$ is the denoised estimate induced by the diffusion prior.

Combining Eqs. (D.3), (D.5) and (D.6) yields a DPS-style reverse process whose guidance is mediated by the neural operator rather than a joint-embedding model. While this decoupled DPS variant consistently improves over FunDPS (Tab. 6), it still inherits the Jensen gap discussed in Rmk. D.1 and is generally worse than DDIS. This shows that decoupling already improves over joint embeddings, while greater gains arise from DAPS naturally enabled by this decoupled design. We therefore include decoupled DPS solely as an ablation to isolate this effect.

D.3 Decoupled Annealing Posterior Sampling (DAPS)

The key motivation behind DAPS is that the two components of posterior sampling—the prior update and the likelihood correction—naturally operate on different variables. The diffusion prior defines a score on the noisy latent variable a_t, whereas the likelihood $p(u_{\text{obs}} \mid a_0)$ is defined on the clean a_0. DPS forces both operations to act on the same noisy a_t and thus requires approximating the intractable $p(a_0 \mid a_t)$. DAPS avoids the conflict by decoupling the updates: it performs the likelihood correction on a clean estimate $\widehat{a}_0(a_t)$, and then re-noises it to obtain the updated a_{t-1}.

Let $p_t(a_t \mid u_{\text{obs}})$ denote the observation-conditioned time-marginal at noise level t. DAPS implements an annealing process with a noise schedule $\{\sigma_t\}$ such that, as t decreases, $p_t(a_t \mid u_{\text{obs}})$ anneals toward the target posterior $p(a_0 \mid u_{\text{obs}})$. At each annealing step, DAPS applies a two-phase sampling based on the current sample a_t: First, it performs Langevin MCMC to approximate sampling from

$$p(a_0 \mid a_t, u_{\text{obs}}) \;\propto\; p(a_0 \mid a_t)\, p(u_{\text{obs}} \mid a_0).$$

Initialized with the clean estimate $a_0^{(0)} = \widehat{a}_0(a_t)$ obtained from the DDPM/DDIM denoiser (Ho et al., 2020; Song et al., 2020a), the Langevin updates take the form:

$$\begin{aligned} a_0^{(j+1)} &= a_0^{(j)} + \eta\, \nabla_{a_0^{(j)}} \log p(a_0^{(j)} \mid a_t) \\ &\quad + \eta\, \nabla_{a_0^{(j)}} \log p(u_{\text{obs}} \mid a_0^{(j)}) + \sqrt{2\eta}\, \epsilon_j, \end{aligned} \tag{D.7}$$

where $\epsilon_j \sim \mathcal{N}(0, I)$, $\eta > 0$ is step size, and $p(a_0^{(j)} \mid a_t)$ is approximated by a Gaussian distribution with variance r_t^2:

$$p(a_0 \mid a_t) \approx \mathcal{N}(a_0; \widehat{a}_0(a_t), r_t^2 I). \tag{D.8}$$

Second, given $a_0^{(N)}$, DAPS samples the next latent a_{t-1} by re-noising $a_0^{(N)}$ under a Gaussian assumption[1]:

$$a_{t-1} \sim p(a_{t-1} \mid a_0^{(N)}) \approx \mathcal{N}\big(a_0^{(N)}, \sigma_{t-1}^2 I\big), \tag{D.9}$$

which yields a_{t-1} approximately distributed according to $p_{t-1}(a_{t-1} \mid u_{\text{obs}})$.

Despite its success in various inverse problems, as benchmarked in (Zhang et al., 2025), DAPS degrades in sparse observation settings, a flaw highlighted by (Yao et al., 2025):

Remark D.2 (Sparse-Guidance Failure of DAPS)**.** The degradation arises from localized gradient updates during the Langevin dynamics phase. When the sparse measurement x_{obs} is drawn from the same space as the prior variable x, i.e., $x_{\text{obs}} = M \odot x + \epsilon$, the likelihood gradient $\nabla_{x_0} \log p(x_{\text{obs}} \mid x_0)$ is non-zero only at observed locations. Consequently, the Langevin process creates a spatially discontinuous intermediate state that is out-of-distribution for the pretrained diffusion model. Sec. 4.2 offers a quantitative analysis for this degradation.

D.3.1 Asymptotic Guarantee for DAPS

Our analysis starts with a statistical interpretation of the DAPS that is not explicitly formulated in the original paper (Zhang et al., 2025). In particular, we show that the ideal DAPS update preserves the observation-conditioned time-marginals and converges asymptotically to the target posterior. In DDIS, the observation model acts in solution space while sampling occurs in coefficient space. The surrogate operator L_ϕ is therefore introduced to transport likelihood information across these spaces. When $L_\phi \to L$, DDIS ensures not only physically meaningful but also dense guidance for the two-step update.

Let $p(a_0 \mid u_{\text{obs}})$ denote the target posterior (Eq. (2.1)). DAPS introduces a decreasing noise schedule $\{\sigma_t\}$, and at noise level t the latent variable a_t follows an observation-conditioned marginal distribution $p(a_t \mid u_{\text{obs}})$. The ideal DAPS update from t to $t-1$ consists of two phases:

(i) **Posterior pullback:** sample $a_0 \sim p(a_0 \mid a_t, u_{\text{obs}})$ through Langevin dynamics (D.7).

(ii) **Re-noising:** sample $a_{t-1} \sim \mathcal{N}(a_0, \sigma_{t-1}^2 I)$, corresponding to (D.9).

Lemma D.1 (Time-Marginal Invariance)**.** If $a_t \sim p(a_t \mid u_{\text{obs}})$, the ideal DAPS update (i)-(ii) yields $a_{t-1} \sim p(a_{t-1} \mid u_{\text{obs}})$.

Proof of Lem. D.1. The DAPS update consists of:

Step 1: Sample $a_0 \sim p(a_0 \mid a_t, u_{\text{obs}})$, **Step 2:** Sample $a_{t-1} \sim p(a_{t-1} \mid a_0)$.

We compute the law of a_{t-1} by marginalizing over all randomness:

$$p(a_{t-1} \mid u_{\text{obs}}) = \iint p(a_{t-1} \mid a_0)\, p(a_0 \mid a_t, u_{\text{obs}})\, p(a_t \mid u_{\text{obs}})\, \mathrm{d}a_0\, \mathrm{d}a_t.$$

Using Bayes' rule for the posterior correction step,

$$p(a_0 \mid a_t, u_{\text{obs}}) = \frac{p(a_t \mid a_0)\, p(a_0 \mid u_{\text{obs}})}{p(a_t \mid u_{\text{obs}})},$$

we substitute and simplify:

$$\begin{aligned} p(a_{t-1} \mid u_{\text{obs}}) &= \iint p(a_{t-1} \mid a_0)\, \frac{p(a_t \mid a_0)\, p(a_0 \mid u_{\text{obs}})}{p(a_t \mid u_{\text{obs}})} p(a_t \mid u_{\text{obs}})\, \mathrm{d}a_0\, \mathrm{d}a_t \\ &= \iint p(a_{t-1} \mid a_0)\, p(a_t \mid a_0)\, p(a_0 \mid u_{\text{obs}})\, \mathrm{d}a_0\, \mathrm{d}a_t \\ &= \int p(a_{t-1} \mid a_0)\, p(a_0 \mid u_{\text{obs}})\, \mathrm{d}a_0, \end{aligned}$$

[1] As Zhang et al. (2025) noted, this Gaussian assumption is optional and adopted for practical efficiency. DAPS also allows exact posterior sampling via diffusion-score estimation at higher computational cost.

where the final equality follows from integrating out a_t. This is exactly the time-$(t-1)$ marginal of the observation-conditioned latent variable. Thus, the two-step update preserves the smoothed posterior marginals $a_{t-1} \sim p(a_{t-1} \mid u_{\text{obs}})$. □

The invariance property leads to an asymptotic guarantee:

Lemma D.2 (Asymptotic Guarantee)**.** As $t \to 0$, the distribution of a_t converges in distribution to the posterior $p(a_0 \mid u_{\text{obs}})$.

The DDIS framework generalizes DAPS by backpropagating the gradient $\nabla_{a_0^{(j)}} \log p(u_{\text{obs}} \mid a_0^{(j)})$ in (D.7) from the solution space to the latent coefficient space through a surrogate operator L_ϕ (Sec. 3). This preserves the statistical structure of the ideal DAPS update while resolving its sparse-guidance failure (Rmk. D.2). When the surrogate matches the true forward map ($L_\phi = L$), the DAPS transition in coefficient space is recovered, so the asymptotic guarantee continues to hold.

E Detailed Derivation of Guidance Attenuation in Joint-Embedding Models

Let the joint coefficient-solution variable be $x := (a, u) \in \mathcal{A} \times \mathcal{U}$. Joint-embedding models learn a probabilistic prior $p(x)$. Unlike DDIS, which specifies a conditional likelihood $p(u_{\text{obs}} \mid a)$ through a forward operator, joint-embedding methods assume a Gaussian observation model defined directly on the clean joint variable x_0:

$$u_{\text{obs}} = Mx_0 + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \widetilde{\sigma}_{\text{obs}}^2 I),$$

where M is a masking operator selecting the observed components of u and $\widetilde{\sigma}_{\text{obs}}^2$ denotes the prescribed observation noise level. This induces the likelihood

$$p_{\text{joint}}(u_{\text{obs}} \mid x_0) = \mathcal{N}\big(Mx_0, \widetilde{\sigma}_{\text{obs}}^2 I\big).$$

For posterior sampling, we consider DPS as adopted by our baselines FunDPS (Yao et al., 2025) and DiffusionPDE (Huang et al., 2024), where the likelihood is instantiated within the diffusion process $\{x_t\}_{t=0}^T$ over the joint variable $x_t = (a_t, u_t)$. Let $\widehat{x}_0(x_t, t)$ denote an estimate of the denoised sample $\mathbb{E}[x_0 \mid x_t]$. The guidance is given by the log-likelihood gradient with respect to x_t, approximating the likelihood via $\widehat{x}_0$ as in (D.4):

$$\begin{aligned} \nabla_{x_t} \log p_{\text{joint}}(u_{\text{obs}} \mid x_t) &\approx \nabla_{x_t} \log p_{\text{joint}}(u_{\text{obs}} \mid \widehat{x}_0(x_t, t)) \\ &= \frac{1}{\widetilde{\sigma}_{\text{obs}}^2} J_{\widehat{x}_0}(x_t, t)^\top M^\top r(x_t, t) \end{aligned} \tag{E.1}$$

where the Jacobian $J_{\widehat{x}_0}(x_t, t) := \nabla_{x_t} \widehat{x}_0(x_t, t)$ and the residual $r(x_t, t) := u_{\text{obs}} - M\widehat{x}_0(x_t, t)$.

Definition E.1 (Scale-free guidance)**.** We define the scale-free guidance as

$$g(x_t, t) := J_{\widehat{x}_0}(x_t, t)^\top M^\top r(x_t, t). \tag{E.2}$$

We consider score-based diffusion models with the following lemma:

Lemma E.1 (Tweedie-type estimator of denoised sample)**.** Let $\{x_t\}_{t=0}^T$ be a diffusion process with clean variable x_0. Given the intermediate state x_t and a score model $s_\theta(x_t, t) \approx \nabla_{x_t} \log p_t(x_t)$, the estimator of the denoised sample $\mathbb{E}[x_0 \mid x_t]$ is

$$\widehat{x}_0(x_t, t) = \frac{1}{\alpha_t}\Big(x_t + \sigma_t^2\, s_\theta(x_t, t)\Big), \tag{E.3}$$

where α_t and σ_t are time-dependent coefficients.

We decompose the denoised estimate and the learned score field into their a- and u-components:

$$\widehat{x}_0(x_t, t) = \begin{pmatrix} \widehat{a}_0(x_t, t) \\ \widehat{u}_0(x_t, t) \end{pmatrix}, \qquad s_\theta(x_t, t) = \begin{pmatrix} s_{\theta,a}(x_t, t) \\ s_{\theta,u}(x_t, t) \end{pmatrix},$$

then we can decompose the guidance into block-form:

Lemma E.2 (Block-form guidance)**.** Let $\widehat{x}_0(x_t, t)$ be defined by the Tweedie estimator (Lem. E.1). Then the likelihood guidance (Def. E.1) under DPS admits the block decomposition

$$g(x_t, t) \propto \begin{pmatrix} \sigma_t^2\, \partial_{a_t} s_{\theta,u}(x_t, t)^\top \\ I + \sigma_t^2\, \partial_{u_t} s_{\theta,u}(x_t, t)^\top \end{pmatrix} r(x_t, t), \tag{E.4}$$

where $r(x_t, t) = u_{\text{obs}} - M\widehat{x}_0(x_t, t)$.

Proof. The Jacobian of the denoised estimate admits the block decomposition

$$J_{\widehat{x}_0}(x_t, t) = \nabla_{x_t} \widehat{x}_0(x_t, t) = \begin{pmatrix} \partial_{a_t} \widehat{a}_0 & \partial_{u_t} \widehat{a}_0 \\ \partial_{a_t} \widehat{u}_0 & \partial_{u_t} \widehat{u}_0 \end{pmatrix}. \tag{E.5}$$

Since M selects the u-component, we have

$$M^\top r(x_t, t) = \begin{pmatrix} 0 \\ r(x_t, t) \end{pmatrix}. \tag{E.6}$$

Substituting (E.5) and (E.6) into the scale-free guidance (Def. E.1) yields

$$\begin{aligned} g(x_t, t) &\propto J_{\widehat{x}_0}(x_t, t)^\top M^\top r(x_t, t) = J_{\widehat{x}_0}(x_t, t)^\top \begin{pmatrix} 0 \\ r(x_t, t) \end{pmatrix} \\ &= \begin{pmatrix} (\partial_{a_t} \widehat{u}_0(x_t, t))^\top\, r(x_t, t) \\ (\partial_{u_t} \widehat{u}_0(x_t, t))^\top\, r(x_t, t) \end{pmatrix}. \end{aligned} \tag{E.7}$$

From Lem. E.1, the u-component satisfies

$$\widehat{u}_0(x_t, t) = \frac{1}{\alpha_t}\Big(u_t + \sigma_t^2\, s_{\theta,u}(x_t, t)\Big). \tag{E.8}$$

Differentiating (E.8) with respect to a_t gives

$$\partial_{a_t} \widehat{u}_0(x_t, t) = \frac{\sigma_t^2}{\alpha_t}\, \partial_{a_t} s_{\theta,u}(x_t, t), \tag{E.9}$$

and differentiating with respect to u_t gives

$$\partial_{u_t} \widehat{u}_0(x_t, t) = \frac{1}{\alpha_t}\Big(I + \sigma_t^2\, \partial_{u_t} s_{\theta,u}(x_t, t)\Big). \tag{E.10}$$

Substituting (E.9)-(E.10) into (E.7) yields (E.4). □

Lemma E.2 shows that the a-component of the guidance depends exclusively on the cross-partial $\partial_{a_t} s_{\theta,u}(x_t, t)$. This is the term that transmits information from the observation space into updates of a in the joint-embedding models. A sufficient condition for guidance attenuation is therefore

$$\partial_{a_t} s_{\theta,u}(x_t, t) \approx 0 \quad \Longrightarrow \quad g_a(x_t, t) \approx 0. \tag{E.11}$$

Equivalently, this characterization can be stated as the following requirement:

Proposition E.1 (Coupling is necessary for effective guidance)**.** Under Eq. (4.1), effective guidance in coefficient space requires the joint-embedding score model s_θ to learn a joint distribution that does not factorize across a_t and u_t, i.e., $p_t(a_t, u_t) \neq p_t(a_t)\, p_t(u_t)$.

Let $\{x_0^{(n)}\}_{n=1}^N$ denote paired training samples of the clean variable x_0. Below we connect the cross-partial to an empirical score learned on the finite samples.

Lemma E.3 (Empirical-score approximation, modified Theorem 3.2 of (Baptista et al., 2025))**.** Let $\{x_0^{(n)}\}_{n=1}^N$ be paired training samples of the clean joint variable $x_0 = (a, u)$. Define the empirical score $s_N(x_t, t) := \nabla_{x_t} \log p_t^{(N)}(x_t)$ where the empirical time-t density $p_t^{(N)}(x_t)$ is an isotropic Gaussian mixture:

$$p_t^{(N)}(x_t) := \frac{1}{N} \sum_{n=1}^{N} \mathcal{N}\big(x_t; \alpha_t x_0^{(n)}, \sigma_t^2 I\big).$$

We assume the trained joint score model $s_\theta(x_t, t)$ satisfies

$$\sup_{x_t} \big\| s_\theta(x_t, t) - s_N(x_t, t) \big\| \leq \varepsilon_N(t), \tag{E.12}$$

with $\varepsilon_N(t)$ small under data-scarcity.

From this point onward, we work with finite-dimensional representations of the coefficient and solution variables obtained via discretization. Accordingly, we treat $a \in \mathbb{R}^{d_a}$, $u \in \mathbb{R}^{d_u}$, and $x = (a, u) \in \mathbb{R}^d$ with $d = d_a + d_u$. The empirical score admits the closed form:

Lemma E.4 (Closed-form empirical score)**.** Let the closed-form isotropic Gaussian kernel be

$$\varphi(r) := (2\pi\sigma^2)^{-d/2} \exp\Big(-\frac{\|r\|_2^2}{2\sigma^2}\Big).$$

Define the mixture responsibilities

$$w_n(x,t) := \frac{\varphi\big(x - \alpha_t x_0^{(n)}\big)}{\sum_{j=1}^N \varphi\big(x - \alpha_t x_0^{(j)}\big)}, \qquad \sum_{n=1}^N w_n(x,t) = 1.$$

Under Lem. E.3, the empirical score admits the closed form

$$s_N(x,t) = \frac{1}{\sigma^2(t)} \sum_{n=1}^N w_n(x,t)\,\big(\alpha_t x_0^{(n)} - x\big). \tag{E.13}$$

Proof. Under Lem. E.3, the empirical time-t density is

$$p_t^{(N)}(x) = \frac{1}{N} \sum_{n=1}^N \varphi\big(x - \alpha_t x_0^{(n)}\big).$$

So the empirical score follows

$$\begin{aligned} s_N(x,t) = \nabla_x \log p_t^{(N)}(x) &= \frac{\sum_{n=1}^N \nabla_x \varphi\big(x - \alpha_t x_0^{(n)}\big)}{\sum_{j=1}^N \varphi\big(x - \alpha_t x_0^{(j)}\big)} \\ &= -\frac{1}{\sigma^2(t)} \sum_{n=1}^N \frac{\varphi\big(x - \alpha_t x_0^{(n)}\big)}{\sum_{j=1}^N \varphi\big(x - \alpha_t x_0^{(j)}\big)} \big(x - \alpha_t x_0^{(n)}\big) \\ &= \frac{1}{\sigma^2(t)} \sum_{n=1}^N w_n(x,t)\,\big(\alpha_t x_0^{(n)} - x\big) \end{aligned}$$

□

We now turn to the empirical score approximation (Lem. E.3, Lem. E.4) to analyze when the coupling condition in Proposition E.1 can be satisfied. Specifically, we characterize the structural source of a-u coupling in the empirical log-density and show that it arises solely through the mixture responsibilities. We formalize this in the following two lemmas. Recall that the empirical score is defined as $s_N(x,t) = \nabla_x \log p_t^{(N)}(x)$ with $p_t^{(N)}(x) = \sum_{n=1}^N \varphi_n(x)$ (Lem. E.3). Accordingly, the a–u cross-block of the empirical score is $\nabla_a \nabla_u \log \sum_{n=1}^N \varphi_n(x)$.

First, we show that this cross-partial depends only on responsibility gradients. Let $\varphi_n(x) := \varphi\big(x - \alpha_t x_0^{(n)}\big)$ and define the mixture responsibilities $w_n(x) := \varphi_n(x)/\sum_{j=1}^N \varphi_n(x)$ as in Lem. E.4.

Lemma E.5 (Cross-block of the empirical score)**.** The cross-block of the empirical score satisfies

$$\nabla_a \nabla_u \log \sum_{n=1}^N \varphi_n(x) = \sum_{n=1}^N \big(\nabla_a w_n(x)\big)\big(\nabla_u \log \varphi_n(x)\big). \tag{E.14}$$

We next provide an auxiliary lemma showing that an individual isotropic Gaussian component induces no a-u coupling.

Lemma E.6. A single isotropic Gaussian mixture component satisfies:

$$\nabla_a \nabla_u \log \varphi_n(x) = 0. \tag{E.15}$$

Proof. Let $(a_c, u_c) \coloneqq \alpha_t x_0^{(n)}$. The log-density of an isotropic Gaussian kernel is:

$$\log \varphi_n(x) = -\frac{1}{2\sigma^2}\|a - a_c\|_2^2 - \frac{1}{2\sigma^2}\|u - u_c\|_2^2 + \text{const},$$

which contains no coupling term of the form $\langle a - a_c,\, u - u_c\rangle$. Therefore,

$$\nabla_u \log \varphi_n(x) = -\frac{1}{\sigma^2}(u - u_c),$$

and differentiating w.r.t. a yields

$$\nabla_a \nabla_u \log \varphi_n(x) = 0.$$

□

Lem. E.5 and Lem. E.6 distinguish the source of a-u coupling. Specifically,

Remark E.1 (Source of a-u coupling in the empirical score)**.** Lem. E.6 shows that a single isotropic Gaussian component induces no a-u coupling. Since

$$\log\Big(\sum_{n=1}^{N} \varphi_n(x)\Big) \neq \sum_{n=1}^{N} \log \varphi_n(x),$$

any nonzero a-u coupling in the empirical log-density $\log p^{(N)}(x) = \log \sum_{n=1}^{N} \varphi_n(x)$ must arise exclusively from the $\log \sum$ operation, i.e., through the dependence of the responsibilities $w_n(x)$.

Using the closed-form empirical score (Lem. E.4), we now make explicit the cross-partial $\partial_a s_{N,u}(x,t)$ that appears in the block-form guidance (Lem. E.2). From Lem. E.4, the u-component of the empirical score is $s_{N,u}(x,t) = \frac{1}{\sigma^2(t)} \sum_{n=1}^{N} w_n(x,t)\,\big(\alpha_t u_0^{(n)} - u\big)$. Since $\partial_a(\alpha_t u_0^{(n)} - u) = 0$, the cross-partial $\partial_a s_{N,u}(x,t)$ becomes

$$\partial_a s_{N,u}(x,t) = \frac{\alpha_t}{\sigma^2(t)} \sum_{n=1}^{N} \big(\partial_a w_n(x,t)\big)\, u_0^{(n)}. \tag{E.16}$$

Eq. (E.16) is an explicit elaboration of (E.14) in Lem. E.5.

Based on Rmk. E.1, we therefore analyze when these gradients vanish or remain nontrivial via responsibility gradients. Recall the responsibilities

$$w_n(x,t) = \frac{\varphi_n(x)}{\sum_{j=1}^{N} \varphi_n(x)} = \frac{\varphi\big(x - \alpha_t x_0^{(n)}\big)}{\sum_{j=1}^{N} \varphi\big(x - \alpha_t x_0^{(j)}\big)}, \qquad \sum_{n=1}^{N} w_n(x,t) = 1. \tag{E.17}$$

Let $Z(x,t) \coloneqq \sum_{j=1}^{N} \varphi_j(x,t)$. Then $w_n(x,t) = \varphi_n(x,t)/Z(x,t)$ with the derivative

$$\begin{aligned}
\partial_a w_n(x,t) &= \frac{(\partial_a \varphi_n) Z - \varphi_n(\partial_a Z)}{Z^2} \\
&= \frac{\varphi_n}{Z}\Big(\frac{\partial_a \varphi_n}{\varphi_n} - \sum_{j=1}^{N} \frac{\varphi_j}{Z}\frac{\partial_a \varphi_j}{\varphi_j}\Big) && \big(\text{By } \partial_a Z = \textstyle\sum \partial_a \varphi_j\big) \\
&= w_n(x,t)\Big(\partial_a \log \varphi_n(x,t) - \sum_{j=1}^{N} w_j(x,t)\, \partial_a \log \varphi_j(x,t)\Big). && \text{(E.18)}
\end{aligned}$$

The following two theorems characterize this behavior through the geometric position of x relative to the training samples $\{x_0^{(n)}\}_{n=1}^{N}$. They separate two cases: a *local* regime, where x is dominated by a single mixture component, and an *overlap* regime, where at least two components have comparable responsibilities.

Case 1 (local regime).

Theorem E.1 (Local dominance implies vanishing responsibility gradients)**.** Fix (x, t) and define responsibilities $w_n(x, t)$ by (E.17). Assume there exists an index $k \in [N]$ such that

$$\sum_{j \neq k} \varphi\big(x - \alpha_t x_0^{(j)}\big) \leq \eta\, \varphi\big(x - \alpha_t x_0^{(k)}\big) \tag{E.19}$$

for some $\eta \in (0, 1)$. Then the responsibility gradients satisfy

$$\big\|\partial_a w_k(x, t)\big\| = O(\eta), \qquad \sum_{j \neq k} \big\|\partial_a w_j(x, t)\big\| = O(\eta), \tag{E.20}$$

which vanishes as $\eta \to 0$.

Proof. Under (E.19),

$$Z(x, t) = \varphi_k(x, t) + \sum_{j \neq k} \varphi_j(x, t) \leq (1 + \eta)\varphi_k(x, t),$$

hence

$$w_k(x, t) = \frac{\varphi_k}{Z} \geq \frac{1}{1 + \eta}, \qquad \sum_{j \neq k} w_j(x, t) = 1 - w_k(x, t) \leq \frac{\eta}{1 + \eta}. \tag{E.21}$$

Define the local bound

$$G(x, t) := \max_{1 \leq j \leq N} \big\|\partial_a \log \varphi_j(x, t)\big\|. \tag{E.22}$$

Then $\partial_a w_n$ (E.18) is bounded by the following two cases:

Case A: $n = k$.

$$\begin{aligned}
\|\partial_a w_k(x, t)\| &= w_k \|\partial_a \log \varphi_k - \sum_j w_j \partial_a \log \varphi_j\| \\
&= w_k \|\sum_{j \neq k} w_j \big(\partial_a \log \varphi_k - \partial_a \log \varphi_j\big)\| \\
&\leq w_k \sum_{j \neq k} w_j \|\big(\partial_a \log \varphi_k - \partial_a \log \varphi_j\big)\| \\
&\leq 2G(x, t)\, w_k \sum_{j \neq k} w_j && \text{(By (E.22))} \\
&\leq 2G(x, t) \frac{\eta}{1 + \eta}. && \text{(By (E.21))}
\end{aligned}$$

Case B: $n \neq k$.

$$\begin{aligned}
\|\partial_a w_j(x, t)\| &= w_j \|\partial_a \log \varphi_j - \sum_m w_m \partial_a \log \varphi_m\| \\
&\leq w_j \Big(\|\partial_a \log \varphi_j\| + \sum_m w_m \|\partial_a \log \varphi_m\|\Big) \\
&\leq 2G(x, t) w_j, \qquad j \neq k. && \text{(By (E.22))}
\end{aligned}$$

Combining bounds from *Case A* and *Case B*,

$$\begin{aligned}
\sum_{n=1}^{N} \|\partial_a w_n\| &= \|\partial_a w_k\| + \sum_{j \neq k} \|\partial_a w_j\| \\
&\leq 2G \frac{\eta}{1 + \eta} + 2G \sum_{j \neq k} w_j \\
&\leq 4G \frac{\eta}{1 + \eta}. && \text{(By (E.21))} \\
&= O(\eta)
\end{aligned}$$

□

Thm. E.1 leads to a sufficient local-dominance condition for guidance attenuation during sampling process.

Corollary E.1.1 (Local dominance implies guidance attenuation)**.** Define $\varphi_n(x,t) := \varphi\big(x - \alpha_t x_0^{(n)}\big)$. If the diffusion state x_t lies in a neighborhood of a single mixture center $\alpha_t x_0^{(k)}$ in the sense that

$$\varphi_k(x_t,t) \gg \varphi_j(x_t,t) \qquad \forall j \neq k, \tag{E.23}$$

the a-component of the scale-free guidance (Def. E.1) attenuates:

$$g_a(x_t,t) \approx 0.$$

Proof. Let η satisfying

$$\sum_{j\neq k} \varphi\big(x - \alpha_t x_0^{(j)}\big) \leq \eta\, \varphi\big(x - \alpha_t x_0^{(k)}\big),$$

then the η must be small due to (E.23). Define the local bound

$$G(x,t) := \max_{1\leq j\leq N} \big\|\partial_a \log \varphi_j(x,t)\big\|.$$

Combining with (E.16) yields

$$\begin{aligned}\big\|\partial_a s_{N,u}(x,t)\big\| &\leq \frac{\alpha_t}{\sigma^2(t)}\Big(\max_{n\in[N]} \|u_0^{(n)}\|\Big) \sum_{n=1}^{N} \|\partial_a w_n(x,t)\| \\ &\leq \frac{4\alpha_t}{\sigma^2(t)}\Big(\max_{n\in[N]} \|u_0^{(n)}\|\Big) G(x,t) \frac{\eta}{1+\eta}.\end{aligned}$$

With small η, the cross-partial of the empirical score vanishes

$$\partial_a s_{N,u}(x,t) \approx 0.$$

From Lem. E.3, the trained score model s_θ approximates the empirical score s_N. Thus, from (E.4), the a-component of the guidance is attenuated

$$g_a(x,t) \propto \sigma_t^2\, \partial_{a_t} s_{\theta,u}(x,t)^\top\, r(x,t) \approx 0.$$

□

The first case (Thm. E.1 and Corollary E.1.1) show that, under local regime, responsibility gradients sufficiently vanish. Next, we show the complementary case: nontrivial responsibility gradients occur only when x lies in an overlap regime.

Case 2 (overlap regime).

Theorem E.2 (Nontrivial responsibility gradients require overlap)**.** Fix (x,t) and define responsibilities $w_n(x,t)$ by (E.17). Let the local bound $G(x,t) := \max_{1\leq j\leq N} \big\|\partial_a \log \varphi_j(x,t)\big\|$ and let $p \in \arg\max_{j\in[N]} w_j(x,t)$ denote an index attaining the maximal responsibility. If there exists an index $n \in [N]$ such that

$$\big\|\partial_a w_n(x,t)\big\| \geq \delta \qquad \text{for some } \delta > 0, \tag{E.24}$$

then there exist $q \neq p$ such that

$$\Big|\big\|x - \alpha_t x_0^{(p)}\big\|_2^2 - \big\|x - \alpha_t x_0^{(q)}\big\|_2^2\Big| \leq 2\sigma^2(t) \log\Big(\frac{1-\tau(\delta)}{\tau(\delta)}\Big), \qquad \tau(\delta) := \frac{\delta}{2G(x,t)(N-1)}, \tag{E.25}$$

so at least two components have comparable mass at (x,t).

Proof. By the identity (E.18),

$$\partial_a w_n = w_n\Big(\partial_a \log \varphi_n - \sum_{j=1}^{N} w_j\, \partial_a \log \varphi_j\Big).$$

Using the definition of local bound and $\sum_j w_j = 1$,

$$\|\partial_a w_n\| \le w_n\Big(\|\partial_a \log \varphi_n\| + \sum_{j=1}^{N} w_j \|\partial_a \log \varphi_j\|\Big) \le 2G\, w_n.$$

Thus (E.24) implies

$$w_n \ge \frac{\delta}{2G}.$$

Since $p \in \arg\max_{j\in[N]} w_j(x,t)$, $w_p(x,t) \ge w_n(x,t) \ge \delta/(2G(x,t))$. Moreover,

$$\sum_{j\neq p} w_j(x,t) = 1 - w_p(x,t) \ge 1 - \frac{\delta}{2G(x,t)}.$$

By the pigeonhole principle, there exists $q \neq p$ such that

$$w_q(x,t) \ge \frac{1 - w_p(x,t)}{N-1} \ge \frac{1}{N-1}\Big(1 - \frac{\delta}{2G(x,t)}\Big).$$

This already yields two non-negligible responsibilities whenever w_p is bounded away from 1. To obtain the explicit $\tau(\delta)$ bound, use instead the standard inequality

$$\big\|\partial_a w_p(x,t)\big\| \le 2G(x,t) \sum_{j\neq p} w_j(x,t) = 2G(x,t)\big(1 - w_p(x,t)\big), \tag{E.26}$$

which follows from the Case A derivation (rewrite the bracket as a sum over $j \neq p$ and use G). If $\|\partial_a w_p(x,t)\| \ge \delta$, then (E.26) implies $1 - w_p(x,t) \ge \delta/(2G(x,t))$, hence there exists $q \neq p$ with

$$w_q(x,t) \ge \frac{1 - w_p(x,t)}{N-1} \ge \frac{\delta}{2G(x,t)(N-1)} = \tau(\delta).$$

Together with $w_p(x,t) \ge 1 - w_p(x,t) \ge \tau(\delta)$, we obtain:

$$w_p(x,t) \ge \tau(\delta), \qquad w_q(x,t) \ge \tau(\delta).$$

For isotropic Gaussian kernels defined in Lem. E.4,

$$\varphi_n(x,t) = (2\pi\sigma^2(t))^{-d/2} \exp\Big(-\frac{|x - \alpha_t x_0^{(n)}|_2^2}{2\sigma^2(t)}\Big).$$

Hence, for any two indices $p \neq q$

$$\log \frac{\varphi_p(x,t)}{\varphi_q(x,t)} = -\frac{1}{2\sigma^2(t)}\Big(|x - \alpha_t x_0^{(p)}|_2^2 - |x - \alpha_t x_0^{(q)}|_2^2\Big). \tag{E.27}$$

If $w_p, w_q \ge \tau$ and $\sum_j w_j = 1$, then

$$\frac{\tau}{1-\tau} \le \frac{w_p(x,t)}{w_q(x,t)} \le \frac{1-\tau}{\tau}. \tag{E.28}$$

Combining (E.27) and (E.28) yields

$$\begin{aligned}
\Big||x - \alpha_t x_0^{(p)}|_2^2 - |x - \alpha_t x_0^{(q)}|_2^2\Big| &= 2\sigma^2(t) \log \frac{\varphi_p(x,t)}{\varphi_q(x,t)} && \text{(By (E.27))}\\
&= 2\sigma^2(t) \log \frac{w_p(x,t)}{w_q(x,t)} && \text{(Since } w_n = \varphi_n / \textstyle\sum_j \varphi_j \text{)}\\
&\le 2\sigma^2(t) \log\Big(\frac{1-\tau}{\tau}\Big), && \text{(By (E.28))}
\end{aligned}$$

which is exactly (E.25). □

Thm. E.2 leads to a necessary overlapping condition for non-vanishing guidance during sampling process.

Corollary E.2.1 (Non-vanishing guidance requires overlap)**.** For the a-component of the scale-free guidance (Def. E.1) to be non-zero by

$$\|g_a(x_t,t)\| > 0,$$

the state x_t must lie in an overlap region of at least two mixture components, in the sense that there exist $p \neq q$ satisfying

$$\left| \|x_t - \alpha_t x_0^{(p)}\|_2^2 - \|x_t - \alpha_t x_0^{(q)}\|_2^2 \right| \lesssim \sigma^2(t).$$

Proof. From Lem. E.2 and Eq. (E.16), non-vanishing guidance implies $\|\partial_a w_n(x,t)\| > 0$ for some n. The claim then follows directly from Thm. E.1. □

A contrapositive statement of Corollary E.2.1 is:

Remark E.2 (Non-overlap implies guidance attenuation)**.** If the diffusion state x_t does not lie in an overlap region of the mixture, then the a-component of the scale-free guidance vanishes by $g_a(x_t, t) \approx 0$.

Corollary E.1.1, Corollary E.2.1, and Rmk. E.2 together yield a geometric interpretation of guidance behavior in joint-embedding diffusion models: (i) far from all mixture centers, guidance attenuates; (ii) even when close to a single mixture center, guidance still attenuates; and (iii) nonzero guidance is possible only when x lies near at least two mixture centers simultaneously. Under data scarcity, the overlap between mixture components may be absent altogether, causing coefficient-space guidance to collapse throughout the sampling process.

F Sample-Complexity Analysis

We analyze the sample complexity of DDIS compared to joint-embedding diffusion models. Let $\mathcal{A}$ and $\mathcal{U}$ denote the coefficient and solution function spaces, and $z \sim p_0(z)$ be base noise. We define three hypothesis classes. The diffusion prior class is the set of generators mapping noise to coefficient fields, $\mathcal{P} := \{G_\theta : \mathcal{Z} \to \mathcal{A}\}_{\theta \in \Theta_P}$ with $a = G_\theta(z)$. The operator (likelihood) class is the set of neural operators approximating the PDE forward map, $\mathcal{L} := \{L_\phi : \mathcal{A} \to \mathcal{U}\}_{\phi \in \Theta_L}$ with $u = L_\phi(a)$. The joint-embedding class is the set of diffusion generators mapping noise directly to coefficient-solution pairs, $\mathcal{J} := \{H_\psi : \mathcal{Z} \to \mathcal{A} \times \mathcal{U}\}_{\psi \in \Theta_J}$ with $(a, u) = H_\psi(z)$. Below, we derive the hypothesis-class inclusion relationships between $\mathcal{P}$, $\mathcal{L}$, and $\mathcal{J}$.

F.1 Hypothesis Class of Joint and Decoupled Models

A joint-embedding generative model aims to represent the joint distribution $p(a, u)$. Since $p(a, u) = p(a)\, p(u \mid a)$, modeling the joint is sufficient to model both the prior $p(a)$ and the likelihood $p(u \mid a)$. Thus, for every prior generator in $\mathcal{P}$ and every operator surrogate in $\mathcal{L}$, there must exist a corresponding joint model in $\mathcal{J}$ whose marginals and conditionals match them. This induces the hypothesis-class inclusion relationships between $\mathcal{P}$, $\mathcal{L}$, and $\mathcal{J}$, elaborated in the following lemmas:

Lemma F.1 (Joint models subsume prior generators)**.**

$$\mathcal{P} \subseteq \mathcal{J}.$$

Proof. Any prior generator $G_\theta \in \mathcal{P}$ maps noise z to coefficients $a \in \mathcal{A}$. A joint model can always reproduce this behavior by ignoring the solution component. For any fixed $u_0 \in \mathcal{U}$, define

$$H_\psi(z) = (G_\theta(z),\, u_0).$$

Then $H_\psi \in \mathcal{J}$, and its marginal matches G_θ, i.e., every prior generator is realizable within $\mathcal{P}$. □

Lemma F.2 (Joint models subsume operator surrogates)**.**

$$\mathcal{L} \subseteq \mathcal{J}.$$

Proof. Any operator surrogate $L_\phi \in \mathcal{L}$ maps coefficients a to solutions u. Since every joint model $H_\psi \in \mathcal{J}$ generates pairs (a, u) together, it must be expressive enough to reproduce the mappings. Choose any base coefficient generator G_{θ_0} and define

$$H_\psi(z) = \big(G_{\theta_0}(z),\, L_\phi(G_{\theta_0}(z))\big).$$

Then $H_\psi \in \mathcal{J}$ with output $u = L_\phi(a)$, i.e., every operator surrogate is realizable within $\mathcal{J}$. □

Lemma F.3 (Joint class inclusion and complexity)**.** Let d_P, d_L, and d_J denote the Rademacher complexities of the prior, operator, and joint hypothesis classes, respectively. Then

$$d_J \geq \max(d_P, d_L). \tag{F.1}$$

Proof. By Lemma F.1 we have $\mathcal{P} \subseteq \mathcal{J}$, and by Lemma F.2 we have $\mathcal{L} \subseteq \mathcal{J}$. Hence $\mathcal{P} \cup \mathcal{L} \subseteq \mathcal{J}$. Rademacher complexity is monotone under inclusion, so

$$d_J = \mathfrak{R}(\mathcal{J}) \;\geq\; \mathfrak{R}(\mathcal{P} \cup \mathcal{L}) \;\geq\; \max\big(\mathfrak{R}(\mathcal{P}), \mathfrak{R}(\mathcal{L})\big) = \max(d_P, d_L).$$

□

F.2 Generalization Bound

We adopt Rademacher complexities to derive generalization bounds of the hypothesis class:

Lemma F.4 (Rademacher generalization bound, Ch. 26 of (Shalev-Shwartz & Ben-David, 2014))**.** Let $\mathcal{H}$ be a hypothesis class with Rademacher complexity d_H, and let $\widehat{h}$ denote an empirical risk minimizer trained on n i.i.d. samples. Then its expected excess risk satisfies

$$\mathbb{E}[\ell(\widehat{h})] - \min_{h \in \mathcal{H}} \mathbb{E}[\ell(h)] = \widetilde{\mathcal{O}}\left(\sqrt{\frac{d_H}{n}}\right). \tag{F.2}$$

Then, based on Lem. F.4, we evaluate the excess risk of DDIS and joint-embedding methods.

Proposition F.1 (Sample complexity of DDIS)**.** Given n_u unpaired samples and n_p paired samples, the DDIS estimation error satisfies

$$\mathrm{err}_{\mathrm{DDIS}} = \widetilde{\mathcal{O}}\left(\sqrt{\frac{d_L}{n_p}} + \sqrt{\frac{d_P}{n_p + n_u}}\right) \xrightarrow[n_u \to \infty]{} \widetilde{\mathcal{O}}\left(\sqrt{\frac{d_L}{n_p}}\right). \tag{F.3}$$

Proof. Apply Lem. F.4 to $\mathcal{P}$ using $n_p + n_u$ samples, and to $\mathcal{L}$ using n_p paired samples, then combine both. □

Proposition F.2 (Sample complexity of joint-embedding methods)**.** Joint-embedding models trained on n_p paired samples satisfy

$$\mathrm{err}_{\mathrm{joint}} = \widetilde{\mathcal{O}}\left(\sqrt{\frac{d_J}{n_p}}\right). \tag{F.4}$$

Remark F.1 (Comparison of sample-complexity guarantees)**.** Combining Proposition F.1 and Proposition F.2 with Lem. F.3, in the imbalanced regime $n_u \gg n_p$ the DDIS bound scales as $\widetilde{\mathcal{O}}(\sqrt{d_L/n_p})$, whereas the joint-embedding bound scales as $\widetilde{\mathcal{O}}(\sqrt{d_J/n_p})$ with $d_J \geq \max(d_P, d_L)$. Thus DDIS yields a strictly more favorable upper bound than joint-embedding models with scarce paired data.

G Proofs of the Failure Modes of Joint Embeddings + DAPS

In this appendix, we provide formal proofs for the covariance-collapse results used in Sec. 4.2. We work in a separable Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$. Let $\delta_{x_i} : \mathcal{H} \to \mathbb{R}$ denote the point evaluation functional at location $x_i \in \Omega$, satisfying $\langle \delta_{x_i}, f \rangle = f(x_i)$ for $f \in \mathcal{H}$.

Starting from the potential U in Eq. (4.8), the functional gradient is

$$\nabla U(x) = C^{-1}(x - x_0) + \frac{1}{\sigma_s^2} \delta_{x_i}(\langle \delta_{x_i}, x \rangle - c). \tag{G.1}$$

Substituting Eq. (G.1) into (4.7) and collecting affine terms yields the (preconditioned) linear SDE

$$dx_t = -\Sigma \left(C^{-1} + \frac{1}{\sigma_s^2} \delta_{x_i} \otimes \delta_{x_i} \right) x_t \, dt + b \, dt + \sqrt{2} \Sigma^{1/2} dW_t, \tag{G.2}$$

for some drift element $b \in \mathcal{H}$ that does not affect the stationary covariance.

Lemma G.1 (Lyapunov Equation for Stationary Covariance)**.** Let $B := C^{-1} + \frac{1}{\sigma_s^2}\delta_{x_i} \otimes \delta_{x_i}$. The stationary covariance operator Σ_∞ of the linear SDE (G.2) satisfies the Lyapunov equation:

$$\Sigma B\, \Sigma_\infty + \Sigma_\infty\, (\Sigma B)^* = 2\Sigma. \tag{G.3}$$

If C and Σ are self-adjoint (hence B is self-adjoint), then the stationary covariance is

$$\Sigma_\infty = B^{-1}. \tag{G.4}$$

Proof. This is a standard fact for preconditioned Langevin dynamics targeting $\pi(x) \propto \exp(-U(x))$. (See Pavliotis (2014), Proposition 3.10.) For completeness, we verify that (G.4) solves (G.3). Since B and Σ are self-adjoint, $(\Sigma B)^* = B\Sigma$. Substituting $\Sigma_\infty = B^{-1}$ into (G.3) gives $\Sigma B B^{-1} + B^{-1} B \Sigma = \Sigma + \Sigma = 2\Sigma$, as required. □

Lemma G.2 (Sherman-Morrison Inversion)**.** Let $B = C^{-1} + \frac{1}{\sigma_s^2}\delta_{x_i} \otimes \delta_{x_i}$ where C is a positive definite covariance operator. Then:

$$B^{-1} = C - \frac{(C\delta_{x_i}) \otimes (C\delta_{x_i})}{\sigma_s^2 + C(x_i, x_i)} \tag{G.5}$$

where $C(x_i, x_i) = \langle \delta_{x_i}, C\delta_{x_i} \rangle$ is the pointwise variance at x_i.

Proof. We apply the Sherman-Morrison formula (Sherman & Morrison, 1950) with base operator C^{-1} and rank-1 perturbation:

$$(A + u \otimes v)^{-1} = A^{-1} - \frac{(A^{-1}u) \otimes (A^{-1}v)}{1 + \langle v, A^{-1}u \rangle} \tag{G.6}$$

After substituting $A = C^{-1}$ and $u \otimes v = \frac{1}{\sigma_s^2}\delta_{x_i} \otimes \delta_{x_i}$, we obtain

$$B^{-1} = C - \frac{(C\delta_{x_i}) \otimes (C\delta_{x_i})}{\sigma_s^2 + \langle \delta_{x_i}, C\delta_{x_i} \rangle} = C - \frac{(C\delta_{x_i}) \otimes (C\delta_{x_i})}{\sigma_s^2 + C(x_i, x_i)}. \tag{G.7}$$

□

Theorem G.1 (Sparse constraint induces correlation shrinkage)**.** Consider the preconditioned Langevin SDE (G.2):

$$dx_t = -\Sigma \left(C^{-1} + \frac{1}{\sigma_s^2}\delta_{x_i} \otimes \delta_{x_i} \right) x_t\, dt + b\, dt + \sqrt{2}\Sigma^{1/2} dW_t,$$

where $C \succ 0$ is the prior covariance operator, $\Sigma \succ 0$ is the noise covariance operator, $\sigma_s^2 > 0$ is the constraint strength, and $b \in \mathcal{H}$ is a constant drift element. Let Σ_∞ denote the stationary covariance operator of this process. Then for any location $x_k \in \Omega$, the covariance function values involving the constrained location x_i satisfy

$$\Sigma_\infty(x_i, x_k) = \Sigma_\infty(x_k, x_i) = \frac{\sigma_s^2}{\sigma_s^2 + C(x_i, x_i)}\, C(x_i, x_k). \tag{G.8}$$

Proof. From Lemma G.1, we have $\Sigma_\infty = B^{-1}$. Using Lemma G.2,

$$\Sigma_\infty = C - \frac{(C\delta_{x_i}) \otimes (C\delta_{x_i})}{\sigma_s^2 + C(x_i, x_i)}. \tag{G.9}$$

Evaluating the covariance function at (x_i, x_k) gives

$$\Sigma_\infty(x_i, x_k) = C(x_i, x_k) - \frac{C(x_i, x_i)\, C(x_i, x_k)}{\sigma_s^2 + C(x_i, x_i)} \tag{G.10}$$

$$= C(x_i, x_k)\, \frac{\sigma_s^2}{\sigma_s^2 + C(x_i, x_i)}. \tag{G.11}$$

By symmetry of Σ_∞, we have $\Sigma_\infty(x_k, x_i) = \Sigma_\infty(x_i, x_k)$ for all x_k. □

H Experimental Setup

H.1 Dataset

Inverse Poisson. We also consider the Poisson equation,

$$\nabla^2 u(x) = a(x), \quad x \in (0,1)^2,$$

under homogeneous Dirichlet conditions $u|_{\partial\Omega} = 0$. Here $a(x)$ is generated by Gaussian random fields $a \sim \mathcal{N}(0, (-\Delta + 9I)^{-2})$. Given sparse samples of $u(x)$, the task is to recover the coefficient field $a(x)$.

Inverse Helmholtz. We study the two-dimensional Helmholtz equation,

$$\nabla^2 u(x) + k^2 u(x) = a(x), \quad x \in (0,1)^2,$$

with $k = 1$ and homogeneous Dirichlet boundary conditions $u|_{\partial\Omega} = 0$. Coefficient fields $a(x)$ are sampled from Gaussian random fields $a \sim \mathcal{N}(0, (-\Delta + 9I)^{-2})$. This setting is particularly challenging due to the oscillatory nature of the Helmholtz solution and the multiscale structure of the GRF prior.

Inverse Navier-Stokes. We further study the two-dimensional incompressible Navier-Stokes equations in vorticity form. Let $u(x,t)$ denote the velocity field and $w(x,t) = \nabla \times u(x,t)$ the corresponding vorticity. The system evolves according to

$$\partial_t w(x,t) + u(x,t) \cdot \nabla w(x,t) = \nu \Delta w(x,t) + f(x), \quad x \in (0,1)^2,\ t \in (0,T],$$
$$\nabla \cdot u(x,t) = 0, \quad x \in (0,1)^2,\ t \in [0,T]$$

with viscosity $\nu = 10^{-3}$ and periodic boundary conditions. The system is initialized by the vorticity field $w(x,0) = a(x)$, where $a(x)$ is sampled from a Gaussian random field $a \sim \mathcal{N}(0,\ 7^{3/2}(-\Delta + 49I)^{-5/2})$ and treated as the unknown coefficient. Given sparse observations of the solution $w(x,T)$ at the terminal time, the task is to recover the initial vorticity $w(x,0)$.

It is worth noting that, due to the loss of information, it is impossible to calculate PDE residual for physics-informed guidance; moreover, the property $\nabla \cdot (\nabla \times u) = 0$ mentioned in Huang et al. (2024) is invalid.

H.2 Details on Decoupled Diffusion Inverse Solver

We summarize the key training hyperparameters in Tab. 7 and the sampling hyperparameters in Tabs. 8 and 9. All experiments are conducted on a single NVIDIA RTX 4090 GPU. Additional details are provided below.

FNO padding and cropping. FNO layers rely on FFT-based spectral convolution and therefore assume periodic boundary conditions. To mitigate boundary artifacts, we pad the input field by p grid cells along each spatial dimension, apply the operator L_ϕ on the padded grid, and crop the output back to the original domain.

Physics-informed training. In physics-informed training, we add a physics-regularization term to the loss function to help the model generalize better by enforcing the PDE residual to be small. In all cases, the neural operator is first trained on (limited) paired data supervision, and then trained with both (limited) data and physics-regularization terms. The physics-regularization term is weighted by $\lambda_{PDE} = 0.1, \lambda_{BC} = 10$. StepLR scheduler is used with a decay of 0.1.

H.3 Details on Flow-based Models

H.3.1 Prior Learning and Hyperparameters

For flow-based baselines, existing work primarily focuses on inverse problems involving only the solution channel u, leveraging a flow-based prior $p(u)$ to address the inverse problem where given u_{obs}. To this end, we extend representative flow-based methods to a joint-embedding setting (a, u) for comparison, and follow the official OFM (Shi et al., 2025) implementation to train a flow-matching prior with an FNO backbone (71.42M). Key model and training hyperparameters for prior learning are summarized in Tab. 10. All models are trained on a single NVIDIA RTX 4090 GPU.

Table 7: Model training hyperparameters.

Component	Hyperparameter	Value
Diffusion prior	Training duration	10M images
	Batch size	32
	Channel base c_{base}	64
	Channels per resolution c_{res}	$[1, 2, 4, 4]$
	Learning rate	1×10^{-4}
	LR ramp-up	5M images
	EMA half-life	0.5M images
	Dropout prob.	0.13
	# params	183M
FNO surrogate	Architecture	FNO with padding
	Fourier modes	$(64, 64)$
	# layers	4
	Hidden channels	64
	In / out channels	$1 \to 1$
	# params	8M

Table 8: Shared experimental setup (all tasks unless specified).

Item	Value
Tasks	Helmholtz / Poisson / Navier–Stokes
Default grid resolution (training)	128^2
Default grid resolution (inference)	Half 64^2 and half 128^2
Observation type	Sparse point observations
# observed points N_{obs}	$500 \approx 3\%$ on 128^2
Observation loss	ℓ_1
FNO padding p	2
RBF initialization scale	0.05
$\sigma_{\min}$	0.002
$\sigma_{\max}$	80
Noise exponent ρ	7
Diffusion step	$N = 5$
Diffusion	$\sigma_{\min} = 0.001, \sigma_{\text{final}} = 0$
Annealing noise range	$\sigma_{\max} = 10,\ \sigma_{\min} = 0.01$
Langevin step size η	0.1
Langevin noise scale τ	10^{-3}

Table 9: DDIS sampling hyperparameters for the three budget groups (Tab. 2).

Task	Time Budget	Anneal (N)	Langevin $(N,\ \text{weights}_1,\ \text{weights}_2)$
Poisson	$\sim$ 16s	100	20, 10, 15
	$\sim$ 32s	200	20, 10, 15
	$\sim$ 128s	200	100, 5, 5
Helmholtz	$\sim$ 16s	100	20, 5, 10
	$\sim$ 32s	200	20, 10, 10
	$\sim$ 128s	200	100, 5, 5
Navier-Stokes	$\sim$ 16s	100	20, 10, 20
	$\sim$ 32s	200	20, 10, 10
	$\sim$ 128s	200	100, 10, 10

H.3.2 Inference Setup

During inference, we evaluate the following flow-based posterior sampling methods, which operate on the same pretrained model described above, and compare them with our decoupled approach. To ensure a fair comparison and verify the quality of the learned prior, we also report performance on the "forward problem," defined as the reconstruction of the full solution field u from partial/masked observations u_{obs}, to align with these methods' original tasks.

ECI-sampling. ECI-sampling (Cheng et al., 2025) applies an extrapolation–correction–interpolation scheme during the flow ODE integration. It enforces hard constraints (e.g., Dirichlet conditions) by replacing observations in intermediate predictions and applying these ECI steps at each ODE step. Due to inherent method-level limitations in robustness, as discussed in App. A.5 and visualized in Fig. 6, we conducted an extensive hyperparameter search (summarized in Tab. 11), but observed only limited performance gains even on the forward problem, and thus report representative configurations in the main results.

OFM Regression. Operator Flow Matching (OFM) (Shi et al., 2025) interprets the function-space flow model as a bijective mapping from a Gaussian process (GP) prior space to the target function space, and formulates exact posterior inference in probability form. In practice, it requires solving a flow ODE alongside Hutchinson trace estimation, then relying heavily on Langevin Dynamics to optimize the latent GP variables. We follow the official implementation, using a learning rate of 1×10^{-3} and an observation noise of 1×10^{-3}; other inference hyperparameters are shown in Tab. 12.

H.3.3 Quantitative Results and Analysis

The quantitative performance is shown in Tab. 12. As these methods are originally designed for single-channel inverse problems, we also report their performance on forward problems for completeness. While OFM yields more competitive results than ECI-sampling, it encounters method-level bottlenecks that limit its practical utility in complex PDE settings:

Computational Efficiency. In practice, while the official OFM implementation suggests 20,000 Langevin steps for convergence, this is computationally prohibitive in our PDE setting, as it would exceed 20 hours per sample. This requirement, when coupled with adaptive ODE solvers and joint (a, u) backpropagation, significantly escalates VRAM consumption and makes standard inference budgets impractical. Qualitative results for OFM across different step counts are provided in Fig. 20.

Numerical Instability. Furthermore, the steep gradients from the data-fidelity term $\|M \odot u - u_{\text{obs}}\|^2/\sigma^2$ can drive samples off the learned prior manifold, leading to frequent numerical instabilities (e.g., NaN/Inf or OOM) under standard GPU budgets. To ensure a statistically meaningful and stable evaluation despite these challenges, we report the OFM regression average performance across 10 independent runs under practical constraints.

H.4 Details on Other Baseline Methods

For Baseline Methods, we compare with the following plug-and-play algorithms that solve PDE inverse problems via a joint-embedding diffusion models, as shown below.

DiffusionPDE. DiffusionPDE (Huang et al., 2024) uses a finite-dimensional diffusion model with a physics-informed DPS formulation, combining observation loss and PDE residuals in its posterior sampling. We use their official checkpoints (54.41M parameters) and maintain the original hyperparameter settings while varying the diffusion sampling steps (500, 1000, and 3000) to report results under different computational budgets in Tab. 2. The results for FNO (Li et al., 2020) and DeepONet (Lu et al., 2019) are taken from the original DiffusionPDE paper.

FunDPS. FunDPS (Yao et al., 2025) uses a joint-embedding diffusion model in function space and represents the prior state-of-the-art for sparse-observation inverse PDE problems. Their model has 54M parameters. We maintain the original hyperparameter settings while varying the diffusion sampling steps to evaluate performance under different budgets.

Table 10: Hyperparameter configurations for the flow-based prior models.

Hyperparameter	**Value**
Model Architecture	
Fourier Modes	32
Hidden Channels	128
MLP Projection Width	128
Visual Channels (a, u)	2
Spatial Dimensions	2D
GP Prior	
Kernel Lengthscale	0.01
Kernel Variance	1.0
Matérn Parameter	0.5
Minimum Noise	1×10^{-4}
Training Setup	
Total Epochs	300
Batch Size	100
Learning Rate	1×10^{-3}
Optimizer	Adam
LRScheduler	StepLR
Scheduler Step Size	50
Scheduler Decay	0.8

Table 11: **Performance comparison of ECI-sampling on the Poisson inverse problem.** $\ell_{2,\mathrm{fwd}}$ and $\ell_{2,\mathrm{inv}}$ denote the relative ℓ_2 error (%) for the forward and inverse problems, respectively. For cases where the model fails to produce physically meaningful results, relative errors exceeding 100% are capped at 100% for clarity.

Method	Euler Steps	ECI Steps	$\ell_{2,\mathrm{inv}}$	$\ell_{2,\mathrm{fwd}}$
ECI-sampling	100	1	94.56	67.47
	200	1	94.83	58.95
	200	5	95.29	42.64
	200	10	95.25	36.17
	400	5	95.18	35.95
	400	10	94.70	30.12
	800	5	94.63	30.04
	800	10	93.81	25.78
	2000	5	93.47	24.72
	2000	10	92.69	22.55
	4000	5	92.70	22.55
	4000	10	91.69	21.27

Table 12: **Performance of flow-based inverse solvers across PDE tasks.** Times (s/sample) are averaged across tasks and normalized by batch size. For entries without explicit Euler step counts, adaptive ODE solvers (e.g., `dopri5` (Dormand & Prince, 1980)) are used. $\ell_{2,\text{fwd}}$ and $\ell_{2,\text{inv}}$ denote the relative ℓ_2 error (%) for the forward and inverse problems, respectively.

Method	Euler Steps	ECI Steps	Langevin Steps	Hutchinson Samples	Poisson		Helmholtz		N-S		GPU	Time (s)
					$\ell_{2,\text{inv}}$	$\ell_{2,\text{fwd}}$	$\ell_{2,\text{inv}}$	$\ell_{2,\text{fwd}}$	$\ell_{2,\text{inv}}$	$\ell_{2,\text{fwd}}$		
ECI	800	5	/	/	94.63	30.04	92.83	25.57	42.36	20.61	RTX 4090	0.20
	2000	5	/	/	93.47	24.72	93.23	23.31	41.68	17.08	RTX 4090	0.38
OFM†	150	/	100	1	51.48	13.10	37.66	9.18	35.74	28.10	RTX 4090	246.05
	150	/	1000	4	31.31	15.26	39.81	29.04	29.78	23.02	RTX 4090	2445.13
	/	/	100	1	71.87	20.15	49.60	12.85	37.57	31.15	2× H100‡	470.44
	/	/	1000	4	47.04	7.41	42.07	8.84	20.98	13.55	2× H100‡	4366.34

† Unstable inference runs producing NaN or Inf values are assigned a relative ℓ_2 error of 100%.

‡ We select these configurations in Tab. 2 based on forward performance, as Euler solvers may introduce overly smooth shortcuts.

I Qualitative Results

In this section, we present qualitative results for forward and inverse problems, providing a visual comparison of representative methods to supplement the quantitative analysis in Tab. 2.

I.1 Inverse Poisson problem

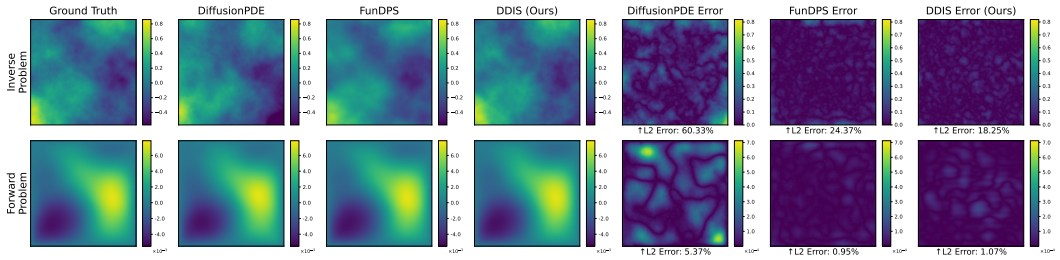

Figure 11: Qualitative comparison of results for the inverse Poisson problem (Time Budget: 16 s).

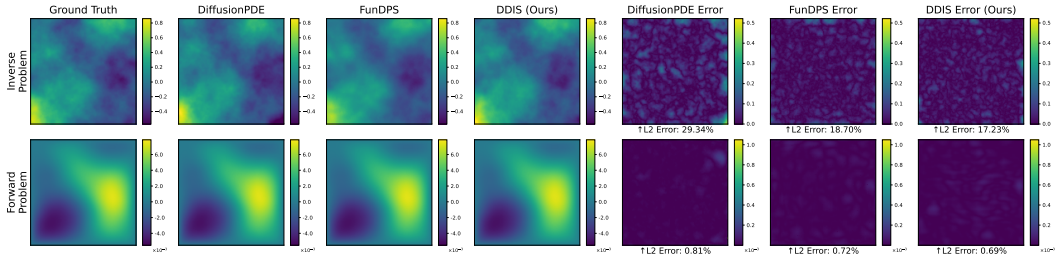

Figure 12: Qualitative comparison of results for the inverse Poisson problem (Time Budget: 32 s).

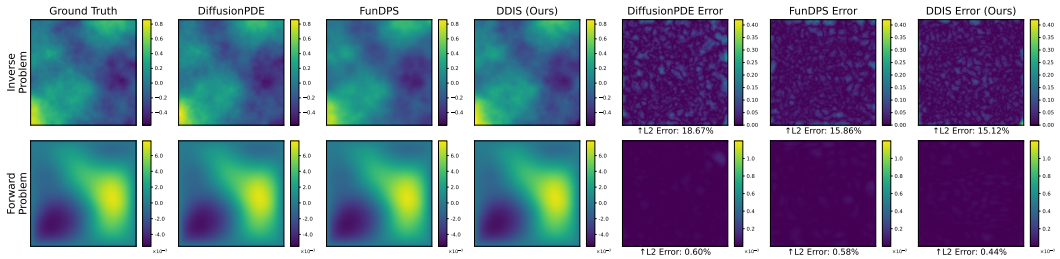

Figure 13: Qualitative comparison of results for the inverse Poisson problem (Time Budget: 128 s).

I.2 INVERSE HELMHOLTZ PROBLEM

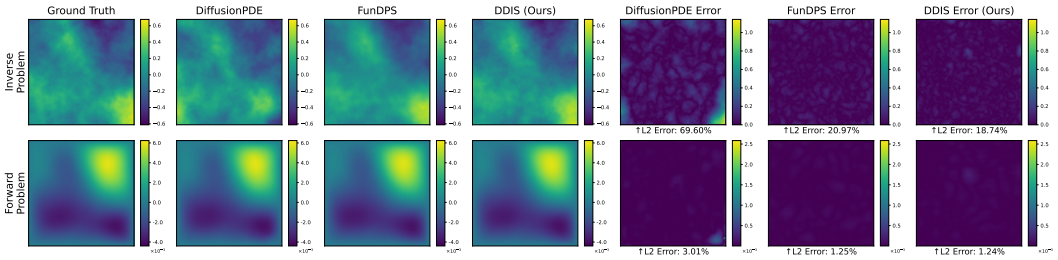

Figure 14: Qualitative comparison of results for the inverse Helmholtz problem (Time Budget: 16 s).

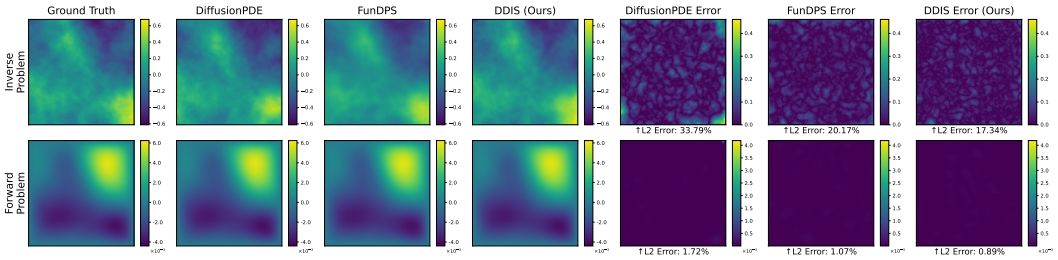

Figure 15: Qualitative comparison of results for the inverse Helmholtz problem (Time Budget: 32 s).

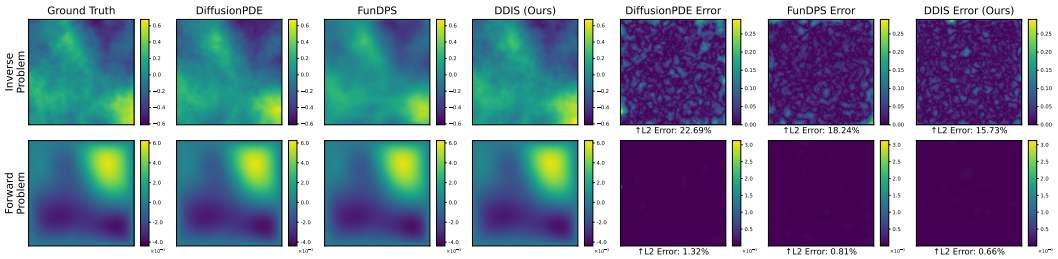

Figure 16: Qualitative comparison of results for the inverse Helmholtz problem (Time Budget: 128 s).

I.3 INVERSE NAVIER-STOKES PROBLEM

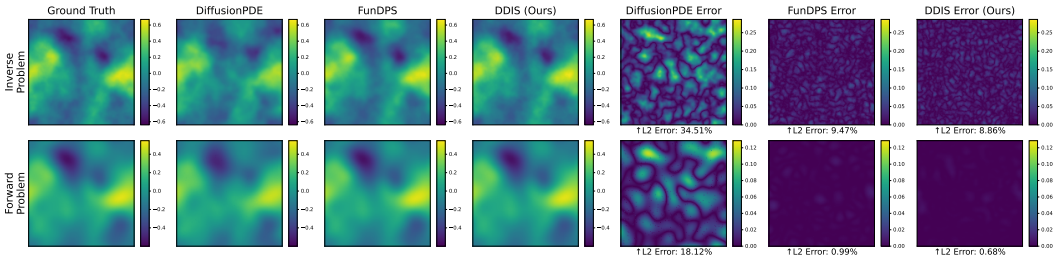

Figure 17: Qualitative comparison of results for the inverse Navier-Stokes problem (Time Budget: 16 s).

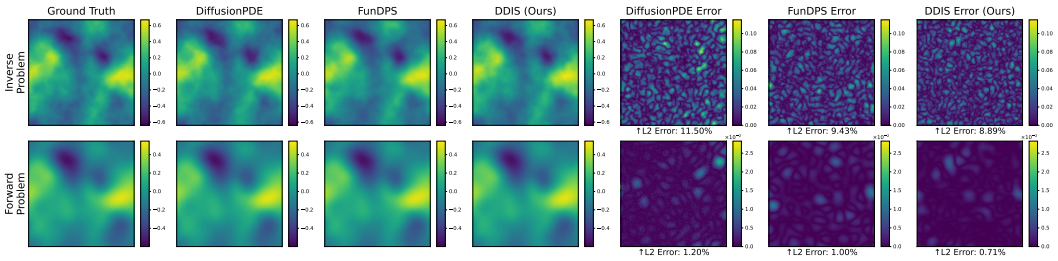

Figure 18: Qualitative comparison of results for the inverse Navier-Stokes problem (Time Budget: 32 s).

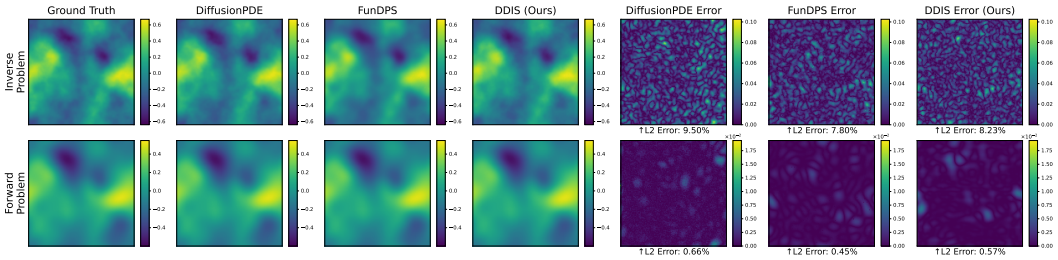

Figure 19: Qualitative comparison of results for the inverse Navier-Stokes problem (Time Budget: 128 s).

I.4 OPERATOR FLOW MATCHING

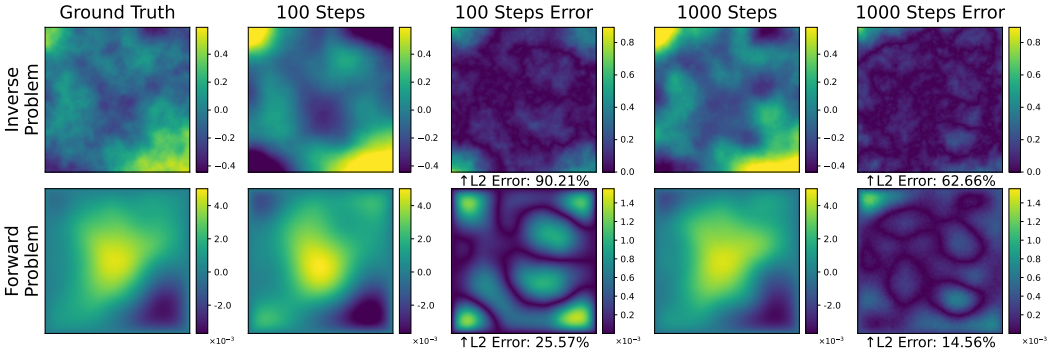

Figure 20: Generation mean with OFM regression across different Langevin steps on the Poisson inverse problem.