Sparsely-Supervised Data Assimilation via Physics-Informed Schrödinger Bridge

Dohyun Bu* **Chanho Kim*** **Seokun Choi** **Jong-Seok Lee†**
Korea Advanced Institute of Science and Technology
{dohyun.bu, chanh99kim, seokunchoi, jongseok.lee}@kaist.ac.kr

Abstract

Data assimilation (DA) for systems governed by partial differential equations (PDE) aims to reconstruct full spatiotemporal fields from sparse high-fidelity (HF) observations while respecting physical constraints. While full-grid low-fidelity (LF) simulations provide informative priors in multi-fidelity settings, recovering an HF field consistent with both sparse observations and the governing PDE typically requires per-instance test-time optimization, which becomes a major bottleneck in time-critical applications. To alleviate this, amortized reconstruction using generative models has recently been proposed; however, such approaches rely on full-field HF supervision during training, which is often impractical in real-world settings. From a more realistic perspective, we propose the *Physics-Informed Conditional Schrödinger Bridge* (PICSB), which transports an informative LF prior toward an auxiliary-energy-defined, observation-conditioned HF endpoint law without any additional inference-time guidance. To enable learning without HF endpoints, PICSB employs an iterative surrogate-endpoint refresh scheme, and directly incorporates PDE residuals into the training objective while enforcing observations via hard conditioning throughout sampling. Experiments on fluid PDE benchmarks demonstrate that PICSB enables extremely fast spatiotemporal field reconstruction while achieving strong accuracy under sparse HF supervision.

1 Introduction

Partial differential equations (PDEs) have long provided fundamental tools for modeling and predicting natural phenomena. A central challenge in many fluid and flow applications is full-field reconstruction: recovering a high-resolution spatiotemporal field from sparse observations while remaining consistent with the governing dynamics, a setting broadly studied under data assimilation (DA) [1, 2]. In such regimes, the available observations provide only partial coverage, and the underlying physical model provides feasibility constraints on recovering the full field [3, 4].

A common way to obtain accurate full-fields is high-fidelity (HF) numerical simulation, but it is often prohibitively time-consuming at the resolutions and regimes of interest [5, 6]. This computational latency is a major bottleneck in time-critical applications, including rapid-update numerical weather forecasting [7], operational ocean prediction [8], and real-time hydrologic forecasting [9]. Practitioners therefore rely on low-fidelity (LF) models or proxy fields—from simplified physics and coarse discretizations to observation-based interpolation—which are fast but systematically biased [10–12].

These constraints make multi-fidelity DA a common practical regime: abundant LF fields provide informative priors, while HF information is available only through sparse observations [13, 14]. However, even in this multi-fidelity setting, classical physics-based DA typically enforces observation and PDE consistency by solving a new constrained inverse problem for each instance, incurring

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

substantial computational cost [15]. This motivates amortized reconstruction, which maps a new LF field and a few HF observations directly to a full HF field without test-time re-optimization [4].

Recent progress toward amortized reconstruction has been driven by generative modeling. Diffusion models, in particular, have become a standard tool for high-dimensional conditional generation, making sparse-to-full field recovery a natural application [16, 17]. In PDE-constrained DA, a prevailing recipe learns a diffusion prior over fully observed HF fields and then produces reconstructions via iterative sampling, while enforcing consistency with sparse observations and the governing PDE via residual-based guidance during sampling [18–21]. While effective, this paradigm typically relies on many sampling steps with inference-time corrections to enforce observations and PDE constraints, increasing inference cost in time-critical settings.

An appealing alternative is provided by Schrödinger bridge (SB) formulations, which learn a stochastic optimal transport between endpoint distributions. For multi-fidelity DA, SB offers a particularly natural geometry by transporting samples from an informative LF field, often requiring substantially fewer sampling steps than diffusion models that start from uninformative noise. Although SB-based methods for PDE field modeling remain relatively scarce, SB-style transport has been studied extensively in broader generative settings such as image synthesis [22] and related bridge formulations [23], with recent PDE-based instances emerging as well [15]. However, existing SB approaches typically still rely on full-field HF supervision to learn the transport, which can be costly in practice because obtaining such supervision requires generating full-grid HF simulation targets [24, 25]. Consequently, amortized multi-fidelity DA under sparse HF supervision remains largely underexplored, particularly for PDE-governed systems that require physically consistent generative models.

In this work, we propose the *Physics-Informed Conditional Schrödinger Bridge Model* (PICSB) for amortized multi-fidelity DA with sparse HF supervision. PICSB models the conditional distribution of the full HF field given an LF input and sparse HF observations, enforcing the observations via hard conditioning. To avoid reliance on full-field HF supervision, we train PICSB solely from physics residuals, using an auxiliary-energy-driven conditional bridge that starts from an informative LF prior and transports it toward an observation-conditioned endpoint law induced by the governing PDE constraints and sparse observations. This physics-residual objective enables guidance-free inference while retaining the efficiency of LF-to-HF transport, yielding fast reconstruction with few sampling steps. Crucially, since HF endpoints are unavailable, we enable bridge learning via an iterative surrogate-endpoint refresh that periodically updates a slowly moving terminal reference distribution. Our contributions are as follows:

- We introduce a conditional SB for amortized reconstruction from LF inputs and sparse HF observations via hard conditioning, enabling fast inference with few sampling steps.
- We characterize the population endpoint selected by PICSB as the LF-proximal minimizer of an auxiliary inverse-PDE energy.
- We empirically show that PICSB achieves competitive accuracy under sparse HF supervision while delivering orders-of-magnitude faster reconstruction than test-time optimization and guidance-based sampling.

2 Related Works

Optimization-based DA. Traditional per-instance DA for PDE-governed systems has primarily relied on numerical methods that solve observation-conditioned inverse problems under physical constraints [26, 27]. To reduce the high computational cost of such approaches, extensive efforts have focused on efficiency-oriented modeling aimed at HF simulations [28, 29]. A related class of approaches replaces explicit numerical solvers with neural-network-based optimization formulations, which parameterize solution fields with neural networks and minimize a composite objective that enforces both observation fitting and governing PDE constraints [30]. More recently, implicit neural representation methods have been explored to reconstruct continuous spatiotemporal fields from sparse observations by directly optimizing latent representations [31].

Amortized Reconstruction. To avoid repeated per-instance optimization, recent work has sought zero-shot DA by learning reusable models that can be directly queried at test time. One line of research learns forward solution operators that map PDE parameters or initial conditions to full fields, enabling fast evaluation without iterative solvers [32–34]. A second line of work targets amortized inverse reconstruction via generative models, typically trained with fully observed HF fields. Many

Table 1: Comparison of learning-based field reconstruction methods for PDE-governed systems. **Zero-shot** indicates amortized reconstruction without test-time re-optimization. **DA** indicates the capability to solve an observation-conditioned inverse problem using sparse HF observations. **Label-Efficient** indicates training under sparse HF supervision, where HF targets are available only at partial spatiotemporal locations. **Guidance-free** indicates that PDE residuals are incorporated directly into training, avoiding inference-time correction.

Method	Zero-shot	DA	Label-Efficient	Guidance-free
PINNs [30]	✗	✓	✓	✓
MMGN [31]	✗	✓	✓	✗
FNO, NOICL [32, 33]	✓	✗	✗	✗
P^{2}INN [34]	✓	✗	✓	✓
DiffusionPDE, ECI, PCFM [18–20]	✓	✓	✗	✗
PIDM, PalSB, PPIDM [35, 15, 17]	✓	✓	✗	✓
PICSB (Ours)	✓	✓	✓	✓

diffusion-based formulations enforce observation and PDE consistency through inference-time residual guidance [18–20], whereas more recent approaches incorporate PDE residuals into the training objective to reduce reliance on inference-time corrections [35, 15, 17].

Limitations. Table 1 summarizes prior learning-based field reconstruction methods for PDE-governed systems in terms of four desirable features: **Zero-shot** inference, **DA** capability, **Label-Efficient** learning under sparse HF supervision, and **Guidance-free** physics integration via training objectives. First, optimization-based methods can assimilate sparse observations while explicitly enforcing the governing physics, but they are not zero-shot and require test-time re-optimization. Second, operator-learning methods enable zero-shot evaluation but largely act as forward surrogates and thus lack DA capability for observation-conditioned reconstruction. Finally, generative inverse methods provide DA capability, but most existing approaches are not label-efficient, as they assume full-field HF supervision; moreover, guidance-based formulations incur additional inference-time correction costs. Consequently, no existing method simultaneously satisfies all four criteria.

3 Preliminaries

We consider a multi-fidelity field reconstruction problem on a fixed discrete grid $\mathcal{X}=\mathbb{R}^n$. $X_0 \in \mathcal{X}$ denotes a LF field defined on the full grid, $X_1 \in \mathcal{X}$ denotes the corresponding latent HF field on the same grid, and $Y \in \mathbb{R}^m$ denotes sparse HF observations with $m \ll n$. The observations are related through a known operator $\mathcal{H} : \mathcal{X} \to \mathbb{R}^m$ such that $Y = \mathcal{H}(X_1)$. Given $Y = y$, we denote by $\pi_0(\cdot \mid y)$ the conditional LF distribution. In a fully supervised bridge setting, one would also have access to full-field HF endpoint samples from the conditional distribution $\pi_1(\cdot \mid y)$. In our sparse-supervision setting, however, such full-field HF endpoints are not empirically available. We therefore specify the desirable HF endpoint through an auxiliary inverse-PDE energy, rather than through observed full-field HF labels.

3.1 Desirable HF endpoint via auxiliary inverse-PDE energy

For a fixed observed pair (x_0, y) where x_0 is a realization of the LF field X_0, let $\mathcal{C}_y := \{x \in \mathcal{X} : \mathcal{H}(x) = y\}$ be the observation-consistent set and $\mathcal{R}_h : \mathcal{X} \to \mathbb{R}^r$ be the discretized PDE residual operator on grid scale h. Following the least-squares finite element viewpoint, a PDE solution can be characterized as a minimizer of a squared residual functional [36]. We therefore define

$$\mathcal{E}_{\lambda,h}(x; x_0, y) := \frac{\lambda_{\mathrm{PDE}}}{2}\|\mathcal{R}_h(x)\|_2^2 + \frac{\lambda_{\mathrm{LF}}}{2}\|x - x_0\|_2^2, \quad x \in \mathcal{C}_y. \tag{1}$$

We set $\mathcal{E}_{\lambda,h}(x; x_0, y) = +\infty$ for $x \notin \mathcal{C}_y$. Here, the first term measures PDE consistency, while the second term anchors the solution to the LF input. Because the energy is finite only on $\mathcal{C}_y$, every finite-energy endpoint satisfies the sparse observations exactly.

The desirable HF endpoint law for a given LF input and sparse observation is then defined as the Gibbs law

$$\pi_{1+}^{\lambda,\beta}(dx \mid x_0, y) = \frac{1}{Z_{\lambda,\beta}(x_0, y)} \exp[-\beta\, \mathcal{E}_{\lambda,h}(x; x_0, y)]\; d\lambda_y(x), \tag{2}$$

where $\beta > 0$ is an inverse temperature, $Z_{\lambda,\beta}$ is the normalizing constant, and λ_y denotes the Lebesgue measure over the unobserved coordinates with the observed coordinates fixed to y. Thus, in the sparse inverse-PDE setting, the target HF endpoint is not an observed full-field label; it is an observation-consistent, PDE residual minimizing, LF-anchored endpoint induced by the auxiliary energy. For compact notation, we write $\pi_{1+}(\cdot \mid y)$ for the marginal endpoint law obtained by averaging the input-conditioned law $\pi_{1+}^{\lambda,\beta}(\cdot \mid x_0, y)$ over $x_0 \sim \pi_0(\cdot \mid y)$.

3.2 Conditional Schrödinger bridge formulation

For a fixed conditioning variable $Y = y$, the conditional SB problem seeks a path measure $\mathbb{P}(\cdot \mid y)$ on $\mathcal{C} = C([0,1], \mathbb{R}^n)$ whose endpoint marginals satisfy $\mathbb{P}_0(\cdot \mid y) = \pi_0(\cdot \mid y)$ and $\mathbb{P}_1(\cdot \mid y) = \pi_1(\cdot \mid y)$ [37]. Specifically, given an uncontrolled Wiener measure $\mathbb{Q}$ as a reference path measure, the conditional SB is defined as

$$\mathbb{P}^\star(\cdot \mid y) = \underset{\mathbb{P}(\cdot \mid y)}{\operatorname{argmin}} \, \mathrm{KL}(\mathbb{P}(\cdot \mid y) \,\|\, \mathbb{Q}), \quad \text{s.t.} \quad \mathbb{P}_0(\cdot \mid y) = \pi_0(\cdot \mid y), \quad \mathbb{P}_1(\cdot \mid y) = \pi_1(\cdot \mid y). \tag{3}$$

In supervised bridge learning, the terminal distribution is usually specified by empirical HF endpoint samples. In our inverse-PDE setting, it is instead specified implicitly by the auxiliary energy in Eq. (1). The SB formulation then provides a principled way to transport probability mass from the LF-informed source distribution $\pi_0(\cdot \mid y)$ toward this desirable endpoint law $\pi_{1+}(\cdot \mid y)$ while remaining close to simple reference dynamics.

3.3 Bridge reference and flow-based parameterization

Under $\mathbb{Q}$, conditioning on the endpoints $(X_0, X_1) = (x_0, x_1)$ yields the Brownian bridge $\mathbb{Q}(\cdot \mid x_0, x_1)$. With noise scale $\varepsilon > 0$, the intermediate state satisfies

$$X_\tau \mid (x_0, x_1) \;\sim\; \mathcal{N}\big((1-\tau)x_0 + \tau x_1,\; \varepsilon\,\tau(1-\tau)\,I\big), \tag{4}$$

which allows direct sampling at an arbitrary intermediate time τ. This property underlies Iterative Markovian Fitting (IMF) and its diffusion-based implementations [22, 38, 39].

To construct an explicit sampler, one may parameterize a conditional path measure via a time-dependent velocity field $v_\tau(\cdot \mid y)$ [40]. Such a parameterization induces the controlled diffusion driven by standard Brownian motion B_τ,

$$dX_\tau = v_\tau(X_\tau \mid y)\, d\tau + \sqrt{\varepsilon}\, dB_\tau, \quad X_0 \sim \pi_0(\cdot \mid y). \tag{5}$$

When paired samples $\{(x_0^i, y^i, x_1^i)\}_{i=1}^N$ are available, IMF methods learn v by matching the drift of the Brownian bridge, leading to the bridge-matching objective given by

$$\min_v \; \mathbb{E}_{i,\tau}\left[\mathbb{E}_{x_\tau^i \sim \mathbb{Q}(\cdot \mid x_0^i, x_1^i)} \left\| v_\tau(x_\tau^i \mid y^i) - \frac{x_1^i - x_\tau^i}{1-\tau} \right\| \right]. \tag{6}$$

In our sparse-supervision setting, the full-field endpoints x_1^i are unavailable; the methodology below therefore uses model-induced surrogate endpoints shaped by the auxiliary energy in Eq. (1).

4 Methodology

We consider the multi-fidelity DA setting in Section 3. The training data consist of pairs $\mathcal{D}_{\text{train}} = \{(x_0^i, y^i)\}_{i=1}^N$, where each $x_0^i \in \mathbb{R}^n$ is a full-grid LF field and $y^i \in \mathbb{R}^m$ denotes sparse HF observations of an unobserved latent HF field $x_1^i \in \mathbb{R}^n$, i.e., $y^i = \mathcal{H}(x_1^i)$. The full-field HF endpoint x_1^i is not used as a training label. Given a new pair (x_0, y) at test time, our goal is to sample plausible HF reconstructions that respect the LF prior information in x_0, satisfy the sparse observations y, and remain physically consistent. We adopt the auxiliary-energy-driven conditional SB viewpoint introduced in Section 3.1–3.2. In the idealized formulation, the conditional bridge transports the LF source law $\pi_0(\cdot \mid y)$ toward the desirable endpoint law $\pi_{1+}(\cdot \mid y)$, rather than toward an empirical full-field HF distribution. In practice, we are given a concrete LF instance x_0 rather than an explicit conditional density $\pi_0(\cdot \mid y)$. Accordingly, our sampler is implemented as an input-conditioned transport process that starts from the provided x_0 and is explicitly conditioned on the sparse observations y, generating a terminal sample $\hat{x}_1 \sim \pi_\theta(\cdot \mid x_0, y)$ whose population target is the desirable endpoint law $\pi_{1+}^{\lambda,\beta}(\cdot \mid x_0, y)$.

4.1 Physics-informed conditional reconstruction

Recall from Section 3.1 that the desirable endpoint is defined on the observation-consistent set $\mathcal{C}_y$ and shaped by the PDE residual operator $\mathcal{R}_h$. Here we specify how these two components are enforced in the sampler. Throughout, we focus on the standard sparse-sensor setting where $\mathcal{H}$ selects a subset of grid coordinates. Let $\odot$ denote element-wise multiplication. We use the coordinate projection onto $\mathcal{C}_y$:

$$\Pi_{\mathcal{C}_y}(x) \;:=\; \mathbf{M} \odot \tilde{y} \;+\; (1 - \mathbf{M}) \odot x, \tag{7}$$

where $\mathbf{M} \in \{0,1\}^n$ is a binary mask indicating the observed coordinates and $\tilde{y} \in \mathbb{R}^n$ denotes the full-grid embedding of the sparse observations. When no confusion arises, we simply write y for $\tilde{y}$.

Due to discretization and numerical approximation, we define the residual floor and δ-feasible physics set as

$$\delta^\star(y; h) \;:=\; \min_{x \in \mathcal{C}_y} \|\mathcal{R}_h(x)\|, \qquad \mathcal{P}^{\delta, y}_{\mathcal{R}_h} \;:=\; \{x \in \mathcal{C}_y : \|\mathcal{R}_h(x)\| \le \delta\}. \tag{8}$$

Our goal is to learn a conditional sampler whose terminal samples remain in $\mathcal{C}_y$ by construction and have small PDE residual. The LF anchoring term in Eq. (1) is implemented through the Brownian-bridge reference initialized from x_0, so that the learned transport is biased toward HF endpoints close to the LF input.

4.2 Learning with surrogate endpoints

We now specify the sampling procedure and the training objective. Our sampler is parameterized by a time-dependent velocity field $v_\theta(x, \tau \mid y)$, inducing the controlled diffusion in Eq. (5). Hard observation feasibility is enforced by projection at every discretized step.

Surrogate bridge reference. Bridge-matching objectives such as Eq. (6) require paired HF endpoints x_1, which are unavailable in our setting. We therefore introduce a surrogate terminal reference $\tilde{\pi}_1(\cdot \mid x_0, y)$, defined as the terminal law induced by a slowly updated static velocity network $v_{\tilde{\theta}}$ under hard conditioning. Concretely, drawing $\tilde{x}_1 \sim \tilde{\pi}_1(\cdot \,|\, x_0, y)$ means running the sampler $\mathrm{Sample}_{\tilde{\theta}}$ initialized from the given LF field x_0 and projected onto $\mathcal{C}_y$ throughout. Given (x_0, y), we sample $\tau \sim \mathrm{Unif}([0,1])$ and draw an intermediate state $\tilde{x}_\tau$ from the Brownian-bridge closed form in Eq. (4) between endpoints $(x_0, \tilde{x}_1)$, followed by projection to ensure $\tilde{x}_\tau \in \mathcal{C}_y$. Starting from $\tilde{x}_\tau$, Algorithm 1 produces a terminal sample $\hat{x}_1 = \mathrm{Sample}_\theta(\tilde{x}_\tau \mid y, \tau)$.

Algorithm 1 Hard-conditioned sampler Sample_θ

Input: initial state x_τ at time $\tau = \tau_{t_0}$; sparse observations y; velocity network v_θ; projection $\Pi_{\mathcal{C}_y}$ in Eq. (7); noise scale ε; step size s.

Output: terminal sample x_{τ_T}.

Initialize: $x_{\tau_{t_0}} \leftarrow \Pi_{\mathcal{C}_y}(x_\tau)$.

For $t = t_0, \ldots, T-1$, draw $z_t \sim \mathcal{N}(0, I)$ and update

$$x_{\tau_{t+1}} \leftarrow \Pi_{\mathcal{C}_y}\big(x_{\tau_t} + s\, v_\theta(x_{\tau_t}, \tau_t \mid y) + \mathbf{1}_{\{t<T-1\}} \sqrt{\varepsilon s}\, z_t\big)\,.$$

Physics-residual training objective. We train the *trainable* sampler v_θ using the PDE-consistency term of the auxiliary energy:

$$\mathcal{L}(\theta) \;:=\; \mathbb{E}_{(x_0,y)\sim\mathcal{D}_{\mathrm{train}}} \mathbb{E}_{\tau\sim\mathrm{Unif}([0,1])} \left[\frac{1}{2}\|\mathcal{R}_h(\hat{x}_1)\|_2^2\right], \tag{9}$$

The observation constraint in Eq. (1) is enforced by hard projection, while the LF anchoring is induced by constructing the bridge from x_0 under the Brownian reference. Thus, the residual objective shapes the terminal samples toward the auxiliary energy defined desirable endpoint without requiring empirical full-field HF labels.

Iterative refresh of $\tilde{\pi}_1$. Because $\tilde{\pi}_1(\cdot \,|\, x_0, y)$ is model-induced, we adopt a periodic refresh scheme with a slowly updated static model, analogous to target-network stabilization [41]. At refresh round k, we (i) sample surrogate endpoints $\tilde{x}_1 \sim \tilde{\pi}_1^{(k)}(\cdot \,|\, x_0, y)$ using $v_{\tilde{\theta}}$ to instantiate a surrogate Brownian-bridge reference, and (ii) update θ by minimizing Eq. (9) under this fixed reference. This yields a distributional update map

$$\tilde{\pi}_1^{(k)}(\cdot \,|\, x_0, y) \;\mapsto\; \tilde{\pi}_1^{(k+1)}(\cdot \,|\, x_0, y), \tag{10}$$

where $\tilde{\pi}_1^{(k+1)}$ denotes the terminal law induced by the updated sampler v_θ. The refresh procedure should therefore be understood as amortizing the auxiliary inverse-PDE solve defined in Section 3.1.

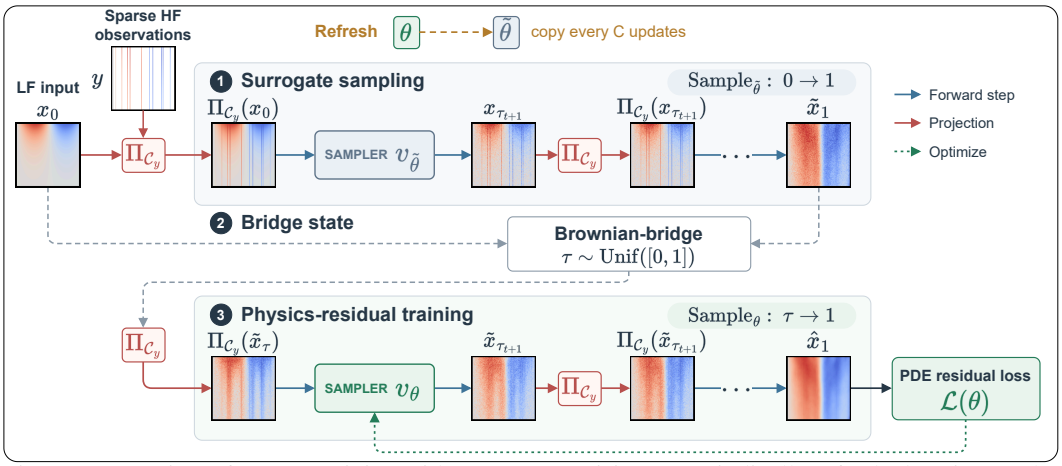

Figure 1: Overview of PICSB training with sparse supervision. A periodically refreshed static sampler with parameters $\tilde{\theta}$ generates surrogate endpoints for bridge construction, while the trainable sampler with parameters θ is optimized using the PDE residual under hard observation projection.

The following theorem characterizes the limiting population endpoint of the PICSB refresh map under a vanishing terminal free-energy gap induced by the auxiliary energy in Eq. (1).

Theorem 4.1 (Population endpoint characterization under a vanishing terminal free-energy gap)**.** *Fix (x_0, y). Assume that hard projection is exact, $\mathcal{C}_y$ is nonempty and closed, $\mathcal{R}_h$ is continuous on $\mathcal{C}_y$, and the residual floor $\delta^\star(y; h)$ in Eq.* (8) *is attained. For fixed $\lambda_{\mathrm{PDE}}, \lambda_{\mathrm{LF}}, \beta > 0$, let $\{\tilde{\pi}_1^{(k)}(\cdot|x_0, y)\}_{k\geq 0}$ be the population terminal laws generated by the PICSB refresh map in Eq.* (10)*. If the terminal free-energy gap of this refresh sequence under $\mathcal{E}_{\lambda,h}$ vanishes, then $\tilde{\pi}_1^{(k)}(\cdot|x_0, y)$ converges to $\pi_{1+}^{\lambda,\beta}(\cdot|x_0, y)$ in KL divergence, and hence also in total variation and weakly. Moreover, in the zero-temperature and residual-dominant limit $\beta \to \infty$ and $\lambda_{\mathrm{PDE}}/\lambda_{\mathrm{LF}} \to \infty$, the population PICSB endpoint law concentrates on*

$$\mathcal{X}_{1+}^\star(x_0, y) := \mathrm{argmin}_x \left\{ \|x - x_0\|_2^2 \,:\, x \in \mathcal{P}_{\mathcal{R}_h}^{\delta^\star, y} \right\}.$$

Theorem 4.1 characterizes the limiting endpoint approached by the population PICSB refresh map under the stated free-energy condition. As the terminal free-energy gap vanishes, the terminal law converges to the Gibbs endpoint law in Eq. (2). The concentration result further shows that, in the inverse-PDE limit, hard projection enforces observation consistency, the PDE residual selects residual-minimizing states in $\mathcal{C}_y$, and the Brownian-bridge geometry supplies the LF anchor that selects the endpoint closest to x_0 when multiple residual minimizers exist. Thus, this population characterization links the PICSB endpoint to an LSFEM [36]-style auxiliary-energy label. A formal proof is given in Appendix A.1.

4.3 PICSB training procedure

Figure 1 summarizes the PICSB training procedure. Training alternates between a *static* sampler $v_{\tilde{\theta}}$, which supplies the current surrogate bridge reference, and a *trainable* sampler v_θ, which is optimized using the PDE residual. At the beginning of each refresh round, the current parameters are copied as $\tilde{\theta} \leftarrow \theta$, and $v_{\tilde{\theta}}$ is held fixed for the following C training updates. Given a training pair (x_0, y), each update follows the three stages.

First, in *surrogate sampling*, the static sampler starts from the LF field x_0 and generates $\tilde{x}_1 = \mathrm{Sample}_{\tilde{\theta}}(x_0 \mid y, 0)$, which serves as the current surrogate HF endpoint in place of an unavailable full-field HF label. The bridge time $\tau \in [0, 1]$ is discretized on a uniform grid $\tau_t = t \cdot s$ for $t = 0, \ldots, T$, with $s = 1/T$. At each sampling step, the controlled drift is advanced, Brownian noise is injected, and the resulting state is projected by $\Pi_{\mathcal{C}_y}$. Consequently, every discretized state remains in $\mathcal{C}_y$. Following common inpainting practice, the velocity network receives the projected state [22, 42] together with the corresponding bridge-time embedding. Second, in *bridge-state construction*, we sample $\tau \sim \mathrm{Unif}([0, 1])$ and draw $\tilde{x}_\tau$ from the Brownian-bridge closed form in Eq. (4) using the endpoint pair $(x_0, \tilde{x}_1)$. The sampled bridge state is then projected as $\tilde{x}_\tau \leftarrow \Pi_{\mathcal{C}_y}(\mathrm{BrownianBridge}(x_0, \tilde{x}_1, \tau))$,

Algorithm 2 Training procedure with iterative $\tilde{\pi}_1$ refresh

Input: D_{train}; refresh period C and rounds L; trainable velocity $v_{\theta^{(0)}}$; projection $\Pi_{\mathcal{C}_y}$; sampler $\text{Sample}_\theta(\cdot)$ in Algorithm 1; $\text{BrownianBridge}(\cdot)$; residual operator $\mathcal{R}_h(\cdot)$.

Output: trained parameters θ.

for $k = 0, 1, \ldots, L-1$ **do**
  (0) Refresh static model: $\tilde{\theta} \leftarrow \theta$.
  for $j = 0, 1, \ldots, C-1$ **do**
    (1) Surrogate endpoint: $\tilde{x}_1 \leftarrow \text{Sample}_{\tilde{\theta}}(x_0 \mid y, 0)$ where $(x_0, y) \sim D_{\text{train}}$.
    (2) Bridge state: $\tau \sim \text{Unif}([0,1])$ and $\tilde{x}_\tau \leftarrow \Pi_{\mathcal{C}_y}\big(\text{BrownianBridge}(x_0, \tilde{x}_1, \tau)\big)$.
    (3) Loss (Eq. (9)) and update: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta)$ where $\hat{x}_1 = \text{Sample}_\theta(\tilde{x}_\tau \mid y, \tau)$.
  end for
end for

so that the trainable transport starts from an observation-consistent intermediate state. Third, in *physics-residual training*, the trainable sampler evolves this bridge state to $\hat{x}_1 = \text{Sample}_\theta(\tilde{x}_\tau \mid y, \tau)$, again enforcing hard projection throughout the trajectory. The parameters θ are then updated by minimizing Eq. (9). Together, these stages realize an amortized auxiliary inverse-PDE solve: the Brownian bridge provides the LF-to-HF transport geometry, hard projection enforces the sparse observations, and the PDE residual supplies the implicit endpoint energy without requiring empirical full-field HF endpoints. This energy-defined endpoint interpretation is aligned with data-to-energy SB learning [43] and is analogous to simulation-based Bayesian learning via diffusion bridges [44]. Algorithm 2 summarizes the three stages together with the periodic refresh of the static sampler.

5 Experiments

We evaluate PICSB on four PDE benchmarks—Burgers, Darcy flow, Gray–Scott, and Kolmogorov flow—in a multi-fidelity DA setting where sparse HF measurements are used both for sparse training supervision and for inference-time conditioning. The benchmarks include fluid-related systems, namely Burgers, Darcy flow, and Kolmogorov flow, as well as the Gray–Scott reaction–diffusion system. They span steady-state and time-dependent systems: Darcy flow is steady-state, while the remaining benchmarks are dynamic; they also cover both scalar HF fields and the multi-channel concentration field in Gray–Scott.

To reflect practical multi-fidelity settings, we consider two LF acquisition scenarios. For Burgers, the HF and LF simulations are initialized from the same analytical initial condition but evolved independently, with the LF field generated using a coarser-grid solver and mismatched viscosity, whereas for Darcy flow, Gray–Scott, and Kolmogorov flow, the LF input is an observation-based coarse proxy constructed from sparse HF measurements using nearest-neighbor or bicubic interpolation, depending on the benchmark. We randomly resample 10% of the relevant grid locations for each instance and use the corresponding HF values as the only HF information available to the model, with full HF fields reserved solely for evaluation. For Burgers, Darcy flow, and Gray–Scott, we use 32 training fields and evaluate on 40 held-out test fields. For Kolmogorov flow, where the available trajectories are limited, we use a 5-fold protocol to secure the same evaluation scale: each fold trains on 32 fields and tests on the remaining 8 fields, and final statistics are aggregated over all 40 held-out evaluations. Detailed PDE definitions, solution-field sources, LF construction procedures, implementation details, and visualization settings are deferred to Appendix B.

5.1 Reference Baselines and Evaluation Metrics

To the best of our knowledge, no off-the-shelf method is directly applicable to our setting, since a valid method must jointly support *zero-shot* reconstruction, observation-conditioned *DA* from sparse HF measurements, and training without full-field HF supervision; see Table 1. We therefore report two reference baselines to contextualize performance, and also include the error of the LF input itself as a reference for the initial LF–HF discrepancy. *PINNs* relax the zero-shot requirement and serve as a strong per-instance optimization reference for assessing the difficulty of each benchmark. *Guidance* is the closest amortized comparison: we adapt a PDE-guided diffusion sampler to the same sparse supervision regime by replacing HF prior training with LF pretraining.

Table 2: Main quantitative comparison over 40 test samples. Error entries report 95% confidence intervals, and wall-time entries report average per-sample inference time. Lower is better; **bold** marks the lowest mean, and † denotes CI overlap with the best method.

Dataset	Metric	*LF*	*PINNs*	*Guidance*	PICSB
Burgers	RMSE ($\times 10^{-2}$)	[9.576, 11.183]	[1.893, 2.842]†	[2.969, 4.102]	**[2.223, 2.381]**
	MAE ($\times 10^{-3}$)	[78.461, 90.933]	[14.110, 21.402]	[17.417, 24.602]	**[5.752, 6.798]**
	RelError (%)	[70.426, 71.428]	[12.445, 22.565]†	[20.177, 25.015]	**[14.889, 19.700]**
	fRMSE ($\times 10^{-4}$)	[9.959, 11.889]	[3.463, 5.093]	[4.520, 6.298]	**[3.069, 3.453]**
	Average wall-time (seconds)	-	4,068	155	**0.026**
Darcy	RMSE ($\times 10^{-3}$)	[1.808, 1.845]	[2.496, 3.212]	[1.519, 1.551]	**[0.121, 0.129]**
	MAE ($\times 10^{-3}$)	[1.296, 1.311]	[1.998, 2.575]	[1.093, 1.103]	**[0.089, 0.094]**
	RelError (%)	[4.405, 4.451]	[6.048, 7.833]	[3.699, 3.745]	**[0.293, 0.313]**
	fRMSE ($\times 10^{-4}$)	[0.199, 0.205]	[0.384, 0.518]	[0.178, 0.185]	**[0.018, 0.020]**
	Average wall-time (seconds)	-	4,647	141	**0.013**
Gray–Scott	RMSE ($\times 10^{-1}$)	[1.159, 1.287]	**[0.273, 0.362]**	[1.240, 1.370]	[0.330, 0.370]†
	MAE ($\times 10^{-2}$)	[5.870, 6.990]	[1.109, 1.667]	[6.598, 7.689]	**[0.859, 1.061]**
	RelError (%)	[17.521, 19.740]	**[4.129, 5.534]**	[18.736, 21.028]	[4.989, 5.693]†
	fRMSE ($\times 10^{-3}$)	[0.537, 0.600]	**[0.167, 0.312]**	[1.022, 1.142]	[0.219, 0.268]†
	Average wall-time (seconds)	-	2,557	498	**0.109**
Kolmogorov	RMSE ($\times 10^{0}$)	[0.668, 0.729]	[2.971, 3.217]	[0.547, 0.587]	**[0.196, 0.220]**
	MAE ($\times 10^{0}$)	[0.286, 0.319]	[2.344, 2.555]	[0.312, 0.338]	**[0.109, 0.124]**
	RelError (%)	[13.497, 14.433]	[60.590, 63.026]	[11.043, 11.665]	**[3.972, 4.299]**
	fRMSE ($\times 10^{-2}$)	[0.159, 0.173]	[3.387, 3.695]	[0.232, 0.247]	**[0.118, 0.144]**
	Average wall-time (seconds)	-	9,527	787	**0.028**

Physics-Informed Neural Networks (*PINNs*) When the *zero-shot* requirement is removed, PINNs provide a natural re-optimization-based DA reference. We parameterize the unknown solution field that takes spatio-temporal coordinates as input. For each test instance, the network is optimized at inference time using a data residual on sparse HF observations and a PDE residual evaluated at collocation points. Thus, PINNs are reported as an infeasible *zero-shot* reference.

Guidance-based sampling (*Guidance*) Guidance is adapted from PDE-guided diffusion sampling [18]. While the original formulation learns a generative prior from full-grid HF fields, we pretrain the prior only on full-grid LF fields. At inference time, the LF-pretrained sampler is steered by gradient-based guidance from two losses: a data residual on the sparse HF observations and a PDE residual over the reconstructed field. This baseline satisfies the same sparse-HF information constraint as PICSB and provides the closest zero-shot DA comparison.

Evaluation metrics Following the evaluation protocol of RealPDEBench [45], we assess reconstruction quality using complementary spatial-domain and spectral-domain errors. In the spatial domain, we report RMSE, MAE, and relative L^2 error between the prediction $\hat{x}_1$ and the held-out HF reference x_1. As a physics-oriented spectral metric, we report fRMSE, which measures reconstruction error in Fourier space and evaluates whether the recovered field preserves physically relevant spatial-frequency content. We additionally report average per-sample inference wall-time.

5.2 Results

Table 2 reports 95% confidence intervals (CIs) for reconstruction and spectral errors together with average inference wall-time, while Figure 2 visualizes representative reconstructions. Additional experimental settings and analyses are reported in Appendix C.

Accuracy. PICSB achieves competitive overall reconstruction accuracy, obtaining the lowest mean errors on Burgers, Darcy flow and Kolmogorov flow, while remaining statistically competitive with the infeasible *zero-shot PINNs* reference on Gray–Scott. On Burgers, Darcy flow, and Kolmogorov flow, PICSB achieves lower mean errors than both *PINNs* and *Guidance* across all metrics, with non-overlapping improvements on Darcy flow and Kolmogorov flow and CI overlap with *PINNs* on some Burgers metrics, showing consistent gains under sparse HF information. On Gray–Scott, PICSB overlaps with *PINNs* on these metrics and achieves the lowest MAE, whereas *Guidance* is consistently worse. The robustness study in Table 7 further shows that PICSB remains competitive under more challenging conditions.

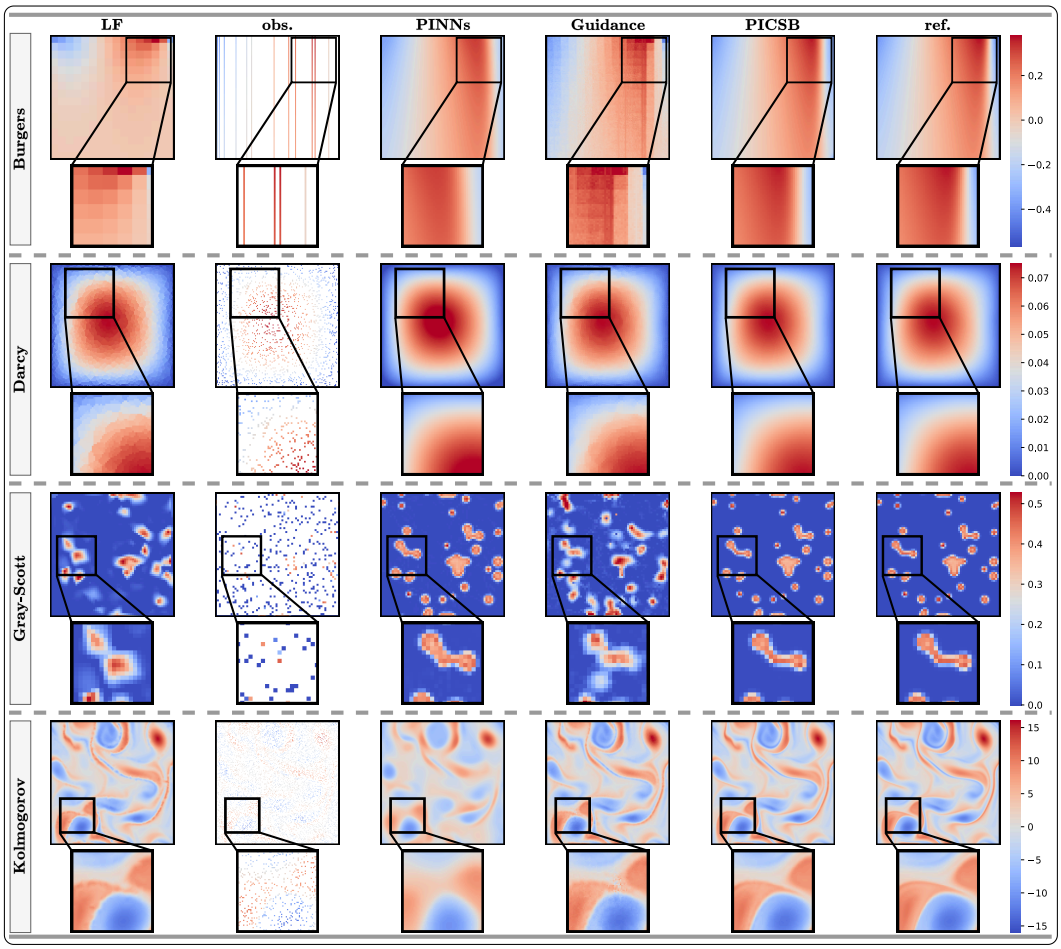

Figure 2: Qualitative comparison of reconstructed fields. We show representative test samples for Burgers and Darcy flow, and snapshots for Gray–Scott and Kolmogorov flow; LF, sparse HF observations (obs.), *PINNs*, *Guidance*, and PICSB are compared with the HF reference (ref.), with zoomed-in regions highlighting local reconstruction quality.

Time Complexity. All wall-times were measured on a single NVIDIA RTX 4090 GPU, except for Gray–Scott, which was evaluated on a single NVIDIA H200 NVL GPU. The largest practical gain of PICSB is inference speed: it produces reconstructions in 0.026–0.109 seconds across benchmarks, compared with $2{,}557$–$9{,}527$ seconds for per-instance *PINNs* optimization and 141–787 seconds for residual-guided diffusion sampling. This orders-of-magnitude reduction enables time-critical DA, where reconstruction must fit within each physical update cycle.

Qualitative visualization. Figure 2 corroborates these results. Compared with the over-smoothed fields from *PINNs* and the LF-biased outputs from *Guidance*, PICSB more faithfully recovers the dominant structures across the benchmarks. It also captures sharp transitions, localized patterns, and vortical structures while reducing interpolation and structural artifacts.

6 Conclusion

We studied sparsely supervised multi-fidelity DA for PDE-governed systems, where full-grid LF fields are available but HF supervision is limited to sparse observations. We proposed PICSB, an auxiliary-energy-driven conditional Schrödinger bridge whose HF endpoint is defined implicitly by observation consistency, PDE residuals, and LF anchoring rather than empirical full-field HF targets, and learned through iterative surrogate-endpoint refresh. Across diverse PDE benchmarks, PICSB achieves strong reconstruction accuracy with sub-second, orders-of-magnitude faster inference than per-instance optimization and inference-time-guided sampling, demonstrating the potential of amortized, guidance-free LF-to-HF transport for time-critical DA.

The present study is limited to settings in which the governing operators are known and sparse HF observations are available for training, with the current implementation restricted to direct coordinate-subset observations. Extending PICSB to settings with unknown or partially known governing operators, nonlinear or indirect observation operators, severe measurement corruption, or data-free training driven solely by the governing PDEs is left for future work.

Acknowledgments and Disclosure of Funding

This research was supported in part by the National Research Foundation of Korea grant funded by the Korean government (MSIT) (Grant numbers RS-2022-NR068758 and RS-2024-00361377), in part by the Korea Evaluation Institute of Industrial Technology grant funded by the Korean government (MOTIR) (Grant number RS-2025-25458052), and in part by the National IT Industry Promotion Agency grant funded by the Korean government (MSIT) (Grant number RS-2026-25621650).

References

[1] Alban Farchi, Marcin Chrust, Marc Bocquet, Patrick Laloyaux, and Massimo Bonavita. Online model error correction with neural networks in the incremental 4d-var framework. *Journal of Advances in Modeling Earth Systems*, 15(9):e2022MS003474, 2023.

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

[3] Michele Buzzicotti. Data reconstruction for complex flows using ai: Recent progress, obstacles, and perspectives. *Europhysics Letters*, 142(2):23001, 2023.

[4] François Rozet and Gilles Louppe. Score-based data assimilation. *Advances in Neural Information Processing Systems*, 36:40521–40541, 2023.

[5] Hao Zhou, Sibo Cheng, and Rossella Arcucci. Multi-fidelity physics constrained neural networks for dynamical systems. *Computer Methods in Applied Mechanics and Engineering*, 420:116758, 2024.

[6] Lei He, Weiqi Qian, Tun Zhao, and Qing Wang. Multi-fidelity aerodynamic data fusion with a deep neural network modeling method. *Entropy*, 22(9):1022, 2020.

[7] Stanley G. Benjamin, Dezsö Dévényi, Stephen S. Weygandt, Kevin J. Brundage, John M. Brown, Georg A. Grell, Dongsoo Kim, Barry E. Schwartz, Tatiana G. Smirnova, Tracy Lorraine Smith, and Geoffrey S. Manikin. An hourly assimilation–forecast cycle: The ruc. *Monthly Weather Review*, 132(2):495 – 518, 2004. doi: 10.1175/1520-0493(2004)132<0495:AHACTR>2.0.CO;2.

[8] Andrew M. Moore, Matthew J. Martin, Santha Akella, Hernan G. Arango, Magdalena Balmaseda, Laurent Bertino, Stefano Ciavatta, Bruce Cornuelle, Jim Cummings, Sergey Frolov, Pierre Lermusiaux, Paolo Oddo, Peter R. Oke, Andrea Storto, Anna Teruzzi, Arthur Vidard, and Anthony T. Weaver. Synthesis of ocean observations using data assimilation for operational, real-time and reanalysis systems: A more complete picture of the state of the ocean. *Frontiers in Marine Science*, Volume 6 - 2019, 2019. ISSN 2296-7745. doi: 10.3389/fmars.2019.00090.

[9] K. Jafarzadegan, P. Abbaszadeh, and H. Moradkhani. Sequential data assimilation for real-time probabilistic flood inundation mapping. *Hydrology and Earth System Sciences*, 25(9): 4995–5011, 2021. doi: 10.5194/hess-25-4995-2021.

[10] Ruijia Niu, Dongxia Wu, Kai Kim, Yi-An Ma, Duncan Watson-Parris, and Rose Yu. Multi-fidelity residual neural processes for scalable surrogate modeling. In *Proceedings of the 41st International Conference on Machine Learning*, pages 38381–38394, 2024.

[11] M. Giselle Fernández-Godino. Review of multi-fidelity models. *Advances in Computational Science and Engineering*, 1(4), 12 2023. doi: 10.3934/acse.2023015.

[12] Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *Siam Review*, 60(3):550–591, 2018.

[13] Jiaqing Kou, Chenjia Ning, and Weiwei Zhang. Transfer learning for flow reconstruction based on multifidelity data. *AIAA Journal*, 60(10):5821–5842, 2022.

[14] Souvik Chakraborty. Transfer learning based multi-fidelity physics informed deep neural network. *Journal of Computational Physics*, 426:109942, 2021.

[15] 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.

[16] Aliaksandra Shysheya, Cristiana Diaconu, Federico Bergamin, Paris Perdikaris, José Miguel Hernández-Lobato, Richard Turner, and Emile Mathieu. On conditional diffusion models for pde simulations. *Advances in Neural Information Processing Systems*, 37:23246–23300, 2024.

[17] Qianxun Xu and Zuchuan Li. Partial physics informed diffusion model for ocean chlorophyll concentration reconstruction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[18] Jiahe Huang, Guandao Yang, Zichen Wang, and Jeong Joon Park. Diffusionpde: Generative pde-solving under partial observation. *Advances in Neural Information Processing Systems*, 37: 130291–130323, 2024.

[19] 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.

[20] Utkarsh Utkarsh, Pengfei Cai, Alan Edelman, Rafael Gomez-Bombarelli, and Christopher Vincent Rackauckas. Physics-constrained flow matching: Sampling generative models with hard constraints. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[21] Pan Du, Meet Hemant Parikh, Xiantao Fan, Xin-Yang Liu, and Jian-Xun Wang. Conditional neural field latent diffusion model for generating spatiotemporal turbulence. *Nature Communications*, 15(1):10416, 2024.

[22] Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A. Theodorou, Weili Nie, and Anima Anandkumar. I^2SB: image-to-image Schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[23] Ziying Li, Xuequan Lu, Xinkui Zhao, Guanjie Cheng, Shuiguang Deng, and Jianwei Yin. Walking the schrödinger bridge: A direct trajectory for text-to-3d generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[24] Tapas Tripura, Akshay Thakur, and Souvik Chakraborty. Multi-fidelity wavelet neural operator surrogate model for time-independent and time-dependent reliability analysis. *Probabilistic Engineering Mechanics*, 77:103672, 2024.

[25] Xingzi Xu, Johann Guilleminot, and Vahid Tarokh. Data generation with optimal experimental design for operator learning. *Computer Methods in Applied Mechanics and Engineering*, 450: 118675, 2026.

[26] Marta D'Elia, Mauro Perego, and Alessandro Veneziani. A variational data assimilation procedure for the incompressible navier-stokes equations in hemodynamics. *Journal of Scientific Computing*, 52(2):340–359, 2012.

[27] Mathias Lemke and Jörn Sesterhenn. Adjoint-based pressure determination from piv data in compressible flows — validation and assessment based on synthetic data. *European Journal of Mechanics - B/Fluids*, 58:29–38, 2016. ISSN 0997-7546. doi: https://doi.org/10.1016/j.euromechflu.2016.03.006.

[28] M. Zauner, V. Mons, O. Marquet, and B. Leclaire. Nudging-based data assimilation of the turbulent flow around a square cylinder. *Journal of Fluid Mechanics*, 937:A38, 2022. doi: 10.1017/jfm.2022.133.

[29] Yuepeng Wang, Jie Li, Wenju Zhao, I.M. Navon, and Guang Lin. Accelerating inverse inference of ensemble kalman filter via reduced-order model trained using adaptive sparse observations. *Journal of Computational Physics*, 496:112600, 2024. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2023.112600.

[30] M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2018.10.045.

[31] Xihaier Luo, Wei Xu, Balu Nadiga, Yihui Ren, and Shinjae Yoo. Continuous field reconstruction from sparse observations with implicit neural networks. In *The Twelfth International Conference on Learning Representations*, 2024.

[32] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

[33] 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. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[34] Woojin Cho, Minju Jo, Haksoo Lim, Kookjin Lee, Dongeun Lee, Sanghyun Hong, and Noseong Park. Parameterized physics-informed neural networks for parameterized PDEs. In *Forty-first International Conference on Machine Learning*, 2024.

[35] 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.

[36] Pavel B Bochev and Max D Gunzburger. *Least-squares finite element methods*. Springer Science & Business Media, 2009.

[37] Yuyang Shi, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. Conditional simulation using diffusion Schrödinger bridges. In James Cussens and Kun Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1792–1802. PMLR, 01–05 Aug 2022.

[38] Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[39] Stefano Peluchetti. Diffusion bridge mixture transports, schrödinger bridge problems and generative modeling. *Journal of Machine Learning Research*, 24(374):1–51, 2023.

[40] Valentin De Bortoli, Iryna Korshunova, Andriy Mnih, and Arnaud Doucet. Schrodinger bridge flow for unpaired data translation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[41] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[42] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, June 2022.

[43] Kirill Tamogashev and Nikolay Malkin. Data-to-energy stochastic dynamics. In *The Fourteenth International Conference on Learning Representations*, 2026.

[44] Byoungwoo Park, Jungwon Choi, Sungbin Lim, and Juho Lee. Stochastic optimal control for diffusion bridges in function spaces. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[45] Peiyan Hu, Haodong Feng, Hongyuan Liu, Tongtong Yan, Wenhao Deng, Tianrun Gao, Rong Zheng, Haoren Zheng, Chenglei Yu, Chuanrui Wang, Kaiwen Li, Zhi-Ming Ma, Dezhi Zhou, Xingcai Lu, Dixia Fan, and Tailin Wu. RealPDEBench: A benchmark for complex physical systems with real-world data. In *The Fourteenth International Conference on Learning Representations*, 2026.

[46] Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam Parker, Miles Cranmer, and Shirley Ho. The well: a large-scale collection of diverse physics simulations for machine learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 44989–45037. Curran Associates, Inc., 2024. doi: 10.52202/079017-1430.

[47] Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. Pdebench: An extensive benchmark for scientific machine learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 1596–1611. Curran Associates, Inc., 2022.

[48] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[49] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022.

A Supplementary for Section 4

A.1 Proof for Theorem 4.1

Theorem 4.1 (Population endpoint selected by PICSB under an auxiliary energy)**.** *Fix (x_0, y). Assume that hard projection is exact, $\mathcal{C}_y$ is nonempty and closed, $\mathcal{R}_h$ is continuous on $\mathcal{C}_y$, and the residual floor $\delta^\star(y; h)$ in Eq.* (8) *is attained. For fixed $\lambda_{\mathrm{PDE}}, \lambda_{\mathrm{LF}}, \beta > 0$, let $\{\tilde{\pi}_1^{(k)}(\cdot|x_0, y)\}_{k\geq 0}$ be the population terminal laws generated by the PICSB refresh map in Eq.* (10)*. If this refresh sequence has vanishing terminal free-energy gap induced by $\mathcal{E}_{\lambda,h}$, then $\tilde{\pi}_1^{(k)}(\cdot|x_0, y)$ converges to $\pi_{1+}^{\lambda,\beta}(\cdot|x_0, y)$ in KL divergence, and hence also in total variation and weakly. Moreover, in the zero-temperature and residual-dominant limit $\beta \to \infty$ and $\lambda_{\mathrm{PDE}}/\lambda_{\mathrm{LF}} \to \infty$, the population PICSB endpoint law concentrates on*

$$\mathcal{X}_{1+}^\star(x_0, y) := \mathrm{argmin}_x \left\{ \|x - x_0\|_2^2 \,:\, x \in \mathcal{P}_{\mathcal{R}_h}^{\delta^\star, y} \right\}.$$

Proof.
Fix (x_0, y), and write

$$F(x) := \|\mathcal{R}_h(x)\|_2^2, \qquad G(x) := \|x - x_0\|_2^2.$$

By exact hard projection, all terminal laws generated by the population sampler are supported on $\mathcal{C}_y$.

Recall the auxiliary energy in Eq. (1):

$$\mathcal{E}_{\lambda,h}(x; x_0, y) = \frac{\lambda_{\mathrm{PDE}}}{2} F(x) + \frac{\lambda_{\mathrm{LF}}}{2} G(x), \qquad x \in \mathcal{C}_y,$$

and

$$\mathcal{E}_{\lambda,h}(x; x_0, y) = +\infty, \qquad x \notin \mathcal{C}_y.$$

Thus every finite-energy endpoint belongs to $\mathcal{C}_y$ and therefore satisfies $\mathcal{H}(x) = y$.

Let λ_y be the Lebesgue measure over the unobserved coordinates with the observed coordinates fixed to y, as in Eq. (2). For a probability measure ν on $\mathcal{C}_y$, write $\nu \ll \lambda_y$ when ν is absolutely continuous with respect to λ_y, i.e., whenever $\lambda_y(A) = 0$ then $\nu(A) = 0$. In this case, the Radon–Nikodym derivative

$$\rho(x) := \frac{d\nu}{d\lambda_y}(x)$$

exists, and we define

$$\mathfrak{F}_{\lambda,\beta}^{x_0,y}(\nu) := \int_{\mathcal{C}_y} \mathcal{E}_{\lambda,h}(x; x_0, y)\, d\nu(x) + \beta^{-1} \int_{\mathcal{C}_y} \log \rho(x)\, d\nu(x).$$

If $\nu \not\ll \lambda_y$, we set

$$\mathfrak{F}_{\lambda,\beta}^{x_0,y}(\nu) = +\infty.$$

The assumption that refresh training is asymptotically optimal for the terminal free energy means

$$\Delta_k := \mathfrak{F}_{\lambda,\beta}^{x_0,y}\left(\tilde{\pi}_1^{(k)}(\cdot|x_0, y)\right) - \inf_{\nu \in \mathcal{P}(\mathcal{C}_y)} \mathfrak{F}_{\lambda,\beta}^{x_0,y}(\nu) \longrightarrow 0.$$

We now identify the minimizer of this free energy. Define the normalizing constant

$$Z_{\lambda,\beta}(x_0, y) := \int_{\mathcal{C}_y} \exp[-\beta \mathcal{E}_{\lambda,h}(x; x_0, y)]\, d\lambda_y(x).$$

Since $\lambda_{\mathrm{LF}} > 0$ and

$$\mathcal{E}_{\lambda,h}(x; x_0, y) \geq \frac{\lambda_{\mathrm{LF}}}{2} \|x - x_0\|_2^2 \quad \text{on } \mathcal{C}_y,$$

the integrand is bounded above by a Gaussian density on the unobserved coordinates. Hence

$$0 < Z_{\lambda,\beta}(x_0, y) < \infty.$$

The desirable endpoint law is therefore

$$\pi_{1+}^{\lambda,\beta}(dx|x_0,y) = \frac{1}{Z_{\lambda,\beta}(x_0,y)} \exp[-\beta \mathcal{E}_{\lambda,h}(x;x_0,y)]\, d\lambda_y(x).$$

Let $\nu \ll \lambda_y$ and let $\rho = d\nu/d\lambda_y$. Then

$$\frac{d\pi_{1+}^{\lambda,\beta}}{d\lambda_y}(x) = Z_{\lambda,\beta}(x_0,y)^{-1} \exp[-\beta \mathcal{E}_{\lambda,h}(x;x_0,y)].$$

Therefore,

$$\begin{aligned}\mathrm{KL}\Big(\nu \,\Big\|\, \pi_{1+}^{\lambda,\beta}(\cdot|x_0,y)\Big) &= \int_{\mathcal{C}_y} \log\left(\frac{d\nu/d\lambda_y}{d\pi_{1+}^{\lambda,\beta}/d\lambda_y}\right) d\nu \\ &= \int_{\mathcal{C}_y} \log \rho(x)\, d\nu + \beta \int_{\mathcal{C}_y} \mathcal{E}_{\lambda,h}(x;x_0,y)\, d\nu + \log Z_{\lambda,\beta}(x_0,y).\end{aligned}$$

Rearranging gives the Gibbs variational identity

$$\mathfrak{F}_{\lambda,\beta}^{x_0,y}(\nu) = -\beta^{-1} \log Z_{\lambda,\beta}(x_0,y) + \beta^{-1}\mathrm{KL}\Big(\nu \,\Big\|\, \pi_{1+}^{\lambda,\beta}(\cdot|x_0,y)\Big).$$

The same identity holds in the extended sense when $\nu \not\ll \lambda_y$, because both the free energy and the KL divergence are then $+\infty$.

Since KL divergence is nonnegative and equals zero only when the two measures are identical, the unique minimizer of $\mathfrak{F}_{\lambda,\beta}^{x_0,y}$ is

$$\pi_{1+}^{\lambda,\beta}(\cdot|x_0,y),$$

and

$$\inf_{\nu \in \mathcal{P}(\mathcal{C}_y)} \mathfrak{F}_{\lambda,\beta}^{x_0,y}(\nu) = -\beta^{-1} \log Z_{\lambda,\beta}(x_0,y).$$

Applying the variational identity to $\nu = \tilde{\pi}_1^{(k)}(\cdot|x_0,y)$ gives

$$\mathrm{KL}\Big(\tilde{\pi}_1^{(k)}(\cdot|x_0,y) \,\Big\|\, \pi_{1+}^{\lambda,\beta}(\cdot|x_0,y)\Big) = \beta\, \Delta_k.$$

Since $\Delta_k \to 0$, we obtain

$$\mathrm{KL}\Big(\tilde{\pi}_1^{(k)}(\cdot|x_0,y) \,\Big\|\, \pi_{1+}^{\lambda,\beta}(\cdot|x_0,y)\Big) \longrightarrow 0.$$

Pinsker's inequality then yields

$$\Big\|\tilde{\pi}_1^{(k)}(\cdot|x_0,y) - \pi_{1+}^{\lambda,\beta}(\cdot|x_0,y)\Big\| \le \sqrt{\frac{1}{2}\mathrm{KL}\Big(\tilde{\pi}_1^{(k)} \,\Big\|\, \pi_{1+}^{\lambda,\beta}\Big)} \longrightarrow 0.$$

Total-variation convergence implies weak convergence. Therefore, under asymptotically optimal refresh training for the terminal free energy, the population terminal laws generated by Algorithm 2 satisfy

$$\mathrm{KL}\Big(\tilde{\pi}_1^{(k)}(\cdot|x_0,y) \,\Big\|\, \pi_{1+}^{\lambda,\beta}(\cdot|x_0,y)\Big) \to 0, \qquad \Big\|\tilde{\pi}_1^{(k)}(\cdot|x_0,y) - \pi_{1+}^{\lambda,\beta}(\cdot|x_0,y)\Big\| \to 0,$$

and hence

$$\tilde{\pi}_1^{(k)}(\cdot|x_0,y) \Rightarrow \pi_{1+}^{\lambda,\beta}(\cdot|x_0,y).$$

It remains to characterize the zero-temperature and residual-dominant limit. Let

$$r := \frac{\lambda_{\mathrm{PDE}}}{\lambda_{\mathrm{LF}}}, \qquad a_\beta := \frac{\beta \lambda_{\mathrm{LF}}}{2}.$$

On $\mathcal{C}_y$, the Gibbs density is proportional to

$$\exp[-a_\beta\{rF(x) + G(x)\}].$$

Let

$$\delta^\star := \delta^\star(y; h) = \min_{x \in \mathcal{C}_y} \|\mathcal{R}_h(x)\|_2.$$

By the definition of the feasible physics set in Eq. (8),

$$\mathcal{P}_{\mathcal{R}_h}^{\delta^\star, y} = \{x \in \mathcal{C}_y : \|\mathcal{R}_h(x)\|_2 \le \delta^\star\}.$$

Since $\delta^\star$ is the minimum residual over $\mathcal{C}_y$, this set is equivalently

$$\mathcal{P}_{\mathcal{R}_h}^{\delta^\star, y} = \left\{x \in \mathcal{C}_y : F(x) = (\delta^\star)^2\right\}.$$

It is nonempty by the attainment assumption, and it is closed because $\mathcal{C}_y$ is closed and F is continuous on $\mathcal{C}_y$. Define

$$g^\star := \min_{x \in \mathcal{P}_{\mathcal{R}_h}^{\delta^\star, y}} G(x).$$

This minimum is attained. Indeed, choose any $x_p \in \mathcal{P}_{\mathcal{R}_h}^{\delta^\star, y}$. It is enough to minimize G over

$$\mathcal{P}_{\mathcal{R}_h}^{\delta^\star, y} \cap \{x : G(x) \le G(x_p)\},$$

which is nonempty, closed, and bounded in finite dimension, hence compact. Therefore

$$\mathcal{X}_{1+}^\star(x_0, y) = \left\{x \in \mathcal{P}_{\mathcal{R}_h}^{\delta^\star, y} : G(x) = g^\star\right\}$$

is nonempty and compact.

For fixed $r > 0$, subtracting the constant $r(\delta^\star)^2$ from the energy does not change the Gibbs law. Define

$$\Phi_r(x) := r\big(F(x) - (\delta^\star)^2\big) + G(x), \qquad x \in \mathcal{C}_y.$$

Then the Gibbs law is proportional to

$$\exp[-a_\beta \Phi_r(x)]\, d\lambda_y(x).$$

Because $F(x) \ge (\delta^\star)^2$ on $\mathcal{C}_y$,

$$\Phi_r(x) \ge G(x).$$

Thus Φ_r is coercive on $\mathcal{C}_y$. Since F and G are continuous, Φ_r is continuous. Hence

$$S_r := \operatorname{argmin}_{x \in \mathcal{C}_y} \Phi_r(x)$$

is nonempty and compact.

We next show zero-temperature concentration for fixed r. Let U be any open neighborhood of S_r in the relative topology of $\mathcal{C}_y$, and let

$$m_r := \min_{x \in \mathcal{C}_y} \Phi_r(x).$$

Because Φ_r is continuous and coercive, its sublevel sets are compact. Therefore

$$\eta_U := \inf_{x \in \mathcal{C}_y \setminus U} \Phi_r(x) - m_r$$

is strictly positive.

Choose $x_r^\star \in S_r$. By continuity of Φ_r, there is a relative ball $B \subset \mathcal{C}_y$ around $x_r^\star$ such that

$$\lambda_y(B) > 0, \qquad \Phi_r(x) \le m_r + \eta_U/4 \quad \text{for } x \in B.$$

Hence the normalizing constant satisfies

$$Z_{r,\beta} := \int_{\mathcal{C}_y} e^{-a_\beta \Phi_r(x)}\, d\lambda_y(x) \ge \lambda_y(B) e^{-a_\beta (m_r + \eta_U/4)}.$$

On the other hand, for $x \in \mathcal{C}_y \setminus U$,

$$\Phi_r(x) \ge m_r + \eta_U.$$

Since Φ_r is coercive,

$$C_r := \int_{\mathcal{C}_y} \exp\left[-\frac{1}{2}(\Phi_r(x) - m_r)\right] d\lambda_y(x) < \infty.$$

For $a_\beta \geq 1$,

$$\begin{aligned}\int_{\mathcal{C}_y \setminus U} e^{-a_\beta \Phi_r(x)}\, d\lambda_y(x) &= e^{-a_\beta m_r} \int_{\mathcal{C}_y \setminus U} e^{-a_\beta(\Phi_r(x)-m_r)}\, d\lambda_y(x) \\ &\leq e^{-a_\beta m_r} e^{-a_\beta \eta_U/2} \int_{\mathcal{C}_y} e^{-\frac{1}{2}(\Phi_r(x)-m_r)}\, d\lambda_y(x) \\ &= C_r e^{-a_\beta(m_r+\eta_U/2)}.\end{aligned}$$

Dividing this upper bound by the lower bound on $Z_{r,\beta}$ gives

$$\pi_{1+}^{\lambda,\beta}(\mathcal{C}_y \setminus U | x_0, y) \leq \frac{C_r}{\lambda_y(B)} \exp[-a_\beta \eta_U/4] \longrightarrow 0 \qquad \text{as } \beta \to \infty.$$

Therefore, for fixed r,

$$\lim_{\beta\to\infty} \pi_{1+}^{\lambda,\beta}(U|x_0, y) = 1$$

for every open neighborhood U of S_r.

We now let $r \to \infty$. Take any $x_r \in S_r$ and any $x^\star \in \mathcal{X}_{1+}^\star(x_0, y)$. Since x_r minimizes Φ_r,

$$\Phi_r(x_r) \leq \Phi_r(x^\star).$$

Because $x^\star \in \mathcal{P}_{\mathcal{R}_h}^{\delta^\star, y}$,

$$F(x^\star) = (\delta^\star)^2.$$

Thus

$$\Phi_r(x^\star) = G(x^\star) = g^\star.$$

Therefore

$$r\big(F(x_r) - (\delta^\star)^2\big) + G(x_r) \leq g^\star.$$

Both terms on the left-hand side are nonnegative, so

$$G(x_r) \leq g^\star, \qquad 0 \leq F(x_r) - (\delta^\star)^2 \leq \frac{g^\star}{r}.$$

The bound $G(x_r) \leq g^\star$ places all such x_r in the compact sublevel set

$$\{x \in \mathcal{C}_y : G(x) \leq g^\star\}.$$

Hence, for any sequence $r_\ell \to \infty$, there exists a subsequence, still denoted r_ℓ, such that

$$x_{r_\ell} \to \bar{x} \in \mathcal{C}_y.$$

By continuity of F,

$$F(\bar{x}) = \lim_{\ell\to\infty} F(x_{r_\ell}) = (\delta^\star)^2.$$

Hence

$$\bar{x} \in \mathcal{P}_{\mathcal{R}_h}^{\delta^\star, y}.$$

By continuity of G,

$$G(\bar{x}) = \lim_{\ell\to\infty} G(x_{r_\ell}) \leq g^\star.$$

Since $g^\star$ is the minimum of G over $\mathcal{P}_{\mathcal{R}_h}^{\delta^\star, y}$, we must have

$$G(\bar{x}) = g^\star.$$

Thus

$$\bar{x} \in \mathcal{X}_{1+}^\star(x_0, y).$$

So every cluster point of minimizers $x_r \in S_r$ lies in $\mathcal{X}_{1+}^\star(x_0, y)$.

Let U be any open neighborhood of $\mathcal{X}_{1+}^\star(x_0, y)$. If S_r were not eventually contained in U, then there would exist $r_\ell \to \infty$ and $x_{r_\ell} \in S_{r_\ell} \setminus U$. By the compactness argument above, a subsequence would converge to a point in $\mathcal{X}_{1+}^\star(x_0, y) \subset U$, contradicting $x_{r_\ell} \notin U$. Hence there exists $r_0 > 0$ such that

$$S_r \subset U, \qquad r \geq r_0.$$

For every $r \geq r_0$, the fixed r zero-temperature concentration result gives

$$\lim_{\beta\to\infty} \pi_{1+}^{\lambda,\beta}(U|x_0, y) = 1.$$

Therefore

$$\lim_{r\to\infty} \lim_{\beta\to\infty} \pi_{1+}^{\lambda,\beta}(U|x_0, y) = 1.$$

Finally, for fixed $(\lambda_{\mathrm{PDE}}, \lambda_{\mathrm{LF}}, \beta)$, we already proved

$$\tilde{\pi}_1^{(k)}(\cdot|x_0, y) \to \pi_{1+}^{\lambda,\beta}(\cdot|x_0, y)$$

in total variation. Hence, for every measurable set U,

$$\left|\tilde{\pi}_1^{(k)}(U|x_0, y) - \pi_{1+}^{\lambda,\beta}(U|x_0, y)\right| \leq \left\|\tilde{\pi}_1^{(k)}(\cdot|x_0, y) - \pi_{1+}^{\lambda,\beta}(\cdot|x_0, y)\right\| \to 0.$$

Combining this with the concentration of $\pi_{1+}^{\lambda,\beta}$, we obtain, for every open neighborhood U of $\mathcal{X}_{1+}^{\star}(x_0, y)$,

$$\lim_{r\to\infty} \lim_{\beta\to\infty} \lim_{k\to\infty} \tilde{\pi}_1^{(k)}(U|x_0, y) = 1.$$

This is the precise meaning of concentration on

$$\mathcal{X}_{1+}^{\star}(x_0, y) = \operatorname*{argmin}_{x\in\mathcal{P}_{\mathcal{R}_h}^{\delta^\star,y}} \|x - x_0\|_2^2.$$

Since $\mathcal{P}_{\mathcal{R}_h}^{\delta^\star,y} \subset \mathcal{C}_y$, every point in the limiting support satisfies $\mathcal{H}(x) = y$, giving zero observation error. Since points in $\mathcal{P}_{\mathcal{R}_h}^{\delta^\star,y}$ attain the residual floor, the PDE residual is minimized over $\mathcal{C}_y$. Finally, the last minimization over $G(x) = \|x - x_0\|_2^2$ shows that, when multiple observation-consistent residual minimizers exist, the limiting endpoint selects the one closest to the LF input x_0. □

A.2 Additional Theoretical Notes — Stability under identifiability

Lemma A.1 (Existence of a localization gap)**.** *Fix y and define $f(x) := \|\mathcal{R}_h(x)\|$. Let $\mathcal{K} \subseteq \mathcal{C}_y$ be a compact set and assume that f is continuous on $\mathcal{K}$ and admits a unique minimizer: there exists a unique $x_1^\star \in \mathcal{K}$ such that*

$$f(x_1^\star) = \min_{x \in \mathcal{K}} f(x) = \delta^\star_{\mathcal{K}}(y; h).$$

For any $\rho \in \big(0, \sup_{z \in \mathcal{K}} \|z - x_1^\star\|\big)$, define the neighborhood

$$\mathcal{U}_\rho := \{x \in \mathcal{K} : \|x - x_1^\star\| \le \rho\}.$$

Then the quantity

$$\bar{\beta}(\rho) \;:=\; \min_{x \in \mathcal{K} \setminus \mathcal{U}_\rho} \Big(f(x) - \delta^\star_{\mathcal{K}}(y; h)\Big)$$

is well-defined and satisfies $\bar{\beta}(\rho) > 0$. Consequently, for any $0 < \beta \le \bar{\beta}(\rho)$,

$$x \in \mathcal{K} \setminus \mathcal{U}_\rho \quad \Longrightarrow \quad f(x) \ge \delta^\star_{\mathcal{K}}(y; h) + \beta.$$

Proof.
Since $\mathcal{K}$ is compact and f is continuous, f attains its minimum on $\mathcal{K}$ by the Weierstrass extreme value theorem, and by assumption the minimizer is unique at $x_1^\star$.

The set $\mathcal{K} \setminus \mathcal{U}_\rho$ is compact. Indeed, the set $\{x \in \mathbb{R}^n : \|x - x_1^\star\| \ge \rho\}$ is closed, and hence

$$\mathcal{K} \setminus \mathcal{U}_\rho = \mathcal{K} \cap \{x : \|x - x_1^\star\| \ge \rho\}$$

is a closed subset of the compact set $\mathcal{K}$, and therefore compact.

Since $f(x) \ge \delta^\star_{\mathcal{K}}(y; h)$ for all $x \in \mathcal{K}$, we have $f(x) - \delta^\star_{\mathcal{K}}(y; h) \ge 0$ on $\mathcal{K} \setminus \mathcal{U}_\rho$. Thus the continuous function $x \mapsto f(x) - \delta^\star_{\mathcal{K}}(y; h)$ attains its minimum on $\mathcal{K} \setminus \mathcal{U}_\rho$, so $\bar{\beta}(\rho)$ is well-defined and satisfies $\bar{\beta}(\rho) \ge 0$.

To show $\bar{\beta}(\rho) > 0$, assume for contradiction that $\bar{\beta}(\rho) = 0$. Then by definition of the minimum, there exists $\bar{x} \in \mathcal{K} \setminus \mathcal{U}_\rho$ such that $f(\bar{x}) = \delta^\star_{\mathcal{K}}(y; h)$. Since $\bar{x} \notin \mathcal{U}_\rho$, we have $\bar{x} \ne x_1^\star$, which contradicts the uniqueness of the minimizer. Hence $\bar{\beta}(\rho) > 0$. □

Proposition A.2 (Stability under identifiability)**.** *Fix y and assume the continuous PDE admits a unique observation-consistent solution $x_1^\star \in \mathcal{C}_y$. Assume that local identifiability holds in a neighborhood of $x_1^\star$, with constants $c > 0$ and $\eta(h)$, and that the residual sublevel sets are localized with constant $\beta > 0$. Then any $\hat{x}_1 \in \mathcal{C}_y$ satisfying $\|\mathcal{R}_h(\hat{x}_1)\| \le \delta$ for some $\delta \le \delta^\star(y; h) + \beta$ obeys*

$$\|\hat{x}_1 - x_1^\star\| \le c(\delta - \delta^\star(y; h)) + \eta(h). \tag{11}$$

Proof.
Fix y and let $x_1^\star \in \mathcal{C}_y$ denote the unique observation-consistent solution. Recall the residual floor on $\mathcal{C}_y$,

$$\delta^\star(y; h) \;:=\; \min_{x \in \mathcal{C}_y} \|\mathcal{R}_h(x)\|.$$

We now make the assumptions in the statement precise.

Assumption A.3 (Local identifiability)**.** There exist $\rho > 0$, a constant $c > 0$, and a function $\eta(h) \ge 0$ with $\eta(h) \to 0$ as $h \to 0$ such that, for the neighborhood

$$\mathcal{U} \;:=\; \{x \in \mathbb{R}^n : \|x - x_1^\star\| \le \rho\},$$

the following holds for all $x \in \mathcal{U} \cap \mathcal{C}_y$:

$$\|x - x_1^\star\| \;\le\; c\big(\|\mathcal{R}_h(x)\| - \delta^\star(y; h)\big) + \eta(h). \tag{12}$$

Deriving a localization constant on a compact relevant set. Let $\mathcal{K} \subseteq \mathcal{C}_y$ be a compact set that contains $\hat{x}_1$, and define

$$\delta^\star_{\mathcal{K}}(y;h) \;:=\; \min_{x\in\mathcal{K}} \|\mathcal{R}_h(x)\|.$$

Assume that $x_1^\star$ is the unique minimizer of $f(x) := \|\mathcal{R}_h(x)\|$ over $\mathcal{K}$. For the same $\rho > 0$ as above, Lemma A.1 yields a constant

$$\beta \;:=\; \bar{\beta}(\rho) \;>\; 0 \quad \text{such that} \quad \inf_{x\in\mathcal{K}\setminus\mathcal{U}} \|\mathcal{R}_h(x)\| \;\geq\; \delta^\star_{\mathcal{K}}(y;h) + \beta.$$

Equivalently, we may record this as the following localization inequality:

$$\inf_{x\in\mathcal{K}\setminus\mathcal{U}} \|\mathcal{R}_h(x)\| \;\geq\; \delta^\star_{\mathcal{K}}(y;h) + \beta. \tag{13}$$

Now let $\hat{x}_1 \in \mathcal{K}$ satisfy $\|\mathcal{R}_h(\hat{x}_1)\| \leq \delta$ for some

$$\delta \;\leq\; \delta^\star_{\mathcal{K}}(y;h) + \beta.$$

We first show that $\hat{x}_1 \in \mathcal{U}$. If $\hat{x}_1 \notin \mathcal{U}$, then $\hat{x}_1 \in \mathcal{K} \setminus \mathcal{U}$ and Eq. (13) gives

$$\|\mathcal{R}_h(\hat{x}_1)\| \;\geq\; \inf_{x\in\mathcal{K}\setminus\mathcal{U}} \|\mathcal{R}_h(x)\| \;\geq\; \delta^\star_{\mathcal{K}}(y;h) + \beta,$$

contradicting $\|\mathcal{R}_h(\hat{x}_1)\| \leq \delta \leq \delta^\star_{\mathcal{K}}(y;h) + \beta$. Therefore $\hat{x}_1 \in \mathcal{U}$.

Since $\hat{x}_1 \in \mathcal{U} \cap \mathcal{C}_y$, we may apply Eq. (12) with $x = \hat{x}_1$:

$$\|\hat{x}_1 - x_1^\star\| \;\leq\; c\big(\|\mathcal{R}_h(\hat{x}_1)\| - \delta^\star(y;h)\big) + \eta(h).$$

Using $\|\mathcal{R}_h(\hat{x}_1)\| \leq \delta$ yields

$$\|\hat{x}_1 - x_1^\star\| \;\leq\; c(\delta - \delta^\star(y;h)) + \eta(h),$$

which proves the claim. □

The intuition regarding the upper bound on the difference between $\hat{x}_1$ and x_1^* is as follows. First, the upper bound tends to increase as the fidelity of the target PDE becomes higher (HF). Conversely, the bound decreases as the spatial resolution is refined or the time interval narrows. We provide a theoretical verification of these intuitions specifically within the context of the Burgers' equation.

Proposition A.4 (Error Bound Stability of 1D Burgers' Equation)**.** *Suppose the observation constraint set $\mathcal{C}_y$ ensures the unique existence of a solution. Let x^* be the true state and $\hat{x}$ be the approximate state. Let $\delta'' = \|\mathcal{R}(\hat{x}) - \mathcal{R}(x^*)\|$ denote the residual difference. Then, there exists a state error bound function $\epsilon(\delta'', PDE_{fidelity}, h)$ such that:*

$$\|\hat{x} - x^*\| < \epsilon(\delta'', PDE_{fidelity}, h) \tag{14}$$

where h is the scaling of grid. The error bound ϵ satisfies:

- $\frac{\partial \epsilon}{\partial \delta''} > 0$*: The bound is monotonically increasing with respect to the residual difference.*
- ϵ *increases as the PDE exhibits higher frequency component*
- $\frac{\partial \epsilon}{\partial h} < 0$*: As the grid resolution increases ($h \to 0$), the bound ϵ tends to decrease.*

Proof.
To establish the stability-based error bound for the 1D Burgers' equation, we analyze the relationship between the state error and the residual difference through the following four steps.

Residual and Jacobian Formulation We consider the discretized residual operator $\mathcal{R}(x)$ on a spatio-temporal grid (i, j):

$$\mathcal{R}(x_{i,j}) = \frac{x_{i,j} - x_{i,j-1}}{\Delta t} + x_{i,j}\frac{x_{i+1,j} - x_{i-1,j}}{2h} - \nu\frac{x_{i+1,j} - 2x_{i,j} + x_{i-1,j}}{h^2} \tag{15}$$

Linearizing $\mathcal{R}$ around the reference state yields the Jacobian matrix $J = \frac{1}{\Delta t}I + A + D$, where D is the symmetric diffusion matrix with $D_{i,i} = \frac{2\nu}{h^2}$ and $D_{i,i\pm1} = -\frac{\nu}{h^2}$, and A is the non-symmetric advection matrix capturing local transport.

Spectral Analysis of Spatial Operators We estimate the lower bound of the minimum eigenvalue of the spatial operator $M = A + D$ using the Gershgorin Circle Theorem. For each row i, the Gershgorin disc is centered at $M_{i,i} = \frac{2\nu}{h^2} + \frac{x_{i+1}-x_{i-1}}{2h}$ with radius $R_i = |-\frac{\nu}{h^2} + \frac{x_i}{2h}| + |-\frac{\nu}{h^2} - \frac{x_i}{2h}|$. Under the refinement condition $h < 2\nu/|x_i|$, the radius reduces to $R_i = 2\nu/h^2$. By incorporating the global spectral property of the Laplacian $\nu(\frac{\pi}{L})^2$, we obtain a refined lower bound:

$$\lambda_{\min}(A + D) \approx \nu \left(\frac{\pi}{L}\right)^2 + \min_i \left(\frac{x_{i+1,j} - x_{i-1,j}}{2h}\right) \tag{16}$$

Norm Inequality and Stability Bound Let $e = \hat{x} - x^*$ be the state error. Applying a first-order Taylor expansion, the residual difference $\delta'' = \|\mathcal{R}(\hat{x}) - \mathcal{R}(x^*)\|$ is approximated by $\|J(x^*) \cdot e\|$. Using the matrix norm property $\|Je\| \geq \sigma_{\min}(J)\|e\|$ and approximating the minimum singular value as $\sigma_{\min}(J) \approx \frac{1}{\Delta t} + \lambda_{\min}(A + D)$, we derive the stability bound.

Final Error Bound and Interpretation Combining the above results, we arrive at the final upper bound for the state error:

$$\|\hat{x} - x^*\| \leq \frac{\delta''}{\frac{1}{\Delta t} + \nu \left(\frac{\pi}{L}\right)^2 + \min_i \left(\frac{x_{i+1,j} - x_{i-1,j}}{2h}\right)} \tag{17}$$

This bound validates our intuition: as $\nu \to 0$ (High Fidelity), the diffusive term vanishes and the bound increases. Conversely, as $\Delta t \to 0$ (Higher Resolution), the temporal term $\frac{1}{\Delta t}$ dominates, effectively suppressing the state error. □

A.3 Additional Theoretical Notes — Self-consistency at a fixed point

Lemma A.5 (Hard observation feasibility)**.** *Suppose* Sample_θ *applies the coordinate projection* $\Pi_{\mathcal{C}_y}$ *at every discretized step. Then its terminal output* $x_1 = \mathrm{Sample}_\theta(x_0 \mid y, 0)$ *satisfies* $x_1 \in \mathcal{C}_y$*, i.e.,* $\mathcal{H}(x_1) = y$ *(if* Sample_θ *is randomized).*

Proof. At each discretized step, the projection $\Pi_{\mathcal{C}_y}$ overwrites the observed coordinates so that the state lies in $\mathcal{C}_y = \{x : \mathcal{H}(x) = y\}$. By induction over the sampling steps, all intermediate states, and in particular the terminal state, belong to $\mathcal{C}_y$. Hence $\mathcal{H}(x_1) = y$. □

Lemma A.6 (Existence of an inner-loop minimizer)**.** *Fix* y *and a reference law* $\tilde{\pi}(\cdot \mid y)$*. Assume* $J_{\tilde{\pi}} : \Theta \to \mathbb{R}$ *is lower semi-continuous and bounded below, where*

$$J_{\tilde{\pi}}(\theta) := \mathbb{E}\big[\|\mathcal{R}_h(\hat{x}_1(\theta))\|\big], \qquad \hat{x}_1(\theta) = \mathrm{Sample}_\theta(\tilde{x}_\tau \mid y, \tau).$$

Assume moreover that there exists $\theta_0 \in \Theta$ *such that the sublevel set*

$$\mathcal{S} := \{\theta \in \Theta : \ J_{\tilde{\pi}}(\theta) \le J_{\tilde{\pi}}(\theta_0)\}$$

is nonempty and compact. Then $\mathrm{argmin}_{\theta\in\Theta} J_{\tilde{\pi}}(\theta)$ *is nonempty.*

Proof. We first restate the assumptions in a form used below.

Lower semi-continuity: for every $\theta \in \Theta$ and every sequence $\{\theta_k\} \subset \Theta$ with $\theta_k \to \theta$, we have

$$J_{\tilde{\pi}}(\theta) \le \liminf_{k\to\infty} J_{\tilde{\pi}}(\theta_k). \tag{18}$$

Bounded below: there exists $L \in \mathbb{R}$ such that

$$J_{\tilde{\pi}}(\theta) \ge L \qquad \text{for all } \theta \in \Theta. \tag{19}$$

Hence $m := \inf_{\theta\in\Theta} J_{\tilde{\pi}}(\theta)$ is finite.

By definition of m, there exists a sequence $\{\theta_n\} \subset \Theta$ such that

$$J_{\tilde{\pi}}(\theta_n) \to m \qquad \text{as } n \to \infty. \tag{20}$$

Since $J_{\tilde{\pi}}(\theta_0) \ge m$, the convergence Eq. (20) implies that there exists $N \in \mathbb{N}$ such that for all $n \ge N$,

$$J_{\tilde{\pi}}(\theta_n) \le J_{\tilde{\pi}}(\theta_0),$$

hence $\theta_n \in \mathcal{S}$ for all $n \ge N$ by definition of $\mathcal{S}$.

Because $\mathcal{S}$ is compact, the tail sequence $\{\theta_n\}_{n\ge N} \subset \mathcal{S}$ admits a convergent subsequence: there exist indices $n_k \ge N$ and $\theta^\star \in \mathcal{S}$ such that

$$\theta_{n_k} \to \theta^\star \qquad \text{as } k \to \infty.$$

Applying lower semi-continuity Eq. (18) to the convergent subsequence gives

$$J_{\tilde{\pi}}(\theta^\star) \le \liminf_{k\to\infty} J_{\tilde{\pi}}(\theta_{n_k}).$$

Since $J_{\tilde{\pi}}(\theta_n) \to m$ along the full sequence, the subsequence also satisfies $J_{\tilde{\pi}}(\theta_{n_k}) \to m$, hence $\liminf_{k\to\infty} J_{\tilde{\pi}}(\theta_{n_k}) = m$. Therefore,

$$J_{\tilde{\pi}}(\theta^\star) \le m.$$

On the other hand, by definition of m as the infimum over Θ, we have $m \le J_{\tilde{\pi}}(\theta^\star)$. Combining the two inequalities yields $J_{\tilde{\pi}}(\theta^\star) = m$.

Thus $\theta^\star$ attains the minimum of $J_{\tilde{\pi}}$ over Θ, i.e., $\theta^\star \in \mathrm{argmin}_{\theta\in\Theta} J_{\tilde{\pi}}(\theta)$, proving that $\mathrm{argmin}_{\theta\in\Theta} J_{\tilde{\pi}}(\theta) \ne \emptyset$. □

Proposition A.7 (Self-consistency at a fixed point)**.** *Consider the sequence $\{\tilde{\pi}_1^{(k)}\}_{k\geq 0}$ generated by Algorithm 2. Assume $\mathcal{R}_h$ is Lipschitz on the relevant support and the model class for v_θ is sufficiently expressive. Assume further that each refresh round performs exact inner-loop optimization. If $\tilde{\pi}_1^{(k)} \equiv \tilde{\pi}_1^{(k+1)} \equiv \tilde{\pi}_1^{\infty}$, then samples $\tilde{x}_1 \sim \tilde{\pi}_1^{\infty}(\cdot \mid y)$ satisfy*

$$\mathcal{H}(\tilde{x}_1) = y \qquad \textit{and} \qquad \mathbb{E}\Big[\|\mathcal{R}_h(\tilde{x}_1)\|\Big] = \inf_{\theta} \mathbb{E}\Big[\|\mathcal{R}_h(\hat{x}_1)\|\Big], \tag{21}$$

where $\hat{x}_1 = \mathrm{Sample}_\theta(\tilde{x}_\tau \mid y, \tau)$ is defined using the fixed reference $\tilde{\pi}_1^{\infty}(\cdot \mid y)$ as in Algorithm 2, with expectation over the sampling randomness.

Proof.
At refresh round $k+1$, the surrogate terminal law $\tilde{\pi}_1^{(k+1)}(\cdot \mid y)$ is, by definition, the distribution of $\tilde{x}_1 = \mathrm{Sample}_{\tilde{\theta}}(x_0 \mid y, 0)$ under hard conditioning (Algorithm 2). Lemma A.5 therefore implies $\mathcal{H}(\tilde{x}_1) = y$ for $\tilde{x}_1 \sim \tilde{\pi}_1^{(k+1)}(\cdot \mid y)$. If $\tilde{\pi}_1^{(k+1)} \equiv \tilde{\pi}_1^{\infty}$, this gives the first claim in Eq. (21).

Fix the reference law $\tilde{\pi}_1^{\infty}(\cdot \mid y)$ and define the fixed-reference objective

$$J_\infty(\theta) \;:=\; J_{\tilde{\pi}_1^{\infty}}(\theta) \;=\; \mathbb{E}\Big[\|\mathcal{R}_h(\hat{x}_1)\|\Big], \qquad \hat{x}_1 = \mathrm{Sample}_\theta(\tilde{x}_\tau \mid y, \tau),$$

where the expectation is taken over all randomness in the sampling procedure under the fixed reference $\tilde{x}_1 \sim \tilde{\pi}_1^{\infty}(\cdot \mid y)$.

By Lemma A.6 (applied with $\tilde{\pi} = \tilde{\pi}_1^{\infty}$), the set $\mathrm{argmin}_{\theta\in\Theta} J_\infty(\theta)$ is nonempty; let $\theta^\star \in \mathrm{argmin}_{\theta\in\Theta} J_\infty(\theta)$. This ensures that, for the fixed reference law $\tilde{\pi}_1^{\infty}(\cdot \mid y)$, the refresh step of Algorithm 2 is well-defined and induces a deterministic update of the surrogate terminal law. A fixed point corresponds to a reference law that is invariant under this update.

Now suppose we are at a fixed point $\tilde{\pi}_1^{(k)} \equiv \tilde{\pi}_1^{(k+1)} \equiv \tilde{\pi}_1^{\infty}$. Then the round-$k$ objective coincides with J_∞ by construction of Algorithm 2. By the exact inner-loop optimization assumption, the update selects a global minimizer of the round-k objective, hence we may take the refreshed parameters $\theta^{(k+1)}$ to satisfy

$$\theta^{(k+1)} \in \mathrm{argmin}_{\theta\in\Theta} J_\infty(\theta).$$

Therefore,

$$\mathbb{E}\Big[\|\mathcal{R}_h(\tilde{x}_1)\|\Big] \;=\; J_\infty(\theta^{(k+1)}) \;=\; \inf_{\theta\in\Theta} J_\infty(\theta) \;=\; \inf_{\theta\in\Theta} \mathbb{E}\Big[\|\mathcal{R}_h(\hat{x}_1)\|\Big],$$

which is exactly the second claim in Eq. (21). □

B Experimental Setup

B.1 Datasets Description

PDEs are defined over a spatial coordinate ξ and a physical time variable γ. For dynamic systems, ξ denotes the spatial coordinate and γ denotes physical time. For steady-state problems, only the spatial coordinate ξ is used. Across benchmarks, the HF reference denotes the target field to be reconstructed: the scalar solution field u for Burgers, Darcy, and 1D Diffusion–Reaction, the two-channel concentration field (A, B) for Gray–Scott, and the vorticity field ω for Kolmogorov.

Dynamic 1D Burgers' Equation. We consider the one-dimensional Burgers' equation on $\xi \in (0, 1)$ over the time horizon $\gamma \in (0, 1]$,

$$\partial_\gamma u(\xi, \gamma) + u(\xi, \gamma)\nabla u(\xi, \gamma) = \nu\, \nabla^2 u(\xi, \gamma),$$

with initial condition $u(\xi, 0) = u_0(\xi)$.

We adapt the Burgers data-generation code used in DiffusionPDE [18], with the solver configurations modified to construct the HF and LF settings described below. For the main Burgers benchmark, we construct the LF input using an independent numerical simulation rather than an observation-derived interpolation proxy. The HF and LF simulations are initialized from the same analytical initial condition, discretized on their respective grids, and subsequently evolved independently. To introduce a controlled cross-fidelity discrepancy, we vary both the numerical resolution and the physical parameter governing diffusion. The HF reference is generated with viscosity $\nu_{\mathrm{HF}} = 10^{-2}$ on a 128×128 space–time grid, whereas the LF trajectories are simulated with the larger viscosity $\nu_{\mathrm{LF}} = 10^{-1}$ on coarser $N \times N$ space–time grids, with $N \in \{16, 32\}$. The results reported in the main text correspond to $N = 32$, while $N = 16$ is reported in the supplementary experiments. Here, N denotes the resolution along both the spatial and temporal dimensions. The larger LF viscosity produces smoother trajectories, while the lower HF viscosity yields sharper gradients, thereby inducing a systematic discrepancy in the solution dynamics in addition to the coarse-grid discretization error.

Before being provided to the model, each LF trajectory is upsampled to the 128×128 resolution of the HF reference. Importantly, the LF field is obtained entirely from its own numerical simulation and does not use sparse HF measurements during its construction. Sparse HF observations are sampled separately from the HF reference and are used only for sparse training supervision and inference-time conditioning. This setting therefore evaluates PICSB with a simulator-based LF prior that differs from the HF target in both discretization resolution and physical fidelity.

For the additional Burgers experiments, following [18], HF trajectories are simulated on a 512×512 space–time grid with $\nu_{\mathrm{HF}} = 0.01$ and downsampled to 128×128, while LF inputs are generated from the same initial condition using $\nu_{\mathrm{LF}} = 0.1$. These experiments are reported in Appendix C.3

2D Darcy Flow. We consider the steady-state Darcy equation on $\xi \in (0, 1)^2$:

$$-\nabla \cdot \big(a(\xi)\, \nabla u(\xi)\big) = f(\xi),$$

where $u(\xi)$ denotes the pressure field and $a(\xi)$ is a spatially varying permeability coefficient. We assume a known governing operator with Dirichlet boundary condition $u(\xi) = 0$ on $\partial(0, 1)^2$ and constant forcing $f(\xi) = 1$. The permeability field $a(\xi)$ is provided for each sample, and all fields are represented on a 128×128 grid following the setup used in prior work [18]. HF references correspond to the PDE solution $u(\xi)$ for the given $a(\xi)$. LF inputs are constructed by applying nearest-neighbor interpolation to the sparse HF measurements, yielding a coarse proxy field.

(2+1)D Gray–Scott Reaction–Diffusion. We use the (2+1)D Gray–Scott dataset introduced in The Well [46]. Here, $(2+1)$D indicates two spatial dimensions (ξ_1, ξ_2) together with one temporal dimension γ. We consider the Gray–Scott system on a periodic spatial domain $\xi = (\xi_1, \xi_2) \in [-1, 1]^2$ over the time horizon $\gamma \in (0, 10000]$:

$$\partial_\gamma A(\xi,\gamma) = \delta_A \nabla^2 A(\xi,\gamma) - A(\xi,\gamma)B(\xi,\gamma)^2 + f\big(1 - A(\xi,\gamma)\big),$$
$$\partial_\gamma B(\xi,\gamma) = \delta_B \nabla^2 B(\xi,\gamma) + A(\xi,\gamma)B(\xi,\gamma)^2 - (f+k)B(\xi,\gamma),$$

where $A(\xi,\gamma)$ and $B(\xi,\gamma)$ denote the scalar concentration fields of two chemical species, and f and k are the feed and kill rates, respectively. We use diffusion coefficients $\delta_A = 2 \times 10^{-5}$ and $\delta_B = 1 \times 10^{-5}$, feed rate $f = 0.098$, and kill rate $k = 0.057$. The original trajectories contain 1001 time frames of 128×128 spatial fields; we downsample them to a $64 \times 64 \times 64$ space–time grid, corresponding to 64 time frames on a 64×64 spatial grid. HF references correspond to the two-channel scalar concentration field (A, B). LF inputs are constructed by subsampling sparse HF concentration observations at each physical time frame and applying bi-cubic interpolation channel-wise to obtain a provisional full-field estimate.

(2+1)D Kolmogorov Flow. We use the turbulent (2+1)D Kolmogorov flow dataset from prior work [35], also adopted in a subsequent field-reconstruction study [15]. Here, $(2+1)$D indicates two spatial dimensions (ξ_1, ξ_2) together with one temporal dimension γ. We consider the vorticity form of the Navier–Stokes dynamics on a periodic spatial domain $\xi = (\xi_1, \xi_2) \in (0, 2\pi)^2$ over the time horizon $\gamma \in (0, 1.25]$:

$$\partial_\gamma \omega(\xi,\gamma) + v(\xi,\gamma) \cdot \nabla\omega(\xi,\gamma) = \frac{1}{Re}\nabla^2\omega(\xi,\gamma) + f(\xi,\gamma),$$

where $\omega(\xi,\gamma)$ denotes the vorticity field and $v(\xi,\gamma)$ is the corresponding velocity field. We use the benchmark configuration from prior work [35], with Reynolds number $Re = 1000$ and forcing $f(\xi,\gamma) = -4\cos(4\xi_2) - 0.1\,\omega(\xi,\gamma)$. Spatial fields are discretized on a 256×256 grid, and we use 40 uniformly spaced time frames over $\gamma \in (0, 1.25]$. HF references are obtained from the provided simulations. LF inputs are constructed by subsampling sparse HF vorticity observations and applying bi-cubic interpolation to obtain a provisional full-field estimate. To evaluate physics consistency, we follow the residual formulation used in prior work [15] and compute PDE residuals from the vorticity equation; under periodic boundary conditions, the velocity field is recovered from vorticity via the stream function ψ:

$$v(\xi,\gamma) = \nabla \times \psi(\xi,\gamma), \quad \psi(\xi,\gamma) = \nabla^{-2}\omega(\xi,\gamma).$$

Dynamic 1D Diffusion–Reaction Equation. We use the Dynamic 1D Diffusion–Reaction dataset from PDEBench [47]. We consider the one-dimensional Diffusion–Reaction equation on $\xi \in (0, 1)$ over the time horizon $\gamma \in (0, 1]$,

$$\partial_\gamma u(\xi,\gamma) = \nu\,\nabla^2 u(\xi,\gamma) + \rho\,u(\xi,\gamma)\big(1 - u(\xi,\gamma)\big),$$

with initial condition $u(\xi, 0) = u_0(\xi)$. HF trajectories are generated by simulating the equation with diffusion coefficient $\nu = 5$ and reaction coefficient $\rho = 5$. HF references correspond to the scalar solution field $u(\xi,\gamma)$. LF inputs are constructed by applying nearest-neighbor interpolation to the sparse HF measurements, yielding a coarse proxy field.

B.2 Implementation Details

PICSB parameterizes the velocity network $v_\theta(x, \mathrm{cond})$ using an AdaIN-conditioned U-Net based on the architecture introduced in prior work [40]. The network takes as input the current state x_{τ_t} and a conditioning vector cond. The conditioning vector $\mathrm{cond} = \mathrm{MLP}(\tau_t)$ is generated by a time-embedding network that encodes the diffusion time. Conditioning is injected through Adaptive Instance Normalization (AdaIN) [48] in the encoder blocks, enabling the drift to vary with diffusion time while keeping the model fully convolutional over the spatial grid.

Input representation. Depending on the benchmark, the input field x_{τ_t} is represented using different tensor encodings. For the Burgers' equation, 1D Diffusion–Reaction equation and Darcy flow, the field is represented as a single channel tensor of shape $x_{\tau_t} \in \mathbb{R}^{B\times 1\times H\times W}$. For Gray–Scott, the field is represented as a tensor of shape $x_{\tau_t} \in \mathbb{R}^{B\times C\times T\times H\times W}$, where C denotes the number of channels of concentration field and T denotes the number of time frames. For the Kolmogorov flow, the temporal dimension is encoded along the channel axis, resulting in a tensor representation $x_{\tau_t} \in \mathbb{R}^{B\times C\times H\times W}$, where C denotes the number of time frames and is set to 40 in our experiments. Sparse observations are represented by a boolean mask $\mathbf{M}$ with the same shape as x_{τ_t}, and the corresponding observed values are denoted by y. Hard conditioning is enforced by overwriting the observed entries at every update according to

$$x_{\tau_t} \leftarrow \mathbf{M} \odot y + (1 - \mathbf{M}) \odot x_{\tau_t},$$

which ensures that the observed locations remain fixed throughout both training and sampling.

AdaIN conditioning. Given an activation tensor $x_{\tau_t} \in \mathbb{R}^{B\times C\times H\times W}$ and a conditioning vector $\mathrm{cond} \in \mathbb{R}^{B\times d_{\mathrm{cond}}}$ (d_{cond} corresponds to the time embedding dimension), AdaIN injects the conditioning information through channel-wise affine transformations. Specifically, the conditioning vector is linearly projected to produce a pair of scale and shift parameters,

$$(\gamma, \beta) = \mathrm{Linear}(\mathrm{cond}) \in \mathbb{R}^{B\times(2C)}, \qquad \gamma, \beta \in \mathbb{R}^{B\times C},$$

which are then reshaped to $(B, C, 1, \ldots, 1)$, i.e., $(B, C, 1, 1, 1)$ for Gray–Scott and $(B, C, 1, 1)$ for the remaining benchmarks, and applied to the normalized activations as

$$\mathrm{AdaIN}(x_{\tau_t}, \mathrm{cond}) = \gamma \odot \mathrm{IN}(x_{\tau_t}) + \beta.$$

Here, $\mathrm{IN}(\cdot)$ denotes instance normalization, which normalizes each channel independently using per-sample spatial statistics. By modulating feature statistics in this manner, AdaIN enables the velocity field to vary smoothly with diffusion time while preserving a fully convolutional structure over the spatial grid.

Network architecture and training hyperparameters. Table 3 summarizes the key architectural and training hyperparameters used for PICSB across all PDE benchmarks. All benchmarks share the same AdaIN-conditioned U-Net backbone, while benchmark-specific settings include the input/output channel dimensionality, noise scale ϵ, learning rate, and gradient clipping threshold. The time-embedding network is implemented as a two-layer MLP with a hidden dimension of 16.

All models are trained using a fixed solver discretization with 10 bridge-time steps and the noise scale specified in Table 3, and the same settings are used during inference-time sampling to ensure consistency between training and inference dynamics. A refresh period is set to 100, meaning that the corresponding refresh operation is performed every 100 training iterations, resulting in a total of 1000 refreshes over $100{,}000$ training iterations. To stabilize optimization under different PDE scalings, gradient clipping is applied during training: for the Burgers, Gray–Scott, and Kolmogorov benchmarks, gradients of both the velocity network and the time-embedding network are clipped with a maximum norm of 1, whereas for the 1D Diffusion–Reaction and Darcy benchmarks—which exhibit different scaling characteristics—a much smaller clipping threshold of 10^{-5} is used. These settings were found to be crucial for stable training across all benchmarks.

Table 3: PICSB architectural and training hyperparameters used for each PDE benchmark. All benchmarks share the same network architecture and training schedule, except for the input/output channel dimensionality and learning rate. The table summarizes the key structural and optimization-related hyperparameters that define the PICSB model.

PDE	Enc. ch.	Dec. ch.	#Enc.	#Dec.	Kernel	d_{cond}	Bridge-time steps	ϵ	LR	Train iters.
Burgers	$128, 256$	$256, 128$	2	2	3×3	8	10	10^{-2}	10^{-3}	100k
Darcy	$128, 256$	$256, 128$	2	2	3×3	8	10	10^{-6}	10^{-6}	100k
Gray–Scott	$128, 256$	$256, 128$	2	2	3×3	8	10	10^{-2}	10^{-4}	100k
Kolmogorov	$128, 256$	$256, 128$	2	2	3×3	8	10	10^{-2}	10^{-3}	100k
Diffusion–Reaction	$128, 256$	$256, 128$	2	2	3×3	8	10	10^{-7}	10^{-4}	100k

In contrast to the baselines, PICSB does not use full-field HF references for hyperparameter tuning or model selection. Hyperparameter tuning based on training-set reconstruction error is performed only for *PINNs* and *Guidance*.

PINNs. We implement the PINNs baseline as a coordinate-based MLP that parameterizes the unknown solution field as $u_\phi(\xi, \gamma)$ (or $u_\phi(\xi)$ for steady-state), and optimizes the network parameters independently for each test instance using Adam for 100k steps (Table 4). The MLP consists of stacked fully connected layers with $\tanh$ activations and Xavier initialization, and its input dimensionality is benchmark-dependent: (ξ, γ) for Burgers and 1D Diffusion–Reaction, (ξ_1, ξ_2) for Darcy, and (ξ_1, ξ_2, γ) for Gray–Scott and Kolmogorov. The output is scalar-valued except for Gray–Scott, where the network predicts the two-channel concentration field (A, B).

At each optimization step, the loss is a weighted sum of an observation-fitting term and a physics residual term, $\lambda_{\text{obs}}\mathcal{L}_{\text{obs}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}}$, where $\mathcal{L}_{\text{phys}}$ is computed by evaluating the governing PDE residual at N_{colloc} collocation points via a dedicated physics-loss routine. For the observation term, we fit the network to the available sparse HF observations. For each benchmark, we fix $\lambda_{\text{obs}} = 1$ and tune λ_{phys} over 10^α with $\alpha \in \{-5, \dots, 1\}$ based on the relative L^2 error on the training set.

For Darcy, Dirichlet boundary conditions are explicitly enforced by sampling N_{bc} boundary coordinates on $\partial(0, 1)^2$ and adding an extra boundary penalty weighted by λ_{bc}. To stabilize optimization under Darcy's scaling, we additionally apply gradient clipping with a maximum norm of 10^{-5} during each training step, whereas the other benchmarks do not require this clipping. After optimization, we recover the full-grid field by evaluating the trained coordinate MLP on a dense coordinate mesh.

Table 4: PINNs architectural and optimization hyperparameters used for each PDE benchmark. The table summarizes the key structural and optimization-related hyperparameters that define the PINNs baseline, with Darcy-specific stabilization settings reported in the last column.

PDE	Hidden dim	Depth	N_{colloc}	λ_{obs}	λ_{phys}	LR	Train iters.	Extra settings
Burgers	512	6	8192	1.0	1.0	10^{-3}	100k	–
Darcy	512	6	8192	1.0	1.0	10^{-3}	100k	λ_{bc}=1, N_{bc}=2048, grad clip (1e-5)
Gray–Scott	128	4	8192	1.0	10^{-5}	10^{-3}	100k	–
Kolmogorov	256	4	8192	1.0	1.0	10^{-3}	100k	–
Diffusion–Reaction	128	4	4096	1.0	10^{-5}	10^{-3}	100k	–

Guidance-based sampling (*Guidance*). We implement an inference-time guidance baseline adapted from PDE-guided diffusion pipelines [18], in which conditional generation is achieved by augmenting an unconditional EDM sampler with gradient-based corrections derived from sparse HF observations and PDE residuals. Since full-grid HF trajectories are unavailable for training in our multi-fidelity setting, the generative prior is pretrained solely on full-grid LF fields, and HF information is incorporated only at inference time through guidance.

Given an EDM noise schedule $(\sigma_{\min}, \sigma_{\max}, \rho)$ and a fixed number of sampling steps, each sampling step first performs a standard EDM update consisting of an Euler step followed by a second-order correction [49], producing a proposal state x_t. Guidance is then applied by differentiating scalar loss functions with respect to the current iterate and updating the state along the resulting gradients.

Across all benchmarks, both the observation loss and the physics loss are computed using ℓ_2 norms of their respective residuals. The observation term penalizes mismatches between the current field and the sparse HF measurements, while the physics term penalizes violations of the governing PDE evaluated on the current field. During sampling, gradients of these losses with respect to x_t are used to steer the trajectory toward observation-consistent and physically plausible solutions. For the 1D Diffusion–Reaction, Gray–Scott, and Kolmogorov benchmarks, guidance is applied in a stage-wise manner. During the early portion of the sampling trajectory, only observation guidance is used to rapidly align the generated samples with the sparse HF measurements. After the first 80% of the total sampling steps, PDE guidance is activated, and the update combines a reduced observation step with an additional physics-based correction.

For 1D Diffusion–Reaction, guidance is applied to the scalar solution field, while for Gray–Scott it is applied jointly to the two concentration channels (A, B) using the Gray–Scott residuals. The corresponding weights $(\lambda_{\text{obs}}, \lambda_{\text{phys}})$ are reported in Table 5.

For Darcy, additional stabilization is required due to the markedly different scaling induced by the elliptic operator and the spatially varying permeability field. In this case, boundary consistency is explicitly enforced during sampling by introducing a boundary residual term defined on a boundary mask. PDE guidance is activated after the first 40% of the sampling steps, and the update initially combines observation, PDE, and boundary guidance. After the first 80% of the sampling steps, only PDE guidance is retained to stabilize late-stage sample generation. Separate step sizes are used for observation, PDE, and boundary guidance, and the corresponding gradients are optionally normalized and clipped to improve robustness under varying state scales.

Since *Guidance* is highly sensitive to residual-gradient scaling, activation schedules, and guidance weights, we tune these hyperparameters separately for each benchmark. For each residual-guidance coefficient, we perform a two-stage search. We first select a scaling factor of the form 10^{α}, with $\alpha \in \{-5, \dots, 5\}$, based on the relative L^2 error on the training set. We then search over a coefficient factor 2^{β}, with $\beta \in \{0, \dots, 6\}$, and use $2^{\beta} \times 10^{\alpha}$ as the final coefficient. The final values are reported in Table 5.

Table 5: Guidance-based sampling hyperparameters used for each PDE benchmark. All benchmarks use the same noise schedule parameters $(\sigma_{\min}, \sigma_{\max}, \rho)$. Observation and PDE guidance are applied in a stage-wise manner, while Darcy additionally employs boundary guidance with separate step sizes.

PDE	Sampling steps	$\sigma_{\min}$	$\sigma_{\max}$	ρ	λ_{obs}	λ_{phys}	λ_{bc}
Burgers	2000	0.002	80	7	320	100	–
Darcy	2000	0.002	80	7	10	10	10
Gray–Scott	2000	0.002	80	7	10000	1000	–
Kolmogorov	2000	0.002	80	7	3.2×10^6	1000	–
Diffusion–Reaction	500	0.002	80	7	320	10^{-1}	–

B.3 Visualization Details for Main Qualitative Comparison

Figure 2 visualizes representative test instances used for the main qualitative comparison. For 1D Diffusion–Reaction and Darcy flow, we show a single held-out test field. For Gray–Scott, we visualize a B-concentration snapshot at $\gamma \approx 3.02 \times 10^3$ seconds from the trajectory horizon $\gamma \in (0, 10^4]$. For Kolmogorov flow, we visualize a vorticity snapshot at $\gamma = 0.64$ from the trajectory horizon $\gamma \in (0, 1.25]$. These snapshot choices are used only for visualization; quantitative metrics are computed over the full evaluated fields.

C Supplementary Experiments

C.1 Extended Main Experimental Results

Table 6 provides the full quantitative results across all evaluated benchmarks and additional Burgers LF resolutions. It extends the main-text comparison by reporting the same reconstruction metrics and inference wall-times for all considered datasets and settings.

Table 6: Main quantitative comparison over 40 test samples. Error entries report 95% confidence intervals, and wall-time entries report average per-sample inference time. For Burgers, the number in parentheses denotes the LF space–time grid resolution N, where the LF simulation is performed on an $N \times N$ grid. Lower is better; **bold** marks the lowest mean, and † denotes CI overlap with the best method.

Dataset	Metric	*LF*	*PINNs*	*Guidance*	PICSB
Burgers (16)	RMSE ($\times 10^{-2}$)	[9.668, 11.303]	**[1.957, 2.665]**	[3.723, 4.986]	[2.489, 2.718]†
	MAE ($\times 10^{-3}$)	[79.557, 92.326]	[14.929, 20.462]	[23.258, 31.769]	**[6.347, 7.357]**
	RelError (%)	[71.149, 72.087]	**[13.589, 18.117]**	[25.758, 30.551]	[16.820, 21.915]†
	fRMSE ($\times 10^{-4}$)	[10.375, 12.424]	[3.533, 4.877]†	[5.609, 7.584]	**[3.467, 3.993]**
	Average wall-time (seconds)	-	3,957	143	**0.027**
Burgers (32)	RMSE ($\times 10^{-2}$)	[9.576, 11.183]	[1.893, 2.842]†	[2.969, 4.102]	**[2.223, 2.381]**
	MAE ($\times 10^{-3}$)	[78.461, 90.933]	[14.110, 21.402]	[17.417, 24.602]	**[5.752, 6.798]**
	RelError (%)	[70.426, 71.428]	[12.445, 22.565]†	[20.177, 25.015]	**[14.889, 19.700]**
	fRMSE ($\times 10^{-4}$)	[9.959, 11.889]	[3.463, 5.093]	[4.520, 6.298]	**[3.069, 3.453]**
	Average wall-time (seconds)	-	4,068	155	**0.026**
Diffusion–Reaction	RMSE ($\times 10^{-2}$)	**[0.288, 0.352]**	[0.355, 1.275]	[0.152, 0.885]†	[0.516, 0.531]
	MAE ($\times 10^{-3}$)	[1.127, 1.507]	[1.946, 7.897]	[0.461, 5.387]†	**[0.795, 0.871]**
	RelError (%)	**[0.309, 0.382]**	[0.372, 1.411]†	[0.153, 0.984]†	[0.553, 0.575]
	fRMSE ($\times 10^{-4}$)	**[0.595, 0.722]**	[1.270, 4.884]	[0.400, 3.306]†	[1.534, 1.588]
	Average wall-time (seconds)	-	4,706	28	**0.012**
Darcy	RMSE ($\times 10^{-3}$)	[1.808, 1.845]	[2.496, 3.212]	[1.519, 1.551]	**[0.121, 0.129]**
	MAE ($\times 10^{-3}$)	[1.296, 1.311]	[1.998, 2.575]	[1.093, 1.103]	**[0.089, 0.094]**
	RelError (%)	[4.405, 4.451]	[6.048, 7.833]	[3.699, 3.745]	**[0.293, 0.313]**
	fRMSE ($\times 10^{-4}$)	[0.199, 0.205]	[0.384, 0.518]	[0.178, 0.185]	**[0.018, 0.020]**
	Average wall-time (seconds)	-	4,647	141	**0.013**
Gray–Scott	RMSE ($\times 10^{-1}$)	[1.159, 1.287]	**[0.273, 0.362]**	[1.240, 1.370]	[0.330, 0.370]†
	MAE ($\times 10^{-2}$)	[5.870, 6.990]	[1.109, 1.667]	[6.598, 7.689]	**[0.859, 1.061]**
	RelError (%)	[17.521, 19.740]	**[4.129, 5.534]**	[18.736, 21.028]	[4.989, 5.693]†
	fRMSE ($\times 10^{-3}$)	[0.537, 0.600]	**[0.167, 0.312]**	[1.022, 1.142]	[0.219, 0.268]†
	Average wall-time (seconds)	-	2,557	498	**0.109**
Kolmogorov	RMSE ($\times 10^{0}$)	[0.668, 0.729]	[2.971, 3.217]	[0.547, 0.587]	**[0.196, 0.220]**
	MAE ($\times 10^{0}$)	[0.286, 0.319]	[2.344, 2.555]	[0.312, 0.338]	**[0.109, 0.124]**
	RelError (%)	[13.497, 14.433]	[60.590, 63.026]	[11.043, 11.665]	**[3.972, 4.299]**
	fRMSE ($\times 10^{-2}$)	[0.159, 0.173]	[3.387, 3.695]	[0.232, 0.247]	**[0.118, 0.144]**
	Average wall-time (seconds)	-	9,527	787	**0.028**

Figure 3 extends the main-text qualitative comparison to all evaluated benchmarks.

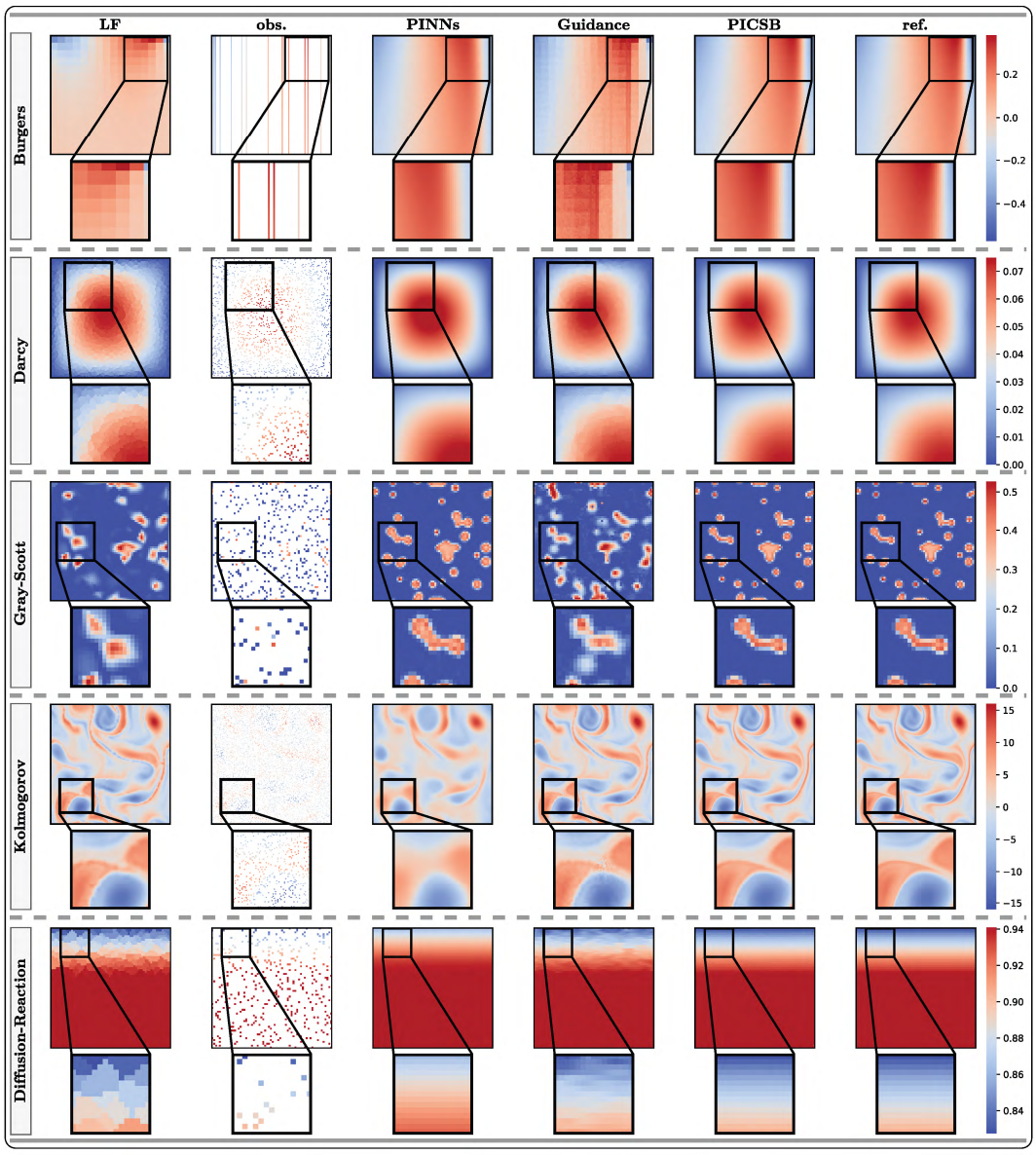

Figure 3: Qualitative comparison of reconstructed fields. We show representative test samples for Burgers, Darcy flow and Diffusion–Reaction, and snapshots for Gray–Scott and Kolmogorov flow; LF, sparse HF observations (obs.), *PINNs*, *Guidance*, and PICSB are compared with the HF reference (ref.), with zoomed-in regions highlighting local reconstruction quality.

Figure 4 further visualizes representative reconstruction errors across the four main benchmarks. The three methods exhibit different error characteristics across the systems. For Burgers and Gray–Scott, both *PINNs* and PICSB reproduce the dominant field structure, while *Guidance* shows more spatially distributed discrepancies and visible artifacts. For Darcy, the *PINNs* and *Guidance* errors remain spread over the domain, whereas the PICSB reconstruction follows the smooth pressure profile with only small residual differences. For Kolmogorov flow, *PINNs* substantially smooths the vortical field and *Guidance* retains noticeable local discrepancies, while PICSB preserves the dominant vortical structures with errors concentrated around finer-scale regions. Overall, the error maps indicate that reconstruction differences are primarily associated with sharp transitions, localized reaction structures, and fine vortical features rather than the global field organization.

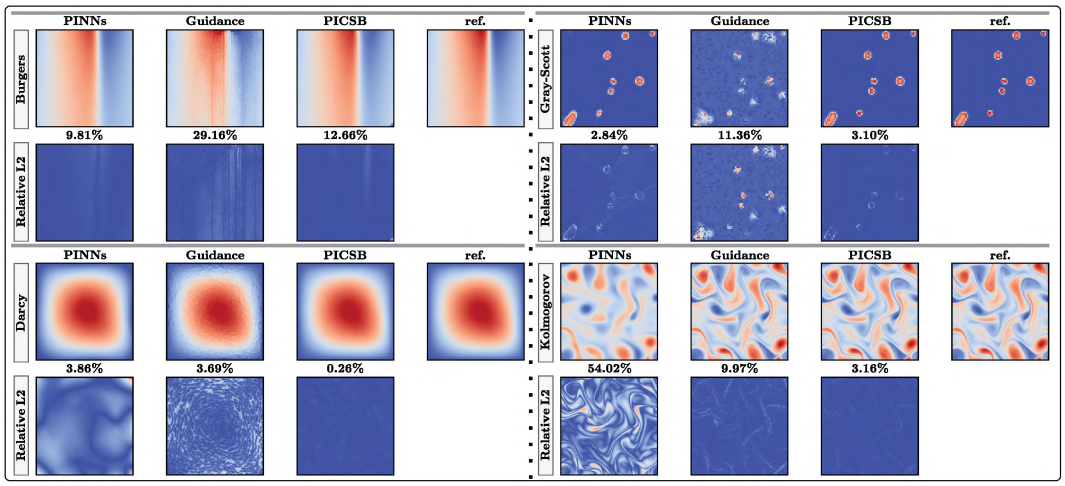

Figure 4: Representative reconstructions and relative L^2 error maps for Burgers, Gray–Scott, Darcy, and Kolmogorov flow. For each benchmark, the top row compares the reconstructed fields with the HF reference, while the bottom row visualizes the corresponding error distribution. Percentages denote the sample-wise relative L^2 errors.

C.2 Additional Experiments under Challenging Settings

Table 7 reports additional experiments under more challenging observation regimes and noisy-training settings. We first consider a globally fixed sensor network, where a single set of sensors is selected on 10% of the spatial grid and reused for all test samples. To further examine diverse spatiotemporal observation patterns, we start from a globally fixed sensor network with temporally consistent observations and introduce intermittent sensor failures over time. Specifically, sensors are placed on 10% of the spatial grid, and among these, up to 10% are randomly deactivated for contiguous durations ranging from 3 to 12 time steps. This breaks the temporal homogeneity of the observations, resulting in sensor-specific missing measurements across time. For noisy HF supervision, we use the same globally fixed sensor configuration and add Gaussian noise to the supervision during training at levels of 5% and 10%, simulating increasing corruption relative to the data scale. All models are trained for 100k iterations.

Table 7: Robustness study on Kolmogorov (8 samples). Lower values indicate better performance for all metrics. **Boldface** denotes the method with the lowest mean value over the test samples

Setting	Method	RMSE	MAE	RelError (%)	fRMSE
globally fixed sensor network	*PINNs*	3.090	2.436	64.012	0.035
	Guidance	0.679	0.398	14.062	0.004
	PICSB	**0.471**	**0.265**	**9.741**	**0.002**
w/o temporal homogeneity	*PINNs*	3.099	2.443	64.215	0.036
	Guidance	0.674	0.395	13.970	0.003
	PICSB	**0.402**	**0.203**	**8.302**	**0.002**
noisy HF supervision (5%)	*PINNs*	3.097	2.441	64.154	0.035
	Guidance	0.693	0.430	14.360	0.004
	PICSB	**0.429**	**0.264**	**8.878**	**0.003**
noisy HF supervision (10%)	*PINNs*	3.093	2.439	64.081	0.035
	Guidance	0.861	0.550	17.846	**0.004**
	PICSB	**0.718**	**0.424**	**14.917**	0.006

C.3 Additional Experiments under Sensor-Location Regimes

To examine more realistic sensing scenarios, we evaluate Burgers, Darcy, and Kolmogorov under three observation regimes defined by how sensor locations are selected and reused across time and samples. For each benchmark, we use 32 training pairs and 4 test pairs, and all experiments are conducted on a single NVIDIA RTX 4090 GPU.

(R1) Resampled observations. For dynamic problems (Burgers and Kolmogorov), we independently resample a random 10% set of spatial sensor locations at each physical time frame. For the steady Darcy problem, this corresponds to sampling a random 10% subset of grid points per sample.

(R2) Per-sample fixed sensor set. For dynamic problems, we draw one random 10% sensor set per trajectory and keep it fixed across all time frames of that trajectory, while different trajectories use different sensor sets. For Darcy, this setting coincides with (R1); thus we do not report a separate Darcy result for (R2).

(R3) Globally fixed sensor network. We choose a single 10% sensor set once and reuse it for all samples. For dynamic problems, the same sensor coordinates are applied at every time frame.

C.3.1 Additional qualitative visualizations

We present additional qualitative visualizations of PICSB. Figure 5 and Figure 6 report reconstruction results for the Burgers and Darcy benchmarks, respectively, while Figures 7–9 visualize the Kolmogorov benchmark at multiple representative temporal frames. All figures report reconstruction results across all regimes (R1–R3) and all test samples (four instances), together with the corresponding HF reference solutions. Reported relative L^2 errors are computed per sample and shown below each reconstructed field.

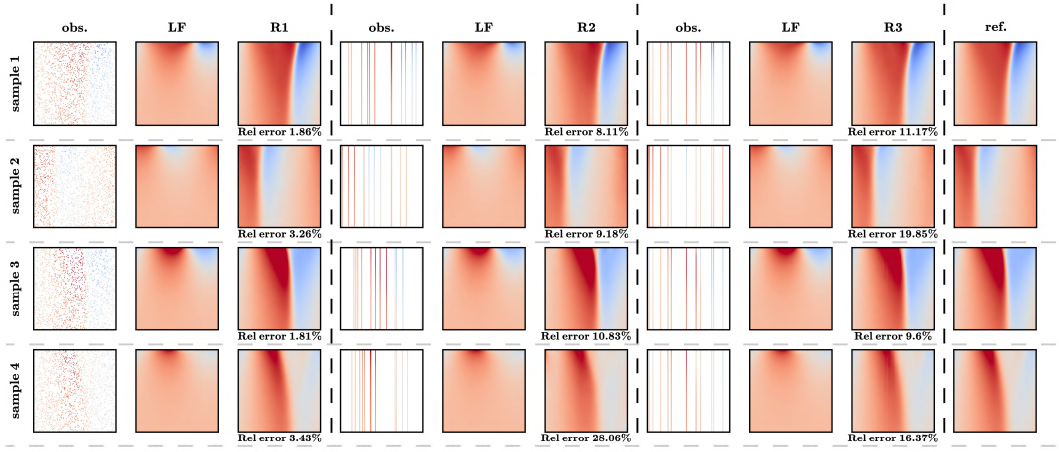

Figure 5: Additional visualization illustrating reconstruction performance for Burgers' equation. Each row corresponds to a different test sample. Columns show the sparse HF observations used for conditioning (obs.), the LF input, reconstructions obtained under different reconstruction regimes (R1–R3), and the HF reference solution (ref.). Relative L^2 errors are reported below each reconstructed field.

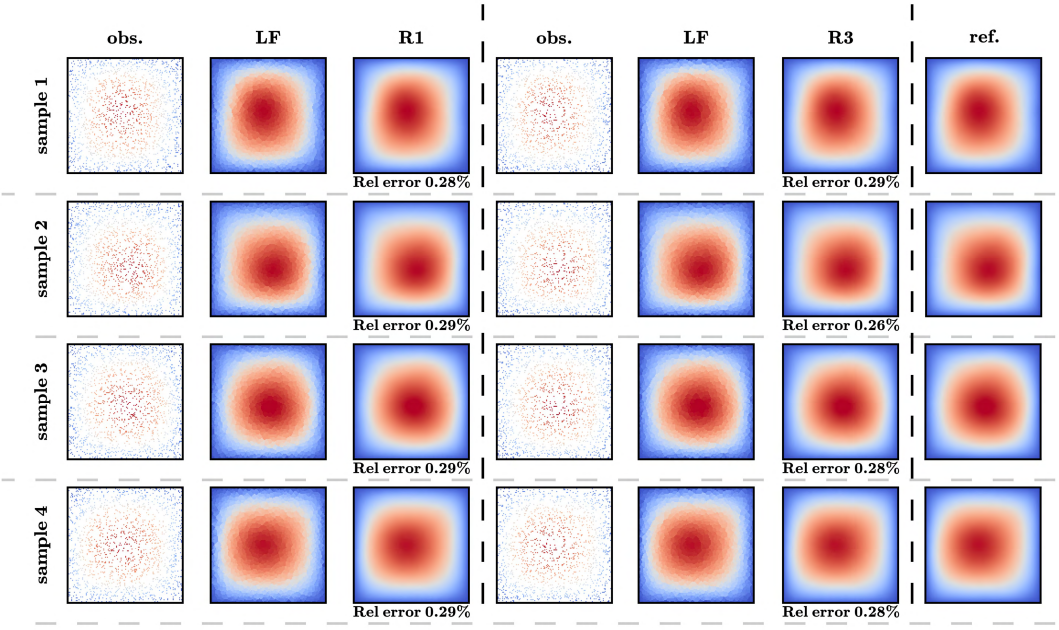

Figure 6: Additional visualization illustrating reconstruction performance for Darcy flow. Each row corresponds to a different test sample. Columns show the sparse HF observations used for conditioning (obs.), the LF input, reconstructions obtained under different reconstruction regimes (R1/R3), and the HF reference solution (ref.). Relative L^2 errors are reported below each reconstructed field.

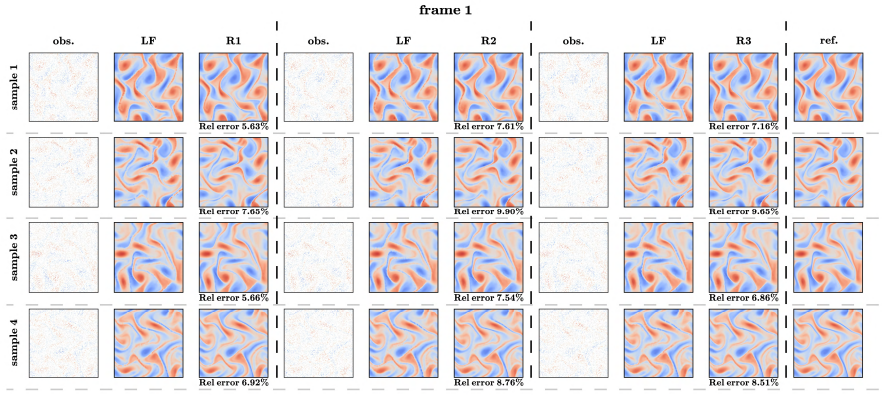

Figure 7: Additional visualization illustrating reconstruction performance for Kolmogorov flow at frame 1. Each row corresponds to a different test sample. Columns show the sparse HF observations used for conditioning (obs.), the LF input, reconstructions obtained under different reconstruction regimes (R1–R3), and the HF reference solution (ref.). Relative L^2 errors are reported below each reconstructed field.

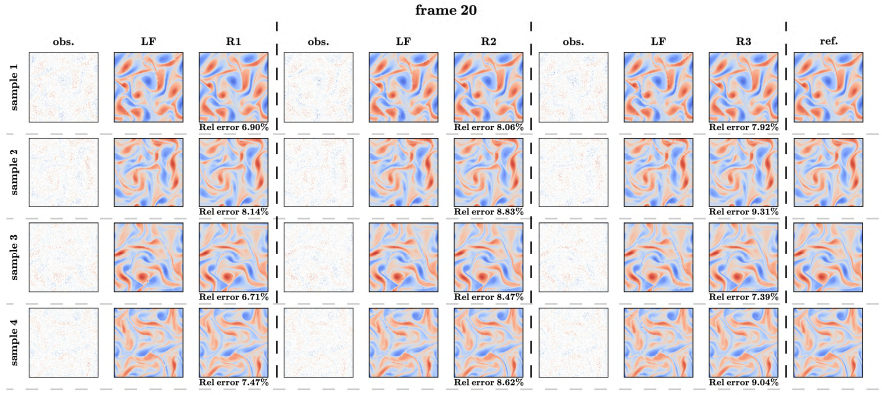

Figure 8: Additional visualization illustrating reconstruction performance for Kolmogorov flow at frame 20. Each row corresponds to a different test sample. Columns show the sparse HF observations used for conditioning (obs.), the LF input constructed from sparse observations via interpolation, reconstructions obtained under different reconstruction regimes (R1–R3), and the HF reference solution (ref.). Relative L^2 errors are reported below each reconstructed field.

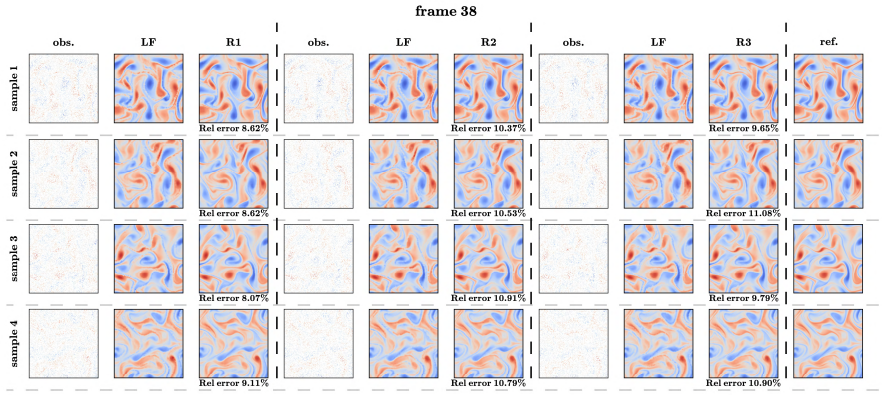

Figure 9: Additional visualization illustrating reconstruction performance for Kolmogorov flow at frame 38. Each row corresponds to a different test sample. Columns show the sparse HF observations used for conditioning (obs.), the LF input constructed from sparse observations via interpolation, reconstructions obtained under different reconstruction regimes (R1–R3), and the HF reference solution (ref.). Relative L^2 errors are reported below each reconstructed field.

C.3.2 Inference dynamics comparison between PICSB and Guidance

This section provides a comparison of inference dynamics between PICSB and the guidance-based diffusion baseline. Figures 10, 11, and 12 visualize the evolution of reconstructed fields for the Burgers, Darcy, and Kolmogorov benchmarks, respectively.

Across all benchmarks, the guidance-based sampling method relies on a long diffusion trajectory with hundreds to thousands of sampling steps before meaningful large-scale structures emerge. During early and intermediate stages of sampling, the reconstructed fields remain dominated by noise, and coherent solution structures typically appear only near the end of the diffusion process.

In contrast, PICSB directly transports the LF prior toward the observation-conditioned HF posterior through a Schrödinger bridge formulation. As a result, physically consistent large-scale structures are recovered within a significantly smaller number of inference steps. Intermediate states produced by PICSB remain stable and structured throughout the inference trajectory, demonstrating a substantial improvement in inference efficiency compared to diffusion-based guidance sampling.

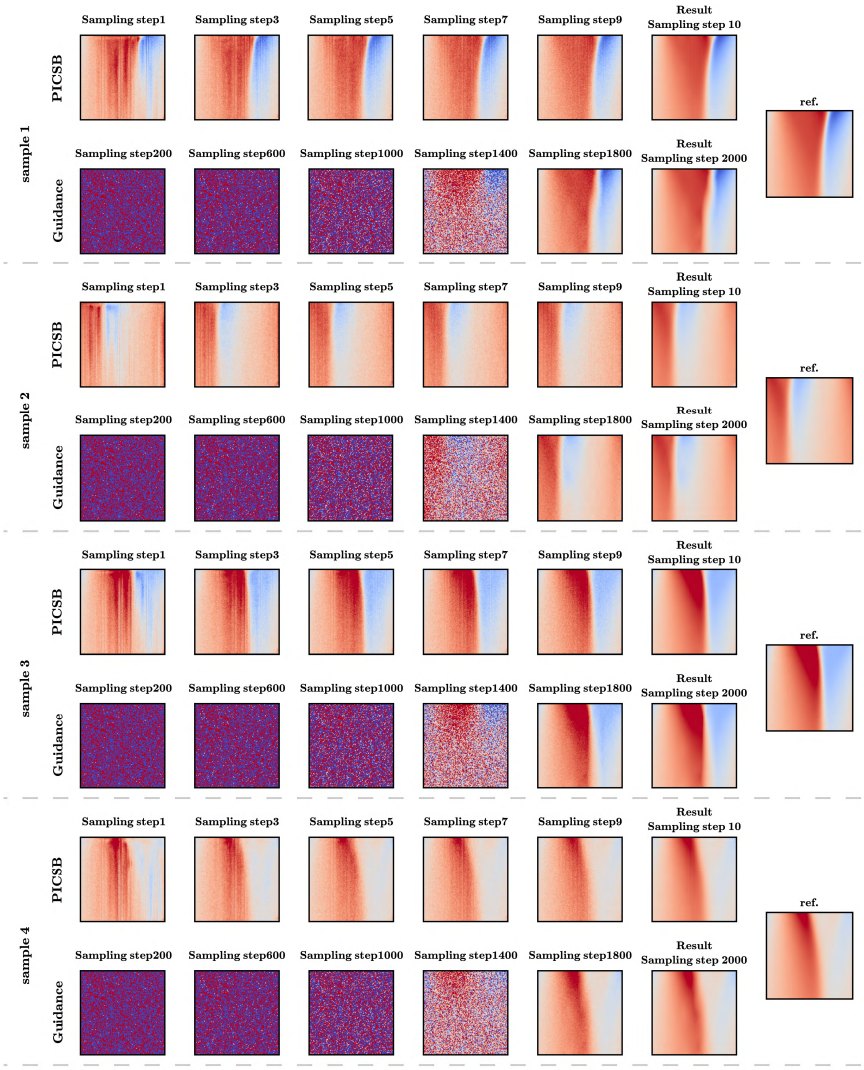

Figure 10: Inference dynamics comparison between PICSB and guidance-based sampling for Burgers' equation.

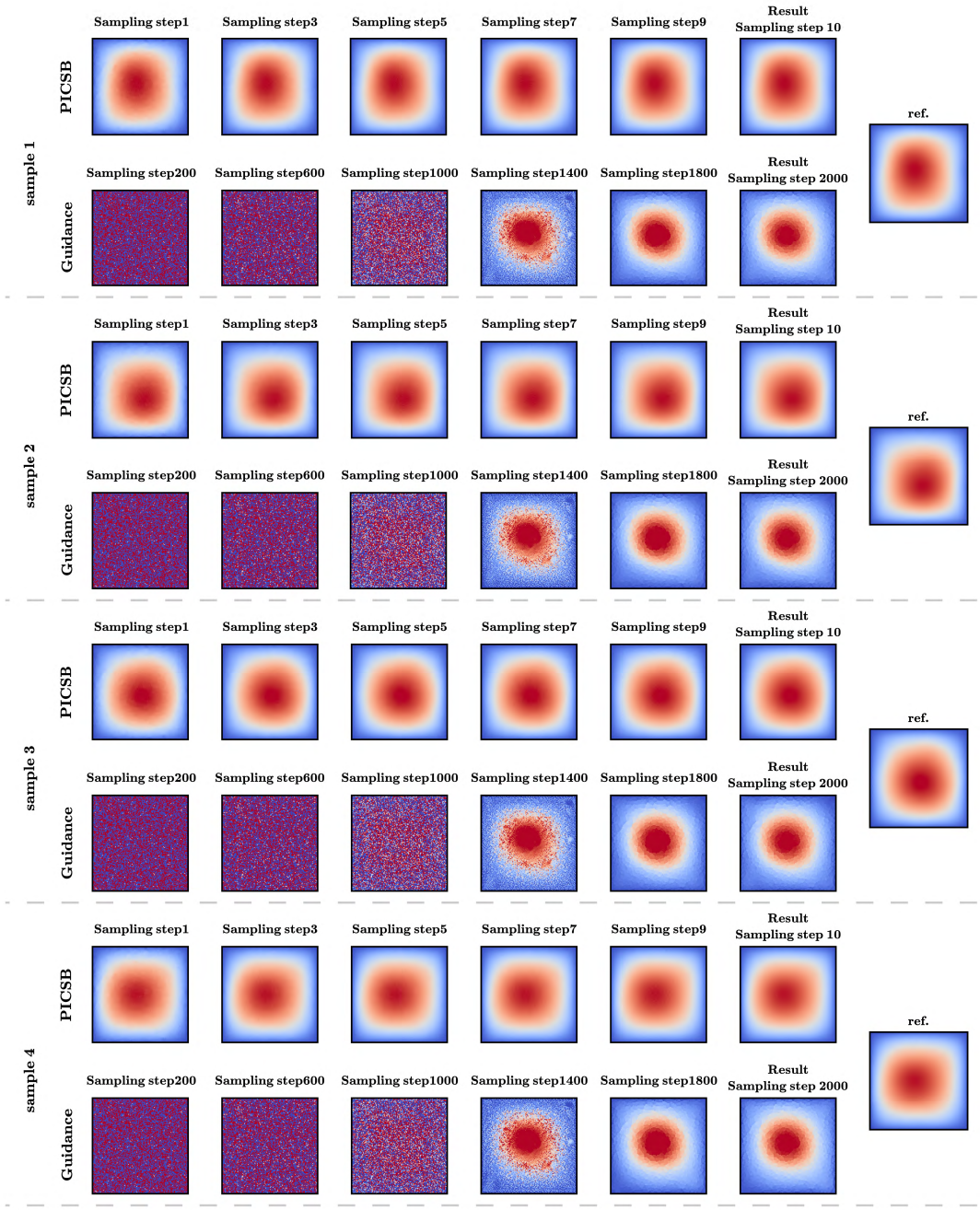

Figure 11: Inference dynamics comparison between PICSB and guidance-based sampling for Darcy flow.

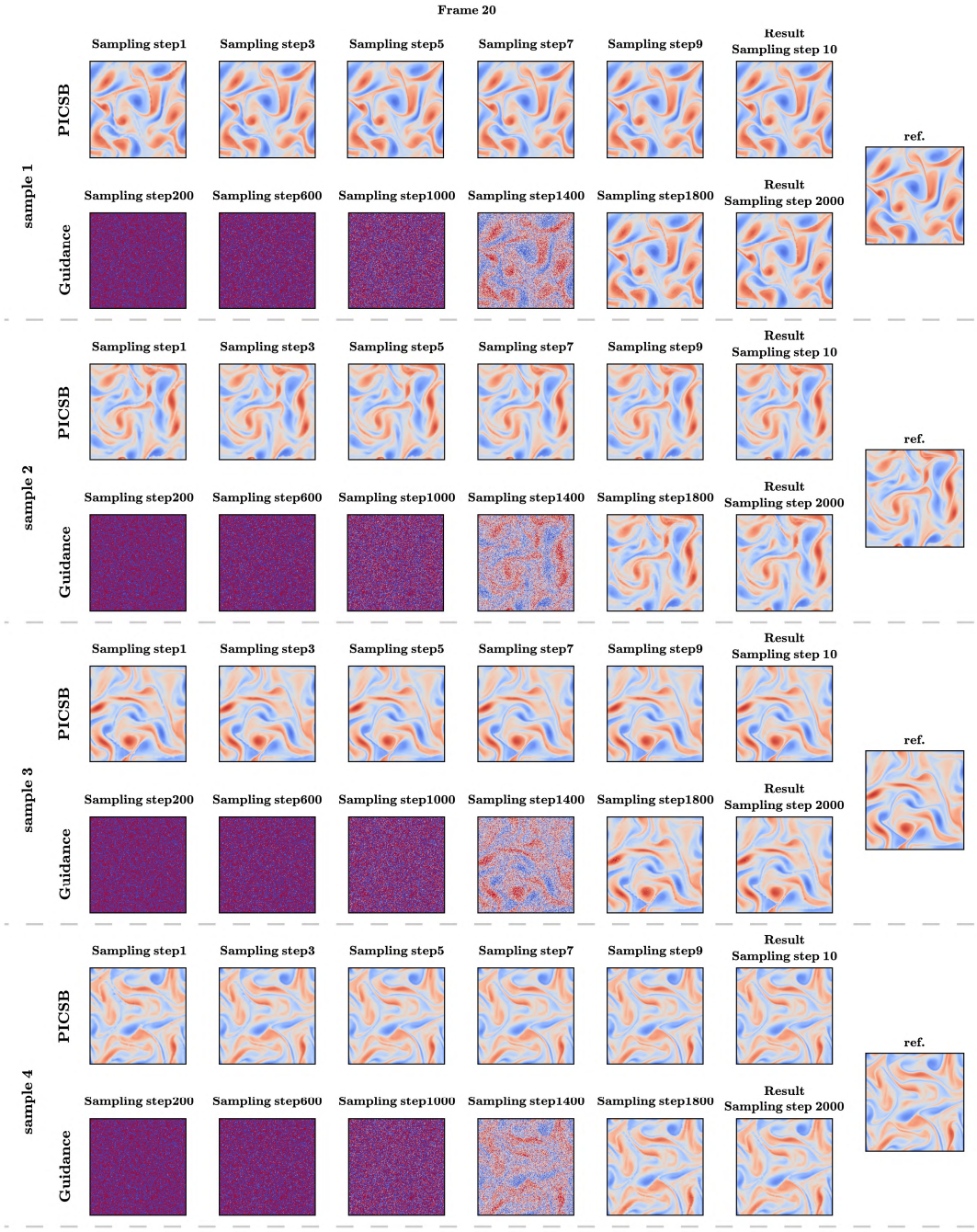

Figure 12: Inference dynamics comparison between PICSB and guidance-based sampling for Kolmogorov flow at frame 20.

C.3.3 Inference with reduced observation availability

In this experiment, we investigate the robustness of PICSB to further reductions in observation availability at inference time, while keeping the LF prior and the trained model fixed. All models are trained using the same LF–HF pairs described in Appendix C.3 and are optimized with sparse HF observations covering 10% of the spatial grid.

At inference time, instead of providing the full 10% observation set used during training, we deliberately subsample this set and condition the model on only a fraction of the available measurements. Specifically, we randomly retain either 50% or 10% of the original observation locations, corresponding to conditioning on 5% and 1% of the full grid, respectively. This setup allows us to isolate the effect of reduced observational information during inference without changing the learned LF-to-HF transport mechanism.

Figure 13 and Figure 14 show reconstruction results for the Burgers' equation across all observation regimes (R1–R3) and all test samples. When 5% of observations are available, reconstruction quality is consistently better than in the 1% setting, indicating that increased observation availability alleviates some of the degradation caused by extreme sparsity. However, the reconstructed fields still exhibit discrepancies from the HF reference, and the overall reconstruction quality remains limited. Moreover, this behavior is primarily observed for Burgers' equation; although results for Darcy flow and Kolmogorov flow are not visualized here, we observed that inference with 5% or fewer observations remains largely unsuccessful for these more complex systems.

These observations suggest that, while PICSB can partially tolerate reduced observation availability at inference time, simply subsampling the observation set without modifying the trained model is insufficient for reliably handling more complex dynamics. Motivated by this limitation, we further investigate whether finetuning can improve inference under severely reduced observation regimes in Section C.8.

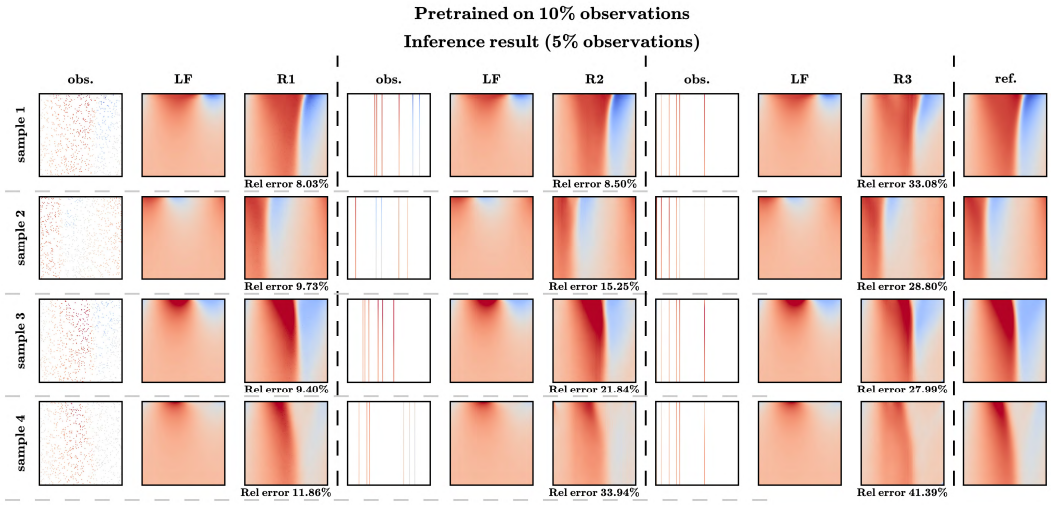

Figure 13: Reconstruction results under 5% sparse observations for the Burgers' equation.

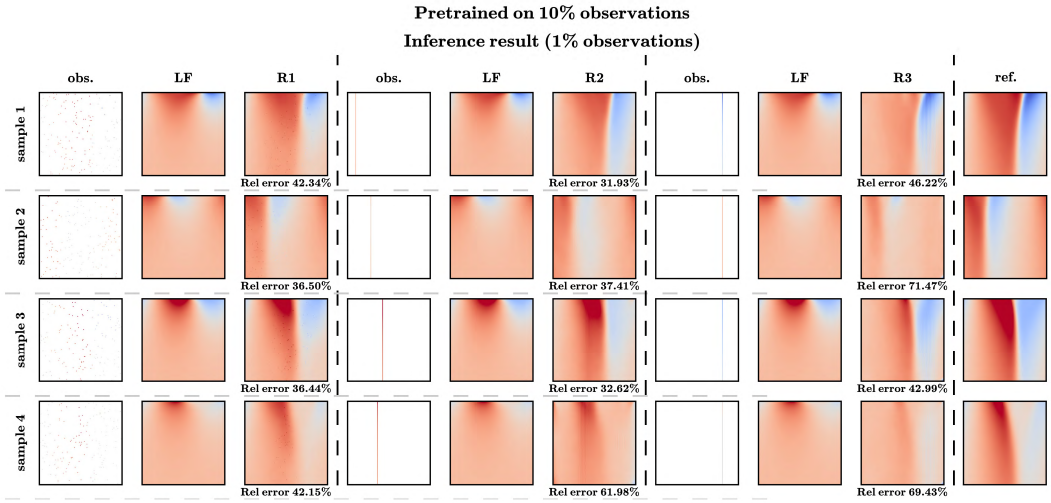

Figure 14: Reconstruction results under 1% sparse observations for the Burgers' equation.

C.4 Hyperparameter Sensitivity

We analyze the sensitivity of PICSB to three key sampling and training hyperparameters: the number of bridge steps, the refresh period, and the Markov noise scale. All experiments are conducted under the globally fixed sensor regime (R3), and each model is trained for 30k iterations. Starting from the default configuration, Refresh $= 100$ and Bridge step $= 10$, with Markov noise 10^{-2} for Burgers and Kolmogorov and 10^{-6} for Darcy, we vary one hyperparameter at a time while keeping the others fixed.

Table 8 reports the resulting relative L^2 errors. Increasing the number of bridge steps degrades performance for Burgers, but improves Darcy and Kolmogorov, suggesting that the optimal bridge discretization depends on the underlying PDE dynamics and LF–HF discrepancy. The Refresh period has a relatively minor impact across all PDEs, with only small variations observed. In contrast, the Markov noise level significantly affects performance, where excessively large noise leads to substantial degradation, while smaller noise levels generally yield better results, particularly for Burgers and Kolmogorov.

Table 8: Hyperparameter sensitivity analysis based on RelError computed over 40 test samples (8 for Kolmogorov) under the R3 observation regime. Lower values indicate better performance.

PDE	Hyperparameter	Value	RelError(%) ↓
Burgers	Bridge step	5	16.55
		10	23.18
		15	28.74
Darcy	Bridge step	5	0.46
		10	0.38
		15	0.30
Kolmogorov	Bridge step	5	9.10
		10	8.22
		15	7.93
Burgers	Refresh period	10	23.43
		100	23.18
		1000	25.62
Darcy	Refresh period	10	0.37
		100	0.38
		1000	0.39
Kolmogorov	Refresh period	10	8.12
		100	8.22
		1000	8.17
Burgers	Markov noise	10^{-3}	12.14
		10^{-2}	23.18
		10^{-1}	33.78
Darcy	Markov noise	10^{-7}	0.36
		10^{-6}	0.38
		10^{-5}	0.66
Kolmogorov	Markov noise	10^{-3}	8.16
		10^{-2}	8.22
		10^{-1}	126.65

C.5 Ablation Study

We further ablate the main components of PICSB on the Kolmogorov benchmark under the R3 observation regime. The bridge component refers to the time discretization of the Brownian bridge process, while conditioning denotes hard projection onto the sparse HF observations at each sampling step. For variants without hard conditioning, observation consistency is instead imposed through an explicit observation fitting loss. All models are trained for 30k iterations and evaluated on 8 test samples.

Table 9 shows that removing either the bridge or conditioning degrades reconstruction accuracy. The degradation is especially severe when hard conditioning is removed, indicating that direct enforcement of sparse HF observations is essential for stable data assimilation. Removing both bridge and conditioning also worsens performance relative to full PICSB, confirming that the bridge dynamics and hard conditioning are complementary components for accurate reconstruction.

Table 9: Ablation study on Kolmogorov (8 samples). Lower is better. **Bold** denotes the best RelError.

Method	RMSE ↓	MAE ↓	RelError (%) ↓	fRMSE ($\times 10^{-3}$) ↓
PICSB	**0.40**	**0.20**	**8.22**	**1.86**
w/o Refresh	0.46	0.26	9.61	2.33
w/o Bridge	0.53	0.27	11.02	2.18
w/o Conditioning (PalSB)	4.03	3.21	83.43	53.50
w/o Bridge & w/o Conditioning	1.64	1.25	34.01	12.99

C.6 Statistical Stability under Repeated Inference

Because PICSB involves stochastic sampling, we evaluate its inference stability by repeating the reconstruction process independently over multiple runs. For each benchmark, we perform 40 independent inference runs and report the standard deviation of the mean evaluation metrics across runs. For Kolmogorov, the same 5-fold evaluation protocol as in the main experiments is used. All experiments follow the R3 observation regime, and the models are trained for 30k iterations.

In addition to metric-level variability, we compute a pixelwise relative standard deviation, defined as $\sigma_i/(|x_i| + \epsilon)$ at each spatiotemporal grid point i, where σ_i is the standard deviation of predictions across repeated runs and x_i is the corresponding HF reference value. Table 10 shows that the standard deviations of all evaluation metrics are extremely small, indicating that PICSB produces stable and consistent reconstructions despite stochastic inference.

Table 10: Statistical stability analysis under repeated inference. We report the standard deviation of mean metrics across 40 independent inference runs. Lower values indicate lower variability.

PDE	RMSE	MAE	RelError(%)	fRMSE	pixelwise
Burgers	9.20×10^{-5}	4.47×10^{-5}	6.53×10^{-4}	1.86×10^{-6}	0.261
Darcy	6.00×10^{-7}	4.90×10^{-7}	1.46×10^{-5}	1.00×10^{-8}	0.005
Kolmogorov	6.26×10^{-5}	3.85×10^{-5}	1.40×10^{-5}	1.41×10^{-6}	0.045

C.7 Inference under Noisy Sparse Observations

In this experiment, we compare **PICSB** with **guidance-based sampling** under inference-time noise corruption in sparse HF observations. For PICSB, we use the same pretrained models as in Appendix C.3, without any retraining or finetuning. At inference time, Gaussian noise is added to the sparse HF observations as

$$\epsilon \sim \mathcal{N}(0, \sigma^2), \quad \sigma = \alpha \cdot \mathrm{std}(\text{data}),$$

where the noise level α is set to 1%, 3%, 5%, 10%, and 15%.

For Darcy and Kolmogorov, the noisy sparse HF observations are interpolated to construct the corresponding LF inputs. Nearest-neighbor interpolation is used for Darcy, and bicubic interpolation is used for Kolmogorov. These noisy LF inputs are then provided to PICSB, while the noisy sparse HF observations are enforced via hard conditioning. For Burgers, we use the original LF inputs described in Appendix B.1 and apply hard conditioning using only the noisy sparse HF observations.

For guidance-based sampling, the same LF-pretrained models are used. During inference, guidance is applied using the noisy sparse HF observations, following the standard guidance formulation without any additional noise-aware modification. Both methods are evaluated under identical noise realizations for fair comparison.

Qualitative analysis. Figures 16–23 provide qualitative comparisons under increasing noise levels. For Burgers, PICSB preserves the global solution structure and smooth transitions even under moderate noise, whereas guidance-based sampling exhibits visible artifacts and amplified noise patterns. For Darcy and Kolmogorov systems, the effect of noisy interpolation is evident for both methods; nevertheless, PICSB produces reconstructions that remain visually coherent and closer to the HF reference, while guidance-based sampling increasingly suffers from noise amplification and loss of structural consistency.

Quantitative results. Figure 15 compares the relative L^2 error under different inference-time noise levels across the considered datasets and observation regimes. As the noise level increases, the reconstruction error generally increases for both PICSB and guidance-based sampling in all settings.

For the Kolmogorov flow, the two methods exhibit largely similar trends with respect to noise level across all sensor regimes (R1–R3). In particular, the relative L^2 error increases at a comparable rate for both PICSB and guidance-based sampling, indicating that the overall sensitivity to observation noise is similar for this system.

In contrast, for the Burgers equation, clearer differences emerge between the two approaches. Except for the per-sample fixed sensor regime (R2), PICSB consistently shows a more moderate increase in relative L^2 error as noise level increases, whereas guidance-based sampling exhibits a steeper error growth. This difference is especially pronounced in the randomly resampled sensor (R1) and globally fixed sensor regimes (R3), where the error of guidance-based sampling grows rapidly with increasing noise, while PICSB maintains a more gradual degradation.

For Darcy flow, the quantitative curves may initially suggest that guidance-based sampling is less sensitive to increasing noise levels, as its relative L^2 error remains nearly constant across noise magnitudes. However, this behavior should not be interpreted as robustness to noise. As illustrated by the qualitative results in Figures 19 and 20, guidance-based sampling fails to produce physically meaningful target samples even at low noise levels. In this setting, the LF prior for Darcy is constructed via nearest-neighbor interpolation from sparse HF observations, resulting in a discretized and piecewise-constant LF representation. When combined with guidance-based sampling, the noisy observation gradients are insufficient to effectively steer the generative process toward a valid HF solution. Consequently, the guidance-based method exhibits consistently high error regardless of the noise level, reflecting a fundamental limitation in leveraging the available observations rather than genuine robustness to noise.

In contrast, PICSB maintains lower reconstruction error with a gradual increase as noise grows, indicating that it is able to utilize noisy sparse observations more effectively through hard conditioning within the learned LF-to-HF transport, even in the presence of discretized LF inputs.

Overall, these results characterize reconstruction behavior under varying inference-time noise levels in sparse observations. As the noise level increases, the relative L^2 error increases for both PICSB

and guidance-based sampling across all datasets and observation regimes. However, the rate at which the error grows differs between the two approaches, depending on the system and observation setting.

Across many configurations, PICSB exhibits a more gradual increase in relative L^2 error as noise increases, whereas guidance-based sampling tends to show a steeper degradation, particularly at higher noise levels. This difference reflects how noisy sparse observations are incorporated during inference: PICSB enforces noisy measurements through hard conditioning within a learned LF-to-HF transport, while guidance-based sampling directly injects gradients derived from noisy observations, which can lead to stronger error amplification as noise increases.

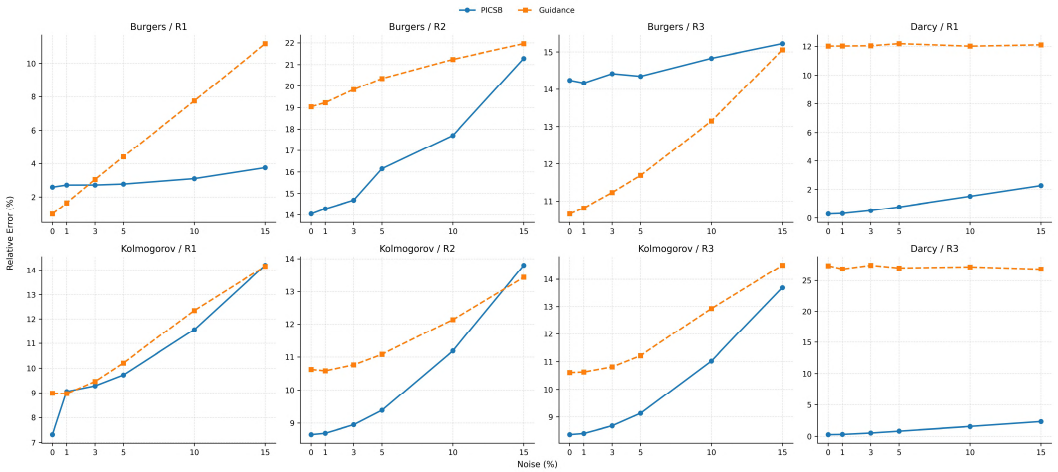

Figure 15: Relative L^2 error under noisy observations across Burgers, Darcy, and Kolmogorov observation regimes.

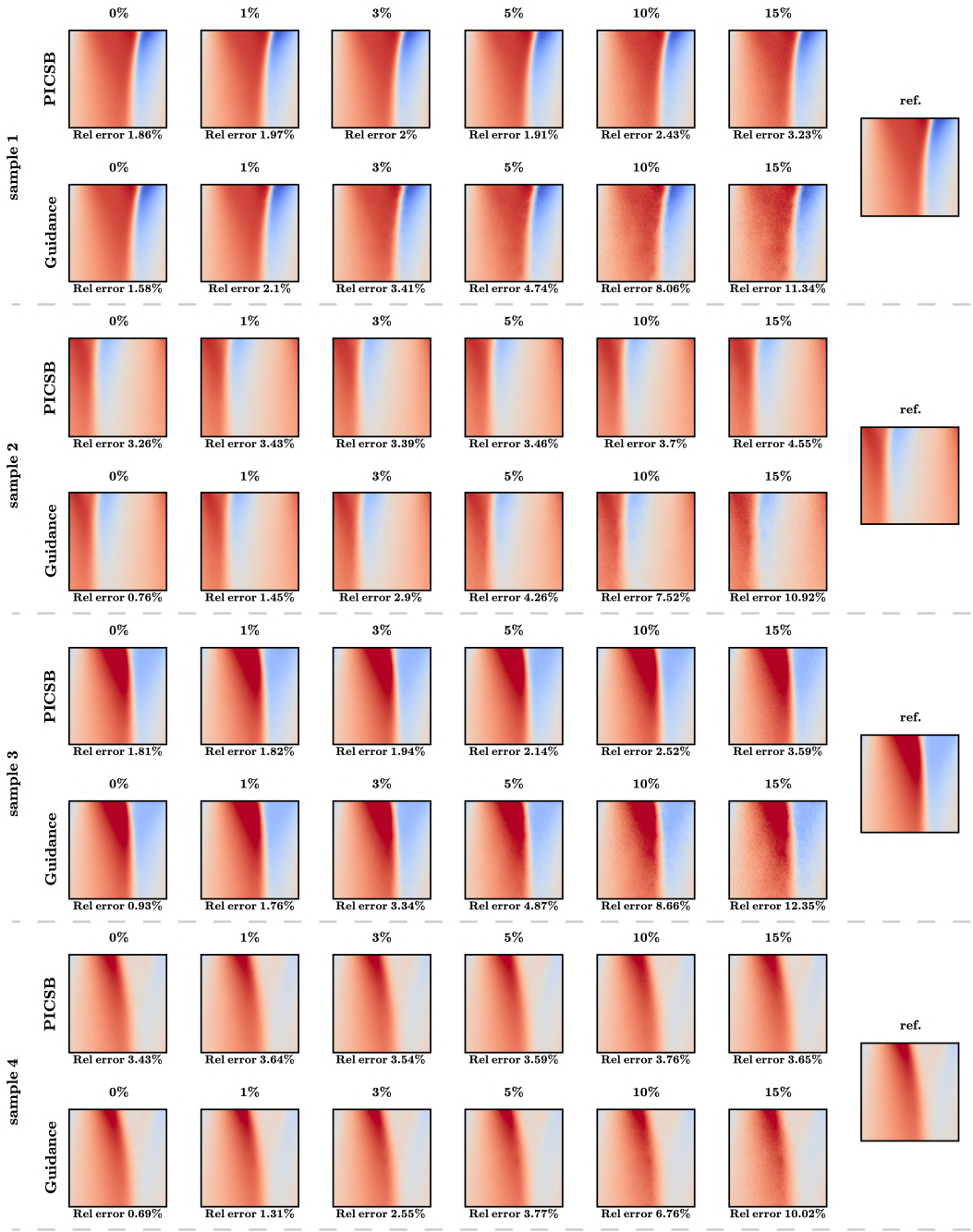

Figure 16: Inference under noisy observations for Burgers' equation with randomly resampled sensors (R1). Each row corresponds to a different test sample, and each column corresponds to a noise level, with the HF reference solution shown in the rightmost column. Relative L^2 errors are reported below each predicted field.

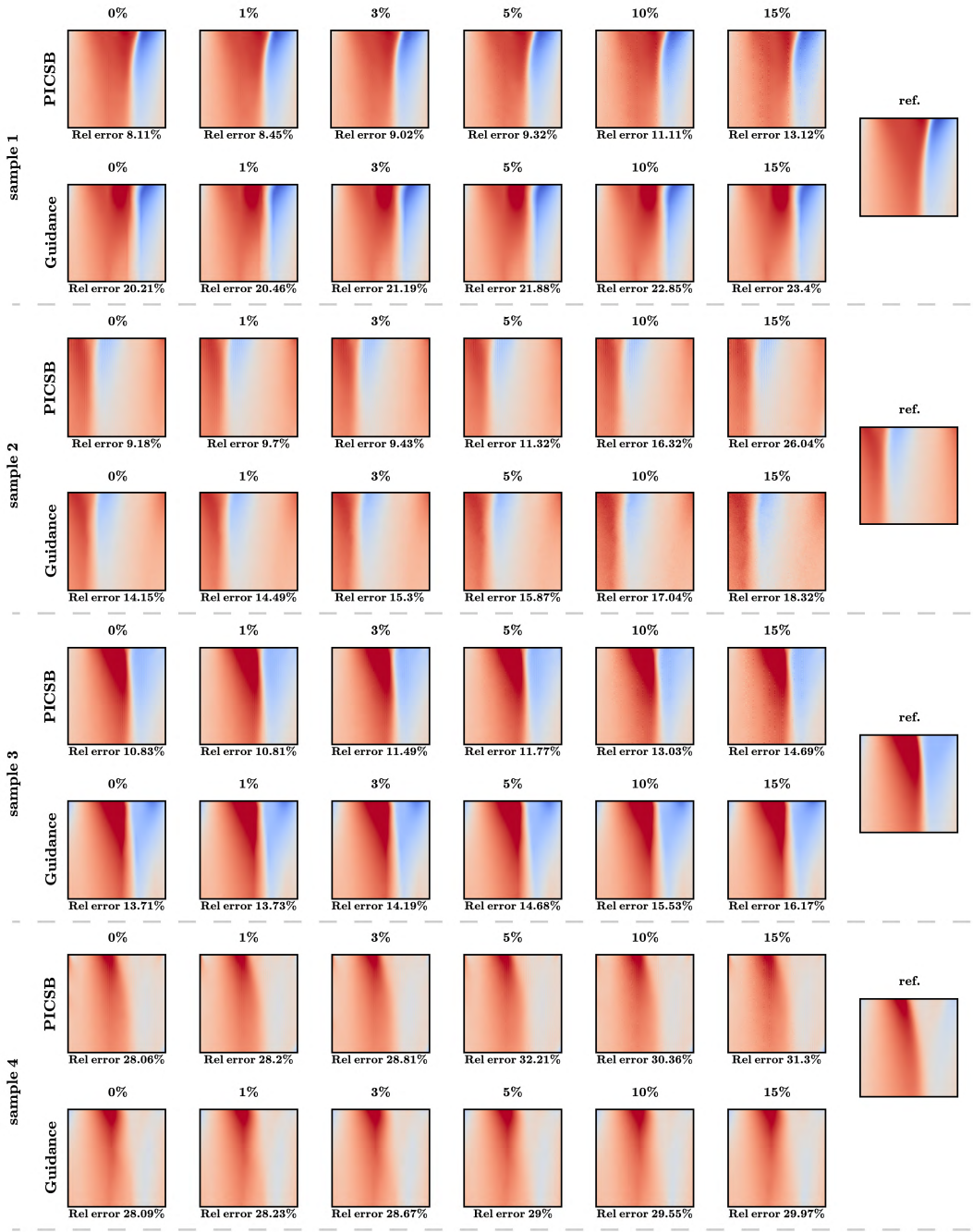

Figure 17: Inference under noisy observations for Burgers' equation with per-sample fixed sensors (R2). Each row corresponds to a different test sample, and each column corresponds to a noise level, with the HF reference solution shown in the rightmost column. Relative L^2 errors are reported below each predicted field.

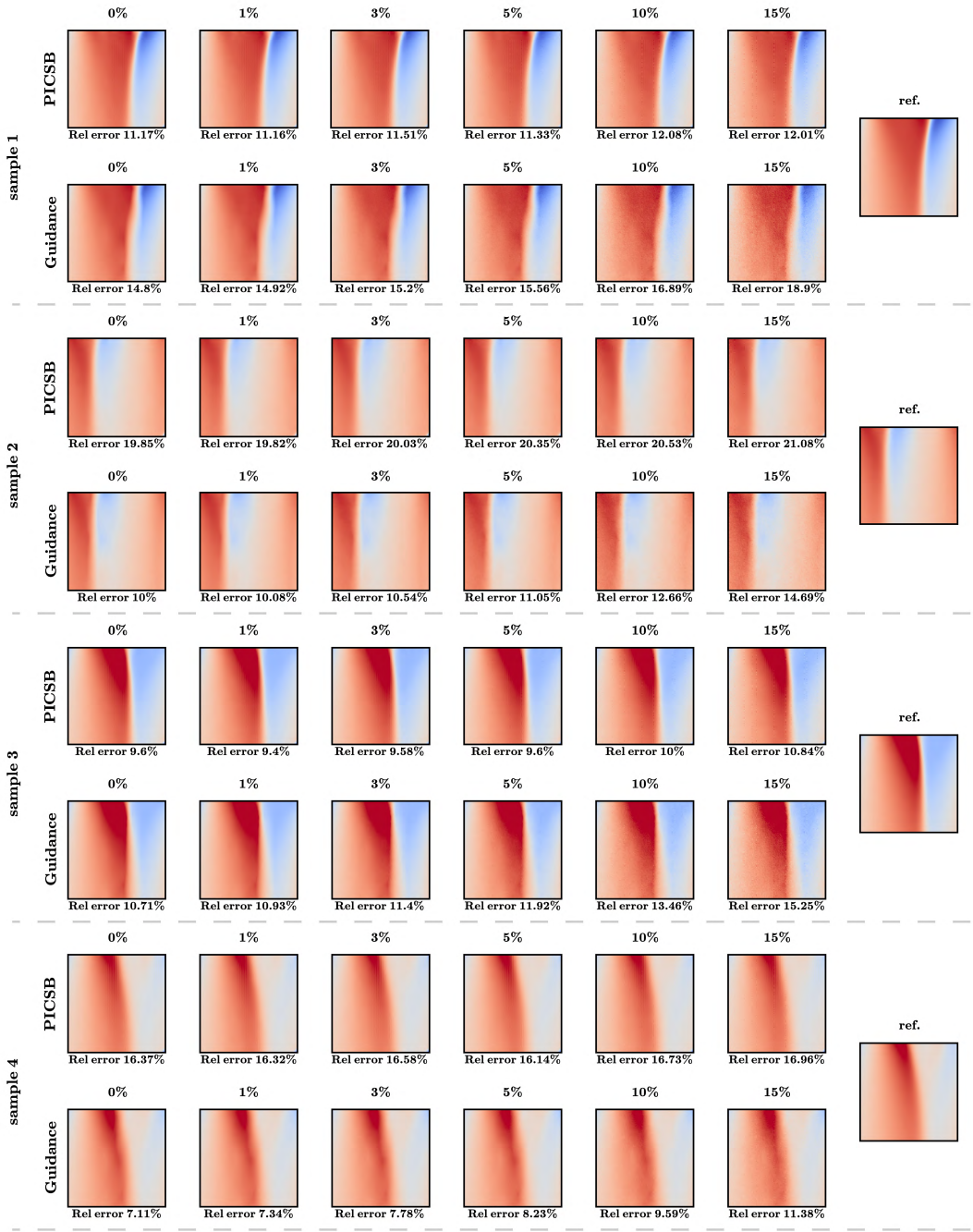

Figure 18: Inference under noisy observations for Burgers' equation with a fixed sensor set (R3). Each row corresponds to a different test sample, and each column corresponds to a noise level, with the HF reference solution shown in the rightmost column. Relative L^2 errors are reported below each predicted field.

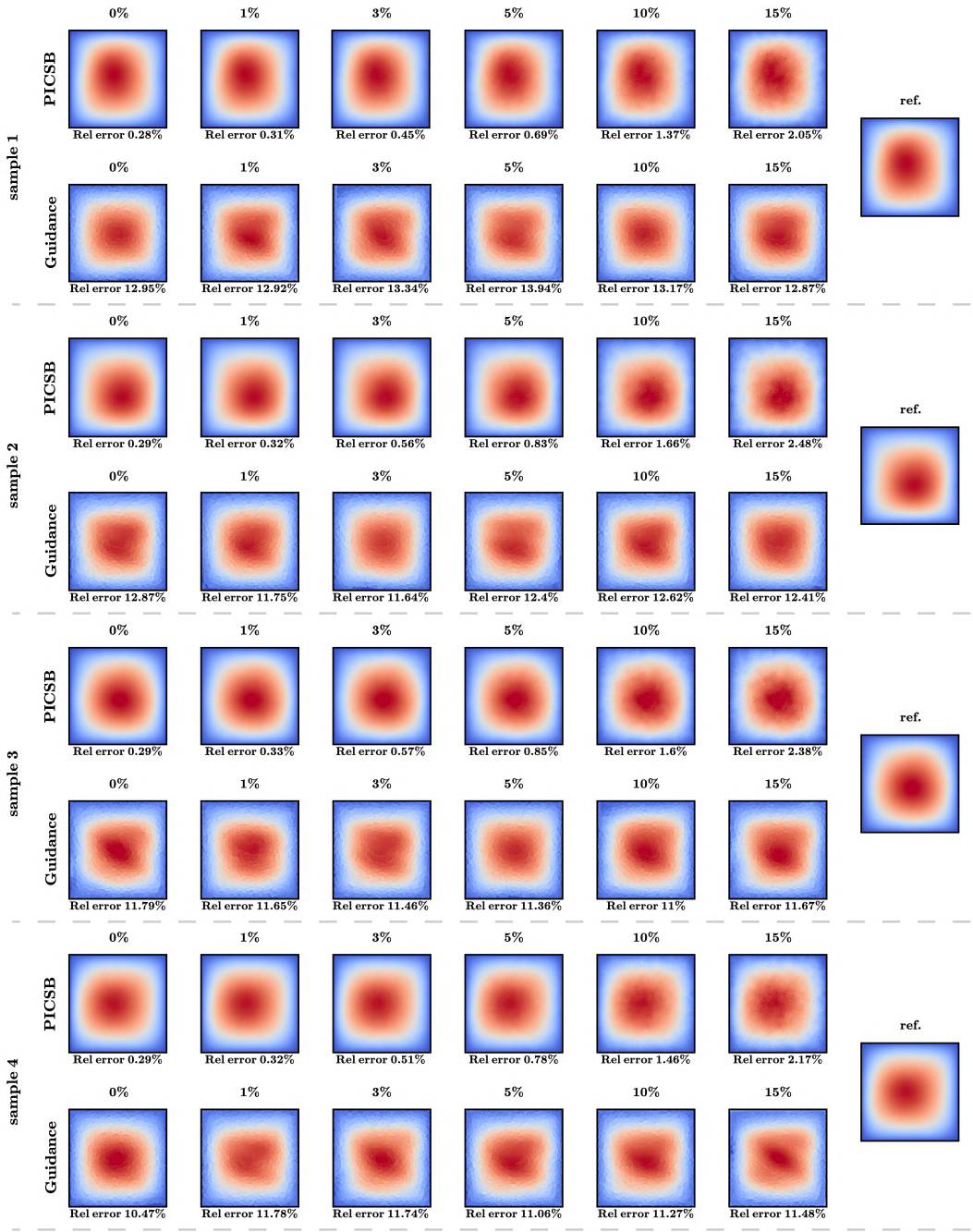

Figure 19: Inference under noisy observations for Darcy flow with randomly resampled sensors (R1). Each row corresponds to a different test sample, and each column corresponds to a noise level, with the HF reference solution shown in the rightmost column. Relative L^2 errors are reported below each predicted field.

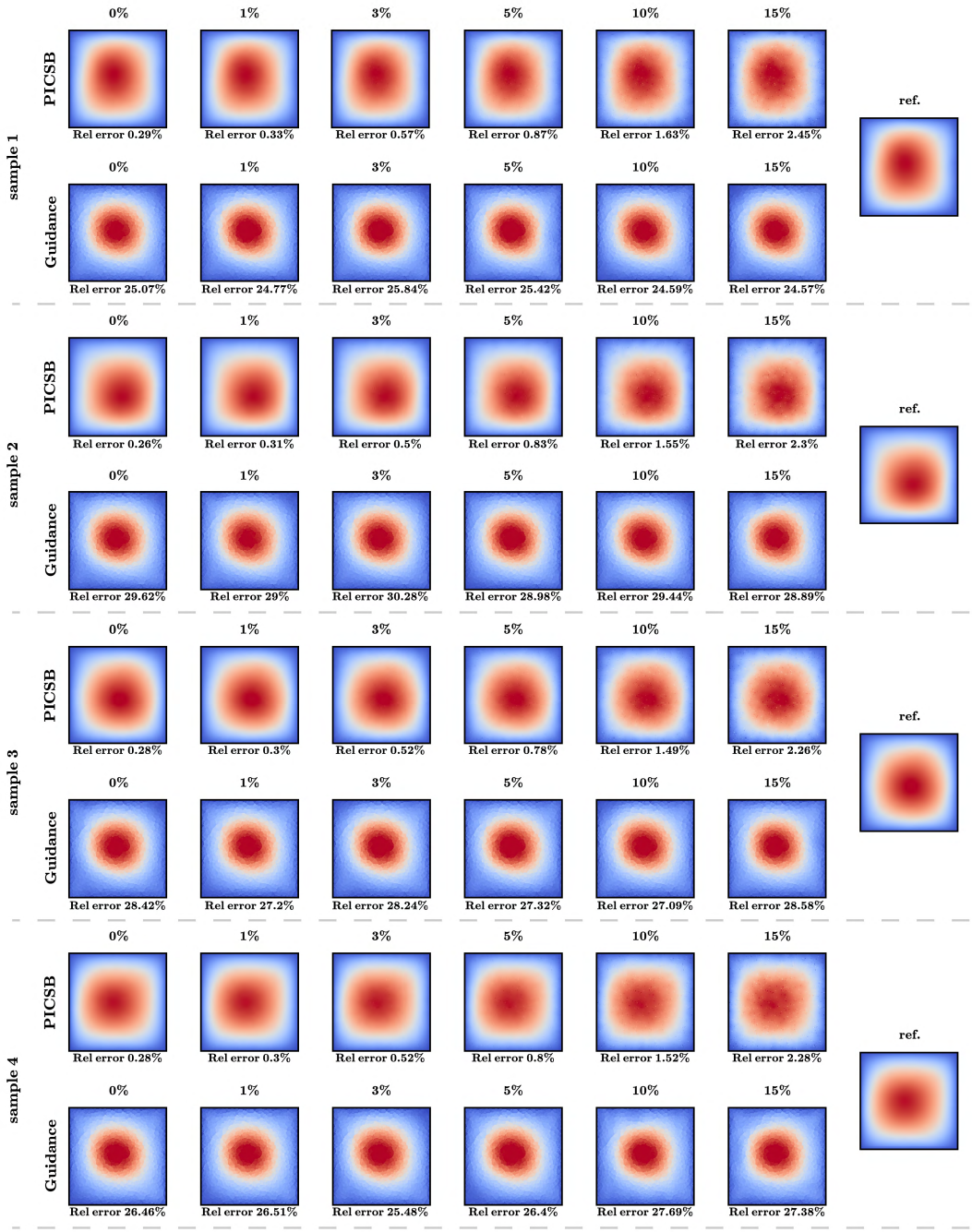

Figure 20: Inference under noisy observations for Darcy flow with a fixed sensor set (R3). Each row corresponds to a different test sample, and each column corresponds to a noise level, with the HF reference solution shown in the rightmost column. Relative L^2 errors are reported below each predicted field.



Figure 21: Inference under noisy sparse observations for Kolmogorov flow with randomly resampled sensors (R1). Each row corresponds to a different test sample, and each column corresponds to a noise level, with the HF reference solution shown in the rightmost column. Relative L^2 errors are reported below each predicted field.

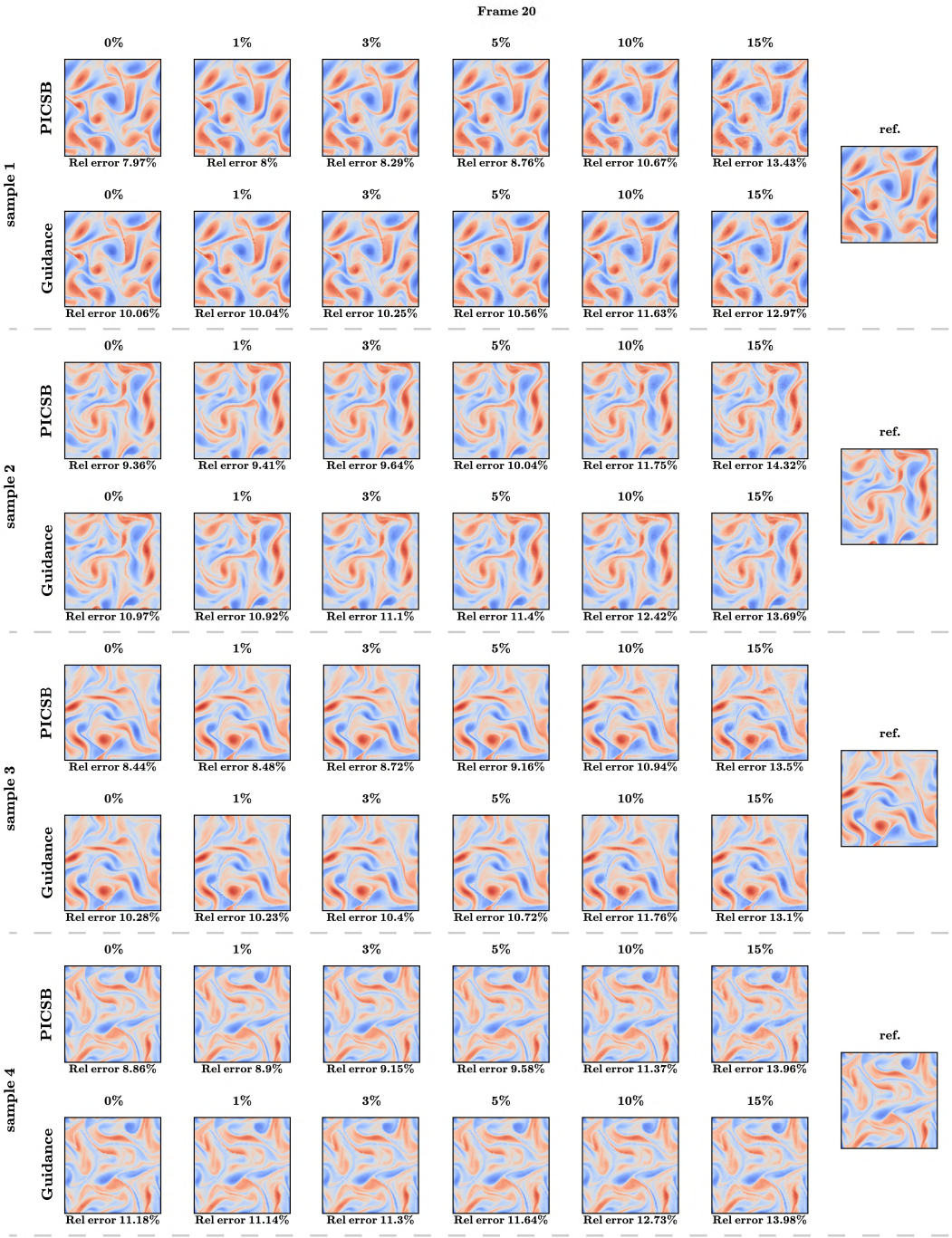

Figure 22: Inference under noisy observations for Kolmogorov flow with per-sample fixed sensors (R2). Each row corresponds to a different test sample, and each column corresponds to a noise level, with the HF reference solution shown in the rightmost column. Relative L^2 errors are reported below each predicted field.

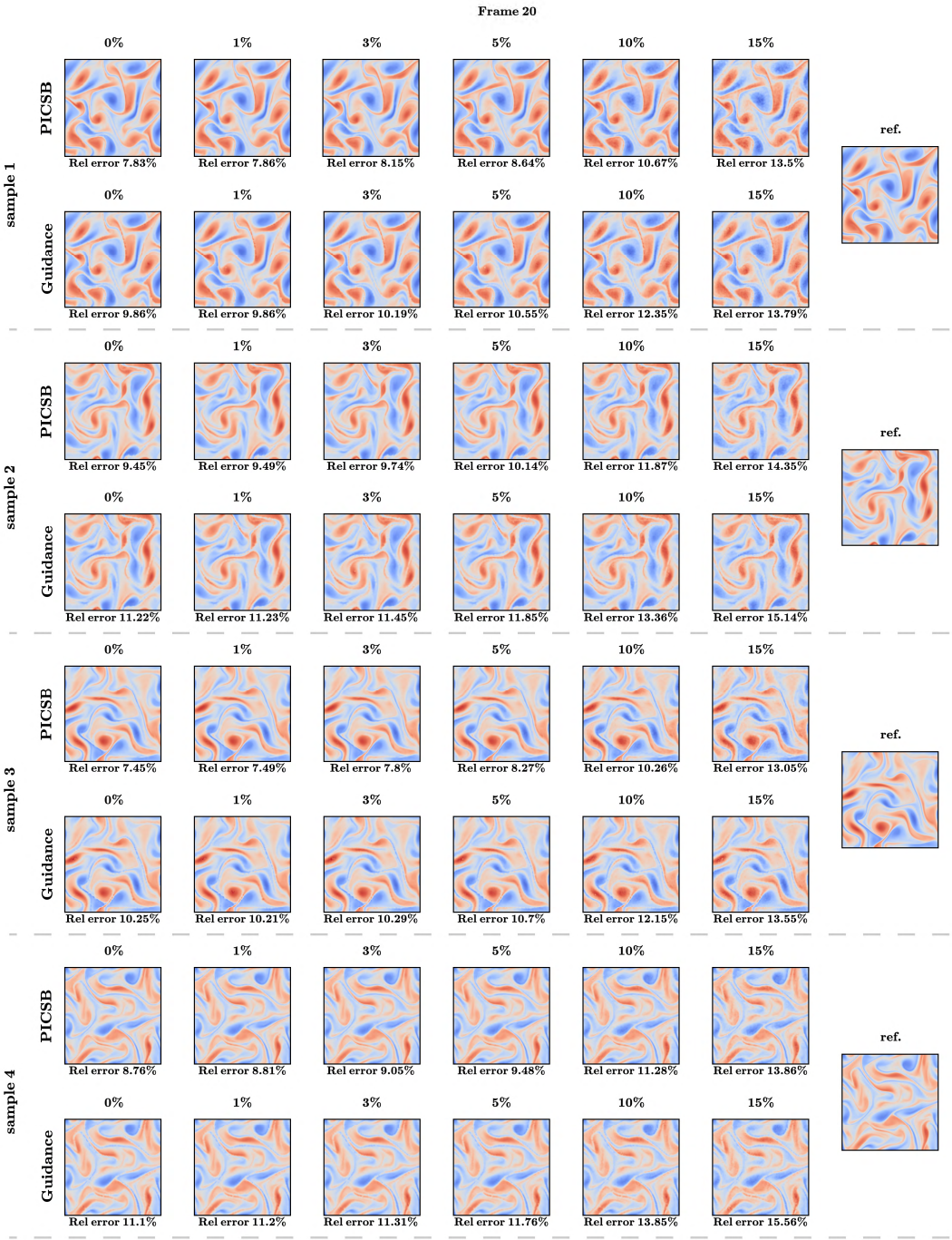

Figure 23: Inference under noisy observations for Kolmogorov flow with a fixed sensor set (R3). Each row corresponds to a different test sample, and each column corresponds to a noise level, with the HF reference solution shown in the rightmost column. Relative L^2 errors are reported below each predicted field.

C.8 Finetuning across Different Levels of Observation Sparsity

In Section C.3.3, we observed that directly applying models trained with 10% sparse observations to substantially sparser inference regimes (e.g., 1% or 5% of the full grid) leads to severe degradation. Inference under such extreme sparsity remains challenging in general when no additional adaptation is performed, particularly for more complex dynamics. To further investigate whether limited adaptation can mitigate this limitation, we conduct additional finetuning experiments across different levels of observation sparsity.

All models are first pretrained using 10% sparse HF observations, following the same configuration as in the main experiments. For Darcy and Kolmogorov, LF inputs are constructed from the 10% sparse HF observations via nearest-neighbor and bi-cubic interpolation, respectively, and used during training. For Burgers, the same LF inputs as in the main experiments are used.

Finetuning Regime 1. Starting from these pretrained models, we first consider finetuning under lower observation availability by subsampling the original 10% observation set. Specifically, we retain either 50% or 10% of the available observation points, corresponding to conditioning at the 5% and 1% observation levels, respectively. For Darcy and Kolmogorov, the resulting observation sets are used to construct the corresponding LF inputs using the same interpolation schemes as in training, while hard conditioning is imposed at the selected sparse locations. For Burgers, the LF inputs remain unchanged, and only the hard conditioning set is reduced accordingly.

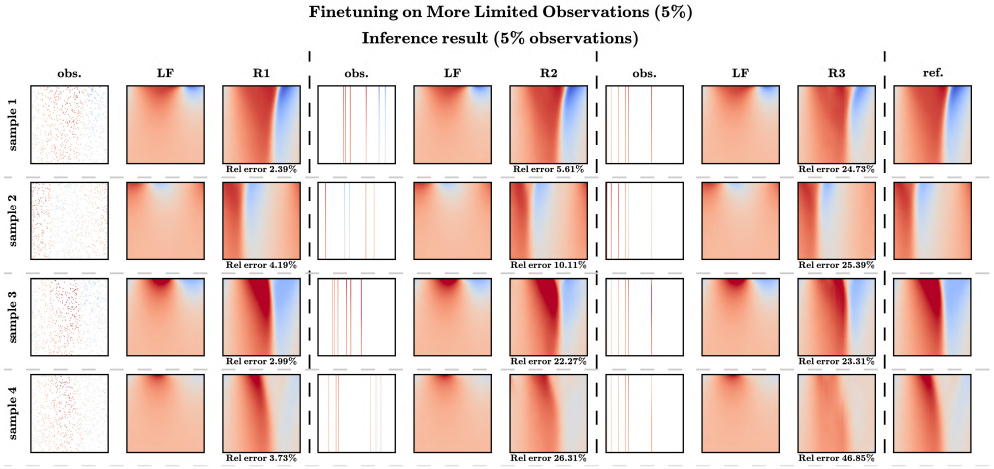

Figure 24: Finetuning and inference results under 5% sparse observations for the Burgers' equation.

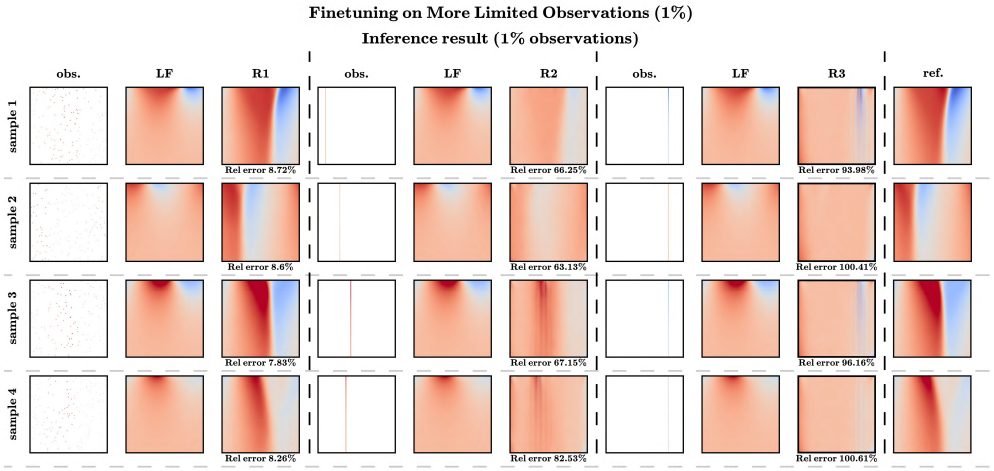

Figure 25: Finetuning and inference results under 1% sparse observations for the Burgers' equation.

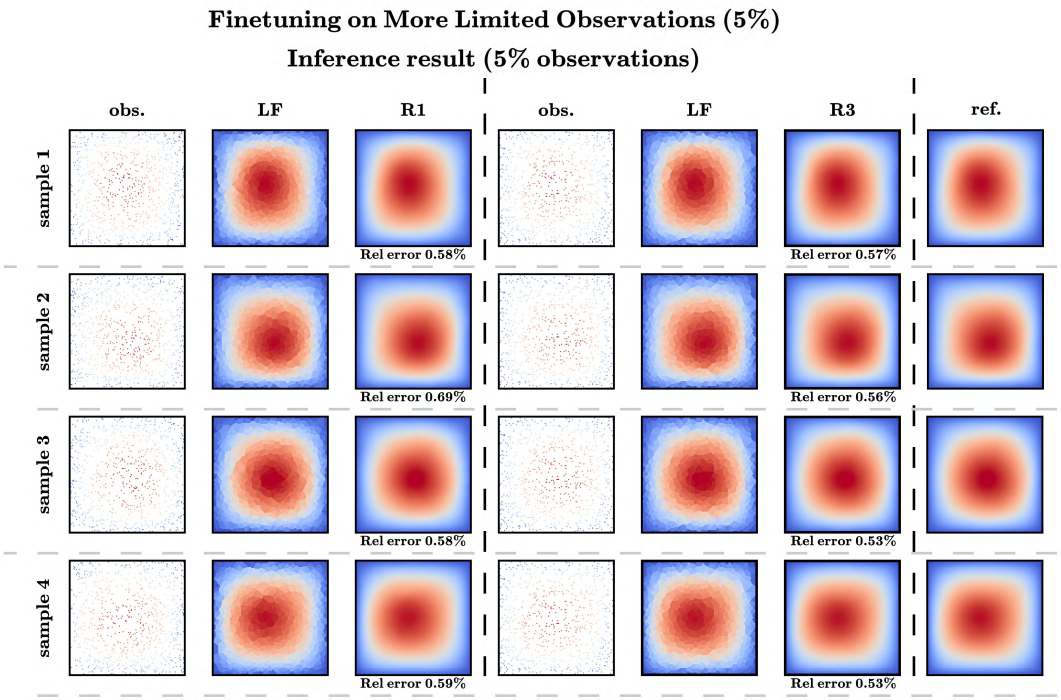

Figure 26: Finetuning and inference results under 5% sparse observations for Darcy flow.

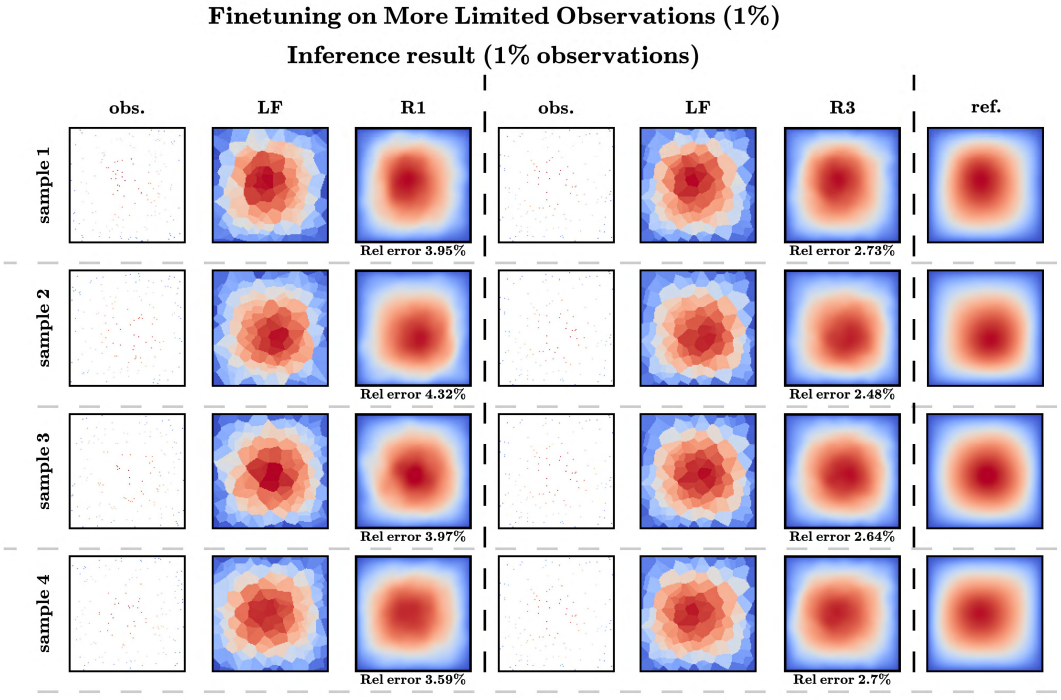

Figure 27: Finetuning and inference results under 1% sparse observations for Darcy flow.

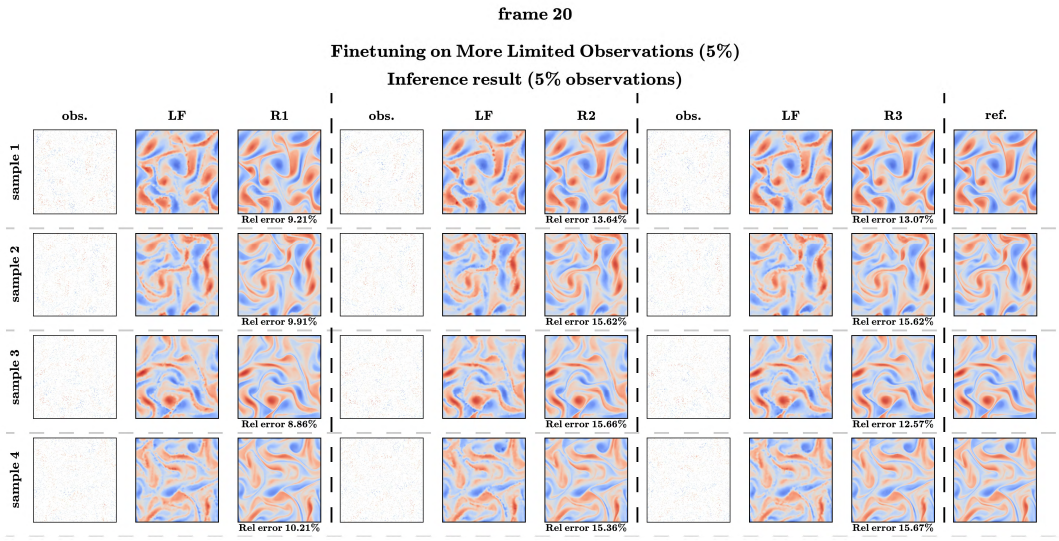

Figure 28: Finetuning and inference results under 5% sparse observations for the Kolmogorov flow.

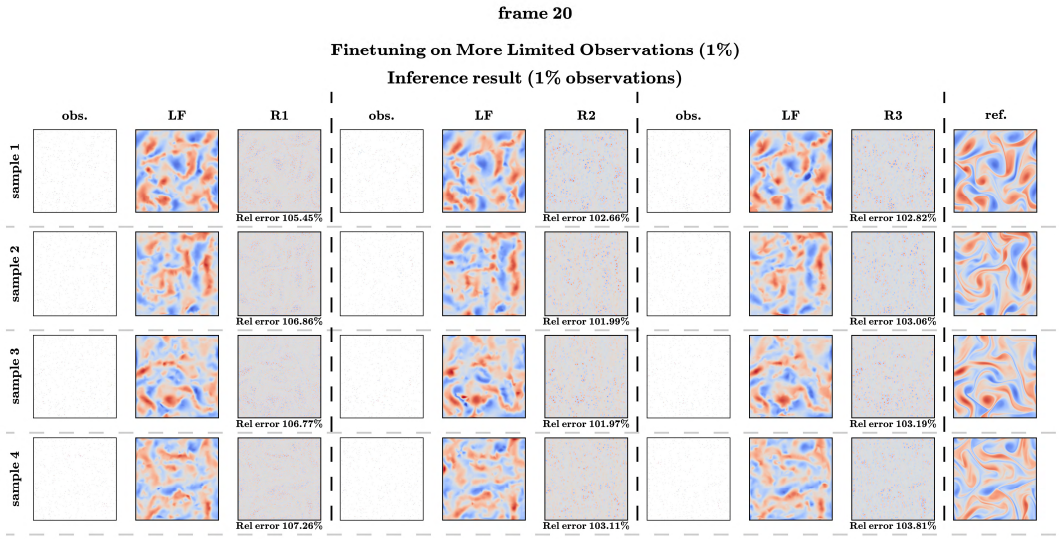

Figure 29: Finetuning and inference results under 1% sparse observations for the Kolmogorov flow.

Finetuning Regime 2. While the finetuning strategy in Regime 1 yields reasonable reconstructions at the 5% observation level, it fails to maintain performance when the observation level is reduced to 1% (see Figures 24–29). To further examine this limitation, we consider an alternative finetuning strategy aimed at improving robustness across different observation sparsity levels.

Starting from the same 10%-pretrained models, finetuning is performed using the same observation datasets as in Regime 1. However, rather than finetuning separately for each observation level, we adopt an *integrated mode* in which finetuning is carried out on a nested collection of observation subsets corresponding to 10%, 5%, and 1% of the full grid. In this setting, the finetuning dataset is formed by aggregating these nested sparsity levels, allowing the model to adapt simultaneously across different observation regimes.

The optimizer state and learning rate settings are identical to those used in Regime 1, with finetuning performed for the same total budget of $20{,}000$ iterations. After finetuning, inference is evaluated separately under 5% and 1% observation regimes.

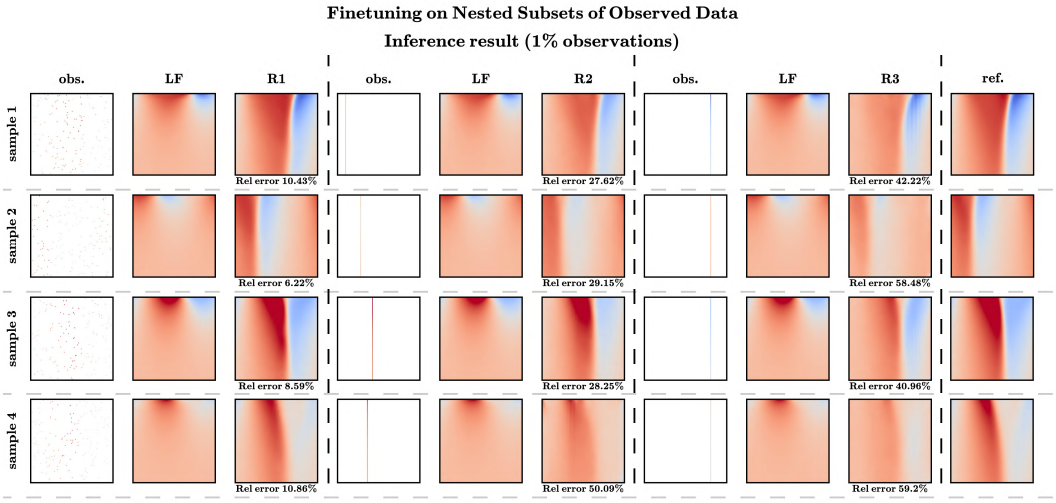

Figure 30: Finetuning and inference results under 1% sparse observations for the Burgers' equation with integrated finetuning.

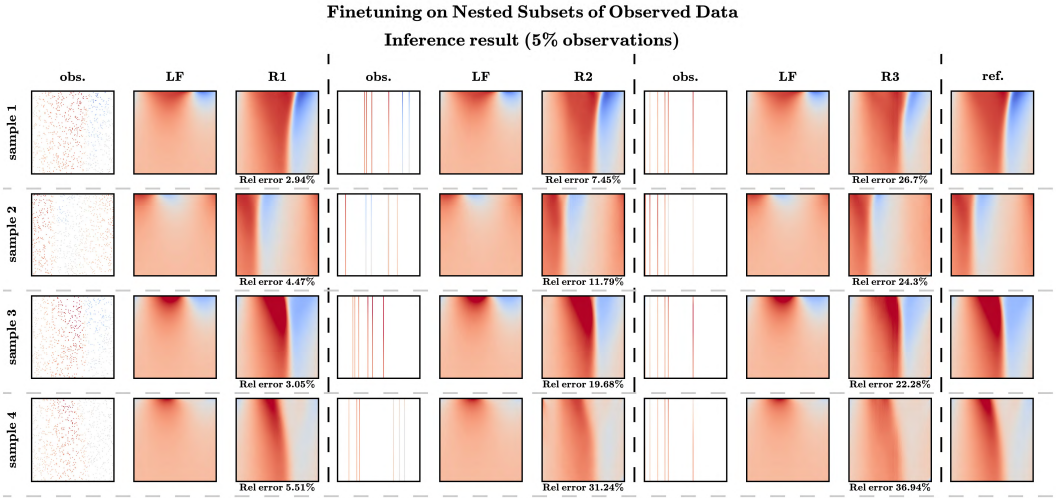

Figure 31: Finetuning and inference results under 5% sparse observations for the Burgers' equation with integrated finetuning.

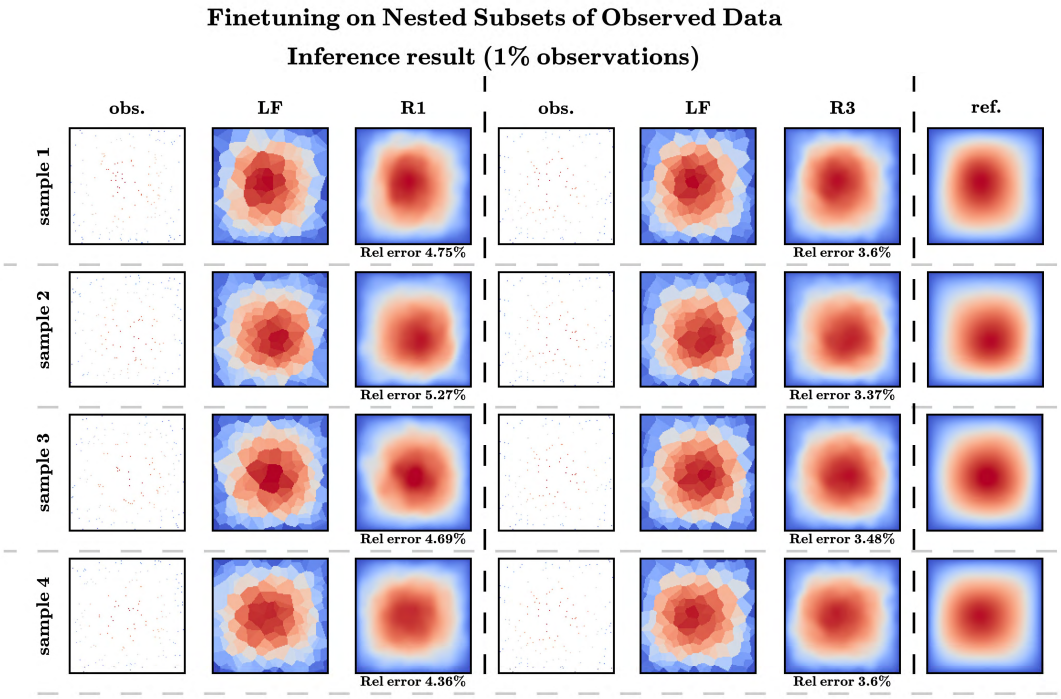

Figure 32: Finetuning and inference results under 1% sparse observations for Darcy flow with integrated finetuning.

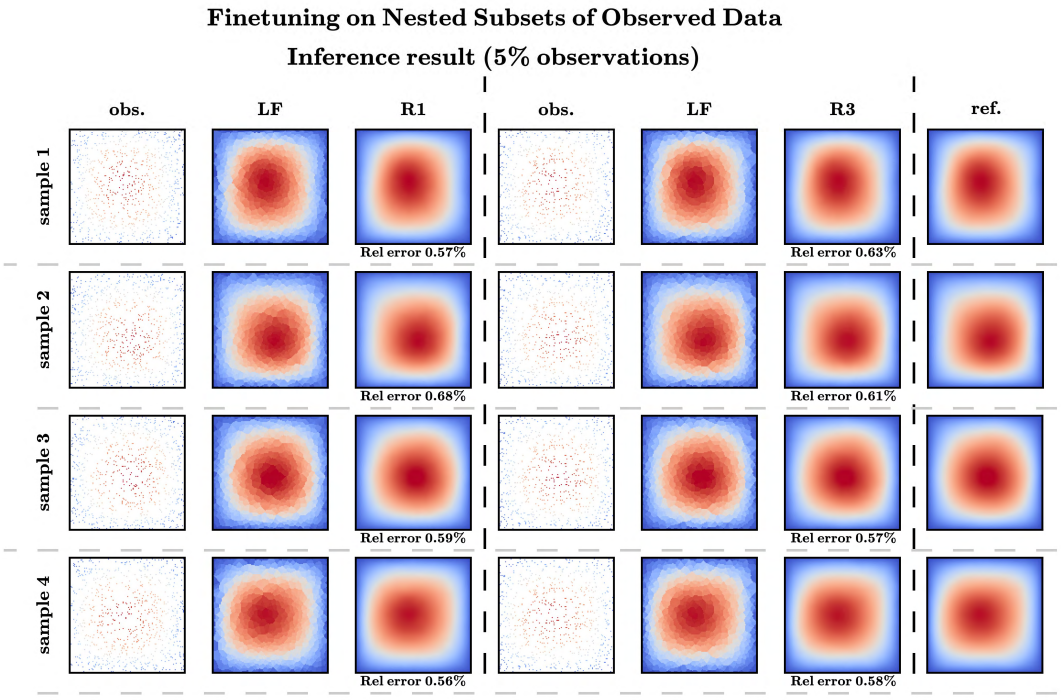

Figure 33: Finetuning and inference results under 5% sparse observations for Darcy flow with integrated finetuning.

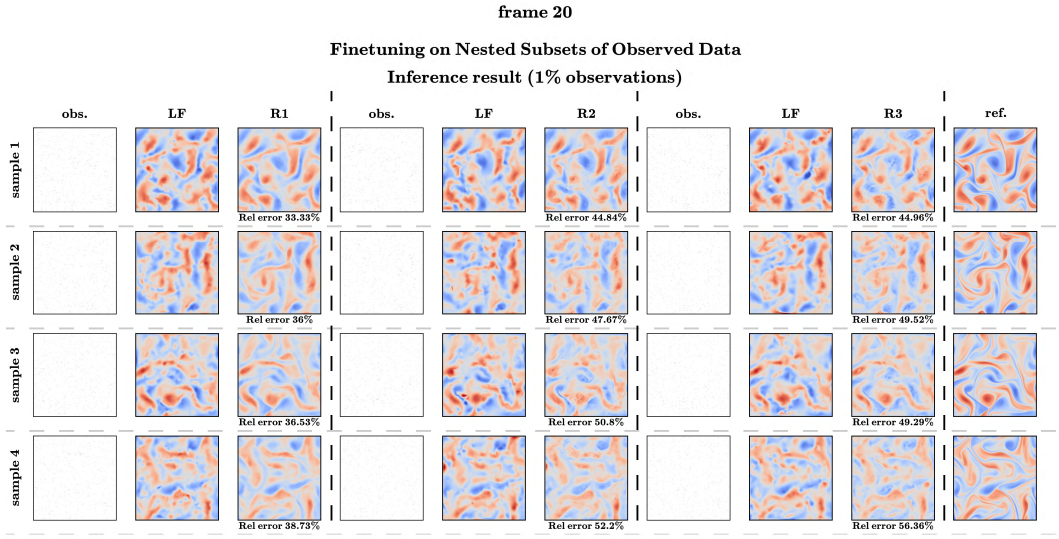

Figure 34: Finetuning and inference results under 1% sparse observations for the Kolmogorov flow with integrated finetuning.

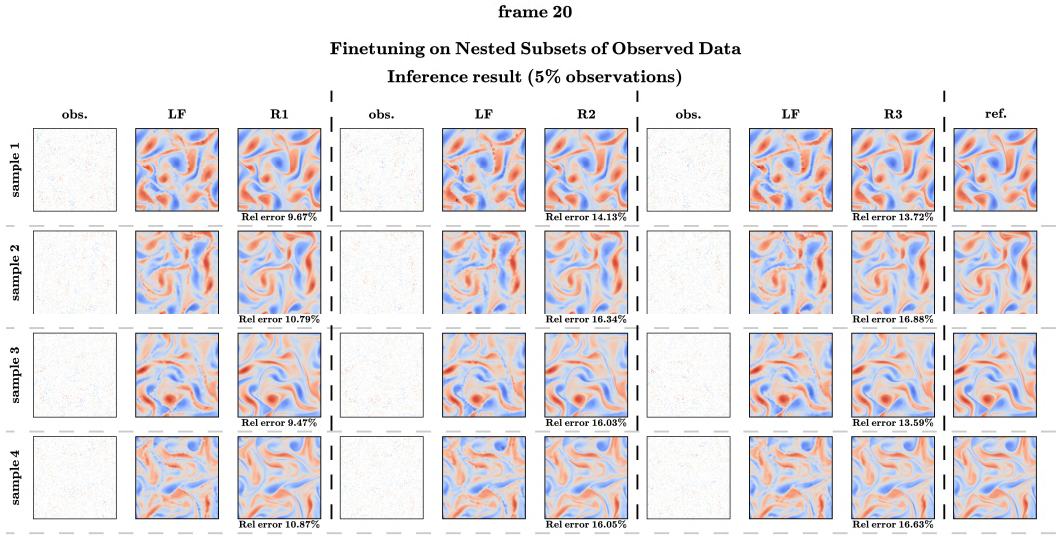

Figure 35: Finetuning and inference results under 5% sparse observations for the Kolmogorov flow with integrated finetuning.

As a result, Figures 31, 33, and 35 show that finetuning under the integrated mode maintains stable reconstruction quality at the 5% observation level. At the more challenging 1% observation level, Figures 30, 32, and 34 demonstrate that the integrated finetuning strategy yields improved robustness relative to finetuning performed separately for each observation level.

Although the reconstructions obtained at 1% sparsity remain imperfect and exhibit non-negligible errors, integrated finetuning enables meaningful recovery in scenarios where per-level finetuning largely fails. Overall, these results indicate that aggregating multiple observation sparsity levels during finetuning can partially alleviate the degradation caused by extreme observation scarcity, while substantial performance gaps persist in such severely underdetermined settings.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction (Section 1) clearly state the paper's main claims, which accurately reflect the contributions and scope of the work.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The main theorem states its assumptions in Sec. 4.2, and the full proof is provided in Appendix A.1. Additional theoretical statements and proofs are provided in Appendix A.2–A.3.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides the algorithmic procedures in Algorithms 1–2 and the experimental protocol in Sec. 5 and Appendix B, including datasets, baselines, metrics, and implementation details.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The datasets used in this paper are publicly available and credited to their original source papers, and the paper specifies the codebases that our implementation builds upon and provides instructions for reproducing the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 5 specifies the benchmark tasks, sparse-observation setting, train/test protocol, metrics, and baselines. Appendix B.1–B.2 provides dataset details, architectures, optimizers, training iterations, and hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 2 reports 95% confidence intervals over all test samples and marks overlap with the best-performing method. Appendix C.3 additionally reports variability over 40 repeated inference runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. 5.2 reports the GPU hardware and inference wall-time.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses PDE simulation benchmarks and does not involve human subjects, personal data, or high-risk deployment. The research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The paper discusses potential positive impacts for time-critical data assimilation applications such as weather, ocean, and hydrologic forecasting. We do not identify direct negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Not applicable, as we do not release any pretrained language or image generation models, and we do not distribute scraped datasets. All data used in this study are sourced from publicly available resources.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing datasets and baseline methods are credited in Appendix B.1–B.2 and the references, and their licenses and terms of use are documented in the accompanying release materials.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Not applicable, because the paper does not introduce a new dataset, benchmark, or standalone released model asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Not applicable, because the work does not involve crowdsourcing or human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: Not applicable, because the work does not involve crowdsourcing or human-subject experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLM assistance was limited to polishing the writing and had no influence on the core methodology, scientific validity, or originality of the work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.