DIFFUSION-BASED MULTI-FIDELITY SURROGATE MODELING VIA NOISE-LEVEL SUPERVISION ALLOCATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning-based surrogate models for physical fields offer an efficient alternative to repeated numerical simulations, but obtaining sufficient high-fidelity training data remains costly. Multi-fidelity modeling addresses this bottleneck by combining limited high-fidelity data with abundant low-fidelity data, often through correction mappings. Rather than learning a low-to-high correction, we propose a novel diffusion-based method that allocates multi-fidelity supervision across noise levels. At selected higher-noise stages, low-fidelity data substitute for part of the high-fidelity supervision. Our method is motivated by the observation that the discrepancy between noisy high- and low-fidelity distributions decreases as the diffusion noise level increases. For deterministic physical responses, we use the diffusion signal-to-noise ratio to quantify this distributional discrepancy and bound the resulting denoising-target bias. Multi-fidelity data are jointly used to train a single conditional diffusion model that predicts physical fields without low-fidelity inputs at inference. On two thermal-field benchmarks, our method reduces the prediction error by up to 90% relative to the single-fidelity model and achieves accuracy comparable to that of existing multi-fidelity models trained with 50 times as many high-fidelity samples.

1 INTRODUCTION

Learning-based surrogate models for physical fields provide an efficient alternative to repeated high-fidelity simulations in computational science and engineering. By learning nonlinear mappings from system parameters, operating conditions, or geometric descriptions to spatially distributed responses, they accelerate design optimization, uncertainty quantification, and real-time prediction. Advances in deep neural networks and neural operators have further improved their ability to represent complex spatial and temporal responses (Lu et al., 2021; Li et al., 2021). However, accurate surrogates typically require substantial high-fidelity training data, whose acquisition through numerical simulations or physical experiments can be costly. Multi-fidelity surrogate modeling addresses this bottleneck by combining limited high-fidelity data with abundant low-fidelity data (Howard et al., 2023). Low-fidelity simulations use coarser discretizations or simplified physical models to reduce the cost of data generation, allowing more input configurations to be sampled within a given computational budget.

The effectiveness of multi-fidelity learning depends on how discrepancies between fidelity levels are handled. A broad class of methods relates low-fidelity predictions to high-fidelity targets through residual correction, autoregressive formulations, or nonlinear mappings (Perdikaris et al., 2017; Meng & Karniadakis, 2020). Other approaches transfer representations through low-fidelity pretraining followed by high-fidelity fine-tuning (Zhang et al., 2023). Existing multi-fidelity strategies typically determine how low-fidelity information is transferred at the model or training-stage level, but do not distinguish whether the same low-fidelity supervision is equally reliable across different learning tasks. We therefore ask whether the usefulness of low-fidelity supervision can itself depend on the learning task. Diffusion models offer a natural setting for exploring this possibility by decomposing distribution learning into denoising tasks at different noise levels (Ho et al., 2020; Song et al., 2021).

Building on this perspective, we propose a diffusion-based multi-fidelity surrogate modeling method via Noise-Level Supervision Allocation (NLSA). Our method is motivated by the attenuation of

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

fidelity differences under Gaussian noise corruption. For high- and low-fidelity responses under the same physical conditions, the forward diffusion process progressively reduces the contribution of the clean responses while increasing the perturbation variance. Consequently, the discrepancy between the corresponding noisy conditional distributions decreases as the noise level increases. This means that an LF response that is too biased for a low-noise denoising task can still provide useful supervision for a higher-noise task. In other words, appropriately selected LF data can serve as a substitute for high-fidelity supervision at sufficiently high noise levels. A small high-fidelity dataset anchors the target response, and both fidelity datasets jointly train a single conditional denoiser. By using noise-dependent fidelity discrepancy as a basis for supervision allocation, NLSA enables inexpensive samples to contribute directly to surrogate learning, with the aim of reducing the number of high-fidelity samples required to achieve a given predictive accuracy.

We provide a theoretical analysis linking fidelity discrepancies to denoising supervision. For deterministic physical responses under the same physical conditions, we derive an exact expression for the conditional KL divergence between their noise-perturbed distributions in terms of the diffusion signal-to-noise ratio and the squared discrepancy between the clean responses. Under the noise-prediction objective, we further bound the deviation between the optimal noise predictor under mixed-fidelity supervision and its high-fidelity counterpart. The conditional KL and the upper bound on squared target deviation decrease with the diffusion SNR, providing a quantitative rationale for restricting low-fidelity supervision to higher-noise denoising tasks.

Experiments on two thermal-field benchmarks (Chen et al., 2021) show that NLSA substantially improves prediction accuracy in the severe high-fidelity-scarcity regime and can match or outperform the single-fidelity model and other multi-fidelity model using far fewer HF training samples. A supervision-boundary ablation and a visualization of noise-corrupted fields further examine the mechanism behind this allocation.

The contributions of this work are threefold:

1. **Noise-level supervision allocation.** We formulate multi-fidelity diffusion learning as the allocation of supervision across denoising tasks with different noise levels. NLSA assigns LF supervision to higher-noise tasks where cross-fidelity discrepancies are attenuated, thereby leveraging inexpensive LF observations to supplement scarce HF supervision.
2. **Noise-dependent characterization of cross-fidelity discrepancy.** We show that diffusion progressively attenuates the discrepancy between noisy HF and LF conditional distributions. For deterministic responses, we further establish a noise-dependent bound on the denoising-target deviation induced by mixed-fidelity supervision.
3. **High-fidelity data efficiency.** Experiments on two thermal-field benchmarks demonstrate substantial error reductions over single-fidelity and multi-fidelity baselines at small high-fidelity data budgets. The supervision-boundary ablation further examines the effect of noise-level allocation.

2 Related Work

Physical-field prediction. Learning-based physical-field prediction provides an efficient alternative to repeated numerical simulation by learning mappings from physical conditions to spatially distributed responses. Grid-based convolutional models have been widely used when physical inputs and outputs admit image-like representations, including temperature-field prediction benchmarks (Chen et al., 2021). Neural operators, such as DeepONet and the Fourier Neural Operator, instead learn mappings between function spaces and have demonstrated strong performance across parametric PDE problems (Lu et al., 2021; Li et al., 2021; Kovachki et al., 2023). Graph- and mesh-based neural simulators further extend learned physical prediction to irregular discretizations (Pfaff et al., 2021; Brandstetter et al., 2022). More recently, diffusion models have been explored for physical-field reconstruction and PDE prediction (Shu et al., 2023; Lippe et al., 2023; Shysheya et al., 2024; Huang et al., 2024). These approaches demonstrate the effectiveness of learning-based models for physical fields, but primarily focus on a single target fidelity. In diffusion training more broadly, noise schedules and SNR-based objectives have been studied to characterize the contribution of different noise levels within a single data distribution (Kingma et al., 2021; Chen, 2023). In contrast, we

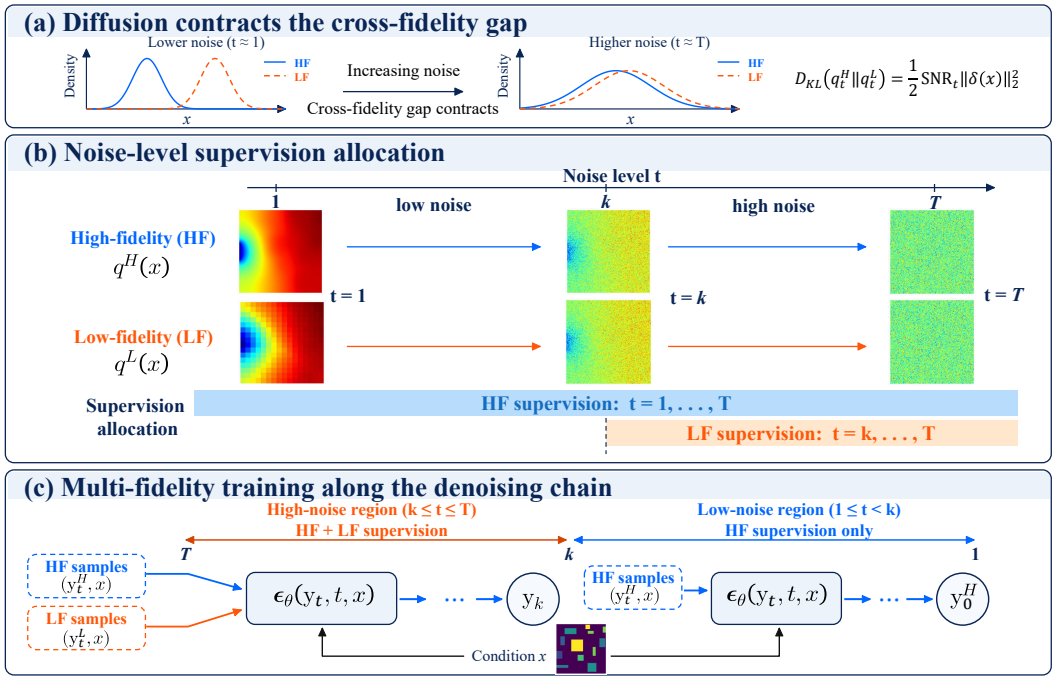


Figure 1: Overview of Noise-Level Supervision Allocation (NLSA). (a) Increasing noise reduces the diffusion SNR and attenuates the discrepancy between noisy HF and LF conditional distributions; the density curves are schematic. (b) HF samples supervise the complete diffusion chain, while LF samples provide supplementary supervision for $t \geq k$. (c) Along the denoising chain, multi-fidelity noisy samples are used to train a shared conditional denoiser with the same noise-prediction objective.

investigate how supervision from multiple fidelity levels can be selectively allocated across diffusion noise levels.

Multi-fidelity surrogate modeling. Multi-fidelity modeling combines limited expensive high-fidelity data with cheaper approximations to reduce surrogate construction cost. Classical approaches model correlations across fidelity levels, for example through autoregressive Gaussian-process formulations (Kennedy & O'Hagan, 2000), while broader multifidelity strategies have been developed for uncertainty propagation, inference, and optimization (Peherstorfer et al., 2018). Learning-based methods extend this idea through nonlinear information fusion and correction mappings (Perdikaris et al., 2017; Meng & Karniadakis, 2020), multi-fidelity operator learning (Howard et al., 2023), and low-fidelity pretraining followed by high-fidelity adaptation (Zhang et al., 2023). In contrast, NLSA treats fidelity as a supervision source indexed by diffusion noise level: LF observations supplement selected denoising tasks during training but are not required as conditioning inputs at inference.

3 Method

We develop a multi-fidelity surrogate modeling framework using a conditional diffusion model and allocate supervision according to the noise levels of its forward process. We first introduce the conditional diffusion formulation, then show that diffusion progressively contracts the cross-fidelity discrepancy, and finally use this property to formulate and analyze the NLSA training objective. Figure 1 provides an overview of this framework.

3.1 Problem Setup and Conditional Diffusion

We consider deterministic physical responses at two fidelity levels, $F_{\mathrm{H}}, F_{\mathrm{L}} : \mathcal{X} \to \mathbb{R}^d$, with datasets $\mathcal{D}_f = \{(\mathbf{x}_i^f, \mathbf{y}_i^f)\}_{i=1}^{N_f}$, where $\mathbf{y}_i^f = F_f(\mathbf{x}_i^f)$ and $f \in \{\mathrm{H}, \mathrm{L}\}$. High-fidelity (HF) simulations are expensive, yielding $N_{\mathrm{H}} \ll N_{\mathrm{L}}$. Our goal is to learn a conditional diffusion surrogate model for F_{H} by leveraging scarce HF observations together with abundant low-fidelity (LF) observations.

For a clean response $\mathbf{y}_0$, the forward diffusion process produces a noisy state at timestep $t \in \{1, \dots, T\}$ as

$$\mathbf{y}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{y}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\{\beta_t\}_{t=1}^{T}$, with $0 < \beta_t < 1$, is a prescribed noise schedule. A conditional network $\boldsymbol{\epsilon}_\theta(\mathbf{y}_t, t, \mathbf{x})$ predicts the injected noise using

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{y}_t, t, \mathbf{x})\|_2^2\right]. \tag{2}$$

This formulation decomposes diffusion learning into denoising tasks at different noise levels. We next examine how diffusion corruption attenuates fidelity discrepancies and use this relationship to determine where LF observations can supplement the scarce HF supervision.

3.2 Diffusion Contracts the Cross-Fidelity Gap

The main difficulty in exploiting low-fidelity data is their systematic bias with respect to the target high-fidelity response. While abundant LF simulations could in principle compensate for the missing HF supervision, directly using them throughout the diffusion process may introduce fidelity-dependent bias, particularly at low noise levels where detailed physical structures remain prominent.

The denoising-task decomposition in Section 3.1 suggests a natural way to revisit this issue: rather than asking whether LF data are globally accurate approximations of HF responses, we ask whether their reliability changes across diffusion noise levels. More specifically, *how does the discrepancy between HF and LF responses evolve under the forward diffusion process?* Figure 1(a) illustrates the central hypothesis: as the diffusion noise level increases, the discrepancy between the noisy HF and LF distributions is progressively attenuated. We now characterize this effect quantitatively.

For a fixed physical condition $\mathbf{x}$, define the clean cross-fidelity discrepancy as

$$\boldsymbol{\delta}(\mathbf{x}) = F_{\text{H}}(\mathbf{x}) - F_{\text{L}}(\mathbf{x}). \tag{3}$$

Applying the same forward diffusion process in Eq. (1) to both fidelities gives

$$q_t^f(\mathbf{y}_t \mid \mathbf{x}) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}F_f(\mathbf{x}), (1-\bar{\alpha}_t)\mathbf{I}\big), \qquad f \in \{\text{H}, \text{L}\}. \tag{4}$$

The two fidelity-specific distributions therefore share the same covariance and differ only in their means. Their Kullback–Leibler divergence admits a closed-form expression.

Proposition 1 (SNR-scaled cross-fidelity discrepancy). For deterministic HF and LF responses corrupted by the same Gaussian forward process,

$$D_{\text{KL}}\left(q_t^{\text{H}} \| q_t^{\text{L}}\right) = \frac{1}{2}\,\text{SNR}_t\,\|\boldsymbol{\delta}(\mathbf{x})\|_2^2, \tag{5}$$

where

$$\text{SNR}_t = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}. \tag{6}$$

Proposition 1 separates the cross-fidelity discrepancy into two factors: the intrinsic physical mismatch $\|\boldsymbol{\delta}(\mathbf{x})\|_2^2$, which is determined by the fidelity pair, and the diffusion-dependent factor SNR_t, which is determined solely by the noise level. Under a standard forward diffusion schedule, $\bar{\alpha}_t$ decreases with t, and therefore SNR_t also decreases monotonically. Consequently, for $t_2 > t_1$,

$$D_{\text{KL}}\left(q_{t_2}^{\text{H}} \| q_{t_2}^{\text{L}}\right) \le D_{\text{KL}}\left(q_{t_1}^{\text{H}} \| q_{t_1}^{\text{L}}\right). \tag{7}$$

Equation (7) makes the contraction effect explicit: the relative attenuation of the cross-fidelity gap along the diffusion chain is scaled entirely by the diffusion SNR, whereas the underlying fidelity discrepancy determines only its overall magnitude. As the forward process moves toward lower-SNR states,

$$D_{\text{KL}}\left(q_t^{\text{H}} \| q_t^{\text{L}}\right) \to 0 \qquad \text{as } \text{SNR}_t \to 0. \tag{8}$$

Importantly, this does not imply that the LF response becomes physically more accurate. The clean discrepancy $\boldsymbol{\delta}(\mathbf{x})$ remains unchanged. What diffusion changes is the distinguishability of the two

fidelities in the noisy space on which the denoising tasks are defined. At low noise levels, fidelity-specific structures remain pronounced, and LF observations may therefore induce substantial bias. At higher noise levels, the same physical discrepancy is increasingly attenuated relative to the injected Gaussian noise, making HF and LF noisy responses progressively better aligned.

This observation provides the key motivation for NLSA. The available HF observations should remain active throughout the entire diffusion chain, since they define the target high-fidelity process. In contrast, LF observations should not be used uniformly across all denoising tasks. Instead, they are introduced only in the higher-noise region, where the cross-fidelity discrepancy has been sufficiently contracted by diffusion.

From this perspective, HF observations supervise the complete diffusion chain. LF observations supplement the additional HF supervision that is unavailable because high-fidelity simulations are costly, and contribute only at selected higher-noise levels. The remaining question is therefore *from which noise level onward their induced discrepancy becomes sufficiently attenuated to justify supplementary supervision*. This motivates the NLSA objective introduced next.

3.3 Noise-Level Supervision Allocation

Motivated by the contraction property above, we introduce a boundary $k \in \{1, \dots, T\}$ that determines when LF supervision is activated. As illustrated in Figure 1(b), the available HF samples are retained over the entire diffusion chain, whereas LF observations provide supplementary supervision only for $t \geq k$. Under a fixed diffusion schedule, the timestep boundary k can equivalently be viewed as a discrete SNR boundary: LF supervision is activated only for states satisfying $\mathrm{SNR}_t \leq \mathrm{SNR}_k$. For fidelity $f \in \{\mathrm{H}, \mathrm{L}\}$, define the timestep-specific denoising risk as

$$\mathcal{R}_{f,t}(\theta) = \mathbb{E}_{\substack{(\mathbf{x},\mathbf{y}^f)\sim\mathcal{D}_f \\ \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta \left(\sqrt{\bar{\alpha}_t}\mathbf{y}^f + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t, \mathbf{x} \right) \right\|_2^2 \right]. \tag{9}$$

The NLSA training objective is

$$\mathcal{L}_{\mathrm{MF}}(\theta; k) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{R}_{\mathrm{H},t}(\theta) + \frac{1}{T-k+1}\sum_{t=k}^{T} \mathcal{R}_{\mathrm{L},t}(\theta) \tag{10}$$

where the first term anchors the reverse process to the target HF system over the complete noise range, while the second supplements higher-noise denoising tasks with abundant LF observations. Both terms are normalized over their respective timestep ranges and optimized under the same noise-prediction objective, with equal weighting between the HF and LF branches.

Equivalently, Eq. (10) can be optimized by sampling

$$t_{\mathrm{H}} \sim \mathrm{Unif}\{1, \dots, T\}, \qquad t_{\mathrm{L}} \sim \mathrm{Unif}\{k, \dots, T\}, \tag{11}$$

and jointly minimizing the corresponding HF and LF denoising losses with a shared conditional network. Figure 1(c) illustrates this multi-fidelity model training scheme: the same conditional denoiser receives HF supervision across all noise levels and additional LF supervision only in the higher-noise region.

The boundary k controls the trade-off between supervision coverage and cross-fidelity bias. A smaller k allows LF observations to supervise more denoising tasks, but also introduces them at larger SNRs where the cross-fidelity discrepancy is stronger. A larger k restricts LF supervision to more strongly corrupted states, tightening the target-bias bound at the expense of a smaller supervised noise range.

The same noise-dependent structure also emerges at the denoising-target level. For a fixed noisy state $\mathbf{y}_t$ and condition $\mathbf{x}$, the fidelity-specific target is

$$\boldsymbol{\epsilon}_f^\star(\mathbf{y}_t, t, \mathbf{x}) = \frac{\mathbf{y}_t - \sqrt{\bar{\alpha}_t}F_f(\mathbf{x})}{\sqrt{1-\bar{\alpha}_t}}. \tag{12}$$

Using Eq. (3), we obtain

$$\|\boldsymbol{\epsilon}_{\mathrm{L}}^\star - \boldsymbol{\epsilon}_{\mathrm{H}}^\star\|_2^2 = \mathrm{SNR}_t \, \|\boldsymbol{\delta}(\mathbf{x})\|_2^2 . \tag{13}$$

Thus, diffusion contracts not only the discrepancy between the noisy HF and LF distributions, but also the discrepancy between their corresponding denoising targets. This target-level discrepancy

also extends to mixed-fidelity training. Under the loss in Eq. (9), the minimizer of the expected mixed-fidelity denoising loss satisfies the following bound.

Corollary 1 (SNR-bounded denoising bias). For $t \geq k$,

$$\|\boldsymbol{\epsilon}^{\star}_{\mathrm{mix}} - \boldsymbol{\epsilon}^{\star}_{\mathrm{H}}\|_2^2 \leq \mathrm{SNR}_t \, \|\boldsymbol{\delta}(\mathbf{x})\|_2^2 \,. \tag{14}$$

The derivation is provided in Appendix A.

Corollary 1 shows that the fidelity-induced deviation under mixed-fidelity supervision admits an SNR-dependent upper bound. As the noise level increases and the SNR decreases, this bound becomes tighter, consistent with the distribution-level contraction in Proposition 1. These results provide a theoretical rationale for restricting LF supervision to higher-noise denoising tasks.

At inference time, no LF simulation or fidelity indicator is required. Starting from Gaussian noise, the learned conditional denoiser follows the standard reverse diffusion process conditioned on $\mathbf{x}$ to generate the target HF response. Multi-fidelity information therefore enters through training supervision rather than inference-time conditioning.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate NLSA on two heat-source-layout temperature-field prediction (HSL-TFP) benchmarks derived from (Chen et al., 2021), referred to as the *simple* and *complex* benchmarks. LF and HF responses are generated on 50×50 and 200×200 grids, respectively, and HF simulation is approximately $25\times$ more expensive than LF simulation. Each dataset provides 2,000 LF training fields. For each HF budget b, all compared methods use the same fixed subset of HF training samples. We evaluate all methods at $b \in \{1, 2, 5, 10, 20, 50, 100, 200\}$, where b denotes the number of distinct HF training samples.

Compared methods. We compare NLSA with a single-fidelity diffusion baseline and two multi-fidelity methods. **SFM** is a single-fidelity surrogate model trained under the standard diffusion framework (Ho et al., 2020). For multi-fidelity learning, **DMFM** uses LF pretraining followed by HF fine-tuning, while **PD-DMFM** uses physics-driven pretraining on unlabeled data before HF fine-tuning (Zhang et al., 2023).

Training configuration. Diffusion models use Adam with a learning rate of 10^{-4} for 6×10^5 iterations. We use $T = 2000$, a linear variance schedule from $\beta_1 = 10^{-6}$ to $\beta_T = 10^{-2}$, and default boundary $k = 100$. HF and LF batch sizes are both 4. Additional training details are provided in Appendix B.

Table 1: Comparison with the single-fidelity baseline across HF budgets. Lower is better, and the best result is shown in bold.

Dataset	Method	Metric	Number of HF training samples b							
			1	2	5	10	20	50	100	200
Simple	SFM	MAE	1.9505	1.2739	1.1753	0.8534	0.4126	0.3465	0.1399	0.1244
		CMAE	2.4059	1.5017	1.3792	1.0059	0.4506	0.3699	0.1305	0.1170
		MT-AE	1.1199	1.0105	0.9643	0.6440	0.5427	0.4175	0.2246	0.2217
	NLSA	MAE	**0.1520**	**0.1183**	**0.1173**	**0.1443**	**0.1009**	**0.1146**	**0.0911**	**0.1017**
		CMAE	**0.1777**	**0.1332**	**0.1164**	**0.1489**	**0.1102**	**0.1189**	**0.1094**	**0.1040**
		MT-AE	**0.2340**	**0.2063**	**0.1281**	**0.1796**	**0.1200**	**0.1565**	**0.1658**	**0.0986**
Complex	SFM	MAE	1.8722	1.8242	1.7625	0.7579	0.4950	0.4590	0.2646	**0.1979**
		CMAE	1.5871	1.6522	1.5733	0.8115	0.5041	0.4698	0.2536	**0.2052**
		MT-AE	1.6379	1.6206	1.2335	1.4987	0.4809	0.4721	0.2611	**0.2170**
	NLSA	MAE	**0.3118**	**0.4546**	**0.3368**	**0.2260**	**0.4088**	**0.2817**	**0.2223**	0.2650
		CMAE	**0.2924**	**0.4386**	**0.3301**	**0.2290**	**0.4016**	**0.1923**	**0.2210**	0.2616
		MT-AE	**0.2690**	**0.5859**	**0.3787**	**0.2168**	**0.4524**	**0.2303**	**0.2160**	0.2745

Evaluation metrics. Following the evaluation protocol for heat-source-layout temperature-field prediction (Chen et al., 2021; Zhang et al., 2023), we report mean absolute error (MAE), component-constrained mean absolute error (CMAE), and maximum-temperature absolute error (MT-AE). All metrics are reported in kelvin, and lower values indicate better accuracy.

4.2 Predictive Performance and HF Data Efficiency

Comparison with single-fidelity baseline. Table 1 demonstrates the benefit of NLSA under limited HF supervision. At $b = 5$, NLSA reduces the MAE from 1.1753 K to 0.1173 K on the simple benchmark, corresponding to a 90.0% reduction, and from 1.7625 K to 0.3368 K on the complex benchmark, with similarly substantial reductions in CMAE and MT-AE. NLSA achieves lower errors than SFM across all reported metrics and HF budgets, except at $b = 200$ on the complex benchmark. These results demonstrate the effectiveness of incorporating LF supervision, particularly when HF data are scarce. The reported NLSA errors are not strictly monotonic across HF training budgets. These fluctuations may reflect differences in the selected HF subsets and optimization stochasticity, whose effects cannot be isolated with a single run per setting. We therefore interpret differences between adjacent budgets cautiously.

Comparison with multi-fidelity baselines. Figure 2 compares NLSA with the two multi-fidelity baselines across HF budgets. On the simple benchmark, NLSA achieves lower MAE, CMAE, and MT-AE than both DMFM and PD-DMFM for all budgets up to $b = 20$. On the complex benchmark, the same three-metric advantage holds for $b \leq 5$. As the HF budget increases, the pretraining-based methods generally achieve lower errors. This phenomenon supports the central motivation of NLSA: LF observations are most valuable when HF supervision is insufficient, while their benefit diminishes as HF coverage becomes abundant. Notably, on the simple benchmark, NLSA with only one HF sample achieves an MAE of 0.1520 K, compared with 0.0938 K and 0.1065 K for DMFM and PD-DMFM trained with 50 HF samples, respectively. Thus, NLSA attains a comparable error scale while using 50 times fewer HF observations.

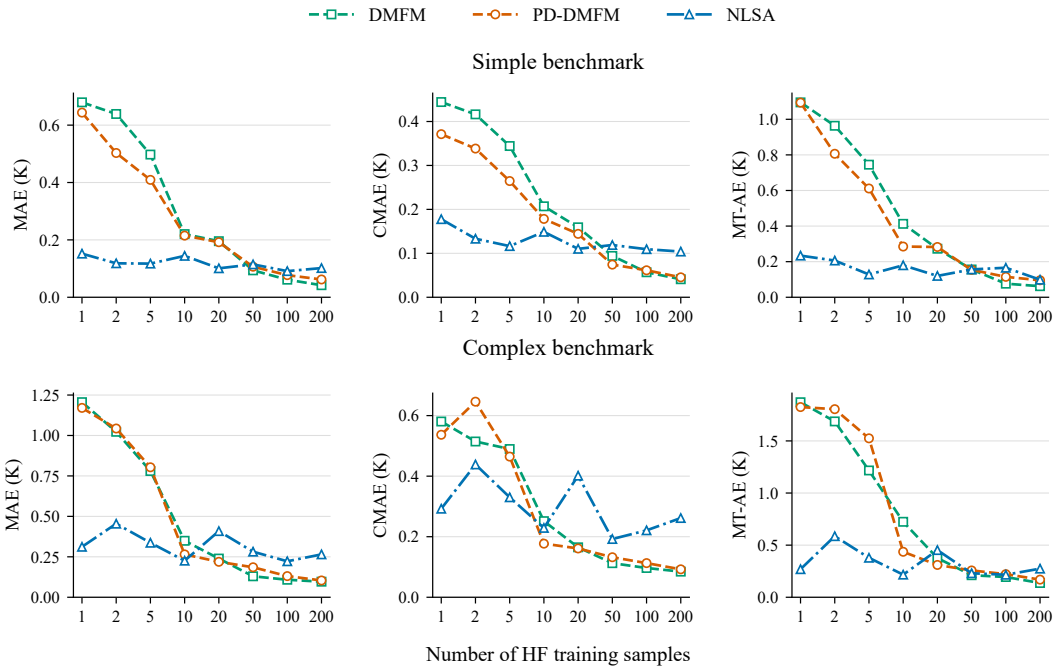


Figure 2: Comparison with multi-fidelity baselines across HF training budgets. All metrics are reported in kelvin, and lower is better. NLSA provides its clearest advantage when HF supervision is severely limited.

HF data efficiency. Beyond the same-budget comparisons above, NLSA reaches comparable or better accuracy with substantially fewer HF observations. On the simple benchmark, NLSA

Table 2: Supervision-boundary ablation on the complex benchmark with $b = 10$ HF samples and $T = 2000$. Bold marks the best result. LF timestep coverage is $100(T - k + 1)/T$. SNR is evaluated at discrete schedule endpoints and rounded to three decimal places.

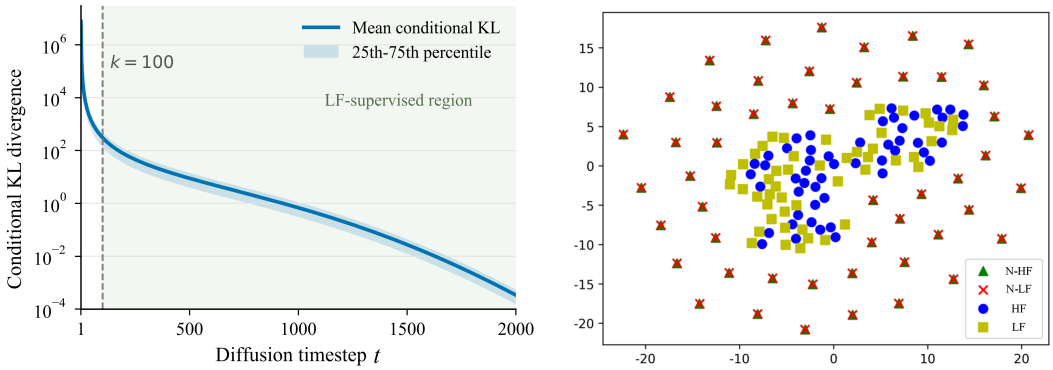
Boundary k	10	20	50	100	200	500	1000
MAE (K)	0.4893	0.4679	0.2736	**0.2260**	0.3256	0.4693	0.5542
LF coverage (%)	99.55	99.05	97.55	95.05	90.05	75.05	50.05
SNR_k	4253.121	1029.990	161.366	39.721	9.531	1.152	0.089

with only five HF samples outperforms SFM trained with 200 HF samples across all three metrics, corresponding to a 97.5% reduction in HF observations. On the complex benchmark, NLSA with ten HF samples similarly outperforms SFM with 100 HF samples, using 90% fewer HF observations.

4.3 Supervision-Boundary and Cross-Fidelity Analysis

Supervision-boundary ablation. We evaluate the supervision boundary on the complex benchmark with $b = 10$ HF training samples, using $k \in \{10, 20, 50, 100, 200, 500, 1000\}$. Table 2 reports the prediction MAE together with the boundary SNR and the fraction of timesteps eligible for LF supervision. The MAE initially decreases with increasing k and then rises, reaching its minimum at $k = 100$.

Increasing k delays the onset of LF supervision to noisier, lower-SNR states, thereby reducing the cross-fidelity discrepancy at the supervision boundary, while simultaneously reducing the range of timesteps that can receive LF supervision. From $k = 10$ to $k = 100$, the boundary SNR decreases by approximately $107\times$, whereas LF timestep coverage decreases by only 4.5 percentage points, from 99.55% to 95.05%. Here, coverage refers to the fraction of diffusion timesteps eligible for LF supervision, rather than the proportion of LF samples. The observed non-monotonic error is consistent with a trade-off between cross-fidelity discrepancy and LF supervision coverage.



(a) Quantitative conditional KL

(b) Qualitative t-SNE visualization

Figure 3: Quantitative and qualitative views of cross-fidelity contraction. (a) Conditional KL along the forward diffusion process on the complex benchmark, evaluated using 200 paired HF/LF temperature fields and the analytical expression in Proposition 1. The solid line shows the mean, the shaded band denotes the 25th–75th percentile, the dashed line marks $k = 100$, and the shaded region indicates timesteps eligible for LF supervision. (b) t-SNE visualization of clean HR/LR fields and their noisy counterparts N-HR/N-LR at $t = 100$. The noisy distributions exhibit substantially greater overlap.

Cross-fidelity contraction analysis. We quantify cross-fidelity contraction using 200 available paired HF/LF temperature fields from the complex benchmark, where each pair corresponds to the same physical condition. After aligning the LF fields to the HF grid and applying the training normalization, we evaluate the analytical conditional KL in Proposition 1 over all diffusion timesteps. Figure 3(a) shows that the mean conditional KL decreases rapidly as the diffusion noise level increases, from 32,788.38 at $t = 10$ to 306.22 at $t = 100$, and further to 0.687 at $t = 1000$. The shaded band denotes the 25th–75th percentile across paired fields. At the default boundary $k = 100$, substantial contraction has already occurred while 95.05% of diffusion timesteps remain eligible for

LF supervision. The absolute KL values depend on the observed HF–LF discrepancies, while their decrease with t follows from the SNR scaling characterized in Proposition 1.

Figure 3(b) provides a complementary low-dimensional visualization. The clean HF and LF samples exhibit clearly different distributions, whereas their noisy counterparts largely overlap, consistent with the cross-fidelity contraction observed in Figure 3 (a).

5 Discussion

The central implication of our results is that the usefulness of LF supervision depends on the diffusion noise level. An LF response can remain useful for diffusion training even when it differs substantially from its HF counterpart at the clean-data level. As the signal-to-noise ratio decreases, the conditional discrepancy between noisy HF and LF responses contracts according to the diffusion SNR, and their corresponding denoising targets become more closely aligned. This suggests a task-dependent notion of fidelity: LF data that are too discrepant to supervise low-noise denoising tasks can still provide useful supervision at higher noise levels. NLSA exploits this structure by allocating LF supervision across noise levels rather than learning a global LF-to-HF correction.

The underlying principle is not specific to thermal-field prediction or to a particular denoising architecture. The analysis applies when deterministic HF and LF responses can be represented in a common output space and subjected to the same Gaussian forward diffusion process. Under these conditions, the cross-fidelity discrepancy can be related directly to the diffusion SNR, providing a principled basis for noise-dependent supervision allocation. Although the present experiments focus on two thermal-field benchmarks, the same principle could be explored for other physical systems and fidelity constructions that admit a common representation.

The current formulation uses a fixed hard boundary k to balance cross-fidelity discrepancy against the coverage of LF supervision. A smaller k exposes more denoising timesteps to LF supervision but also includes higher-SNR states with larger cross-fidelity discrepancy, whereas a larger k restricts LF supervision to noisier states at the cost of reduced coverage. At present, k is selected empirically, and the non-monotonic ablation suggests that an appropriate boundary may depend on the particular fidelity pair and physical problem. A natural extension is therefore to determine k adaptively from the estimated cross-fidelity discrepancy, for example by activating LF supervision once the discrepancy falls below a prescribed tolerance. Beyond such a hard boundary, NLSA could be generalized to a soft allocation scheme in which the contribution of LF supervision varies continuously with the diffusion timestep, SNR, or estimated cross-fidelity discrepancy. A further direction is to derive a principled allocation rule from an objective or error bound that jointly accounts for fidelity-induced target deviation and the benefit of additional LF supervision. Such a formulation could yield either an adaptive boundary or a continuous weighting schedule, replacing the empirically chosen fixed boundary with a discrepancy-aware supervision strategy.

6 Conclusion

We proposed NLSA, a diffusion-based multi-fidelity surrogate framework that allocates supervision across diffusion noise levels. HF observations supervise the full denoising process, while LF observations supplement higher-noise tasks where cross-fidelity discrepancies are reduced. For deterministic responses, we showed that the discrepancy between noisy HF and LF conditional distributions, as well as the corresponding denoising-target deviation, scales with the diffusion SNR. Experiments on two thermal-field benchmarks show that NLSA is particularly effective when HF supervision is scarce, substantially improving prediction accuracy and HF data efficiency. These results demonstrate that diffusion noise level provides a principled mechanism for determining when LF data can serve as useful supervision.

AI Use Statement

Large Language Models were used to assist with manuscript writing, language polishing, and formatting. All research ideas, methodological design, theoretical derivations, experimental design, interpretation of results, and scientific conclusions were developed and verified by the authors. All

AI-assisted content was reviewed by the authors, who take full responsibility for the accuracy and integrity of the submitted work.

Ethics Statement

We affirm that all authors have read and adhere to the ICLR Code of Ethics. This work focuses on multi-fidelity surrogate modeling for simulated physical fields and does not involve human or animal subjects, sensitive personal data, or privacy risks. The experiments are based entirely on numerical simulation data and do not involve proprietary user data. We are not aware of any immediate ethical risks specific to the proposed method. As with surrogate modeling methods more broadly, deployment in safety-critical scientific or engineering applications should be accompanied by appropriate validation of prediction accuracy and reliability.

Reproducibility Statement

To support reproducibility of the reported results, we provide the implementation of NLSA, training scripts, evaluation code. Code will be made public upon acceptance. The main paper specifies the proposed supervision-allocation method, diffusion configuration, HF training budgets, compared methods, and evaluation metrics. Additional implementation details, data and evaluation protocols, diffusion training and sampling settings, and additional qualitative results are provided in the appendices.

References

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations*, 2022.

Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023.

Xianqi Chen, Xiaoyu Zhao, Zhiqiang Gong, Jun Zhang, Weien Zhou, Xiaoqian Chen, and Wen Yao. A deep neural network surrogate modeling benchmark for temperature field prediction of heat source layout. *Science China Physics, Mechanics & Astronomy*, 64(11), 2021. doi: 10.1007/s11433-021-1755-6.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Amanda A. Howard, Mauro Perego, George Em Karniadakis, and Panos Stinis. Multifidelity deep operator networks for data-driven and physics-informed problems. *Journal of Computational Physics*, 493:112462, 2023. doi: 10.1016/j.jcp.2023.112462. URL https://doi.org/10.1016/j.jcp.2023.112462.

Jiahe Huang, Guandao Yang, Zichen Wang, and Jeong Joon Park. DiffusionPDE: Generative PDE-solving under partial observation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4140. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb3878c1dcbfff9ee95d5d033e5f5942-Abstract-Conference.html.

Marc C. Kennedy and Anthony O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000. doi: 10.1093/biomet/87.1.1.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Zongyi Li, Nikola 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. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Phillip Lippe, Bas Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d529b943af3dba734f8a7d49efcb6d09-Abstract-Conference.html.

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

Xuhui Meng and George Em Karniadakis. A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse PDE problems. *Journal of Computational Physics*, 401:109020, 2020. doi: 10.1016/j.jcp.2019.109020. URL https://doi.org/10.1016/j.jcp.2019.109020.

Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018. doi: 10.1137/16M1082469.

Paris Perdikaris, Maziar Raissi, Andreas Damianou, Neil D. Lawrence, and George Em Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2198):20160751, 2017. doi: 10.1098/rspa.2016.0751. URL https://doi.org/10.1098/rspa.2016.0751.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021.

Dule Shu, Zijie Li, and Amir Barati Farimani. A physics-informed diffusion model for high-fidelity flow field reconstruction. *Journal of Computational Physics*, 2023. doi: 10.1016/j.jcp.2023.111972. URL https://arxiv.org/abs/2211.14680.

Aliaksandra Shysheya, Cristiana Diaconu, Federico Bergamin, Paris Perdikaris, José Miguel Hernández-Lobato, Richard E. Turner, and Emile Mathieu. On conditional diffusion models for PDE simulations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0732. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2974844555dc383ea16c5f35833c7a57-Abstract-Conference.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yunyang Zhang, Zhiqiang Gong, Weien Zhou, Xiaoyu Zhao, Xiaohu Zheng, and Wen Yao. Multi-fidelity surrogate modeling for temperature field prediction using deep convolution neural network. arXiv preprint arXiv:2301.06674, 2023. URL https://arxiv.org/abs/2301.06674.

A Theoretical Proofs

A.1 Proof of the cross-fidelity contraction

Fix a physical input $\mathbf{x}$. For $f \in \{\mathrm{H}, \mathrm{L}\}$, deterministic responses corrupted at discrete timestep t have conditional distributions

$$q_t^f(\mathbf{y}_t \mid \mathbf{x}) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t} F_f(\mathbf{x}), (1-\bar{\alpha}_t)\mathbf{I}\big), \qquad \boldsymbol{\delta}(\mathbf{x}) = F_{\mathrm{H}}(\mathbf{x}) - F_{\mathrm{L}}(\mathbf{x}). \tag{A.1}$$

The covariance is positive definite because $0 < \bar{\alpha}_t < 1$. The equal-covariance Gaussian KL identity therefore yields

$$\begin{aligned} D_{\mathrm{KL}}(q_t^{\mathrm{H}} \| q_t^{\mathrm{L}}) &= \frac{1}{2}\left(\sqrt{\bar{\alpha}_t}\boldsymbol{\delta}(\mathbf{x})\right)^\top \left[(1-\bar{\alpha}_t)\mathbf{I}\right]^{-1}\left(\sqrt{\bar{\alpha}_t}\boldsymbol{\delta}(\mathbf{x})\right) \\ &= \frac{\bar{\alpha}_t}{2(1-\bar{\alpha}_t)}\|\boldsymbol{\delta}(\mathbf{x})\|_2^2 = \frac{1}{2}\,\mathrm{SNR}_t\,\|\boldsymbol{\delta}(\mathbf{x})\|_2^2. \end{aligned} \tag{A.2}$$

Under $0 < \beta_t < 1$, $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ decreases with t. Since $a/(1-a)$ is increasing for $0 < a < 1$, $\mathrm{SNR}_t = \bar{\alpha}_t/(1-\bar{\alpha}_t)$ decreases with t, so the KL divergence is non-increasing. This proves Proposition 1 and Eq. (7), including $\boldsymbol{\delta}(\mathbf{x}) = \mathbf{0}$, when both divergences vanish. For nonzero discrepancy, their ratio equals the SNR ratio.

A.2 Squared-loss denoising-target bias

At a fixed $(\mathbf{y}_t, t, \mathbf{x})$, specifying the fidelity determines the injected noise:

$$\boldsymbol{\epsilon}_f^\star = \frac{\mathbf{y}_t - \sqrt{\bar{\alpha}_t}F_f(\mathbf{x})}{\sqrt{1-\bar{\alpha}_t}}, \qquad \boldsymbol{\epsilon}_{\mathrm{L}}^\star - \boldsymbol{\epsilon}_{\mathrm{H}}^\star = \sqrt{\mathrm{SNR}_t}\,\boldsymbol{\delta}(\mathbf{x}). \tag{A.3}$$

For $t \geq k$, let $w(\mathbf{y}_t, \mathbf{x}, t) = \Pr(f = \mathrm{L} \mid \mathbf{y}_t, \mathbf{x}, t)$ under the population mixture induced by Eq. (10). The squared-loss population optimum is the conditional mean:

$$\boldsymbol{\epsilon}_{\mathrm{mix}}^\star = (1-w)\boldsymbol{\epsilon}_{\mathrm{H}}^\star + w\boldsymbol{\epsilon}_{\mathrm{L}}^\star, \qquad \boldsymbol{\epsilon}_{\mathrm{mix}}^\star - \boldsymbol{\epsilon}_{\mathrm{H}}^\star = w\sqrt{\mathrm{SNR}_t}\,\boldsymbol{\delta}(\mathbf{x}). \tag{A.4}$$

Thus, wherever the conditional mixture is defined, $0 \leq w \leq 1$ gives

$$\|\boldsymbol{\epsilon}_{\mathrm{mix}}^\star - \boldsymbol{\epsilon}_{\mathrm{H}}^\star\|_2^2 = w^2\,\mathrm{SNR}_t\,\|\boldsymbol{\delta}(\mathbf{x})\|_2^2 \leq \mathrm{SNR}_t\,\|\boldsymbol{\delta}(\mathbf{x})\|_2^2, \tag{A.5}$$

which proves Corollary 1. The result characterizes the fidelity-induced target bias; it does not determine the optimal supervision boundary k, which also depends on the coverage of LF supervision.

B Reproducibility Details

B.1 Benchmarks and Evaluation Protocol

The heat-source-layout temperature-field prediction (HSL-TFP) task maps heat-source spatial configurations and heat-generation conditions to a two-dimensional steady-state temperature field. The simple and complex benchmarks are derived from Chen et al. (2021). The simple benchmark contains 20 square components, each measuring $0.01\,\mathrm{m} \times 0.01\,\mathrm{m}$ with heat-source intensity $10{,}000\,\mathrm{W/m^2}$. The complex benchmark contains 12 rectangular components with different dimensions and heat-source intensities ranging from $4{,}000$ to $20{,}000\,\mathrm{W/m^2}$. Both use a domain with side length $0.1\,\mathrm{m}$, thermal conductivity $1\,\mathrm{W/(m\,K)}$, and a heat sink held at $298\,\mathrm{K}$; the remaining boundary is insulated.

For one test field, let $\widehat{Y}_{ij}$ and Y_{ij} denote the predicted and reference HF temperatures at pixel (i,j), and let N be the total number of evaluated pixels. The three metrics are

$$\mathrm{MAE} = \frac{1}{N}\sum_{i,j}|\widehat{Y}_{ij} - Y_{ij}|, \tag{B.1}$$

$$\mathrm{CMAE} = \frac{\sum_{i,j} M_{ij}|\widehat{Y}_{ij} - Y_{ij}|}{\sum_{i,j} M_{ij}}, \tag{B.2}$$

$$\text{MT-AE} = \left|\max_{i,j}\widehat{Y}_{ij} - \max_{i,j} Y_{ij}\right|. \tag{B.3}$$

Here $M_{ij} = 1$ in heat-source component-covered regions and $M_{ij} = 0$ elsewhere. All three metrics are in kelvin (K), and lower values indicate better predictions.

B.2 Diffusion Training and Sampling

SFM and NLSA use the same diffusion configuration in Table 3; only NLSA uses the LF supervision branch. When fewer than four distinct HF samples are available, samples are repeated within mini-batches without increasing the number of distinct HF labels. Sampling starts from Gaussian noise and proceeds through the standard DDPM ancestral reverse process, conditioned on the physical input. All experiments were implemented in PyTorch and run on an NVIDIA RTX 3090 GPU with 24 GB memory.

Table 3: Reported diffusion training and sampling settings.

Setting	Value
Initialization	Orthogonal
Optimizer	Adam
Learning rate	10^{-4}
Training iterations	6×10^5
HF / LF batch size	4 / 4
Forward / reverse steps	2000 / 2000
Variance schedule	Linear, $\beta_1 = 10^{-6}$ to $\beta_T = 10^{-2}$
Default boundary	$k = 100$
Tested boundaries	10, 20, 50, 100, 200, 500, 1000
Sampling	Standard DDPM ancestral sampling

B.3 Multi-Fidelity Baseline Training

DMFM pretrains on LF observations and then fine-tunes on the available HF observations. PD-DMFM instead uses physics-driven pretraining on unlabeled layouts before HF fine-tuning. Both use Kaiming initialization and the Ranger optimizer with cosine annealing and warm restarts over 150 epochs. Table 4 lists the learning rates, schedule, and batch sizes. All experiments were implemented in PyTorch and run on an NVIDIA RTX 3090 GPU with 24 GB memory.

Table 4: Training settings for the multi-fidelity baselines.

Setting	Reported value
Training epochs	150
Initial backbone learning rate, pretraining	0.01
Initial backbone learning rate, fine-tuning	0.001
Initial projection-head learning rate	0.01
Learning-rate schedule	Cosine annealing with warm restarts
Minimum learning rate	10^{-7}
Initial restart period	10 epochs
Restart-period multiplier	2
Pretraining batch size	16
Fine-tuning batch size	1

C Additional Qualitative Results

Figures 4 and 5 present temperature-field predictions and corresponding spatial absolute-error maps on the simple and complex benchmarks, respectively.

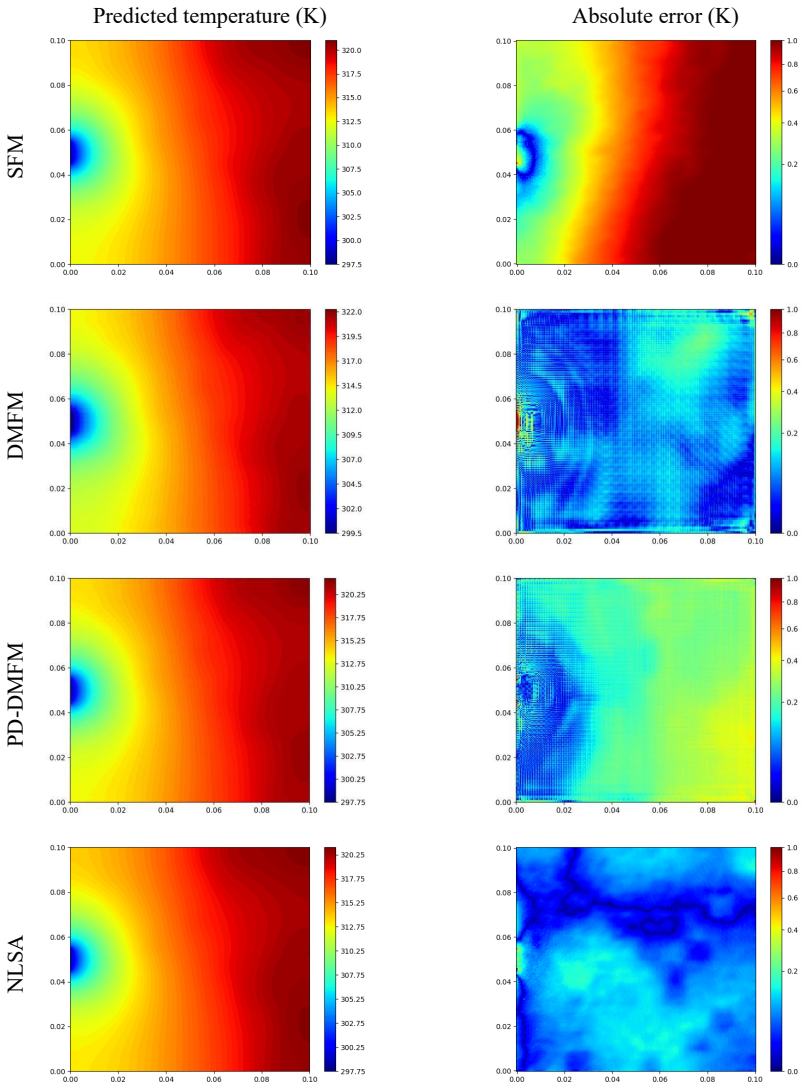


Figure 4: Predicted temperature fields (left) and absolute-error maps (right) on the simple benchmark at $b = 50$, for the same test configuration. Error panels share a 0–1 K scale with square-root normalization and saturation above 1 K. Coordinates are in meters.

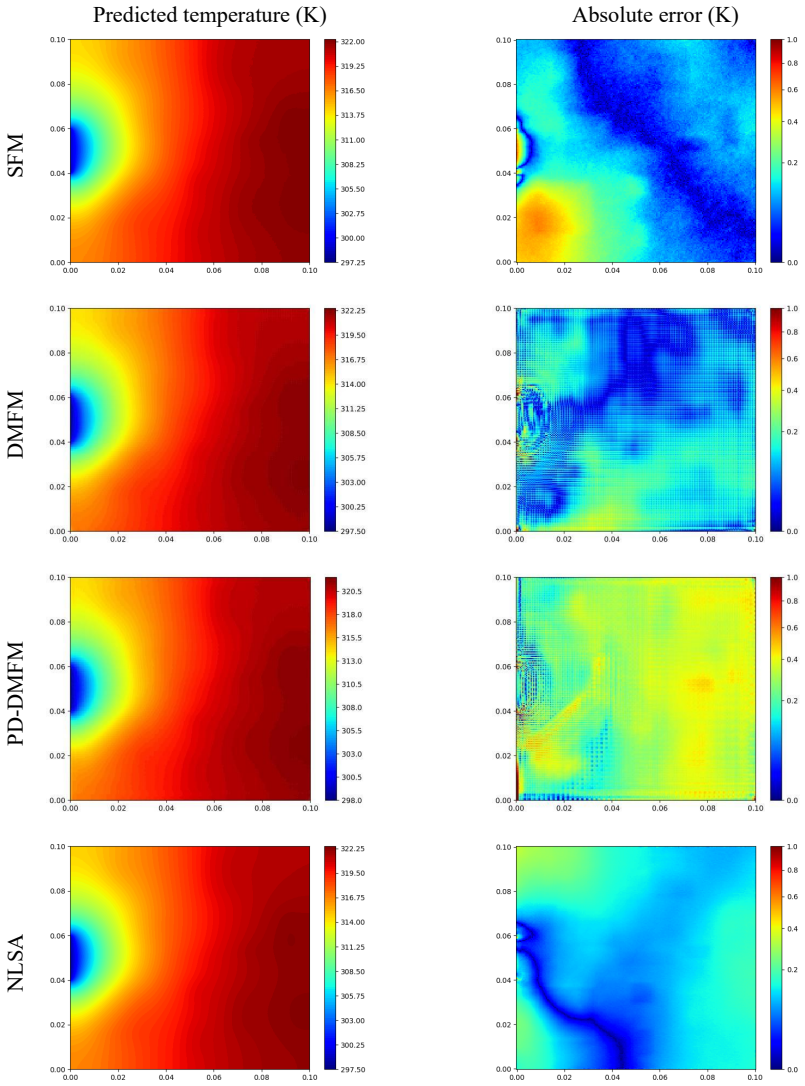


Figure 5: Predicted temperature fields (left) and absolute-error maps (right) on the complex benchmark at $b = 50$, for the same test configuration. Error panels share a 0–1 K scale with square-root normalization and saturation above 1 K. Coordinates are in meters.