FINITE-STEP LAW ERROR AND LEARNED SCORE RISK IN CAHN–HILLIARD–COOK DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical simulators are increasingly used to supply training data for score-based generative models. A simulator can be stable and exactly conservative while still sampling the wrong distribution, and a score model trained on its output inherits that error. We give a quantitative account of this chain for Cahn–Hilliard–Cook (CHC) dynamics, from integrator to sampled law to learned score. For linearized CHC, the discrete Lyapunov equation gives the per-mode bias of an energy-stable IMEX step in closed form (up to $4.6\times$ excess high-frequency variance), and a discrete fluctuation–dissipation covariance removes it at no cost. Under a shared Gaussian corruption, the excess target risk of learning the simulated law is a covariance-weighted relative Fisher divergence; we derive its exact fixed-time form and a risk-elasticity identity that predicts sub-quadratic growth at practical step sizes and a larger penalty for under- than over-dispersion. On a nonlinear 15×15 CHC Gibbs field at fixed physical time, coarser steps increase endpoint error and held-out target-DSM excess for a CNN and an FNO in all ten paired seeds, following the predicted pattern: law error $\propto h^{1.0}$ and risk exponents falling from about 1.9 to 1.5. Generation and inverse-problem tests show where these gains carry over and where the sampler decides the outcome.

1 INTRODUCTION

Score-based diffusion models usually corrupt data with designed Gaussian noise, but structured and physics-driven forward processes are now well established (Scimeca et al., 2025; Singhal et al., 2024; Holzschuh et al., 2023). Whenever a numerical simulator enters this pipeline—as a forward process or as the source of training samples—its finite-step implementation raises a question upstream of learning: if the simulator samples a slightly wrong distribution, how much does the learned score pay for it? Numerical stability and exact conservation do not answer this question. A simulator can satisfy both and still supply biased data. We answer it quantitatively: the integrator determines the sampled law in closed form, and the sampled law determines the population score risk through an exact identity whose growth rate we characterize.

We therefore separate four properties that are often conflated: **structural fidelity** (the implemented constraints hold), **measure fidelity** (the simulator samples its prescribed invariant law, or the intended law at a fixed horizon), **learning fidelity** (a score trained on the simulated data performs well on the target), and **downstream utility** (generation or inference improves). CHC is a convenient testbed because mass conservation is directly checkable, its linearization admits an exact covariance analysis, and a small nonlinear Gibbs field is tractable enough for a held-out learning test. Exact mass conservation, which CHC provides by construction, says nothing about the other three properties, so we examine each one in turn.

Our main experiment keeps physical simulation and learning corruption apart. Let π be the prescribed zero-mass Gibbs law and $\widetilde{\pi}_{h,T}$ the endpoint law after physical time T with simulation step h. We train noise predictors on fields from either π or $\widetilde{\pi}_{h,T}$ using the same Gaussian variance-preserving (VP) corruption, indexed by a separate learning time τ, and evaluate all of them on held-out reference fields with shared corruption draws. The comparison thus isolates how a change in the training distribution affects target score risk. We hold $T = 1$ fixed and vary h, so every model sees fields produced at the same physical horizon (Fig. 4).

Our contributions are:

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

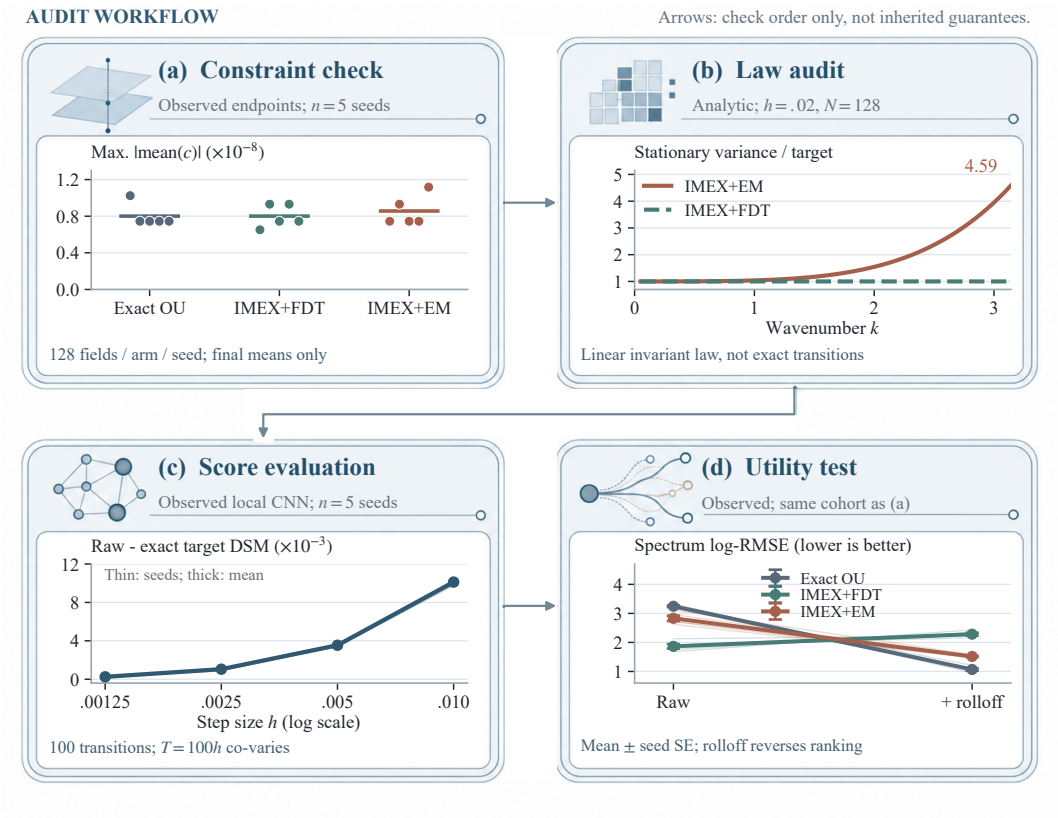

Figure 1: The four checks in the order a simulator-trained model should pass them, with evidence from our experiments. (a) Structural fidelity: exact mass conservation (maximum absolute spatial mean over 128 final fields per arm and seed, $n = 5$). (b) Measure fidelity: analytic stationary variance ratio of the linear chain, biased under IMEX+EM and exact under IMEX+FDT. (c) Learning fidelity: target-DSM excess grows with step size (fixed-transition study; the fixed-T comparison is Fig. 4). (d) Downstream utility: generated-spectrum error, whose ranking depends on the sampler's rolloff. Arrows give the order of checks; each panel is a separate experiment, and header glyphs are illustrative.

- **A closed-form measure bias and a free repair.** For linearized CHC, the discrete Lyapunov equation gives the per-mode stationary bias of the IMEX step in closed form for both post-update and in-solve noise placements. The bias reaches $4.6\times$ at the analyzed step and grows with resolution. Replacing the continuum noise amplitude by the discrete fluctuation–dissipation covariance removes it for either placement without changing the deterministic step (Fig. 2).
- **An exact law-to-risk theory.** Under a shared Gaussian corruption, the excess target risk of learning the simulated law is a covariance-weighted relative Fisher divergence. For the linear control we derive its exact fixed-time form, its h^2 asymptote, and a risk-elasticity identity (Proposition 1) that explains why the risk grows sub-quadratically at practical steps and why under-dispersion is penalized more than over-dispersion. A corollary gives a network-accuracy condition under which trained models provably preserve the population ordering.
- **Confirmation on a nonlinear field with two backbones.** On a 15×15 CHC Gibbs field at fixed $T = 1$, endpoint error and held-out target-DSM excess increase with step size for a CNN and an FNO in all ten paired seeds, and follow the growth pattern predicted by the theory (Figs. 3 and 4). Generation and inverse tests show that correct covariance helps exact-Gaussian inference, while generated-sample rankings are governed by the sampler, which gives a concrete order in which to validate a simulator-trained model.

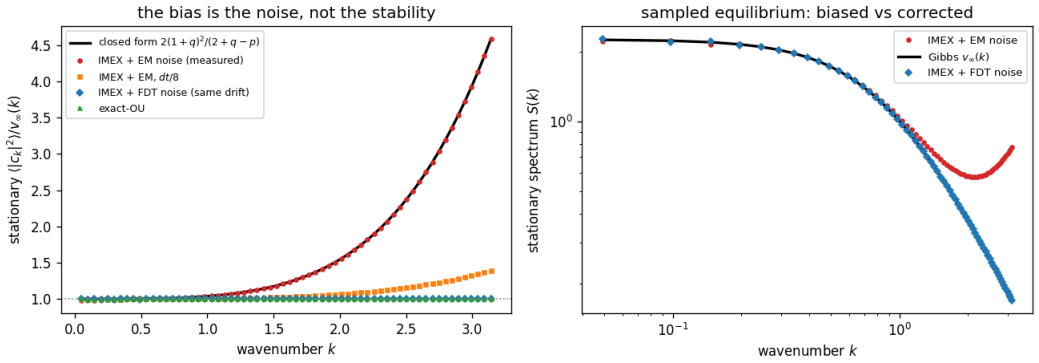

Figure 2: Energy stability and measure preservation are independent. *Left*: the energy-stable IMEX step with post-update Euler–Maruyama noise (red) follows the closed-form bias $2(1+q)^2/(2+q-p)$ (black) and over-weights high k by up to $4.6\times$; the *same* drift with discrete-FDT noise (blue diamonds) is exact for this linear target, while a smaller step (orange) reduces but does not remove the bias. *Right*: the sampled equilibrium spectrum. Naive IMEX (red) turns up at high k where the Gibbs spectrum (black) decays; the FDT-corrected scheme (blue) tracks Gibbs.

2 Related Work

Score diffusion and physics-defined forwards. Denoising diffusion and score SDEs reverse designed Gaussian corruptions (Ho et al., 2020; Song et al., 2021; Karras et al., 2022); function-space models define field-valued Gaussian reference measures (Kerrigan et al., 2023; Lim et al., 2025), and score Fokker–Planck regularization enforces consistency of learned scores (Lai et al., 2023). Structured corruptions include heat, cold, soft, spectral, anisotropic-SPDE, Kuramoto, multiplicative, and count-valued variants (Rissanen et al., 2023; Hoogeboom & Salimans, 2024; Bansal et al., 2023; Daras et al., 2022; Scimeca et al., 2025; Holl et al., 2026; Song et al., 2025; Gruhlke et al., 2026; E. Santos et al., 2025), alongside nonlinear local DSM (Singhal et al., 2024), phase-space forwards (Dockhorn et al., 2022; Pandey & Mandt, 2023), SMDP reversal (Holzschuh et al., 2023), measure-preserving dynamics (Zhang et al., 2024), and prior-constrained models (Zhou et al., 2025). We ask a complementary question: whether a finite-step physical simulator samples its prescribed measure before learning, and what the learned score pays when it does not. Cahn–Hilliard, Allen–Cahn, and Model-C dynamics (Cahn & Hilliard, 1958; Allen & Cahn, 1979; Hohenberg & Halperin, 1977) provide the testbed, with a brine-rejection-inspired coupling (Worster, 1997; Feltham et al., 2006) as a downstream operator.

Inverse problems as a downstream test. DPS, ΠGDM, DDRM, and variational methods condition diffusion priors on measurements (Chung et al., 2024; Song et al., 2023; Kawar et al., 2022; Mardani et al., 2024); PDE-aware reconstruction and assimilation also use simulators during inference (Huang et al., 2024; Rozet & Louppe, 2023; Huynh et al., 2026; Li et al., 2025). Our melt map is a simulator in the SBI sense (Cranmer et al., 2020) used as a transfer test; we compare DPS with regularized, oracle-Gaussian (Laumont et al., 2022), and conditional-score (Batzolis et al., 2021; Baldassari et al., 2023) baselines, following work on calibration (Talts et al., 2020; Hermans et al., 2022; Lemos et al., 2023) and sampler sensitivity (Crafts & Villa, 2025; Zach et al., 2026; MOUFAD et al., 2025; Zhang et al., 2025).

Discrete fluctuation–dissipation. Energy stability, including that of convex-splitting CH schemes (Eyre, 1998), does not imply that a stochastic discretization preserves the invariant law; discrete fluctuation–dissipation and invariant-measure error are classical for SDE/SPDE integrators (Bréhier, 2014; Abdulle et al., 2014; Leimkuhler & Matthews, 2015; Delong et al., 2013; Pazner et al., 2019; Deng et al., 2025), and score matching has been linked to entropy production (Ding et al., 2026). Building on the standard covariance, Tweedie, and Metropolis (Bou-Rabee & Vanden-Eijnden, 2010) tools, we derive the modal bias of the CHC–IMEX split, its exact score-risk consequences, and test how they carry over to learning.

3 PRELIMINARIES

Score-based diffusion. Generation reverses a forward diffusion using the score $\nabla_x \log p_\tau(x)$, learned by denoising score matching (DSM); for a VP process $p_\tau(x_\tau \mid x_0) = \mathcal{N}(\alpha_\tau x_0, \beta_\tau^2 I)$ and DSM reduces to predicting the injected noise. In our main experiment this corruption is applied to fields produced by the physical simulation.

Cahn–Hilliard–Cook (CHC) dynamics. A conserved scalar field c evolves by $\partial_t c = M\nabla^2\big(\delta\mathcal{F}/\delta c\big) + \xi$, with Ginzburg–Landau free energy $\mathcal{F}[c] = \int \frac{\kappa}{2}|\nabla c|^2 + f(c)\,dx$, mobility M, and conservative thermal noise ξ satisfying the fluctuation–dissipation theorem (FDT). With a double-well f the field phase-separates and coarsens. With a temperature-dependent Landau curvature $f'' = a(\vartheta)$, raising the temperature ϑ past ϑ_c melts the ordered state toward a near-Gaussian equilibrium. The conservative ∇^2 freezes the $k{=}0$ (mean) mode, so total mass is exactly conserved.

4 METHOD

4.1 PHYSICAL TRAINING LAWS AND A SHARED LEARNING TASK

The nonlinear testbed is a periodic 15×15 field on $\mathcal{S} = \{x \in \mathbb{R}^{225} : \mathbf{1}^\top x = 0\}$ with prescribed Gibbs law π. Write $\widetilde{K}_h$ for the implemented CHC transition and $\widetilde{\pi}_{h,T} = \pi_{\text{init}}\widetilde{K}_h^{T/h}$ for its endpoint law, with integer T/h. We start from equilibrium, $\pi_{\text{init}} = \pi$, approximated in practice by independently validated HMC samples; any departure of $\widetilde{\pi}_{h,T}$ from π is then due to the simulator.

For training data X drawn from either π or $\widetilde{\pi}_{h,T}$, the learning task is

$$Y_\tau = \alpha_\tau X + \beta_\tau\,\epsilon, \qquad \alpha_\tau = e^{-\tau/2}, \quad \beta_\tau^2 = 1 - e^{-\tau},$$
$$\epsilon \sim \mathcal{N}_{\mathcal{S}}(0, I_{\mathcal{S}}), \quad \tau \sim \nu = \text{LogUniform}[.001, 8].$$

Physical time T controls data generation; the learning time τ controls only the subsequent corruption. Let $\widehat{f}_{h,T}$ and $\widehat{f}_{\text{ref}}$ be noise predictors trained on simulated endpoints and on HMC reference fields, respectively. We evaluate both on the same independent group of HMC reference fields with shared (τ, ϵ) draws and report

$$\Delta_{h,T} = \widehat{\mathcal{R}}_{\text{ref}}(\widehat{f}_{h,T}) - \widehat{\mathcal{R}}_{\text{ref}}(\widehat{f}_{\text{ref}}), \qquad \mathcal{R}_\pi(f) = c_{\text{mse}}\,\mathbb{E}_{X\sim\pi,\tau,\epsilon}\|f(Y_\tau,\tau) - \epsilon\|_2^2,$$

where $c_{\text{mse}} = 1/225$ averages over the ambient pixels. Positive $\Delta_{h,T}$ means that training on the simulated endpoint law worsens prediction on the prescribed target; figures label the HMC-trained comparator "exact".

4.2 AUXILIARY LINEAR-CHC FORWARD

Linearizing CHC about the melted equilibrium ($a \equiv a_{\text{high}} > 0$) gives, per Fourier mode, an Ornstein–Uhlenbeck process $dc_k = -\lambda_k c_k\,d\tau + \sigma_k\,dW$ with $\lambda_k = Mk^2(a_{\text{high}} + \kappa k^2)$ and $\sigma_k^2 = 2Mk^2 k_{\text{B}}\vartheta$, i.e. a mode-wise VP-SDE with the Ornstein–Zernike stationary spectrum

$$v_\infty(k) = \frac{\sigma_k^2}{2\lambda_k} = \frac{k_{\text{B}}\vartheta}{a_{\text{high}} + \kappa k^2}. \tag{1}$$

Since $\lambda_0 = \sigma_0 = 0$, mass is conserved (final mean residuals $\sim 10^{-8}$, Fig. 31a). Its analytic kernel supplies DSM pairs without time stepping for the downstream generation and inverse tests, which use a circular FiLM CNN and Tweedie ancestral sampling with an optional high-k rolloff (Appendix D).

5 FINITE-STEP MEASURE BIAS

A general covariance check. Consider a stable linear stochastic step

$$x^{n+1} = Gx^n + \eta^n, \qquad \eta^n \sim \mathcal{N}(0, Q), \qquad \rho(G) < 1. \tag{2}$$

For any positive-definite target covariance C, the Gaussian law $\mathcal{N}(0, C)$ is invariant under Eq. (2) if and only if

$$Q = C - GCG^\top \succeq 0. \tag{3}$$

This is the classical discrete Lyapunov, or discrete fluctuation–dissipation, condition (Appendix A). If a continuum-time noise covariance is inserted without integrating it consistently with G, the stationary covariance instead solves $V = GVG^\top + Q$: deterministic stability guarantees that some stationary covariance exists, not that it equals C. From an arbitrary initial covariance V_0,

$$V_n = G^n V_0 (G^\top)^n + \sum_{j=0}^{n-1} G^j Q (G^\top)^j. \tag{4}$$

Equation (3) therefore keeps $V_n = C$ when $V_0 = C$, but it does not reproduce the exact continuous-time transition from an arbitrary V_0; kernel equality additionally requires matching drift and noise covariance.

Energy-stable $\neq$ measure-preserving. Take one nonzero Fourier mode of the melted linearized dynamics, $dc_k = -\lambda_k c_k\, dt + \sigma_k\, dW$, with $\lambda_k = Mk^2(a + \kappa k^2)$ and, by the fluctuation–dissipation theorem, $\sigma_k^2 = 2Mk^2\, k_\mathrm{B}\vartheta$ (Cook, 1970), so that $v_\infty(k)$ is given by Eq. (1). Our simulator uses an energy-stable IMEX deterministic update (Eyre, 1998), with the stiff κk^4 term implicit and the ak^2 term explicit, followed by an additive Euler–Maruyama increment at the continuum amplitude. With M=1, write $P_k = ak^2$ and $Q_k = \kappa k^4$, so $\lambda_k = P_k + Q_k$, and let $p = P_k\, dt$, $q = Q_k\, dt$. Then

$$g_k = \frac{1-p}{1+q}, \qquad c_k^{n+1} = g_k c_k^n + \xi_k, \quad \mathrm{Var}(\xi_k) = \Sigma_k = \sigma_k^2 dt = 2(p+q)v_\infty(k). \tag{5}$$

The stationary variance of this recursion is $V_k = \Sigma_k/(1-g_k^2)$, and since $1 - g_k^2 = (p+q)(2+q-p)/(1+q)^2$,

$$\boxed{\frac{V_k}{v_\infty(k)} = \frac{2(1+q)^2}{2+q-p}}. \tag{6}$$

For a nonzero mode, $2 + q - p > 0$ is equivalent to stability, $|g_k| < 1$. The ratio tends to one as $dt \to 0$ but grows with q. At dt=0.02, a=0.8, and κ=1 it is 1.17 at k=1.5, 2.37 at k=2.5, and 3.95 near k=3, rising to 4.6 at k=π.

The mismatch is in the noise, not the drift. Convex-splitting and IMEX schemes were designed to guarantee deterministic free-energy decay, not to sample a Gibbs measure. Keeping the same amplification g_k but using the discrete fluctuation–dissipation covariance

$$\Sigma_k^{\mathrm{FDT}} = v_\infty(k)(1 - g_k^2) \tag{7}$$

gives $V_k = v_\infty(k)$ exactly at every k, with the deterministic step unchanged. For this linear system, energy stability and measure accuracy are independent. The bias also depends on where the noise enters. Passing the same increment through the implicit solve, $c_k^{n+1} = [(1-p)c_k^n + \xi_k]/(1+q)$, gives $V_k/v_\infty(k) = 2/(2+q-p)$ instead, which *under*-weights high k (0.75 at k=2.5 and 0.53 at k=π for the same parameters; 8192 simulated chains agree within 0.001). The two placements err in opposite directions, and the discrete FDT covariance applied to the effective increment corrects both; Proposition 1 below shows that the two directions are not equally costly for learning. Appendix C derives the general post-update relation (Eq. (18)) and compares drift updates and placements in Table 2. Stable integrators with biased invariant measures are classical (Delong et al., 2013; Pazner et al., 2019; Leimkuhler & Matthews, 2015; Roberts & Tweedie, 1996); what is specific here is the closed-form modal bias for this CHC–IMEX split and its consequence for learning.

Numerical confirmation (Fig. 2). With 8192 per-mode chains started at the stationary law, IMEX with post-update EM noise follows Eq. (6) (measured 1.184, 2.375, 3.927 near $k = 1.5, 2.5, 3$; maximum 4.6×), discrete-FDT noise gives at most 1.008, and $dt/8$ still leaves 1.4×. Because $q = \kappa k^4 dt$ grows with resolution, grid refinement at fixed step raises the maximum bias from 1.016 at N=32 to 63.6 at N=256; in a 512-field test the high-k spectrum is inflated up to 15× and the correlation length shrinks from 3.8 to 2.7, while discrete FDT stays exact (Appendix C).

The bias matters whenever a time-stepped physical SDE defines the terminal law or generates the training data; for non-Gaussian targets, Metropolis correction or invariant-measure integrators play the role of Eq. (7).

5.1 From Sampled Law to Score Risk

The following population identity is a standard consequence of the conditional-expectation projection and Gaussian Tweedie formula; we use it to quantify what a law error costs in score risk. Let π and $\widetilde{\pi}$ be prescribed and implemented endpoint laws on the modeled subspace, and pass both through the same Gaussian corruption $Y_\tau = A_\tau X + L_\tau \epsilon$ with linear A_τ, L_τ, $\epsilon \sim \mathcal{N}_{\mathcal{S}}(0, I_{\mathcal{S}})$, $\tau \sim \nu$, and X, ϵ, τ independent. Write $p_\tau, \widetilde{p}_\tau$ for the corrupted densities and $f^\star_\pi, f^\star_{\widetilde{\pi}}$ for the unrestricted population minimizers of the respective DSM risks. Then

$$\mathcal{R}_\pi(f^\star_{\widetilde{\pi}}) - \mathcal{R}_\pi(f^\star_\pi) = c_{\mathrm{mse}}\, \mathbb{E}_{\tau\sim\nu} \int p_\tau(y) \left\| L_\tau^\top (\nabla \log p_\tau(y) - \nabla \log \widetilde{p}_\tau(y)) \right\|_2^2 dy \geq 0, \tag{8}$$

a covariance-weighted relative Fisher divergence. Under the regularity and injectivity conditions of Appendix A, distinct endpoint laws give strictly positive excess. For our zero-mass VP task, $A_\tau = \alpha_\tau I_{\mathcal{S}}$ and $L_\tau = \beta_\tau I_{\mathcal{S}}$; in ambient coordinates the noise covariance $\beta_\tau^2(I - \mathbf{1}\mathbf{1}^\top/225)$ is singular in $\mathbb{R}^{225}$ but positive definite on $\mathcal{S}$.

A fixed-time Gaussian risk law. In the linear control, Eq. (8) can be evaluated in closed form. Initialize each mode at its target variance $v_\infty(k)$, fix physical time T, and take $n = T/h$ IMEX steps. With $B_k(h) = 2(1 + Q_k h)^2/[2 + (Q_k - P_k)h]$, the ratio in Eq. (6) at $dt = h$, the endpoint variance satisfies

$$v_{h,T,k} - v_\infty(k) = v_\infty(k)[B_k(h) - 1]\left[1 - g_k(h)^{2T/h}\right]. \tag{9}$$

For a fixed finite set of modes and $h = T/n \to 0$, this equals $d_k(T)h + O(h^2)$ with $d_k(T) = v_\infty(k)(P_k + 3Q_k)(1 - e^{-2\lambda_k T})/2$. Substituting into Eq. (8) under the shared VP corruption gives

$$\mathcal{R}_\pi(f^\star_{h,T}) - \mathcal{R}_\pi(f^\star_\pi) = K(T)h^2 + O(h^3), \qquad K(T) = c_{\mathrm{mse}} \mathbb{E}_\tau \sum_k \frac{\beta_\tau^2 \alpha_\tau^4 d_k(T)^2}{(\alpha_\tau^2 v_\infty(k) + \beta_\tau^2)^3}. \tag{10}$$

A first-order error in the endpoint covariance thus costs second order in population score risk (Appendix B). The next result describes the growth away from the asymptote, which is the regime of practical step sizes.

Proposition 1 (Risk elasticity). *Let $\pi = \mathcal{N}(0, C)$ and $\widetilde{\pi} = \mathcal{N}(0, V)$ be diagonal in a common real orthonormal basis of $\mathcal{S}$, with $u_k = v_k - c_k$, $s_{\tau,k} = \alpha_\tau^2 c_k + \beta_\tau^2$, and $\widetilde{s}_{\tau,k} = \alpha_\tau^2 v_k + \beta_\tau^2$. Under the shared VP corruption,*

$$\Delta_{\mathrm{pop}} = \mathcal{R}_\pi(f^\star_{\widetilde{\pi}}) - \mathcal{R}_\pi(f^\star_\pi) = c_{\mathrm{mse}}\, \mathbb{E}_\tau \sum_k \frac{\beta_\tau^2 \alpha_\tau^4 u_k^2}{s_{\tau,k}\, \widetilde{s}_{\tau,k}^2}. \tag{11}$$

Each summand increases with $|u_k|$, with elasticity $\partial \log(\cdot)/\partial \log |u_k| = 2s_{\tau,k}/\widetilde{s}_{\tau,k}$, and is bounded by $c_{\mathrm{mse}}\beta_\tau^2/s_{\tau,k}$ when $u_k > 0$. Consequently, for any smooth family $V(h)$,

$$\frac{d \log \Delta_{\mathrm{pop}}}{d \log h} = \mathbb{E}_w \left[\frac{2s_{\tau,k}}{\widetilde{s}_{\tau,k}} \varepsilon_k(h) \right], \qquad \varepsilon_k(h) = \frac{d \log |u_k(h)|}{d \log h}, \tag{12}$$

where $\mathbb{E}_w$ averages over (τ, k) with weights proportional to the summands of Eq. (11).

The proof is a direct differentiation (Appendix B). Equation (12) separates the two factors that set the growth of the risk: the order ε_k of the law error, fixed by the integrator, and the discount $2s/\widetilde{s}$, fixed by how strongly the error inflates the corrupted variance where the risk concentrates. Three consequences follow. First, as $h \to 0$ we have $\widetilde{s} \to s$, and a first-order integrator ($\varepsilon_k \to 1$) recovers the h^2 law of Eq. (10). Second, an over-dispersed simulator ($u_k > 0$, as with post-update noise) has $2s/\widetilde{s} < 2$: the risk grows sub-quadratically in the law error and each summand saturates. Third, an under-dispersed simulator ($u_k < 0$, as with in-solve noise) has $2s/\widetilde{s} > 2$ and pays more risk per unit of squared variance error.

Figure 3 evaluates these expressions exactly on the 15×15 grid of the main experiment. Over $h = .00125$–$.01$, the law-error order stays at $\varepsilon \approx 1.01$–1.05 while the weighted discount falls from 1.90 to 1.48, which reproduces the exact risk exponents $1.92 \to 1.55$ (Fig. 3b); $\Delta_{\mathrm{pop}}/h^2$ at $h = .01$ is 58% of $K(1)$. The discount is strongest exactly where the risk sits: at $h = .01$, 99% of the excess comes from the upper half of the wavenumbers and 85% from learning times $\tau \lesssim 0.1$ (Fig. 3c). In-solve noise has about $4\times$ smaller variance error at these steps and 8–$12\times$ smaller population excess, yet costs 0.12–0.14 units of risk per unit of squared variance error against 0.06–0.10 for post-update noise, as the third consequence predicts.

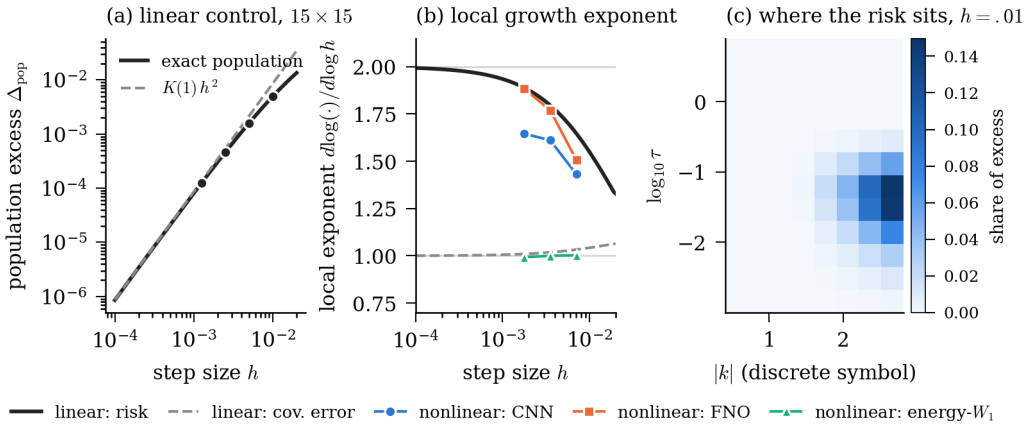

Figure 3: From law error to score risk. *(a)* Exact population excess (Eqs. 9, 8) for a 15×15 linear CHC control (a=.8, κ=1, temperature .3, T=1) against its $K(1)h^2$ asymptote; dots mark the four experimental steps. *(b)* Local growth exponents. Linear control: risk (solid) and summed covariance error (dashed). Nonlinear experiment (Stage A/B arm means, Fig. 4): pairwise exponents $\log_2$ of successive ratios, plotted at geometric-mean steps. *(c)* Share of the linear excess at h=.01 by wavenumber and learning time. Panel (b) compares growth patterns across the two systems.

Finite networks. Trained networks do not attain the population minimizers, so measured contrasts need not follow Eq. (8). Let $\delta = f^\star_{\widetilde{\pi}} - f^\star_\pi$, $e_{\widetilde{\pi}} = \widehat{f}_{\widetilde{\pi}} - f^\star_{\widetilde{\pi}}$, and $e_\pi = \widehat{f}_\pi - f^\star_\pi$, and define $\langle u, v\rangle_\pi = c_{\mathrm{mse}}\mathbb{E}_{\tau\sim\nu, Y\sim p_\tau}[u^\top v]$ with $\|u\|^2_\pi = \langle u, u\rangle_\pi$. Then

$$\mathcal{R}_\pi(\widehat{f}_{\widetilde{\pi}}) - \mathcal{R}_\pi(\widehat{f}_\pi) = \|\delta\|^2_\pi + 2\langle \delta, e_{\widetilde{\pi}}\rangle_\pi + \|e_{\widetilde{\pi}}\|^2_\pi - \|e_\pi\|^2_\pi. \tag{13}$$

The first term is the population excess; the others depend on how well each network fits its own training law. The decomposition yields a simple accuracy condition for the population ordering to survive training.

Corollary 1 (Ordering guarantee)**.** *If* $\|e_{\widetilde{\pi}}\|_\pi + \|e_\pi\|_\pi < \|\delta\|_\pi = \Delta^{1/2}_{\mathrm{pop}}$, *then* $\mathcal{R}_\pi(\widehat{f}_{\widetilde{\pi}}) > \mathcal{R}_\pi(\widehat{f}_\pi)$. *More generally, for two simulated laws with population excesses* $\Delta < \Delta'$ *and network errors* e, e', *the trained models preserve the ordering whenever* $\|e\|_\pi + \|e'\|_\pi < \Delta'^{1/2} - \Delta^{1/2}$.

The corollary follows from Eq. (13) and the triangle inequality (Appendix B). It turns the population theory into a statement about trained models in units of per-pixel RMS noise-prediction error. In the linear control, $\Delta^{1/2}_{\mathrm{pop}}$ ranges from 0.011 at $h = .00125$ to 0.071 at $h = .01$, and adjacent steps differ by 0.010–0.031. Held-out contrasts with shared draws estimate the left-hand side of Eq. (13) directly for trained checkpoints.

Nonlinear invariant-measure controls. The unadjusted bias persists beyond the linear case. On a zero-mass 15×15 double-well CHC Gibbs field, 100 unadjusted steps at $dt = .0025$ give a free-energy-density W_1 of .02143, against .00064 with Metropolis correction and an independent-sample floor of .00071, with agreement across interface and corner-density statistics, smaller grids, and higher-dimensional controls (Appendix Tables 31–35). In nonlinear double-well CHC the coarse step over-weights the high-k tail while leaving the bulk morphology intact, which is the over-dispersed, high-wavenumber error that Proposition 1 identifies as the dominant source of learning risk; an underdamped non-diagonal control shows that correcting the measure improves score fidelity (Appendix C).

6 Nonlinear Learning at Fixed Physical Time

Our main experiment asks whether the simulator step size affects held-out target DSM when the physical horizon of data generation is fixed. We first measure the endpoint law error (Stage A), then train and evaluate score models on the resulting endpoints (Stage B).

Target, initial fields, and simulation. The periodic 15×15 Gibbs field has unit spacing, quadratic coefficient $a = -.7$, quartic coefficient $u = 1$, gradient coefficient $\kappa = 1$, temperature .3, and mobility 1. For each of ten data seeds, all four arms start from the same 1,536 HMC reference fields. The unadjusted stabilized CHC transition (stabilizer 7.35) runs to $T = 1$ with $h = .00125, .0025, .005, .010$, i.e. $800, 400, 200, 100$ transitions. Coarse-step noise increments are normalized sums of the fine increments, so all arms follow one Brownian path on $[0, 1]$. The held-out reference fields are disjoint from the starting fields. We measure endpoint law error by the Wasserstein-1 distance between free-energy-density distributions (energy-W_1).

Matched training and evaluation. Each endpoint dataset trains one circular CNN (231,649 parameters) and one periodic FNO (231,349 parameters) on the shared zero-mass VP task of Section 4.1, with 6,000 Adam updates, batch size 128, and learning rate 2×10^{-4}. Within a backbone and seed, all arms share initialization, batch indices, learning times, and Gaussian noise with the reference-trained comparator. Five evaluation streams of 512 held-out fields each are shared across arms and backbones and averaged within seed; the independent unit is the data/training seed ($n = 10$). We report slopes against $\log_2(h/.00125)$ and largest-minus-smallest-step contrasts, with Student-t_9 and whole-seed bootstrap intervals.

Design. The fixed-time study builds on earlier fixed-transition experiments (100 transitions, so $T = 100h$), which show the same positive trends, and shares their ten-seed cohort and reference-trained models. It was specified before its own runs, with the learning comparison (Stage B) following the endpoint comparison (Stage A); the CNN and FNO are paired observations on this cohort (Appendix F, Table 34).

Endpoint law (Stage A). At fixed $T = 1$, energy-W_1 increased from .010879 at the finest step to .086751 at the coarsest, roughly doubling with each doubling of h. The slope was $+.024926$ (95% CI $[+.024879, +.024972]$), and all ten seed slopes and high–low contrasts were positive. The narrow interval reflects the shared starting fields and Brownian paths (Appendix Table 42).

Held-out target risk (Stage B). Target-DSM excess over the reference-trained model increased with step size for both backbones: from $+.000422$ to $+.010883$ for the CNN and from $+.000276$ to $+.009882$ for the FNO. Slopes were $+.003410$ (CNN) and $+.003128$ (FNO), with 95% Student-t intervals $[+.003326, +.003494]$ and $[+.003012, +.003243]$; high–low contrasts were $+.010461$ and $+.009606$. All ten seed slopes and contrasts were positive for each backbone. Descriptively, the FNO slope was smaller than the CNN slope in every seed (Appendix Table 43). In the fixed-transition cohort the corresponding excess is about a 2% relative increase in target DSM at the largest step (Appendix Table 41).

Growth pattern. On a log–log scale (from arm means), energy-W_1 grows as $h^{1.00}$, while the DSM excess grows with pairwise exponents $1.88, 1.77, 1.51$ (FNO) and $1.65, 1.61, 1.43$ (CNN); the exact linear theory gives $1.89, 1.79, 1.64$ over the same steps (Fig. 3b). The nonlinear field thus shows both signatures of Proposition 1: learned risk grows faster than the first-order law error, and its exponent falls below two as the step grows, as expected for the over-dispersed high-k error of this scheme. The FNO tracks the linear exponents closely. The ordering across steps is preserved in every seed for both backbones, the behavior that Corollary 1 guarantees for sufficiently accurate networks.

6.1 Structural and Downstream Controls

Structural and numerical scope. The linear update fixes the zero Fourier mode; final mean residuals across five seeds range from 6.52×10^{-9} to 1.12×10^{-8} (Fig. 31a). The coupled inverse operator is a synthetic stress test rather than a validated physical solver; its benchmark checks are reported in Appendices D.3 and D.2.

Generated samples. Generated fields recover phase separation, correlation length, and bimodality ($\mathrm{kurt} \approx -1.7$ in 1D, -1.5 in 2D; Appendix Fig. 38), unlike the linear forward. For matched exact-OU, discrete-FDT, and EM training (five paired 4,000-update seeds, the independent unit throughout this subsection; protocols in Appendix D), raw spectrum log-RMSE is $3.245 \pm .027$, $1.865 \pm .076$,

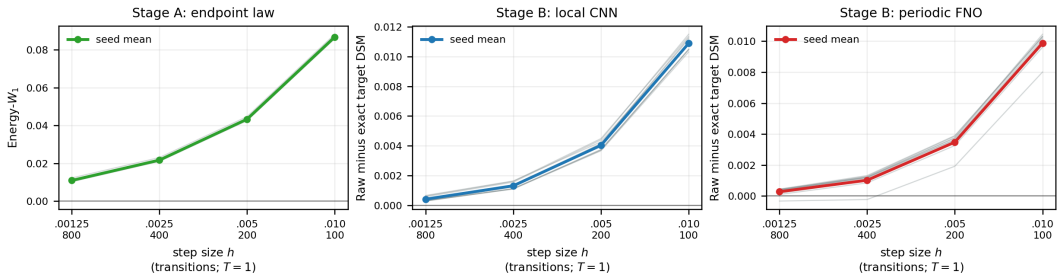

Figure 4: Main nonlinear experiment: physical simulation followed by shared VP score learning. Every arm has $T = 1$; $h = .00125, .0025, .005, .010$ uses $800, 400, 200, 100$ transitions. *Left*: endpoint energy-W_1 increases with step size (Stage A, $n = 10$). *Middle/right*: held-out target-DSM excess increases for the CNN and the FNO trained on the corresponding endpoints (Stage B). "Exact" denotes the HMC reference-trained comparator. Gray curves are seed means (five evaluation streams averaged within seed); all ten seed slopes and high–low contrasts are positive for each backbone.

and $2.830 \pm .091$. A common high-k rolloff reverses the ranking to 1.068, 2.284, and 1.517, while DSM barely changes and the inverse-consistency and closest-to-truth RMSE intervals cross zero (Appendix Table 18). Across $\Delta t = .005, .010, .020$, the raw EM-minus-FDT gap grows with the terminal bias ($+.303$, $+.648$, $+.966$, all intervals positive) and persists to 12,000 updates, while rolloff reverses every gap (Appendix Fig. 16, Table 15). The measure bias is thus visible in raw generated spectra in proportion to its size, and the final ranking is set jointly with the sampler's postprocessing.

Inverse problems. Coupled Allen–Cahn/CHC dynamics define a differentiable melt map $M : c \mapsto \phi_{\text{final}}$; DPS differentiates its data misfit and is compared with Tikhonov, TV, double-well, and oracle-Gaussian Langevin baselines (Algorithm 1, Appendix H). With joint energy as the primary metric (16 ground-truth fields, 24 chains, 120 DPS steps; stress tests in Appendix E), learned priors retain bimodality while classical and oracle-Gaussian baselines blur (Appendix Figs. 36–37). Joint energy favors the cosine arm on ID/OOD morphology and physics under high noise, mismatch, and sparse blur, while other metrics and the strongest guidance setting change the ordering, so inverse performance is best compared per regime and metric.

7 Discussion and Conclusion

Finite-step simulation can conserve mass exactly while changing the distribution it supplies to score learning, and this change can be predicted and removed. The discrete Lyapunov equation gives the bias in closed form and the discrete FDT covariance $Q = C - GCG^\top$ removes it at no cost; the population risk identity, its fixed-time form, and the risk-elasticity identity give the learning cost, and the nonlinear experiment reproduces its signatures with two backbones in every seed. The practical recipe is to check the law the simulator actually samples, repair it where possible, and evaluate the deployed sampler and task directly.

Scope. The nonlinear study uses one 15×15 Gibbs target, two backbones, and a shared ten-seed cohort, with HMC reference fields and energy-W_1 as the endpoint measure; the exact risk theory is derived for linear-Gaussian laws, and the nonlinear comparison is at the level of growth patterns. Extending the finite-network analysis to larger grids, other integrators, and real data is a natural next step.

AI USE STATEMENT

This AI use statement distinguishes writing assistance from research ideation and execution.

Writing assistance. AI tools assisted with manuscript preparation and language polishing. An image-generation tool produced the decorative background and conceptual header glyphs of Fig. 31; labels and connectors were rendered separately, and quantitative panels were plotted from archived measurements and analytic calculations.

Research ideation and execution. AI tools assisted with hypothesis refinement, literature review, experimental-methodology feedback, checks of theoretical derivations and claim boundaries, code and synthetic-data-pipeline implementation, scripted experiments and audits, and result interpretation. The authors checked sources and derivations, approved the designs and runs, validated measurements and artifacts, reviewed all AI-assisted work, made the scientific decisions, and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Appendix D and the anonymous supplement provide exact configurations, commands, seeds, checkpoint hashes, per-run outputs, and automated consistency checks.

REFERENCES

Assyr Abdulle, Gilles Vilmart, and Konstantinos C. Zygalakis. High order numerical approximation of the invariant measure of ergodic sdes. *SIAM Journal on Numerical Analysis*, 52(4):1600–1622, 2014. doi: 10.1137/130935616. URL https://doi.org/10.1137/130935616.

Samuel M. Allen and John W. Cahn. A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta Metallurgica*, 27(6):1085–1095, 1979. ISSN 0001-6160. doi: https://doi.org/10.1016/0001-6160(79)90196-2. URL https://www.sciencedirect.com/science/article/pii/0001616079901962.

Nirmal Baishnab, Ethan Herron, Aditya Balu, Soumik Sarkar, Adarsh Krishnamurthy, and Baskar Ganapathysubramanian. 3d multiphase heterogeneous microstructure generation using conditional latent diffusion models. *Digital Discovery*, 4(11):3175–3190, 11 2025. ISSN 2635-098X. doi: 10.1039/d5dd00159e. URL https://doi.org/10.1039/d5dd00159e.

Lorenzo Baldassari, Ali Siahkoohi, Josselin Garnier, Knut Solna, and Maarten V. de Hoop. Conditional score-based diffusion models for bayesian inference in infinite dimensions. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 24262–24290. Curran Associates, Inc., 2023. doi: 10.52202/075280-1055. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4c79c359b3c5f077c0b955f93cb0f53e-Paper-Conference.pdf.

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 41259–41282. Curran Associates, Inc., 2023. doi: 10.52202/075280-1789. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/80fe51a7d8d0c73ff7439c2a2554ed53-Paper-Conference.pdf.

Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional Image Generation with Score-Based Diffusion Models. *arXiv e-prints*, art. arXiv:2111.13606, November 2021. doi: 10.48550/arXiv.2111.13606.

Nawaf Bou-Rabee and Eric Vanden-Eijnden. Pathwise accuracy and ergodicity of metropolized integrators for sdes. *Communications on Pure and Applied Mathematics*, 63(5):655–696, 2010. doi: https://doi.org/10.1002/cpa.20306. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20306.

Charles-Edouard Bréhier. Approximation of the invariant measure with an Euler scheme for stochastic PDEs driven by space-time white noise. *Potential Analysis*, 40(1):1–40, 2014. ISSN 1572-929X. doi: 10.1007/s11118-013-9338-9. URL https://doi.org/10.1007/s11118-013-9338-9.

John W. Cahn and John E. Hilliard. Free energy of a nonuniform system. i. interfacial free energy. *The Journal of Chemical Physics*, 28(2):258–267, 02 1958. ISSN 0021-9606. doi: 10.1063/1.1744102. URL https://doi.org/10.1063/1.1744102.

Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems, 2024. URL https://arxiv.org/abs/2209.14687.

H.E Cook. Brownian motion in spinodal decomposition. *Acta Metallurgica*, 18(3):297–306, 1970. ISSN 0001-6160. doi: https://doi.org/10.1016/0001-6160(70)90144-6. URL https://www.sciencedirect.com/science/article/pii/0001616070901446.

Evan Scope Crafts and Umberto Villa. Benchmarking diffusion annealing-based bayesian inverse problem solvers. *IEEE Open Journal of Signal Processing*, 6:975–991, 2025.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. doi: 10.1073/pnas.1912789117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1912789117.

Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alexandros G. Dimakis, and Peyman Milanfar. Soft diffusion: Score matching for general corruptions, 2022. URL https://arxiv.org/abs/2209.05442.

Steven Delong, Boyce E. Griffith, Eric Vanden-Eijnden, and Aleksandar Donev. Temporal integrators for fluctuating hydrodynamics. *Phys. Rev. E*, 87:033302, Mar 2013. doi: 10.1103/PhysRevE.87.033302. URL https://link.aps.org/doi/10.1103/PhysRevE.87.033302.

Nan Deng, Yibo Wang, and Wanrong Cao. Ergodicity and invariant measure approximation of the stochastic cahn-hilliard equation via an explicit fully discrete scheme, 2025. URL https://arxiv.org/abs/2512.01621.

Xuehao Ding, H. T. Quan, and Yuhai Tu. Stochastic thermodynamics of score matching in diffusion models, 2026. URL https://arxiv.org/abs/2606.17252.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=CzceR82CYc.

Javier E. Santos, Agnese Marcato, Roman Colman, NIcholas Lubbers, and Yen Ting Lin. Discrete spatial diffusion: Intensity-preserving diffusion modeling. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 126425–126458. Curran Associates, Inc., 2025. doi: 10.52202/085713-4216. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/b7e104e2a26480d316f672de76da18f1-Paper-Conference.pdf.

David J. Eyre. Unconditionally gradient stable time marching the cahn-hilliard equation. *MRS Proceedings*, 529:39, 1998. doi: 10.1557/PROC-529-39.

D. L. Feltham, N. Untersteiner, J. S. Wettlaufer, and M. G. Worster. Sea ice is a mushy layer. *Geophysical Research Letters*, 33(14), 2006. doi: https://doi.org/10.1029/2006GL026290. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2006GL026290.

Robert Gruhlke, Valentin Resseguier, and Merveille Talla. Multiplicative diffusion models: Beyond gaussian latents. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 108324–108412, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b05cd0262afc98ce2536bba9629545ee-Paper-Conference.pdf.

Michael C. Heiber, Andrew A. Herzing, Lee J. Richter, and Dean M. DeLongchamp. Charge transport and mobility relaxation in organic bulk heterojunction morphologies derived from electron tomography measurements. *Journal of Materials Chemistry C*, 8(43):15339–15350, 11 2020. ISSN 2050-7526. doi: 10.1039/d0tc03087b. URL https://doi.org/10.1039/d0tc03087b.

Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, Volodimir Begy, and Gilles Louppe. A crisis in simulation-based inference? beware, your posterior approximations can be unfaithful. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=LHAbHkt6Aq.

Andrew A. Herzing, Lee J. Richter, and Ian M. Anderson. 3d nanoscale characterization of thin-film organic photovoltaic device structures via spectroscopic contrast in the tem 1. *The Journal of Physical Chemistry C*, 114(41):17501–17508, 10 2010. ISSN 1932-7447. doi: 10.1021/jp105883p. URL https://doi.org/10.1021/jp105883p.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

P. C. Hohenberg and B. I. Halperin. Theory of dynamic critical phenomena. *Rev. Mod. Phys.*, 49: 435–479, Jul 1977. doi: 10.1103/RevModPhys.49.435. URL https://link.aps.org/doi/10.1103/RevModPhys.49.435.

Sascha Holl, Jente Vandersanden, Gurprit Singh, and Hans-Peter Seidel. Score-based generative modeling through anisotropic stochastic partial differential equations, 2026. URL https://arxiv.org/abs/2605.08976.

Benjamin Holzschuh, Simona Vegetti, and Nils Thuerey. Solving inverse physics problems with score matching. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 61888–61922. Curran Associates, Inc., 2023. doi: 10.52202/075280-2706. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c2f2230abc7ccf669f403be881d3ffb7-Paper-Conference.pdf.

Emiel Hoogeboom and Tim Salimans. Blurring diffusion models, 2024. URL https://arxiv.org/abs/2209.05557.

Jiahe Huang, Guandao Yang, Zichen Wang, and Jeong Joon Park. Diffusionpde: Generative pde-solving under partial observation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 130291–130323. Curran Associates, Inc., 2024. doi: 10.52202/079017-4140. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/eb3878c1dcbfff9ee95d5d033e5f5942-Paper-Conference.pdf.

Phuoc Toan Huynh, Ruth Y. López Fajardo, Guannan Zhang, Lili Ju, and Feng Bao. A score-based diffusion model approach for adaptive learning of stochastic partial differential equation solutions. *Journal of Computational Physics*, 556:114814, 2026. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2026.114814. URL https://www.sciencedirect.com/science/article/pii/S0021999126001646.

A.M. Jokisaari, P.W. Voorhees, J.E. Guyer, J. Warren, and O.G. Heinonen. Benchmark problems for numerical implementations of phase field models. *Computational Materials Science*, 126:139–151, 2017. ISSN 0927-0256. doi: https://doi.org/10.1016/j.commatsci.2016.09.022. URL https://www.sciencedirect.com/science/article/pii/S0927025616304712.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022. doi: 10.52202/

068431-1926. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a98846e9d9cc01cfb87eb694d946ce6b-Paper-Conference.pdf.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23593–23606. Curran Associates, Inc., 2022. doi: 10.52202/068431-1714. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/95504595b6169131b6ed6cd72eb05616-Paper-Conference.pdf.

Gavin Kerrigan, Justin Ley, and Padhraic Smyth. Diffusion generative models in infinite dimensions, 2023. URL https://arxiv.org/abs/2212.00886.

Chieh-Hsin Lai, Yuhta Takida, Naoki Murata, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. FP-diffusion: Improving score-based diffusion models by enforcing the underlying score fokker-planck equation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18365–18398. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/lai23d.html.

Rémi Laumont, Valentin De Bortoli, Andrés Almansa, Julie Delon, Alain Durmus, and Marcelo Pereyra. Bayesian imaging using plug & play priors: When langevin meets tweedie. *SIAM Journal on Imaging Sciences*, 15(2):701–737, 2022. doi: 10.1137/21M1406349. URL https://doi.org/10.1137/21M1406349.

Benedict Leimkuhler and Charles Matthews. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*. Interdisciplinary Applied Mathematics. Springer, United Kingdom, May 2015. ISBN 978-3-319-16374-1. doi: 10.1007/978-3-319-16375-8.

Pablo Lemos, Adam Coogan, Yashar Hezaveh, and Laurence Perreault-Levasseur. Sampling-based accuracy testing of posterior estimators for general inference. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19256–19273. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/lemos23a.html.

Zeyu Li, Hongkun Dou, Shen Fang, Wang Han, Yue Deng, and Lijun Yang. Physics-aligned field reconstruction with diffusion bridge. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 58864–58900, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9421261e06f1a63a352b068f1ac90609-Paper-Conference.pdf.

Jae Hyun Lim, Nikola B. Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, Christopher Pal, Arash Vahdat, and Anima Anandkumar. Score-based diffusion models in function space. *Journal of Machine Learning Research*, 26(158):1–62, 2025. URL http://jmlr.org/papers/v26/23-1472.html.

Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 28027–28053, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/7711026951f3dc3afb2be427bf98bc73-Paper-Conference.pdf.

S. Maus, M. Schneebeli, and A. Wiegmann. An x-ray micro-tomographic study of the pore space, permeability and percolation threshold of young sea ice. *The Cryosphere*, 15(8):4047–4072, 2021. doi: 10.5194/tc-15-4047-2021. URL https://tc.copernicus.org/articles/15/4047/2021/.

Badr MOUFAD, Yazid Janati el idrissi, Lisa Bedin, Alain Oliviero Durmus, randal douc, Eric Moulines, and Jimmy Olsson. Variational diffusion posterior sampling with midpoint guidance. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 95170–95214, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ed524bb14de1b52c8522b977ded241d3-Paper-Conference.pdf.

Kushagra Pandey and Stephan Mandt. A complete recipe for diffusion generative models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4238–4249, 2023. doi: 10.1109/ICCV51070.2023.00393.

W. Pazner, N. Trask, and P.J. Atzberger. Stochastic discontinuous galerkin methods (sdgm) based on fluctuation-dissipation balance. *Results in Applied Mathematics*, 4:100068, 2019. ISSN 2590-0374. doi: https://doi.org/10.1016/j.rinam.2019.100068. URL https://www.sciencedirect.com/science/article/pii/S2590037419300688.

Bonnie Raffel, Alexandra Touzeau, Harald Sodemann, Benjamin Allen Lange, Anca Cristea, Dmitry V Divine, and Mats A Granskog. Stratigraphy of sea-ice cores collected in the Eurasian basin during CAATEX 2019 cruise. PANGAEA, 2023. doi: 10.1594/PANGAEA.955715. URL https://doi.org/10.1594/PANGAEA.955715. In: Raffel, B et al. (2023): Physical and chemical properties of different summer sea-ice types, sea water and precipitation in the Eurasian basin during the CAATEX 2019 cruise [dataset bundled publication]. PANGAEA, https://doi.org/10.1594/PANGAEA.954891.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation, 2023. URL https://arxiv.org/abs/2206.13397.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. ISSN 13507265. URL http://www.jstor.org/stable/3318418.

François Rozet and Gilles Louppe. Score-based data assimilation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 40521–40541. Curran Associates, Inc., 2023. doi: 10.52202/075280-1763. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7f7fa581cc8a1970a4332920cdf87395-Paper-Conference.pdf.

Luca Scimeca, Thomas Jiralerspong, Berton Earnshaw, Jason Hartford, and Yoshua Bengio. Learning what matters: Steering diffusion via spectrally anisotropic forward noise, 2025. URL https://arxiv.org/abs/2510.09660.

Raghav Singhal, Mark Goldstein, and Rajesh Ranganath. What's the score? automated denoising score matching for nonlinear diffusions. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

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

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

Yue Song, Andy Keller, Sevan Brodjian, Takeru Miyato, Yisong Yue, Pietro Perona, and Max Welling. Kuramoto orientation diffusion models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 4047–4076. Curran Associates, Inc., 2025. doi: 10.52202/085713-0143. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/062649fadaed89b5bcdc74359db8a35b-Paper-Conference.pdf.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration, 2020. URL https://arxiv.org/abs/1804.06788.

M. Grae Worster. Convection in Mushy Layers. *Annual Review of Fluid Mechanics*, 29:91–122, January 1997. doi: 10.1146/annurev.fluid.29.1.91.

Martin Zach, Youssef Haouchat, and Michael Unser. A statistical benchmark for diffusion-posterior-sampling algorithms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 108984–109021, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b1449228e8841aa6844e516e043d107c-Paper-Conference.pdf.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20895–20905, 2025. doi: 10.1109/CVPR52734.2025.01946.

Weitong Zhang, Chengqi Zang, Liu Li, Sarah Cechnicka, Cheng Ouyang, and Bernhard Kainz. Stability and generalizability in sde diffusion models with measure-preserving dynamics. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 81606–81644. Curran Associates, Inc., 2024. doi: 10.52202/079017-2593. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/948b1c9d660d7286dd767cd07dabd487-Paper-Conference.pdf.

Zihan Zhou, Xiaoxue Wang, and Tianshu Yu. Generating physical dynamics under priors. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 79515–79549, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/c573258c38d0a3919d8c1364053c45df-Paper-Conference.pdf.

A Proof of the covariance audit

Suppose $x^n \sim \mathcal{N}(0, C)$ and that x^n and η^n are independent. Then x^{n+1} is Gaussian with zero mean and covariance

$$\mathrm{Cov}(x^{n+1}) = GCG^\top + Q. \tag{14}$$

Consequently, $\mathcal{N}(0, C)$ is invariant if and only if $C = GCG^\top + Q$, which is equivalent to Eq. (3). Conversely, if Eq. (3) holds and $Q \succeq 0$, the same calculation proves invariance. For any supplied Q and $\rho(G) < 1$, the unique stationary covariance is

$$V = \sum_{j=0}^{\infty} G^j Q (G^\top)^j, \tag{15}$$

which is the unique solution of $V = GVG^\top + Q$. The scalar relation used for each CHC Fourier mode follows by setting $G = g_k$, $C = v_\infty(k)$, and $Q = \Sigma_k$.

Proof and conditions for Eq. (8). Identify the finite-dimensional modeled subspace $\mathcal{S}$ with $\mathbb{R}^r$ through an orthonormal basis. Let π and $\widetilde{\pi}$ be Borel probability laws on $\mathcal{S}$, use the same measurable families of linear maps $A_\tau, L_\tau : \mathcal{S} \to \mathcal{S}$ and time law ν for both, sample X, ϵ, τ independently, and assume that $L_\tau L_\tau^\top \succ 0$ for ν-almost every τ. Gaussian convolution then gives positive smooth densities $p_\tau, \widetilde{p}_\tau$. We additionally require the displayed cross-risk to be finite; injectivity of A_τ is needed only for the final strictness claim.

For fixed τ, conditional expectation is the L^2 projection, hence

$$\mathcal{R}_{\pi,\tau}(f) - \mathcal{R}_{\pi,\tau}(f_\pi^\star) = c_{\mathrm{mse}}\, \mathbb{E}_{p_\tau} \|f - f_\pi^\star\|_2^2, \qquad f_\pi^\star(y,\tau) = \mathbb{E}_\pi[\epsilon \mid Y_\tau = y]. \tag{16}$$

Differentiating the Gaussian mixture on $\mathcal{S}$ gives

$$\nabla \log p_\tau(y) = -L_\tau^{-\top} \mathbb{E}_\pi[\epsilon \mid Y_\tau = y], \qquad f_\pi^\star(y,\tau) = -L_\tau^\top \nabla \log p_\tau(y), \tag{17}$$

and the same identity holds with π replaced by $\widetilde{\pi}$. Substituting $f = f_{\widetilde{\pi}}^\star$ into Eq. (16) and integrating over ν proves Eq. (8). Its integrand vanishes only when the two positive densities have the same score; on the connected space $\mathcal{S}$, normalization then gives $p_\tau = \widetilde{p}_\tau$. Finally, wherever A_τ is injective, equality of these Gaussian convolutions implies $\pi = \widetilde{\pi}$: their characteristic functions satisfy $\varphi_{p_\tau}(t) = \varphi_\pi(A_\tau^\top t) e^{-t^\top L_\tau L_\tau^\top t/2}$, so the nonzero Gaussian factor can be divided out and $A_\tau^\top$ ranges over $\mathcal{S}$. Thus, if A_τ is injective for ν-almost every τ, distinct endpoint laws give strictly positive population excess. This proof concerns a fixed-horizon endpoint marginal under the audited initial law, not the full nonlinear Markov kernel, path law, intermediate marginals, dose monotonicity, or finite-network ordering.

B Proof of the fixed-time Gaussian risk law

For the stable linear IMEX control in the main text, use h for the physical step and write $P_k = ak^2$, $Q_k = \kappa k^4$, $\lambda_k = P_k + Q_k$, and $c_k = v_\infty(k) = k_{\mathrm{B}}\vartheta/(a + \kappa k^2)$. The scalar update has

$$g_k(h) = \frac{1 - P_k h}{1 + Q_k h}, \qquad B_k(h) = \frac{2(1 + Q_k h)^2}{2 + (Q_k - P_k)h},$$

and its stationary variance is $c_k B_k(h)$. If the initial variance is c_k, the finite-horizon covariance recursion gives

$$v_{h,T,k} = g_k(h)^{2T/h} c_k + \left[1 - g_k(h)^{2T/h}\right] c_k B_k(h),$$

which proves Eq. (9).

Fix a finite nonzero mode set, $T > 0$, and $h = T/n \to 0$. Taylor expansion gives

$$B_k(h) - 1 = \frac{P_k + 3Q_k}{2} h + O(h^2), \qquad \log g_k(h) = -\lambda_k h + O(h^2).$$

Consequently,

$$v_{h,T,k} - c_k = d_k(T) h + O(h^2), \qquad d_k(T) = c_k \frac{P_k + 3Q_k}{2} (1 - e^{-2\lambda_k T}).$$

For the VP kernel, set $\alpha_\tau = e^{-\tau/2}$, $\beta_\tau^2 = 1 - e^{-\tau}$, $s_{\tau,k} = \alpha_\tau^2 c_k + \beta_\tau^2$, and $\widetilde{s}_{h,T,\tau,k} = \alpha_\tau^2 v_{h,T,k} + \beta_\tau^2$. Gaussian conditional expectation gives $f_\pi^\star = \beta_\tau S_\tau^{-1} y$ and $f_{h,T}^\star = \beta_\tau \widetilde{S}_{h,T,\tau}^{-1} y$. Evaluating their squared difference under the target corrupted law yields the exact modal formula

$$\mathcal{R}_\pi(f_{h,T}^\star) - \mathcal{R}_\pi(f_\pi^\star) = c_{\mathrm{mse}} \mathbb{E}_\tau \sum_k \frac{\beta_\tau^2 \alpha_\tau^4 (v_{h,T,k} - c_k)^2}{s_{\tau,k} \widetilde{s}_{h,T,\tau,k}^2}.$$

The target covariance $s_{\tau,k}$, rather than $\widetilde{s}_{h,T,\tau,k}$, in the first factor follows because risk is evaluated on π. The VP time law in the paper is log-uniform on $[.001, 8]$; all denominators are then uniformly positive. Substituting the variance expansion into the last display gives Eq. (10); the finite mode set makes the $O(h^3)$ remainder uniform and integrable. This completes the proof.

Proof of Proposition 1. Equation (11) is the modal formula above with $u_k = v_k - c_k$ and $\widetilde{s} = s + \alpha^2 u$. Fix (τ, k) and write $\phi(u) = u^2/(s\,\widetilde{s}^2)$. Since $du/d\log|u| = u$ for either sign of u,

$$\frac{\partial \log \phi}{\partial \log |u|} = 2 - \frac{2\alpha^2 u}{s + \alpha^2 u} = \frac{2s}{\widetilde{s}} > 0.$$

For $u > 0$, $\widetilde{s} \geq \alpha^2 u$ gives $\beta^2 \alpha^4 \phi \leq \beta^2/s$. For a smooth family $V(h)$, differentiate $\log \Delta_{\mathrm{pop}} = \log \mathbb{E}_\tau \sum_k T_{\tau,k}$, where $T_{\tau,k}$ is the summand, and use $d \log T_{\tau,k}/d \log h = (2s/\widetilde{s})\, \varepsilon_k$; the weights $T_{\tau,k}/\sum T$ give Eq. (12). As $h \to 0$, $u_k \to 0$ and $\widetilde{s} \to s$, so the discount tends to 2. For the 15×15 control of Fig. 3, a central finite difference of $\log \Delta_{\mathrm{pop}}$ agrees with Eq. (12) to four decimals at every experimental step (1.9197, 1.8488, 1.7291, 1.5499).

Proof of Corollary 1. By Eq. (13), $\mathcal{R}_\pi(\widehat{f}_{\widetilde{\pi}}) - \mathcal{R}_\pi(\widehat{f}_\pi) = \|\delta + e_{\widetilde{\pi}}\|_\pi^2 - \|e_\pi\|_\pi^2 \geq (\|\delta\|_\pi - \|e_{\widetilde{\pi}}\|_\pi)^2 - \|e_\pi\|_\pi^2$, which is positive when $\|e_{\widetilde{\pi}}\|_\pi + \|e_\pi\|_\pi < \|\delta\|_\pi$. For two simulated laws, write $\mathcal{R}_\pi(\widehat{f}) - \mathcal{R}_\pi(f_\pi^\star) = \|\delta + e\|_\pi^2$ for each; then $\|\delta' + e'\|_\pi \geq \|\delta'\|_\pi - \|e'\|_\pi > \|\delta\|_\pi + \|e\|_\pi \geq \|\delta + e\|_\pi$ under the stated condition, with $\|\delta\|_\pi^2 = \Delta$ and $\|\delta'\|_\pi^2 = \Delta'$.

Matrix form and numerical checks. For general, not necessarily commuting, Gaussian laws $\mathcal{N}(0, C)$ and $\mathcal{N}(0, V)$ on $\mathcal{S}$, let $S_\tau = A_\tau C A_\tau^\top + L_\tau L_\tau^\top$, $\widetilde{S}_\tau = A_\tau V A_\tau^\top + L_\tau L_\tau^\top$, and $D_\tau = \widetilde{S}_\tau^{-1} - S_\tau^{-1}$. The Bayes noise predictors are $L_\tau^\top S_\tau^{-1} y$ and $L_\tau^\top \widetilde{S}_\tau^{-1} y$, and Eq. (8) becomes

$$\mathcal{R}_\pi(f_V^\star) - \mathcal{R}_\pi(f_C^\star) = c_{\mathrm{mse}}\, \mathbb{E}_{\tau \sim \nu}\, \mathrm{tr}\big[L_\tau^\top D_\tau S_\tau D_\tau^\top L_\tau\big],$$

which is zero if and only if $V = C$ when L_τ is invertible on a set of positive ν-measure. With a real Fourier basis each real degree of freedom is counted once; an rFFT implementation must weight each non-Nyquist mode by two. We checked the formulas numerically. For the 1D linear control of Fig. 2 (N=128, L=128, a=.8, κ=1, $k_{\mathrm{B}}\vartheta$=1.8, T=1): (i) the direct covariance recursion matches Eq. (9) to relative error 2×10^{-13} for $h \in \{.00125, \ldots, .02\}$; (ii) Gauss–Legendre quadrature in $\log \tau$ with 100–800 nodes agrees to 13 digits; (iii) a Monte Carlo estimate with the analytic predictors at h=.01 gives .005626 against the exact .005623; (iv) $\Delta_{\mathrm{pop}}/h^2$ increases from 56.2 at h=.01 to 128.2 at h=1/6400, approaching $K(1)$=130.5; and (v) discrete-FDT noise gives exactly zero excess. For a fixed non-diagonal 6×6 example the trace formula gives .03121 and a 4×10^5-sample direct estimate .03131.

Pre-asymptotic exponents. Figure 3 uses the 15×15 periodic grid with the discrete Laplacian symbol $\hat{K} = 4\sin^2(k_x/2) + 4\sin^2(k_y/2)$ in place of $|k|^2$ and c_{mse}=1/225. Over h = .00125, .0025, .005, .01 the pairwise risk exponents are 1.89, 1.79, 1.64 and the covariance-error exponents 1.01, 1.02, 1.03; changing $a \in \{.3, .8\}$ or the temperature in $\{.3, 1.8\}$ moves the risk exponents by at most 0.02. Table 1 compares them with exponents computed from the arm means of the nonlinear experiments. These are post-hoc descriptive summaries without intervals; seed-level exponents would require the per-seed contrasts.

system	quantity	h: .00125→.0025	.0025→.005	.005→.01
linear 15×15 (exact)	covariance error	1.01	1.02	1.03
linear 15×15 (exact)	population risk	1.89	1.79	1.64
nonlinear, fixed T=1	energy-W_1	0.99	1.00	1.00
nonlinear, fixed T=1	FNO target-DSM excess	1.88	1.77	1.51
nonlinear, fixed T=1	CNN target-DSM excess	1.65	1.61	1.43
nonlinear, 100 steps	energy-W_1	1.03	1.01	1.01
nonlinear, 100 steps	FNO target-DSM excess	2.00	1.79	1.52
nonlinear, 100 steps	CNN target-DSM excess	1.79	1.64	1.47

Table 1: Pairwise growth exponents $\log_2(x_{2h}/x_h)$. Nonlinear rows use arm means from Tables 41 and 43.

C Additional finite-step scope controls

Relation to discrete fluctuation–dissipation. The covariance identity is classical (Delong et al., 2013; Pazner et al., 2019). For any scalar one-step linear update with amplification g_k and a post-update Euler–Maruyama increment $\Sigma_k^{\mathrm{EM}} = \sigma_k^2 dt = 2\lambda_k v_\infty dt$,

$$\frac{V_k}{v_\infty(k)} = \frac{2\lambda_k dt}{1 - g_k^2}. \tag{18}$$

Table 2 evaluates this specific post-update-noise construction for representative drift amplifications. It does not imply that every implementation of implicit Euler or Crank–Nicolson uses this noise placement. With the stated placement, even exact deterministic amplification $g_k = e^{-\lambda_k dt}$ is biased unless the noise covariance is also integrated exactly. For an implicit-Euler spot check at k=2.5, EM noise gives 2.454 (analytic 2.457) and discrete-FDT noise gives 1.005.

Stable integrators sampling biased invariant measures are classical (Leimkuhler & Matthews, 2015; Roberts & Tweedie, 1996). Our contribution is the closed-form modal bias for this CHC–IMEX split and its field-level consequence in a physics-defined generative forward. Because the terminal measure is part of the generative construction, the bias enters the training target or terminal sampler unless corrected.

Resolution and stiffness scope. A frozen 960-combination analytic sweep holds physical length fixed while varying resolution, step, stiffness, curvature, and temperature (Appendix Table 3 and Fig. 6). At $dt = .02$, $a = .8$, and $\kappa = 1$, the maximum bias rises from 1.016 at $N = 32$ to 63.6 at $N = 256$; every tested fixed-step refinement and stiffness slice is strictly increasing. The dimensionless ratio is temperature-independent, whereas discrete FDT remains exactly one throughout. Thus spatial refinement without a commensurate step/noise redesign can amplify rather than cure measure error.

Field-level consequence (Fig. 5). For 512 two-dimensional melted-equilibrium fields, the IMEX+EM scheme produces a high-k upturn in $S(k)$ (up to $15\times$ the correct value) and correlation length ξ=2.7, versus 3.8 for the measure-exact ensemble. Discrete-FDT noise matches the exact-OU spectrum mode for mode.

Matrix-valued generality control. Beyond conserved Fourier-diagonal CHC, we audit an underdamped Langevin oscillator with non-diagonal drift and target Gibbs covariance $C = I$. Two backward-Euler substeps give $\rho(G) = .806$. Discrete FDT preserves C; continuum noise passed through the same solve gives $C_{\mathrm{EM}} = \left[\begin{smallmatrix} .841 & .075 \\ .075 & .751 \end{smallmatrix}\right]$, with temperature eigenvalue ratios .708, .883 and spurious q–p correlation .094. Across five paired noise-predictor seeds, BE+EM has higher relative score MSE than BE+FDT (difference .00054, 95% interval [.00003, .00111]), yet its colder channel is easier to denoise (.385 vs. .420 MSE), matching the quadrature-exact oracle ordering (Fig. 18, Appendix Table 19). Thus correcting the measure improves score fidelity while a biased channel can define an easier task.

Nonlinear scope control. We repeat the equilibrium experiment for nonlinear double-well CHC with 256 two-dimensional fields over the same physical time and $dt \in \{0.02, 0.02/4, 0.02/8\}$.

drift update	g_k	$B(2.5)$	$B(3)$
explicit Euler	$1-\lambda dt$	1.79	8.47^*
implicit Euler	$1/(1+\lambda dt)$	2.46	4.06
Crank–Nicolson	$\frac{1-\lambda dt/2}{1+\lambda dt/2}$	2.08	3.54
exact drift	$e^{-\lambda dt}$	2.13	3.63
IMEX split	$\frac{1-p}{1+q}$	2.37	3.95
IMEX, noise inside solve	$\frac{1-p}{1+q}$	0.75	0.58
any + discrete FDT	—	1.00	1.00

Table 2: Stationary-measure bias $B(k) = V_k/v_\infty(k)$ for representative drift amplifications combined with the same post-update EM increment (dt=0.02). The "noise inside solve" row passes the same increment through the implicit operator, $c^{n+1} = [(1-p)c^n + \xi]/(1+q)$, giving $B = 2/(2+q-p)$; 8192 simulated chains give 0.746 and 0.575 (analytic 0.746, 0.575). The discrete-FDT covariance (7) removes the bias for each linear update. "Exact drift" is not exact OU unless its noise covariance is integrated consistently. *Explicit EM approaches instability near Nyquist.

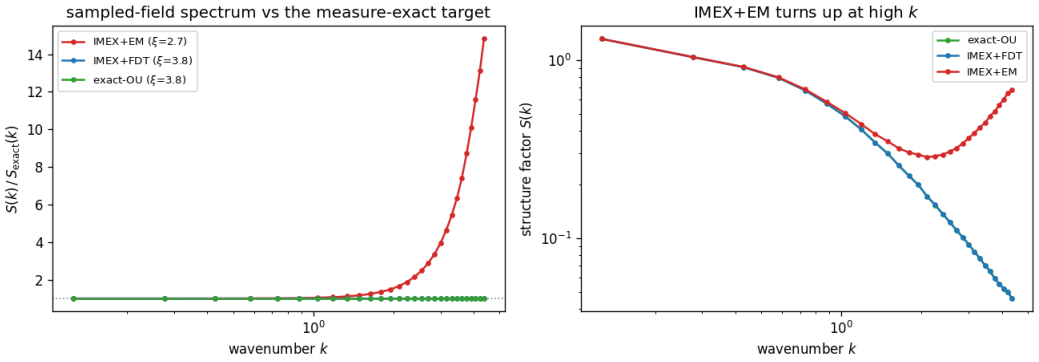

Figure 5: Field-level consequence. Sampling 512 2D melted-equilibrium fields by time-stepping the SDE: relative to the measure-exact ensemble (*left*, ratio of structure factors), the energy-stable IMEX with Euler–Maruyama noise (red) over-weights high k by up to $15\times$ and yields a 30%-too-short correlation length (ξ=2.7 vs 3.8); the FDT-corrected scheme (blue) matches the exact-OU target mode for mode (*right*, spectra).

Relative to the finest run, the coarse scheme over-weights only the extreme high-k tail (up to roughly $6\times$), while all three runs give $\xi = 17.1$ and the displayed samples retain similar bulk morphology (Fig. 27). This negative scope result matters: the linear counterexample establishes non-preservation, but a large error in physically salient nonlinear observables does not follow automatically.

D Reproducibility Details

D.1 Coupled Salt–Ice Operator and Internal Checks

For the inverse problem we couple a non-conserved ice phase ϕ ($\phi = +1$ ice, -1 liquid) to the conserved salt field c through the free energy

$$\mathcal{F}[\phi, c] = \int \tfrac{\kappa_\phi}{2}|\nabla\phi|^2 + \tfrac{a}{4}(\phi^2-1)^2 + \tfrac{\kappa_c}{2}|\nabla c|^2 + \tfrac{\alpha}{2}c^2 - \beta\, c\, h(\phi)\, dx, \tag{19}$$

N	$dt = .00125$	.0025	.005	.01	.02
32	1	1	1	1.01	1.02
64	1.01	1.03	1.05	1.1	1.21
128	1.19	1.39	1.81	2.69	4.59
256	4.45	8.24	16	31.9	63.6

Table 3: Maximum modal stationary-variance ratio for IMEX drift plus post-update Euler–Maruyama noise at fixed physical domain length $L = 128$, $a = .8$, and $\kappa = 1$. Fixed-step refinement exposes higher wavenumbers and increases the bias; the discrete-FDT covariance gives ratio one in every cell.

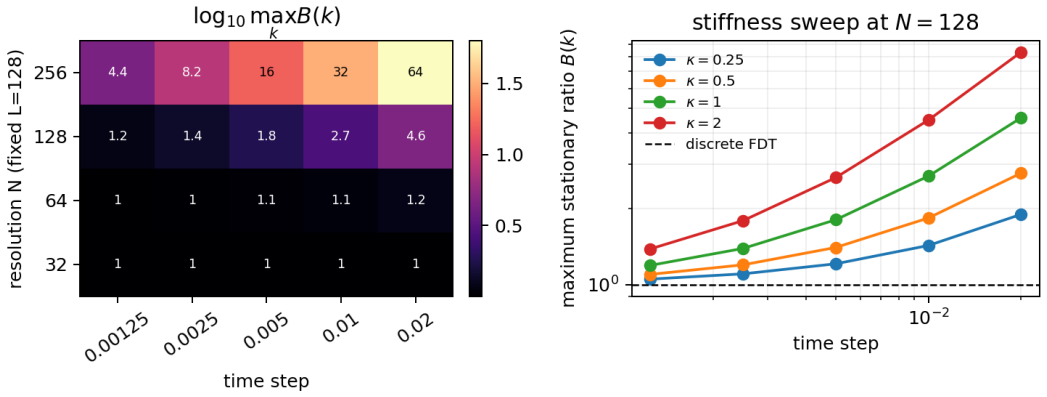

Figure 6: Frozen CHC discretization scope map. *Left*: at fixed physical length, fixed-step refinement exposes stiffer modes and can sharply increase the maximum stationary bias. *Right*: the same bias grows with both step and gradient stiffness at $N = 128$. The discrete-FDT covariance stays at ratio one for all 960 tested combinations; temperature scales absolute covariance but cancels from this ratio.

with liquid indicator $h(\phi) = \frac{1}{2}(1 - \phi)$ (brine rejection: salt prefers liquid). The dynamics are Allen–Cahn for the non-conserved ice and Cahn–Hilliard for the conserved salt,

$$\partial_\tau \phi = -M_\phi \tfrac{\delta \mathcal{F}}{\delta \phi} + \xi_\phi, \qquad \partial_\tau c = M_c \nabla^2 \tfrac{\delta \mathcal{F}}{\delta c} + \xi_c, \tag{20}$$

each with its own FDT noise (non-conserved for ϕ, conserved for c). Note the salt free energy is purely quadratic ($\frac{\alpha}{2}c^2$, no double well of its own): the bimodal "brine channel" structure in the data is *induced* by coupling to the bimodal ϕ, i.e. it is a spatially correlated field statistic, not a pointwise value statistic. This motivates a learned *field* prior: a pointwise double-well penalty matches preferred values but not spatial correlations. Salt depresses the local freezing point, so as temperature rises the melt operator $M : c \mapsto \phi_{\text{final}}$—all-ice relaxed under the salt-tilted Allen–Cahn to a graded steady state—melts high-salt regions first. Internal tests check mass conservation to $\sim 10^{-8}$ (T1), non-conservation of ϕ (T2), monotone energy dissipation (T3), the FDT spectrum (T4), and brine segregation $\mathrm{corr}(\phi, c) \approx -0.95$, in 1D, 2D, and 3D.

All stochastic components use explicit seeded random-number generators. The anonymous supplement contains the data-generation, training, sampling, evaluation, and plotting scripts together with exact commands and recorded outputs. The main text reports the grid sizes, model widths, training iterations, reverse-step counts, physical constants, ensemble sizes, and hardware. Grids are $N{=}128$ (1D), 64^2 (2D), and 32^3 (3D). The score networks use 48–64 channels and up to six FiLM blocks, trained for 4000–6000 iterations in the main comparison; reverse sampling uses 120–160 steps. GPU runs use FP32 on one 24-GB NVIDIA RTX 4090; the underdamped MLP control runs on CPU. The recorded environment is Python 3.11.9, PyTorch 2.6.0 (CUDA 12.4 build), NumPy 2.4.6, and SciPy 1.17.1. The OPV audit sets the cuBLAS workspace before CUDA initialization and enables PyTorch deterministic algorithms plus deterministic cuDNN; every formal row records execution version `torch_deterministic_v1`. Confirmatory energy scores use every off-diagonal ensemble pair (implementation version `exact_energy_u_stat_v1`); unconditional scores additionally average every generated–held-out cross pair and divide by the square root of field dimen-

sion. The paired learning control uses one shared 384-field dataset (288 training and 96 held out), identical initial weights and random training streams within each seed, five training seeds, 4000 iterations, a 64-channel six-block network, and 128 fixed-evaluation generated fields. Its registered dose-response extension holds these choices fixed at $\Delta t = .005, .010, .020$, pairs weights, batches, times, and Gaussian innovations, and uses 160 reverse steps. Raw spectral log-RMSE is primary; rolloff spectral error and raw energy are secondary. The .020 endpoint is the disclosed outcome-known anchor, so the positive label additionally requires both new cells, both adjacent increments, their intervals, and every leave-one-seed-out mean to be positive. Its fixed inverse evaluation uses the same 16 truths and observation noises across training seeds, 24 chains per truth, 120 DPS steps, and 40 melt steps. The matched cosine-VP control uses the same training budget and architecture and normalizes only at the network interface by the training-split standard deviation. Both inverse arms use reconstruction–fresh-noise ancestral transitions with matched sampler randomness, guidance, rolloff, clipping, chains, and steps. Shared physics constants are $M{=}\kappa{=}1$, $a_{\text{high}}{=}0.8$, and $kT{=}1.8$. The frozen downstream stress suite reuses those five training seeds and evaluates four new regimes with eight fixed truths, 24 chains per truth, and 120 DPS steps: observation-noise standard deviation .05; melt observations generated at $\beta = 1.8$ for 60 steps but inverted with $\beta = 1.5$ for 40 steps; OOD morphologies generated with 2500 rather than 8000 simulator steps; and a circular Gaussian blur of width two followed by a fixed 25% Bernoulli mask and noise standard deviation .02. Joint posterior energy score was frozen as primary before these runs. A direction is promoted only when paired training-seed and two-level seed–truth 95% bootstrap intervals exclude zero with the same sign. The separate guidance sweep reports every predeclared $\zeta_0 \in \{.5, 1, 2, 4\}$ using five seeds, four truths, 16 chains, and 80 DPS steps; no coefficient is selected after observing outcomes.

Task-specific conditional controls. We train the same 64-channel, six-block cosine-VP backbone with a zero-initialized observation branch. Initialization, batches, times, and salt noise are seed-paired; observation noise has an isolated stream. A registered nine-channel extension adds the mask, operator family, and parameters, trains uniformly over three melt and three blur configurations, and holds out the test blur pair. Both use matched ancestral sampling without a likelihood gradient and trade prior reuse for paired task data; neither is a forward-schedule arm. The frozen conditional control uses 288 training and 96 held-out salt–melt pairs, five seeds, 4000 iterations, batch size 128, learning rate 2×10^{-4}, and the same 16 ID truths, 24 chains, and 120 reverse steps. Its primary contrast is joint energy score against each DPS arm. Directional language requires both the paired-seed and two-level seed–truth 95% intervals to exclude zero with the same sign. Without retraining, operator metadata, masks, or renormalization, the fixed ID model is also evaluated on all four eight-truth stress regimes; the sparse-blur result is explicitly an asymmetric measurement-channel shift. The operator-aware extension keeps the five-seed, 4000-update, 24-chain, and 120-step budgets but trains one zero-initialized nine-channel branch uniformly over three melt and three sparse-blur configurations. Operator choices, masks, and observation noise use isolated deterministic generators. The held-out blur pair $(\sigma{=}2, .25)$ is absent from training; its exact operator metadata is supplied at test time. The cross-operator positive label requires both melt and sparse-blur primary contrasts to favor the model under both paired-seed and two-level intervals. The paired-data extension keeps the same architecture, five seeds, 4000 updates, and fixed ID evaluation while using nested prefixes of 72 and 144 pairs; the 288-pair point reuses the original formal result. The scalar salt normalization is computed from the full 288-field training split, while observation normalization uses only the available prefix. All three points and all contrasts against both DPS arms are retained. The underdamped control uses unit mass, stiffness, friction, and reference temperature; each learned level comprises two backward-Euler substeps of size .2. Its double-well phase-space data use temperature .16. A three-hidden-layer, 128-unit MLP is trained for 3000 iterations on 8192 samples with batch size 512, using five paired initializations and random streams. Fixed evaluation uses 2048 held-out points at physical time 2.0; the exact marginal score and posterior mean are computed by one-dimensional quadrature. The separate exact-posterior sampler control uses the same three audited 20-level kernels without a learned network. It evaluates five paired sampler seeds with 1024 samples per method against 8192 fixed IID posterior samples. A 1601-point grid on $q \in [-3, 3]$ computes both prior and posterior marginal scores; all-pair energy score is primary and all four predeclared guidance coefficients are reported. Paired intervals and leave-one-seed-out means govern directional language. For the high-dimensional sampler control, the cosine epsilon predictor at $t_\star = .05$ induces the smoothed-prior score $s_\star(c) = -\epsilon_\theta(\sqrt{\bar\alpha_\star}c/d, t_\star)/(\sigma_\star d)$, where d is the training standard deviation and $\sigma_\star^2 = (1-\bar\alpha_\star)/\bar\alpha_\star$. We run $c_{j+1} = c_j + \eta[s_\star(c_j) + \nabla \log p(y \mid c_j)] + \sqrt{2\eta}\,\xi_j$ for 400 steps after a 120-step unconditional initialization. Training-only, two-start stability calibra-

run	$F(10^5)$	vs MEU.	vs FEn.	vs nominal	mass drift
128^2, $dt = .1$	24.037	0.0188	0.0555	0.0000	1.1e-16
128^2, $dt = .05$	24.040	0.0732	0.0904	0.0703	1.1e-16
256^2, $dt = .1$	19.081	0.1064	0.1447	0.1049	1.1e-16

Table 4: External reproduction control on NIST/PFHub Benchmark 1a. Log-RMSE uses 64 frozen times; lower is better. MEU. is the method-matched public MEUMAPPS Fourier-spectral trajectory, FEn. is the public FEniCS trajectory, and 'vs nominal' exposes time/space sensitivity rather than a continuum guarantee.

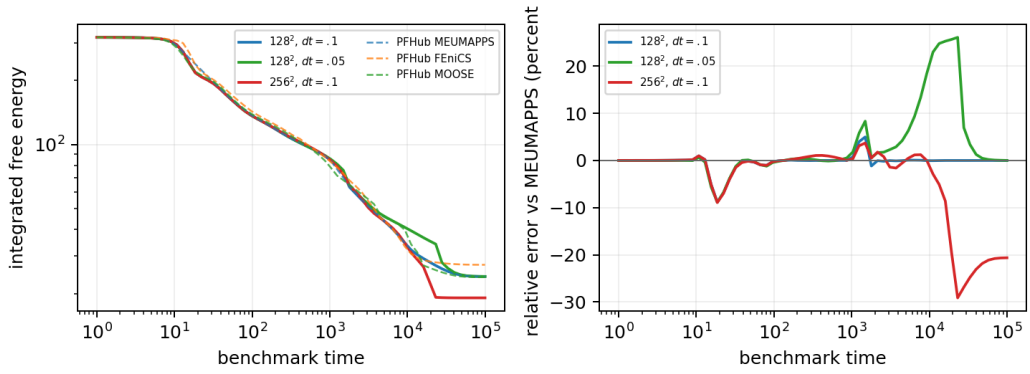

Figure 7: Frozen NIST/PFHub Benchmark 1a control. *Left*: the nominal Fourier run tracks the method-matched public MEUMAPPS energy curve, with FEniCS and MOOSE shown as independent public references. *Right*: complete refinements reveal substantial transient sensitivity; every run is retained.

tion fixes $\eta = 5 \times 10^{-5}$ without a truth metric. Five checkpoint seeds, eight truths, and 24 chains are reused for ID melt, mismatch, and sparse blur; initialization and Markov noises have independent fixed streams. This is a smoothed, ULA-discretized approximate posterior, not an exact reference. A secondary complete curve fixes $t_\star \in \{.02, .05, .10\}$, the first four matched truths, and the same step size. Its protocol was frozen after the formal .05 result was known but before evaluating .02 or .10; all levels are retained and no empirical winner is selected.

D.2 External deterministic solver control

NIST/PFHub Benchmark 1a version 1 is a periodic 200×200 Cahn–Hilliard spinodal decomposition benchmark (Jokisaari et al., 2017). Before running it, we froze the published initial condition and parameters, a double-precision 128^2, $dt = .1$ semi-implicit Fourier run, complete $dt = .05$ and 256^2 refinements, 64 log-spaced evaluation times from 1 to 10^5, and snapshots at 10^3, 10^4, and 10^5. Three public energy files and the PFHub specification are pinned by SHA-256 and repository commit. The nominal log-RMSE against the method-matched MEUMAPPS trajectory is .01877, versus a maximum public-pair discrepancy of .06871, and its final energy 24.0374 lies in the public $[23.8508, 27.2831]$ range. All three runs conserve mass to 1.1×10^{-16} and have no recorded energy increase. However, the time-refined trajectory differs from nominal by .07029 log-RMSE, and the space-refined trajectory differs by .10489 and ends at energy 19.0811 in a different stripe morphology. The control therefore reproduces a public, method-matched discretized trajectory: same-discretization reproduction, not continuum convergence or real-data validation. It is neither a learned-model nor a stochastic-measure benchmark.

Benchmark 2a couples one conserved Cahn–Hilliard concentration to four non-conserved Allen–Cahn variants. Before any full-horizon run, we froze the public version-2 notebook, five trajectories and source provenance, a four-MOOSE pointwise-median primary target, and nominal/time/space-refined configurations. The nominal run fails the public-primary targets: log-RMSE .6781>.1904 and final energy $434.038 \notin [167.019, 199.784]$. Its guarded increase also fails, .0571>9.12×10^{-5}; mass and completeness pass. The retained time/space refinements give $(F, \mathrm{RMSE}) =$

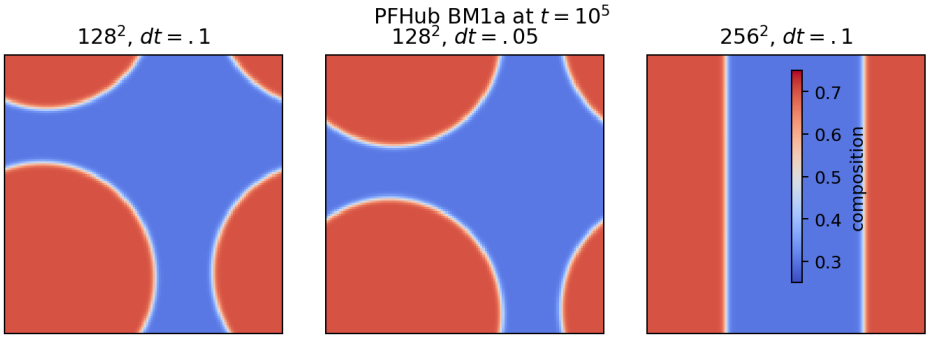

Figure 8: PFHub Benchmark 1a fields at $t = 10^5$. The time-refined run has nearly the same final energy as nominal but a shifted interface, whereas the 256^2 run reaches a lower- energy stripe. Public-trajectory reproduction at one discretization is therefore not reported as continuum convergence.

Run	Grid	Δt	Final F	Med. log-RMSE
nominal	128^2	0.1	434.038	0.6781
time refined	128^2	0.05	375.052	0.5629
space refined	256^2	0.1	434.022	0.6774

Table 5: Frozen NIST/PFHub Benchmark 2a coupled CH/AC control. The primary target is the pointwise median of four public MOOSE trajectories; all refinements are retained.

$(375.052, .5629)/(434.022, .6774)$, while the latter reduces the maximum guarded increase to 1.91×10^{-9}. All runs and snapshots are retained. Thus this mixed/negative control is an external numerical diagnostic, not validation of the coupled solver.

Post-hoc source comparability audit. The frozen rule above is unchanged, but the preserved public decks do not define one like-for-like target (Table 6). Hilty, Guan, and Wu integrate chemical plus concentration-gradient energy; an unlinked Greenquist repository candidate also includes four η-gradient terms. Guan changes the concentration perturbation from .05 to .01, and Wu changes the specified x/y field to a y-only expression. The three-deck partial observable still gives nominal/time/space log-RMSE $.5127/.3134/.5124$. Because initial fields differ and the full-energy candidate is not curve-linked, this explanatory audit cannot rescue, replace, or reverse the frozen decision. Leaving out any one primary curve still gives full-energy log-RMSE .6533–.7035, above each retained-set tolerance (.1168–.1904), so no single public curve drives the discrepancy.

Table 6: Post-hoc source audit of the four primary public MOOSE Benchmark 2a trajectories. *Match* refers only to the specified concentration initial field; *the repository deck is not linked to the uploaded CSV by the public metadata.

Source	ϵ_c	Concentration IC	Integrated gradient fields
Hilty	0.05	match	c only
Guan	0.01	deviates	c only
Wu	0.05	deviates	c only
Greenquist*	0.05	match	$c + 4\eta$

Registered post-hoc time refinement. After the frozen nominal result and the .05 plateau were known, but before any .025 step, we registered a first-order prediction and then ran 128^2, $dt = .025$ through 10^5. Full energy ends at 346.628; the registered prediction misses by .31%, the successive-difference ratio is 2.075 (near the first-order value 2), and the finest-level Richardson limit is 318.205, still above the mixed public final range. Partial energy ends at 234.159, with ratio 1.756, prediction error -2.22%, and limit 190.540. The latter falls inside the mixed public range, but the

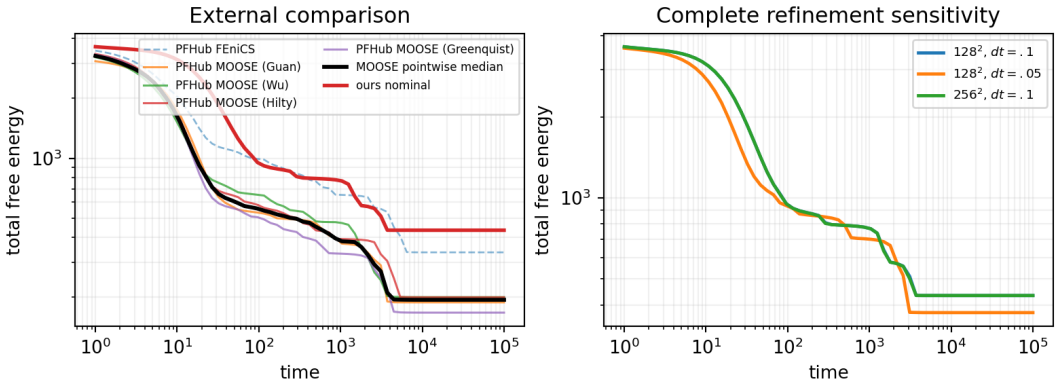

Figure 9: Frozen NIST/PFHub Benchmark 2a control. *Left*: the nominal trajectory lies above the four-MOOSE primary community target; FEniCS is a disclosed secondary curve. *Right*: time refinement changes the late energy substantially, whereas space refinement nearly overlays nominal. The failed frozen decision is retained.

Δt	ϕ rel. L_2	c rel. L_2
0.02000	1.216e-02	4.258e-04
0.01000	6.185e-03	1.998e-04
0.00500	3.131e-03	9.225e-05
0.00250	1.588e-03	5.356e-05
0.00125	8.133e-04	5.097e-05

Table 7: Registered post-hoc error to the independent $N = 1024$ finite-difference BDF reference. The frozen formal decision against $N = 512$ remains failed.

source audit shows it is not a like-for-like target. This post-hoc diagnosis is not promotion eligible and cannot alter the frozen failed decision.

D.3 Independent cross-discretization check of the testbed operator

Before running it, we froze a smooth $T = 1$ trajectory of the manuscript's one-dimensional two-field Model-C equations and compared the imported production Torch/Fourier step to an independent periodic finite-difference method-of-lines solve using variable-step BDF. The production path and a separate float64 Fourier implementation agree to relative L_2 $6.23 \times 10^{-6}/6.40 \times 10^{-6}$ for ϕ/c; both $N = 256, 512$ BDF solves succeed with mass drift below 9×10^{-18}. Against the frozen $N = 512$ target, the finest $dt = .0025$ Fourier errors are $.170\%/.022\%$, but the formal rule remains `failed_independent_crosscheck_rule`: the c error decreases through $dt = .005$ and then rises from 2.035×10^{-4} to 2.226×10^{-4}.

After retaining that failure, a registered diagnostic added an $N = 1024$ BDF reference and $dt = .00125$ (Table 7). The independent reference difference drops by about four from $256 \to 512$ to $512 \to 1024$, consistent with the finite-difference space error that exceeded the failed increment. Relative errors to $N = 1024$ decrease at all five time steps; three adjacent Fourier self-difference ratios are 1.962–1.990 for ϕ and 1.979–1.994 for c. First-order Richardson errors are $7.97 \times 10^{-5}/6.36 \times 10^{-5}$. This diagnosis is not promotion eligible and does not change the formal label. It supports smooth deterministic numerical consistency, not physical, stochastic, inverse, PFHub-2a, or real-data validation.

D.4 Frozen public sea-ice morphology sanity check

Before computing any real–synthetic descriptor, we froze a narrow audit using the public CAATEX 2019 sea-ice-core section photographs (Raffel et al., 2023). The pinned source PDF contains normal- and polarized-light photographs; deterministic layout and color rules extract 37 normal-light transverse sections from five stations, of which 36 pass the frozen segmentation QC. For each section,

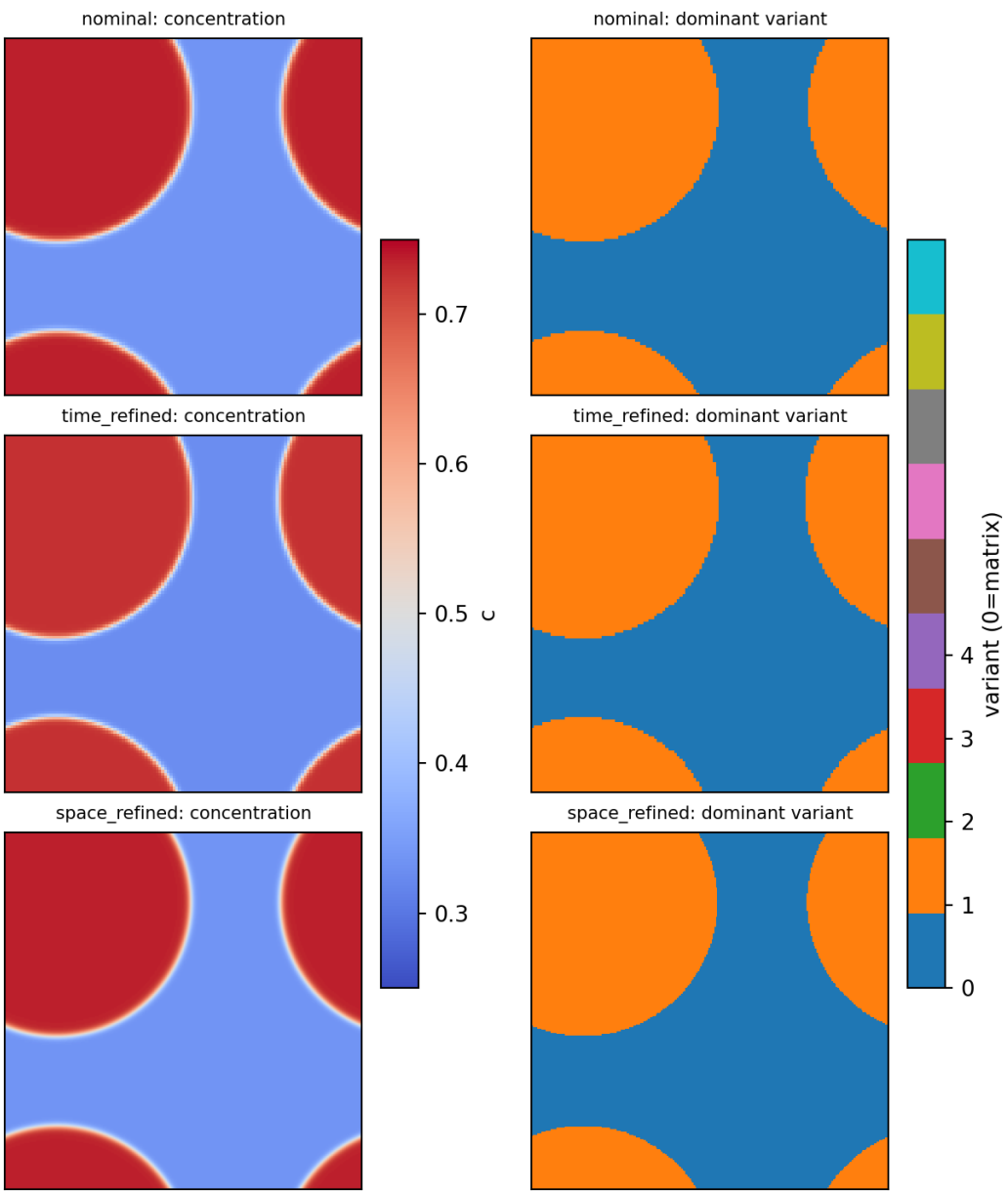

Figure 10: Complete Benchmark 2a final fields for all frozen configurations. Concentration and dominant-variant morphology are similar at this display scale despite the time-step sensitivity of the reported full free energy.

its optically dark-inclusion fraction is matched by thresholding each of the 96 held-out synthetic salt fields at its own corresponding quantile. The comparison therefore conditions on foreground fraction and tests geometry only.

The frozen rule fails (Table 8). No usable section falls inside the synthetic 2.5–97.5% interval for perimeter or component density, and no station median is covered for either descriptor: real optical inclusions are substantially more fragmented. Largest-component share covers 9/36 sections and 0/5 station medians; gradient anisotropy covers 30/36 and 3/5. All nine registered crop/illumination settings retain the perimeter- and component-density failures. The photographs have no pixel-wise salt or brine labels—dark pixels can include air, cracks, preparation artifacts, and illumination residuals—so this is a failed matched-fraction optical-geometry sanity check, not physical, coupled-operator, inverse, salt-label, or real-data validation.

descriptor	section coverage	station coverage	rule
perimeter density	0/36 (0.0%)	0/5	fail
component density	0/36 (0.0%)	0/5	fail
largest-component share	9/36 (25.0%)	0/5	fail
gradient anisotropy	30/36 (83.3%)	3/5	fail

Table 8: Frozen CAATEX matched-fraction morphology sanity check. Real images are unlabeled optically dark inclusions; this is not brine, inverse, or physical validation.

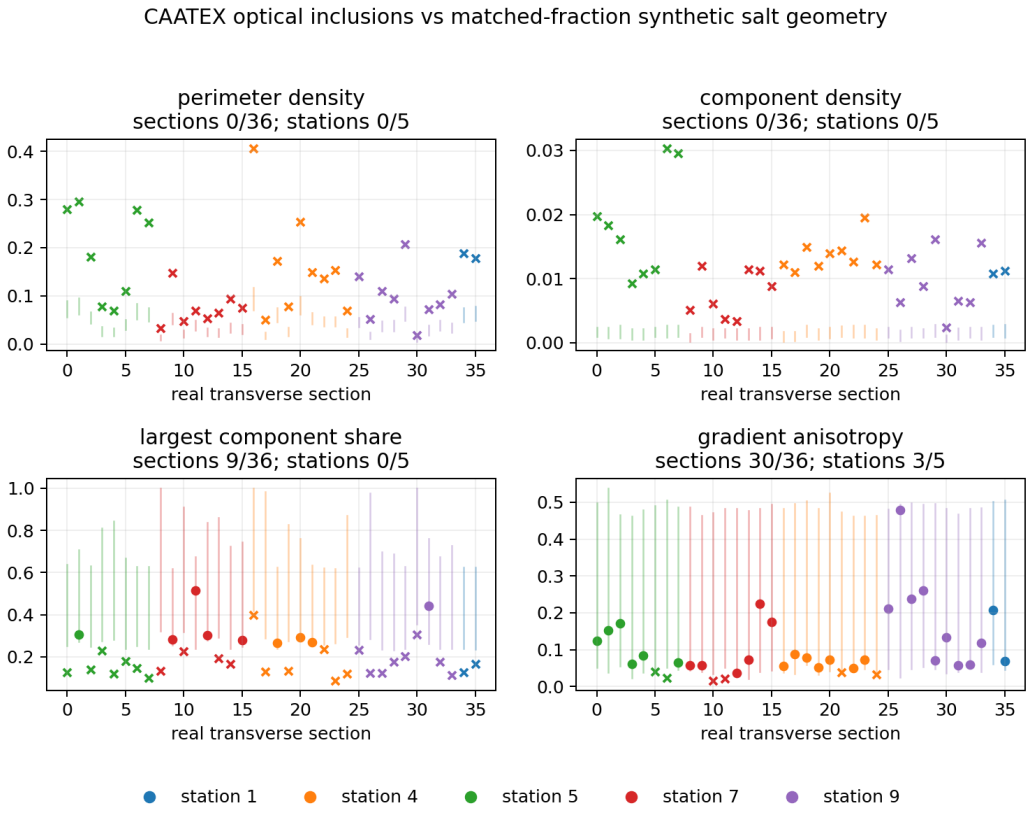

Figure 11: Frozen CAATEX matched-fraction geometry check. Vertical intervals are the 2.5–97.5% ranges across 96 held-out synthetic fields after matching each section's foreground fraction; points are real-section descriptors (circles covered, crosses not covered). Real optical inclusions have consistently higher perimeter and component density.

descriptor	slice coverage	core coverage	rule
scaled perimeter density	0/43 (0.0%)	1/5	fail
scaled component density	24/43 (55.8%)	3/5	fail
largest-component share	7/43 (16.3%)	0/5	fail
gradient anisotropy	36/43 (83.7%)	4/5	pass
segmentation QC	43/45 slices; all samples ≥ 2		pass

Table 9: Post-hoc frozen scale-normalized micro-CT open-pore morphology audit. Intervals are 2.5–97.5% across 96 porosity-matched held-out synthetic fields. This audit cannot supersede the failed frozen CAATEX control.

descriptor	slice coverage	core coverage	diagnostic rule
scaled perimeter density	2/43 (4.7%)	0/5	fail
scaled component density	29/43 (67.4%)	3/5	fail
largest-component share	7/43 (16.3%)	0/5	fail
gradient anisotropy	36/43 (83.7%)	4/5	pass

Table 10: Post-result edge-connected-foreground exclusion diagnostic. It cannot promote or replace the frozen micro-CT result.

D.5 Post-hoc public micro-CT open-pore audit

After the frozen CAATEX audit failed, but before comparing descriptors in this dataset, we registered a second, explicitly post-hoc audit using public micro-CT volumes of young sea ice from Adventfjorden (Maus et al., 2021). The samples were centrifuged before imaging: connected brine open to the sample boundary was expelled, leaving an air-filled pore network. The source study reports accurate air–ice segmentation but substantial uncertainty for residual brine; accordingly, we analyze only dark open-air pores and make no residual-brine or salt-concentration claim.

Outcome-blind metadata rules select three normalized depths in each of five cores, three axial slices per segment, for 45 registered slices. We apply the published two-pixel median and $\sigma = 1.5$ Gaussian filters, one pooled Otsu threshold per segment, and area downsampling; 43/45 slices pass frozen QC. Since the CT voxel length has no defensible mapping to the synthetic lattice, each real slice is compared with 96 held-out synthetic masks at matched foreground fraction, and perimeter and component densities are nondimensionalized by the mask's periodic two-point $1/e$ correlation length.

The frozen post-hoc rule fails (Table 9). Scale-normalized perimeter covers 0/43 slices and 1/5 core medians; 38 slices fall below and five above their synthetic intervals. Scaled component density covers 24/43 and 3/5. Largest-component share covers only 7/43 and 0/5, with 35 slices below the synthetic interval, indicating that the real open-pore masks are typically split across smaller components. Gradient anisotropy alone passes, at 36/43 and 4/5. None of the nine registered crop/filter settings attains the pass label.

Visual inspection then identified exterior air at an irregular sample boundary (Fig. 13). We registered a non-promotion-eligible diagnostic that removes every native foreground component touching a crop edge. It retains all three formal failures: coverage becomes 2/43 and 0/5 for scaled perimeter, 29/43 and 3/5 for scaled component density, and remains 7/43 and 0/5 for largest-component share (Table 10). Thus the observed mismatch is not explained by that boundary contamination alone. This is a narrow geometry-domain-gap result from one field campaign, not physical calibration, coupled-operator validation, or real-data inversion; it does not replace the confirmatory CAATEX failure.

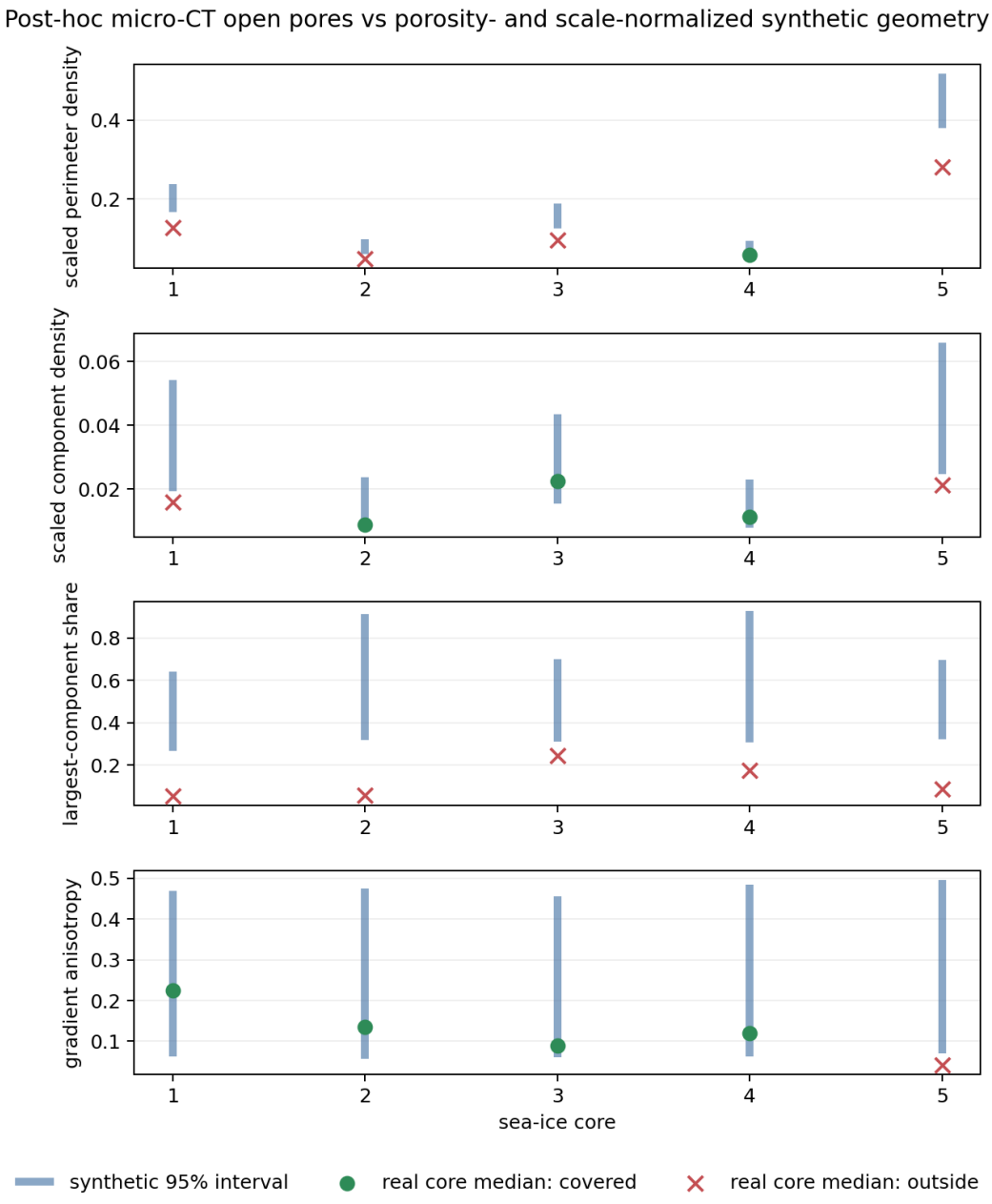

Figure 12: Post-hoc micro-CT core medians (points) against 2.5–97.5% intervals from 96 porosity-matched held-out synthetic masks after correlation-length normalization. Green circles are covered; red crosses are outside. Only gradient anisotropy meets both frozen slice- and core-coverage rules.

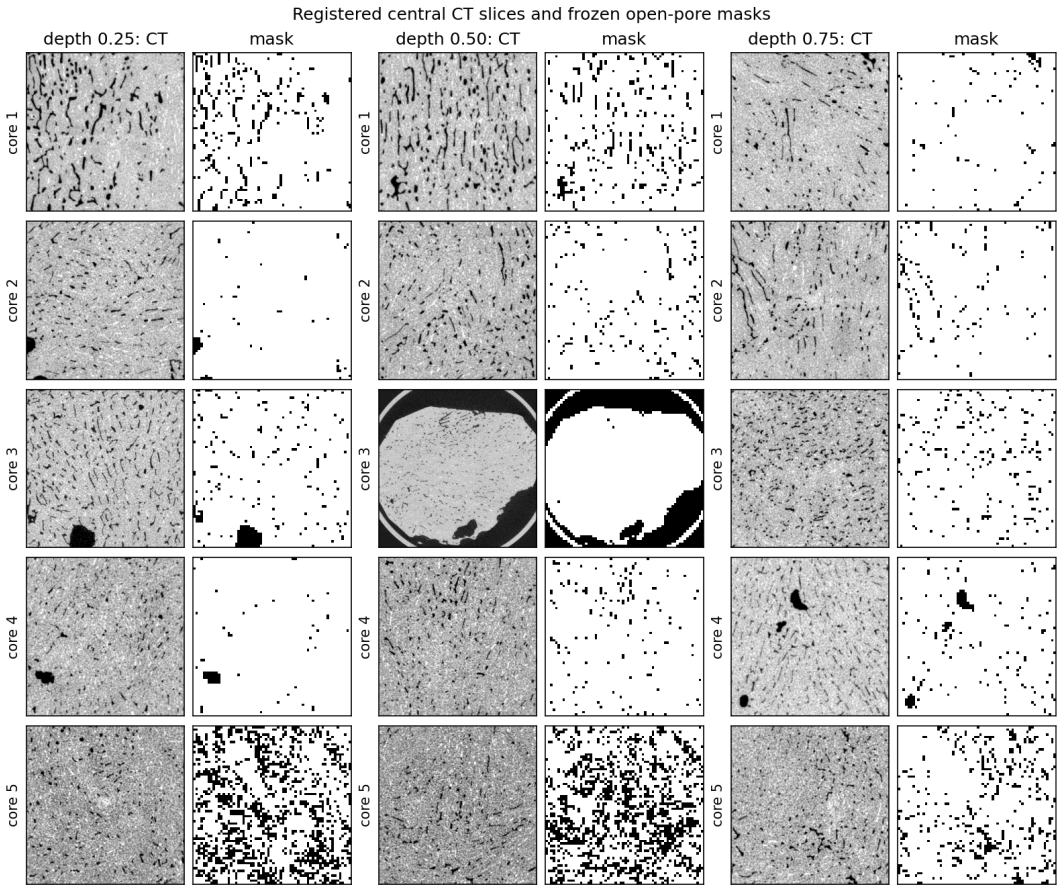

Figure 13: Registered middle axial CT slices and frozen Otsu open-pore masks at the three normalized depths in each core. The conspicuous exterior-air region in core 3, depth 0.50, motivated only the registered post-result edge-component diagnostic; it did not alter the formal masks or label.

D.6 FROZEN ZERO-SHOT INVERSE ON EXPERIMENTAL OPV TOMOGRAPHY

We next test a directly observed inverse task without retraining. The public MicroGen3D release contains binary donor–acceptor morphologies reconstructed by energy-filtered electron tomography from spin-cast P3HT:PCBM films made with chlorobenzene (CB) and dichlorobenzene (DCB) (Baishnab et al., 2025; Heiber et al., 2020; Herzing et al., 2010). Although the release exposes 21,421 subvolumes, they are overlapping crops from only these two parent tomograms. Before reconstruction, we therefore froze a coordinate-only greedy selection of 45 mutually non-overlapping source-plane blocks (22 CB, 23 DCB), with the central axial slice primary and $z \in \{16, 48\}$ as sensitivities. CB and DCB are the two experimental units; blocks and slices are repeated measurements, so we report no pseudo-replicated experimental confidence interval.

Each binary slice is mean-centered and scaled to the original synthetic training standard deviation, isolating geometry while satisfying the conserved prior's zero-mode support. We then reuse the already frozen sparse-blur observation: Gaussian width 2, 25% Bernoulli sampling, and observed-pixel noise .02. The five paired physics/cosine checkpoints retain 24 chains, 120 reverse steps, and all sampler settings. Tikhonov and TV weights are chosen only on the first 16 original synthetic held-out fields from the registered eight-value grid; the selected values are .1 and .03, respectively. The primary transfer rule requires one learned arm to beat both baselines in reconstruction-mean RMSE for every checkpoint seed on both parents, improve five-seed orientation-invariant Dice on both, and retain the RMSE direction after pooling all three axial slices.

The rule fails (Table 11). On CB, TV/Tikhonov central RMSE is $.279/.283$, versus .303 for cosine DPS and .352 for physics DPS; on DCB it is $.234/.244$, versus .270 and .328. All-slice sensitivity retains the ordering. Dice does too: the stronger classical value is $.829/.890$ on CB/DCB, versus $.795/.852$ for cosine and $.707/.765$ for physics. Thus cosine transfers better than the physics-defined forward, but neither learned synthetic prior beats a tuned classical estimator. This is a zero-shot real-observation domain-gap result under one generic operator, not validation of OPV constitutive physics or of the salt–ice melt operator.

Parent	Method	Central RMSE ↓	Central Dice ↑	All-slice RMSE ↓
CB	Tikhonov	0.283	0.829	0.283
CB	TV	0.279	0.825	0.279
CB	Physics DPS	0.352 [0.351,0.355]	0.707	0.353 [0.351,0.355]
CB	Cosine DPS	0.303 [0.300,0.305]	0.795	0.304 [0.301,0.306]
DCB	Tikhonov	0.244	0.890	0.243
DCB	TV	0.234	0.889	0.232
DCB	Physics DPS	0.328 [0.327,0.329]	0.765	0.328 [0.327,0.329]
DCB	Cosine DPS	0.270 [0.268,0.271]	0.852	0.268 [0.267,0.270]

Table 11: Frozen zero-shot inverse audit on two public experimental P3HT:PCBM tomograms. Forty-five non-overlapping source-plane blocks provide repeated measurements; CB and DCB are the only experimental units. Learned entries are five-checkpoint means with seed ranges in brackets for RMSE. Central Dice is orientation-invariant. The all-slice column pools the registered axial slices $z \in \{16, 32, 48\}$. Both tuned classical baselines beat both learned priors on both parents, so the frozen transfer rule fails.

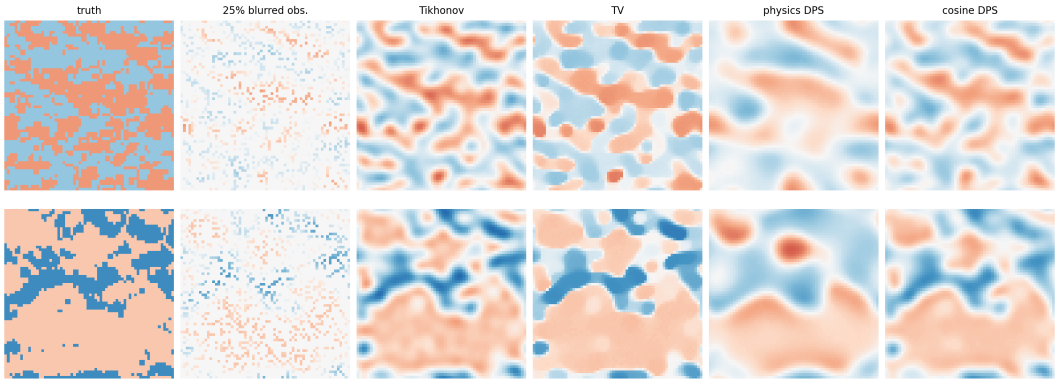

Figure 14: Representative frozen central-slice reconstructions from the two parent OPV tomograms (CB top, DCB bottom). The display uses the first coordinate-selected block and checkpoint seed zero; the formal result aggregates all 45 non-overlapping blocks, five checkpoint seeds, and three registered axial slices. Both learned posterior means are smoother and less faithful than the tuned classical reconstructions.

D.7 MATCHED EXACT-GAUSSIAN POSTERIOR-UTILITY BOUNDARY

The practical failures above do not mean that target-measure fidelity is useless. Consider $x \sim \mathcal{N}(0, C)$, $y = Hx + \epsilon$, $\epsilon \sim \mathcal{N}(0, R)$, and a linear estimator Ay. With $S = HCH^\top + R$, its population squared-error risk is

$$\mathcal{R}(A) = \operatorname{tr}(C) - 2\operatorname{tr}(AHC) + \operatorname{tr}(ASA^\top).$$

The true-prior posterior mean has $A_* = CH^\top S^{-1}$, and completing the square gives

$$\mathcal{R}(A) - \mathcal{R}(A_*) = \operatorname{tr}\left[(A - A_*)S(A - A_*)^\top\right] \geq 0.$$

Thus a correct covariance is guaranteed to minimize posterior-mean risk when the model is matched and inference is exact; this standard Bayes projection property supplies a positive boundary for our empirical non-implication claim.

We froze a deterministic CHC modal control before evaluation. Truths use the 64 nonzero $N = 128$ target variances, while the misspecified prior uses the analytic stationary covariance of the same-drift EM-noise scheme at $dt = .02$. Four fixed identity/blur and noise regimes use exact diagonal posterior means. The excess-risk identity is verified to 10^{-12}, and ten paired CPU float64 Monte Carlo seeds (65,536 vectors each) agree with analytic risks within 1%. Target covariance lowers mean risk in all four regimes: the relative EM excess is $.03\%$, 6.03%, 4.36%, and $.09\%$ (Table 12); every paired interval and leave-one-seed-out direction is positive. Correct-prior coverage is $.950$ and expected posterior KL is zero; the EM prior over-covers up to $.972$. This is a matched linear-Gaussian guarantee, not a claim about learned scores, approximate DPS, nonlinear posteriors, or real observations.

regime	correct risk	EM risk	excess	EM coverage
identity, $\sigma = .02$	0.00039963	0.000399735	0.03%	0.950
identity, $\sigma = .50$	0.168742	0.178909	6.03%	0.958
blur 1, $\sigma = .10$	0.118557	0.123721	4.36%	0.967
blur 2, $\sigma = .02$	0.200228	0.200409	0.09%	0.972

Table 12: Frozen matched exact-Gaussian inverse control. Risk is analytic posterior-mean MSE averaged over all 64 nonzero CHC modes; lower is better. The measure-correct prior has nominal 0.950 coverage and zero expected posterior KL in every regime. The EM column uses the stationary covariance induced by the same-drift continuum-amplitude noise. All paired Monte Carlo intervals for EM-minus-correct MSE exclude zero.

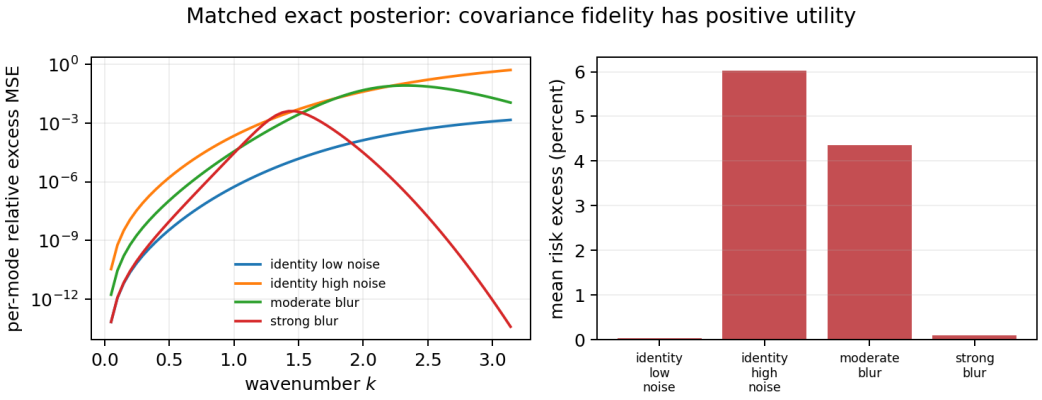

Figure 15: Frozen matched exact-Gaussian utility control. Left: per-mode relative excess posterior-mean MSE from the EM-biased stationary covariance. Right: analytic mean excess over all 64 nonzero modes. Correct covariance is Bayes-optimal in every regime, but its practical gain is small when observations are nearly noise-free or erase the affected high modes.

Δt	EM bias	FDT raw	EM raw	EM−FDT raw [95% CI]
0.005	0.520	2.754±0.027	3.057±0.028	+0.303 [+0.271,+0.335]
0.010	1.090	2.309±0.041	2.957±0.055	+0.648 [+0.615,+0.680]
0.020	2.296	1.865±0.076	2.830±0.091	+0.966 [+0.915,+1.016]

Table 13: Registered terminal-measure-bias dose response. Values are five-seed means ± seed standard errors; intervals are paired seed bootstraps. The $\Delta t = 0.020$ row is the disclosed outcome-known anchor.

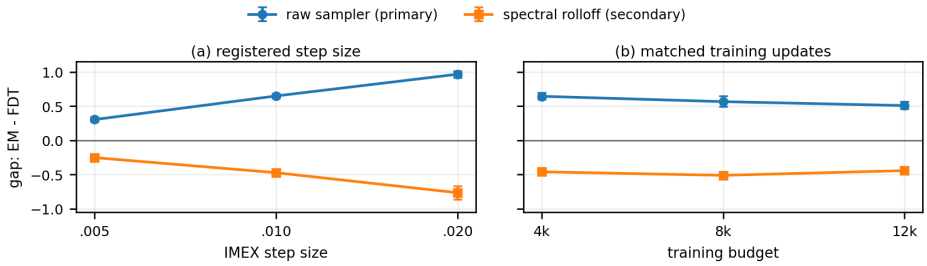

Figure 16: Registered linear audits: the raw EM–FDT gap increases with step size (left) and attenuates but remains resolved through 12,000 matched updates at $\Delta t = .010$ (right). Rolloff reverses every comparison; bars are paired five-seed intervals and outcome-known anchors are disclosed.

Registered terminal-measure-bias dose response. The three-step result receives the frozen label `measure_bias_learning_dose_response`. Table 13 reports the primary raw spectral metric; Fig. 16 retains the opposite secondary rolloff direction. This is a fixed synthetic CHC propagation result, not a universal causal law or evidence of real-data or inverse benefit. Each arm uses its own analytic stationary covariance, excluding a shared-terminal mismatch.

The adjacent raw-gap increments are $+.345\,[+.332, +.355]$ and $+.318\,[+.240, +.382]$; both registered intervals and every leave-one-seed-out mean are positive.

Registered training-budget saturation. At $\Delta t = .010$, the five-seed paired raw EM–FDT gap receives the frozen label `measure_bias_gap_compute_persistent`: both new-budget intervals and every leave-one-seed-out mean are positive. The 4,000-update row was an outcome-known deterministic replication check; 8,000 and 12,000 updates were new. The pre-outcome amendment changed only the failed byte-level anchor-replication procedure, before any new-budget training or outcome existed.

The per-seed raw gaps at 4,000 updates are $\{.608, .650, .743, .629, .585\}$; at 8,000, $\{.523, .711, .454, .606, .531\}$; and at 12,000, $\{.447, .546, .476, .610, .459\}$. The 12,000-minus-4,000 paired change is $-.136\,[-.207, -.069]$, licensing only finite-budget attenuation. Secondary metrics disagree: rolloff gaps are $-.461\,[-.509, -.417]$, $-.513\,[-.565, -.454]$, and $-.444\,[-.466, -.418]$, while raw-energy gaps are unresolved at 4,000 and 8,000 ($-.000078\,[-.000272, +.000142]$ and $-.000114\,[-.000303, +.000069]$) and tiny but positive at 12,000 ($+.000131\,[+.000071, +.000190]$). Validation and training-tail losses remain closely matched; all generated-field summaries and replayable cell records are retained in the supplement. This audit rules out disappearance under the tested continuation of one ScoreNet/Adam trajectory family, not an asymptotic or architecture-independent gap.

Post-hoc evaluation-stream sensitivity. We repeated every training-budget evaluation on four new frozen streams, retaining the original outcome-known stream as seed zero. Across the resulting five streams, the raw gaps remain $+.643\,[+.604, +.696]$, $+.566\,[+.498, +.647]$, and $+.504\,[+.454, +.565]$; all train-seed bootstrap lower bounds, leave-one-train-seed-out means, stream means, and leave-one-stream-out means are positive.

The stream-mean ranges are only $[+.642, +.645]$, $[+.563, +.568]$, and $[+.502, +.507]$. Rolloff still reverses the gaps ($-.461\,[-.509, -.419]$, $-.512\,[-.561, -.457]$, and $-.440\,[-.463, -.414]$),

updates	FDT raw	EM raw	EM−FDT raw [95% CI]
4,000	2.321±0.048	2.964±0.061	+0.643 [+0.602,+0.694]
8,000	1.961±0.033	2.525±0.069	+0.565 [+0.497,+0.648]
12,000	1.786±0.045	2.293±0.043	+0.507 [+0.458,+0.567]

Table 14: Registered training-budget saturation audit at $\Delta t = .010$. Values are five-seed means $\pm$ seed standard errors; intervals are paired seed bootstraps. The 4,000-update row is a disclosed outcome-known deterministic replication gate.

updates	mean gap	train-seed 95% CI	eval-stream range
4,000	+0.643	[+0.604,+0.696]	[+0.642,+0.645]
8,000	+0.566	[+0.498,+0.647]	[+0.563,+0.568]
12,000	+0.504	[+0.454,+0.565]	[+0.502,+0.507]

Table 15: Disclosed post-hoc evaluation-stream sensitivity. Gaps are raw spectral log-RMSE (EM minus FDT). The training-seed interval treats the five evaluation streams as within-seed Monte Carlo replicates; evaluation seed zero was outcome-known and seeds one through four were new.

whereas raw-energy and rolloff-energy intervals cross zero at every budget. This disclosed sensitivity does not upgrade the registered decision or establish architecture, sampler, real-data, or inverse-task robustness.

Post-hoc network-capacity sensitivity. A disclosed 32-channel/four-block ScoreNet (231,649 parameters) retains positive raw gaps at 4,000/8,000/12,000 updates: $+.583\,[+.525, +.641]$, $+.506\,[+.431, +.562]$, and $+.451\,[+.400, +.494]$. All frozen evaluation-stream and leave-one-out checks pass at 8,000 and 12,000 updates, giving `capacity_robust_persistent`; rolloff remains reversed ($-.179, -.308, -.315$), and all raw- and rolloff-energy intervals cross zero.

All five evaluation streams were frozen before the sensitivity training outcomes. This post-hoc result is bounded to the reference 64-channel/six-block (1,364,929-parameter) and sensitivity 32-channel/four-block (231,649-parameter) ScoreNets in the fixed CHC design. The smaller-minus-reference gap interval crosses zero at every budget; neither superiority/equivalence nor arbitrary-architecture, sampler, real-data, or inverse-task robustness follows.

Post-hoc architecture-family sensitivity. Selected after the original ScoreNet result and after the within-family capacity audit had begun, this audit replaces the local ScoreNet with a periodic FNO (width 18, four blocks, 16 modes per direction, time embedding 64; 1,346,207 real parameters versus 1,364,929). Every training seed, budget, evaluation stream, metric, and decision rule was frozen before formal FNO training. Its 4,000/8,000/12,000-update raw gaps are $+.723\,[-.260, +1.269]$, $+5.234\,[+1.162, +13.295]$, and $+1.204\,[+1.170, +1.241]$; evaluation-stream ranges are $[-.486, +1.184]$, $[+3.717, +6.890]$, and $[+1.203, +1.206]$. Thus 4,000 updates are unresolved, whereas both frozen new-budget checks pass, giving `architecture_family_robust_persistent`. At 8,000 updates one training-seed mean is 21.336 versus 1.138–1.331 for the other four, so the claim concerns direction, not a stable effect size.

Unlike the ScoreNet checks, FNO rolloff gaps keep the raw direction: $+.889\,[+.029, +1.328]$, $+4.892\,[+1.233, +12.157]$, and $+1.242\,[+1.222, +1.262]$. Raw-energy gaps are also positive, while rolloff-energy intervals cross zero at every budget. FNO-minus-reference gap intervals cross zero at 4,000 updates but are positive at 8,000/12,000 ($+4.669\,[+.634, +12.652]$ and $+.700\,[+.626, +.770]$). This disclosed post-hoc result covers only one local ScoreNet and one near-parameter-matched periodic FNO in the fixed CHC design; it establishes neither stable effect magnitude nor robustness to untested backbones, samplers, real data, or inverse tasks.

updates	mean gap	train-seed 95% CI	eval-stream range
4,000	+0.583	[+0.525,+0.641]	[+0.582,+0.585]
8,000	+0.506	[+0.431,+0.562]	[+0.504,+0.508]
12,000	+0.451	[+0.400,+0.494]	[+0.450,+0.453]

Table 16: Disclosed post-hoc second-capacity sensitivity using a 32-channel, four-block ScoreNet (231,649 parameters). Gaps are raw spectral log-RMSE (EM minus FDT); the training-seed interval averages five frozen evaluation streams within seed.

updates	mean gap	train-seed 95% CI	eval-stream range
4,000	+0.723	[-0.260,+1.269]	[-0.486,+1.184]
8,000	+5.234	[+1.162,+13.295]	[+3.717,+6.890]
12,000	+1.204	[+1.170,+1.241]	[+1.203,+1.206]

Table 17: Disclosed post-hoc architecture-family sensitivity using a periodic FNO (width 18, four blocks, 16 modes per direction; 1,346,207 parameters). Gaps are raw spectral log-RMSE (EM minus FDT); the training-seed interval averages five frozen evaluation streams within seed.

metric	exact OU	IMEX+FDT	IMEX+EM
DSM val. loss↓	0.00567 ± 0.00002	0.00571 ± 0.00005	0.00570 ± 0.00003
spectral log-RMSE↓	3.245 ± 0.027	1.865 ± 0.076	2.830 ± 0.091
+ rolloff↓	1.068 ± 0.041	2.284 ± 0.047	1.517 ± 0.021
inverse clTru↓	0.037 ± 0.002	0.037 ± 0.003	0.037 ± 0.001

Table 18: Paired 5-seed run (4000 training iterations per arm). Exact OU and IMEX+FDT preserve the same terminal measure; IMEX+FDT and IMEX+EM share the same drift. All checkpoints are re-evaluated with identical held-out fields, terminal noise, observations, and sampler randomness. Values are mean $\pm$ seed standard error.

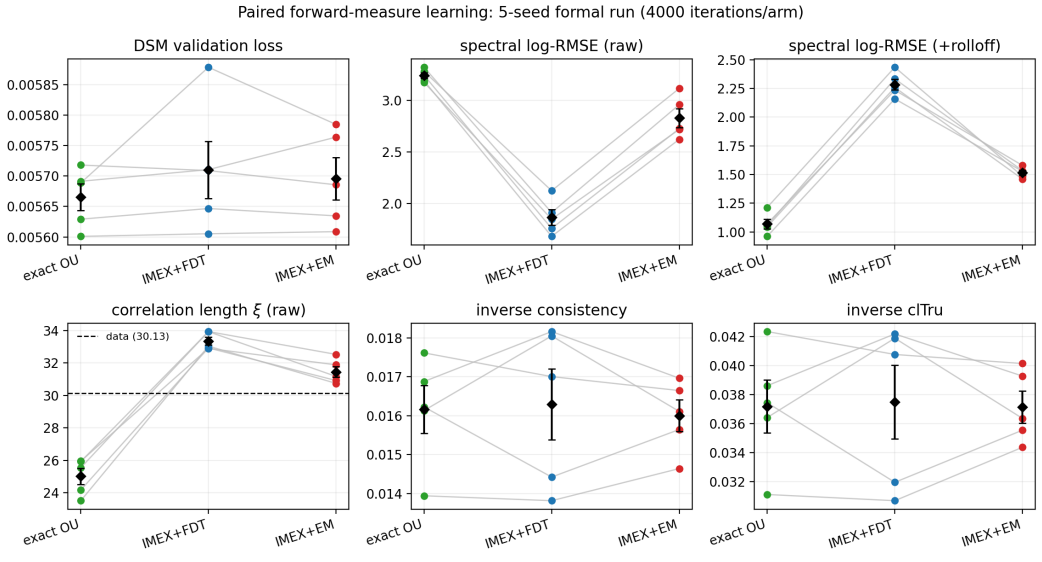

Figure 17: Paired five-seed, 4000-iteration bridge from invariant-measure bias to learning. Gray lines pair the same training seed; diamonds show means with seed standard errors. The discrete-FDT arm has lower raw generated-spectrum error than exact OU and same-drift EM, but the common high-k rolloff reverses the ranking. DSM and inverse contrasts do not resolve a stable ordering.

metric	exact OU	BE+FDT	BE+EM
ϵ val. loss↓	0.102 ± 0.000	0.082 ± 0.000	0.114 ± 0.000
relative score MSE↓	0.00136 ± 0.00021	0.00114 ± 0.00006	0.00168 ± 0.00029
denoising MSE↓	0.414 ± 0.001	0.420 ± 0.000	0.385 ± 0.001
well-sign accuracy↑	0.680 ± 0.001	0.674 ± 0.002	0.708 ± 0.000

Table 19: Paired five-seed underdamped-Langevin control (3000 iterations per arm). BE+FDT and BE+EM share the same stable matrix drift. Discrete FDT preserves the target Gibbs covariance and improves learned-score accuracy, whereas the colder biased BE+EM channel makes the fixed-time denoising task easier. Values are mean $\pm$ seed standard error.

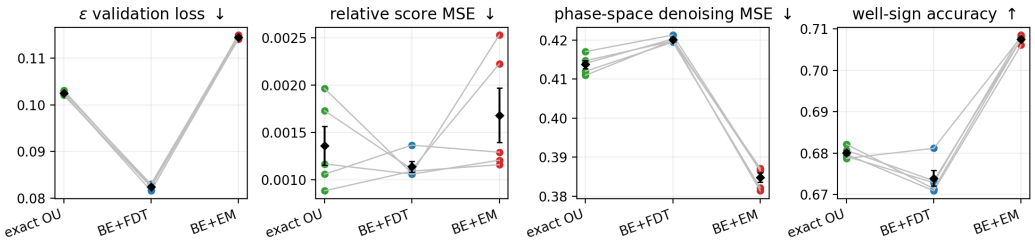

Figure 18: Non-diagonal underdamped-Langevin control. Gray lines pair five training seeds; diamonds are means with seed standard errors. Discrete FDT gives the lowest mean score error, whereas the colder biased BE+EM channel is easier to denoise and classify by double-well sign. Measure fidelity, learned-score fidelity, and task difficulty are distinct even beyond conserved Fourier-diagonal CHC.

regime	physics OU	cosine VP	cosine−physics [95% CI]
melt, ID	0.903 ± 0.008	0.879 ± 0.001	-0.024 [-0.036,-0.008]
melt, high noise	1.448 ± 0.026	1.713 ± 0.004	+0.265 [+0.214,+0.316]
melt, mismatch	5.436 ± 0.086	11.538 ± 0.006	+6.102 [+5.947,+6.250]
melt, OOD morphology	2.448 ± 0.115	0.880 ± 0.002	-1.568 [-1.783,-1.399]
sparse blur	1.068 ± 0.019	1.188 ± 0.001	+0.121 [+0.096,+0.160]

Table 20: Predeclared multi-regime stress test. The primary outcome is joint posterior energy score (lower is better); values are mean ± training-seed standard error, and contrasts use paired training-seed bootstrap intervals. Two-level seed–truth intervals and all secondary metrics are reported in the anonymous supplement.

metric	physics OU	cosine VP
consistency↓	0.01560 ± 0.00049	0.00877 ± 0.00000
clTru↓	0.036 ± 0.001	0.01562 ± 0.00002
energy↓	0.903 ± 0.008	0.879 ± 0.001
variogram↓	0.00134 ± 0.00003	8.75e-04 ± 2.0e-06
RR-ECD↓	0.314 ± 0.080	0.229 ± 0.010
marginal ECE↓	0.119 ± 0.020	0.645 ± 0.000
mass drift↓	4.99e-09 ± 3.9e-11	5.08e-04 ± 1.8e-06

Table 21: Physics-defined exact-OU versus cosine-VP priors under paired 5-seed, 4000-iteration training and fixed inverse evaluation (16 truths, 24 chains, 120 DPS steps). Values are mean ± seed standard error.

E Downstream stress and guidance details

The pointwise primary energy-score ordering is heterogeneous across the five regimes: paired-seed, two-level seed–truth, and leave-one-seed-out analyses give the same direction in each regime. Secondary metrics expose different axes. Cosine has lower observation inconsistency in all five regimes, whereas physics has lower marginal ECE and mass drift in all five. The per-chain truth RMSE favors cosine except under melt mismatch; variogram score favors cosine for ID, OOD morphology, and sparse blur, physics for mismatch, and is unresolved under high noise. RR-ECD resolves only sparse blur in favor of cosine; the remaining regimes are unresolved under the predeclared rule.

Post-hoc familywise sensitivity. The primary rules above treat each predeclared regime or coefficient as a separate estimand. As a multiplicity sensitivity, Bonferroni-adjusted percentile-bootstrap intervals at familywise 95% retain four of five stress directions; the ID two-level interval becomes unresolved. All four guidance directions remain resolved. This post-hoc analysis does not replace the pointwise primary intervals or license a universal schedule ranking.

Conditional-score transfer. On the ID melt task, conditional joint energy is $.628$ versus $.903/.879$ for physics/cosine DPS, hence 28.6% below the stronger reusable-prior arm. The ID-trained conditional model is evaluated without retraining, operator metadata, masks, or test-time optimization. It improves on both reusable-prior DPS arms for OOD morphology, but not for high noise; under model mismatch it lies between physics and cosine DPS. Sparse blur changes the measurement channel itself, so the extreme failure there diagnoses missing operator conditioning rather than a symmetric ranking of the three methods.

Operator-aware multi-operator control. Before training, we froze one shared conditional model over three melt and three sparse-blur configurations. Its nine-channel branch receives the observation, mask, known operator family, and normalized parameters; the tested $(\sigma{=}2, .25)$ blur pair is absent from training. On melt, its energy is 0.708 versus $.903/.879$ for physics/cosine DPS—19.5% below cosine DPS—but $.080$ above the single-operator conditional model; every paired and two-level interval resolves these directions. On held-out sparse blur, operator awareness reduces the single-operator failure from 142.853 to 1.390, yet physics/cosine DPS remain lower at $1.068/1.188$. Operator-aware minus physics is $+.322$ (paired $[+.149, +.524]$; two-level $[+.144, +.522]$), with every leave-one-seed-out mean positive. This specialization–transfer tradeoff is interpolation within a trained operator family, not unseen-family transfer.

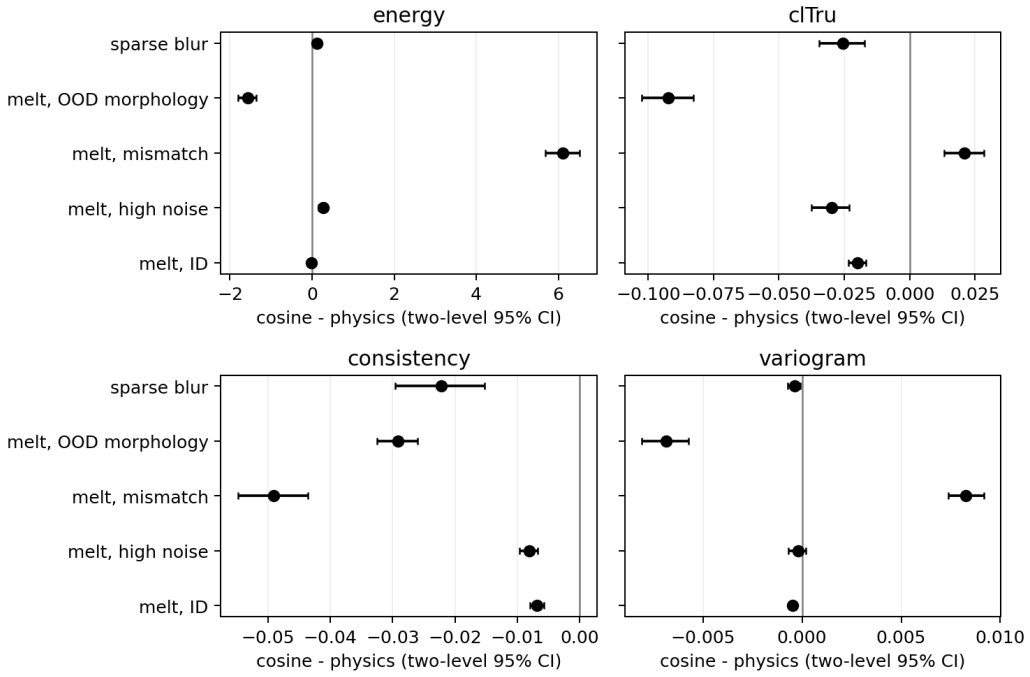

Figure 19: Primary stress-test contrasts (cosine minus physics). Points and bars are paired training-seed means and 95% intervals; outlined bars show the two-level seed–truth intervals. Negative values favor cosine and positive values favor physics. Both uncertainty levels agree in all five predeclared regimes.

ζ_0	physics cons.	physics energy	cosine cons.	cosine energy
0.5	0.17811	14.912	0.12714	8.963
1	0.14417	11.061	0.02800	1.137
2	0.04876	2.373	0.00883	0.820
4	0.01103	0.806	0.00977	0.900

Table 22: Guidance sensitivity on the in-distribution melt inverse. Every predeclared coefficient is shown; lower consistency and energy are better. The curve uses five paired training seeds, four truths, 16 chains, and 80 DPS steps.

family	setting	adjusted two-level interval	decision
stress	melt, ID	$[-0.0532, +0.0059]$	unresolved
stress	melt, high noise	$[+0.1220, +0.3778]$	physics better
stress	melt, mismatch	$[+5.5372, +6.6516]$	physics better
stress	melt, OOD morphology	$[-1.8864, -1.2966]$	cosine better
stress	sparse blur	$[+0.0413, +0.1922]$	physics better
guidance	$\zeta_0 = 0.5$	$[-6.8295, -4.9781]$	cosine better
guidance	$\zeta_0 = 1$	$[-10.6968, -9.0992]$	cosine better
guidance	$\zeta_0 = 2$	$[-2.2134, -0.9425]$	cosine better
guidance	$\zeta_0 = 4$	$[+0.0655, +0.1232]$	physics better

Table 23: Post-hoc Bonferroni familywise sensitivity for the primary energy contrasts. Intervals use paired training-seed and paired seed–truth resampling; the table shows the adjusted two-level intervals. The five stress regimes and four guidance coefficients are corrected as separate families. This analysis does not replace the predeclared pointwise intervals.

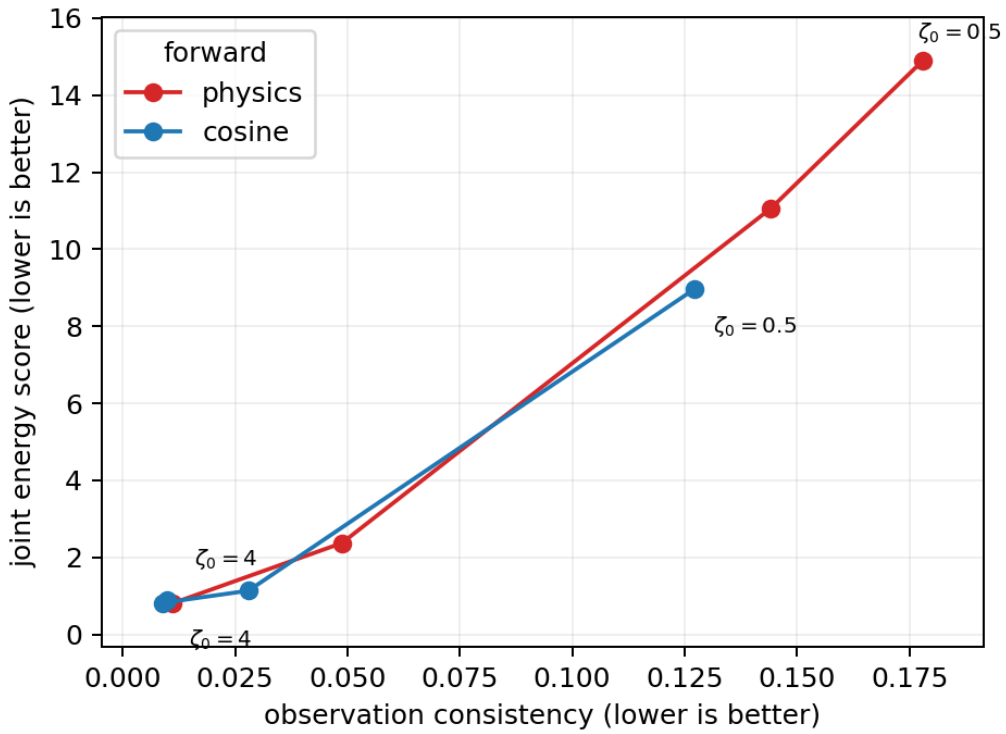

Figure 20: Complete predeclared guidance curve on the ID melt inverse. Points follow $\zeta_0 = .5, 1, 2, 4$ for each forward; lower consistency and energy are better. The mean nondominated set contains cosine at $\zeta_0 = 2$ and physics at $\zeta_0 = 4$; all points and inferential intervals are reported rather than selecting a coefficient post hoc.

method	cons.	energy	clTru	ECE	mass
physics DPS	0.0156 ± 0.0005	0.903 ± 0.008	0.036 ± 0.001	0.119 ± 0.020	4.99e-09 ± 3.9e-11
cosine DPS	0.0088 ± 0.0000	0.879 ± 0.001	0.0156 ± 0.0000	0.645 ± 0.000	5.08e-04 ± 1.8e-06
conditional	0.0099 ± 0.0000	0.628 ± 0.002	0.0146 ± 0.0001	0.320 ± 0.003	8.59e-04 ± 2.3e-04

Table 24: Task-specific conditional baseline on the ID melt inverse. Values are mean ± training-seed standard error over five seeds; all metrics are lower better. Physics and cosine use unconditional priors with DPS, whereas conditional sampling is amortized and uses no likelihood gradient or test-time optimization.

regime	physics DPS	cosine DPS	conditional
melt, high noise	1.448 ± 0.026	1.713 ± 0.004	2.290 ± 0.007
melt, mismatch	5.436 ± 0.086	11.538 ± 0.006	11.255 ± 0.043
melt, OOD morphology	2.448 ± 0.115	0.880 ± 0.002	0.684 ± 0.002
sparse blur	1.068 ± 0.019	1.188 ± 0.001	142.853 ± 1.612

Table 25: Frozen transfer of the ID-trained conditional model. Joint posterior energy score is lower better. The conditional model receives no adaptation, operator metadata, or mask; sparse blur is therefore a measurement-channel shift, not a symmetric method-ranking experiment.

regime	physics DPS	cosine DPS	single-op cond.	operator-aware cond.
melt, ID	0.903 ± 0.008	0.879 ± 0.001	0.628 ± 0.002	0.708 ± 0.066
sparse blur, held-out config.	1.068 ± 0.019	1.188 ± 0.001	142.853 ± 1.612	1.390 ± 0.088

Table 26: One operator-aware conditional model across two observation families. Joint posterior energy score is lower better; values are mean ± training-seed standard error over five seeds. The sparse-blur parameter pair is absent from the training bank, while the operator family and exact test metadata are available.

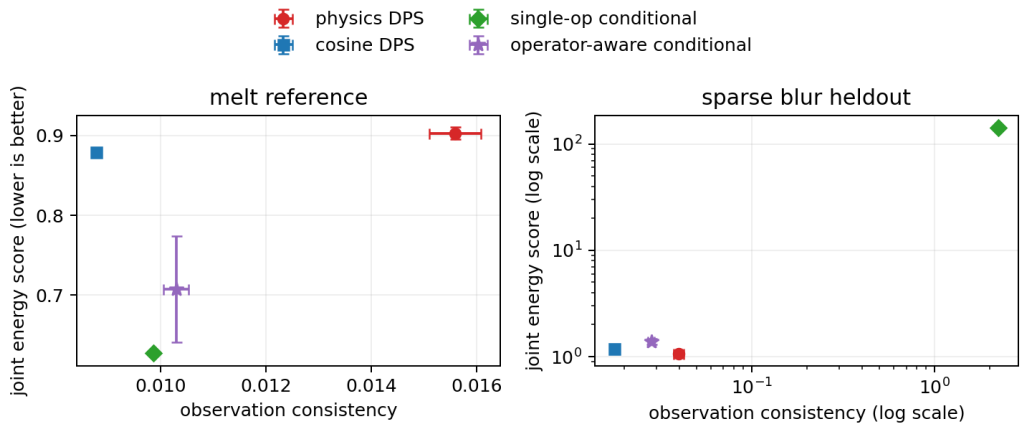

Figure 21: Frozen multi-operator conditional control. The shared operator-aware model retains the melt benefit but pays a specialization penalty (*left*). On held-out sparse blur it removes the single-channel catastrophe but does not beat either reusable-prior DPS arm (*right*; both axes logarithmic). Bars are training-seed standard errors.

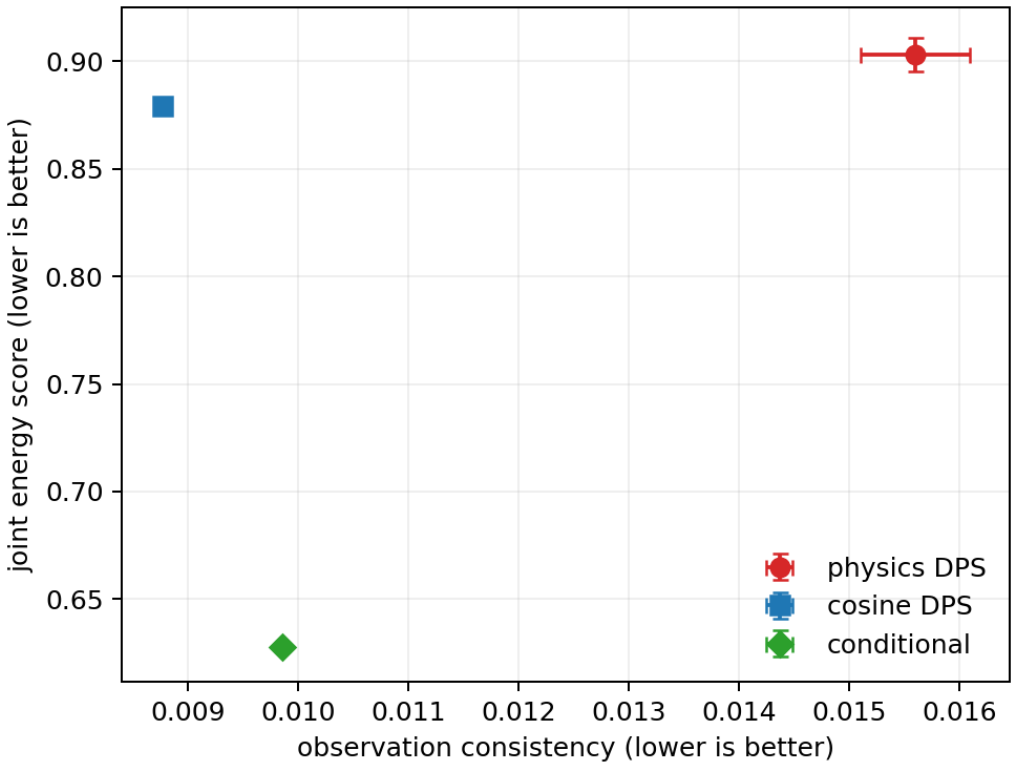

Figure 22: In-distribution task-specific conditional control. Lower is better on both axes. The conditional score improves joint energy relative to both reusable-prior DPS arms without a likelihood gradient, while cosine DPS retains slightly lower observation inconsistency. Bars show training-seed standard errors over five paired seeds.

Paired-data efficiency. With 4000 updates held fixed, the conditional energy scores are 0.640, 0.631, and 0.628 for 72, 144, and 288 paired fields, respectively (Table 27). At 72 fields, conditional-minus-physics is -0.2627 (paired interval $[-0.2852, -0.2298]$; two-level interval $[-0.2930, -0.2270]$), and conditional-minus-cosine is -0.2390 ($[-0.2522, -0.2210]$; $[-0.2526, -0.2214]$); every leave-one-seed-out mean preserves both directions. These are nested paired-data points with a shared scalar salt normalization, not a fixed-epoch study or an interpolated sample-complexity estimate.

Exact nonlinear posterior control. For $p(q,p) \propto \exp(-[((q^2-1)^2/4) + p^2/2]/0.16)$ and the sign-ambiguous observation $y = q^2 + \epsilon$ with $y = 1$ and $\epsilon \sim \mathcal{N}(0, 0.15^2)$, one-dimensional quadrature supplies the exact prior and posterior marginal scores. The IID posterior energy floor is 0.6306. The oracle-conditional ancestral chains give 0.6384–0.6461, exposing a smaller reverse-chain discretization error. With the prior score also exact, DPS at the predeclared main coefficient $\zeta_0 = 2$ gives 1.1497–1.2896; its paired excess over the corresponding oracle-conditional chain is 0.504–0.649, with every interval and leave-one-seed-out mean positive. This low-dimensional result isolates a guidance/sampler failure mode; it does not identify the high-dimensional salt–melt posterior.

High-dimensional learned-prior Langevin control. The frozen PnP-ULA comparison is deliberately not an overall-winner test. Cosine DPS has lower energy on ID melt, whereas PnP-ULA has lower energy under mismatch and sparse blur; both uncertainty levels and every leave-one-seed-out mean agree in all three regimes. PnP-ULA has worse observation consistency and clTru throughout, while its marginal ECE is lower throughout; variogram and RR-ECD directions are mixed. The chain diagnostics stabilize by 200 iterations for ID and mismatch; sparse-blur observation RMSE continues decreasing to 400, so the frozen endpoint is retained rather than selected post hoc.

Complete PnP smoothing-scale sensitivity. The first-four-truth comparison is not stable to smoothing on ID melt or sparse blur. Both uncertainty levels and every leave-one-seed-out mean favor PnP-ULA at $t_\star = .02$ but cosine DPS at $.10$ in those regimes; the known formal $.05$ point favors DPS on ID and PnP-ULA on sparse blur. Melt mismatch favors PnP-ULA at all three levels, with contrasts $-.964$, -1.446, and -1.875. This is a sensitivity result, not a post-hoc replacement of the formal operating point.

Registered non-Gaussian Gibbs audit. For $U(x) = (x^2-1)^2/4$ at $T = .16$, the convex-splitting proposal $(y-x)/h = -y^3 + x + \sqrt{2T/h}\,\xi$ is deterministically energy stable. Ten seeds of 131,072 chains start exactly from the quadrature-normalized Gibbs law. At the registered $h = .10$, 100-step endpoint, W_1 is $.02446$ unadjusted versus $.00239$ after standard MH correction of the same proposal and $.00236$ for the independent IID floor. The paired unadjusted-minus- corrected interval is $[+.02095, +.02308]$, all leave-one-seed-out means are positive, and the corrected acceptance rate is $.970$. The full $h = .02$–$.20$ curve is retained: this extends the stability–measure separation beyond Gaussian targets, but is a scalar Gibbs control and not a new Metropolis method or nonlinear CHC-field result.

Registered 32-dimensional non-Gaussian audit. We next fix $x_j = z_j + 0.35 z_{j-1}$ periodically for 32 IID double-well latent variables z_j. The target is nonlinear and nonseparable in the reported coordinates x, yet the invertible mixing supplies exact target draws and proposal densities. At the same registered $h = .10$ and 100 transitions, pooled latent W_1 is $.02507$ unadjusted, $.00178$ after one joint standard MH decision, and $.00180$ for the IID floor. The paired difference is $.02329$ with interval $[+.02249, +.02405]$. In reported coordinates, sliced-W_1 unadjusted-minus-corrected is $+.00134$ with interval $[+.00091, +.00176]$; every leave-one-seed-out mean is positive and corrected-to-IID ratios are $.986/1.005$ for the two metrics. Acceptance is $.777$ and full deterministic recomputation passes. This is a coupled-coordinate exact-target control, not a nonlinear CHC-field, local-interaction, learned, inverse, or new Metropolis result.

paired fields	val. loss	consistency	energy	clTru
72	0.0272 ± 0.0002	0.0099 ± 0.0000	0.640 ± 0.009	0.0147 ± 0.0002
144	0.0272 ± 0.0002	0.0099 ± 0.0000	0.631 ± 0.002	0.0146 ± 0.0000
288	0.0273 ± 0.0003	0.0099 ± 0.0000	0.628 ± 0.002	0.0146 ± 0.0001

Table 27: Fixed-update paired-data efficiency of the task-specific conditional score. Values are mean ± training-seed standard error over five seeds; all metrics are lower better. The 72/144-field sets are frozen nested prefixes; the 288-field row reuses the original formal conditional result.

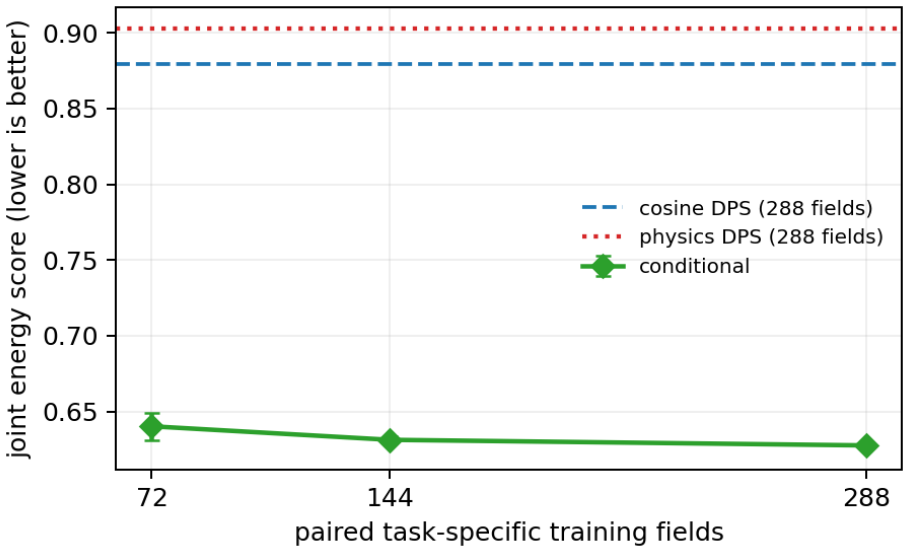

Figure 23: Fixed-update conditional paired-data curve. Diamonds and bars are five-seed means and standard errors; horizontal lines are the two reusable-prior DPS baselines. Every frozen training-set size is shown.

Registered nonlinear nearest-neighbor exact-target audit. To remove the linear and nonlocal-inverse restrictions, we register the triangular chain $x_0 = z_0$ and $x_j = z_j + 0.35x_{j-1} + 0.25\tanh(x_{j-1})$ for 32 IID double-well z_j. Its unit-Jacobian inverse makes the exact target energy a sum of nonlinear nearest-neighbor terms. At $h = .10$ and 100 transitions, pooled local-residual W_1 is .02506 unadjusted, .00196 after one joint standard MH decision, and .00153 for the IID floor. The paired difference is $+.02310$ with interval $[+.02274, +.02342]$. Reported-coordinate sliced-W_1 unadjusted-minus-corrected is $+.00198$ with interval $[+.00081, +.00308]$; every leave-one-seed-out mean is positive and corrected-to-IID ratios are $1.280/1.040$. Acceptance is .777 and full deterministic recomputation passes. This constructed local exact-target control is not a CHC field, conservation result, standard physical phase-field equilibrium, learned or inverse result, or new Metropolis method.

Table 28: Exact nonlinear posterior control. Values are energy score against a fixed IID posterior reference (lower is better), mean ± training-free sampler-seed standard error over five paired seeds. Every frozen guidance point is shown.

Method	Exact OU	BE+FDT	BE+EM
IID posterior floor		0.6306 ± 0.0001	
Oracle conditional	0.6384 ± 0.0002	0.6461 ± 0.0004	0.6409 ± 0.0003
DPS $\zeta_0 = 0.5$	0.6918 ± 0.0010	0.7287 ± 0.0008	0.6858 ± 0.0009
DPS $\zeta_0 = 1$	0.8830 ± 0.0026	0.7795 ± 0.0023	0.7992 ± 0.0044
DPS $\zeta_0 = 2$	1.2560 ± 0.0093	1.1497 ± 0.0116	1.2896 ± 0.0097
DPS $\zeta_0 = 4$	2.0363 ± 0.0260	2.0127 ± 0.0228	1.6218 ± 0.0033

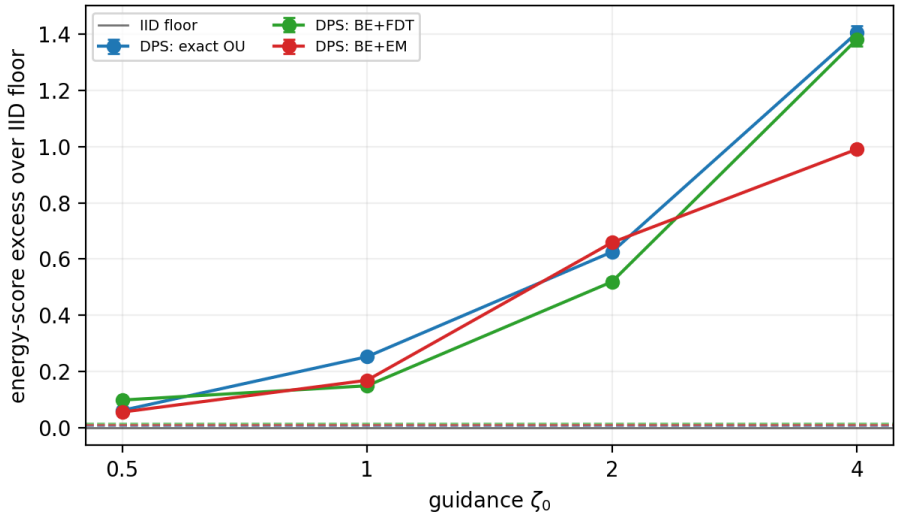

Figure 24: Complete exact-posterior guidance control. Solid curves show DPS energy-score excess over the IID posterior floor; dashed colored lines are the corresponding oracle-conditional ancestral chains. All four frozen coefficients are retained.

regime	cosine DPS	PnP-ULA	PnP−DPS [95% CI]
melt, ID	0.878 ± 0.001	0.905 ± 0.003	+0.027 [+0.023,+0.032]
melt, mismatch	11.538 ± 0.006	10.101 ± 0.014	-1.437 [-1.450,-1.424]
sparse blur	1.188 ± 0.001	1.059 ± 0.002	-0.130 [-0.134,-0.126]

Table 29: Frozen high-dimensional alternative-sampler control. The primary outcome is joint posterior energy score (lower is better); values are mean ± training-seed standard error and contrasts use paired-seed intervals. Two-level seed–truth intervals and all secondary outcomes are in the supplement.

regime	$t_\star$	cosine DPS	PnP-ULA	PnP−DPS [95% CI]
melt, ID	0.02	0.884 ± 0.001	0.636 ± 0.001	-0.248 [-0.250,-0.246]
	0.05	0.884 ± 0.001	0.908 ± 0.003	+0.024 [+0.020,+0.029]
	0.10	0.884 ± 0.001	1.551 ± 0.005	+0.667 [+0.657,+0.676]
melt, mismatch	0.02	11.625 ± 0.007	10.661 ± 0.019	-0.964 [-0.986,-0.936]
	0.05	11.625 ± 0.007	10.179 ± 0.014	-1.446 [-1.459,-1.433]
	0.10	11.625 ± 0.007	9.750 ± 0.005	-1.875 [-1.885,-1.864]
sparse blur	0.02	1.179 ± 0.001	0.903 ± 0.005	-0.277 [-0.287,-0.268]
	0.05	1.179 ± 0.001	1.054 ± 0.003	-0.125 [-0.131,-0.120]
	0.10	1.179 ± 0.001	1.526 ± 0.004	+0.347 [+0.337,+0.355]

Table 30: Complete learned-prior PnP-ULA smoothing-scale sensitivity. The primary outcome is joint energy score (lower is better); every frozen level is reported. The $t_\star = .05$ row reuses the first four fields of the formal eight-truth result; other rows were unseen when this curve was frozen.

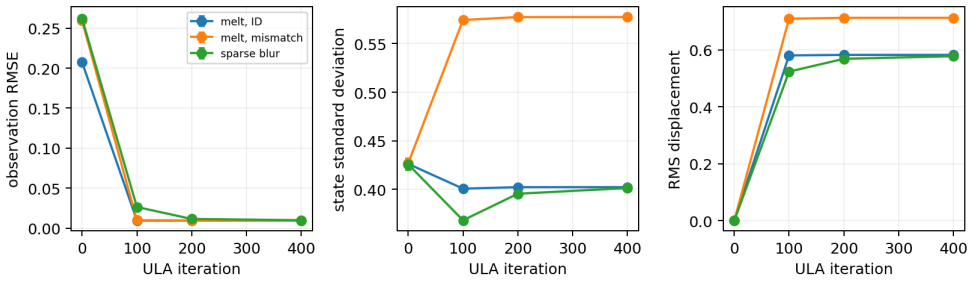

Figure 25: Learned-prior PnP-ULA chain diagnostics at every predeclared checkpoint. Curves are means with training-seed standard errors over the same eight truths. ID and mismatch statistics plateau by 200 steps; sparse-blur observation RMSE continues improving through the frozen 400-step endpoint. These are convergence diagnostics, not exact-posterior tests.

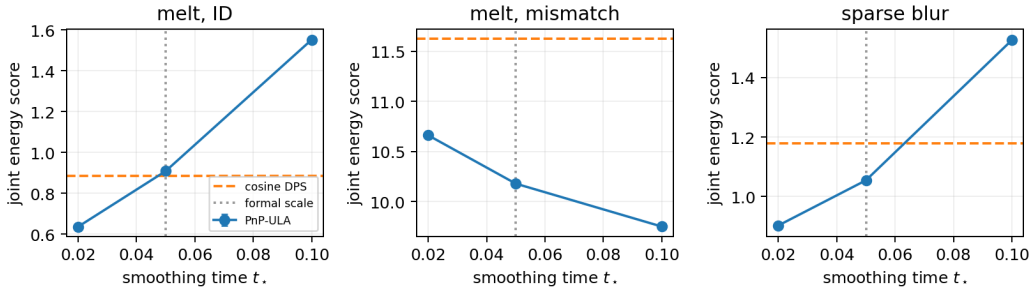

Figure 26: Complete learned-prior smoothing-scale curve on the first four matched truths. Points and bars are five-seed means and standard errors; dashed lines are the matched cosine DPS baselines and the vertical dotted line is the known formal scale. Lower is better.

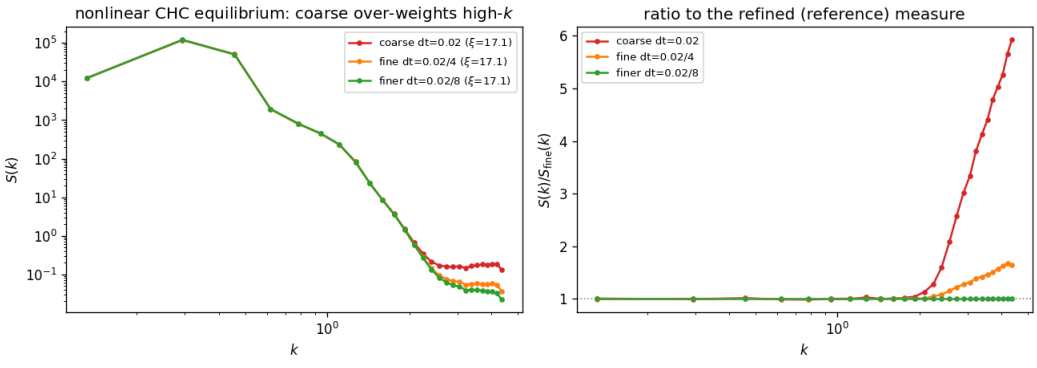

Figure 27: Nonlinear CHC scope control. Refining $dt = 0.02 \to 0.02/8$ changes only the extreme high-k spectral tail; the measured correlation length remains $\xi = 17.1$. Thus the linear counterexample does not imply a large bulk-morphology error.

h	steps	unadjusted W_1	same proposal + MH	acceptance
0.02	1	0.0027 ± 0.0004	0.0027 ± 0.0004	0.997
0.02	100	0.0062 ± 0.0003	0.0032 ± 0.0004	0.997
0.05	1	0.0033 ± 0.0003	0.0026 ± 0.0004	0.988
0.05	100	0.0135 ± 0.0002	0.0027 ± 0.0005	0.988
0.10	1	0.0059 ± 0.0003	0.0028 ± 0.0004	0.971
0.10	100	0.0245 ± 0.0003	0.0024 ± 0.0005	0.970
0.20	1	0.0134 ± 0.0002	0.0026 ± 0.0004	0.934
0.20	100	0.0450 ± 0.0003	0.0024 ± 0.0003	0.934

Table 31: Frozen nonlinear Gibbs-measure control. The deterministic convex-splitting proposal is energy stable; the standard Metropolis–Hastings correction uses the same proposal and restores the registered target by detailed balance. Values are mean ± seed standard error over ten seeds with 131,072 chains per seed.

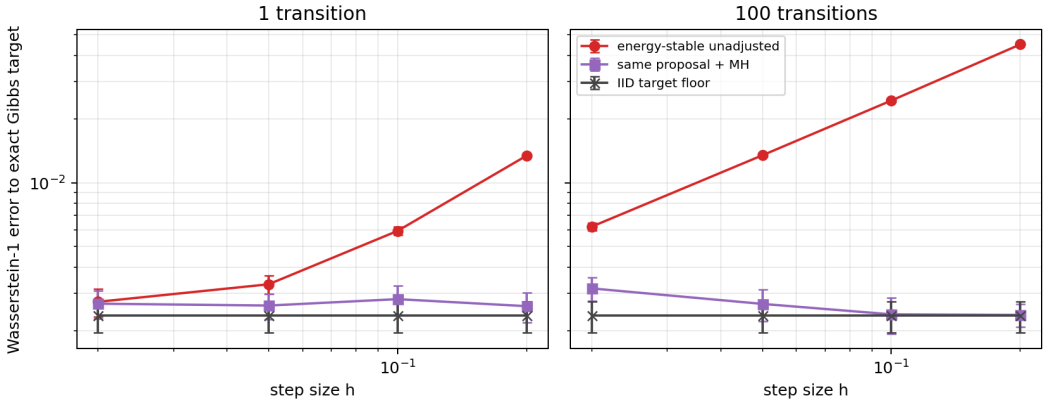

Figure 28: Frozen non-Gaussian Gibbs control. The energy-stable unadjusted proposal departs from the exact target as step size and transition count grow; standard MH correction of the same proposal remains at the IID finite-sample floor. Bars are ten-seed standard errors.

h	steps	raw pooled	MH pooled	raw sliced	MH sliced	acc.
0.02	1	0.0018	0.0018	0.0194	0.0194	0.973
0.02	100	0.0052	0.0019	0.0183	0.0181	0.973
0.05	1	0.0025	0.0017	0.0194	0.0194	0.907
0.05	100	0.0133	0.0016	0.0193	0.0189	0.905
0.10	1	0.0054	0.0018	0.0195	0.0195	0.776
0.10	100	0.0251	0.0018	0.0200	0.0186	0.777

Table 32: Frozen 32-dimensional non-Gaussian control. The target is nonseparable in reported coordinates under the fixed local mixing $x_j = z_j + 0.35 z_{j-1}$. The energy-stable proposal is shown unadjusted and with standard joint MH correction.

h	steps	raw residual	MH residual	raw sliced	MH sliced	acc.
0.02	1	0.0019	0.0018	0.0210	0.0210	0.973
0.02	100	0.0056	0.0021	0.0210	0.0211	0.973
0.05	1	0.0026	0.0019	0.0210	0.0210	0.903
0.05	100	0.0132	0.0021	0.0215	0.0210	0.905
0.10	1	0.0052	0.0018	0.0208	0.0208	0.777
0.10	100	0.0251	0.0020	0.0236	0.0216	0.777

Table 33: Frozen nonlinear nearest-neighbor exact-target control. The target energy is a sum of local terms under the registered triangular map. The energy-stable proposal is shown unadjusted and with standard joint MH correction.

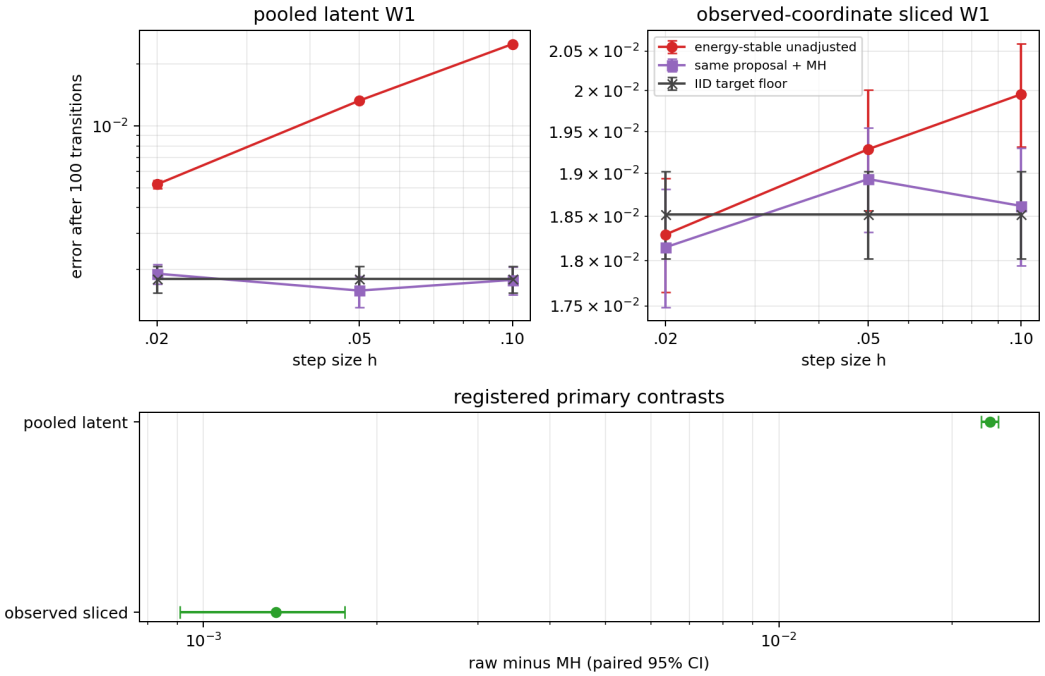

Figure 29: Frozen 32-dimensional non-Gaussian control. The unadjusted energy-stable proposal departs from the exact target as step size grows, while joint standard MH remains at the IID floor. The bottom panel reports the predeclared paired primary contrasts with 95% intervals.

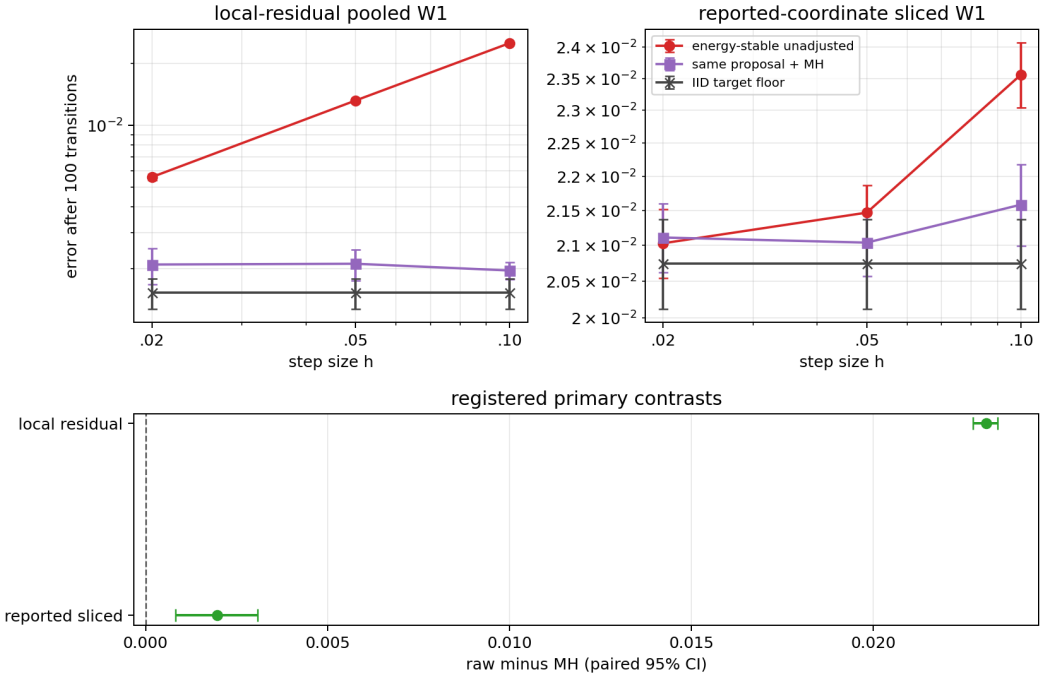

Figure 30: Frozen nonlinear nearest-neighbor exact-target control. The energy-stable unadjusted proposal departs from both the local-residual and reported-coordinate target laws as step size grows, while joint standard MH remains near the IID floor. The bottom panel reports the predeclared paired primary contrasts with 95% intervals.

F NONLINEAR LEARNING EXPERIMENTS AND PROVENANCE

The main paper foregrounds the completed fixed-$T = 1$ experiment. This section retains the earlier fixed-transition discovery, backbone replication, and sampler diagnosis, including their original numerical results and outcome-dependent design history. The nonlinear learning task uses CHC endpoint fields as data for a separate, common zero-mass Gaussian VP corruption; physical horizon T and learning time τ have distinct roles. Table 34 summarizes cohort and checkpoint reuse.

Reordering the presentation does not change registered estimands or convert reused observations into independent replications.

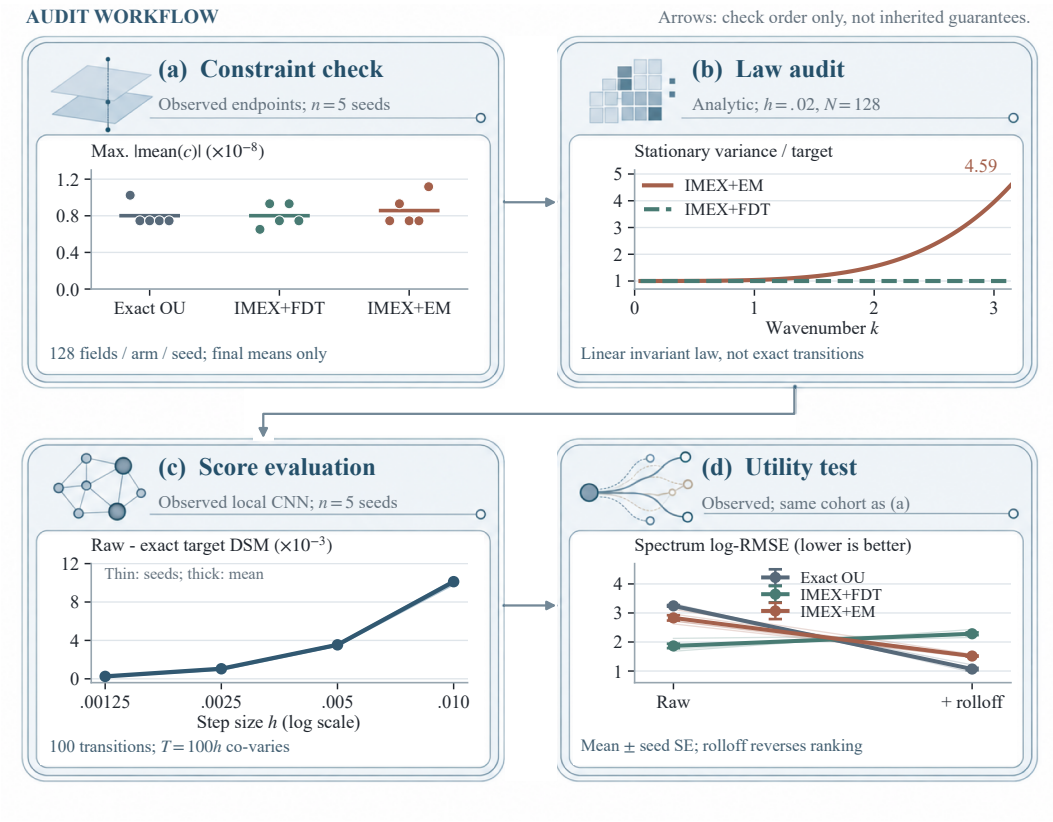

Figure 31: Supporting audit overview with evidence from separate experiments. Arrows indicate check order, not inherited guarantees or a single experimental pipeline. Header glyphs are conceptual, not data or model architecture. (a) Observed maximum absolute spatial mean over 128 final fields per arm and seed ($n = 5$); bars denote seed means, not a drift threshold. (b) Analytic stationary variance ratios for the specified stable linear chain. (c) Initial five-seed CNN discovery with 100 transitions, so $T = 100h$ varies; thin lines are seeds and the thick line is their mean. (d) Observed spectral error for the same seeds as (a), with means and seed SEs; rolloff reverses the arm ordering. Panel (c) is supporting evidence; the main fixed-time comparison is Main-paper Fig. 4.

Experiment	Design	Relationship to the main result
Initial CNN study	$n = 5$; 100 steps	Separate data/training seeds; $T = 100h$ varies.
FNO replication and paired CNN	Shared $n = 10$; 100 steps	FNO cohort new relative to the initial study; paired CNN uses the same data.
Fixed-time Stage A	Same $n = 10$; $T = 1$	Reuses reference fields; endpoints generated on shared fine Brownian paths. Designed after the earlier results.
Fixed-time Stage B	Same $n = 10$; $T = 1$	Run only after a positive Stage A; new endpoint-trained models, reused reference-trained models.

Table 34: History of the nonlinear learning experiments. All rows use the same finite Gibbs target; the five evaluation streams per seed are Monte Carlo replicates. The fixed-time experiment reported in the main text shares the cohort of the earlier rows.

Registered two-dimensional nonlinear mass-conserving CHC scale control. We discretize the periodic double-well CHC free energy on a finite 15×15 grid and work in its 224-dimensional orthonormal zero-mass subspace. Two separately tuned HMC kernels provide disjoint starting, target, IID-floor, and cross-kernel reference groups; every split-$\widehat{R}$ is at most 1.01 and cross-kernel errors stay at the internal finite-sample floors. At the registered $dt = .0025$, 100-step cell, raw/MH/IID W_1 is $.02143/.00064/.00071$ for free-energy density and $.04219/.00118/.00138$ for squared-interface density. The energy raw-minus-MH interval is $[+.02054, +.02105]$. Two local morphology statistics agree: sign-interface W_1 is $.02592/.00345/.00428$ and corner-density W_1 is $.01681/.00239/.00273$. All four registered intervals and every leave-one-seed-out mean are positive; acceptance is .642, mass error is below 10^{-12}, and all ten seeds replay byte-identically. This result has frozen label `nonlinear_chc_15x15_scale_robustness`. It establishes finite-grid scale robustness, not continuum convergence, resolved large-domain morphology, thermodynamic-limit behavior, learned/inverse benefit, correctness for arbitrary CHC discretizations, or a new Metropolis method.

Smaller-grid and one-dimensional support. The separately frozen finite 7×7 field agrees: energy-density raw/MH/IID $W_1 = .04183/.00073/.00053$ with interval $[+.04082, +.04141]$; squared-interface, sign-interface, and corner-density values are $.08842/.00175/.00119$, $.05133/.00278/.00307$, and $.02113/.00181/.00197$. Its frozen label is `nonlinear_chc_2d_field_measure_and_morphology_separation`. This is supporting finite-grid evidence, not the lead result.

The earlier finite-dimensional 1D CHC Gibbs-field result agrees: energy-density raw/MH/IID $W_1 = .06538/.00138/.00153$, energy contrast interval $[+.06309, +.06492]$, and interface-density $W_1 = .12113/.00264/.00278$. It is supporting dimension-sensitivity evidence, not the lead result.

Registered nonlinear transition-bias-to-target-DSM dose response. The preceding scale, single-dose CNN/FNO, and sampler results were outcome-known. Before any mode of this new experiment was run or any scientific outcome was inspected, we found that the v1 seed labels 5–9

h	steps	raw energy	MH energy	IID energy	MH acc.
0.00125	1	0.0010	0.0007	0.0007	0.856
0.00125	100	0.0105	0.0007	0.0007	0.854
0.00250	1	0.0029	0.0007	0.0007	0.636
0.00250	100	0.0214	0.0006	0.0007	0.642
0.00500	1	0.0095	0.0007	0.0007	0.277
0.00500	100	0.0429	0.0008	0.0007	0.280
0.01000	1	0.0287	0.0007	0.0007	0.023
0.01000	100	0.0867	0.0008	0.0007	0.024

Table 35: Frozen finite 15×15 nonlinear mass-conserving CHC Gibbs-field scale control. Values are ten-seed mean Wasserstein-1 errors of free-energy density against an independent HMC reference. The standard MH arm corrects the same stabilized IMEX-plus-continuum-noise proposal; IID is a disjoint HMC finite-sample floor.

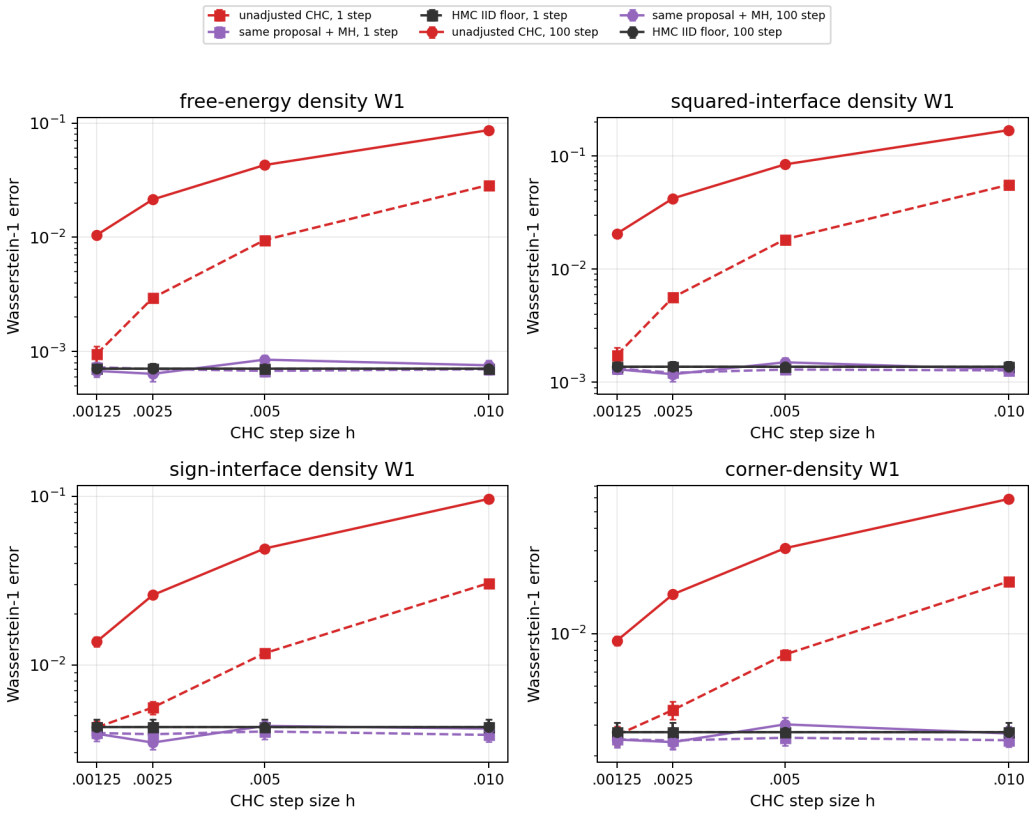

Figure 32: Frozen finite 15×15 nonlinear mass-conserving CHC Gibbs-field scale control. The unadjusted stabilized IMEX proposal departs from independent HMC references as step size and transition count grow; standard MH correction of the same proposal remains at the IID reference floor. Registered energy, squared-interface, sign-interface, and corner-density contrasts are all positive.

h	Energy-W_1	Raw−exact DSM	Relative
0.00125	0.010449	+0.000251	+0.052%
0.00250	0.021351	+0.001046	+0.219%
0.00500	0.043104	+0.003540	+0.741%
0.01000	0.086754	+0.010119	+2.119%
Slope	–	+0.003210	t [+0.003135,+0.003284]
High−low	–	+0.009868	t [+0.009641,+0.010096]

Table 36: Frozen nonlinear CHC target-DSM dose response. Energy-W_1 and DSM entries are grand means over five new training/data seeds; DSM first averages five evaluation streams within seed. The slope uses $x = \log_2(h/0.00125)$. Student-t intervals use four degrees of freedom; paired seed-bootstrap intervals are slope [+0.003160,+0.003254] and endpoint [+0.009714,+0.010001].

would reuse HMC random seeds from the outcome-known scale control; the frozen v2 protocol therefore uses new training/data seeds 10–14 and HMC seeds $2031081500 + s$. It pairs a new exact-target training arm with four new raw 100-transition arms at $h \in \{.00125, .0025, .005, .01\}$, sharing initial states, innovations, network initialization, and training randomness within seed. Five held-out target-DSM evaluation streams are averaged within each seed; they are Monte Carlo replicates rather than independent units. The exact-target arm uses independent HMC target fields and is not the earlier MH-endpoint arm.

Across the four $\log_2$-spaced doses, raw endpoint energy-W_1 error against the held-out target is $.010449, .021351, .043104, .086754$, while raw-trained-minus-exact-trained target DSM is $+.000251, +.001046, +.003540, +.010119$ (cellwise relative means $+.052\%, +.219\%, +.741\%, +2.119\%$). The seed-level slope is $+.00320984$ with Student-t_4 95% CI $[+.00313547, +.00328420]$ and paired 20,000-draw seed-bootstrap CI $[+.00316035, +.00325419]$; high-minus-low is $+.00986829$ with corresponding intervals $[+.00964079, +.01009580]$ and $[+.00971420, +.01000051]$. The five seed slopes are $+.003172, +.003240, +.003283, +.003130, +.003224$; the five descriptive evaluation-stream slopes are $+.003054, +.003278, +.003207, +.003036, +.003475$, and both leave-one-seed/stream-out families are positive. The predeclared exact sign test is one-sided $p = .03125$ and two-sided $p = .0625$.

From earliest to latest, the four time-quartile gap vectors are $(.000354, .000366, .000221, .000063)$, $(.001116, .002252, .000751, .000093)$, $(.003209, .008300, .002607, .000153)$, and $(.007728, .024143, .008669, .000269)$. Thus the seed-level dose trend is not a claim of uniform effect over diffusion time or every seed–stream cell. All 125 evaluation records and 100 paired contrasts are retained in the machine-readable bundle; formal and independent replay JSON are byte-identical, dataset tensors and checkpoint-state hashes match, and the frozen label is `nonlinear_transition_bias_increases_target_dsm_risk_with_step`. All preregistered upstream and positive learned-dose criteria passed; sign-reversed alternative-branch predicates are retained with their exact identifiers in the machine-readable summary.

This result is limited to one finite 15×15 conserved Gibbs target, one local CNN, and fixed 100-transition implementations. It establishes neither continuum order nor architecture/sampler independence, generation or inverse-task improvement, practical large-scale MH, or real-data validity.

Frozen nonlinear learned-score bridge. The preceding unlearned 15×15 raw–MH separation was outcome-known; before any learned outcome, we froze five HMC dataset/training seeds, five new evaluation streams, a 32-channel four-block circular ScoreNet, and a common analytic zero-mass VP learning problem. Each arm uses 1,536 endpoint fields from shared exact-reference starts and proposal innovations; evaluation uses a disjoint 1,536-field HMC target group. MH-trained target DSM MSE is $.477009$ versus $.478238$ for raw-trained, giving raw-minus-MH $+.001229$ $[+.001124, +.001333]$, or $+.257\%$ relative to raw DSM. A disclosed post-hoc small-n sensitivity gives Student-t_4 CI $[+.001070, +.001388]$; five positive seed means imply exact one-sided $p = .03125$ and two-sided $p = .0625$. All five training-seed means ($+.001045$ to $+.001405$), all five evaluation-stream means ($+.001042$ to $+.001412$), and both leave-one-out families are positive, yielding `nonlinear_target_repair_improves_learned_score`. Formal and independent replay JSON, dataset tensors, and checkpoint state hashes match exactly.

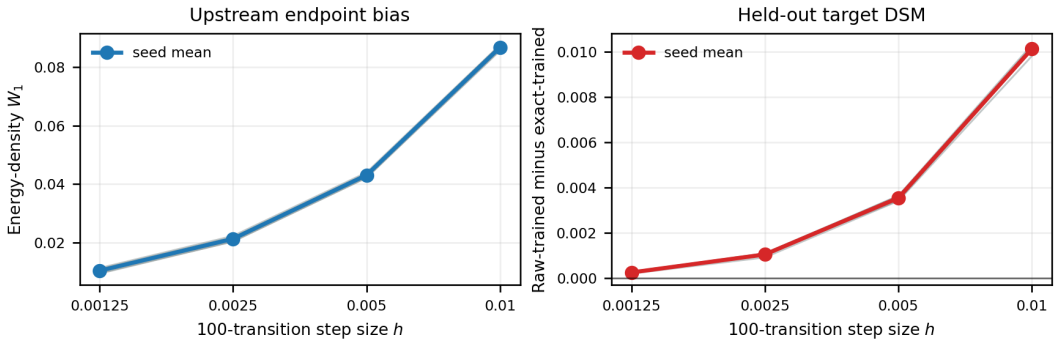

Figure 33: Initial nonlinear implementation-dose audit. Gray curves show five paired training/data seeds and their mean. At 100 transitions, increasing $\log_2$-spaced h and exposure $T = 100h$ raises endpoint energy-W_1 and raw-minus-exact target DSM. Five evaluation streams are Monte Carlo replicates, not additional independent units.

outcome	raw mean	MH mean	raw−MH [95% CI]
target DSM MSE	0.478238	0.477009	+0.001229 [+0.001124,+0.001333]
generated energy W1	–	–	-0.010756 [-0.011609,-0.010009]
generated interface W1	–	–	-0.022177 [-0.023294,-0.020782]
generated sign-interface W1	–	–	-0.012528 [-0.014160,-0.010382]
generated corner W1	–	–	-0.010253 [-0.011503,-0.008778]
generated site W1	–	–	-0.002887 [-0.004656,-0.000973]

Table 37: Frozen nonlinear learned-score bridge. Lower is better. The predeclared target-DSM contrast favors MH-trained scores, but every secondary generated-W_1 contrast is negative (raw-trained lower) under the common 100-step reconstruction–fresh-noise sampler. Thus the experiment supports a target score-risk improvement, not a generated-fidelity or inverse-task improvement.

This bridge is finite and narrow: one nonlinear Gibbs target, one proposal/MH correction, one primary network and optimizer, and no inverse or real-data task. The sampler-dependent secondary reversal is retained rather than used to upgrade the primary score-risk claim: under the original sampler, all five tested generated-W_1 contrasts reverse.

Post-hoc periodic-FNO architecture transfer. After the local-CNN bridge and sampler diagnosis were outcome-known, we froze a non-isomorphic periodic FNO sensitivity before any FNO scientific metric. It reuses the five exact dataset bundles, paired random streams, 6,000-update Adam problem, and five target-DSM evaluation streams. The three-block width-18 FNO retains all independent 15×15 Fourier modes and has 231,349 parameters versus 231,649 for the local ScoreNet. Its MH-trained target DSM MSE is $.477866$ versus $.478757$ for raw-trained, giving raw-minus-MH $+.000891$ $[+.000568, +.001192]$, or $+.186\%$ relative to raw DSM. The post-hoc Student-t_4 CI is $[+.000395, +.001387]$; five positive seed means again give exact one-sided $p = .03125$ and two-sided $p = .0625$. All five training-seed means ($+.000292$ to $+.001393$), all five evaluation-stream means ($+.000660$ to $+.001046$), and both leave-one-out families are positive, yielding `nonlinear_target_repair_transfers_to_periodic_fno`. Formal and independent replay JSON are byte-identical, and all dataset and checkpoint-state hashes match.

This disclosed sensitivity transfers only the target-risk direction across the two tested backbones at near-matched parameter scale. It does not establish stable magnitude, capacity invariance, arbitrary-backbone independence, improved generation, or sampler robustness.

Before scientific unblinding of the separate fresh periodic-FNO dose cohort, we froze a validation-only portability amendment for its registered float32 zero-mass check. CUDA and CPU reductions must each pass the unchanged registered threshold but need not be bit-identical; the amendment changes no scientific artifact, estimand, decision rule, or label.

Backbone	Parameters	Raw−MH target DSM	Status
Local ScoreNet	231,649	$+0.001229\,[+0.001124, +0.001333]$	outcome-known
Periodic FNO	231,349	$+0.000891\,[+0.000568, +0.001192]$	robust positive

Table 38: Disclosed post-hoc nonlinear architecture-transfer sensitivity. Intervals bootstrap five paired training-seed means after averaging five frozen evaluation streams. Positive raw-minus-MH values favor MH-corrected training data.

h	Energy-W_1	Raw−exact DSM	Relative
0.00125	0.010395	+0.000249	+0.052%
0.00250	0.021263	+0.000997	+0.206%
0.00500	0.042869	+0.003439	+0.709%
0.01000	0.086320	+0.009848	+2.028%
Slope	–	+0.003124	t [+0.003009,+0.003238]
High−low	–	+0.009599	t [+0.009249,+0.009948]

Table 39: Fresh periodic-FNO target-DSM dose response on ten independent new training/data seeds. DSM first averages five Monte Carlo evaluation streams within seed. The slope uses $x = \log_2(h/0.00125)$. Student-t intervals use nine degrees of freedom; paired seed-bootstrap intervals are slope [+0.003030,+0.003217] and endpoint [+0.009314,+0.009884].

Fresh periodic-FNO four-dose result. Portable strict validation assigns the formal label `fresh_periodic_fno_transition_bias_increases_target_dsm_risk` and reproduces the formal/replay evidence before manuscript admission. The five evaluation streams are Monte Carlo replicates averaged within each of the ten independent training/data seeds; neither streams nor replay copies enlarge n.

In Table 39, "Relative" denotes the relative cell mean for the four dose rows; in the final slope and high–low rows the same column instead reports the registered Student-t_9 95% interval. The clean aggregate positive label does not imply a uniform-time effect. Here a "time quartile" is one of four equal bins in $\log \tau$ over $\tau \in [.001, 8]$, with edges approximately $.001, .009457, .089443, .845897, 8$ (the first three bins are right-open and the last includes its right endpoint). The final reported log-time-bin slope and endpoint are positive for seven of ten seeds, whereas the first three bins are positive for ten of ten. The final-bin four-dose mean contrasts are -1.30×10^{-6}, $+.61 \times 10^{-6}$, $+13.33 \times 10^{-6}$, and $+75.28 \times 10^{-6}$. These supporting fixed-transition results therefore do not establish an effect uniform over learning time.

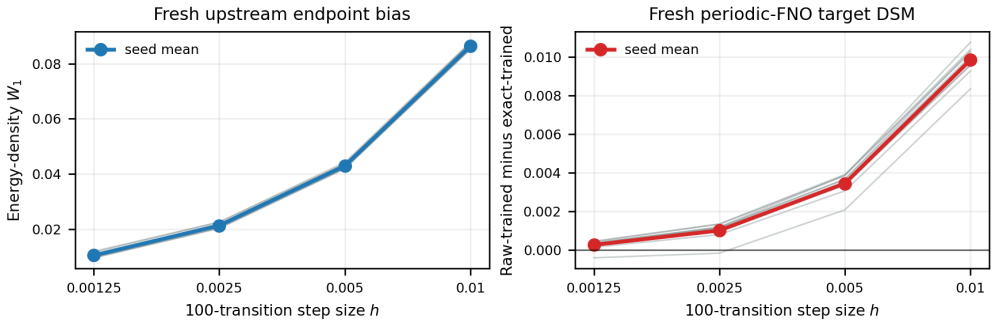

Figure 34: Fresh periodic-FNO four-dose replication. *Left*: upstream endpoint energy-W_1 against the held-out Gibbs target. *Right*: raw-trained-minus-exact-trained held-out target-DSM. Curves use ten independent new training/data seeds; five evaluation streams are averaged within seed. The result concerns one fixed periodic FNO, one finite target, and fixed 100-transition implementations, not arbitrary backbones, generated-sample quality, or inverse utility.

Frozen sampler/step robustness diagnosis. Because the preceding 100-step reconstruction–fresh-noise reversal was outcome-known, we froze a post-hoc grid before evaluating any other cell: the same ten checkpoints, five new generation streams, 512 fields per arm, 50/100/200 reverse levels, and the original fresh-noise update plus generalized DDIM at $\eta \in \{0, 1\}$. The original update is named literally and is not a canonical DDPM transition. Its 100-step stream-0 tensors and metrics reproduce exactly; all 225 formal records and 225 independent replay records are byte-identical and pass strict validation.

The two predeclared primary energy-W_1 contrasts under deterministic DDIM are $+.013848$ $[+.002625, +.022577]$ at 100 steps and $+.017661$ $[+.008096, +.024928]$ at 200. The former is unresolved because training-seed 0 remains negative ($-.006515$) despite positive stream and leave-one-out means; the latter passes every seed, stream, interval, and leave-one-out sign rule. Hence the frozen label is `ddim_generated_energy_sampler_sensitive_or_unresolved`, not a selected positive operating point. Fresh-noise energy contrasts remain robustly negative at all three step counts; stochastic DDIM is negative/unresolved/positive at 50/100/200, respectively.

This diagnostic covers one finite 15×15 target, three related VP updates, and three step counts. It establishes sampler/step sensitivity in this grid, not arbitrary-sampler, PDE, architecture, inverse-task, real-data, or large-scale-MH robustness.

Prospectively frozen post-outcome paired implementation-dose audit. This prospectively frozen post-outcome local CNN replication and paired CNN/FNO cohort uses a new $n = 10$ cohort on two tested backbones; the backbone observations share the same cohort and are not independent confirmations. The old n=5 discovery remains separate and is never pooled with the new cohort. Training/data seed is the only independent unit. Five evaluation streams are Monte Carlo replicates, while formal/replay is deterministic verification; neither streams, architectures, nor replay copies are additional independent units. Each arm uses fixed 100 transitions and 6,000 training updates on one finite 15×15 target. Consequently $h = .00125, .0025, .005, .010$ co-varies with $T = 100h = .125, .25, .5, 1$: this is a fixed-transition-count joint step-size/exposure implementation dose, not a pure step-size intervention.

Shared upstream endpoint energy-W_1 increases from .010395 to .086320. CNN raw-minus-exact target-DSM gaps increase from $+.000356$ to $+.010643$, and FNO gaps from $+.000249$ to $+.009848$. Both tested CNN and FNO show the robust positive direction, and the six-component IUT passed. All ten seed slopes and endpoints are positive for both learned components (two-sided exact sign $p = .001953$ each); CNN/FNO slopes are $+.003346/+.003124$ and high–low contrasts are $+.010287/+.009599$. The complete four-dose curves are reported in Table 41 and Fig. 35.

method	steps	raw W1	MH W1	raw−MH [95% CI]
fresh_noise	50	0.065375	0.075779	-0.010404 [-0.011082,-0.009857]
fresh_noise	100	0.065738	0.076504	-0.010766 [-0.011670,-0.009962]
fresh_noise	200	0.065886	0.076797	-0.010911 [-0.012295,-0.009807]
ddim_eta0	50	0.017837	0.010538	+0.007299 [-0.006139,+0.019973]
ddim_eta0	100*	0.020113	0.006265	+0.013848 [+0.002625,+0.022577]
ddim_eta0	200*	0.022674	0.005013	+0.017661 [+0.008096,+0.024928]
ddim_eta1	50	0.007627	0.012534	-0.004907 [-0.008635,-0.002337]
ddim_eta1	100	0.013578	0.007390	+0.006188 [+0.002465,+0.008440]
ddim_eta1	200	0.017283	0.004239	+0.013045 [+0.009819,+0.015335]

Table 40: Frozen sampler/step diagnosis of generated free-energy fidelity. Lower W1 is better and positive raw-minus-MH favors the MH-trained score. Stars mark the two primary cells. Robust sign requires the paired interval, all five training-seed and generation-stream means, and both leave-one-out families to agree; a positive interval alone is insufficient.

h	Upstream W_1	CNN Δ	FNO Δ	FNO−CNN	CNN rel. (%)	FNO rel. (%)
0.00125	0.010395	+0.000356	+0.000249	-0.000107	+0.073	+0.052
0.00250	0.021263	+0.001231	+0.000997	-0.000234	+0.254	+0.206
0.00500	0.042869	+0.003834	+0.003439	-0.000395	+0.790	+0.709
0.01000	0.086320	+0.010643	+0.009848	-0.000795	+2.194	+2.028

Table 41: Paired local-CNN/FNO target-DSM dose response on ten shared training/data seeds. Five evaluation streams are averaged within seed before inference. Student-t intervals use nine degrees of freedom and the common paired whole-seed bootstrap uses 20,000 draws. The formal label is `paired_two_backbone_transition_bias_increases_target_dsm_risk`. The two backbone observations share data/evaluation blocks and are not independent confirmations. FNO-minus-CNN interactions are secondary descriptive only and license neither equivalence nor superiority.

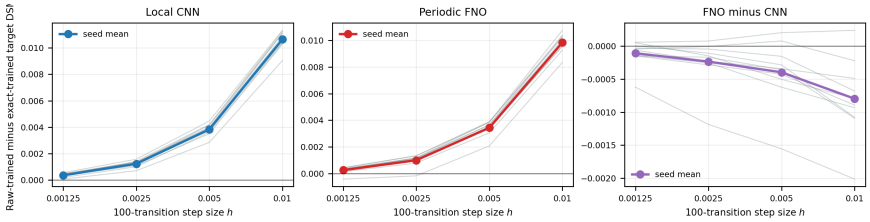

Figure 35: Post-outcome paired local-CNN/FNO implementation-dose audit on ten shared training/data seeds. Each backbone's gray curves are seed-level raw-minus-exact target-DSM gaps and colored curves are cohort means. The two tested backbones share the same cohort and are not independent confirmations; fixed 100 transitions make h and exposure $T = 100h$ co-vary.

The signed FNO-minus-CNN interactions are $\gamma = -.000223$ for slope and $\zeta = -.000688$ for high–low contrast. Gamma and zeta are secondary descriptive only and do not enter the joint label. A zero-crossing interval would not be equivalence, and no equivalence margin or directional interaction gate was registered. The mandatory exact-target baseline is also secondary; its FNO-minus-CNN mean is $+.000805$. These quantities license neither CNN/FNO superiority nor equivalence. The formal label is `paired_two_backbone_transition_bias_increases_target_dsm_risk`.

The result is not architecture-independent and not magnitude-stable. It is restricted to the two tested backbones, the same paired cohort, one finite target, and the registered fixed-transition implementations. It supplies no generation or inverse result and establishes neither sampler independence,

Table 42: Post-outcome-selected, prospectively frozen fixed-$T = 1$ nonlinear-CHC control on one finite target with $n = 10$ source/data seeds.

Arm	h	Transitions	T	Dose	Energy-W_1
fixedT_h00125	0.00125	800	1	0	0.01087911
fixedT_h00250	0.0025	400	1	1	0.021647215
fixedT_h00500	0.005	200	1	2	0.04328787
fixedT_h01000	0.01	100	1	3	0.086750875

Estimand	Mean	Student-t 95% CI	Bootstrap 95% CI	Signs $+/-/0$
Slope	0.024925595	[0.0248792, 0.02497199]	[0.024888581, 0.024964219]	10/0/0
High-minus-low	0.075871766	[0.075729534, 0.076013997]	[0.075758454, 0.075990187]	10/0/0

High-minus-low and both intervals are registered raw inference fields, not differences recomputed from rounded arm means. Secondary metrics are descriptive only.
Formal label: `nonlinear_chc_fixed_time_energy_w1_increases_with_step`. This fixed-$T = 1$ finite-target result establishes neither continuum/order convergence nor causal mediation.

continuum or scaling behavior, practical large-scale MH, causal mediation, arbitrary-PDE generalization, nor real-data validity.

Fixed-physical-time Stage A. The common-$T = 1$ unlearned control couples every arm through the same fine Brownian path and treats the source/data seed as the independent unit ($n = 10$). Table 42 reports registered raw means, Student-t intervals, whole-seed bootstrap intervals, and exact-sign counts; neither endpoint contrasts nor intervals are reconstructed from rounded arm means. The formal label is `nonlinear_chc_fixed_time_energy_w1_increases_with_step`. In the Python 3.11 artifact environment, a disclosed post-outcome read-only portability audit retains every original source, tensor, and formal–replay gate and permits only finite floating-point reconstruction differences below $10^{-15} + 10^{-12} \max(|a|, |b|)$; all nonfloating fields, decisions, labels, hashes, and receipt bindings remain exact, and no scientific byte is changed.

Fixed-physical-time Stage B. Stage B was triggered only by the exact positive Stage A label and reuses its arm definitions and $n = 10$ seed cohort. The fixed CNN and fixed FNO are a six-component hierarchical analysis, not independent confirmations. The post-outcome validation-only repair changed no scientific bytes and performed no retraining; it corrected only the registered witness serialization and removal of duplicate physical checkpoint hashes from role-neutral comparison. Γ and ζ are descriptive only, not superiority or equivalence tests. The formal label is `nonlinear_chc_fixed_time_two_backbone_dose_increases_target_dsm_risk`.

Table 43: Conditional post-outcome fixed-$T = 1$ learned response in the same $n = 10$ cohort, given the exact positive Stage A gate.

Arm	h	Transitions	T	Dose	Stage A W_1	CNN DSM	FNO DSM
fixedT_h00125	0.00125	800	1	0	0.01087911	0.00042209566	0.00027611196
fixedT_h00250	0.0025	400	1	1	0.021647215	0.0013209385	0.0010195452
fixedT_h00500	0.005	200	1	2	0.04328787	0.0040349305	0.003476603
fixedT_h01000	0.01	100	1	3	0.086750875	0.010882877	0.0098824823

Component / estimand	Mean	Student-t 95% CI	Bootstrap 95% CI	Signs $+/-/0$
CNN slope	0.0034096336	[0.0033256304, 0.0034936369]	[0.0033407147, 0.0034788429]	10/0/0
CNN high-minus-low	0.010460781	[0.010209115, 0.010712448]	[0.010255135, 0.01066894]	10/0/0
FNO slope	0.0031276169	[0.0030123282, 0.0032429056]	[0.0030209593, 0.0032061989]	10/0/0
FNO high-minus-low	0.0096063703	[0.0092522655, 0.0099604752]	[0.0092784218, 0.0098464371]	10/0/0
Γ slope (desc.)	-0.00028201675	[-0.00037636623, -0.00018766728]	[-0.00036605293, -0.00021260197]	0/10/0
ζ endpoint (desc.)	-0.00085441113	[-0.001143534, -0.00056528824]	[-0.0011153467, -0.00064563334]	0/10/0

Arms are consumed from the admitted Stage A summary; the Stage B upstream raw panel and inference objects equal Stage A exactly under the registered field map. High-minus-low values and intervals are registered fields, not rounded-value recomputations.

Γ and ζ are descriptive only, not decision-driving, superiority, or equivalence analyses. This was post-outcome validation only, with no retraining and no scientific-byte modification.

Formal label: `nonlinear_chc_fixed_time_two_backbone_dose_increases_target_dsm_risk`. The conditional same-cohort fixed-$T = 1$ result establishes neither causal mediation, architecture independence, stable magnitude, nor continuum/order convergence.

G Detailed scope boundaries

Our melt operator is inspired by freezing-point depression but is not a mushy-layer simulation: it has no temperature/enthalpy field or latent-heat (Stefan) condition, no gravity-driven brine drainage, and uses periodic boundaries. We therefore use it only as a coupled conserved/non-conserved phase-field testbed. The grids are small and the domains span only a few correlation lengths, so morphology statistics carry finite-size error. The measure-preservation result is derived and verified for the linearized melted regime and for a specific operator splitting in which noise is added after the IMEX drift; the classical discrete-FDT identity is broader, but other stochastic placements have different effective covariances. In nonlinear double-well CHC, refinement changes the spectral tail; the measured correlation length is unchanged and displayed samples remain similar, so we do not claim a large nonlinear generation effect. Finally, the high-dimensional inverse suite remains synthetic and uses approximate DPS and one learned-prior PnP-ULA family tested at three smoothing scales and one fixed calibrated step size. The deterministic NIST/PFHub control reproduces a public, method-matched Benchmark 1a trajectory, but full time and space refinements materially change the transient and, for space refinement, the late morphology. It does not validate the coupled salt–melt operator, the stochastic nonlinear invariant measure, or physical observations. It covers one coupling/time-step mismatch, one OOD coarsening shift, and one sparse-blur operator, but not real observations, broad families of model error, or broad practical posterior-sampler families. The exact-posterior control is two-dimensional and diagnostic; it removes learned-score error but does not validate the salt–melt posterior. We test one specialized and one operator-aware conditional architecture, not a broad amortized-inference family. The latter sees both operator families and exact metadata yet still loses on the held-out sparse-blur configuration; additional physical systems, operator families, and architectures are needed for a stronger downstream claim. The zero-shot OPV inverse audit covers only two parent tomograms and one sparse-blur operator. Its non-overlapping blocks improve spatial accounting but do not create additional experimental replicates; it establishes a case-study domain gap, not population-level failure. For the underdamped control, the coarse ancestral sampler under-disperses momentum even with the quadrature-exact score, so its sample metrics are retained only as diagnostics; the reported claims use score error and fixed-time posterior-mean denoising instead.

Method	consist.↓ (noise 0.01)	kurt (data −1.5)	clTru↓ (mean)
Diffusion prior	**0.012**	**−1.39**	**0.024**
Tikhonov (point)	0.041	−1.42	0.121
PnP-ULA (Gaussian prior)	0.068	−0.33	0.344

Table 44: 2D coupled salt→melt inverse under the two-field testbed. A learned diffusion prior recovers the non-Gaussian morphology; Tikhonov blurs, while the optimal Gaussian-prior PnP-ULA cannot represent bimodality (kurt → 0). TV and double-well point baselines are evaluated separately in the 1D test in Appendix Fig. 36. Physics-defined and cosine-VP schedules are compared under matched conditions in Table 21.

H Posterior-sampling algorithm and auxiliary inverse comparison

Classical inverse methods. Tikhonov and total-variation regularized adjoints and PDE-constrained optimization are standard point estimators for such inverse problems. We use them, a hand-coded double-well prior, and the Gaussian-prior Langevin posterior sampler (Laumont et al., 2022) as baselines.

Diffusion posterior sampling. Given an observed melt ϕ_{obs} we sample the posterior $p(c \mid \phi_{\text{obs}})$ over salt fields using the *same* score as prior—no retraining, no PDE penalty. At each reverse step we form $\hat{c}_0$ as above and add a guidance term that descends the melt data-fit through the differentiable operator M,

$$c \leftarrow c_{\text{uncond}} - \zeta \nabla_c \|\phi_{\text{obs}} - M(\hat{c}_0(c))\|^2, \tag{21}$$

with a per-sample step ζ and gradient-norm clipping for stability (Alg. 1). Running K chains yields a posterior ensemble. As baselines we use adjoint point estimators minimizing $\|\phi_{\text{obs}} - M(c)\|^2$ plus a Tikhonov ($\|c\|^2$), total-variation ($\|\nabla c\|_1$), or double-well ($\sum(c^2 - a^2)^2$) regularizer, and a semi-implicit Gaussian-prior Langevin posterior (PnP-ULA) whose oracle Gaussian prior uses the data power spectrum. A separate sampler control reuses the learned cosine denoiser inside PnP-ULA rather than DPS.

Algorithm 1 Coupled inverse via diffusion posterior sampling

Require: score s_θ, melt operator M, observation ϕ_{obs}, schedule $\{\tau_i\}$, chains K
1: draw $c \sim \mathcal{N}(0, \operatorname{diag} v_\infty)$ (melted prior), K chains
2: **for** $i = 1$ to T **do**
3: $\quad \hat{n} \leftarrow s_\theta(c, \tau_i)$; form $\hat{c}_0$ (Tweedie, capped amp, high-k rolloff)
4: $\quad r \leftarrow \|\phi_{\text{obs}} - M(\hat{c}_0)\|^2$; $g \leftarrow \nabla_c r$ (clip norm)
5: $\quad c_{\text{uncond}} \leftarrow$ ancestral step from $\hat{c}_0$ to τ_{i+1}
6: $\quad c \leftarrow c_{\text{uncond}} - (\zeta_0/\sqrt{r})\, g$; subtract mean to preserve mass
7: **end for**
8: **return** K posterior salt fields

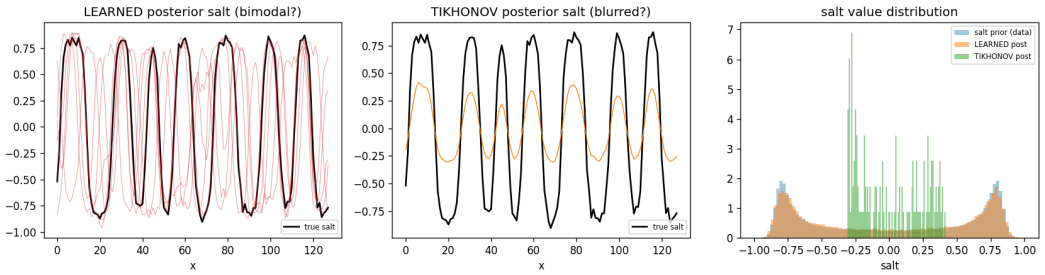

Figure 36: The coupled inverse testbed (1D). A diffusion prior fills the melt null space with bimodal brine matching the data; Tikhonov is a unimodal blur missing the wells. This is a property of the non-Gaussian *prior*, not of the physics forward (Table 21).

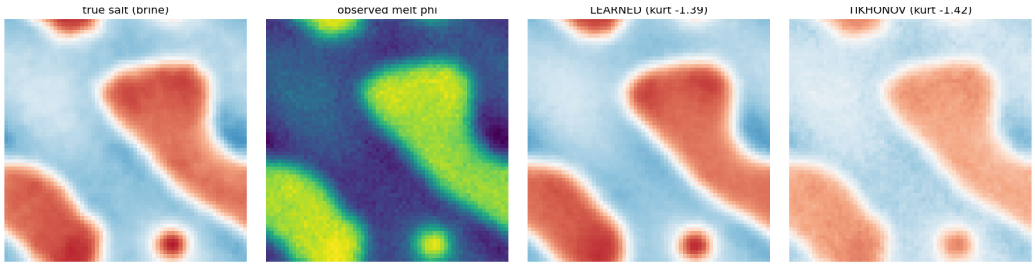

Figure 37: 2D coupled inverse under the two-field physics: the diffusion-prior recovery is sharp and consistent with the synthetic testbed; Tikhonov regresses to the mean. The same qualitative result holds for the cosine-VP prior compared in Table 21.

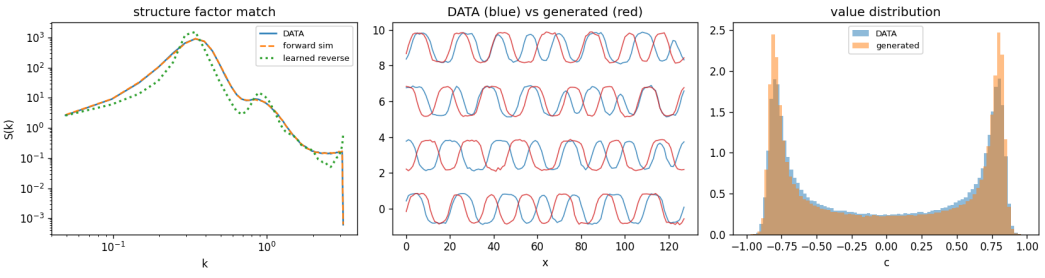

Figure 38: Physics-defined generation reproduces phase-separated morphologies and conserves mass by construction.