NUMERICALLY CONVERGED, PHYSICALLY WRONG: REFERENCE VALIDITY FOR SAMPLERS OF DIFFUSION PHYSICS EMULATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based emulators of physical systems integrate a learned reverse process for every frame of a rollout, and cheaper samplers are judged by how closely they reproduce a fine-step reference. We ask when this yardstick can be trusted. First, we make it precise: a local error budget is the largest reverse step, at each noise level, whose endpoint stays within a tolerance of the reference. Measured with a censoring-aware protocol on pretrained autoregressive conditional diffusion models, budgets predict held-out step failures, transfer to rollout states, and follow an accuracy-limited rather than CFL-like law. Second, in a frozen evaluation on held-out physical conditions with 200-step rollouts scored against ground truth, fidelity to the reference ranks 20 coarse schedules like physical accuracy on transonic cylinder flow (Spearman $\rho = 0.80$ and 0.997 for two checkpoints) but in reverse on isotropic turbulence ($\rho = -0.88$ and -0.92). The inversion appears only after about 50 frames, is absent over the upstream 100-frame evaluation horizon, and has a specific cause: the deterministic reference is numerically converged yet retains less than half of the true kinetic energy. On the same time grids, stochastic updates lower the physical error from 0.70 to 0.45, match the model's native sampler, and restore the agreement. On an accurate incompressible-wake emulator, all samplers sit within 0.014 of the reference and the choice is moot. Budget-based schedule selection, frozen before testing, ties but does not beat direct pilot evaluation at a matched calibration budget. Before fidelity to a reference guides deployment, the reference itself must be validated on physical observables over the deployment horizon.

1 INTRODUCTION

Diffusion models are increasingly used as emulators of physical systems: an autoregressive conditional diffusion model generates the next state of a flow from the previous states, and repeating this step produces a long simulation (Kohl et al., 2026; Lippe et al., 2023; Lienen et al., 2024; Price et al., 2025). Every frame requires a numerical solve of the learned reverse process, so the *sampler* (time grid, network evaluations per frame, deterministic or stochastic update) sets the cost of the emulator. To lower it, practitioners borrow schedules from image generation (Karras et al., 2022; Lu et al., 2022) and judge a cheaper sampler by how closely it reproduces an expensive fine-step reference, as one judges a coarse ODE solver by its discretization error.

This paper asks one question: *when a diffusion emulator is sampled with fewer steps, does agreement with a fine-step reference tell us whether the rollout is still physically accurate?* A useful answer says which sampler to deploy and whether that choice can be trusted over the horizon for which the emulator is used. We answer it for pretrained autoregressive conditional diffusion models (ACDMs; Kohl et al., 2026) of transonic cylinder flow (Tra), isotropic turbulence (Iso) and incompressible wake flow (Inc), without retraining any model.

Reference fidelity can be measured precisely. We define a *local error budget*: the largest reverse step at each noise level whose endpoint stays within a tolerance of the fine reference (Section 4). Measured with a censoring-aware protocol, the budgets are reproducible and predictive. Steps below the budget passed in every held-out test, including real physical conditions and states visited during

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.

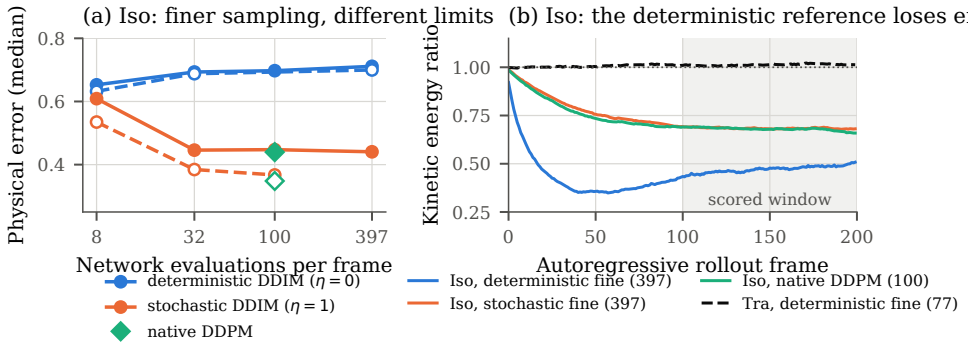

Figure 1: Converged but wrong. (a) Iso: median physical error of 200-step rollouts (frames 100–200, 14 held-out slices) against network evaluations per frame. Deterministic DDIM (blue; Karras grids at 8 and 32, uniform-t at 100, fine reference at 397) moves toward its fine reference as steps increase and becomes physically worse. Stochastic DDIM on the same grids (orange) becomes physically better and reaches the model's native DDPM sampler (green). Filled and open markers: checkpoints 00 and 01; checkpoint 01 lacks the two costliest cells. (b) Kinetic energy of the rollout relative to ground truth (median, checkpoint 00). The deterministic Iso reference loses more than half of the energy; on Tra the deterministic reference keeps the true energy.

autoregressive rollouts, and a schedule's largest budget overshoot predicts how far its whole rollout deviates from the reference. The budget law is accuracy-limited rather than CFL-like.

Reference fidelity is not physical fidelity. In a frozen evaluation on held-out physical conditions, with 200-step rollouts scored against simulation and DNS ground truth (Section 5), fidelity to the reference ranks 20 coarse schedules like physical accuracy on Tra (Spearman $\rho = 0.80$ and 0.997 for two checkpoints) and in reverse on Iso ($\rho = -0.88$ and -0.92). On Iso, raising the Karras schedule from 8 to 32 evaluations per frame brings the rollout four to five times closer to the reference and makes its physics worse (Figure 1a, blue).

The reference is converged but wrong. Refining the deterministic reference barely changes it on the development condition, yet late in the rollout it retains less than half of the true kinetic energy, with a deficit at every wavenumber (Figure 1b; Section 6). A same-grid decomposition isolates the cause. Adding the DDPM posterior noise at every step ($\eta = 1$) lowers the physical error from 0.70 to 0.45 on an identical time grid, matches the model's native ancestral sampler, and makes fidelity and physics agree again: finer stochastic sampling is physically better (Figure 1a, orange). The deterministic path fails, not the resolution of its time grid.

Budgets alone do not choose a better sampler. Budget-based schedule selection, frozen before testing, ties but does not beat direct evaluation of candidate samplers on a few pilot trajectories at a matched calibration budget (Section 7). Most of its cost is the fine reference itself, and its choices coincide with what the two schedules it rolls out for calibration (its "anchors") already imply.

Contributions.

- Censoring-aware local error budgets for trained diffusion emulators, validated on held-out conditions and rollout states; a falsified CFL law and an accuracy-limited one.
- A pre-registered physical validation: reference fidelity predicts physical accuracy on Tra, inverts it on Iso (invisibly over short horizons), and is moot on Inc, where every sampler matches an accurate reference.
- A mechanism: the deterministic reference drifts in energy while numerically converged; stochastic sampling on the same grid removes the drift.
- A decision-level evaluation and a rule: validate the reference on physical observables over the deployment horizon before trusting a fidelity-based audit.

All claims concern inference with pretrained checkpoints of one architecture family. Analyses fixed before the main predictions are reported separately from diagnostics specified after the frozen outcomes were observed.

2 Related Work

Diffusion samplers. Sampling a diffusion model integrates a reverse-time SDE or its probability-flow ODE (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021b; Anderson, 1982); DDIM (Song et al., 2021a) interpolates between the deterministic flow ($\eta = 0$) and the DDPM ancestral update ($\eta = 1$). EDM (Karras et al., 2022), DPM-Solver (Lu et al., 2022), UniPC (Zhao et al., 2023) and learned solvers (Frankel et al., 2025) reduce the number of evaluations. For images, stochastic steps are known to correct errors of an imperfect denoiser (Karras et al., 2022); SDE samplers contract accumulated error while ODE samplers have smaller discretization error (Xu et al., 2023); and guarantees for stochastic samplers need only an L^2 score error (Chen et al., 2023b), whereas probability-flow analyses rely on corrector steps (Chen et al., 2023a). This distinction decides whether a fine deterministic reference is a valid yardstick.

Diffusion emulators of physical systems. ACDMs emulate turbulent and transonic flows (Kohl et al., 2026); PDE-Refiner (Lippe et al., 2023), generative 3D turbulence (Lienen et al., 2024) and GenCast (Price et al., 2025) extend diffusion to PDE rollouts and weather. Lazy Diffusion (Sambamurthy & Chattopadhyay, 2025) and bias-constrained schedules (Le Cleï et al., 2026) design noise schedules for PDE emulation, scale-dependent schedules exist for images (Ziashahabi et al., 2024), and long autoregressive rollouts are prone to drift (McCabe et al., 2023). We keep the trained models fixed and ask how the choice of inference sampler should be validated.

Evaluating scientific surrogates. Physical surrogates must preserve physical diagnostics rather than appearance (Brunton et al., 2020; Pope, 2000), and weak baselines and reporting choices can overstate progress (Takamoto et al., 2022; McGreivy & Hakim, 2024). Our evaluation fixes splits, candidates, metrics, tolerances and decision rules before the main predictions, compares against direct evaluation at a matched budget, and reports abstentions and failures.

Step-size control. Classical solvers choose steps from local error estimates (Hairer & Wanner, 1996; Kloeden & Platen, 1992), CFL conditions tie steps to resolved scales (Courant et al., 1928; LeVeque, 2002); the reverse process passes through distinct regimes (Raya & Ambrogioni, 2023; Biroli et al., 2024; Sclocchi et al., 2025; Li & Chen, 2024), which error-aware samplers exploit (Kong et al., 2026).

3 Setting

Emulators. An ACDM (Kohl et al., 2026) is a denoising diffusion probabilistic model (Ho et al., 2020) of the next physical state x_{n+1} given the two previous states. A U-Net receives the noised target with the conditioning frames, noised to the same level; a rollout samples x_{n+1}, appends it to the conditioning and repeats, with prescribed parameters (Mach number for Tra, Reynolds number for Inc) as a constant channel reset after every frame. We use the released checkpoints, two independently trained seeds per data set (00 and 01), at 128×64; Iso has $T = 100$ diffusion steps, Tra and Inc $T = 20$. No model is retrained.

Data. Tra: transonic cylinder flows at Mach 0.5–0.9 (41 simulations). Iso: two-dimensional slices of a three-dimensional isotropic-turbulence DNS from the Johns Hopkins Turbulence Database (Perlman et al., 2007; Li et al., 2008), correlated samples of one realization. Inc: incompressible wake flows at Reynolds numbers 100–1000 (91 simulations). Calibration, selection and test conditions are disjoint, and every Tra and Inc condition excluded by the upstream training script is reserved for testing (Table 4). For Inc we report the calibration side of the protocol (the eight selection conditions; Appendix G), not held-out test conditions.

Samplers. The models use a variance-preserving forward process $x_t = \alpha(t)x_0 + \sigma(t)\epsilon$ with $\alpha = \sqrt{\bar{\alpha}}$ and $\sigma = \sqrt{1 - \bar{\alpha}}$; we report noise levels as $\sigma_{\text{eff}} = \sigma/\alpha$ and evaluate fractional t by interpolating

$\log \bar{\alpha}$. One DDIM(η) step (Song et al., 2021a) from t_a to $t_b < t_a$ computes $\hat{x}_0 = (x_a - \sigma_a \hat{\epsilon})/\alpha_a$ and

$$x_b = \alpha_b \hat{x}_0 + \sqrt{\sigma_b^2 - s^2}\,\hat{\epsilon} + s\,z, \qquad s = \eta\,\frac{\sigma_b}{\sigma_a}\sqrt{1 - \frac{\alpha_a^2}{\alpha_b^2}}, \qquad z \sim \mathcal{N}(0, I). \tag{1}$$

With $\eta = 0$ the update is deterministic and discretizes the probability-flow ODE; with $\eta = 1$ on the integer grid it is the DDPM ancestral update with posterior variance $\tilde{\beta}_t$, which we call the native DDPM sampler; the last step to the clean state is deterministic. A sampler is a time grid plus η: grids uniform in t, in σ_{eff} or in $\log \sigma_{\text{eff}}$ ("log") and Karras grids (exponent 7; Karras et al., 2022) at 4, 8, 16 and 32 network evaluations per frame (NFE), plus four "response" grids whose steps are equal multiples of the local budget of Section 4. The *fine reference* is deterministic DDIM with spacing $\Delta t = 0.25$ in the model's time index (397 evaluations per frame for Iso, 77 for Tra and Inc). All deterministic candidates of a trajectory share its initial and conditioning noise.

Two kinds of fidelity. Let S_c be the mean radial power spectrum of the (fluid-masked) velocity field of a rollout produced by sampler c, averaged over frames 100–200 of a 200-frame rollout, and let K_c be its mean kinetic energy over the same frames. With S_*, K_* for the ground truth and S_f, K_f for the fine reference,

$$e_{\text{phys}}(c) = \max\left(\frac{\|S_c - S_*\|_2}{\|S_*\|_2}, \frac{|K_c - K_*|}{|K_*|}\right), \qquad e_{\text{ref}}(c) = \max\left(\frac{\|S_c - S_f\|_2}{\|S_f\|_2}, \frac{|K_c - K_f|}{|K_f|}\right). \tag{2}$$

e_{phys} is the *physical error* and e_{ref} the *reference deviation*. A rollout fails at tolerance $\tau \in \{0.05, 0.1, 0.2, 0.4\}$ if $e_{\text{phys}} > \tau$ or its output is not finite (a *hard failure*); 40% is an operational threshold, not engineering accuracy. State errors, temporal spectra and autocorrelation are reported but enter no decision.

Protocol. Splits, candidates, observables, tolerances and the selection rules of Section 7 were fixed before any main prediction, and selection policies were frozen and hash-bound before any test rollout (Appendix B). Each test condition is rolled out from two true frames for 200 frames with two checkpoints, two noise seeds and, for Tra and Iso, two start windows; these repetitions share source conditions and are not independent, and Iso slices share one DNS realization; analyses specified after the frozen outcomes are labelled post hoc.

4 Local Error Budgets Make Reference Fidelity Measurable

Definition. For a state x at noise level $\sigma = \sigma_{\text{eff}}(t)$, let $U_h(x)$ be one deterministic DDIM step to $\sigma - h$ and $R_h(x)$ the fine reference path over the same interval. For an endpoint error $e(h; x)$ between $U_h(x)$ and $R_h(x)$ and a tolerance τ, the local error budget is

$$\Delta\sigma_{\max}(\sigma; \tau) = \sup\{h :\ e(h'; x) \le \tau \text{ for all measured } h' \le h\}, \tag{3}$$

the largest contiguous passing trial step from states reached by integrating the reverse process; right-censoring (the largest trial passes, so $\Delta\sigma_{\max}$ is a lower bound) is recorded and no parametric law is fitted. A budget curve aggregates states per noise bin (median, or 10th percentile for a conservative curve), and a time grid $\sigma_0 > \cdots > \sigma_N$ is scored by its largest overshoot $R = \max_i(\sigma_i - \sigma_{i+1})/\Delta\sigma_{\max}(\sigma_i)$. Two endpoint errors are used: the radial-spectrum error of the first velocity channel on Iso (original protocol, Appendix C) and the relative ℓ_2 error of the normalized state ($\tau = 0.05$) on eight real conditions per data set.

Budgets predict local failures on held-out states. With spectrum budgets (median over conditions) calibrated on two Iso checkpoints, steps below the budget passed on all fresh initial conditions, steps at the budget passed at 90.6–97.9%, and steps at $1.2\times$ the budget at only 1.8–20.1% (Table 5). The state budgets of the physical campaign are conservative by construction (10th percentile, trial steps that double) and transfer to real conditioning and to rollout states (Table 9, frames 0, 50 and 150 pooled; Figure 4, frame 150): steps at $0.5\times$ and $0.8\times$ the budget passed in every case for true, fine-rollout and Karras-8-rollout conditioning, steps at the budget passed at 95.6–100%, and at $1.5\times$ the pass rate fell to 0.3–1.3% on Tra. On Iso it was 62.5% for every conditioning state because at five noise levels the recorded budget (0.32σ) sits one trial step below the true limit (between 0.48σ

and 0.64σ); right-censoring (32% of Iso and 52–53% of Tra budgets; Table 8) affects only the lowest levels.

The budget law is accuracy-limited, not CFL-like. Across three tolerances and two Iso checkpoints (2737 measurements, no monotonicity violation) a two-regime law beats a single exponent by $\Delta\mathrm{AIC} \geq 47$, with exponents 0.656 ± 0.050 and 1.046 ± 0.046 around a knot at $\sigma_{\mathrm{eff}} \approx 0.86$ where the denoiser saturates. A strict CFL rule $\Delta\sigma_{\max} \propto \lambda_{\max}^{-1}$ fails ($\lambda_{\max}\Delta\sigma_{\max}$ varies 2.5–3.7×, $\lambda_{\max}\Delta\sigma_{\max}^2$ only 1.25–1.42×; local defects have order $p \approx 2.4$): the limit is endpoint accuracy (Appendix D).

Budgets predict the reference deviation of whole rollouts. Over the eight log and Karras schedules of the physical campaign, the overshoot R ranks schedules by the median reference deviation e_{ref} of their 200-step test rollouts with Spearman $\rho = 0.976$ on Tra (both checkpoints) and 0.81 and 0.905 on Iso; on a stress set of 120 schedule groups, $R > 1$ flagged every group whose median spectral endpoint failed, with 10 conservative false positives (Table 6).

Everything in this section measures fidelity to the fine reference (Equation 3), and on that question the budgets work. On Tra, R also ranks the eight schedules by physical error ($\rho = 0.976$). On Iso it does not ($\rho = -0.02$ and -0.24). The next section shows why this is not a failure of the budgets.

5 Does Reference Fidelity Predict Physical Accuracy?

We roll out every held-out test condition for 200 frames with the fine reference and 20 coarse samplers (16 conventional grids, four response grids) and score each rollout against ground truth (Equation 2): 3,864 Tra and 2,352 Iso test rollouts (23 simulations and 14 slices; two checkpoints, noise seeds and windows each), none non-finite or, on Tra, with non-positive density or pressure. Inc is scored on its eight selection conditions (672 rollouts; Appendix G).

On Tra, fidelity is a good proxy. No coarse schedule has a lower median physical error than the fine reference in the scored window (Table 1), although the margins are small (≤ 0.003) and several schedules tie it pairwise. The 20 coarse schedules are ranked by reference deviation closely ($\rho = 0.80$) or almost exactly (0.997) as by physical error (Figure 2a). The physical error of about 0.2 is a floor that none of the 21 deterministic samplers goes below; whether stochastic sampling would lower it on Tra was not tested.

On Iso, the ranking inverts. The more closely a schedule follows the fine reference, the worse its physics ($\rho = -0.88$ and -0.92; Figure 2b), and 19 of the 20 coarse schedules have a lower median physical error than the reference itself. Within the Karras family, increasing the evaluations per frame from 8 to 32 reduces the reference deviation from 0.35 to 0.07 (checkpoint 00) and from 0.23 to 0.06 (checkpoint 01), while the physical error rises from 0.65 to 0.69 and from 0.63 to 0.69. The fine reference itself has physical error 0.71 and 0.70.

The inversion is invisible over short horizons. Scoring the same rollouts on other frame windows (post hoc; Figure 2c, Table 15) shows that on Iso fidelity and physics agree almost perfectly over frames 0–20 ($\rho = 0.994$ for both checkpoints; no coarse schedule beats the reference), weaken over frames 20–60 (0.29 and 0.52), and invert over frames 50–100 (-0.93 and -0.38) and 100–200 (-0.88 and -0.92). Over the upstream 100-frame evaluation horizon the correlation is still positive (0.38 and 0.62), so the upstream evaluation would have certified the Iso reference. On Tra ρ stays positive in every window (0.53–1.00).

On Inc, the choice hardly matters. The Inc reference is accurate (median physical error 0.021 and 0.010, energy ratio 1.00; Table 22), every coarse schedule lands within 0.014 of it (7 and 11 of 20 below it, by at most 0.006), and the fidelity–physics rank correlation is near zero (0.25 and -0.26) because sampler deviations (0.0002–0.004 for the Karras grids) are well below the emulator's own error: the reference passes any validity check and the cheapest schedule serves as well.

The budgets themselves stay valid along the rollout (Section 4); on Iso what changes is the relation between the reference and the physics, examined next.

Table 1: Held-out physical validation of the fine deterministic reference and 20 coarse schedules (200-step rollouts, frozen protocol). Physical error is $\max$(relative ℓ_2 error of the late-window mean velocity spectrum, relative error of late-window mean kinetic energy) against ground truth over frames 100–200. ρ is the Spearman rank correlation, over the 20 coarse schedules, between the median deviation from the fine reference and the median physical error, computed on frames 0–20 and 100–200. E ratio is the fine reference's late-window kinetic energy relative to truth (median). Rollouts per candidate combine source conditions, windows and two noise seeds; they are not independent trajectories.

Data	Ckpt	Sources	Rollouts	Fine reference error		Best coarse schedule	ρ(fidelity, physics)		E ratio
				median	q90	median	0–20	100–200	fine
Tra	00	23	92	0.214	0.410	0.216 (unif.-σ-16)	+0.995	+0.802	1.02
	01	23	92	0.203	0.306	0.205 (Karras-32)	+1.000	+0.997	1.04
Iso	00	14	56	0.711	0.828	0.532 (unif.-σ-16)	+0.994	−0.875	0.47
	01	14	56	0.700	0.734	0.503 (unif.-σ-32)	+0.994	−0.925	0.44

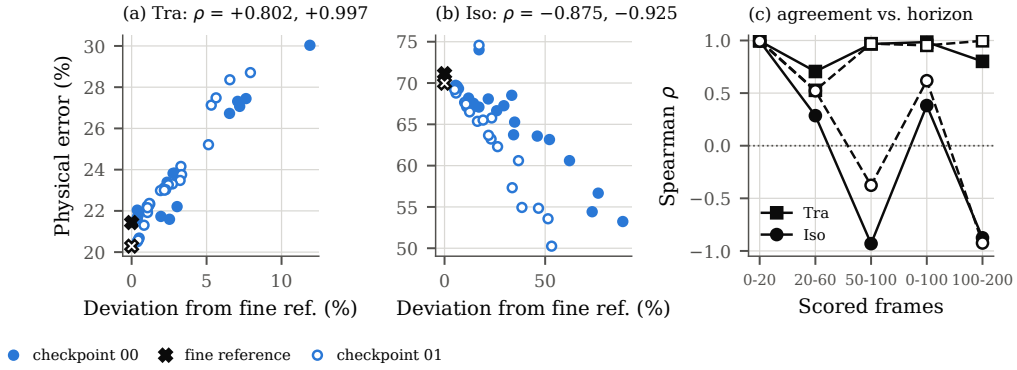

Figure 2: Reference fidelity versus physical accuracy. (a, b) Each point is one of 20 coarse deterministic schedules: median deviation from the fine reference against median physical error (frames 100–200, in percent); crosses mark the fine reference. Filled and open markers: checkpoints 00 and 01; ρ is the Spearman rank correlation per checkpoint. (c) The same rank correlation on different frame windows of the same rollouts (post hoc). Iso agrees over the first 20 frames, is still positive over the upstream 100-frame horizon, and inverts after about frame 50.

6 Why the Reference Fails: Converged but Wrong

The reference is numerically converged. On the development condition of each data set, refining the deterministic reference from $\Delta t = 0.25$ to 0.125 changes its physical error by at most 0.001 on Tra and 0.0005 on Iso (Table 17), while the Iso physical error itself is 0.74–0.81.

The converged path loses energy. Over frames 100–200 the deterministic Iso reference retains a median 47% and 44% of the true kinetic energy (checkpoints 00 and 01); the Tra reference retains 102% and 104% (Table 1). The Iso deficit covers every wavenumber (Figure 3a), a loss of amplitude rather than a redistribution between scales, and it starts at the first frame: the single-step spectral error from true conditioning is 0.057 and 0.046 for the reference, against 0.011 and 0.005 for the native DDPM sampler, and grows through the rollout (Figure 3b).

Coarse schedules win by cancellation. With $u = (S_f - S_*)/\|S_*\|$ the reference's spectral error and $v = (S_c - S_f)/\|S_*\|$ a coarse schedule's deviation from it, the coarse spectral error is exactly $\|u\|^2 + \|v\|^2 + 2\langle u, v\rangle$. For Iso Karras-8 the cross term is negative in 56 of 56 and 54 of 56 rollouts, with median $\cos(u, v) = -0.999$ and -1.000 (Table 16): the coarse deviation points almost exactly

against the reference's error, so being far from a wrong reference is what makes these schedules better; on Tra the cosines are mixed and the cross terms small.

Stochasticity, not refinement of the deterministic grid, decides. The native DDPM sampler of the Iso model uses 100 ancestral steps. Compared on the same test slices, it has physical error 0.44 and 0.35, well below the reference's 0.71 and 0.70. That comparison changes stochasticity and grid at once, so we specified a same-grid decomposition (after the frozen outcomes, before its own runs) with DDIM(η) steps (Equation 1) sharing the initial and conditioning noise of the frozen rollouts (Table 2). On an identical 100-step grid, switching from $\eta = 0$ to $\eta = 1$ lowers the median physical error from 0.698 to 0.447 and improves all 28 paired trajectories. The $\eta = 0$ result is within 0.013 of the 397-step reference (0.711): refining the deterministic grid changes almost nothing. With $\eta = 1$ the error decreases with the number of evaluations and saturates at the native sampler's level (0.61, 0.45, 0.45 and 0.44 at 8, 32, 100 and 397 evaluations on checkpoint 00) and matches the native DDPM sampler (0.44 and 0.35). For 32 or more evaluations the late-window kinetic energy ratio rises from 0.44–0.49 to 0.62–0.68; at 8 evaluations the stochastic rollout improves the spectrum while its energy ratio falls from 0.56 to 0.48.

With a valid reference, fidelity and physics agree again. Under $\eta = 1$, a closer approach to the (stochastic) fine sampler is also physically better. The $\eta = 1$ Karras-32 and uniform-100 rollouts deviate from the $\eta = 1$ fine rollouts by only 1.9% and 1.4% in their late spectra and 8.5% and 4.0% in energy, while the $\eta = 0$ schedules are 39–52% and 16–27% away (Table 18). The local error budgets and the fidelity yardstick behave as intended once the path they measure against is physically right.

Interpretation. With an exact denoiser the probability-flow ODE and the reverse SDE share their marginals (Song et al., 2021b); with a learned denoiser, deterministic integration carries the score error forward while stochastic updates partly contract it (Karras et al., 2022; Xu et al., 2023), and autoregressive conditioning can compound a small per-frame amplitude loss. Our experiments isolate determinism as the controlling factor for this model but do not prove the compounding mechanism. Even stochastic sampling leaves an Iso error of 0.35–0.45 and no sampler meets the 40% tolerance at the 90th percentile; we attribute this to the emulator itself on this underdetermined data set.

7 What Budgets Can and Cannot Decide

Frozen selection rules. For each data set, checkpoint, calibration size $n \in \{2, 4, 8\}$, aggregation (90th percentile or maximum) and tolerance, two policies were frozen before any test rollout over eight log and Karras schedules with 4–32 evaluations (Appendix F). The *budget audit* measures budgets on n calibration conditions, rolls out the fine reference and two *anchor* schedules (log-8, Karras-16) on $\min(n, 4)$ selection conditions on Iso and n elsewhere, predicts each candidate's physical error by composing the reference's error with the deviation of the nearest anchor of at least equal overshoot, and selects the cheapest candidate predicted within tolerance. The *direct pilot* spends the same sample evaluations rolling out candidates, cheapest first, on the same conditions and selects the cheapest one observed within tolerance. Either may abstain.

Neither rule dominates. On Iso every policy abstains, for both rules and all 48 settings of each (Table 3). The audit abstains because the fine reference's own physical error on the selection conditions (0.71–0.77) exceeds every tolerance; the direct pilot abstains because no queried candidate passes, consistent with the test data, where no candidate meets any tolerance at the 90th percentile. The pilot reaches its abstention at 26% of the audit's cost.

Tra. At the main setting ($n = 8$, $\tau = 0.4$) the audit selects log-8 for both checkpoints (test failure rates 10.9% and 6.5%) and the direct pilot selects Karras-4 and log-4 (9.8% and 14.1%); paired source-level differences are $+1.1$ and -7.6 percentage points with resampling intervals $[-3.3, 6.5]$ and $[-18.5, 0]$ (Table 19, Figure 7). Over 1,000 deployment rollouts the audit's choice costs more (2.28 against 1.36 million sample evaluations for checkpoint 00) because it selects an 8-evaluation schedule; selected schedules behave alike on model-held-out and trained sources (Table 21).

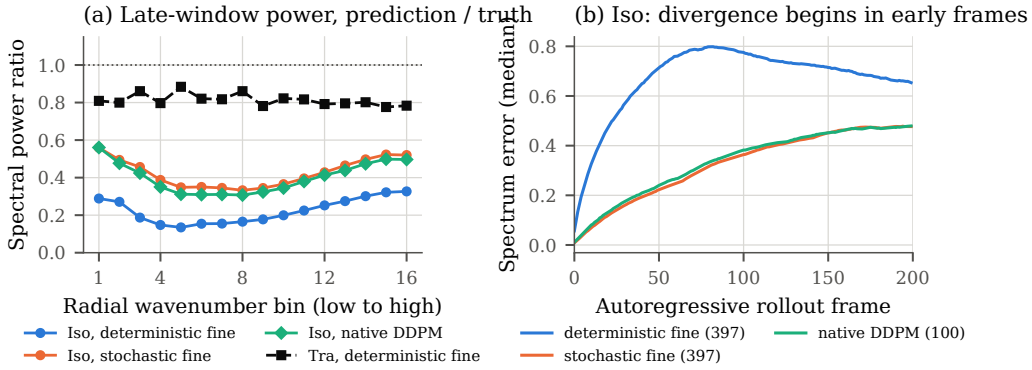

Figure 3: How the deterministic reference fails (checkpoint 00). (a) Late-window spectral power of the rollout relative to truth in 16 radial wavenumber bins: the deterministic Iso reference is deficient at every scale; stochastic sampling recovers part of the deficit. The Tra reference is about 20% low in fluctuation power at every scale (the 0.2 error floor of Section 5) while its total kinetic energy, dominated by the mean flow, is preserved. (b) Iso: median per-frame spectral error. The deterministic reference departs from the truth from the first frames, several times faster than stochastic samplers.

Table 2: Iso: stochasticity, not step count, controls the reference's physical error. Median physical error (frames 100–200) of DDIM(η) rollouts on the 14 Iso test slices; $\eta = 0$ is the deterministic update, $\eta = 1$ adds the DDPM posterior noise at every step on the same time grid. The native DDPM row is the unmodified ancestral sampler of the pretrained model. Same-grid pairing (uniform-t, 100 evaluations, checkpoint 00): $\eta = 1$ improves 28/28 paired trajectories, median change -0.262. "—": cell not run; the protocol limits the two costliest cells (uniform-t at 100 with $\eta = 0$, fine with $\eta = 1$) to checkpoint 00 and one noise seed (28 rollouts); all other cells use 56 rollouts. E ratio: late-window kinetic energy relative to truth.

Time grid, evaluations/frame	$\eta = 0$		$\eta = 1$		E ratio (ckpt 00)	
	ckpt 00	ckpt 01	ckpt 00	ckpt 01	η=0	η=1
Karras, 8	0.653	0.632	0.609	0.535	0.56	0.48
Karras, 32	0.693	0.688	0.446	0.384	0.49	0.63
uniform-t, 100	0.698	—	0.447	0.367	0.48	0.66
fine Δt=0.25, 397	0.711	0.700	0.441	—	0.47	0.68
native DDPM, 100			0.440	0.348		0.68

Where the audit helps. With sparse calibration ($n = 2$, $\tau = 0.2$) the audit abstains on both Tra checkpoints while the direct pilot accepts log-8 and Karras-4, which fail on 63.0% and 65.2% of test rollouts; with four or more conditions the pilot abstains too, and a rule that always abstains would also have avoided these failures.

The budget curve rarely changes the decision. Among the 14 Tra policies in which any rule selected a schedule, the audit's choice equals the one implied by its two anchor rollouts alone in 13, and nine of its ten selections are one of the two anchors. The fine reference accounts for 72.6% (Tra) and 85.9% (Iso) of the audit's calibration cost, the local budgets for 4.8% and 8.9%. With ground-truth pilots available, direct evaluation is a strong and cheaper baseline for choosing a schedule; the budgets' value is to measure where a grid departs from its reference and predict unsafe steps, relative to a reference whose validity must be established first.

A reference-validity check. Our results suggest one rule for any use of a fine reference as a yardstick, including reporting deviation from a many-step solution: compare the reference with ground truth on physical observables (kinetic energy, spectra) on a few held-out trajectories over the horizon of use. Here the check passes Tra (energy ratio 1.02–1.04; no cheaper sampler has a lower median error) and fails Iso (0.47 and 0.44; 19 of 20 cheaper samplers beat the reference), and a

Table 3: Frozen budget-selection policies on held-out test conditions. Each policy was frozen from calibration and selection data before any test rollout; parentheses give the test failure rate of the chosen schedule (fraction of test rollouts whose physical error exceeds τ). "Anchors only" applies the audit's composition rule to the two measured anchor schedules without the local budget curve. Costs are calibration sample-network evaluations in thousands; the direct pilot is capped at the audit's budget. Share of the audit's calibration cost spent on the fine reference (n=8): Tra 72.6%, Iso 85.9%.

Data	Ckpt	Budget audit	Anchors only	Direct pilot	Cost (k NFE)	
					audit	pilot
main: n=8, τ=0.4, 90th-percentile rule						
Tra	00	log-8 (10.9%)	log-8 (10.9%)	Karras-4 (9.8%)	679	563
Tra	01	log-8 (6.5%)	log-8 (6.5%)	log-4 (14.1%)	679	563
Iso	00	abstain	abstain	abstain	1478	384
Iso	01	abstain	abstain	abstain	1478	384
sparse: n=2, τ=0.2, 90th-percentile rule						
Tra	00	abstain	abstain	log-8 (63.0%)	170	141
Tra	01	abstain	abstain	Karras-4 (65.2%)	170	141
Iso	00	abstain	abstain	abstain	706	192
Iso	01	abstain	abstain	abstain	706	192

stochastic reference restored the yardstick where it failed (Section 6). The frozen audit paid for this comparison but used the reference's error only as an additive term, never as a validity check. The rule is drawn from our results, not validated prospectively.

8 Discussion and Limitations

What generalizes. In principle any fidelity yardstick built on a deterministic reference inherits the reference's physical bias, and autoregressive conditioning can amplify a small per-frame bias into a large drift; we observed this on one data set. The diagnostic that exposed it, a same-grid comparison of deterministic and stochastic updates scored on physical observables over a long horizon, needs no retraining and cost a few hundred rollouts (three GPU-hours) here.

Limitations. All models are ACDMs from one code base, two checkpoints per data set; the reference failure is observed on Iso and not on Tra or Inc, and the Inc evidence is from the protocol's selection conditions rather than held-out ones. Other denoisers, parameterizations or learned solvers may behave differently. Iso is one DNS realization with correlated slices. The scored observables are late-window spectra and energy; temporal power spectra stay far from the truth even for the Tra reference (median relative error about 0.9), so a passed tolerance does not certify the dynamics. Budgets are defined for deterministic steps; against a stochastic reference they would need distributional endpoints. The horizon, decomposition, stochasticity and anchors-only analyses are post hoc (the stochasticity protocol was fixed before its own runs), and the reference-validity check is a recommendation, not a validated procedure.

9 Conclusion

Local error budgets make reference fidelity a precise, predictive yardstick for diffusion-emulator samplers. Whether it is the right yardstick depends on the reference: on transonic cylinder flow the deterministic reference is as accurate as any sampler tested and fidelity tracks physics; on isotropic turbulence the same kind of reference is numerically converged, loses half of the kinetic energy and turns fidelity into an inverse proxy until stochastic updates on the same grid restore it. Numerical convergence is not physical validity, which must be checked on physical observables over the horizon of use.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting sections of this paper and to improve the clarity, organization, and readability of the writing. We did not use generative AI tools to run the experiments. We reviewed and revised all AI-assisted text and checked technical claims, citations, and descriptions of experiments and results against the underlying work and original sources. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experiments are inference-only on the publicly released ACDM checkpoints and data (Kohl et al., 2026); the models, data sets and samplers are specified in Section 3 and Appendix A. The frozen protocol (splits, candidate samplers, observables, tolerances and the two selection rules) is given in Appendix B, the local-budget measurement in Appendix C, the stochasticity decomposition in Appendix E, the Inc setup in Appendix G and the compute in Appendix I. The supplementary material contains the code: the measurement, rollout and analysis scripts of both protocols, the protocol files (splits, candidate samplers, observables, tolerances and selection rules), the frozen selection manifests with input hashes, and the scripts that generate every figure and table of this paper from the result tables.

REFERENCES

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. doi: 10.1016/0304-4149(82)90051-5.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024. doi: 10.1038/s41467-024-54281-3.

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

Steven L. Brunton, Bernd R. Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. *Annual Review of Fluid Mechanics*, 52:477–508, 2020. doi: 10.1146/annurev-fluid-010719-060214.

Chubin Chen, Jiashu Zhu, Xiaokun Feng, Nisha Huang, Chen Zhu, Meiqi Wu, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Stochastic self-guidance for training-free enhancement of diffusion models. In *International Conference on Learning Representations*, 2026.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ODE is provably fast. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68552–68575, 2023a. doi: 10.52202/075280-2998.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023b.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Richard Courant, Kurt Friedrichs, and Hans Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Mathematische Annalen*, 100:32–74, 1928.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16222–16239, 2023.

Eric Frankel, Sitan Chen, Jerry Li, Pang Wei Koh, Lillian J. Ratliff, and Sewoong Oh. S4S: Solving for a fast diffusion model solver. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17542–17574, 2025.

Uriel Frisch. *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press, Cambridge, UK, 1995.

Jayesh K. Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized PDE modeling. *Transactions on Machine Learning Research*, 2023.

Ernst Hairer and Gerhard Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer-Verlag, Berlin, Heidelberg, 2 edition, 1996.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Philipp Holl, Vladlen Koltun, and Nils Thuerey. Learning to control PDEs with differentiable physics. In *International Conference on Learning Representations*, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*, volume 23 of *Applications of Mathematics*. Springer-Verlag, Berlin, Heidelberg, 1992.

Dmitrii Kochkov, Jamie A. Smith, Ayya Alieva, Qing Wang, Michael P. Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21):e2101784118, 2021. doi: 10.1073/pnas.2101784118.

Georg Kohl, Li-Wei Chen, and Nils Thuerey. Benchmarking autoregressive conditional diffusion models for turbulent flow simulation. *Neural Networks*, 199:108641, 2026. doi: 10.1016/j.neunet.2026.108641.

A. N. Kolmogorov. The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. *Doklady Akademii Nauk SSSR*, 30(4):299–303, 1941. English translation: Proceedings of the Royal Society of London A 434, 9–13 (1991), doi:10.1098/rspa.1991.0075.

Inho Kong, Sojin Lee, Youngjoon Hong, and Hyunwoo J. Kim. Error as signal: Stiffness-aware diffusion sampling via embedded Runge–Kutta guidance. In *International Conference on Learning Representations*, 2026.

Constantin Le Cleï, Nils Thuerey, and Xiaoxiang Zhu. Bias-constrained diffusion schedules for PDE emulations: Reconstruction error minimization and efficient unrolled training. *arXiv preprint arXiv:2604.08357*, 2026.

Randall J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, Cambridge, UK, 2002.

Marvin Li and Sitan Chen. Critical windows: Non-asymptotic theory for feature emergence in diffusion models. In *International Conference on Machine Learning*, 2024.

Yi Li, Eric Perlman, Minping Wan, Yunke Yang, Charles Meneveau, Randal Burns, Shiyi Chen, Alexander Szalay, and Gregory Eyink. A public turbulence database cluster and applications to study Lagrangian evolution of velocity increments in turbulence. *Journal of Turbulence*, 9:N31, 2008. doi: 10.1080/14685240802376389.

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

Marten Lienen, David Lüdke, Jan Hansen-Palmus, and Stephan Günnemann. From zero to turbulence: Generative modeling for 3D flow simulation. In *International Conference on Learning Representations*, 2024.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, 2023.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, 2022.

Michael McCabe, Peter Harrington, Shashank Subramanian, and Jed Brown. Towards stability of autoregressive neural operators. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=RFfUUtKYOG.

Nick McGreivy and Ammar Hakim. Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations. *Nature Machine Intelligence*, 6 (10):1256–1269, 2024. doi: 10.1038/s42256-024-00897-5.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171, 2021.

Eric Perlman, Randal Burns, Yi Li, and Charles Meneveau. Data exploration of turbulence simulations using a database cluster. In *Proceedings of the 2007 ACM/IEEE Conference on Supercomputing (SC '07)*, pp. 1–11, 2007. doi: 10.1145/1362622.1362654.

Stephen B. Pope. *Turbulent Flows*. Cambridge University Press, Cambridge, UK, 2000.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044): 84–90, 2025. doi: 10.1038/s41586-024-08252-9.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

François Rozet and Gilles Louppe. Score-based data assimilation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 40521–40541, 2023. doi: 10.52202/075280-1763.

Anish Sambamurthy and Ashesh Chattopadhyay. Lazy diffusion: Mitigating spectral collapse in generative diffusion-based stable autoregressive emulation of turbulent flows. *arXiv preprint arXiv:2512.09572*, 2025.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8459–8468, 2020.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1): e2408799121, 2025. doi: 10.1073/pnas.2408799121.

Dule Shu, Zijie Li, and Amir Barati Farimani. A physics-informed diffusion model for high-fidelity flow field reconstruction. *Journal of Computational Physics*, 478:111972, 2023. doi: 10.1016/j.jcp.2023.111972.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2256–2265, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Kimberly Stachenfeld, Drummond B. Fielding, Dmitrii Kochkov, Miles Cranmer, Tobias Pfaff, Jonathan Godwin, Can Cui, Shirley Ho, Peter Battaglia, and Alvaro Sanchez-Gonzalez. Learned simulators for turbulence. In *International Conference on Learning Representations*, 2022.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1596–1611, 2022. doi: 10.52202/068431-0117. Datasets and Benchmarks Track.

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

Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi Jaakkola. Restart sampling for improving generative processes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76806–76838, 2023. doi: 10.52202/075280-3356.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):105:1–105:39, 2024. doi: 10.1145/3626235.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Amir Ziashahabi, Baturalp Buyukates, Artan Sheshmani, Yi-Zhuang You, and Salman Avestimehr. Frequency domain diffusion model with scale-dependent noise schedule. In *IEEE International Symposium on Information Theory (ISIT)*, 2024. doi: 10.1109/ISIT57864.2024.10619452.

A Models and Samplers

Models. Each ACDM (Kohl et al., 2026) is a conditional DDPM with a U-Net denoiser whose input stacks two conditioning frames and the noised target: 15 channels for Tra (velocity, density, pressure, Mach number), 12 for Inc (velocity, pressure, Reynolds number) and 12 for Iso (three velocity components, pressure). Conditioning frames c are noised to the current diffusion level during training and sampling. In our DDIM runs they enter as $\alpha(t)c+\sigma(t)\xi$ with one fixed Gaussian tensor ξ per frame, so that a candidate sampler and the reference see identical conditioning. Fields are normalized with the released statistics and denormalized before scoring.

DDIM(η) and the DDPM posterior. In Equation 1, on the integer grid $t_a = t$, $t_b = t - 1$, we have $\alpha_a^2/\alpha_b^2 = \bar{\alpha}_t/\bar{\alpha}_{t-1} = 1 - \beta_t$, so for $\eta = 1$

$$s^2 = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\,\beta_t = \tilde{\beta}_t,$$

the posterior variance used by the released ancestral sampler (the native DDPM sampler). The runner asserts this identity numerically before any rollout (at $t = 50$: 0.0497767 in both expressions) and asserts that $\eta = 0$ reproduces the frozen deterministic rollouts exactly (maximum absolute deviation 0.0).

Time grids. A grid with N evaluations visits N noise levels from $t = T - 1$ down to $t = 0$ and ends with a deterministic jump to the clean state. Uniform-t spaces the levels linearly in t; uniform-σ and log space σ_{eff} linearly and geometrically between $\sigma_{\text{eff}}(T-1)$ and $\sigma_{\text{eff}}(0)$; Karras grids use $\sigma_i = (\sigma_{\max}^{1/\rho} + \frac{i}{N-1}(\sigma_{\min}^{1/\rho} - \sigma_{\max}^{1/\rho}))^\rho$ with $\rho = 7$. Levels are mapped back to fractional t by exactly inverting the piecewise-linear $\log\bar{\alpha}$. Response grids place levels at equal quantiles of $\int d\sigma/\Delta\sigma_{\max}(\sigma)$, using the eight-condition budget curve. The fine reference uses $\Delta t = 0.25$ from $T - 1$ to 0.

Noise. Initial and conditioning noise are drawn per trajectory from a generator keyed by data set, source, window and noise seed, independent of batching and of the candidate, so all deterministic candidates of a trajectory share them. For $\eta = 1$ the additional posterior noise comes from one generator per (cell, checkpoint, noise seed) with a fixed batch order. The native DDPM runs use the released sampler without modification and do not share random numbers with the DDIM runs.

B Frozen Physical-Validation Protocol

Order of commitments. The physical protocol (horizon, reference spacing, candidate grids, windows, observables, tolerances and hard-failure definition) and the split file were written before any main prediction. One change was made before any main prediction: the Tra split was revised so that every simulation excluded by the upstream training script is a test condition. The decision protocol was written before any selection outcome. Selection policies were computed from calibration and selection data only and frozen into a manifest whose hash every test record carries; each test record is time-stamped after the freeze. All 96 policies per data set (two rules times 48 settings), all candidates, all conditions and all failures are reported; sampler details are in Appendix A and post-hoc analyses in Appendix E.

Observables. Velocity fields are multiplied by the fluid mask (1 in the fluid, 0 in obstacles). The spatial spectrum is the mean Fourier power of the masked velocity in 16 radial wavenumber bins excluding the mean; on Tra and Inc it is a diagnostic of an obstacle-domain field, not an isotropic spectrum. Kinetic energy is $\frac{1}{2}|v|^2$, weighted by density on Tra, averaged over fluid cells. The primary error of Equation 2 uses means over frames 100–200. A rollout is a hard failure if any output is not finite or, on Tra, if more than 10^{-4} of fluid space–time cells have non-positive density or pressure; hard failures count as failures at every tolerance. None occurred on Tra or Iso. Secondary metrics (state errors over frames 0–20 and 100–200, the error of the late mean field, the temporal power spectrum and velocity autocorrelation) are reported and enter no decision.

Table 4: Physical-condition splits, fixed before any main prediction. Model-held-out: test sources excluded by the upstream training script. Each rollout conditions on two true frames and generates the stated horizon autoregressively; stride is the physical frame spacing of the pretrained model.

Data	Calibration	Selection	Test	Model-held-out	Windows	Horizon	Stride
Tra	8	8	23	8	2	200	2
Iso	8	4	14	all (slices held out)	2	200	1

Budget audit. For candidate c, the audit feature is $R(c) = \max_i \Delta\sigma_i/\Delta\sigma_{\max}(\sigma_i)$ over non-terminal steps, with log-linear interpolation of the 10th-percentile budget curve of the n calibration conditions at tolerance 0.05, and R is undefined outside the calibrated support. The two anchors (log-8, Karras-16) are rolled out with the fine reference on the selection conditions. Sorting anchors by R, the predicted reference deviation of c is the running maximum of anchor deviations up to the first anchor with $R \geq R(c)$; a candidate beyond both anchors is unsupported. The predicted physical error is $q_f + (1 + q_f)\,\hat{e}_{\mathrm{ref}}(c)$, where q_f is the fine reference's physical-error quantile on the selection conditions. This composition is a heuristic without a coverage guarantee. The audit selects the cheapest candidate predicted within tolerance, with ties broken by pool order, and abstains otherwise, or if the reference or the relevant anchor had a hard failure.

Direct pilot. The direct pilot queries the pool in the fixed order log-4, Karras-4, log-8, Karras-8, log-16, Karras-16, log-32, Karras-32 on the same selection conditions and stops before the next query would exceed the audit's calibration cost. It selects the cheapest queried candidate whose observed physical-error quantile is within tolerance and that had no hard failure.

Costs and uncertainty. Costs are sample-network evaluations actually spent: for the audit, the local budget measurement, the fine reference and the two anchors; for the direct pilot, its queries. GPU seconds are amortized from batch timings and reported only as a secondary figure. Windows, noise seeds and checkpoints share source conditions, so intervals are computed by resampling source conditions (Tra) or contiguous blocks of slices (Iso) and are descriptive; Iso has one DNS realization and no interval describes new-realization risk.

C Local Error Budgets: Measurement and Validation

Original spectrum protocol (Iso). For a start level σ and trial step h, the endpoint error is the relative ℓ_2 distance between the radially averaged spectra of the first velocity channel of $U_h(x_\sigma)$ and $R_h(x_\sigma)$. Starts are off-grid noisy states obtained by interpolating $\log\bar{\alpha}$; conditioning noise is fixed within an initial condition; trial steps are bracketed adaptively. Every per-condition measurement is stored with left-, right-, cap- and boundary-censoring flags, and a noise bin is analyzed only with enough uncensored measurements, low censoring, low monotonicity violation and non-trivial state change; the same gates are used at every tolerance. For held-out prediction, the two checkpoints' curves are combined by the median or the minimum, and multiples $0.5, 0.8, 1, 1.2, 1.5$ of the budget are tested on fresh initial conditions with shared noise; endpoints outside the noise support are infeasible and excluded from denominators.

The first-channel spectrum tolerance tracks state-level errors (Table 7), and budget decisions largely agree when failure is defined by state error instead (Table 10).

Campaign protocol (real conditions). For each data set and checkpoint, 16 noise levels are spaced geometrically over the model's support. At each level, trial endpoints are placed at fractions 0.02, 0.04, 0.08, 0.16, 0.32, 0.64 and 1 of the available interval, clamped to the trained grid, and compared with the fine reference by the relative ℓ_2 error of the normalized state (tolerances 0.02, 0.05, 0.1 and 0.2; 0.05 is primary). States are reached by integrating the reverse process from Gaussian noise, conditioned on real frame pairs at rollout frames 0, 50 and 150 of each calibration condition. The budget of a measurement is its largest contiguous passing prefix; the curve is the 10th percentile over conditions, windows and frames, for nested sets of 2, 4 and 8 calibration conditions. No parametric law is fitted.

Table 5: Held-out safe-step prediction. Budgets calibrated from the response curve transfer to fresh held-out initial conditions: below-budget steps pass uniformly, at-budget steps mostly pass, and $1.2\times$ steps collapse. The $< 1\times$ column aggregates $0.5\times$ and $0.8\times$ multipliers. Records report feasible IC-level endpoint tests for $< 1 \times /1 \times /1.2\times$; the final column reports feasible/infeasible multiplier–sigma grid cells. Infeasible endpoints are metadata-only and are excluded from denominators.

Seed	Budget	$< 1\times$ pass	$1\times$ pass	$1.2\times$ pass	Records	Grid
`_00`	median	100.0%	96.5%	20.1%	1024/512/448	31/9
`_00`	min-seed	100.0%	97.9%	14.8%	1024/512/384	30/10
`_01`	median	100.0%	90.6%	1.8%	1024/512/448	31/9
`_01`	min-seed	100.0%	94.3%	3.9%	1024/512/384	30/10

Table 6: Schedule-level safety audit from response-budget overshoot in the diversified stress audit. Each schedule group is one seed, schedule family, and NFE setting on the Iso interval. The audit flags a schedule as unsafe when $R_{\max} = \max_i \Delta\sigma_i/\Delta\sigma_{\max}(\sigma_i) > 1$. It spans 120 groups with $R_{\max}$ from 0.23 to 1.08, flags all median-failing groups, and produces conservative false positives near the threshold.

Target	TP	FP	TN	FN	Precision	Recall
Median spectrum failure	20	10	90	0	0.667	1.000
Majority IC failure	20	10	90	0	0.667	1.000
Any IC failure	30	0	78	12	1.000	0.714

D THE BUDGET LAW: TWO REGIMES, NOT CFL

These results come from the original Iso spectrum protocol (three tolerances, two checkpoints; Appendix C). The budget curve is fit by single-power, free-knot piecewise and constrained power laws after gating (Appendix C). A single exponent is rejected in every run; the low- and high-noise exponents are close to $2/3$ and 1; the Kolmogorov-inspired $4/3$ is rejected; and the knot coincides with a denoiser saturation crossing that does not move with the tolerance (Table 11, Figure 5).

To test a CFL explanation we finite-difference the DDIM update field along Fourier perturbations and take the maximum gain as a stiffness proxy $\lambda_{\max}$. A strict CFL law $\Delta\sigma_{\max} \propto \lambda_{\max}^{-1}$ would make $\lambda_{\max}\Delta\sigma_{\max}$ constant; it varies by $3.74\times$ and $2.53\times$ across noise levels, whereas $\lambda_{\max}\Delta\sigma_{\max}^2$ varies by only $1.25\times$ and $1.42\times$. Two-half-step defects are of order $p_D \approx 2.4$ and direct endpoint errors of order 2.4–3.0 (Table 12, Figure 6). The budget is an endpoint-accuracy limit whose coefficient changes at the saturation transition, not a stability limit. The protocol also transfers to other conditioning regimes and model variants (Table 13), and the budgets change by at most 1.8% when the reference spacing is refined from 1.0 to 0.25 (Table 14).

E HORIZON, ENERGY AND STOCHASTICITY ANALYSES

All analyses in this appendix re-score or decompose rollouts after the frozen outcomes were observed, except the stochasticity cells, whose protocol was written before they were run. None changes a frozen policy.

Stochasticity decomposition. Cells: $\eta = 1$ at Karras-8, Karras-32 and uniform-t-100 on both checkpoints and both noise seeds; $\eta = 1$ with the fine grid and $\eta = 0$ with uniform-t-100 on checkpoint 00 and one noise seed (for cost). $\eta = 0$ Karras and fine cells are the frozen test rollouts. The pre-registered readings were: determinism dominates if the $\eta = 1$ 100- and 397-step medians fall below 0.5 and within 0.1 of the native sampler; the update scheme dominates if $\eta = 1$ stays within 0.1 of $\eta = 0$ on matched grids. The first reading holds (Table 2). The cells took 2.85 A100-hours (392 rollouts). Table 18 compares every cell with the $\eta = 1$ fine rollouts of the same trajectories statistically; stochastic paths are not compared pathwise.

Table 7: Threshold/metric calibration on held-out safe-step records ($n = 7808$). The primary first-channel radial spectrum threshold tracks state-space and channel-level degradation, supporting its use as a scientific endpoint proxy rather than a purely arbitrary fitting knob.

Diagnostic pair	Correlation
Budget multiplier vs. spectrum/threshold ratio, Spearman	0.958
Budget multiplier vs. state L2, Spearman	0.967
Spectrum/threshold ratio vs. state L2, Spearman	0.929
Spectrum/threshold ratio vs. state L2, Pearson	0.870
Spectrum/threshold ratio vs. max channel L2, Spearman	0.929

Table 8: Censoring of the local-budget measurements (eight calibration conditions, local tolerance 0.05). A right-censored measurement passed at the largest trial step, so its budget is a lower bound. No measurement was left-censored and none recovered after failing.

Data	Ckpt	Measurements	Right-censored	Left-censored	Non-monotone
Tra	00	768	405 (53%)	0	0
Tra	01	768	403 (52%)	0	0
Iso	00	768	247 (32%)	0	0
Iso	01	768	246 (32%)	0	0

F Decision Results

No safety guarantee. The frozen rules aim at a 90th-percentile tolerance and carry no coverage guarantee. At the main Tra setting, the audit approves schedule–condition groups that fail at the median (Table 20), the selected log-8 schedule has at least one failing repetition on 3 of 23 source conditions for both checkpoints, and its temporal power spectrum error remains near 0.9 even when the primary error passes. Abstention is not a proof of infeasibility, and a selection is not a certificate.

G Inc: Calibration-Side Results

Inc follows the same frozen protocol as Tra and Iso (Appendix B): 55 simulations staged from the released archive, one 200-step window per simulation starting at frame 800, calibration and selection sets of eight training-included simulations interleaved across the Reynolds range, and a test set of 38 simulations including all 20 that the upstream training script excludes. The results reported here are the eight selection conditions, on which the policies of Section 7 are frozen; they are calibration-side data of the protocol and are scored exactly as the test data of Tra and Iso.

The Inc emulator is accurate: the fine reference's physical error is 0.021 and 0.010 for the two checkpoints, its late-window kinetic energy matches the truth, and refining it from $\Delta t = 0.25$ to 0.125 on the development condition changes the error by less than 0.001. Every coarse schedule, down to four evaluations per frame, lands within 0.014 of the reference (within 0.006 on checkpoint 01), and the sampler deviations of the Karras grids (0.0002–0.004) are well below the emulator's error of 0.010–0.021. The rank correlation between fidelity and physics is therefore near zero and not informative: the reference passes the validity check of Section 7, and the cheapest sampler is as good as the reference. Budget calibration on Inc produced complete response grids for both checkpoints; 74–76% of the calibrated budgets are right-censored, so the Inc model tolerates larger local steps than the trial grid probes.

H Further Related Work

Diffusion modelling and sampling. Denoising diffusion models descend from nonequilibrium-thermodynamics (Sohl-Dickstein et al., 2015), denoising score matching (Vincent, 2011) and the reverse-time SDE view (Anderson, 1982; Song et al., 2021b); improved noise schedules and learned variances (Nichol & Dhariwal, 2021) and the survey of Yang et al. (2024) cover the design space. Guidance (Dhariwal & Nichol, 2021; Ho & Salimans, 2022; Epstein et al., 2023; Chen et al., 2026),

Table 9: Local-budget transfer to real conditioning and to autoregressive rollout states (frames 0, 50 and 150 pooled; four test sources per data set, all windows, two checkpoints). Entries: steps meeting the 5% local relative-state tolerance / feasible steps, at multiples of the measured budget. Budgets are the 10th-percentile contiguous passing prefix of the eight-condition calibration, measured on a trial grid whose steps double; where the next trial step failed, the true limit lies between $1\times$ and $2\times$ the recorded budget, which is why steps above $1\times$ can pass.

Data	Conditioning state	$0.5\times$	$0.8\times$	$1.0\times$	$1.2\times$	$1.5\times$
Tra	true	672/672	576/576	750/768	375/384	5/384
	fine rollout	672/672	576/576	742/768	371/384	2/384
	Karras-8 rollout	672/672	576/576	744/768	368/384	1/384
Iso	true	672/672	576/576	768/768	337/528	240/384
	fine rollout	672/672	576/576	734/768	358/528	240/384
	Karras-8 rollout	672/672	576/576	746/768	380/528	240/384

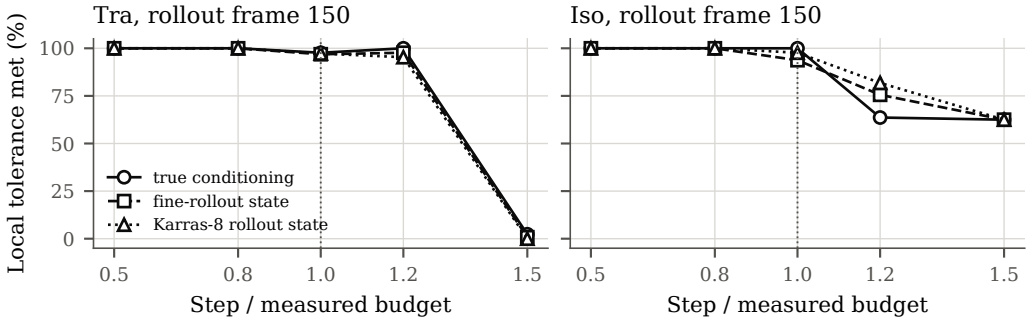

Figure 4: Local-budget transfer at rollout frame 150 (four test sources per data set, both checkpoints): fraction of steps meeting the local tolerance at multiples of the measured budget, when conditioning on true frames, on a fine-reference rollout, or on a Karras-8 rollout.

editing (Meng et al., 2022) and inverse-problem samplers (Chung et al., 2023) modify the reverse process for a downstream objective; our samplers are unmodified DDIM(η) integrations of the trained reverse process, and the question is only how their outputs should be validated.

Learned simulators and autoregressive rollouts. Neural surrogates for PDEs range from graph-network simulators, whose rollouts are stabilized by training noise (Sanchez-Gonzalez et al., 2020), to learned coarse models of turbulence (Stachenfeld et al., 2022; Kochkov et al., 2021), neural operators (Li et al., 2021) and U-Net families (Gupta & Brandstetter, 2023); the pushforward trick (Brandstetter et al., 2022) and PDE-Refiner (Lippe et al., 2023) counter the error accumulation of autoregressive prediction that we also observe. Diffusion models enter fluid problems as reconstruction priors (Shu et al., 2023) and as data-assimilation posteriors (Rozet & Louppe, 2023); in all of these, agreement with a numerical reference on spectra and statistics is the usual yardstick, which is the yardstick this paper examines.

Physical priors and step-size laws. Differentiable physics (Holl et al., 2020) and data-driven surrogates (Brunton et al., 2020) motivate physically grounded evaluation of learned simulators. The $4/3$ exponent tested in Appendix D was a Kolmogorov-inspired candidate (Kolmogorov, 1941; Frisch, 1995); the measured budgets reject it together with the strict CFL analogy (Courant et al., 1928; LeVeque, 2002).

I Compute

All physical-validation runs used one NVIDIA A100 (SXM4, 40 GB) with PyTorch 2.7. The Tra and Iso campaign (development checks, budget calibration, selection, test and transfer probes) took 22.55 GPU-hours of queue time; the native Iso control took 1.08 GPU-hours, the stochasticity cells

Table 10: Do local-budget decisions agree when failure is defined by another metric? Held-out local-step records (Iso, $n = 3904$ per checkpoint; threshold matched on checkpoint 00 to the spectrum pass rate at tolerance 0.2 and applied unchanged to checkpoint 01, with $\pm 20\%$ sensitivity). Recall and agreement of the budget flag and of the spectrum criterion against state-based failure. Spectrum–state Spearman correlations on other data sets: Inc 0.855, Tra 0.895.

Failure metric	Threshold scale	Budget flag		Spectrum criterion	
		recall	agreement	recall	agreement
state ℓ_2	0.8	0.929	0.984	0.905	0.959
	1.0	1.000	0.997	0.972	0.971
	1.2	1.000	0.890	0.973	0.870
max-channel ℓ_2	0.8	0.928	0.983	0.904	0.959
	1.0	1.000	0.998	0.972	0.972
	1.2	1.000	0.912	0.975	0.892

Table 11: Multi-threshold response-law summary. Std. is descriptive across six threshold-by-seed runs.

Quantity	Value	Interpretation
Monotonicity violations	$0/2737$	clean endpoint response
Minimum ΔAIC vs. single power	47.4	single exponent rejected
Mean low-regime exponent	0.656 ± 0.050	compatible with $2/3$
Mean high-regime exponent	1.046 ± 0.046	compatible with 1
Maximum single-run exponent	1.135	rejects $4/3$
Aux saturation crossing	0.856 ± 0.008	threshold-stable knot proxy

2.85 GPU-hours, and the Inc calibration side (development check, budget calibration and selection rollouts) 2.25 GPU-hours. The original budget-law and mechanism experiments are listed in Table 23. Throughput is about 1.7 ms per sample and network evaluation at batch size 28 or more.

Table 12: Mechanism diagnostics. The probes reject a first-order CFL law but are consistent with an endpoint-accuracy interpretation whose coefficient changes near the denoiser saturation transition.

Diagnostic	Observation	Interpretation
Strict CFL product	$\lambda\Delta\sigma$ varies by $3.74\times$ (seed _00) and $2.53\times$ (seed _01)	$\Delta\sigma_{\max} \propto \lambda^{-1}$ does not collapse the measured budgets
Quadratic accuracy proxy	$\lambda\Delta\sigma^2$ varies only $1.25\times$ and $1.42\times$	a second-order local-accuracy budget organizes the stiffness probe substantially better
Local defect order	two-half-step defect gives median $p_D = 2.41/2.44$; direct endpoint fits rise from $p \approx 2.4$ to 3.0	the endpoint tolerance is accuracy-limited, not a first-order stability boundary
Saturation transition	denoiser saturation crossings at $\sigma = 0.847/0.865$	the coefficient/regime change aligns with the empirical response-curve knot

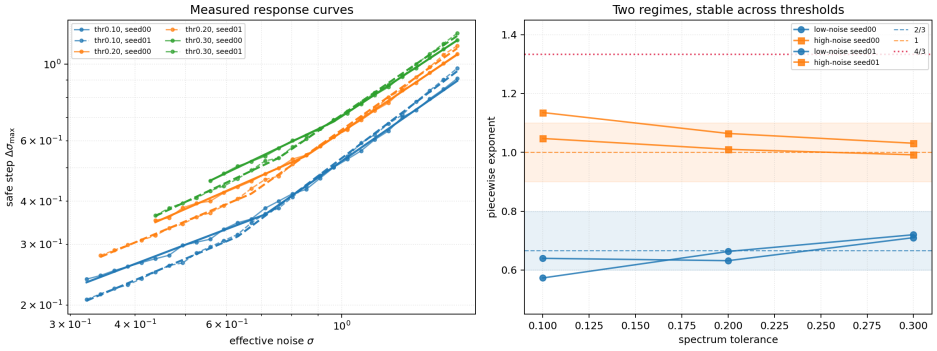

Figure 5: Multi-threshold budget curves on Iso (original protocol): two regimes with low-noise exponents near $2/3$ and high-noise exponents near 1; $4/3$ is not supported ($0/2737$ monotonicity violations; minimum ΔAIC 47.4 against a single power law).

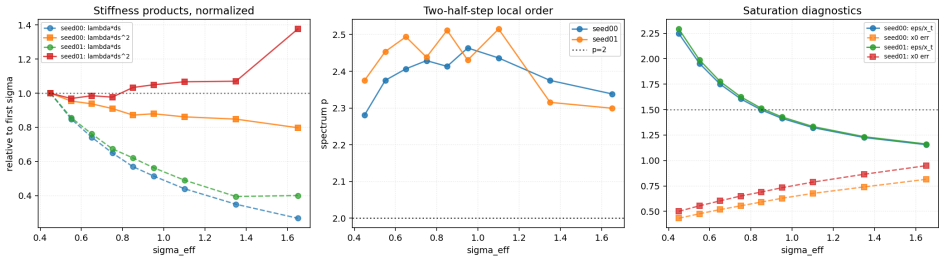

Figure 6: Mechanism probes across both checkpoints: the strict CFL product $\lambda_{\max}\Delta\sigma_{\max}$ is not constant, the quadratic accuracy product $\lambda_{\max}\Delta\sigma^2_{\max}$ is much flatter, local defects are near second order, and the saturation crossing aligns with the knot.

Table 13: Cross-condition response-curve checks beyond the main Iso checkpoints. NCN/Iso validates a low-regime response law; Inc/Tra seed01 checks show that the protocol transfers to different conditioning regimes, although the exact exponent is not universal.

Variant	Seed	Threshold	Bins kept	Mono. viol.	Slope b	Gate
NCN/Iso	_00	0.01	18/18	0.00%	0.6956	pile-flag
NCN/Iso	_01	0.01	18/18	0.00%	0.6849	pass
Inc	_01	0.02	13/20	0.31%	0.863	uncens.-flag
Tra	_01	0.05	20/20	0.16%	1.132	pass

Table 14: Fine-step reference convergence sanity check. We rerun five representative Iso seed00 response bins with three deterministic DDIM reference step sizes. The table reports median $\Delta\sigma_{\max}$ over 16 ICs per bin. All bins are uncensored for all three references; the intentionally small n_{IC} makes this an appendix reference check, not a replacement for the main response-curve runs.

σ_{eff}	$\Delta\sigma_{\max}^{dt=1.0}$	$\Delta\sigma_{\max}^{dt=0.5}$	$\Delta\sigma_{\max}^{dt=0.25}$	rel. diff 1.0 vs 0.25	rel. diff 0.5 vs 0.25
0.527	0.393	0.386	0.386	1.80%	0.00%
0.703	0.495	0.491	0.487	1.76%	0.88%
0.937	0.629	0.625	0.620	1.45%	0.72%
1.250	0.781	0.776	0.776	0.64%	0.00%
1.667	1.086	1.086	1.086	0.00%	0.00%
Median absolute relative difference				1.45%	0.00%
Maximum absolute relative difference				1.80%	

Table 15: Physical error (median, checkpoint 00) recomputed on alternative frame windows of the same rollouts (post hoc, descriptive). The frozen protocol scores frames 100–200; the upstream evaluation horizon for Iso is 100 frames. The stochastic fine cell has one noise seed (28 rollouts); all other rows have two (56).

Data	Sampler	0–20	20–60	50–100	0–100	100–200
Tra	fine (det.)	0.118	0.181	0.185	0.174	0.214
	Karras-8 (det.)	0.137	0.186	0.205	0.190	0.231
	Karras-32 (det.)	0.120	0.179	0.183	0.174	0.216
Iso	fine (det.)	0.359	0.649	0.778	0.624	0.711
	Karras-8 (det.)	0.461	0.757	0.728	0.679	0.653
	Karras-32 (det.)	0.380	0.673	0.782	0.642	0.693
	fine (η=1)	0.076	0.216	0.308	0.229	0.441
	Karras-32 (η=1)	0.091	0.252	0.346	0.259	0.446
	native DDPM	0.086	0.238	0.344	0.255	0.440

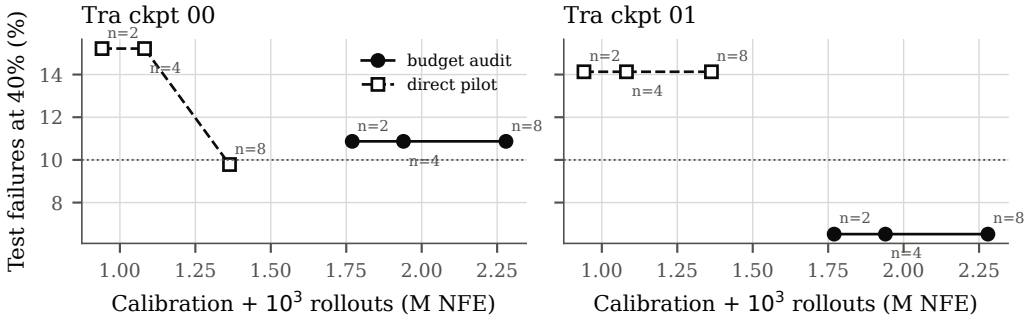

Figure 7: Test failure rate at tolerance 0.4 against total cost (calibration plus 1,000 deployment rollouts) for the frozen budget audit and the direct pilot at calibration sizes $n = 2, 4, 8$ (90th-percentile rule). The dotted line is the nominal 10% target. Iso is omitted because every policy abstains.

Table 16: Spectral error decomposition of coarse schedules relative to the fine deterministic reference. With $u = (S_f - S_*)/\|S_*\|$ and $v = (S_c - S_f)/\|S_*\|$ for late-window mean spectra of the coarse (c) and fine (f) samplers and truth ($*$), the coarse spectral error satisfies $\|u + v\|^2 = \|u\|^2 + \|v\|^2 + 2\langle u, v\rangle$ exactly. A negative cross term means the coarse deviation offsets part of the reference's error. Columns: rollouts with negative cross term; rollouts where the coarse schedule beats the fine reference on the primary error; median cosine between u and v.

Data	Ckpt	Schedule	$\langle u, v\rangle < 0$	coarse better	median $\cos(u, v)$
Tra	00	Karras-8	35/92	35/92	+0.754
Tra	00	Karras-32	53/92	53/92	-0.580
Tra	00	unif.-σ-16	51/92	48/92	-0.251
Tra	01	Karras-8	34/92	24/92	+0.821
Tra	01	Karras-32	36/92	29/92	+0.672
Tra	01	unif.-σ-16	35/92	28/92	+0.817
Iso	00	Karras-8	56/56	56/56	-0.999
Iso	00	Karras-32	56/56	56/56	-0.999
Iso	00	unif.-σ-16	56/56	56/56	-1.000
Iso	01	Karras-8	54/56	54/56	-1.000
Iso	01	Karras-32	53/56	53/56	-1.000
Iso	01	unif.-σ-16	56/56	56/56	-0.999

Table 17: Reference refinement on the development condition (one source, checkpoint 00, one noise seed). Physical error: primary error against truth; deviation: primary error against the $\Delta t = 0.25$ reference used throughout. Refining the deterministic reference changes the physical error far less than the physical error itself: on Iso the reference is numerically converged but not physically accurate.

Data	Reference	NFE	Window start	Physical error	Deviation from $\Delta t = 0.25$
Tra	$\Delta t = 0.125$	153	0	0.161	0.0012
	$\Delta t = 0.125$	153	500	0.070	0.0033
	$\Delta t = 0.25$	77	0	0.162	0.0000
	$\Delta t = 0.25$	77	500	0.069	0.0000
	$\Delta t = 0.5$	39	0	0.164	0.0017
	$\Delta t = 0.5$	39	500	0.070	0.0027
	$\Delta t = 1$	20	0	0.165	0.0039
	$\Delta t = 1$	20	500	0.071	0.0148
Iso	$\Delta t = 0.125$	793	0	0.809	0.0025
	$\Delta t = 0.125$	793	500	0.739	0.0013
	$\Delta t = 0.25$	397	0	0.808	0.0000
	$\Delta t = 0.25$	397	500	0.739	0.0000
	$\Delta t = 0.5$	199	0	0.807	0.0061
	$\Delta t = 0.5$	199	500	0.738	0.0028
	$\Delta t = 1$	100	0	0.804	0.0229
	$\Delta t = 1$	100	500	0.736	0.0101

Table 18: Statistical deviation of every sampler from the stochastic fine rollouts ($\eta = 1$, $\Delta t = 0.25$, 397 evaluations) on the same 28 Iso trajectories (checkpoint 00, noise seed 2411): median relative ℓ_2 distance between late-window mean spectra and median relative difference of late-window kinetic energy. Stochastic rollouts are compared as statistics of the same physical condition, not pathwise.

Sampler	n	Late spectrum	Late energy
$\eta = 1$, uniform-t-100	28	0.014	0.040
$\eta = 1$, Karras-32	28	0.019	0.085
$\eta = 1$, Karras-8	28	0.264	0.242
native DDPM, 100	28	0.023	0.036
$\eta = 0$, uniform-t-100	28	0.512	0.255
$\eta = 0$, fine (397)	28	0.525	0.270
$\eta = 0$, Karras-32	28	0.501	0.244
$\eta = 0$, Karras-8	28	0.387	0.160
$\eta = 0$, log-32	28	0.507	0.246
$\eta = 0$, log-8	28	0.430	0.178
$\eta = 0$, uniform-t-32	28	0.463	0.220
$\eta = 0$, uniform-t-16	28	0.393	0.160

Table 19: Paired comparison of the frozen selections at tolerance 0.4 (90th-percentile rule): difference in test failure rate (audit minus direct pilot, percentage points) and a descriptive 95% interval from resampling source conditions. Repetitions share sources; the interval describes observed source variability, not a population risk.

Data	Ckpt	n	Audit	Direct pilot	Difference (pp)	Interval
Tra	00	2	log-8	log-4	-4.3	[-12.0, +3.3]
Tra	00	4	log-8	log-4	-4.3	[-12.0, +3.3]
Tra	00	8	log-8	Karras-4	+1.1	[-3.3, +6.5]
Tra	01	2	log-8	log-4	-7.6	[-18.5, +0.0]
Tra	01	4	log-8	log-4	-7.6	[-18.5, +0.0]
Tra	01	8	log-8	log-4	-7.6	[-18.5, +0.0]

Table 20: Schedule–condition groups approved by the frozen budget audit (main setting: $n = 8$, 90th-percentile rule, $\tau = 0.4$) that fail on test: groups whose median repetition fails and groups with any failing repetition, out of all failing groups, with the number of rejected passing groups. Groups are correlated; counts are descriptive.

		Median-failing groups		Any-failing groups		
Data	Ckpt	missed	rejected passing	missed	rejected passing	Approved
Tra	00	10/16	63	15/27	57	62.5%
Tra	01	1/6	64	10/23	56	62.5%
Iso	00	0/112	0	0/112	0	0.0%
Iso	01	0/112	0	0/112	0	0.0%

Table 21: Selected schedules on test sources that the upstream training script excluded ("held out") versus sources it trained on (main setting, $n = 8$, 90th-percentile rule, $\tau = 0.4$). Held-out sources are model extrapolation conditions; zero observed failures on 32 correlated repetitions is not a guarantee.

Data	Ckpt	Rule	Sources	n	Rollouts	Median error	Failures
Tra	00	audit	trained	15	60	0.242	16.7%
Tra	00	audit	held out	8	32	0.211	0.0%
Tra	00	direct pilot	trained	15	60	0.278	13.3%
Tra	00	direct pilot	held out	8	32	0.231	3.1%
Tra	01	audit	trained	15	60	0.233	8.3%
Tra	01	audit	held out	8	32	0.227	3.1%
Tra	01	direct pilot	trained	15	60	0.291	15.0%
Tra	01	direct pilot	held out	8	32	0.252	12.5%

Table 22: Preliminary Inc results from the completed calibration side of the frozen protocol: 200-step rollouts on the eight *selection* conditions (two checkpoints, two noise seeds; 16 rollouts per candidate and checkpoint, 672 in total), scored against ground truth exactly as in Table 1. These conditions were used to freeze the selection policies and are not held-out test conditions; the frozen held-out evaluation (38 test simulations, 20 of them model-held-out) is running and will be reported. Columns as in Table 1; "beating" counts coarse schedules with a lower median physical error than the fine reference. Development-condition refinement of the fine reference: $\Delta t = 0.125$: 0.022; $\Delta t = 0.25$: 0.022; $\Delta t = 0.5$: 0.021; $\Delta t = 1$: 0.021.

Data	Ckpt	Sources	Fine ref. error		Best coarse schedule	Beating	ρ	E ratio
			median	q90	median	fine	100–200	fine
Inc (selection)	00	8	0.021	0.023	0.015 (unif.-σ-32)	7/20	+0.253	1.00
	01	8	0.010	0.014	0.005 (unif.-σ-32)	11/20	−0.256	1.00

Table 23: Representative compute resources. All runs are inference-only on pretrained checkpoints; no model training is performed. Wall times are representative A100 inference-only runs when available.

Experiment	GPU	Seeds/runs	Size	Wall time
Schedule failure audit	A100	2 Iso	32 ICs, 12 schedules, 5 NFEs	10.3 min total
Threshold/metric calibration	CPU	2 Iso seeds	7808 held-out records	seconds
Inc response	A100	1 seed	20 bins, 32 ICs/bin	27.8 min
Tra response	A100	1 seed	20 bins, 32 ICs/bin	23.2 min
Held-out safe steps	A100	2 Iso	8 sigmas, 64 ICs, 2 policies	about 1 h/seed