TARGET FIDELITY IS NOT NUMERICAL FIDELITY SIGNED BIAS–DISCRETIZATION ALIGNMENT AT EQUAL NFE

Anonymous authors
Paper under double-blind review

ABSTRACT

At equal numbers of score evaluations (NFE), Heun can converge faster to a fixed learned probability-flow ODE while Euler remains closer to the target. We prove this separation for a smooth Gaussian VP construction: numerical errors have exact orders one and two, yet Euler has strictly smaller target W_2 error at every sufficiently large even budget. The result persists in a one-dimensional neighborhood of smooth nonlinear fields, with a uniform budget threshold. The target-error gap decays as M^{-1} and the relative advantage vanishes. Signed alignment between continuous-flow bias and Euler's leading defect determines the eventual coupled ranking. In DSM-trained GMMs, the joint ranking persists beyond the earliest checkpoint: at 2400 and 6000 updates, respectively, 6/15 and 11/15 fixed networks retain the opposite coupled target ranking throughout a common 128–2048 NFE normal-order diagnostic window, while Heun has smaller numerical error in every comparison. Oracle-assisted alignment estimates do not outperform simple target-gap extrapolation. A controlled 32-dimensional study provides additional coupled evidence; an unpaired kernel-embedding test finds no statistically certified MMD reversal. These results establish a locally robust distributional existence theorem and a controlled learned-score mechanism; image-scale relevance remains open.

1 INTRODUCTION

Probability-flow ODEs turn diffusion sampling into deterministic numerical integration and make high-order solvers a natural route to fewer neural-network evaluations (Song et al., 2021; Karras et al., 2022; Lu et al., 2022; Zhao et al., 2023). Classical order says that, once asymptotic behavior is visible, Euler's numerical error decays linearly in the step size while Heun's decays quadratically (Hairer et al., 1993). This fact answers a numerical-fidelity question: which discrete trajectory is closer to the continuous trajectory defined by the learned score? It does not by itself answer the target-fidelity question: which discrete endpoint is closer to the desired distribution under the same number of score evaluations? Our main result is not an order-collapse claim: Euler and Heun can retain these nominal orders while their target-error ranking is reversed.

The distinction matters because the continuous trajectory is already biased when the score is imperfect. Let $X_0^\star$ be an exact-score endpoint, $\bar{X}_0^\theta$ the exact endpoint under an approximate score, and $X_{0,m}^{M,\theta}$ the endpoint of solver m with NFE budget M. Under a shared terminal sample, define

$$B := \bar{X}_0^\theta - X_0^\star, \qquad E_m(M) := X_{0,m}^{M,\theta} - \bar{X}_0^\theta. \tag{1}$$

The coupled target RMS has the exact geometry

$$\boxed{R_m^2(M) = F_c^2 + D_m^2(M) + 2\rho_m(M) F_c D_m(M),} \tag{2}$$

where $F_c = \|B\|_{L_2}$, $D_m = \|E_m\|_{L_2}$, and, when $F_c D_m > 0$, $\rho_m = \langle B, E_m \rangle_{L_2} / (F_c D_m)$. If either norm vanishes, the cross term is zero and we set $\rho_m = 0$ by convention. Throughout, R_m denotes error under this specified coupling; its equality with a distributional distance requires an additional argument. The elementary identity shows how reducing D_m can remove a beneficial cancellation. In particular, F_c is a limiting error, not a finite-budget lower floor.

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.

For Euler (E) and Heun (H), the comparison has the exact decision form

$$R_{\mathrm{H}}^2 - R_{\mathrm{E}}^2 = \underbrace{2(C_{\mathrm{H}} - C_{\mathrm{E}})}_{P_{\mathrm{align}}} - \underbrace{(D_{\mathrm{E}}^2 - D_{\mathrm{H}}^2)}_{S_{\mathrm{num}}}, \qquad C_m := \langle B, E_m \rangle. \tag{3}$$

When $S_{\mathrm{num}} > 0$, Heun is numerically more faithful, yet it loses against the target exactly when the alignment penalty P_{align} exceeds the numerical saving S_{num}.

We start with a variance-preserving (VP) probability flow for Gaussian data with a smooth affine score error. It gives exact finite-M recurrences and an optimal one-dimensional coupling. We then prove that the normal-order reversal survives small smooth nonlinear perturbations, without requiring Gaussian endpoints. This realized signed interaction can be obscured by bounds that average over centered score perturbations.

Our contributions are as follows. They concern target-error ranking conditional on a fixed learned field; they do not claim that the nominal numerical order of Euler or Heun collapses.

- **Locally robust W_2 reversal at ordinary numerical orders.** A fixed smooth VP score field has $D_{\mathrm{E}} \asymp M^{-1}$ and $D_{\mathrm{H}} \asymp M^{-2}$ but smaller Euler target W_2 error at every sufficiently large even M (Theorem 1). Proposition 2 extends both rankings to a smooth nonlinear function neighborhood with a uniform budget threshold.
- **Signed ranking and vanishing relative advantage.** Negative alignment between continuous-flow bias and Euler's leading defect yields simultaneous $D_{\mathrm{H}} < D_{\mathrm{E}}$ and $R_{\mathrm{E}} < R_{\mathrm{H}}$ (Corollary 6). Both target errors approach the same nonzero limit, so a persistent strict ranking need not yield a fixed-percentage gain. Exact affine recurrences and an RKHS extension specify the finite-budget geometry and metric boundaries.
- **Fixed-checkpoint order and mechanism tests.** We jointly measure numerical orders and coupled target gaps over seven budgets for all 15 original GMM runs at 800, 2400, and 6000 updates. The later checkpoints retain reversals in a common normal-order diagnostic window, without a monotone training-maturity trend. Simple extrapolation baselines, reference audits, curve-wise bands, and higher-dimensional and unpaired tests delimit the evidence.

The exact construction and its nonlinear extension establish realizability and local stability, not prevalence under training. Synthetic learned-score tests examine the mechanism; the unpaired distributional gate is negative. We do not propose a low-cost selector or establish image-model relevance. Complete proofs appear in the appendix.

2 Related Work and Scope

Fast deterministic diffusion samplers. Probability-flow ODEs give deterministic diffusion samplers (Song et al., 2021). EDM separates preconditioning, schedules, and integration (Karras et al., 2022); DPM-Solver and UniPC exploit diffusion structure (Lu et al., 2022; Zhao et al., 2023). We use classical methods to isolate the mechanism, not to propose a new sampler. Equation 3 also applies to diffusion-specific methods, but its practical effect there remains untested. Jiao et al. (2025) already study signed first-order defects and report competitive image sampling. Their Theorem 3 concerns cancellation between two discretization defects; ours concerns a defect's interaction with flow bias and two opposite distance rankings. Appendix A makes this distinction explicit.

Convergence with approximate scores. General Wasserstein analyses propagate initialization, score, and discretization errors through probability-flow ODEs (Gao & Zhu, 2025). Beyler & Bach (2025) give Wasserstein bounds for Euler and Heun, while Huang et al. (2025) analyze arbitrary-order exponential Runge–Kutta schemes with additive score and discretization terms. End-to-end theory further decomposes learning, optimization, and reverse-time discretization errors (Huang et al., 2026); complementary score-geometry work shows that raw L_2 score magnitude need not determine marginal fidelity (Khelifa et al., 2026). Zhou (2026) independently shows that forward-time score accuracy does not certify numerical stability, reinforcing the distinction between score quality and solver behavior. Our point is different and finite-budget: even after numerical stability and lower discretization error are established, a realized signed endpoint interaction can reverse target-error ordering.

Table 1: Selected statements in the closest analyses. The distinction is the joint conclusion, not the first observation that a lower-order sampler can win.

Result	Quantity and conclusion
Salajegheh & Mudur (2026), Thm. 5.1 / Cor. 5.2	True-field one-step error bound and equal-NFE global interaction-scale comparison.
Jiao et al. (2025), Thm. 3	Cancellation of the leading defects of two first-order discretizations of the same learned ODE.
Theorem 1 and Proposition 2 here	Exact first/second learned-flow orders and opposite target W_2 ranking for every sufficiently large even NFE; persistence in a one-dimensional C_b^5 neighborhood with one uniform threshold.

Discretization–learning interaction and order collapse. Salajegheh & Mudur (2026) already analyze local and global interaction at equal NFE; neither global analysis nor fixing a learned field distinguishes our work by itself. Table 1 isolates the additional joint conclusion. Appendix A gives the proposition-level comparison and explains why the spatial-derivative sufficient condition is different, without claiming a contradiction of their theory.

Gaussian and spectral analyses. Pierret & Galerne (2025) derive exact Gaussian diffusion flows and Wasserstein errors for arbitrary numerical schemes, including fixed-compute comparisons; their Gaussian-microtexture experiment also reports that Euler–Maruyama can be more faithful than Heun under an approximate neural score. Our contribution is the equal-NFE signed decision criterion, its cancellation and fixed-percentage consequences, and its extension to distributional mean embeddings, rather than the first empirical observation that lower order can win. Hurault et al. (2025) provide a fine-grained Gaussian analysis with affine score perturbations. Their perturbations are centered and independent across time and the result averages over them, which removes the realization-conditional linear interaction studied here; they analyze Euler rather than equal-NFE Euler–Heun ranking. Spectral transfer, schedule design, and parameterization-aware Gaussian discretization are developed by Benita et al. (2025) and Hurault et al. (2026). Our spectral experiment therefore diagnoses finite-budget onset rather than proposing new parameterization-aware spectral theory.

Mixtures and score-error accumulation. Recent work analyzes score-error accumulation in Gaussian mixtures and derives stability-aware coordinates or grids (He et al., 2026; Gao et al., 2026). Our mixture experiment instead stress-tests the signed endpoint decomposition; it does not provide a new GMM stiffness bound, grid, or transferable certificate.

3 Setup: Target Error Is a Vector Geometry

3.1 VP probability flow with affine score error

Consider the constant-rate VP SDE

$$\mathrm{d}X_t = -\beta X_t\,\mathrm{d}t + \sqrt{2\beta}\,\mathrm{d}W_t, \qquad X_0 \sim \mathcal{N}(0,\lambda), \tag{4}$$

with $\beta, T, \lambda > 0$. Its marginal variance and exact score are

$$v_t = 1 + (\lambda - 1)e^{-2\beta t}, \qquad s_t^\star(x) = -x/v_t. \tag{5}$$

We use the deterministic affine perturbation

$$\widehat{s}_t(x) = s_t^\star(x) + \delta(t)x + \zeta(t). \tag{6}$$

This perturbation is a tractable proxy for a learned score residual, not a claim that neural-network errors are affine. Along a fixed exact trajectory, a smooth nonlinear residual admits a first-order local expansion in the state; the multiplicative and additive terms in Equation 6 isolate the corresponding state-dependent and state-independent components while preserving an exactly solvable flow. The trained-score experiments below then test the same endpoint geometry without prescribing the residual field. With reverse clock $u = T - t$, the approximate-score probability flow is

$$\frac{\mathrm{d}Y_u}{\mathrm{d}u} = q(u)Y_u + r(u), \quad q(u) = \beta\left[1 + \delta(T-u) - v_{T-u}^{-1}\right], \quad r(u) = \beta\zeta(T-u). \tag{7}$$

We initialize exactly with $Z := Y_0 \sim \mathcal{N}(0, v_T)$. Thus finite T is not itself an error source in this model. Exact-score integration maps Z to $X_0^\star = k_\star Z$, where $k_\star = \sqrt{\lambda/v_T}$.

The approximate continuous endpoint is $\bar{X}_0^\theta = k_\theta Z + m_\theta$ with

$$k_\theta = k_\star \exp\left(\beta \int_0^T \delta(t)\,\mathrm{d}t\right), \tag{8}$$

$$m_\theta = \int_0^T \beta\zeta(t)\sqrt{\lambda/v_t}\exp\left(\beta\int_0^t \delta(a)\,\mathrm{d}a\right)\mathrm{d}t. \tag{9}$$

The additive perturbation is needed for a non-degenerate range of alignments in one dimension: with centered, purely multiplicative error, B and E_m are collinear and $\rho_m \in \{-1, 1\}$ whenever both are nonzero.

3.2 Equal NFE, coupled RMS, and W_2

An NFE counts one evaluation of $\widehat{s}_t$. Euler therefore takes M steps, whereas Heun and explicit midpoint take $M/2$ steps; RK4 takes $M/4$. We compare endpoints under the same Z. This makes $R_m = \|X_{0,m}^{M,\theta} - X_0^\star\|_{L_2}$ a coupled RMS, not automatically a distributional distance.

For an affine numerical endpoint $X_{0,m}^{M,\theta} = k_m Z + m_m$, the geometry is explicit:

$$\begin{aligned} F_c^2 &= v_T(k_\theta - k_\star)^2 + m_\theta^2, \\ D_m^2 &= v_T(k_m - k_\theta)^2 + (m_m - m_\theta)^2, \\ C_m &= v_T(k_\theta - k_\star)(k_m - k_\theta) + m_\theta(m_m - m_\theta). \end{aligned} \tag{10}$$

The one-dimensional Gaussian distance is

$$W_2^2\Big(\mathcal{L}(X_{0,m}^{M,\theta}), \mathcal{N}(0,\lambda)\Big) = m_m^2 + v_T\big(|k_m| - k_\star\big)^2. \tag{11}$$

Hence $R_m = W_2$ when $k_m \geq 0$. We report coupled RMS throughout and invoke W_2 only in the orientation-preserving regime.

The shared-Z construction always yields $W_2(\mathcal{L}(X_{0,m}^{M,\theta}), \mathcal{L}(X_0^\star)) \leq R_m(M)$, since it is one admissible coupling. Yet ordering the costs of two admissible couplings does not order the corresponding optimal-transport distances. Theorem 1 closes this gap in one dimension: increasing endpoint maps make the shared-latent construction the optimal monotone (quantile) coupling. Outside that argument, the learned-GMM results are coupled mechanism tests; the unpaired MMD gate separately asks whether the ranking survives for an unpaired distributional discrepancy.

4 Normal-Order Reversal and Its Local Stability

4.1 A fixed smooth model with a persistent Wasserstein reversal

The conditional ranking result has a distributional realization with one score field fixed across every budget. Here both leading numerical coefficients are nonzero, so the stated orders are asymptotic equivalents rather than only upper bounds.

Theorem 1 (Fixed-field, normal-order W_2 reversal)**.** *Fix $\lambda \in (0,1)$, $\beta, T > 0$, and $k_\star = \sqrt{\lambda/v_T} < 1$. Choose any fixed $a \in (\log k_\star, 0)$, set $q = a/T$, and define the smooth score perturbation*

$$\delta_q(t) = q/\beta - 1 + v_t^{-1}, \qquad \zeta(t) = 0. \tag{12}$$

The learned reverse flow is $Y' = qY$. Write P_θ, P_{E}^M, and P_{H}^M for its continuous and numerical endpoint laws. Then

$$\begin{aligned} D_{\mathrm{E}}(M) &\sim c_{\mathrm{E}} M^{-1}, & c_{\mathrm{E}} &= \sqrt{v_T}e^a a^2/2 > 0, \\ D_{\mathrm{H}}(M) &\sim c_{\mathrm{H}} M^{-2}, & c_{\mathrm{H}} &= -2\sqrt{v_T}e^a a^3/3 > 0. \end{aligned} \tag{13}$$

There exists M_0, depending only on the fixed parameters, such that every even $M \geq M_0$ satisfies

$$W_2(P_{\mathrm{H}}^M, P_\theta) < W_2(P_{\mathrm{E}}^M, P_\theta),$$
$$W_2(P_{\mathrm{E}}^M, P^\star) < W_2(P_{\mathrm{H}}^M, P^\star). \tag{14}$$

On these budgets $D_m = W_2(P_m^M, P_\theta)$ and $R_m = W_2(P_m^M, P^\star)$ exactly. Moreover $R_{\mathrm{H}} - R_{\mathrm{E}} \sim c_{\mathrm{E}}/M$, whereas $R_{\mathrm{E}}/R_{\mathrm{H}} \to 1$.

The interval for a is open; the field is not retuned with M. Positive Gaussian scale maps make the common-latent coupling optimal. Appendix H.2 proves the result and checks $(\beta, T, \lambda, a) = (1, 1, 0.5, -0.2)$.

Proposition 2 (Nonlinear local stability of the two rankings)**.** *Fix the parameters of Theorem 1 and perturb the reverse field to $f_g(u, x) = qx + g(u, x)$, with $g \in C_b^5([0, T] \times \mathbb{R})$. We use $\|g\|_{C_b^5} = \max_{i+j\leq 5} \sup_{u,x} |\partial_u^i \partial_x^j g(u, x)|$. There exist $\varepsilon_0 > 0$ and one even M_0 such that, for every fixed $\|g\|_{C_b^5} < \varepsilon_0$ and every even $M \geq M_0$, both inequalities in Equation 14 hold with the perturbed continuous and numerical laws. Uniformly over this neighborhood,*

$$D_{\mathrm{E}} = c_{\mathrm{E}}(g)M^{-1} + O(M^{-2}), \qquad D_{\mathrm{H}} = c_{\mathrm{H}}(g)M^{-2} + O(M^{-3}), \tag{15}$$

where both coefficients have strictly positive uniform lower bounds. Also $R_{\mathrm{H}} - R_{\mathrm{E}} = c_R(g)M^{-1} + O(M^{-2})$, with $\inf_g c_R(g) > 0$, and $R_{\mathrm{E}}/R_{\mathrm{H}} \to 1$.

The endpoint laws need not be Gaussian: small-step numerical maps remain increasing, so their shared-latent coupling is optimal in one dimension. Appendix G supplies uniform defect expansions and a negative alignment margin, not merely continuity at each fixed M. This is local stability in a strong smoothness norm, not a high-dimensional W_2 or training-error guarantee.

The nonlinear class also has a pointwise-in-time score interpretation. The reverse-clock VP identity gives

$$\widehat{s}_g(t, x) = (q/\beta - 1)x + g(T - t, x)/\beta,$$
$$p_{g,t}(x) \propto \exp\left\{ \tfrac{1}{2}(q/\beta - 1)x^2 + \beta^{-1} \int_0^x g(T - t, y)\, \mathrm{d}y \right\}.$$

Since $q < 0$ and g is bounded, $p_{g,t}$ is a normalizable one-dimensional density with a negative quadratic log-density tail. We do not assume that $(p_{g,t})_t$ is the marginal curve of a single VP forward process. More importantly, Appendix G proves uniform defect remainders and $\langle B_g, A_{\mathrm{E}}(g) \rangle \leq -\gamma_0/2$, yielding $(R_{\mathrm{H}}^g)^2 - (R_{\mathrm{E}}^g)^2 \geq \gamma_0/M - K_R/M^2$. Together with the numerical-order and monotonicity bounds, these estimates give one common even-budget existence threshold over the entire neighborhood, rather than a budget-by-budget continuity statement. Appendix Table 3 also checks five fixed nonlinear amplitudes over seven budgets: all 35 comparisons preserve both rankings, with final-interval orders near one and two. This checks one nonlinear direction, not the full neighborhood or a certified radius.

4.2 Why the two rankings separate

The construction has a simple proof: the endpoint scales eventually satisfy $k_\star < k_{\mathrm{E}} < e^a < k_{\mathrm{H}}$, while the two numerical errors have nonzero first- and second-order coefficients. The nonlinear extension requires more than continuity at a fixed budget. Uniform defect remainders preserve a negative bias–defect inner product throughout a small function neighborhood; monotone maps make the coupling optimal in one dimension.

More generally, if $F_c > 0$, $E_{\mathrm{E}} = L_{\mathrm{E}}/M + O_{L_2}(M^{-2})$ and $E_{\mathrm{H}} = O_{L_2}(M^{-2})$, expansion of the target norms gives

$$R_{\mathrm{H}} - R_{\mathrm{E}} = -\frac{\langle B, L_{\mathrm{E}} \rangle}{F_c M} + O(M^{-2}). \tag{16}$$

Negative alignment therefore makes Euler closer to the target even when Heun is more faithful to the learned flow. Positive alignment gives the opposite target ranking. Both target errors approach F_c, so strict eventual ranking is compatible with a vanishing relative advantage. Appendix B states the general ranking theorem, its normal-order corollary, the fixed-percentage budget bound, and the affine recurrences. These supporting calculations are not claimed as a new numerical method.

Equation 16 also gives a curve-level signature: on a common budget window, D_{E} and D_{H} should exhibit orders one and two, while $M(R_{\mathrm{H}} - R_{\mathrm{E}})$ approaches $-\langle B, L_{\mathrm{E}} \rangle / F_c$. A one-budget reversal is insufficient because the $O(M^{-2})$ remainder can still control the sign. We therefore inspect numerical orders and persistent target-gap signs jointly, rather than treat isolated target rankings as evidence for the asymptotic mechanism.

5 Experiments

All reported synthetic experiments use fixed pseudorandom seeds and have regeneration scripts. We count actual score evaluations: Euler/Heun/midpoint/RK4 cost $1/2/2/4$ evaluations per step. Continuous scalar quantities use analytic formulas plus Gauss–Legendre quadrature; reference resolutions and doubling audits are specified per experiment in Appendix J.

5.1 All later checkpoints: ordinary orders and opposite rankings

To test whether the joint ranking is confined to the earliest checkpoint, we extend all 15 original runs at 2400 and 6000 updates to the same seven budgets, retaining every curve and the same 512 initial samples. Heun has smaller learned-flow error in all 210 comparisons. Figure 1 displays every later run, including non-reversals, and the unscaled effect distribution. Table 2 reports the common numerical-only diagnostic window, 128–2048 NFE: Euler and Heun observed orders stay within $[0.8, 1.2]$ and $[1.8, 2.2]$ on every doubling interval, while 6/15 and 11/15 curves retain smaller Euler coupled target error throughout. These are point-order diagnostics, not asymptotic or order-confidence guarantees.

Across all seven budgets, reversal counts at 800, 2400, and 6000 updates are 8/15, 5/15, and 10/15, respectively: there is no monotone maturity trend. The 30 later curves share 15 training trajectories, and later updates do not establish optimization convergence. Their largest relative target excess falls below 0.2% at 2048 NFE, consistent with convergence to a common biased endpoint, not with loss of Heun's higher order or a general practical advantage for Euler. Appendix J.8 reports all curves, effect sizes, reference checks, and the numerical-only window rule.

5.2 Earlier checkpoint and mechanism checks

At checkpoint 800, the same 15 original runs retain Heun's numerical advantage at all seven budgets, while 8 have the opposite coupled target ranking throughout. Figure 7 in Appendix J.7 illustrates the first run in protocol order; all runs, including non-reversals, are retained. Replay of archived metrics has maximum discrepancy 0.0e+00, an implementation-consistency check rather than an independent replication.

An oracle-assisted estimate of Euler's leading defect from 32 and 64 NFE predicts 59/60 larger-budget gap signs. Direct gap extrapolation matches that sign count with lower RMSE. At the later checkpoints, alignment predicts 117/120 signs, whereas a two-term gap fit predicts 118/120 with lower RMSE (Appendix J.8). These dependent comparisons support leading-defect consistency, not predictive superiority or a deployable solver selector.

5.3 Training-generated score errors at the original budgets

The original suite uses small time-conditioned score networks trained by variance-weighted denoising score matching: 3 two-dimensional shared-covariance GMMs, 5 seeds per model, checkpoints at 800, 2400, and 6000 updates, and NFE budgets 32, 64, and 128. These are the same training runs as above, not additional replications. Mean exact-score validation relative RMSE falls from 0.517 to 0.196, so the checkpoints span materially different approximation quality.

Heun has lower continuous-flow discretization error in all 135 comparisons. Even so, the alignment penalty exceeds the numerical saving in 77 cases (57.0%), so Heun has larger coupled target error. At least one reversal occurs for every mixture and checkpoint, and 38 have paired terminal-path bootstrap intervals wholly above one.

Grouping by model, training seed, and checkpoint gives 45 three-budget trajectories; 23 retain both rankings throughout. Table 5 in Appendix J.6 summarizes their medians, not independent replications.

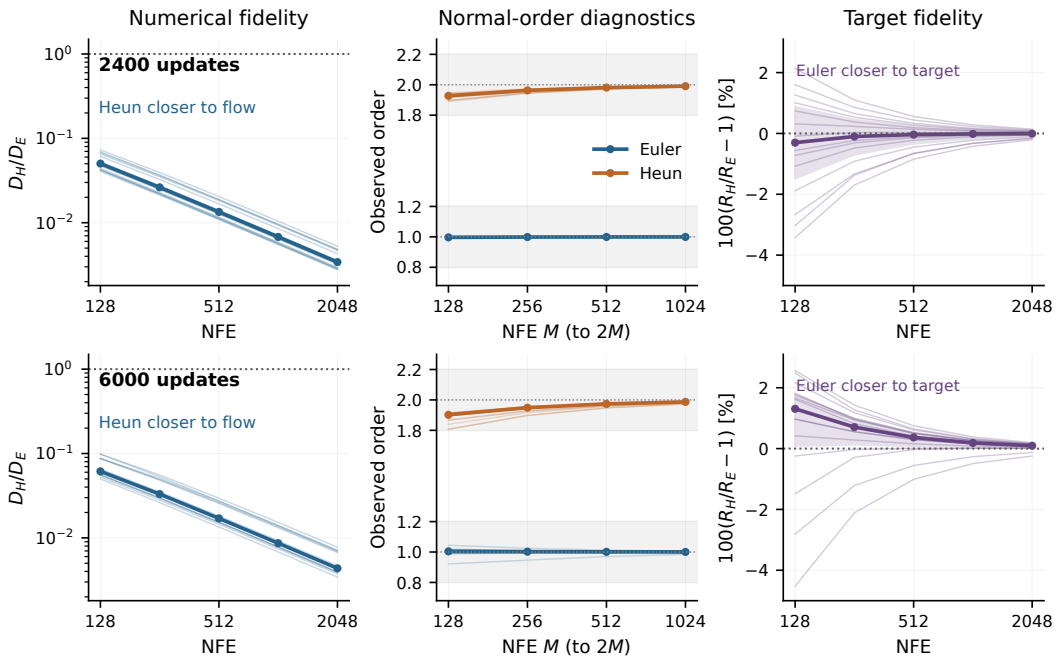

Figure 1: All 15 original runs at each later checkpoint, with no target-based filtering. Columns show the learned-flow error ratio, observed Euler/Heun orders, and unscaled relative coupled target excess. Positive excess favors Euler. Thin lines retain every run; thick lines are medians. Purple shading is the across-run interquartile range; gray bands mark the declared order tolerances. Neither is a confidence band. The common 128–2048 NFE window uses numerical diagnostics only. Order points at M use $M \to 2M$, ending at 1024.

Table 2: Later checkpoints, 15 fixed networks each. The common window is $128, 256, 512, 1024, 2048$ NFE; final orders use $1024 \to 2048$. Bands are approximate 95% seven-budget bands within each network, restricted here to the common window, not simultaneous across networks. Target excess is $100(R_{\mathrm{H}}/R_{\mathrm{E}} - 1)$, maximized over all 15 networks.

Checkpoint (updates)	2400	6000
$D_{\mathrm{H}} < D_{\mathrm{E}}$, all seven budgets	15/15	15/15
$R_{\mathrm{E}} < R_{\mathrm{H}}$, all seven budgets	5/15	10/15
$R_{\mathrm{E}} < R_{\mathrm{H}}$, common window	6/15	11/15
Band lower bound > 1, common window	5/15	8/15
Final observed order, E	0.999–1.001	0.985–1.006
Final observed order, H	1.986–1.993	1.973–1.989
Max. target excess at 128 NFE	2.137%	2.562%
Max. target excess at 2048 NFE	0.141%	0.194%

The effects are usually modest but recur across the tested grid. Among reversals, the H/E ratio has median 1.017, 90th percentile 1.040, and maximum 1.099; 52 cases exceed a 1% degradation and 4 exceed 5%. An empirical sliced-W_2 diagnostic ranks Heun worse in 72 cases, including 64 with projection-bootstrap intervals above one. Figure 6 relates both effect sizes to the exact decision boundary. Reference doubling changes endpoint RMS by at most 2.0e-06 for the learned flow and 2.8e-09 for the target.

5.4 A CONTROLLED 32-DIMENSIONAL TEACHER–STUDENT BRIDGE

To test whether the mechanism survives aggregation across dimensions while keeping every endpoint observable, we use an analytic 32-dimensional, four-component GMM score as the teacher and train three student networks by the same DSM objective. Across three checkpoints and three NFE budgets,

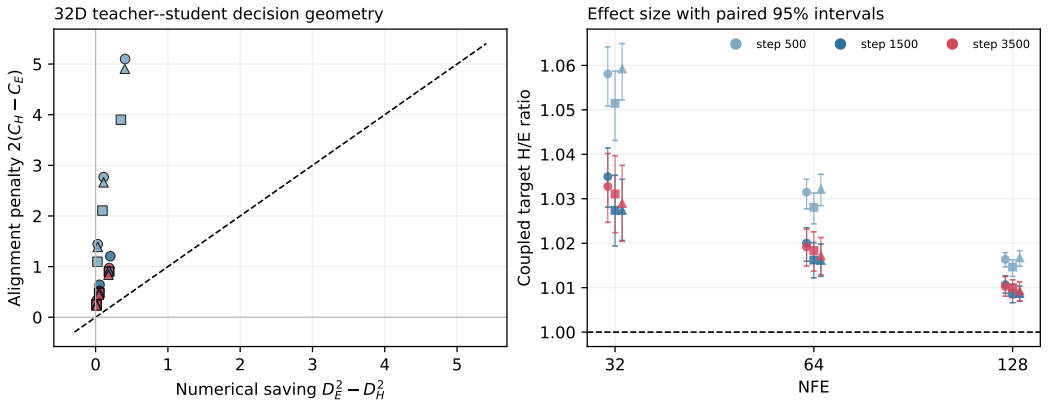

Figure 2: Controlled 32D analytic-teacher/DSM-student experiment. Left: every comparison lies above the exact reversal boundary. Right: coupled H/E effect sizes with paired terminal-path 95% bootstrap intervals; colors denote checkpoints and markers denote training seeds.

all 27 comparisons satisfy $D_{\mathrm{H}} < D_{\mathrm{E}}$ but $R_{\mathrm{H}} > R_{\mathrm{E}}$, and every paired 95% interval excludes one. The alignment penalty therefore exceeds the numerical saving at every tested budget.

The coupled H/E ratio has median 1.019 and maximum 1.059; 23 cases exceed 1% and 3 exceed 5%. Empirical sliced-W_2 ratios have median 1.044 and maximum 1.112. Figure 2 summarizes these results. This experiment establishes a higher-dimensional realization with a learned student, not performance on image models.

5.5 Unpaired distributional gate

The same additive decomposition holds exactly for kernel mean embeddings, giving an MMD analogue without target pairing (Theorem 7, Appendix). This does not establish a corresponding FID decomposition. We apply it to the same kind of trained GMM score without pairing generated samples to target samples. Before examining rankings, we fix three kernels—linear, a KID-style cubic polynomial, and a three-scale RBF mixture whose median bandwidth is estimated from target samples only. For each of three checkpoints and NFE 32, 64, and 128, we draw 1024 independent target observations and 1024 common-latent generated paths. We use block U-statistics with block size 64 and 1000 bootstrap resamples; the high-NFE learned-flow reference is doubled on a held-out subset. This is a screening gate rather than a high-power test: the modest coupled effects can be harder to detect after unpairing, so the null result is not interpreted as evidence that the mechanism fails distributional evaluation.

Heun is numerically more faithful in all 27 solver–kernel comparisons. Only 3 polynomial-kernel point estimates reverse the target ranking, and none has a 95% interval wholly above zero; the linear and RBF kernels show no point reversal. A linear-Gaussian calibration with a known population gap yields 92.0% empirical coverage for nominal 95% percentile intervals over 100 repetitions; this small calibration does not precisely diagnose coverage. The maximum reference-doubling RMS change is 5.0e-07. This gate therefore provides no statistically certified evidence that coupled reversals transfer automatically to MMD. An image-scale test should predeclare its feature map, kernel, sample size, and decision margin.

5.6 Designed controls and transfer boundary

Appendix J gives outcome-selected magnitude controls, not prevalence estimates. An affine grid shows that $F_c + D_m$ is not order-preserving. Exact-score tests isolate order onset: at $\lambda = 10^{-4}$, raw VP time shows no second-order onset through 8192 NFE, whereas log-variance time reaches it at 32 NFE. Scalar perturbations transferred to a GMM do not preserve solver labels. Alignment must be assessed in the model being sampled.

6 LIMITATIONS AND CONCLUSION

These results do not constitute a practical solver selector. Theorem 1 gives a distributional existence result, and Proposition 2 extends it locally to non-Gaussian endpoints in one dimension. Neither guarantees reversal under small training L_2 score error or general high-dimensional transport. The learned-GMM evidence remains coupled RMS, not automatically W_2. Alignment extrapolation does not outperform simple target-gap baselines on this suite. The RKHS identity is general, but its empirical use depends on the feature map, kernel, sample size, and reference; the negative MMD gate limits that extension.

The matched cases are search-selected, and the learned-score tests use synthetic GMMs and small networks; repeated rows are neither independent nor prevalence estimates. Later checkpoints rule out confinement to the earliest tested stage, not dependence on model scale or incomplete training. Small high-budget target gaps are consistent with both methods approaching the same biased flow: persistent ranking is not a nonvanishing practical gain, and smoothness alone does not determine the signed interaction. The 1024-sample unpaired gate may be underpowered, and paired RMS ratios do not determine MMD power. Image models, diffusion-specific integrators, adaptive or stochastic samplers, and perceptual metrics require a separate predeclared, reference-audited study.

A convincing image-scale follow-up should fix checkpoints before comparing solvers, use a common NFE window selected without target rankings, audit the learned-flow reference by resolution doubling, retain every run, and report absolute alongside relative gaps. These controls would distinguish signed bias–defect cancellation from optimization noise, reference error, or metric variance; without them, a lower FID or KID value for Euler would not identify the mechanism studied here.

Solver order controls D_m, but target error depends on $B + E_m$: alignment can reverse Euler and Heun despite lower Heun error. Benchmarks should therefore distinguish learned-flow from target fidelity.

AI USE STATEMENT

Generative AI tools were used to help develop theoretical models and mathematical claims, assist with proof drafting, design and implement the synthetic experiments, interpret numerical outputs, search and summarize literature, prepare figures, and draft and edit the manuscript. No human-subject or proprietary data were supplied to these tools. AI-assisted code and numerical claims were checked using exact identities, unit tests, empirical-order tests, paired, projection, and block bootstraps, and reference-tolerance audits; complete formulas and deterministic regeneration code accompany the draft. The authors are responsible for verifying all AI-assisted mathematical, numerical, and textual content before submission and for the final claims and artifacts.

ETHICS STATEMENT

This work uses only synthetic Gaussian and Gaussian-mixture distributions and raises no direct human-subject, privacy, or dataset-consent concerns. The main research-integrity risk is overgeneralizing a controlled error mechanism to learned image models. We address that risk by stating the coupling, orientation, model, and statistical limitations explicitly and by reporting the negative transfer result in the GMM stress test.

REPRODUCIBILITY STATEMENT

The appendices contain complete proofs, including the nonlinear extension in Appendix G. Appendix J specifies coefficients, NFE counts, seeds, references, and bootstrap protocols. The code regenerates the JSON results, LaTeX macros, and figures; headline numbers and the new diagnostic tables are imported from generated files.

REFERENCES

Roi Benita, Miki Elad, and Joseph Keshet. Spectral analysis of diffusion models with application to schedule design. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0077.

Eliot Beyler and Francis Bach. Convergence of deterministic and stochastic diffusion-model samplers: A simple analysis in Wasserstein distance, 2025. URL https://arxiv.org/abs/2508.03210.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

Weiguo Gao, Ming Li, Lei Shi, and Hanfei Zhou. A quantitative approximation framework for flow distillation in diffusion models, 2026. URL https://arxiv.org/abs/2606.03820.

Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow ODEs of diffusion models in Wasserstein distances. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1009–1017. PMLR, 2025. URL https://proceedings.mlr.press/v258/gao25c.html.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://www.jmlr.org/papers/v13/gretton12a.html.

Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer, second edition, 1993. doi: 10.1007/978-3-540-78862-1.

Baoxiang He, Valentio Iverson, Shuai Li, Cheng Chen, and Bo Jiang. The accumulation of score estimation error in diffusion models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://openreview.net/forum?id=YoGDpLRqN8.

Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Fast convergence for high-order ODE solvers in diffusion probabilistic models, 2025. URL https://arxiv.org/abs/2506.13061.

Jinshu Huang, Yiming Jiang, and Chunlin Wu. From score learning to discretized sampling: An end-to-end generalization analysis of diffusion models, 2026. URL https://arxiv.org/abs/2607.23226.

Samuel Hurault, Matthieu Terris, Thomas Moreau, and Gabriel Peyré. From score matching to diffusion: A fine-grained error analysis in the Gaussian setting, 2025. arXiv:2503.11615v3, revised May 2026.

Samuel Hurault, Thomas Moreau, and Gabriel Peyré. Geometry-aware discretization error of diffusion models, 2026. URL https://arxiv.org/abs/2605.08392.

Yuchen Jiao, Na Li, Changxiao Cai, and Gen Li. Are first-order diffusion samplers really slower? a fast forward-value approach, 2025. URL https://arxiv.org/abs/2512.24927.

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*, volume 35, 2022. doi: 10.52202/068431-1926.

Naïl B. Khelifa, Richard E. Turner, and Ramji Venkataramanan. Diffusion models observe only gradients: A geometric perspective on score matching errors, 2026. URL https://arxiv.org/abs/2606.06179.

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*, volume 35, 2022. doi: 10.52202/068431-0418.

Emile Pierret and Bruno Galerne. Diffusion models for Gaussian distributions: Exact solutions and Wasserstein errors. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49355–49381. PMLR, 2025. URL https://proceedings.mlr.press/v267/pierret25a.html.

Farzad Salajegheh and Sudhir Mudur. When higher order hurts: Pre-asymptotic order collapse in generative ODE sampling. In *Computer Vision – ECCV 2026*, volume 17060 of *Lecture Notes in Computer Science*, pp. 151–167. Springer, 2026. doi: 10.1007/978-3-032-37092-1_9. URL https://doi.org/10.1007/978-3-032-37092-1_9.

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

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*, volume 36, 2023. doi: 10.52202/075280-2170.

Yiwei Zhou. Score accuracy along the forward diffusion does not certify numerical stability in diffusion sampling, 2026. URL https://arxiv.org/abs/2607.08757.

A Proposition-Level Comparison with Discretization–Learning Interaction

The comparison below uses the full conference paper of Salajegheh & Mudur (2026), Sections 3–5, and supplementary Sections B and D. Both papers fix a learned field and account for method cost at equal NFE. The distinction is the stated quantity and ranking condition.

Object	Salajegheh–Mudur	Present result
Error reference	Theorem 5.1, Eq. (6): true-field one-step error bound	D_m: learned-flow endpoint error; R_m: coupled target endpoint error
Ranking	Corollary 5.2, Eq. (7): global interaction-scale comparison	Theorem 1: two strict endpoint W_2 inequalities
Condition	Spatial derivative magnitudes in their Assumption 2	Sign of $\langle B, L_{\mathrm{E}}\rangle$; positive-scale Gaussian realization
Budget regime	Corollary 5.4: crossover conditions	Every sufficiently large even M for one fixed field
Global analysis	Theorem 5.3: manifold exit; supplement B.3: accumulated interaction	Nominal learned-flow orders with reversed target ordering

There is also a concrete separation from their displayed spatial-curvature condition. In Theorem 1, the ideal reverse field is $f^\star(x,u) = q_\star(u)x$ and the learned field is qx. Its residual $\varepsilon(x,u) = (q - q_\star(u))x$ has $\partial_x^2 \varepsilon = 0$ and $\sup_u |\partial_x \varepsilon| > 0$. Consequently the condition $(2h)^2 C_2(\varepsilon) > hC_1(\varepsilon)$ in their Eq. (7), with their stated spatial-derivative definitions, is not satisfied, whereas our strict target reversal holds. This separates that sufficient mechanism from ours. It does not assert a contradiction of their theorem, equivalence of their manifold hypotheses with Gaussian full support, or the absence of other backward-error interactions: their time augmentation can introduce mixed derivatives even when spatial curvature vanishes.

Signed defects are not new in themselves. Theorem 3 of Jiao et al. (2025) compares idealized forward-value and DDIM discretizations of the same learned ODE and shows that their first-order endpoint defects sum to $O(M^{-2})$. This is defect–defect cancellation. In contrast, our condition $\langle B, L_{\mathrm{E}}\rangle < 0$ concerns flow-bias–defect interaction; Theorem 1 realizes two reversed W_2 rankings despite exact first- and second-order convergence. The cited theorem does not state this joint conclusion. We claim neither the first signed-error analysis nor the first observation of lower-order image-quality gains.

B Supporting Geometry and General Ranking

B.1 Magnitude non-identifiability

Proposition 3 (Scalar magnitudes do not identify solver ranking)**.** *For any $F > D_{\mathrm{E}} > D_{\mathrm{H}} > 0$, there are two endpoint geometries with the same triple $(F_c, D_{\mathrm{E}}, D_{\mathrm{H}}) = (F, D_{\mathrm{E}}, D_{\mathrm{H}})$ and opposite Euler–Heun target rankings.*

Proof. Let e be a unit vector and set $B = Fe$. With $(E_{\mathrm{E}}, E_{\mathrm{H}}) = (-D_{\mathrm{E}}e, +D_{\mathrm{H}}e)$, Euler wins because $F - D_{\mathrm{E}} < F + D_{\mathrm{H}}$. Reversing both signs makes Heun win because $F - D_{\mathrm{H}} < F + D_{\mathrm{E}}$. The three norms are unchanged. Thus no rule receiving only these magnitudes can rank both geometries correctly. □

This proposition is Hilbert-space geometry rather than a diffusion realizability claim. Proposition 9 supplies smooth VP instances near both extreme alignments; the learned-score experiment tests whether the same decision geometry appears without prescribing the error field.

B.2 Affine endpoint recurrences

Theorem 4 (Exact equal-NFE Gaussian geometry)**.** *Let δ, ζ be continuous and let M be even. On a uniform reverse-time grid, both Euler and Heun have affine endpoints $k_m Z + m_m$. For Euler,*

set $K_{\mathrm{E}} = M$, $h_{\mathrm{E}} = T/M$; for Heun, set $K_{\mathrm{H}} = M/2$, $h_{\mathrm{H}} = 2T/M$. With $q_n = q(nh_m)$ and $r_n = r(nh_m)$, initialize $k_0 = 1, m_0 = 0$ and update

$$k_{n+1} = A_n k_n, \qquad m_{n+1} = A_n m_n + d_n, \tag{17}$$

where

$$A_n^{\mathrm{E}} = 1 + h_{\mathrm{E}} q_n, \qquad d_n^{\mathrm{E}} = h_{\mathrm{E}} r_n, \tag{18}$$

$$A_n^{\mathrm{H}} = 1 + \tfrac{h_{\mathrm{H}}}{2}(q_n + q_{n+1}) + \tfrac{h_{\mathrm{H}}^2}{2} q_{n+1} q_n, \qquad d_n^{\mathrm{H}} = \tfrac{h_{\mathrm{H}}}{2}(r_n + r_{n+1} + h_{\mathrm{H}} q_{n+1} r_n). \tag{19}$$

The resulting F_c, D_m, C_m, R_m are exactly given by Equations 2 and 10. If $k_m \geq 0$, the coupled target error is the Gaussian W_2 in Equation 11; otherwise the latter uses $|k_m|$ and can be smaller than the coupled RMS.

The theorem is a direct but useful finite-M reduction: solver selection becomes Euclidean geometry in the coefficient plane $(\sqrt{v_T}k, m)$. Its proof, including the score and time-direction conventions, is in Appendix E.

B.3 Eventual ranking versus fixed-percentage advantage

Classical consistency gives $D_{\mathrm{E}} = O(M^{-1})$ and $D_{\mathrm{H}} = O(M^{-2})$ under sufficient smoothness. Equal NFE introduces the factor four in Heun's leading term because its step size is $2T/M$.

For a fixed improvement threshold $\gamma \in (0, 1)$, define the budgets at which Heun improves on Euler by at least that percentage as

$$\mathcal{W}_\gamma := \{M \in 2\mathbb{N} : R_{\mathrm{H}}(M) \leq (1-\gamma) R_{\mathrm{E}}(M)\}. \tag{20}$$

The experiments use $\gamma = 0.2$.

Theorem 5 (Alignment controls eventual ranking; fixed-percentage advantage is bounded)**.** *Fix the learned field, time horizon, target, and terminal coupling independently of M. Let $B = \bar{X}_0^\theta - X_0^\star$ with $F_c = \|B\|_{L_2} > 0$. Suppose, along even M,*

$$E_{\mathrm{E}}(M) = \frac{L_{\mathrm{E}}}{M} + O_{L_2}(M^{-2}), \qquad E_{\mathrm{H}}(M) = \frac{4L_{\mathrm{H}}}{M^2} + O_{L_2}(M^{-3}). \tag{21}$$

Then

$$R_{\mathrm{E}}(M) = F_c + \frac{\langle B, L_{\mathrm{E}} \rangle}{F_c M} + O(M^{-2}), \tag{22}$$

$$R_{\mathrm{H}}(M) = F_c + \frac{4\langle B, L_{\mathrm{H}} \rangle}{F_c M^2} + O(M^{-3}). \tag{23}$$

If $\langle B, L_{\mathrm{E}} \rangle > 0$ (respectively < 0), Heun (respectively Euler) has smaller target error for all sufficiently large allowed budgets. Yet $\mathcal{W}_\gamma$ is bounded for every $\gamma \in (0, 1)$. More precisely, if $D_{\mathrm{E}} \leq a/M$ and $D_{\mathrm{H}} \leq b/M^2$, every $M \in \mathcal{W}_\gamma$ satisfies

$$M \leq \frac{(1-\gamma)a + \sqrt{(1-\gamma)^2 a^2 + 4\gamma F_c b}}{2\gamma F_c}. \tag{24}$$

Corollary 6 (Normal-order separation of numerical and target rankings)**.** *Under the assumptions of Theorem 5, suppose $L_{\mathrm{E}} \neq 0$. Then, for all sufficiently large even M,*

$$D_{\mathrm{H}}(M) < D_{\mathrm{E}}(M). \tag{25}$$

If $\langle B, L_{\mathrm{E}} \rangle < 0$, the same fixed learned field also satisfies

$$D_{\mathrm{H}}(M) < D_{\mathrm{E}}(M) \quad \textit{and} \quad R_{\mathrm{E}}(M) < R_{\mathrm{H}}(M) \tag{26}$$

for every sufficiently large even M. If $\langle B, L_{\mathrm{E}} \rangle > 0$, the first inequality remains true while the target ranking is $R_{\mathrm{H}}(M) < R_{\mathrm{E}}(M)$. In either case, $R_{\mathrm{E}}(M)/R_{\mathrm{H}}(M) \to 1$. The Euler error has exact order one; if additionally $L_{\mathrm{H}} \neq 0$, Heun has exact order two. Otherwise its error is of still higher order under Equation 21.

If $L_{\mathrm{E}} \neq 0$, the theorem's alignment statement is literal:

$$\rho_{\mathrm{E}}(M) \longrightarrow \frac{\langle B, L_{\mathrm{E}} \rangle}{F_c \|L_{\mathrm{E}}\|_{L_2}}. \tag{27}$$

Thus "eventually smaller" and "smaller by a fixed percentage" are different claims. When Euler approaches the continuous endpoint from the target side, its $O(M^{-1})$ error can remain beneficial even though $D_{\mathrm{H}} \ll D_{\mathrm{E}}$; both target errors still converge to F_c, so their ratio converges to one. The same leading-direction argument applies to any unequal orders $p < q$ when the order-p leading defect is nonzero. Appendix F proves the theorem and gives the leading defect functions for affine flows.

The distinction from an effective-order or order-collapse explanation is important. Here the learned field is fixed, and the theorem assumes the ordinary global orders $D_{\mathrm{E}} = O(M^{-1})$ and $D_{\mathrm{H}} = O(M^{-2})$ rather than showing that either order deteriorates. Corollary 6 makes the joint claim explicit: if $\langle B, L_{\mathrm{E}} \rangle < 0$, the same fixed model has $D_{\mathrm{H}} < D_{\mathrm{E}}$ but $R_{\mathrm{E}} < R_{\mathrm{H}}$ for all sufficiently large allowed budgets. This is a target-ranking statement, not a claim that Heun has lost second-order convergence.

C Conventions and Continuous Affine Flow

We collect the sign conventions because probability-flow derivations are especially vulnerable to a missing time reversal or factor of two. For Equation 4, the forward drift and diffusion variance are $f(t, x) = -\beta x$ and $g^2(t) = 2\beta$. The forward probability-flow ODE is

$$\dot{x}_t = f(t, x_t) - \tfrac{1}{2} g^2(t) \widehat{s}_t(x_t) = -\beta x_t - \beta \widehat{s}_t(x_t). \tag{28}$$

Substituting Equation 6 and then using $Y_u = x_{T-u}$ gives Equation 7. Under exact score, $q_\star(u) = \beta(1 - v_{T-u}^{-1})$ and $r_\star = 0$.

For the general affine reverse ODE, define the transition factor

$$\Phi(u, s) := \exp\left(\int_s^u q(w)\, \mathrm{d}w \right). \tag{29}$$

Variation of constants gives

$$Y_u = \Phi(u, 0) Z + \int_0^u \Phi(u, s) r(s)\, \mathrm{d}s. \tag{30}$$

The exact-score scale follows from

$$\int_0^T \beta(1 - v_{T-u}^{-1})\, \mathrm{d}u = \frac{1}{2} \log \frac{\lambda}{v_T}, \tag{31}$$

and is $k_\star = \sqrt{\lambda / v_T}$. For the approximate score, splitting off $\beta\delta$ yields Equation 8. In the offset integral, change variables from u to $t = T - u$ and use the propagation from physical time t to zero:

$$\Phi(T, T - t) = \sqrt{\lambda / v_t} \exp\left(\beta \int_0^t \delta(a)\, \mathrm{d}a \right). \tag{32}$$

This proves Equation 9.

D Distributional Extension without Target Pairing

The coupled identity is not the only setting in which the signed interaction is exact. Let k be a positive-definite kernel with reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, and let $\mu_k(P)$ denote the kernel mean embedding of a distribution P (Gretton et al., 2012). Write $P^\star$ for the target distribution, P_θ for the endpoint distribution of the continuous approximate-score flow, and P_m^M for its numerical endpoint. Define

$$B_k := \mu_k(P_\theta) - \mu_k(P^\star), \qquad E_{m,k} := \mu_k(P_m^M) - \mu_k(P_\theta). \tag{33}$$

Theorem 7 (Distributional alignment and finite-sample certificate)**.** *Suppose* $\mathbb{E}_{X\sim P}\sqrt{k(X,X)} < \infty$ *for every distribution above. Then*

$$\mathrm{MMD}_k^2(P_m^M, P^\star) = \|B_k\|_{\mathcal{H}}^2 + \|E_{m,k}\|_{\mathcal{H}}^2 + 2\langle B_k, E_{m,k}\rangle_{\mathcal{H}}. \tag{34}$$

Consequently, the decision decomposition in Equation 3 holds verbatim with RKHS quantities and requires no pairing between target and generated samples.

If additionally $k(x,x) \le \kappa^2$*, let empirical mean embeddings be formed from* n_j *independent observations from each of* J *distributions and put*

$$\epsilon_j := \frac{\kappa}{\sqrt{n_j}}\left(1 + \sqrt{2\log(J/\delta)}\right). \tag{35}$$

For two numerical distributions P_a, P_b*, with probability at least* $1-\delta$*,*

$$|\widehat{\Delta}_{b,a} - \Delta_{b,a}| \le 4\kappa(\epsilon_a + \epsilon_b + 2\epsilon_\star), \tag{36}$$

$$\widehat{\Delta}_{b,a} := \widehat{\mathrm{MMD}}_k^2(P_b, P^\star) - \widehat{\mathrm{MMD}}_k^2(P_a, P^\star), \tag{37}$$

$$\Delta_{b,a} := \mathrm{MMD}_k^2(P_b, P^\star) - \mathrm{MMD}_k^2(P_a, P^\star). \tag{38}$$

Thus an empirical gap whose magnitude exceeds the right-hand side certifies its population sign. The bound is conservative rather than sample-optimal.

The population identity applies to any kernel; a characteristic kernel makes MMD a metric on the relevant distribution class. Kernel Inception Distance is a polynomial-kernel MMD on Inception features (Bińkowski et al., 2018), but the identity does not extend to FID, whose covariance term has Bures geometry. The finite-sample certificate is more restrictive: a polynomial kernel is not globally bounded on an unbounded feature space, so applying the stated McDiarmid bound to KID requires a uniform population feature-support bound (or an explicitly imposed feature clipping/truncation rule). Our synthetic polynomial-kernel gate is therefore interpreted through its block bootstrap intervals rather than as an unconditional application of the bounded-kernel certificate. Appendix D.1 proves the theorem and shows that changing the intermediate reference distribution shifts numerical saving and alignment penalty by the same amount, leaving their decision difference invariant.

The certificate uses ordinary empirical mean embeddings (a V-statistic MMD). For lower finite-sample bias, our exploratory experiment also reports the corresponding U-statistic decomposition: equal-index cross terms are omitted between solver outputs generated from the same terminal latent. The algebraic decision identity remains exact for that estimator, while uncertainty is computed from independent sample blocks rather than from dependent entries of the kernel matrix.

D.1 Proofs and reference invariance

Proof of Theorem 7. The integrability assumption ensures that every Bochner mean embedding exists. By Equation 33,

$$\mu_k(P_m^M) - \mu_k(P^\star) = B_k + E_{m,k}. \tag{39}$$

Taking the squared RKHS norm proves Equation 34; subtracting the identities for two solvers gives the decision decomposition.

For the certificate, first consider one distribution P. Its empirical mean embedding $\widehat{\mu}_P = n^{-1}\sum_{i=1}^n k(X_i,\cdot)$ satisfies

$$\mathbb{E}\|\widehat{\mu}_P - \mu_P\|_{\mathcal{H}} \le \frac{\kappa}{\sqrt{n}}. \tag{40}$$

Replacing one observation changes this norm by at most $2\kappa/n$. McDiarmid's inequality and a union bound over the J distributions therefore give $\|\widehat{\mu}_j - \mu_j\|_{\mathcal{H}} \le \epsilon_j$ simultaneously with probability at least $1-\delta$. Because all population and empirical mean embeddings have norm at most κ,

$$\left|\|\widehat{\mu}_P - \widehat{\mu}_Q\|_{\mathcal{H}}^2 - \|\mu_P - \mu_Q\|_{\mathcal{H}}^2\right| \le 4\kappa(\epsilon_P + \epsilon_Q). \tag{41}$$

Applying this bound to $(P_b, P^\star)$ and $(P_a, P^\star)$ and adding the two errors proves Equation 36. □

Lemma 8 (Reference invariance and numerical-fidelity audit)**.** *Let* Q *be any intermediate reference distribution with mean embedding* μ_Q*. Defining* $B_Q = \mu_Q - \mu_\star$ *and* $E_{m,Q} = \mu_m - \mu_Q$ *leaves the target decision gap unchanged. If* $\|\mu_Q - \mu_\theta\| \le \eta$*, then*

$$\left|\left(\|E_{a,Q}\|^2 - \|E_{b,Q}\|^2\right) - \left(\|E_{a,k}\|^2 - \|E_{b,k}\|^2\right)\right| \le 2\eta\|E_{a,k} - E_{b,k}\|. \tag{42}$$

Proof. For any Q, $B_Q + E_{m,Q} = \mu_m - \mu_\star$, so the decomposed target gap is identical. Put $z = \mu_Q - \mu_\theta$. Then $E_{m,Q} = E_{m,k} - z$, and expanding the two squared norms proves Equation 42. The alignment penalty changes by exactly the same signed amount, as it must because their difference is reference invariant. □

E Proof of the Exact Equal-NFE Geometry

Proof of Theorem 4. Suppose the current numerical state is $Y_n = k_n Z + m_n$. Euler applies

$$Y_{n+1} = Y_n + h(q_n Y_n + r_n) = (1 + hq_n)Y_n + hr_n, \tag{43}$$

which gives Equation 18 by matching the coefficient of Z and the constant term.

For Heun, the predictor is

$$\widetilde{Y}_{n+1} = (1 + hq_n)Y_n + hr_n. \tag{44}$$

The corrector is

$$Y_{n+1} = Y_n + \frac{h}{2}\big(q_n Y_n + r_n + q_{n+1}\widetilde{Y}_{n+1} + r_{n+1}\big) \tag{45}$$

$$= \left[1 + \frac{h}{2}(q_n + q_{n+1}) + \frac{h^2}{2} q_{n+1} q_n\right] Y_n + \frac{h}{2}(r_n + r_{n+1} + hq_{n+1} r_n), \tag{46}$$

which is Equation 19. Induction from $k_0 = 1, m_0 = 0$ proves that both endpoints are affine in the same Z.

Now

$$B = (k_\theta - k_\star)Z + m_\theta, \tag{47}$$

$$E_m = (k_m - k_\theta)Z + (m_m - m_\theta). \tag{48}$$

Because $\mathbb{E}Z = 0$ and $\mathbb{E}Z^2 = v_T$, expanding $\mathbb{E}B^2$, $\mathbb{E}E_m^2$, and $\mathbb{E}[BE_m]$ gives Equation 10; expanding $\mathbb{E}(B + E_m)^2$ gives Equation 2.

Finally, for one-dimensional Gaussians,

$$W_2^2(\mathcal{N}(m_1, \sigma_1^2), \mathcal{N}(m_2, \sigma_2^2)) = (m_1 - m_2)^2 + (\sigma_1 - \sigma_2)^2. \tag{49}$$

The numerical standard deviation is $|k_m|\sqrt{v_T}$ and the target standard deviation is $k_\star\sqrt{v_T} = \sqrt{\lambda}$, proving Equation 11. If $k_m \geq 0$, the shared-Z map is monotone and is the optimal quantile coupling. If $k_m < 0$, it reverses orientation and need not be optimal. The same argument shows that D_m equals the W_2 distance to the approximate continuous endpoint when both k_m and k_θ are nonnegative. More precisely,

$$R_m^2 - W_2^2 = 4\sqrt{\lambda v_T}\,(-k_m)_+, \tag{50}$$

so the coupled RMS always upper-bounds this one-dimensional W_2 and equals it exactly in the nonnegative-scale regime. □

F Asymptotic Defects, Ranking, and Worth

F.1 Leading global defects

Let $y' = q(t)y + r(t)$ on $[0, T]$, with $q, r \in C^3$, and let $y(t; z)$ be the exact trajectory from z. A one-step Taylor expansion gives the Euler numerical-minus-exact local defect

$$-\frac{h^2}{2} y''(t; z) + O(h^3). \tag{51}$$

Transporting local defects by $\Phi(T, t)$ and summing the Riemann sum yields

$$Y_{\mathrm{E}}(T) - Y(T) = h\widetilde{L}_{\mathrm{E}}(z) + O_{L_2}(h^2), \qquad \widetilde{L}_{\mathrm{E}}(z) = -\frac{1}{2}\int_0^T \Phi(T, t) y''(t; z)\,\mathrm{d}t. \tag{52}$$

For Heun, expanding the predictor and corrector to third order gives local defect $h^3\ell_{\mathrm{H}}(t,y)+O(h^4)$ with

$$\ell_{\mathrm{H}}(t,y)=\left(\frac{q''}{12}-\frac{q^3}{6}\right)y+\frac{r''}{12}+\frac{q'r}{6}-\frac{q^2r}{6}-\frac{qr'}{6}. \tag{53}$$

Therefore

$$Y_{\mathrm{H}}(T)-Y(T)=h^2\widetilde{L}_{\mathrm{H}}(z)+O_{L_2}(h^3),\qquad \widetilde{L}_{\mathrm{H}}(z)=\int_0^T\Phi(T,t)\ell_{\mathrm{H}}(t,y(t;z))\,\mathrm{d}t. \tag{54}$$

The standard stability and Riemann-sum argument is uniform over z in L_2 because each exact and numerical endpoint is affine in Gaussian z and the coefficients and derivatives are bounded. With $h_{\mathrm{E}}=T/M$ and $h_{\mathrm{H}}=2T/M$, Equations 52–54 have the equal-NFE form in Equation 21 with $L_{\mathrm{E}}=T\widetilde{L}_{\mathrm{E}}$ and $L_{\mathrm{H}}=T^2\widetilde{L}_{\mathrm{H}}$. If a leading coefficient vanishes, the method is superconvergent for that trajectory and the nominal leading order must be replaced; none of our theorems assumes it is nonzero unless an asymptotic equivalent is stated.

F.2 Proof of Theorem 5

Proof. The L_2 norm is Fréchet differentiable at every nonzero B, and for $e\to 0$,

$$\|B+e\|_{L_2}=\|B\|_{L_2}+\frac{\langle B,e\rangle_{L_2}}{\|B\|_{L_2}}+O(\|e\|_{L_2}^2). \tag{55}$$

Substituting the two expansions in Equation 21 proves Equations 22 and 23. A nonzero M^{-1} term dominates any M^{-2} term, giving the eventual ranking.

Both consistent methods converge to the same approximate endpoint, so $R_{\mathrm{E}}(M),R_{\mathrm{H}}(M)\to F_c$. For $\gamma>0$,

$$R_{\mathrm{H}}(M)-(1-\gamma)R_{\mathrm{E}}(M)\longrightarrow\gamma F_c>0, \tag{56}$$

which proves that $\mathcal{W}_\gamma$ is bounded.

For the quantitative statement, the reverse and ordinary triangle inequalities give

$$F_c-D_{\mathrm{H}}\le R_{\mathrm{H}}\le(1-\gamma)R_{\mathrm{E}}\le(1-\gamma)(F_c+D_{\mathrm{E}}). \tag{57}$$

Hence every worth budget satisfies

$$\gamma F_c\le D_{\mathrm{H}}+(1-\gamma)D_{\mathrm{E}}\le b/M^2+(1-\gamma)a/M. \tag{58}$$

Multiplying by M^2 and taking the positive root of the resulting quadratic proves Equation 24. □

F.3 Proof of Corollary 6

Proof. From Equation 21 and $L_{\mathrm{E}}\neq 0$,

$$D_{\mathrm{E}}(M)=\frac{\|L_{\mathrm{E}}\|_{L_2}}{M}+O(M^{-2}),\qquad D_{\mathrm{H}}(M)=O(M^{-2}). \tag{59}$$

Hence $D_{\mathrm{H}}(M)/D_{\mathrm{E}}(M)\to 0$, proving the strict numerical inequality for all sufficiently large even M. The two alternatives for the target ranking are exactly the sign alternatives in Theorem 5; the ratio conclusion follows from $R_{\mathrm{E}}(M),R_{\mathrm{H}}(M)\to F_c>0$. □

G Nonlinear Local Stability of Normal-Order Reversal

G.1 Uniform leading-defect expansions

We prove Proposition 2. All parameters $a,q,T,v_T,k_\star$ are fixed as in Theorem 1. The neighborhood is in the affine function space $qx+C_b^5([0,T]\times\mathbb{R})$; the unbounded linear base field is not itself required to lie in C_b^5. Let $X_g(u;z)$ solve $X_g'=f_g(u,X_g)$ from z, and write

$$B_g=X_g(T;Z)-k_\star Z,\qquad F_g=\|B_g\|_{L_2},\qquad Z\sim\mathcal{N}(0,v_T).$$

The numerical errors below are relative to $X_g(T;Z)$, not the target. We use $A_{\mathrm{H}} = 4L_{\mathrm{H}}$ to absorb the equal-NFE factor in Equation 21.

For $f = f_g$, define its material acceleration and a second-derivative term by

$$\mathcal{A} = f_u + f_x f, \qquad \mathcal{K} = f_{uu} + 2f_{ux}f + f_{xx}f^2. \tag{60}$$

The exact flow's third derivative is $\mathcal{K} + f_x\mathcal{A}$. Taylor expansion of a numerical step minus an exact step gives $h^2 d_{\mathrm{E}} + O(h^3)$ for Euler and $h^3 d_{\mathrm{H}} + O(h^4)$ for Heun, where

$$d_{\mathrm{E}} = -\mathcal{A}/2, \qquad d_{\mathrm{H}} = \mathcal{K}/12 - f_x\mathcal{A}/6. \tag{61}$$

Indeed Heun has third-order term $h^3\mathcal{K}/4$, fixing the sign in the second expression. Define

$$\begin{aligned} J_g(T,u;z) &= \exp\left(\int_u^T f_x(s, X_g(s;z))\,\mathrm{d}s\right), \\ A_{\mathrm{E}}(g)(z) &= T\int_0^T J_g(T,u;z) d_{\mathrm{E}}(u, X_g(u;z))\,\mathrm{d}u, \\ A_{\mathrm{H}}(g)(z) &= 4T^2\int_0^T J_g(T,u;z) d_{\mathrm{H}}(u, X_g(u;z))\,\mathrm{d}u. \end{aligned} \tag{62}$$

The factor four is necessary because Heun uses step size $2T/M$.

We first establish that, on any fixed ball $\|g\|_{C_b^5} \le \varepsilon \le 1$, constants independent of g give

$$\|E_{\mathrm{E}}^g(M) - A_{\mathrm{E}}(g)/M\|_{L_2} \le K_{\mathrm{E}}/M^2, \qquad \|E_{\mathrm{H}}^g(M) - A_{\mathrm{H}}(g)/M^2\|_{L_2} \le K_{\mathrm{H}}/M^3. \tag{63}$$

Here and below bounds hold for sufficiently small steps with a uniform threshold. The fields have uniform linear growth and global spatial Lipschitz constant at most $|q| + 1$. Exact trajectories, numerical trajectories, and internal stages are bounded by $C(1 + |z|)$ on $[0, T]$. Their local Taylor remainders are bounded by a fixed polynomial in $|z|$ times the required power of h, uniformly in g.

Stability constants independent of the initial state. Set $L = |q| + 1$. Denote a step starting at time u by $\Psi_m^{u,h}$. The Euler predictor has Lipschitz factor $1 + hL$. Applying the same spatial bound at time $u + h$ to the two Heun predictors gives

$$\begin{aligned} &\mathrm{Lip}(\Psi_{\mathrm{E}}^{u,h}) \le 1 + hL \le e^{hL}, \\ &\mathrm{Lip}(\Psi_{\mathrm{H}}^{u,h}) \le 1 + \tfrac{hL}{2}\{1 + (1 + hL)\} = 1 + hL + \tfrac{1}{2}h^2L^2 \le e^{hL}. \end{aligned} \tag{64}$$

The product of propagation factors over steps of total length at most T is thus bounded by e^{LT}, uniformly in g and z. For a uniform mesh $h = T/N$, let $e_n = X_{m,n} - X_g(u_n;z)$ be the numerical grid error and $\tau_{m,n} = \Psi_m^{u_n,h}(X_g(u_n;z)) - X_g(u_{n+1};z)$. Uniform smoothness and the trajectory bound give $|\tau_{m,n}| \le C_{\mathrm{loc}} h^{p+1} P(|z|)$, where $p = 1$ for Euler and $p = 2$ for Heun. Starting from $e_0 = 0$, the unexpanded stability recursion yields

$$|e_n| \le \sum_{j=0}^{n-1} e^{L(n-1-j)h}|\tau_{m,j}| \le C_{\mathrm{loc}} T e^{LT} h^p P(|z|). \tag{65}$$

The only initial-state dependence is polynomial; no factor $\exp(CP(|z|))$ occurs. Spatial Lipschitz continuity supplies the propagation bound, while the C_b^5 hypothesis supplies the local defect and remainder expansions. The former does not replace the latter or extend this proposition to arbitrary Lipschitz fields.

For completeness, the one-step recursion and its expansion about the exact trajectory imply

$$\begin{aligned} e_{n+1} = [1 + hf_x(u_n, X_g(u_n;z))]e_n + h^{p+1}d_m(u_n, X_g(u_n;z)) \\ + O\big(h^2|e_n|P(|z|) + h|e_n|^2P(|z|) + h^{p+2}P(|z|)\big), \end{aligned} \tag{66}$$

where P can change between bounds but is always a fixed polynomial. Substitute the independently obtained bound Equation 65 into this remainder, rather than applying Gronwall to a coefficient

involving $P(|z|)$. The accumulated remainder is $O(h^{p+1}P(|z|))$. The leading linear recursion is the Riemann-sum approximation to

$$U'_m = f_x(u, X_g(u;z))U_m + d_m(u, X_g(u;z)), \qquad U_m(0) = 0,$$

with uniform $O(hP(|z|))$ error after division by h^p. Variation of constants yields Equation 62.

To make the initial-state integrability explicit, put $S = 1 + |z|$ and restrict $h \le h_0 := \min\{1, T\}$. Along the exact trajectory, $|X'| + |X''| \le CS$, $|X'''| \le CS^2$, and $|X^{(4)}| \le CS^3$. Lower-order Taylor bounds give $|\tau_{\mathrm{E},n}| \le Ch^2S$ and $|\tau_{\mathrm{H},n}| \le Ch^3S^2$, hence Equation 65 gives $|e_{\mathrm{E},n}| \le ChS$ and $|e_{\mathrm{H},n}| \le Ch^2S^2$. No condition on $h|z|$ is needed: all derivatives of g used in these bounds are globally bounded. The step maps satisfy $|\partial_x^2 \Psi_m^{u,h}| \le Ch$, and, on the exact trajectory, $|\partial_x \Psi_{\mathrm{H}}^{u,h} - (1 + hf_x)| \le Ch^2S$. Thus the residual in Equation 66 is bounded by Ch^3S^2 for Euler and $C(h^4S^3 + h^5S^4)$ for Heun. The coefficient equations above satisfy $|U''_{\mathrm{E}}| \le CS^2$ and $|U''_{\mathrm{H}}| \le CS^3$. Comparing the normalized error e_n/h^p with $U_m(u_n)$ using the same propagation factor e^{LT} yields

$$\begin{aligned} |e_{\mathrm{E},N} - hU_{\mathrm{E}}(T)| &\le Ch^2S^2, \\ |e_{\mathrm{H},N} - h^2U_{\mathrm{H}}(T)| &\le C(h^3S^3 + h^4S^4) \le Ch^3S^4. \end{aligned} \tag{67}$$

All constants here are independent of g, z, h on the unit perturbation ball. If $\mu_r = \|(1 + |Z|)^r\|_{L_2}$, then $\mu_r^2 \le 2^{2r-1}\{1 + (2r-1)!!\, v_T^r\} < \infty$. Using $h_{\mathrm{E}} = T/M$, $h_{\mathrm{H}} = 2T/M$, $A_{\mathrm{E}} = TU_{\mathrm{E}}(T)$ and $A_{\mathrm{H}} = 4T^2U_{\mathrm{H}}(T)$ proves Equation 63; one may take $K_{\mathrm{E}} = C_{\mathrm{E}}T^2\mu_2$ and $K_{\mathrm{H}} = 8C_{\mathrm{H}}T^3\mu_4$. The fourth and eighth Gaussian moments suffice for these conservative remainder bounds. The C_b^5 assumption remains a convenient sufficient condition, not a claimed minimal one.

G.2 A NEGATIVE ALIGNMENT MARGIN ON THE ENTIRE NEIGHBORHOOD

At $g = 0$, the coefficients are

$$B_0 = (e^a - k_\star)Z, \qquad A_{\mathrm{E}}(0) = -\tfrac{1}{2}e^a a^2 Z, \qquad A_{\mathrm{H}}(0) = -\tfrac{2}{3}e^a a^3 Z. \tag{68}$$

All three have nonzero norm and

$$\alpha_0 := \langle B_0, A_{\mathrm{E}}(0)\rangle = -\tfrac{1}{2}v_T(e^a - k_\star)e^a a^2 < 0. \tag{69}$$

Write $\gamma_0 = -\alpha_0 > 0$.

If $\|g\|_{C_b^5} \le \varepsilon \le 1$, variation of constants gives

$$X_g(u;z) - e^{qu}z = \int_0^u e^{q(u-s)} g(s, X_g(s;z))\,\mathrm{d}s.$$

Since $q < 0$, the absolute difference is at most $T\varepsilon$, uniformly in z. Also

$$J_g(T, u; z) = e^{q(T-u)} \exp\left(\int_u^T g_x(s, X_g(s;z))\,\mathrm{d}s\right),$$

so $|J_g - J_0| \le C\varepsilon$. The difference of the acceleration terms along the two trajectories is bounded by $C\varepsilon(1 + |z|)$; that of the $\mathcal{K}$ terms is bounded by $C\varepsilon(1 + |z|^2)$. Applying these estimates in Equation 62 gives

$$\|B_g - B_0\|_{L_2} + \|A_{\mathrm{E}}(g) - A_{\mathrm{E}}(0)\|_{L_2} + \|A_{\mathrm{H}}(g) - A_{\mathrm{H}}(0)\|_{L_2} \le C\varepsilon. \tag{70}$$

Cauchy–Schwarz also yields $|\alpha_g - \alpha_0| \le C'\varepsilon$ for $\alpha_g = \langle B_g, A_{\mathrm{E}}(g)\rangle$. Choosing ε_0 sufficiently small, uniformly on the resulting ball we have

$$F_g \ge F_0/2 > 0, \qquad \|A_{\mathrm{E}}(g)\|_{L_2} \ge e_{\min} > 0, \qquad \|A_{\mathrm{H}}(g)\|_{L_2} \ge h_{\min} > 0, \qquad \alpha_g \le -\gamma_0/2. \tag{71}$$

These quantities also have finite uniform upper bounds. The norm expansions in Equation 15 follow from Equation 63, with $c_m(g) = \|A_m(g)\|_{L_2}$.

Expanding the squared target norms now gives, for a uniform constant K_R,

$$(R_{\mathrm{H}}^g)^2 - (R_{\mathrm{E}}^g)^2 = -2\alpha_g/M + O(M^{-2}) \ge \gamma_0/M - K_R/M^2. \tag{72}$$

This is positive for $M \geq 2K_R/\gamma_0$. For the numerical ranking, let $h_{\max}$ bound $\|A_{\mathrm{H}}(g)\|_{L_2}$. Taking $M \geq \max\{1, 2K_{\mathrm{E}}/e_{\min}\}$ and $M > 2(h_{\max} + K_{\mathrm{H}})/e_{\min}$ ensures

$$D_{\mathrm{E}} \geq e_{\min}/(2M) > (h_{\max} + K_{\mathrm{H}})/M^2 \geq D_{\mathrm{H}}.$$

Finally $R_{\mathrm{H}} + R_{\mathrm{E}} = 2F_g + O(M^{-1})$ uniformly, so

$$R_{\mathrm{H}} - R_{\mathrm{E}} = -\frac{\alpha_g}{F_g M} + O(M^{-2}), \qquad R_{\mathrm{E}}/R_{\mathrm{H}} \longrightarrow 1. \tag{73}$$

Thus $c_R(g) = -\alpha_g/F_g$ has a positive uniform lower bound. Continuity at a fixed M alone would not prove a common threshold for all large budgets: the gap shrinks as M^{-1}. The uniform margin and remainders establish the required quantifiers.

G.3 Monotone maps and the distributional conclusion

Set $L = |q| + \varepsilon_0$. The exact flow is strictly increasing in z because $J_g > 0$. Euler's one-step derivative is at least $1 - hL$. Writing

$$A = f_x(u, x), \qquad B = f_x(u + h, x + hf(u, x)),$$

Heun's derivative is $1 + \frac{h}{2}[A + B(1 + hA)] \geq 1 - hL - (hL)^2/2$. It is positive whenever $hL \leq 1/2$. Therefore $M \geq 4TL$ suffices for both numerical endpoint maps to be increasing. The target map $k_\star z$ is increasing as well. In one dimension their common-latent coupling is the optimal quantile coupling, giving exactly

$$D_m = W_2(P_m^M, P_g), \qquad R_m = W_2(P_m^M, P^\star).$$

For an explicit aggregation of the thresholds, set $M_{\mathrm{exp}} = 2\lceil T/h_0 \rceil$, so both step sizes lie in the expansion range. Let $F_{\min} = \inf_g F_g > 0$, $a_{\max} = \sup_g \|A_{\mathrm{E}}(g)\|_{L_2}$, and $Q = a_{\max} + K_{\mathrm{E}} + h_{\max} + K_{\mathrm{H}}$. The threshold $M_{\mathrm{den}} = \max\{1, 2Q/F_{\min}\}$ keeps the target-norm denominators uniformly away from zero. Choose any even M_0 strictly greater than

$$\max\left\{M_{\mathrm{exp}},\, M_{\mathrm{den}},\, \frac{2K_R}{\gamma_0},\, \frac{2K_{\mathrm{E}}}{e_{\min}},\, \frac{2(h_{\max} + K_{\mathrm{H}})}{e_{\min}},\, 4T(|q| + \varepsilon_0)\right\}. \tag{74}$$

Every quantity is finite and independent of g in the chosen neighborhood. This is an existence bound, not a numerically certified radius or budget for the experiments, and its constants may depend on the fixed base-model parameters. This proves Proposition 2.

The associated approximate VP score is

$$\widehat{s}_g(t, x) = (q/\beta - 1)x + g(T - t, x)/\beta.$$

Its linear coefficient is negative and g is bounded, so at each time it is the score of a normalizable density with a negative quadratic log-density tail. We do not require these approximate densities to be the marginals of a single exact VP forward process. Nor does the proof establish a high-dimensional optimal coupling, stability in training L_2 error alone, or a numerically certified radius ε_0.

G.4 Independent nonlinear numerical check

For $T = \beta = 1$, $\lambda = 0.5$, $a = -0.2$, we use

$$g(u, x) = \eta\{[1 + 0.25\sin(2\pi u)]\tanh x + 0.25\sin(\pi u)\}. \tag{75}$$

Each amplitude defines a single fixed field for all seven budgets $32, 64, \ldots, 2048$. Direct Euler/Heun integrations are checked against a separate RK4 integration of the exact-flow and coefficient equations. Table 3 reports alignment and the last doubling-interval orders. All 35 comparisons satisfy both strict inequalities. This finite check verifies formulas in one nonlinear direction; it does not prove a function-neighborhood statement or certify that the amplitude equals its C_b^5 norm.

Expectations use 96-point Gauss–Hermite quadrature, checked against 48 points. The largest change among D_m, R_m is 5.812×10^{-11}. The augmented RK4 reference uses 8192 *steps*, checked against 4096 steps (four stages per step, not the same NFE convention as the GMM references); the maximum endpoint RMS change is 5.015×10^{-15}. At the last budget the relative discrepancies of ME_{E} and $M^2 E_{\mathrm{H}}$ from their transported coefficients are at most 1.450×10^{-4}. The zero-amplitude coefficients are additionally checked against Equation 68.

Table 3: Fixed nonlinear fields: negative alignment and normal observed orders on 1024–2048 NFE. All seven tested budgets reverse the target ranking in every row.

η	$\langle B_g, A_{\mathrm{E}}(g)\rangle$	$\widehat{p}_{\mathrm{E}}$	$\widehat{p}_{\mathrm{H}}$
-0.02	-0.00125452	1.000091	2.000042
-0.01	-0.00129047	1.000090	2.000102
0.00	-0.00131923	1.000087	2.000213
0.01	-0.00134108	1.000082	2.000084
0.02	-0.00135631	1.000074	2.000029

H Perfect Cancellation and Constant-Field Geometry

H.1 Pointwise cancellation constructions

The ranking reversal is not limited to numerical search.

Proposition 9 (Perfect cancellation under a smooth VP score perturbation)**.** *Fix $\lambda \in (0,1)$ and any sufficiently large even budget M_0. There is a smooth, purely multiplicative perturbation $\delta(t)$ for which M_0-step Euler has zero target W_2 error, $\rho_{\mathrm{E}} = -1$, and Heun has positive target error with $\rho_{\mathrm{H}} = 1$, even while $D_{\mathrm{H}} < D_{\mathrm{E}}$. There is another smooth perturbation for which Heun is exact, $\rho_{\mathrm{H}} = -1$, and Euler has positive error with $\rho_{\mathrm{E}} = 1$. Both strict rankings persist on an open neighborhood of perturbations.*

Unlike Theorem 1, this proposition allows the perturbation to depend on M_0 and asserts exact cancellation only at that budget.

For the construction, choose

$$\delta_q(t) = q/\beta - 1 + v_t^{-1}, \qquad \zeta(t) = 0, \tag{76}$$

which turns Equation 7 into $Y' = qY$. Setting

$$q_{\mathrm{E}} = \frac{M_0}{T}\big(k_\star^{1/M_0} - 1\big), \tag{77}$$

$$q_{\mathrm{H}} = \frac{M_0}{2T}\left[-1 + \sqrt{2k_\star^{2/M_0} - 1}\right] \tag{78}$$

makes Euler or Heun, respectively, land exactly on $k_\star$. At $T = \beta = 1$, $\lambda = 0.5$, and $M_0 = 100$, the chosen-method residual is 7.5e-16 in double precision. In the Euler-exact construction, Heun's numerical error is only 0.004 times Euler's, but removing Euler's cancellation makes the target error worse. The full inequalities are proved in Appendix H.

H.2 Proof of Theorem 1

Proof. The perturbation in Equation 12 is smooth since $v_t > 0$. It gives $q(u) = a/T$, $r(u) = 0$, and the endpoint coefficients

$$k_\theta = e^a, \qquad k_{\mathrm{E}} = (1 + a/M)^M, \qquad k_{\mathrm{H}} = (1 + 2a/M + 2a^2/M^2)^{M/2}. \tag{79}$$

For sufficiently large M the Euler factor is positive. For $x < 0$, $1 + x < e^x < 1 + x + x^2/2$, where the first inequality is used only for a positive left side. Applying these inequalities at $x = a/M$ and $x = 2a/M$ yields $0 < k_{\mathrm{E}} < k_\theta < k_{\mathrm{H}}$. Since $k_{\mathrm{E}} \to e^a > k_\star$, eventually

$$0 < k_\star < k_{\mathrm{E}} < k_\theta < k_{\mathrm{H}}. \tag{80}$$

For example, putting $b = -a > 0$ and $g = a - \log k_\star > 0$, the bound

$$0 < a - M\log(1 + a/M) \le \frac{b^2}{2(M - b)} \quad (M > b)$$

shows $k_{\mathrm{E}} > k_\star$ whenever $M > b + b^2/(2g)$. This explicit threshold establishes the chain, not yet the numerical-error comparison.

The logarithmic expansions

$$M\log(1+a/M) = a - \frac{a^2}{2M} + O(M^{-2}),$$
$$\frac{M}{2}\log(1+2a/M+2a^2/M^2) = a - \frac{2a^3}{3M^2} + O(M^{-3}) \tag{81}$$

give Equation 13 and $D_{\mathrm{H}}/D_{\mathrm{E}} \to 0$. All endpoint laws are centered Gaussians obtained by positive scaling of $Z \sim \mathcal{N}(0, v_T)$, so their W_2 distances equal the absolute scale differences times $\sqrt{v_T}$. The chain and numerical-error ratio prove both strict inequalities. On the same budgets,

$$R_{\mathrm{H}} - R_{\mathrm{E}} = \sqrt{v_T}(k_{\mathrm{H}} - k_{\mathrm{E}}) = D_{\mathrm{E}} + D_{\mathrm{H}} \sim c_{\mathrm{E}}/M.$$

Both target errors converge to $F_c = \sqrt{v_T}(e^a - k_\star) > 0$, proving the ratio limit. □

The verification script evaluates this fixed model at $16, 32, \ldots, 2048$ NFE and saves the exact endpoint metrics alongside the trained-score results. This finite check verifies implementation; the quantifier over every sufficiently large even budget follows from the proof.

H.3 Proof of Proposition 9

Proof of Proposition 9. The perturbation in Equation 76 is smooth because $v_t > 0$ and makes $q(u) \equiv q$, $r(u) = 0$. Thus

$$k_\theta = e^{qT}, \quad k_{\mathrm{E}}(M) = (1+qT/M)^M, \quad k_{\mathrm{H}}(M) = \left(1 + \frac{2qT}{M} + \frac{1}{2}\left(\frac{2qT}{M}\right)^2\right)^{M/2}. \tag{82}$$

For $q = q_{\mathrm{E}}$ in Equation 77, $k_{\mathrm{E}}(M_0) = k_\star$ by construction. Since $q_{\mathrm{E}} < 0$ and $(1+x/n)^n < e^x$ for $x < 0$ in the positive-factor regime,

$$k_{\mathrm{E}}(M_0) = k_\star < k_\theta. \tag{83}$$

Let $z = 2q_{\mathrm{E}}T/M_0 < 0$. For sufficiently large M_0, $z \in (-1, 0)$ and

$$1 + z + z^2/2 > e^z. \tag{84}$$

Raising both sides to $M_0/2$ gives $k_{\mathrm{H}}(M_0) > k_\theta > k_\star$. The continuous-flow endpoint bias and Euler numerical error point in opposite coefficient directions, so $\rho_{\mathrm{E}} = -1$ and $R_{\mathrm{E}} = 0$; the Heun error points in the same direction as the bias, so $\rho_{\mathrm{H}} = 1$ and $R_{\mathrm{H}} > 0$. Taylor expansion gives $D_{\mathrm{H}} = O(M_0^{-2})$ while $D_{\mathrm{E}} = F_c = O(M_0^{-1})$, hence $D_{\mathrm{H}} < D_{\mathrm{E}}$ for sufficiently large M_0.

For the second construction, write $a = k_\star^{2/M_0}$. For sufficiently large M_0, $2a - 1 > 0$. Choosing $z = -1 + \sqrt{2a-1}$ solves $1 + z + z^2/2 = a$, so $q = q_{\mathrm{H}} = zM_0/(2T)$ gives $k_{\mathrm{H}}(M_0) = k_\star$. The standard inequalities for Euler and the second-order stability polynomial give

$$k_{\mathrm{E}}(M_0) < k_\theta < k_{\mathrm{H}}(M_0) = k_\star. \tag{85}$$

Thus Heun cancels the continuous-flow endpoint bias and Euler amplifies it. All strict inequalities and the endpoint coefficients depend continuously on δ, so each nonzero ranking persists under sufficiently small perturbations in the uniform norm. □

H.4 A convex constant-affine model

The same construction with constant r gives a continuous geometric taxonomy. Let $q < 0$, $a = r/q$, and consider $Y' = qY + r$. The continuous and numerical offsets obey

$$m_\theta = a(k_\theta - 1), \qquad m_m = a(k_m - 1). \tag{86}$$

Consequently

$$R_m^2 = G(k_m) := v_T(k_m - k_\star)^2 + a^2(k_m - 1)^2, \tag{87}$$

a strictly convex quadratic with minimizer

$$k_{\mathrm{opt}} = \frac{v_T k_\star + a^2}{v_T + a^2}. \tag{88}$$

Moreover

$$\tfrac{1}{2}G'(k_\theta) = v_T(k_\theta - k_\star) + a^2(k_\theta - 1) \tag{89}$$

is the continuous-bias inner product with the common coefficient direction of numerical error.

To make the budget paths explicit, put $x = -qT > 0$. On the shared positive-factor branch $M > x$,

$$k_{\mathrm{E}}(M) = (1 - x/M)^M, \qquad k_{\mathrm{H}}(M) = (1 - 2x/M + 2x^2/M^2)^{M/2}. \tag{90}$$

For Euler, with $y = x/M \in (0, 1)$,

$$\frac{\mathrm{d}}{\mathrm{d}M} \log k_{\mathrm{E}}(M) = \log(1 - y) + \frac{y}{1 - y} > 0. \tag{91}$$

For Heun, set $w = 2x/M \in (0, 2)$ and $P(w) = 1 - w + w^2/2$. Writing $\phi(w) = \log P(w)/w$, the numerator of $\phi'(w)$ is $N(w) = wP'(w)/P(w) - \log P(w)$, where

$$N(0) = 0, \qquad N'(w) = \frac{w^2(1 - w/2)}{P(w)^2} > 0. \tag{92}$$

Thus increasing M increases $k_{\mathrm{E}}(M)$ and decreases $k_{\mathrm{H}}(M)$, both toward $k_\theta = e^{-x}$. If k_{opt} lies on one method's covered coefficient interval, strict convexity of G makes that method's target error decrease until the path crosses k_{opt} and then increase toward F_c; otherwise it is monotone. This constant-field model gives a closed-form version of the non-monotone mechanism visible in Figure 3. We do not infer that a general time-varying worth set is a single interval.

I Stability, Asymptotic Onset, and Spectra

For exact score and $\lambda < 1$, the reverse-time scalar field is $q_\star(u) = \beta(1 - v_{T-u}^{-1}) \le 0$. We quantify finite-budget order by

$$\widehat{p}_{\mathrm{H}}(M) := \log_2\left(\frac{D_{\mathrm{H}}(M)}{D_{\mathrm{H}}(2M)}\right). \tag{93}$$

The reported observed-order onset is the first tested budget M for which both $\widehat{p}_{\mathrm{H}}(M)$ and $\widehat{p}_{\mathrm{H}}(2M)$ lie in $[1.8, 2.2]$. This is an operational two-interval criterion, not a proof that all later remainders are negligible.

The endpoint layer can be resolved by reparameterizing physical time with normalized log variance. With $\Delta = \log(v_T/\lambda)$, define

$$s(t) := \frac{\log(v_T/v_t)}{\Delta}, \qquad 0 \le s \le 1, \tag{94}$$

so $s = 0$ at terminal time and $s = 1$ at data time. A uniform s grid flattens the exact scalar coefficient to $-\Delta/2$ after time change. In multiple dimensions we construct one shared grid from $\lambda_{\min}$; it therefore regularizes the stiffest direction without flattening every eigendirection simultaneously.

Observed order is a finite diagnostic. From Equation 21, define the leading Heun coefficient $B_{\mathrm{H}} := \|L_{\mathrm{H}}\|_{L_2}$, so that $D_{\mathrm{H}}(M) = 4B_{\mathrm{H}}/M^2 + O(M^{-3})$ when $B_{\mathrm{H}} > 0$. A theoretical relative-remainder onset can then be defined by

$$M_{\mathrm{asymp,H}}(\eta) := \min\left\{M_0 : \left|\frac{M^2 D_{\mathrm{H}}(M)}{4B_{\mathrm{H}}} - 1\right| \le \eta \text{ for every even } M \ge M_0\right\}. \tag{95}$$

If $|D_{\mathrm{H}} - 4B_{\mathrm{H}}/M^2| \le K_{\mathrm{H}}/M^3$, then an even rounding of $K_{\mathrm{H}}/(4\eta B_{\mathrm{H}})$ is sufficient. If $B_{\mathrm{H}} = 0$, the nominal second-order onset is undefined and a higher leading order must be identified.

For a contractive scalar field $q(u) \le 0$, define stepwise contractivity as $|A_n| \le 1$ for every homogeneous step factor. If $L = \sup_u |q(u)|$, the standard real-axis conditions give the sufficient requirements

$$h_{\mathrm{E}}L \le 2 \quad \text{and} \quad h_{\mathrm{H}}L \le 2 \tag{96}$$

for Euler and Heun, respectively; Euler orientation preservation uses the stronger $h_{\mathrm{E}}L \le 1$. Under equal NFE these translate to different M thresholds because $h_{\mathrm{H}} = 2h_{\mathrm{E}}$. They are stability conditions, not proofs that the global leading term dominates its remainder.

Table 4: Search-selected scalar cases at 128 NFE. Similar scalar magnitudes do not determine the target-error ordering; the alignment signs reverse it.

Case	F_c	D_{E}	D_{H}	ρ_{E}	ρ_{H}	$R_{\mathrm{H}}/R_{\mathrm{E}}$
A	0.053	0.034	0.074	0.98	-0.95	0.33
B	0.047	0.030	0.052	-0.58	0.58	2.31

For commuting d-dimensional Gaussian covariance with eigenvalues λ_i, all endpoint maps separate by direction. The squared coupled quantities sum:

$$F_c^2 = \sum_i F_i^2, \quad D_m^2 = \sum_i D_{m,i}^2, \quad C_m = \sum_i C_{m,i}. \tag{97}$$

When all denominators are nonzero,

$$\rho_m = \frac{\sum_i \rho_{m,i} F_i D_{m,i}}{F_c D_m}, \tag{98}$$

so aggregate alignment is not an unweighted average of spectral alignments. For exact scores $F_c = 0$, we use only D_m and observed order.

To verify Equation 94, let $\Delta = \log(v_T/\lambda) > 0$. Along reverse time,

$$\frac{\mathrm{d}v}{\mathrm{d}u} = -2\beta(1-v), \qquad \frac{\mathrm{d}u}{\mathrm{d}s} = \frac{\Delta v}{2\beta(1-v)}. \tag{99}$$

Multiplying by $q_\star = -\beta(1-v)/v$ gives $q_\star \, \mathrm{d}u/\mathrm{d}s = -\Delta/2$, as claimed.

J Complete Experimental Protocol

J.1 Software, NFE accounting, and checks

The archived experiments use Python 3.11, NumPy, and Matplotlib; the fixed-checkpoint extension was reproduced with Python 3.9, NumPy 2.0.2, and Matplotlib 3.9.4. The learned-score networks are implemented in pure NumPy with manual backpropagation, rather than PyTorch or JAX. One call to a scalar or vector score routine counts as one NFE for the whole batch. Euler, Heun, midpoint, and RK4 use $1, 2, 2, 4$ evaluations per step, so an NFE budget M uses $M, M/2, M/2, M/4$ steps. Unit tests cover exact-score recovery, the coupled identity, the orientation-preserving W_2 equality, scalar Euler/Heun orders, both closed-form cancellations, GMM score and integrator checks, log-variance coordinates, invalid budgets, the vectorized variable-time GMM score, DSM training finiteness and improvement, score-network support beyond two dimensions, the new fixed-field closed forms and order constants, and curve-band reproducibility. The maximum reported identity residual is 1.6e-13.

J.2 Comparable magnitudes, opposite rankings

We search a six-parameter smooth family by applying the following three-coefficient profile independently to the multiplicative error δ and the additive error ζ; the six coefficients are therefore $(\delta_0, \delta_1, \delta_2, \zeta_0, \zeta_1, \zeta_2)$:

$$c(t) = c_0 + c_1 \cos(\pi t/T) + c_2 \sin(2\pi t/T) \tag{100}$$

for $c \in \{\delta, \zeta\}$ and for two orientation-preserving examples at $\lambda = 0.03$, $T = 3$, and $M = 128$. The search is an illustrative construction, not a random benchmark: from 12,000 fixed-seed coefficient draws, we require opposite alignment signs and then match $(\log F_c, \log D_{\mathrm{E}}, \log D_{\mathrm{H}})$. Across the selected cases, the ratios of $F_c, D_{\mathrm{E}}, D_{\mathrm{H}}$ are 1.13, 1.13, and 1.42.

Figure 3 shows the full NFE trajectories. In Case A, Euler amplifies the continuous-flow endpoint bias while Heun cancels it at the focus budget, giving a 0.33 ratio. In Case B the signs reverse and Euler wins by more than a factor of two. These matched cases test magnitude non-identifiability; at the focus budget they do not test the stronger $D_{\mathrm{H}} < D_{\mathrm{E}}$ reversal, which is established by Proposition 9 and Theorem 1 and the trained-score experiments in Section 5. Target error can increase with more NFE as a useful cancellation disappears.

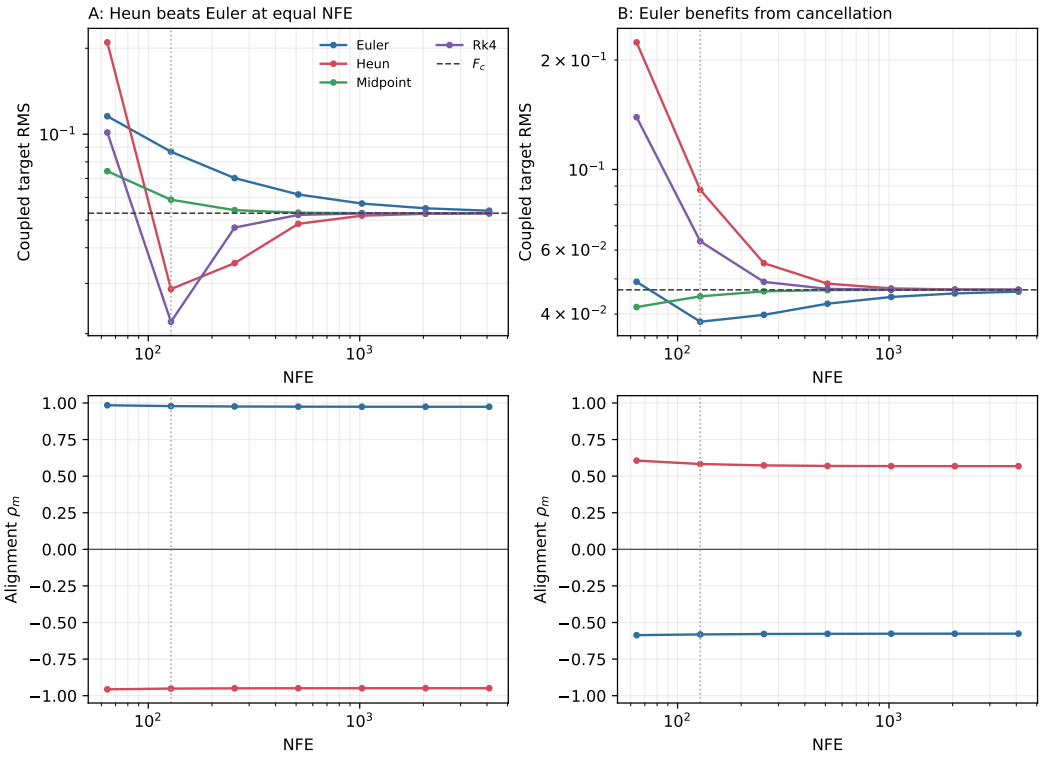

Figure 3: Coupled target RMS (top) and alignment (bottom). The dotted line marks 128 NFE; the dashed line is F_c. Curves at coarse budgets are coupled errors; equality with W_2 additionally requires a nonnegative numerical scale. Higher order can reduce numerical error yet approach the biased continuous flow from the less favorable side.

J.3 Matched scalar search protocol

We set $\beta = 1$, $T = 3$, $\lambda = 0.03$, $M = 128$, and use Equation 100 independently for δ and ζ. With seed 20270820, 12,000 coefficient vectors are drawn uniformly from

$$\delta_0 \in [-0.25, 0.25], \qquad \delta_1, \delta_2 \in [-0.9, 0.9], \tag{101}$$

$$\zeta_0 \in [-0.55, 0.55], \qquad \zeta_1, \zeta_2 \in [-0.7, 0.7]. \tag{102}$$

We retain positive Euler and Heun scales. Case A must satisfy $R_{\mathrm{H}}/R_{\mathrm{E}} \le 0.78$, $\rho_{\mathrm{E}} \ge 0.5$, and $\rho_{\mathrm{H}} \le -0.5$; Case B uses the reversed alignment inequalities and $R_{\mathrm{H}}/R_{\mathrm{E}} \ge 1.28$. Among these candidates we minimize squared Euclidean distance in $(\log F_c, \log D_{\mathrm{E}}, \log D_{\mathrm{H}})$. The selected coefficients are

Case	δ_0	δ_1	δ_2	ζ_0	ζ_1	ζ_2
A	0.07755	-0.86479	-0.70581	-0.02315	0.66741	0.17774
B	-0.07462	0.51270	0.05946	-0.08413	-0.07420	0.66341

Because selection uses the outcomes, these examples establish existence and illustrate geometry; they are not an unbiased estimate of how often either solver wins.

J.4 Phase grid

The 61×61 grid uses $\delta = (\delta_0, 0.58, -0.22)$ with $\delta_0 \in [-0.24, 0.24]$ and $\zeta = (\zeta_0, -0.42, 0.31)$ with $\zeta_0 \in [-0.5, 0.5]$. All other scalar settings match the search. We compare

$$R_{\mathrm{H}} \le 0.8 R_{\mathrm{E}} \quad \text{and} \quad F_c + D_{\mathrm{H}} \le 0.8(F_c + D_{\mathrm{E}}). \tag{103}$$

The actual rule marks 3019 points and the surrogate marks zero; all 3721 numerical scales are positive.

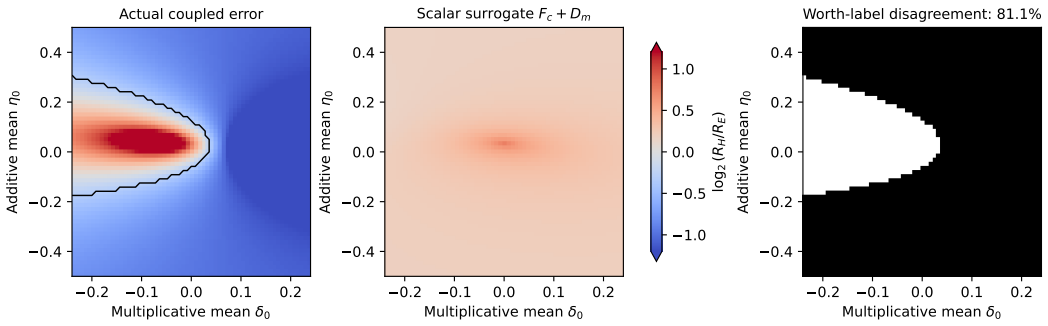

Figure 4: Actual and scalar-surrogate Heun/Euler ratios at 128 NFE. Black contours mark the 20% worth boundary; black in the right panel indicates label disagreement. The reported percentage describes this grid only.

J.5 Spectral experiment

The scalar eigenvalues are $10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$ and the NFE budgets are $8, 16, \ldots, 8192$. The 32-dimensional spectra use condition numbers $10, 10^2, 10^3, 10^4$ and three shapes: one small eigenvalue, a geometric spectrum, and 16 unit plus 16 small eigenvalues. A raw grid is uniform in VP time; the log-variance grid is built from $\lambda_{\min}$. We report the first two-interval onset from Equation 93. "Not detected" means only that the criterion does not hold through 8192 NFE; it is not a proved lower bound on the true onset.

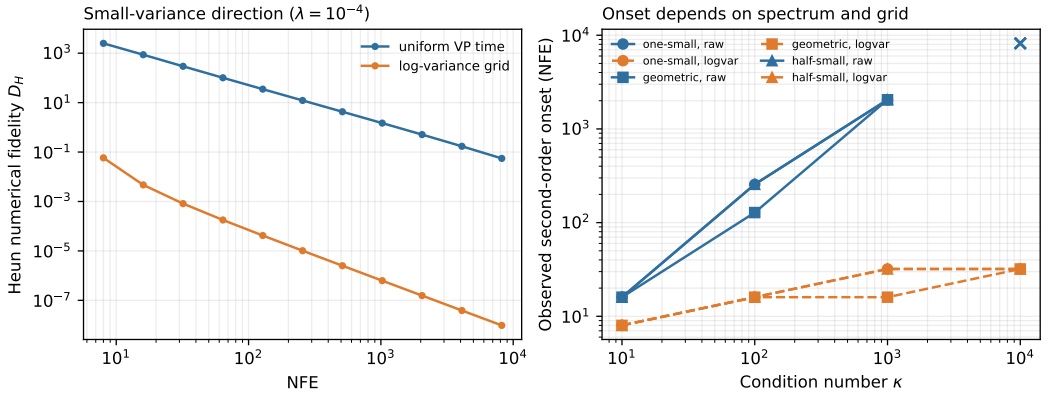

Figure 5: Exact-score Heun numerical error and observed second-order onset. Crosses denote onset not detected through 8192 NFE. A log-variance grid mitigates the raw endpoint coefficient, but a shared multidimensional grid remains spectrum dependent.

J.6 DSM-trained score networks

We use three two-dimensional shared-diagonal-covariance GMMs. The triangular mixture has means $(3, 0), (-1.5, 2.6), (-1.5, -2.6)$, weights $(0.4, 0.3, 0.3)$, and diagonal covariance $(0.12, 0.35)$; the asymmetric mixture has means $(2.8, 0.2), (-2.1, 1.7), (-1.1, -2.5)$, weights $(0.55, 0.25, 0.20)$, and diagonal covariance $(0.18, 0.48)$; the separated bimodal mixture has means $(-2.7, -0.2), (2.5, 0.3)$, weights $(0.52, 0.48)$, and diagonal covariance $(0.10, 0.55)$. The VP parameters are $\beta = 1$, terminal time $T = 3$, and stationary variance 1; terminal states are sampled from the exact finite-T forward marginal. For each model, five seeds train a two-hidden-layer tanh score network of width 32 by variance-weighted denoising score matching. Its linear-growth skip is the state-linear map

$$g_{\text{skip}}(x, t) = W_{\text{skip}}[x/x_s;\ sx/x_s;\ s^2 x/x_s;\ s;\ \sin(\pi s);\ \cos(\pi s)], \qquad s = t/T,$$

with $x_s = 3$; the remaining two hidden layers use tanh activations. Batch size is 512; checkpoints are saved after 800, 2400, and 6000 Adam updates. Adam uses learning rate 2×10^{-3}, $(\beta_1, \beta_2) =$

$(0.9, 0.999)$, $\epsilon = 10^{-8}$, no weight decay, and elementwise gradient clipping to $[-10, 10]$. Validation relative RMSE uses 4096 fresh samples and the analytic marginal score, but analytic scores are not training targets.

For each checkpoint, 512 paired terminal samples are integrated at 32, 64, and 128 NFE. An RK4 solve at 4096 NFE under the analytic score defines the target endpoint; RK4 at 512 NFE under the learned score defines the approximate continuous endpoint. On 64 paths, doubling these budgets changes endpoint RMS by at most 2.8e-09 and 2.0e-06, respectively. Empirical sliced-W_2 uses 256 fixed random projections. Every reported reversal satisfies $D_{\mathrm{H}} < D_{\mathrm{E}}$ and is checked through

$$(R_{\mathrm{H}}^2 - R_{\mathrm{E}}^2) - \{2(C_{\mathrm{H}} - C_{\mathrm{E}}) - (D_{\mathrm{E}}^2 - D_{\mathrm{H}}^2)\} = 0 \tag{104}$$

up to floating-point residual. The 135 rows are a full factorial stress-test suite over models, seeds, checkpoints, and NFE budgets; because rows share models, networks, and terminal samples, their fraction is descriptive and has no binomial interpretation.

For each row, a paired percentile interval for $R_{\mathrm{H}}/R_{\mathrm{E}}$ is computed from 500 nonparametric resamples of the 512 shared terminal paths. The sliced-W_2 ratio interval analogously resamples the 256 fixed projection costs 500 times. Of the 77 coupled reversals, 38 have a 95% interval wholly above one; 52 exceed a 1% effect and 4 exceed 5%. These row-wise intervals assess stability to the sampled paths or projections, are not multiplicity-adjusted simultaneous confidence statements, and do not include uncertainty over mixture selection or training seeds.

Table 5: Fixed learned-field trajectories in the original DSM experiment. Each trajectory keeps model, training seed, and checkpoint fixed and varies only NFE; entries are medians over all trajectories, not independent replications.

NFE	median $D_{\mathrm{H}}/D_{\mathrm{E}}$	median $R_{\mathrm{H}}/R_{\mathrm{E}}$
32	0.152	1.005
64	0.090	1.005
128	0.050	1.003

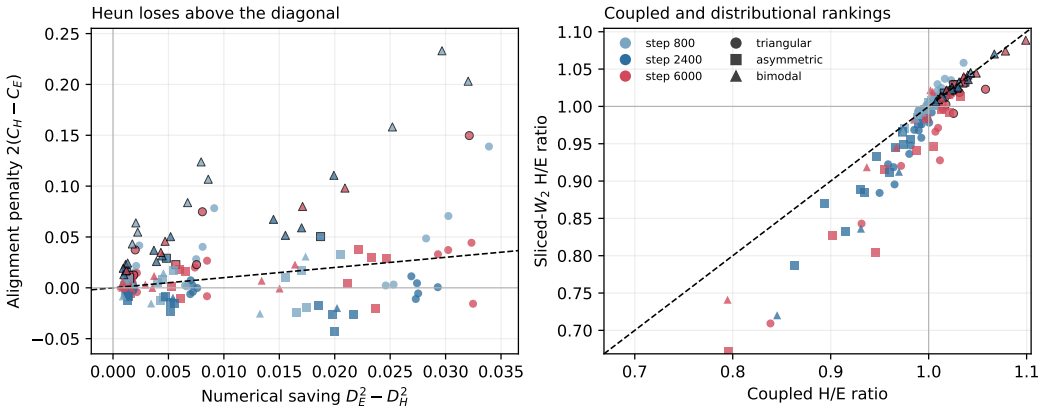

Figure 6: DSM-trained two-dimensional score networks at the original budgets. Left: every point has $S_{\mathrm{num}} > 0$; points above the dashed boundary reverse the target ranking. Right: coupled and empirical sliced-W_2 H/E ratios. Black outlines mark coupled reversals whose paired 95% interval lies above one; colors denote checkpoints and markers denote GMMs.

J.7 Extended fixed-checkpoint diagnostics

The extension uses all 15 original training runs (three GMMs, five seeds each) at checkpoint 800. This checkpoint is the earliest in the existing protocol. No curve is excluded based on its ranking. The original 32–128 NFE outcomes had already been inspected; the four additional budgets test extrapolation conditional on these same trained models and terminal samples. Checkpoints are

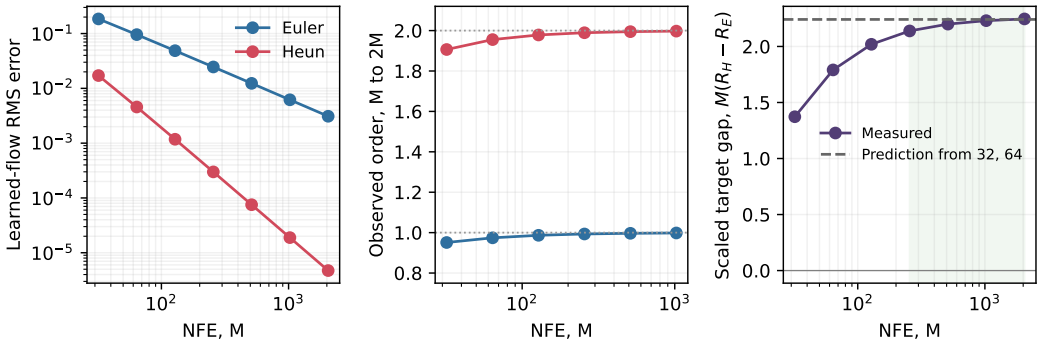

Figure 7: One fixed DSM checkpoint across equal-NFE budgets. Left: learned-flow RMS errors. Middle: adjacent-budget numerical orders. Right: the rescaled coupled target gap stays positive and approaches the coefficient predicted from 32 and 64 NFE; shading marks larger-budget holdouts. This illustrative run is first in protocol order, not an independent confirmatory selection. All 15 runs, including non-reversals, appear in Appendix J.7.

reconstructed with the original DSM seed and parameters. Re-evaluation at the old budgets with the old references matches the archived $D_{\mathrm{E}}, D_{\mathrm{H}}, R_{\mathrm{E}}, R_{\mathrm{H}}$ values to a maximum absolute discrepancy of 0.0e+00. Parameter hashes are saved in the result JSON.

Each run uses the same 512 terminal samples at every budget. The learned-flow reference is RK4 at 4096 NFE, checked against 8192 NFE on all paths; the exact-score target uses 8192 NFE, checked against 4096. The reference-doubling difference is a convergence diagnostic rather than a rigorous error bound. We report it relative to the smallest Heun error to expose potential reference-floor contamination.

For $m = \mathrm{E}, \mathrm{H}$, the observed order is $\widehat{p}_m(M) = \log_2[D_m(M)/D_m(2M)]$. Smaller D_{H} alone does not establish second order. The right panel of Figure 7 instead tests the leading target-gap prediction from Theorem 5. With $M_c = 32$, estimate

$$\widehat{L}_{\mathrm{E}} = 4M_c E_{\mathrm{E}}(2M_c) - M_c E_{\mathrm{E}}(M_c), \qquad \widehat{g}(M) = -\frac{\langle B, \widehat{L}_{\mathrm{E}} \rangle}{F_c M}. \tag{105}$$

This cancels a common M^{-2} term in the defect if a finer expansion exists; the assumptions of Theorem 5 alone guarantee only $\widehat{L}_{\mathrm{E}} = L_{\mathrm{E}} + O(M_c^{-1})$. The estimate and references are fixed before evaluating signs at $256, 512, 1024, 2048$. We retain the incorrect prediction as well as the 59 correct ones. These 60 comparisons are dependent and the prediction requires oracle score information, so the count is neither a binomial accuracy estimate nor a deployable selection guarantee.

Post-hoc extrapolation baselines. Let $G(M) = R_{\mathrm{H}}(M) - R_{\mathrm{E}}(M)$ and $m = 32$. Fitting $G(M) \approx c_1/M + c_2/M^2$ to the same two calibration budgets gives

$$\widehat{c}_1 = 4mG(2m) - mG(m), \qquad \widehat{c}_2 = 2m^2\{G(m) - 2G(2m)\}. \tag{106}$$

Table 6 compares $\widehat{c}_1/M$ and $\widehat{c}_1/M + \widehat{c}_2/M^2$ with alignment extrapolation and with copying one calibration sign (scaling its gap by m/M). Both direct extrapolations match the alignment sign count and have lower RMSE on this suite. They use oracle target endpoints but no learned-flow reference, whereas alignment needs both references. These are post-hoc, dependent comparisons, not significance tests or evidence of an image-deployable predictor.

To quantify path-sample uncertainty over an entire curve, resample the 512 path indices jointly across all seven budgets for 2000 bootstrap replicates. Set $\ell_j = \log(R_{\mathrm{H}}/R_{\mathrm{E}})$ at budget j, estimate its bootstrap standard deviation s_j, and take the 95th percentile c of $\max_j |\ell_j^* - \ell_j|/s_j$. The band is $[\exp(\ell_j - cs_j), \exp(\ell_j + cs_j)]$. This is an approximate simultaneous bootstrap band over budgets conditional on a network. It is not adjusted across the 15 networks and does not include training uncertainty. The original 45 checkpoint curves comprise only 15 training runs, since checkpoints from one run are correlated.

Table 6: Extrapolation to 256–2048 NFE across the same 15 checkpoints. All 60 comparisons share models and samples. RMSE is for the unscaled target gap.

Method	Correct signs	Gap RMSE
Copy 32-NFE sign; scale gap	45/60	9.549×10^{-4}
Copy 64-NFE sign; scale gap	53/60	4.450×10^{-4}
Alignment extrapolation	59/60	1.153×10^{-4}
Direct gap: $\widehat{c}_1/M$	59/60	7.717×10^{-5}
Direct gap: $\widehat{c}_1/M + \widehat{c}_2/M^2$	59/60	4.338×10^{-5}

J.8 Later-checkpoint extension and common numerical window

All original model–seed pairs are retained at 2400 and 6000 updates. The shared budgets, samples, RK4 references, and within-network bootstrap construction are as in Appendix J.7. Frozen protocol and source hashes, model-parameter hashes, and raw endpoints accompany all 30 curves. Reconstruction reproduces the checkpoint-800 parameter hashes; replay at 32, 64, and 128 NFE with the old references has maximum metric discrepancy zero. Independent recomputation from the saved endpoints also has discrepancy zero. These are implementation-consistency checks, not independent replications.

For each curve we select the earliest suffix with at least two doubling intervals, smaller Heun numerical error at every budget, and all subsequent point orders in $[0.8, 1.2]$ for Euler and $[1.8, 2.2]$ for Heun. The intersection of these 30 suffixes is $128, 256, 512, 1024, 2048$. This rule never reads target rankings. It does not account for uncertainty in the observed orders. Target-ratio bands still use all seven budgets when calculating the critical value; their restrictions to the common window are not recalibrated after observing target outcomes. They condition on each network, do not cover all networks jointly, and do not estimate a training-population reversal rate.

The full-seven-budget band lies above one for 5/15 and 3/15 curves at the two later checkpoints, respectively; over the common window these counts are 5/15 and 8/15. Thus pointwise reversal counts must not be read as significance counts. Mean validation score relative RMSE improves from 0.2889 to 0.1956 between the later checkpoints, but the full checkpoint reversal sequence $8, 5, 10$ is not monotone. Figures 9–11 retain non-reversals and display the finite-budget order diagnostics.

The largest learned-flow reference-doubling change is 7.403×10^{-10} in endpoint RMS; its maximum ratio to a curve's smallest Heun error is 4.710×10^{-5}. The largest target-reference change is 4.367×10^{-9}. These checks diagnose reference resolution, not rigorous reference error. At 128 NFE, the largest Heun relative target excess is 2.137% and 2.562% at 2400 and 6000 updates; at 512 NFE it is 0.557% and 0.751%, and at 2048 NFE 0.141% and 0.194%. The larger 32-NFE excesses (6.705% and 9.911%) lie outside the common normal-order window and cannot be combined with orders measured elsewhere as one joint observation.

Table 7: Later-checkpoint extrapolation from 32/64 to 256–2048 NFE. All 120 comparisons are dependent. All methods require oracle targets; alignment additionally requires a learned-flow reference. The simple two-term fit is stronger on both displayed criteria.

Method	Correct signs	Gap RMSE
Alignment	117/120	0.000254
Direct leading term	117/120	0.00022
Direct two-term fit	118/120	0.000163

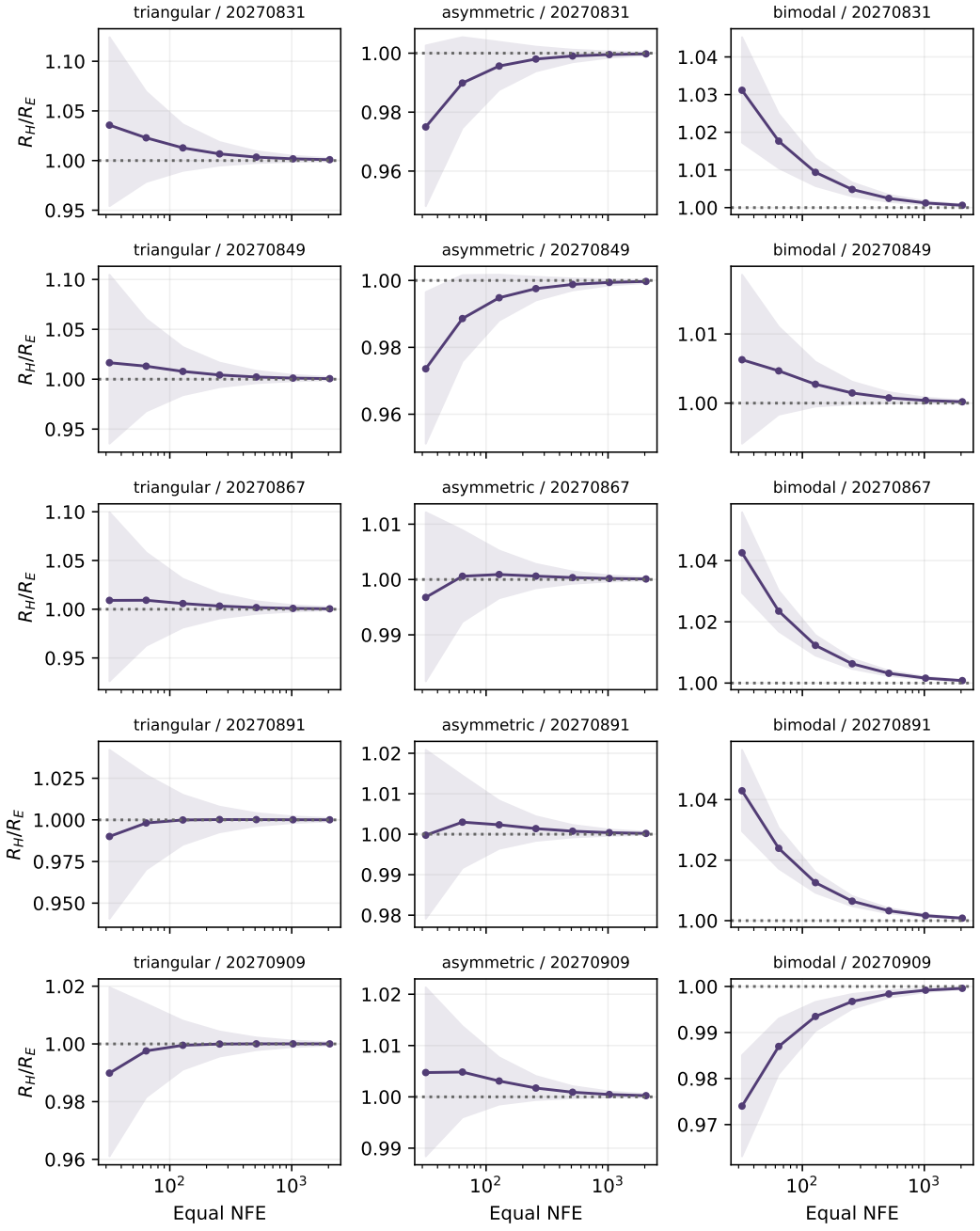

Figure 8: All fixed-checkpoint curves: columns are GMMs and rows are seeds in protocol order. Ratios above one favor Euler in coupled target RMS; shaded bands are approximate curve-wise 95% bootstrap bands over the seven budgets. All 15 runs have lower Heun numerical error; 8 have pointwise target reversal at every tested budget, and 3 have their entire band above one. Different panels share terminal samples within a GMM and are not independent population draws.

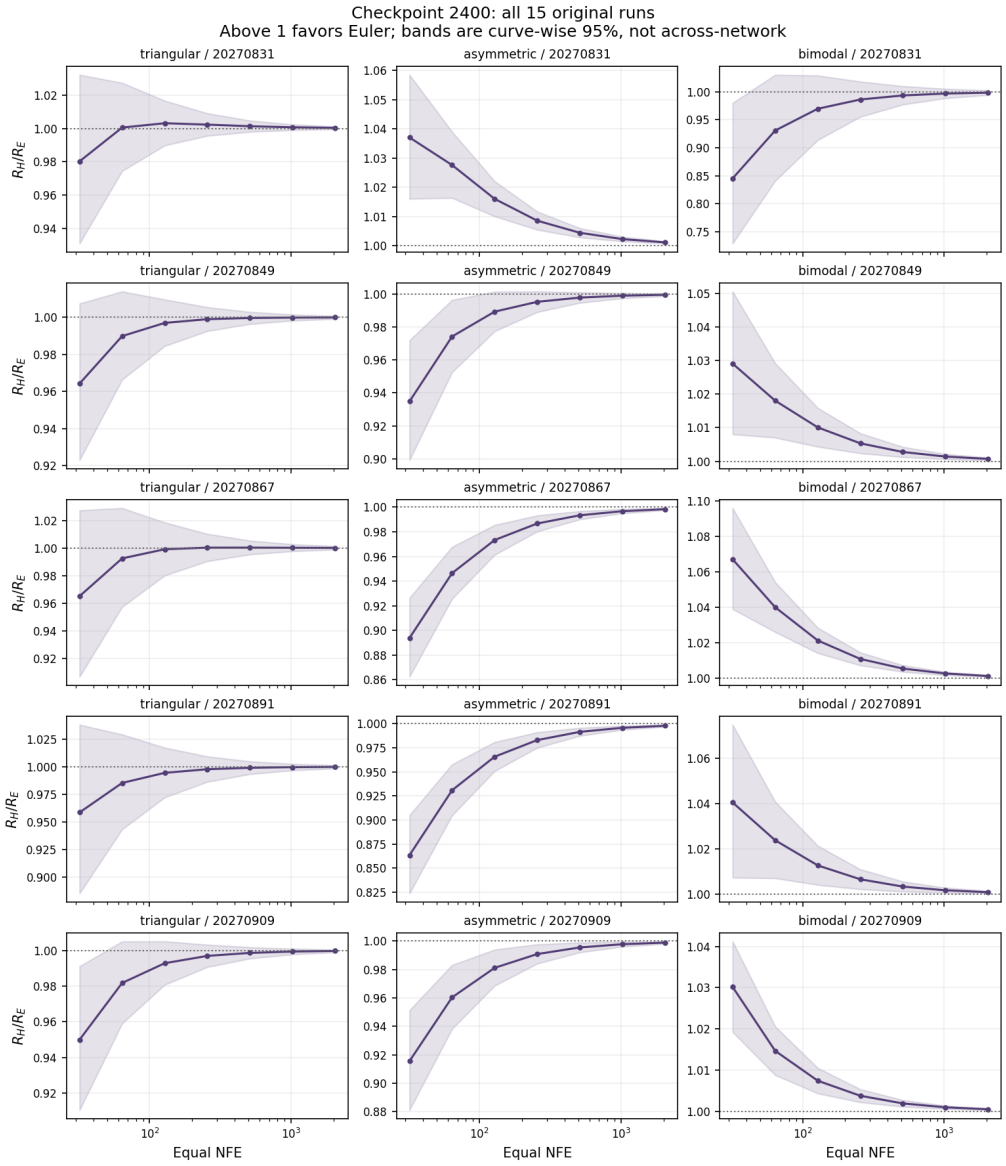

Figure 9: All 15 checkpoint-2400 curves in protocol order. Shading gives within-network seven-budget bands, not across-network coverage. Ratios above one favor Euler in coupled target RMS.

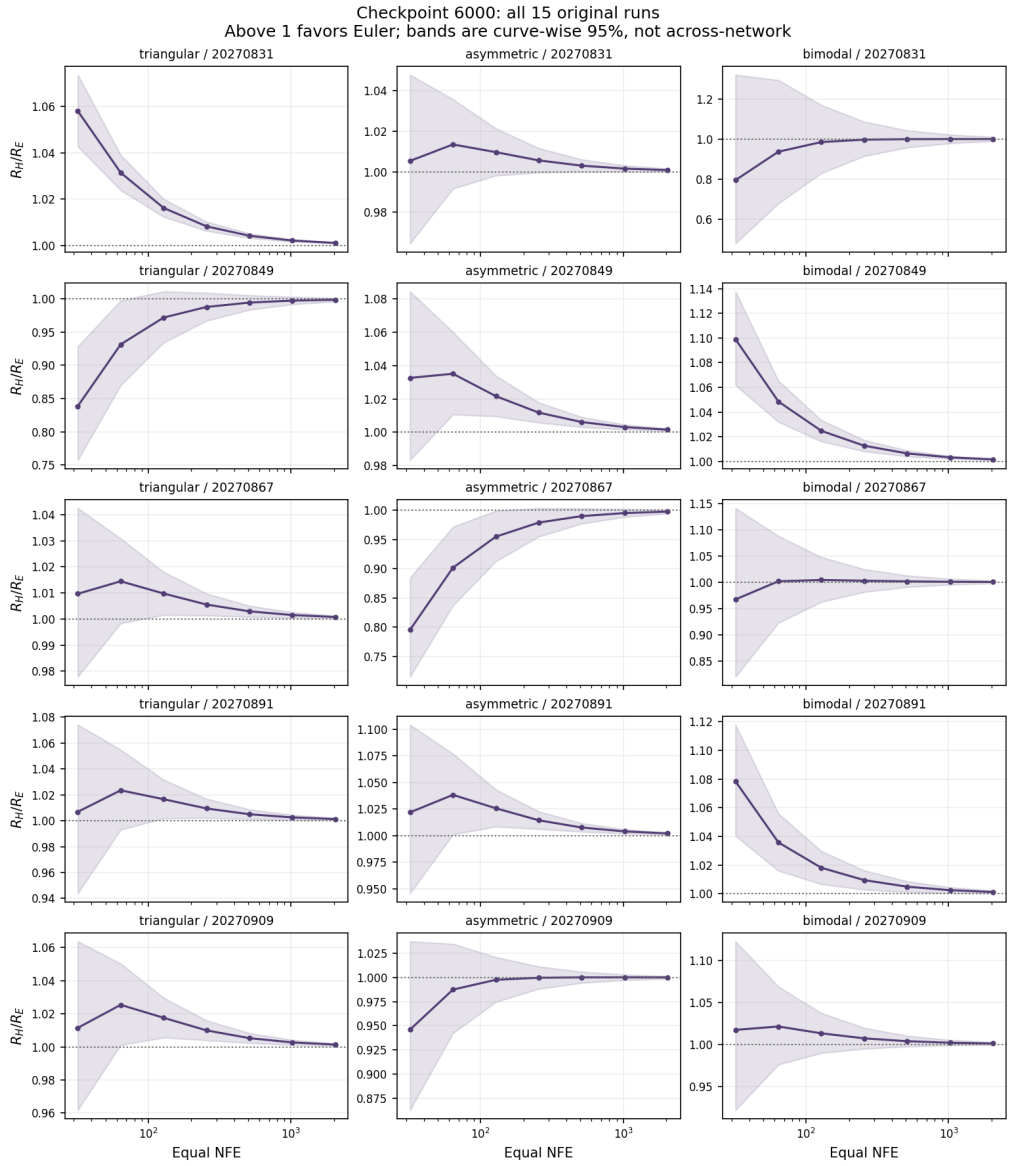

Figure 10: All 15 checkpoint-6000 curves, with the same interpretation as Figure 9. Checkpoints from a common training run and panels sharing initial samples are dependent.

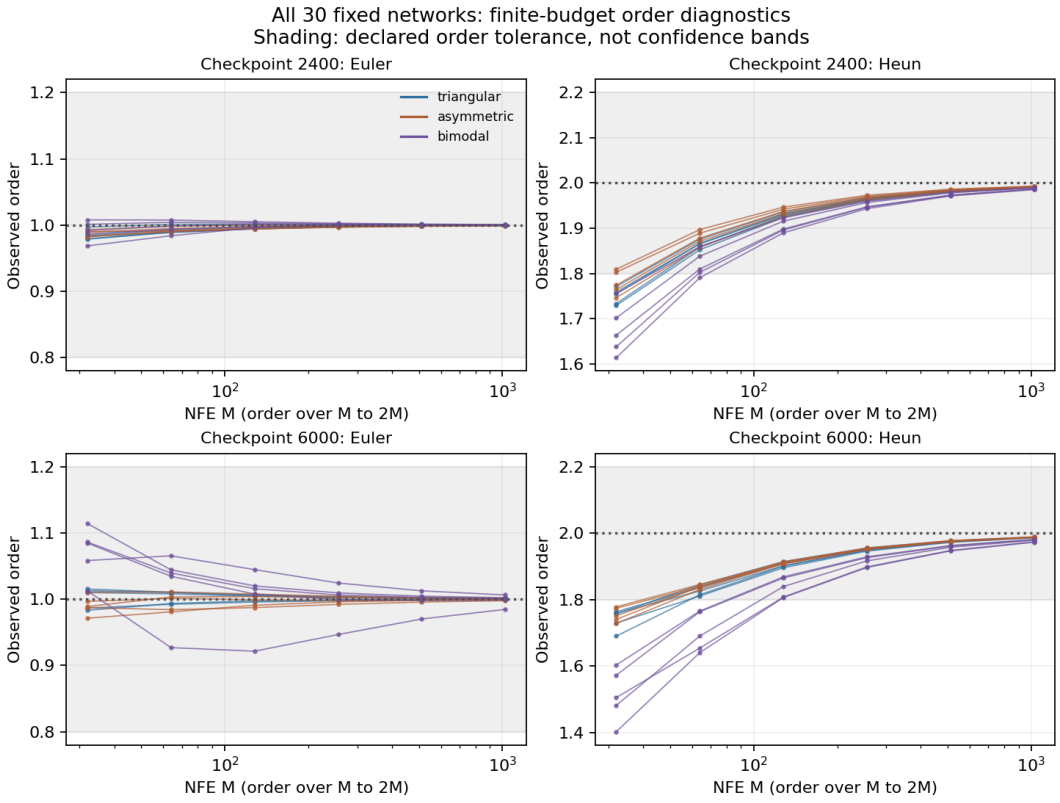

Figure 11: Numerical orders for all 30 later-checkpoint curves. The horizontal axis is the lower NFE of each doubling interval. Shading marks the declared point-order tolerances, not confidence bands.

J.9 CONTROLLED 32-DIMENSIONAL TEACHER–STUDENT EXPERIMENT

The teacher is a four-component, 32-dimensional GMM with weights $(0.34, 0.27, 0.22, 0.17)$. With seed 20279820, a Gaussian matrix is orthogonalized by QR; the first four orthonormal directions are each multiplied by 6, so every component mean has Euclidean norm 6. The shared diagonal covariance has entries geometrically spaced from 0.12 to 0.85. Thus the analytic marginal score remains available in every dimension, but is used only for validation and the target reference, not as a student training label.

The VP parameters are $\beta = 1$, terminal time $T = 3$, and stationary variance 1; terminal states are sampled from the exact finite-T forward marginal. Three seeds train two-hidden-layer tanh networks of width 48 by variance-weighted DSM with batch size 256, learning rate 2×10^{-3}, Adam $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, no weight decay, and elementwise gradient clipping to $[-10, 10]$. Checkpoints are taken after 500, 1500, and 3500 Adam updates. For each checkpoint, 256 paired terminal paths are evaluated at 32, 64, and 128 NFE. The analytic-teacher target uses RK4 at 2048 NFE and the student continuous endpoint uses RK4 at 512 NFE; doubling these budgets on 32 paths changes endpoint RMS by at most 3.9e-08 and 1.5e-08, respectively. Empirical sliced-W_2 uses 512 fixed random projections. Coupled intervals resample the 256 paths and sliced intervals resample projection costs, each for 750 percentile-bootstrap repetitions. They are row-wise stability diagnostics, not multiplicity-adjusted simultaneous confidence statements.

All 27 rows satisfy $D_{\mathrm{H}} < D_{\mathrm{E}}$ and reverse the coupled target ranking, and all 27 paired 95% intervals lie above one. The mean validation relative score RMSE falls from 0.396 at the first checkpoint to 0.334 at the last. Because the rows reuse one teacher, terminal sample, and training design, these results demonstrate a controlled higher-dimensional realization rather than a prevalence estimate or an image-model proxy.

J.10 GAUSSIAN-MIXTURE STRESS TEST

The GMM has means $(3, 0), (-1.5, 2.6), (-1.5, -2.6)$, weights $(0.4, 0.3, 0.3)$, and shared diagonal covariance $\mathrm{diag}(0.12, 0.35)$. We draw 2048 exact samples from its $T = 3$ forward marginal with seed 20270820. The exact marginal score is computed by mixture responsibilities. The affine additive profile points in direction $(1, -0.35)$. Exact-score and approximate-score continuous endpoints use RK4 at 8192 NFE. Doubling the reference budget on the first 256 samples changes the approximate endpoint RMS by at most 1.5e-10.

At 64 NFE, the paired results are

Profile	Solver	D_m	ρ_m	R_m	$R_{\mathrm{H}}/R_{\mathrm{E}}$
A	Euler	0.09968	-0.2537	1.60518	1.03
	Heun	0.08843	0.1757	1.64542	
B	Euler	0.27924	-0.1386	1.93623	1.01
	Heun	0.03131	0.2732	1.96385	

The 95% intervals in the main text are percentile intervals from 1000 nonparametric resamples of the 2048 paired terminal paths. They do not include training-seed variation. We also compute sliced W_2 with 128 fixed random projections as a secondary diagnostic, but make no theorem-level claim from it.

J.11 UNPAIRED RKHS GATE

The unpaired gate uses the first GMM above with $\beta = 1$, $T = 3$, and stationary variance 1; terminal states follow the exact finite-T forward marginal. We draw 1024 independent data-time target samples with seed 20270821 and 1024 terminal samples with seed 20270822; the score network is trained with seed 20270823, batch size 512, width 32, and checkpoints at 800, 2400, and 6000 Adam updates. Each checkpoint is evaluated at 32, 64, and 128 NFE. The learned-flow reference uses RK4 at 512 NFE and is doubled to 1024 NFE on 128 held-out paths. We predeclare the linear, KID-style cubic polynomial, and three-scale RBF kernels; the RBF bandwidth is estimated from target samples only. For each comparison, independent target and generated observations are partitioned into blocks of

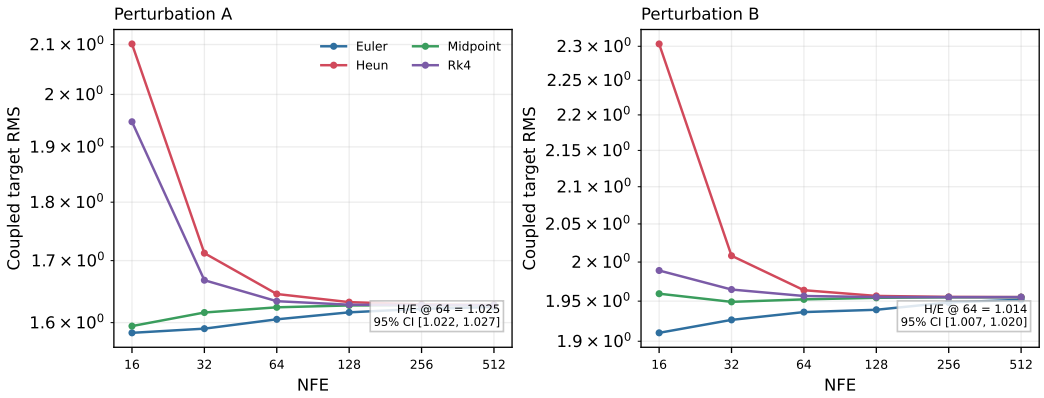

Figure 12: GMM coupled target RMS. Both transferred profiles mildly favor Euler at 64 NFE, even though Heun has lower numerical error. The scalar labels therefore do not transfer, while the signed interaction continues to explain the within-model ordering.

64, and the block means are resampled 1000 times to form nominal 95% percentile intervals. The calibration experiment uses the same nominal level over 100 repetitions. Because the gate has only 16 blocks, its intervals are a conservative screening diagnostic rather than a claim of adequate power for all small effects.

K Claim and Reproducibility Audit

Claim	Evidence	Boundary
Fixed-field reversal with exact first/second numerical orders	Theorem 1	Exact W_2 for positive-scale Gaussian maps
Nonlinear local stability of both rankings	Proposition 2	One-dimensional C_b^5 neighborhood; uniform threshold, not a training guarantee
Alignment determines eventual coupled ranking	Theorem 5, Corollary 6	Requires nonzero bias and nonzero leading alignment
Observed orders coexist with trained-score reversal	Figure 1, Table 2	Finite-budget diagnostics; three checkpoints of 15 training runs, no monotone maturity trend
Alignment agrees with larger-budget behavior	Eq. 105, Table 6	Oracle diagnostic; simple gap extrapolation matches signs and has lower RMSE
Fixed 20% worth vanishes at large budget when $F_c > 0$	Theorem 5	Does not say the ordinary ranking must switch
Raw spectral stiffness delays observed order	Figure 5	Diagnostic; parameterization-aware spectral theory is prior work
Alignment remains relevant in one GMM	Appendix J.10	One synthetic mixture; scalar labels do not transfer
Training-generated errors exhibit reversals	Figure 6, Appendix J.6	Three synthetic 2D GMMs; no prevalence or image-model claim
The signed decomposition extends to unpaired MMD	Theorem 7	Mean-embedding identity; not a claim that MMD itself is new
Coupled reversal need not transfer to MMD	Unpaired distributional gate in Section 5	Three polynomial-kernel point reversals, none certified; no linear/RBF reversal
The reversal survives 32D aggregation with a learned student	Figure 2, Appendix J.9	One controlled analytic teacher; no architecture- or dataset-level generalization

The base regeneration entry point is `python experiments/run_all.py`. The fixed-field and nonlinear extensions have separate entry points `run_fixed_field_validation.py`

and `run_revision_checks.py`; their `--from-results` options rebuild derived artifacts. Later checkpoints are generated by `run_mature_checkpoint_validation.py`; `export_later_checkpoint_tables.py` verifies frozen source and raw-state hashes and recomputes endpoint metrics before exporting their tables. Image-validation code is preparatory and is not a source of any empirical claim in this manuscript.