When Higher Order Hurts: Pre-Asymptotic Order Collapse in Generative ODE Sampling
A Theory of Discretization–Learning Interaction

Farzad Salajegheh[1] and Sudhir Mudur[1]

Concordia University, Montreal, Canada {farzad.salajegheh, sudhir.mudur}@concordia.ca

Abstract. Higher-order ODE solvers are widely believed to improve sample quality in diffusion and flow-based generative models at a fixed number of function evaluations (NFE). We show that this assumption can fail: at practical step sizes, higher-order methods can systematically underperform first-order Euler. The mechanism is a *multiplicative interaction* between discretization and learning error. Applying backward error analysis to a learned field $\hat{f} = f + \varepsilon$ reveals interaction terms $h^k \nabla^k \varepsilon$ that dominate classical truncation error at practical NFEs in diffusion and flow models, causing a solver of nominal order p to behave as lower order—*pre-asymptotic order collapse*. Our analysis yields a predictive crossover step size $h^\star$ characterizing when higher order helps or hurts. Validation on toy manifolds and pretrained diffusion and flow models confirms the predicted hierarchy: Euler is best at low NFE, Heun overtakes at moderate budgets, and RK4 can surpass Heun at higher NFE. Building on this theory, we propose an adaptive sampler that dynamically selects solver order based on the local roughness score $R = h/h^\star$, estimated at zero extra NFE, and often improves sample quality across models and datasets.

Keywords: Generative ODEs · Diffusion models · Numerical integration · Pre-asymptotic order collapse · Adaptive solvers

1 Introduction

ODE-based generative models [20, 22, 32] produce samples by integrating a learned velocity field from noise to data. At inference time, this reduces to a numerical integration problem: *which solver, and at what step size, should be used to obtain the best samples under a fixed function-evaluation (NFE) budget?* Recent advances have developed higher-order solvers tailored to diffusion ODEs [14,23,36,37], achieving strong sample quality with as few as 10–20 NFEs.

Empirical paradox. Despite the theoretical advantage of higher-order methods, practitioners frequently observe that in low-compute regimes (10–20 NFEs), second-order methods such as Heun can underperform first-order Euler on pretrained diffusion and rectified-flow models. These degradations occur not only under the equal-NFE convention (which inflates the step size of higher-order solvers), but can also arise when step sizes are matched. Thus, increasing solver

order does not necessarily improve sample quality in practice. This behavior is widely observed in practice but lacks a principled theoretical explanation grounded in numerical analysis.

We provide such an explanation using *backward error analysis* (BEA) [12], the classical framework for understanding numerical integrators. The key distinction from standard ODE analysis is that the vector field in generative modeling is *learned*. We model it explicitly as $\hat{f} = f + \varepsilon$, where f is a smooth target field and ε denotes structured learning error. Applying BEA to $\hat{f}$ reveals that discretization and learning error couple *multiplicatively* in the modified equation, generating interaction terms of the form $h^k \nabla^k \varepsilon$, i.e., products of step size and higher-order derivatives of the learning error.

These interaction terms are absent from classical BEA, which assumes access to an exact vector field. They are likewise not modeled in prior BEA analyses of diffusion ODEs [11][1], nor in convergence analyses that treat learning and discretization errors additively [3, 7]. By decomposing $\hat{f} = f + \varepsilon$, our analysis tracks how solver truncation interacts with derivatives of the learning error.

When the learned field is spatially rough—particularly off the data manifold— these interaction terms can dominate classical truncation error in the pre-asymptotic regime relevant to low-NFE sampling. Consequently, a solver of nominal order p can exhibit an effectively lower convergence order for practically relevant step sizes. We term this phenomenon *pre-asymptotic order collapse*.

Importantly, this mechanism does not rely solely on equal-NFE conventions. Even at equal step size, interaction terms involving $\nabla^p \varepsilon$ may dominate classical truncation error. Our analysis yields an explicit crossover step size $h^\star$ that predicts when higher-order solvers provide genuine gains and when increasing order instead amplifies model-induced error.

Relation to classical order reduction. Classical numerical analysis shows that if a vector field lacks sufficient smoothness, a Runge–Kutta method of order p cannot attain order exceeding the field's regularity [12]. However, this theory treats the vector field as a monolithic object. In learned generative models, the field naturally decomposes as $\hat{f} = f + \varepsilon$, where f is smooth and ε captures structured approximation error. Our contribution differs in three respects. **(i) Structural decomposition.** We separate f and ε, identifying the regularity bottleneck as residing entirely in the learning error and exposing explicit interaction terms coupling $\nabla^k \varepsilon$ with the smooth dynamics. **(ii) Predictive crossover.** Rather than merely stating that attainable order is limited, we derive a quantitative crossover step size $h^\star$ that predicts solver preference in practice as a function of derivative norms of ε. **(iii) Geometric amplification.** In generative models, learning error is structured: it is small near the data manifold but grows off it. This geometry induces an accelerated off-manifold drift mechanism (Theorem 5.3) without analogue in generic order-reduction theory. Empirically, this

[1] Gao et al. analyze backward error for diffusion ODE solvers under the assumption that the learned model accurately represents the probability-flow ODE; learning error is not explicitly modeled.

regime aligns with early and mid sampling phases where learned fields exhibit higher curvature.

1.1 Contributions

We make five main contributions: **C1. Structural decomposition:** We analyze learned generative dynamics via $\hat{f} = f + \varepsilon$, applying BEA to isolate interaction terms between solver truncation and learning error (Sec. 4). **C2. Pre-asymptotic order-collapse theorem:** We show effective convergence order is limited by the smoothness of ε, providing conditions where Heun underperforms Euler at equal NFE (Sec. 5). **C3. Geometric amplification:** We derive an explicit crossover step size $h^\star$ depending on higher-order derivatives of ε and manifold geometry, identifying when increasing solver order amplifies model-induced error (Sec. 5). **C4. Interaction-aware diagnostic:** We propose a roughness score $R = h/h^\star$, computable at zero extra NFE, for adaptive solver-order selection (Sec. 6). **C5. Experimental validation:** We validate our theory on toy manifolds (Supplementary) and show crossover behavior at scale using pretrained EDM and Rectified Flow models on CIFAR-10, CelebA-HQ, and ImageNet (Sec. 7).

In short, higher solver order helps only when the learned vector field is smooth enough; otherwise, interaction between discretization and learning error collapses the effective order. Overall, our results suggest solver order should be chosen based on the regularity structure of the learned field, rather than nominal asymptotic order alone.

2 Related Work

Diffusion and flow models as generative ODEs. Score-based generative models [32] unify diffusion processes under SDEs and introduce the probability-flow ODE, enabling deterministic sampling with identical marginals. DDPM [13] and DDIM [31] provide the canonical discrete and deterministic foundations, with DDIM highlighting degradation under imperfect scores. Flow Matching [20] trains generative ODEs via vector-field regression, while Rectified Flow [22] promotes straighter transport through reflow, reducing curvature—the type of derivative structure measured by our roughness diagnostic. Stochastic Interpolants [1] unify these approaches in a continuous-time bridge framework. Scalable Interpolant Transformers (SiT) [25], built on DiT [27], provide ImageNet-scale implementations. All share the same inference mechanism: numerical integration of a learned velocity field, to which our decomposition $\hat{f} = f + \varepsilon$ applies.

Fast ODE solvers for diffusion and flow sampling. DPM-Solver [23] and DPM-Solver++ [24] derive higher-order exponential-integral solvers tailored to diffusion ODEs, observing instabilities at large guidance scales. DPM-Solver-v3 [39] reduces first-order error at very low NFEs using empirical model statistics. UniPC [37], DEIS [36], and PNDM [21] adapt predictor–corrector and multistep strategies, with PNDM invoking manifold structure. GENIE [9] computes higher-order derivatives via Jacobian–vector products to construct improved solvers. EDM [14] reports diminishing returns of higher-order methods at very low NFEs, and AMED-Solver [40] learns corrected step directions in this regime.

Backward error analysis and neural ODE numerics. Backward error analysis (BEA) via modified equations [12] and B-series [4] forms the theoretical foundation of our work; classical BEA assumes an exact vector field. Gao *et al.* [11] apply BEA to diffusion ODE samplers and optimize time-step schedules, but do not model learning error, so interaction terms $h^k\nabla^k\varepsilon$ do not appear. Barrett and Dherin [2] use modified equations to study implicit regularization in optimization. Neural ODEs [5] and their numerical treatment [15] motivate adaptive integration in learned dynamics.

Train–inference mismatch and sampling drift. Daras *et al.* [8] formalize sampling drift under imperfect scores; Ning *et al.* [26] analyze exposure bias and error compounding. Li *et al.* [18] study inference-time step adjustment, and Lin *et al.* [19] identify training–inference discrepancies in noise schedules. These works support the view that learning error grows off the data manifold.

Schedule optimization and adaptive solvers. Align Your Steps [30], Xue *et al.* [35], and LD3 [33] optimize time-step placement for fixed solvers. DC-Solver [38] and BELM [34] analyze correction and truncation behavior in multistep schemes. Kong *et al.* [16] adapt diffusion sampling using embedded Runge–Kutta pairs and local error estimates. S4S [10] learns custom few-step solvers.

Convergence theory for generative ODE samplers. Chen *et al.* [7] and Lee *et al.* [17] establish polynomial-time convergence under L^2-accurate scores. Chen *et al.* [6] analyze probability-flow ODE sampling with correctors. De Bortoli [3] proves convergence under a manifold hypothesis, and Pidstrigach [29] studies support recovery. Pfarr *et al.* [28] derive Wasserstein-2 guarantees for higher-order SDE samplers. These analyses treat learning and discretization errors additively; in contrast, we identify multiplicative interaction terms $h^k\nabla^k\varepsilon$ that can induce deterministic pre-asymptotic order collapse.

3 Problem Setup

3.1 Generative ODE Sampling

We consider a generative ODE $\dot{x} = f(x,t)$ transporting a prior π_0 to a data distribution π_T:

$$\frac{dx}{dt} = f(x,t), \qquad x(0) \sim \pi_0, \qquad x(T) \sim \pi_T. \tag{1}$$

In practice, f is unknown. We have a learned approximation $\hat{f}(x,t)$ with error $\varepsilon(x,t) := \hat{f}(x,t) - f(x,t)$. Samples are generated by applying a numerical solver Φ_h to $\hat{f}$: $x_{n+1} = \Phi_h(x_n, t_n; \hat{f})$.

Equal-NFE convention. We fix total NFE $= N$: Euler takes N steps of size $h = T/N$; Heun takes $N/2$ steps of $2h$; RK4 takes $N/4$ steps of $4h$.

3.2 The Data Manifold and Structured Error

The data distribution π_T is assumed to concentrate near a compact C^3 submanifold $\mathcal{M} \subset \mathbb{R}^d$ of intrinsic dimension $m < d$. Within the tubular neighborhood $\mathcal{M}_\rho := \{x : \mathrm{dist}(x,\mathcal{M}) < \rho\}$ for $\rho < \mathrm{reach}(\mathcal{M})$, every point has a unique

nearest-point projection $\pi_{\mathcal{M}}(x)$ and admits a tangential/normal decomposition $v = P_\tau(x)v + P_\nu(x)v$ for any vector v.

We make the following assumptions.

Assumption 1 (Structured Learning Error) *There exist constants $\varepsilon_{\text{train}} \geq 0$ and $L_\varepsilon \geq 0$ such that for all $x \in \mathcal{M}_\rho$, $t \in [0,T]$:*

$$\|\varepsilon(x,t)\| \leq \varepsilon_{\text{train}} + L_\varepsilon \cdot \text{dist}(x, \mathcal{M}). \tag{A1}$$

Assumption 2 (Derivative Regularity of ε) *The learning error $\varepsilon(\cdot,t)$ is C^{p+1} on $\mathcal{M}_\rho$ (a reasonable approximation for standard smooth neural network parametrizations), with bounded spatial derivative norms $C_k(\varepsilon) := \sup_{x\in\mathcal{M}_\rho,\, t\in[0,T]} \|\nabla_x^k \varepsilon\|_{\text{op}} < \infty$ for $k = 1,\ldots,s$, where $\|\cdot\|_{\text{op}}$ denotes the tensor operator norm. The integer $s \in \{1,\ldots,p\}$ is the **practical regularity index**: $\nabla^k\varepsilon$ exists for all k under common smooth neural parametrizations, but $C_k(\varepsilon)$ for $k > s$ is so large that the corresponding interaction terms dominate for all practical h*—pre-asymptotic order collapse *rather than classical order reduction, which would require actual derivative non-existence. Although s is not directly observable, its effect is operationalized through the roughness score $R = h/h^*$ (Sec. 6), which estimates the ratio $C_2(\hat{f})/C_1(\hat{f})$ from finite differences of $\hat{f}$ along the trajectory at zero extra NFE.*

Assumption 3 (True Field Regularity) *$f(\cdot,t)$ is C^{p+1} on $\mathcal{M}_\rho$ with bounded derivatives $C_k(f) < \infty$ for $k = 0,\ldots,p+1$.*

Assumption 4 (Manifold Approximate Invariance) *For $x \in \mathcal{M}_\rho$ and $t \in [0,T]$, the true flow satisfies $\text{dist}(\varphi_t^f(x), \mathcal{M}) \leq C_{\mathcal{M}} \cdot \text{dist}(x, \mathcal{M})$ while $\varphi_{[0,t]}^f(x) \subset \mathcal{M}_\rho$, for some $C_{\mathcal{M}} \geq 1$. This is a local-in-tube Lipschitz stability condition on the normal component of f; it holds, for example, whenever $\|P_\nu \nabla f\|_{\text{op}} \leq \ln C_{\mathcal{M}}/T$ on $\mathcal{M}_\rho$.*

Remark 3.1 *Assumption 1 captures the key modeling intuition: neural networks trained on data from $\mathcal{M}$ have small error on $\mathcal{M}$ but no training signal off $\mathcal{M}$, leading to error that grows with distance. Assumption 2 captures that a small but spatially rough error can have large derivatives—the mechanism for pre-asymptotic order collapse.*

Remark 3.2 (Time-dependent support) *The assumptions can be extended to a time-dependent tube $\mathcal{M}_\rho(t)$; since BEA is local (one-step), bounds apply at each t with parameters $\varepsilon_{\text{train}}(t)$, $L_\varepsilon(t)$, $C_k(\varepsilon)(t)$. The roughness profile is model-dependent: for EDM, $R(t)$ rises by several orders of magnitude toward the data end ($t \to T$); for Rectified Flow, roughness peaks at the noise end ($t \approx 0$) and decreases—the profile is effectively inverted. Both behaviors are empirically observed in Sec. 7.3.*

Under the above assumptions, sampling with $\hat{f}$ amounts to integrating a perturbed vector field whose derivatives may be significantly rougher than those of f. We next show that this derivative roughness fundamentally alters the pre-asymptotic behavior of higher-order solvers.

4 Backward Error Analysis for Learned Fields

4.1 Modified Equations via Taylor Matching

For notational simplicity we present the modified equations in autonomous form; the non-autonomous case $\hat{f}(x,t)$ is handled via standard augmentation $z=(x,t)$, $\dot{z}=(\hat{f}(x,t),1)$—see Remark 4.1 below.

We seek a modified vector field $\tilde{f}$ such that the solver's one-step map $\Phi_h(x;\hat{f})$ equals the exact flow of $\tilde{f}$ up to order r:

$$\Phi_h(x;\hat{f}) = \varphi_h^{\tilde{f}}(x) + \mathcal{O}(h^{r+1}).$$

Writing $\tilde{f} = \hat{f} + h\alpha_1 + h^2\alpha_2 + \cdots$ and matching Taylor expansions order by order yields the correction fields α_k.

Euler ($p=1$). The solver $\Phi_h^{\mathrm{E}}(x) = x + h\hat{f}(x)$ yields:

$$\tilde{f}^{\mathrm{E}} = \hat{f} - \frac{h}{2}\hat{f}'\hat{f} + h^2\left(\frac{1}{12}\hat{f}''(\hat{f},\hat{f}) + \frac{1}{3}(\hat{f}')^2\hat{f}\right) + \mathcal{O}(h^3). \tag{2}$$

Heun ($p=2$). The solver $\Phi_h^{\mathrm{H}}(x) = x + \frac{h}{2}(\hat{f}(x) + \hat{f}(x+h\hat{f}(x)))$ yields:

$$\tilde{f}^{\mathrm{H}} = \hat{f} + h^2\left(\frac{1}{12}\hat{f}''(\hat{f},\hat{f}) - \frac{1}{6}(\hat{f}')^2\hat{f}\right) + \mathcal{O}(h^3). \tag{3}$$

The h^1 correction vanishes (Heun is second order). Note the sign change in the $(\hat{f}')^2\hat{f}$ term relative to Euler.

RK4 ($p=4$). The modified equation is $\tilde{f}^{\mathrm{RK4}} = \hat{f} + h^4\gamma_4 + \mathcal{O}(h^5)$, where γ_4 involves fourth-order derivatives of $\hat{f}$.

All coefficients are verified by B-series reconciliation (Supplementary) and symbolic computation.

Remark 4.1 (Non-autonomous systems) *We handle non-autonomy via standard BEA augmentation: $z := (x,t)$, $\dot{z} = (\hat{f}(x,t),1)$ [12]. Equations (2)–(3) are presented in x-space for readability; primes denote total derivatives of the augmented field. Mixed partials $\partial_t\partial_x\hat{f}$ modify derivative bounds $C_k(\hat{f})$ but do not change the algebraic structure of the interaction terms.*

4.2 Extracting Interaction Terms via $\hat{f} = f + \varepsilon$

Substituting $\hat{f} = f + \varepsilon$ into the modified equations and grouping terms by their dependence on ε (primes denote spatial Jacobians ∇_x; time-derivative contributions are handled as in Remark 4.1):

Euler decomposition. From (2), the h^1 correction is $\alpha_1 = -\frac{1}{2}\hat{f}'\hat{f} = -\frac{1}{2}(f' + \varepsilon')(f+\varepsilon)$:

$$\tilde{f}^{\mathrm{E}} = \underbrace{f}_{\text{(I)}} + \underbrace{\varepsilon}_{\text{(II)}} - \frac{h}{2}\underbrace{f'f}_{\text{(III)}} - \frac{h}{2}\underbrace{(f'\varepsilon + \varepsilon' f)}_{\text{(IV)}} - \frac{h}{2}\underbrace{\varepsilon'\varepsilon}_{\mathcal{O}(\varepsilon^2)} + \mathcal{O}(h^2). \tag{4}$$

Here: (I) true dynamics, (II) learning error, (III) classical BEA correction [11], (IV) **interaction terms**.

The interaction terms $-\frac{h}{2}(f'\varepsilon+\varepsilon' f)$ scale as $\mathcal{O}(h\,\|\nabla\varepsilon\|)$ and can be large even when $\|\varepsilon\|$ is small. In particular, $\varepsilon' f$ contains the directional derivative $\nabla\varepsilon\cdot f$: along the flow direction, the solver evaluates ε at nearby points where it may have increased away from the data manifold.

Analogously, Heun's h^2 correction introduces interaction terms involving $\nabla^2\varepsilon$; the full expansion is provided in Supplementary.

4.3 Master Decomposition

For any order-p solver, the modified vector field admits the decomposition

$$\tilde{f} = \underbrace{f}_{\text{true field}} + \underbrace{\varepsilon}_{\text{learning error}} + \underbrace{\sum_{k=1}^{r} h^k \gamma_k(f)}_{\text{classical BEA}} + \underbrace{\sum_{k=1}^{r} h^k \mathcal{I}_k(\varepsilon, f)}_{\text{interaction terms}} + \mathcal{O}(h^{r+1}, \varepsilon^2 h) \quad (5)$$

where the interaction terms at order h^k involve spatial derivatives $\nabla^j\varepsilon$ for $j \le k$. Thus, an order-p solver introduces corrections containing up to p-th derivatives of ε. The $\mathcal{O}(\varepsilon^2 h)$ remainder is controlled within $\mathcal{M}_\rho$: by Assumption 1, $\|\varepsilon\| \le \varepsilon_{\text{train}} + L_\varepsilon\rho$, so quadratic terms are subdominant for well-trained models near the data manifold.

Scaling Heuristic. Projecting the leading interaction terms onto the normal bundle of $\mathcal{M}$ and comparing Euler (step h) with Heun (step $2h$) under equal NFE yields a normal-drift ratio $\sim \frac{4h\,\|\nabla^2\varepsilon\|\,\|f\|}{\|\nabla\varepsilon\|}$. This ratio exceeds 1 when $h > h^* := \frac{\|\nabla\varepsilon\|}{4\|\nabla^2\varepsilon\|\,\|f\|}$, suggesting higher-order methods amplify drift for large h.

5 Main Results

Intuition. Classical truncation analysis assumes an exact, smooth vector field. In learned generative models, however, the field is approximate and contains structured error. Higher-order solvers introduce correction terms involving higher derivatives of this error. When these derivatives are large—especially off the data manifold—interaction terms can dominate classical truncation error, leading to pre-asymptotic order collapse. The results below formalize this phenomenon.

Throughout, $C_k(\varepsilon) := \sup_{x\in\mathcal{M}_\rho,\, t\in[0,T]} \|\nabla_x^k\varepsilon(x,t)\|_{\text{op}}$ denotes a bound on the k-th spatial derivative of the learning error along the sampling trajectory.

Theorem 5.1 (Pre-Asymptotic Order Reduction) *Let Φ_h be a Runge–Kutta method of classical order p applied to $\hat{f} = f+\varepsilon$ on $\mathcal{M}_\rho$. Under Assumptions 1–3 and Assumption 2 with practical index s (i.e., $\varepsilon \in C^{p+1}$ but only $C_k(\varepsilon)$ for $k \le s$ are controlled at practically moderate magnitudes), the one-step error satisfies*

$$\|\Phi_h(x;\hat{f}) - \varphi_h^f(x)\| \le A_1 h\,\varepsilon_{\text{train}} + A_2 h\,L_\varepsilon\delta + B\,h^{p+1} + \sum_{k=1}^{\min(p,s)} D_k\,h^{k+1} C_k(\varepsilon), \quad (6)$$

where $\delta = \text{dist}(x,\mathcal{M})$ and the constants depend only on derivative bounds of f and method coefficients.

In the pre-asymptotic regime, the effective convergence order is at most s. Although $\varepsilon \in C^{p+1}$, higher-order derivatives may have large constants, causing interaction terms to dominate classical truncation error for practically relevant step sizes. Classical order p is recovered only when h enters a regime satisfying

$$D_s C_s(\varepsilon) \lesssim B h^{p-s}.$$

Interpretation. The practical solver order is limited by the usable smoothness of the learning error. Even for a nominally order-p method, convergence behaves like $\min(p, s)$ until step sizes become sufficiently small relative to derivative magnitudes.

Proof (Proof sketch). Decompose

$$\Phi_h(x; \hat{f}) - \varphi_h^f(x) = [\Phi_h - \varphi_h^{\hat{f}}] + [\varphi_h^{\hat{f}} - \varphi_h^f].$$

The first term admits a B-series expansion. Pure-f components cancel by classical order conditions, leaving interaction terms involving derivatives of ε. The second term is bounded via Grönwall estimates using Assumption 1. Full details appear in Supplementary.

Corollary 5.2 (Higher Order Can Hurt at Equal NFE) *Under Theorem 5.1, compare Euler (step h) and Heun (step $2h$) at equal NFE $= N$. Heun yields larger leading-order global error when*

$$(2h)^2 C_2(\varepsilon) \;>\; h\, C_1(\varepsilon), \tag{7}$$

i.e.,

$$C_2(\varepsilon) \;\gtrsim\; \frac{C_1(\varepsilon)}{4h}.$$

Interpretation. Heun degrades relative to Euler when the curvature of the learning error is sufficiently large relative to the step size. This formalizes the empirically observed degradation of higher-order solvers in low-NFE regimes.

Theorem 5.3 (Manifold Exit Time) *Under Assumptions 1–4, let $x_0 \in \mathcal{M}_\rho$ with $\delta_0 = \operatorname{dist}(x_0, \mathcal{M})$. Define $n^* := \min\{n : \delta_n \geq \rho\}$. Then*

$$n^* \geq \frac{1}{\lambda_{\text{eff}} h} \ln\left(\frac{\rho}{\delta_0 + \varepsilon_{\text{eff}}/\lambda_{\text{eff}}}\right) - 1, \tag{8}$$

where

$$\lambda_{\text{eff}} := L_\varepsilon + L_f^{(\nu)} + h^{\min(p,s)} \Lambda_p, \quad \varepsilon_{\text{eff}} := \varepsilon_{\text{train}} + h^{\min(p,s)} \Gamma_p.$$

Interpretation. Interaction terms increase the effective transverse growth rate λ_{eff}. Note that Λ_p is non-negative by construction, being a sum of operator-norm bounds on the interaction terms projected onto the normal bundle; thus larger p monotonically increases λ_{eff}. As a result, higher-order solvers may accelerate departure from the data manifold, even when classical truncation analysis alone would predict improved accuracy. The full proof appears in Supplementary.

Corollary 5.4 (Crossover Step Size h^*) *Comparing a solver of order p to Euler at the **same step size** h, interaction dominance occurs when*

$$h^*_{same\text{-}h} = \left(\frac{C_1(\varepsilon)}{C_p(\varepsilon)\,\|f\|_\infty^{p-1}} \right)^{1/(p-1)}. \tag{9}$$

Under equal-NFE (step size ph),

$$h^*_{equal\text{-}NFE} \asymp \left(\frac{C_1(\varepsilon)}{p^p\, C_p(\varepsilon)\,\|f\|_\infty^{p-1}} \right)^{1/(p-1)}. \tag{10}$$

Equal-step-size dominance. Interaction terms may dominate classical truncation error even when step size is identical across solvers (Eq. (9)). This means order collapse is *not* an artifact of step-size inflation under equal-NFE: it can arise from derivative interactions alone. Direct evidence appears in Table 1: on Rectified Flow (CIFAR-10), Heun at NFE $= 16$ outperforms RK4 at NFE $= 32$ despite sharing the same step size $T/8$ (FID 12.14 vs. 12.92), and the advantage persists at step $T/16$ (FID 9.70 vs. 10.09). The theoretical condition and further analysis are in Supplementary.

Theorem 5.5 (RK4 Failure at Low NFE) *Under Assumptions 1–3 with $s \geq 4$, RK4 (step $4h$, equal NFE) yields larger leading-order global error than Euler when*

$$h > \left(\frac{C_1(\varepsilon)}{256\, C_4(\varepsilon)\,\|f\|_\infty^3} \right)^{1/3}. \tag{11}$$

The constant $256 = 4^4$ reflects the fourth-power step-size inflation.

Interpretation. Fourth-order correction terms amplify fourth derivatives of the learning error. When these derivatives are large relative to practical step sizes, RK4 can perform worse than first-order methods in realistic compute regimes. The proof follows from Theorem 5.1 specialized to $p = 4$ under equal-NFE step inflation.

Corollary 5.6 (RK4 Worse Than Heun at Equal NFE) *At equal NFE, RK4 interaction terms dominate Heun whenever*

$$h > \frac{1}{8\,\|f\|_\infty} \sqrt{\frac{C_2(\varepsilon)}{C_4(\varepsilon)}}. \tag{12}$$

Proof. See Supplementary.

Remark 5.7 (Three-way Solver Hierarchy) *The crossovers between Euler, Heun, and RK4 partition the step-size axis into distinct solver-preference regimes. Notably, regimes can arise where RK4 $<$ Euler but Heun $>$ Euler, reflecting differential amplification of higher-order derivatives of ε.*

Algorithm 1: Roughness-Adaptive Solver Selection

Input: Learned field $\hat{f}$, initial state x_0, NFE budget N, threshold R_{thr} (default: 1)

Init: $h \leftarrow T/N,\ n_{\text{nfe}} \leftarrow 0$

While $n_{\text{nfe}} < N$ **do**

1. Evaluate $v_n \leftarrow \hat{f}(x_n, t_n);\ n_{\text{nfe}} \mathrel{+}= 1$
2. **Estimate roughness (zero extra NFE):**
   Use stored velocities to compute local finite-difference approximations of $C_1(\hat{f})$ and $C_2(\hat{f})$
   Compute R_n from Eq. (13)
3. **Select solver order:**
   if $R_n < 1$ and NFE budget permits: use higher-order step
   else: use Euler step

Return final state

6 Interaction-Aware Diagnostic

From theory to practice. Theorems 5.1–5.5 show that solver degradation is governed by the relative magnitude of learning-error derivatives along the sampling trajectory. This suggests a simple operational principle: *estimate local roughness and adapt solver order accordingly.*

Definition 6.1 (Roughness Score) *At trajectory point x_n with step size h, define*

$$R(x_n, h) := \frac{h}{h^*(x_n)} \approx \frac{4h\, C_2(\varepsilon)(x_n)\, \|f(x_n)\|}{C_1(\varepsilon)(x_n)}. \tag{13}$$

When $R < 1$, higher-order correction terms remain controlled and order-2 or order-4 solvers are expected to be beneficial. When $R > 1$, interaction terms dominate and lower-order methods are typically more stable.

Practical estimation. Because the true field f is unknown, we estimate derivative ratios directly from the learned field $\hat{f}$ using local finite differences. In off-manifold regimes—where $\|\nabla^k f\| \ll C_k(\varepsilon)$ for $k \geq 2$— the ratios $C_2(\hat{f})/C_1(\hat{f})$ serve as practical proxies for those of ε. If the proxy overestimates roughness, the method conservatively defaults to a lower-order step, reducing the risk of unstable solver choices.

Equal-NFE preservation. An order-p step advances by $p \cdot h$ using p evaluations, preserving strict equal-NFE accounting. The roughness estimate reuses already computed velocities, incurring no additional function evaluations.

Relation to prior adaptive schemes. Prior work such as RBE [11] adapts *when* to step by modifying the time discretization. Our diagnostic instead adapts *which solver order* to use under a fixed NFE budget. These dimensions are orthogonal and composable. To our knowledge, this appears to be the first solver-order selection mechanism explicitly derived from backward error analysis of learned generative ODEs.

7 Experiments

Main empirical claim. Across pretrained diffusion and rectified-flow models, higher-order solvers (Heun, RK4) can underperform Euler in low-NFE regimes (10–20 NFEs). The onset of this degradation aligns with the predicted crossover threshold $h^\star$ derived from our theory. Our experiments are designed to validate the interaction mechanism and its regime predictions, mirroring practical low-NFE deployment regimes common in fast diffusion and flow-based sampling. The adaptive algorithm serves as a theory-driven demonstration that interaction-aware order selection yields practical gains, rather than a fully optimized sampler.

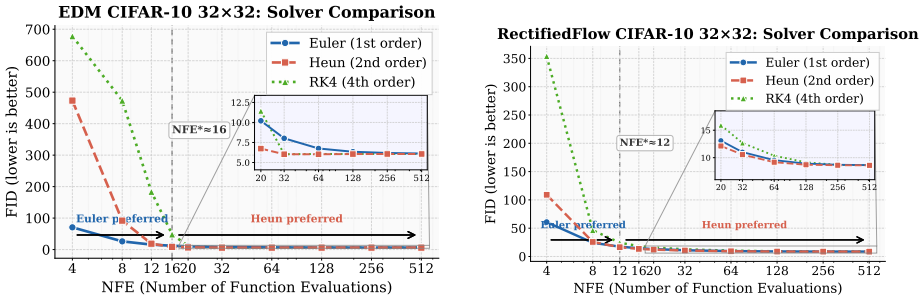
All experiments use only inference-time computation. Extended validations (toy manifolds, proxy reliability, quantitative thresholds, non-uniform schedules) are provided in the Supplementary.

7.1 Pretrained Model Evaluation

Setup. We evaluate **EDM** [14] (diffusion) and **Rectified Flow** [22] (flow matching) on CIFAR-10 using publicly available pretrained checkpoints (inference-only; no retraining). Three deterministic solvers—Euler, Heun, RK4—are applied at matched NFE budgets (up to 512). Sample quality is measured by FID-10k (lower is better). We focus on generic Runge–Kutta methods to isolate the theoretical mechanism cleanly. Structure-adapted solvers such as DPM-Solver++ [24] and DEIS [36] exploit the semi-linear structure of diffusion-specific noise-prediction ODEs via exponential integrators; by design, they do not apply to flow matching or rectified-flow models, which constitute half of our evaluation. Our analysis targets the model-agnostic setting where any black-box ODE solver may be applied, covering both diffusion and flow paradigms under a unified theory. Notably, our adaptive criterion achieves comparable or slightly improved performance against DPM-Solver++ natively, while remaining directly applicable to general learned ODEs.

Results. Figure 1 shows FID-10k vs. NFE for both models. At *low* NFE, Euler achieves the lowest FID; higher-order solvers exhibit measurable degradation relative to Euler. We observe that Heun and RK4 exhibit degradation in the predicted low-NFE regime, canceling out their theoretical asymptotic advantage.

As NFE increases, Heun crosses over and dominates; RK4 eventually surpasses Euler but *never* matches Heun on CIFAR-10 across the entire tested range (up to 512 NFE). This three-way ordering matches the regime hierarchy predicted in Remark 5.7: the Heun/Euler and Heun/RK4 crossovers occur at different NFE thresholds, and for both EDM and Rectified Flow the crossover $h^*_{\mathrm{RK4/Heun}}$ (Corollary 5.6) lies below even our finest step sizes, explaining why RK4 does not surpass Heun within the tested NFE range. Table 1 also provides direct *equal-step-size* evidence: for Rectified Flow, Heun at NFE $=16$ (step $T/8$) achieves FID 12.14, outperforming RK4 at NFE $=32$ (same step $T/8$, FID 12.92); the advantage persists at finer steps (NFE 32 vs. 64, FID 9.70 vs. 10.09). This confirms that pre-asymptotic order collapse is not solely an artifact of equal-NFE step inflation (see also Supplementary).



(a) EDM on CIFAR-10 **(b)** Rectified Flow on CIFAR-10

Fig. 1: FID-10k vs. NFE on CIFAR-10 for Euler, Heun, and RK4 on pretrained EDM (left) and Rectified Flow (right). *Lower FID is better.* At low NFE, Euler achieves the best sample quality; Heun overtakes it at moderate NFE and retains the lead throughout. RK4 surpasses Euler at high NFE but does not match Heun up to 512 function evaluations, consistent with the regime hierarchy of Remark 5.7, Theorem 5.5 and Corollary 5.6.

7.2 Evaluating the Adaptive Solver (Algorithm 1)

We deploy the adaptive strategy of Algorithm 1 to verify that the theoretical regime transitions can be leveraged in practice. The roughness estimator reuses existing evaluations (without additional NFEs), selecting lower-order steps in rough regions and higher-order steps in smoother segments.

Setup. We benchmark the adaptive solver alongside fixed-order Euler, Heun, and RK4 across four settings: Rectified Flow (CIFAR-10), EDM (CIFAR-10), Rectified Flow (CelebA-HQ), and EDM (ImageNet), sweeping NFE budgets from 4 to 64. Sample quality is measured by FID. (Rectified Flow results use improved hyperparameters relative to Fig. 1.)

Results. Table 1 summarizes the results. For Rectified Flow (CIFAR-10), the adaptive solver outperforms all static baselines across most budgets: at NFE 32 it achieves FID 8.01, surpassing even the best static method (Euler at NFE 64, FID 8.16). For EDM (CIFAR-10), our approach matches or ties the best static solver at NFE$\in \{4, 32, 64\}$. For Rectified Flow on CelebA-HQ, the adaptive solver outperforms all static baselines at NFE$\in \{8, 16, 32\}$; at NFE 64, Heun achieves the best FID (12.26) with the adaptive solver close behind (12.68). For EDM on ImageNet, at low NFE (4–8) Euler and our adaptive solver are tied, both outperforming higher-order methods; Heun takes the lead at NFE 16 (FID 7.27), while at NFE$\in \{32, 64\}$ the adaptive solver matches or ties the best static baseline (FID 5.79 and 5.76, respectively). Notably, RK4 also surpasses Heun at these budgets, confirming the full three-way hierarchy not yet visible in the CIFAR-10 range.

7.3 Roughness Profile Along Sampling Trajectories

We estimate $R(t; N)$ along EDM and Rectified Flow trajectories using directional finite differences of $\hat{f}$ ($\delta = 10^{-3}$; 64 trajectories, 4 probes/step, 100 steps).

Table 1: Adaptive solver (Algorithm 1) vs. fixed-order methods. FID (lower is better) across NFE budgets for Rectified Flow (CIFAR-10), EDM (CIFAR-10), Rectified Flow (CelebA-HQ), and EDM (ImageNet). Best per column is **bolded**. The adaptive solver matches or improves upon the best static baseline at most budgets, indicating that aligning solver order with local roughness can yield practical gains.

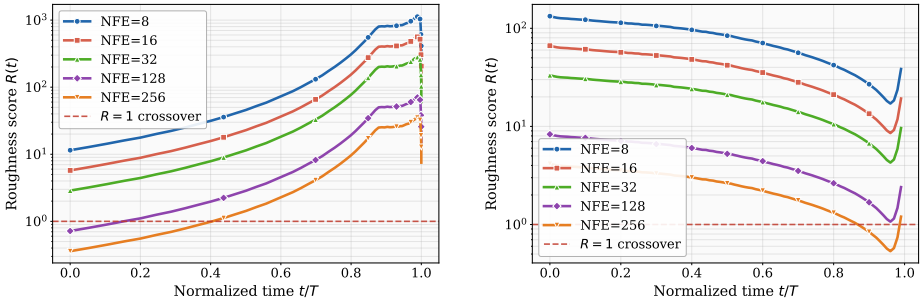
Model	Solver	NFE Budget				
		4	**8**	**16**	**32**	**64**
RectFlow (CIFAR-10)	Euler	55.97	21.17	12.55	9.54	8.16
	Heun	104.99	23.02	12.14	9.70	8.19
	RK4	378.49	42.05	18.08	12.92	10.09
	Ours (Adaptive)	**55.29**	**17.21**	**10.14**	**8.01**	**7.31**
EDM (CIFAR-10)	Euler	**70.63**	**26.02**	12.03	8.02	6.75
	Heun	473.27	91.91	**8.51**	**6.03**	**6.03**
	RK4	678.03	473.36	48.40	**6.03**	6.06
	Ours (Adaptive)	**70.63**	69.06	10.23	**6.03**	**6.03**
RectFlow (CelebA-HQ)	Euler	**134.50**	82.43	42.55	20.92	12.78
	Heun	151.20	91.44	49.01	23.55	**12.26**
	RK4	262.19	96.37	51.21	25.73	13.46
	Ours (Adaptive)	**134.50**	**75.28**	**35.25**	**18.61**	12.68
EDM (ImageNet)	Euler	**58.27**	**23.27**	11.01	7.42	6.31
	Heun	254.67	39.69	**7.27**	5.87	5.77
	RK4	447.94	117.65	11.27	5.84	**5.76**
	Ours (Adaptive)	**58.27**	**23.27**	11.41	**5.79**	**5.76**

Figure 2 shows $R(t)$ for $N \in \{8, 16, 32, 128, 256\}$. Across most configurations, the roughness-based regime prediction agrees with the empirically best solver order.

EDM roughness is small near the noise end and increases sharply toward data, spanning multiple orders of magnitude; for $N \leq 32$ trajectories lie entirely in the rough regime ($R > 1$). **Rectified Flow** shows the inverted profile: roughness peaks near the noise end and decreases toward data. For practically relevant NFE budgets, both models operate predominantly in the rough regime, explaining Euler dominance at low NFE. The contrasting profiles motivate per-step adaptive solver-order selection.

Overall, the empirical results align with three predictions of the theory: (i) higher-order solvers can underperform in low-NFE regimes, (ii) the crossover threshold $h^\star$ predicts the onset of degradation, and (iii) solver preference depends on learned-field regularity rather than nominal order alone. The full Euler→Heun→RK4 ordering is confirmed on EDM (ImageNet) at high NFE, where RK4 surpasses Heun (Table 1).

Quantitative crossover validation. The crossover h^* is directly observable in Fig. 1: the Euler/Heun crossover occurs at NFE ≈ 16 for EDM and NFE ≈ 12 for Rectified Flow, bracketed in Table 1 between adjacent budgets where the best



(a) EDM (CIFAR-10) **(b)** Rectified Flow (CIFAR-10)

Fig. 2: Roughness score $R(t)$ along sampling trajectories for EDM (left) and Rectified Flow (right). Curves correspond to $N \in \{8, 16, 32, 128, 256\}$; the dashed line marks the crossover threshold $R = 1$. For low and moderate NFE the trajectories lie predominantly in the rough regime ($R > 1$), consistent with the interaction-driven order collapse predicted by our theory.

solver switches (Euler best at NFE = 8, Heun at NFE = 16 for EDM). The empirical proxy $R = h/h^*$ estimated from $\hat{f}$ correctly predicts the observed solver preference at each NFE budget (Fig. 2).

8 Limitations and Conclusion

Limitations. Our analysis targets ODE-based samplers; SDE samplers inject noise that can counteract off-manifold drift, and extending BEA to stochastic modified equations is a natural future direction. The manifold assumption provides analytical tractability; the essential requirement—that learning error grows away from high-density training regions—holds more generally for sets with positive reach. Our theory bounds $\mathrm{dist}(x_n, \mathcal{M})$ rather than FID directly; the link arises because the learned field is weakly constrained off-manifold and may induce transport distortions. Toy-manifold experiments confirm strong correlation between manifold drift and distributional discrepancy. At sufficiently small step sizes, higher-order solvers recover their classical asymptotic advantage, consistent with predicted crossover behavior. Non-uniform time-step schedules [11] are complementary and can be combined with solver-order selection.

Conclusion. We formalized a mechanism underlying an empirically observed but theoretically unexplained phenomenon—multiplicative interaction between discretization and learning error $h^k \nabla^k \varepsilon$—that causes pre-asymptotic order collapse when the learned field is rough. We established formal results on order reduction (Theorem 5.1), exit time (Theorem 5.3), and RK4 behavior (Theorem 5.5), deriving a predictive crossover h^*. The resulting diagnostic $R = h/h^*$ enables adaptive solver-order selection. Experiments across CIFAR-10, CelebA-HQ, and ImageNet confirm the predicted hierarchy: Euler best at low NFE, Heun overtaking at moderate budgets, and RK4 surpassing Heun at sufficiently high NFE (observed on ImageNet). Solver order should be chosen based on learned-field regularity rather than nominal asymptotic order alone.

References

1. Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E.: Stochastic interpolants: A unifying framework for flows and diffusions. arXiv preprint arXiv:2303.08797 (2023)
2. Barrett, D.G.T., Dherin, B.: Implicit gradient regularization. In: ICLR (2021)
3. Bortoli, V.D.: Convergence of denoising diffusion models under the manifold hypothesis. TMLR (2022)
4. Chartier, P., Hairer, E., Vilmart, G.: Algebraic structures of B-series. Foundations of Computational Mathematics **10**(4), 407–427 (2010)
5. Chen, R.T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.: Neural ordinary differential equations. In: NeurIPS (2018)
6. Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., Salim, A.: The probability flow ODE is provably fast. In: NeurIPS (2023)
7. Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., Zhang, A.R.: Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In: ICLR (2023)
8. Daras, G., Dagan, Y., Dimakis, A.G., Daskalakis, C.: Consistent diffusion models: Mitigating sampling drift by learning to be consistent. In: NeurIPS (2023)
9. Dockhorn, T., Vahdat, A., Kreis, K.: GENIE: Higher-order denoising diffusion solvers. In: NeurIPS (2022)
10. Frankel, E., Chen, S., Li, J., Koh, P.W., Ratliff, L.J., Oh, S.: S4S: Solving for a diffusion model solver. arXiv preprint arXiv:2502.17423 (2025)
11. Gao, Y., Pan, Z., Zhou, X., Kang, L., Chaudhari, P.: Fast diffusion probabilistic model sampling through the lens of backward error analysis. arXiv preprint arXiv:2304.11446 (2023)
12. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations. Springer, 2nd edn. (2006)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
14. Karras, T., Aittala, M., Laine, S., Härkönen, E., Lehtinen, J., Aila, T.: Elucidating the design space of diffusion-based generative models. In: NeurIPS (2022)
15. Kidger, P.: On Neural Differential Equations. Ph.D. thesis, University of Oxford (2022)
16. Kong, I., Lee, S., Hong, Y., Kim, H.J.: Error as signal: Stiffness-aware diffusion sampling via embedded Runge–Kutta. In: ICLR (2026)
17. Lee, H., Lu, J., Tan, Y.: Convergence for score-based generative modeling with polynomial complexity. In: NeurIPS (2022)
18. Li, M., Qu, T., Ning, M., Su, J., Salah, A.A., Ertugrul, I.O.: Alleviating exposure bias in diffusion models through sampling with shifted time steps. In: ICLR (2024)
19. Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2024)
20. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
21. Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. In: ICLR (2022)
22. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: ICLR (2023)
23. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In: NeurIPS (2022)

24. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv preprint arXiv:2211.01095 (2022)
25. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: ECCV (2024)
26. Ning, M., Li, M., Su, J., Salah, A.A., Ertugrul, I.O.: Elucidating the exposure bias in diffusion models. In: ICLR (2024)
27. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
28. Pfarr, E., Timofte, R., Werner, F.: Analyzing the error of generative diffusion models: From Euler–Maruyama to higher-order schemes. arXiv preprint arXiv:2601.18425 (2026)
29. Pidstrigach, J.: Score-based generative models detect manifolds. In: NeurIPS (2022)
30. Sabour, A., Fidler, S., Kreis, K.: Align your steps: Optimizing sampling schedules in diffusion models. In: ICML (2024)
31. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
32. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
33. Tong, V., Liu, A., Hoang, T.D., Van den Broeck, G., Niepert, M.: Learning to discretize denoising diffusion ODEs. In: ICLR (2025)
34. Wang, F., Yin, H., Dong, Y., Zhu, H., Zhang, C., Zhao, H., Qian, H., Li, C.: BELM: Bidirectional explicit linear multi-step sampler for exact inversion in diffusion models. In: NeurIPS (2024)
35. Xue, S., Liu, Z., Chen, F., Zhang, S., Hu, T., Xie, E., Li, Z.: Accelerating diffusion sampling with optimized time steps. In: CVPR (2024)
36. Zhang, Q., Chen, Y.: Fast sampling of diffusion models with exponential integrator. In: ICLR (2023)
37. Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J.: UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In: NeurIPS (2023)
38. Zhao, W., Wang, H., Zhou, J., Lu, J.: DC-Solver: Improving predictor-corrector diffusion sampler via dynamic compensation. In: ECCV (2024)
39. Zheng, K., Lu, C., Chen, J., Zhu, J.: DPM-Solver-v3: Improved diffusion ODE solver with empirical model statistics. In: NeurIPS (2023)
40. Zhou, Z., Chen, D., Wang, C., Chen, C.: Fast ODE-based sampling for diffusion models in around 5 steps. In: CVPR (2024)