SECOND-ORDER GEOMETRY OF GENERATIVE PATHS: FRAGILITY AND TARGETED REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative probability paths can agree at their endpoints and in first-order transport summaries yet react differently to coarse sampling or small perturbations. We characterize the missing information at second order. For a dynamic path (μ_t, v_t), a metric-compatible derivative on the moving $L^2(\mu_t)$ bundle yields an exact Frenet decomposition of material acceleration into schedule-dependent speed variation and reparameterization-invariant bending. A first-order blindness theorem shows that endpoints, path length, and kinetic action may all remain fixed while bending energy becomes arbitrarily large, excluding any uniform upper certificate for bending energy based only on those summaries. A discrete-to-continuum error analysis then yields a fixed-budget evaluation density, while a spectral law identifies locally most-expanding perturbation directions. We also establish consistency of the material finite-difference estimator used at scale. Controlled constant-speed flows isolate bending; in CIFAR-10 Rectified Flow++, acceleration risk adds information beyond velocity and path length, and an independent 20,480-sample confirmation shows that it places the same refinement budget more effectively than random, uniform, or velocity allocation. In image editing, amplifying leading expansion modes increases non-target DINOv2, LPIPS, and RMSE drift relative to an orthogonal placebo. The results support two operational second-order risk channels associated with realized generative fragility, rather than a universal quality score or a faster sampler.

1 INTRODUCTION

A generator can have a strong high-precision endpoint and still degrade when sampled with only a few function evaluations. The same trajectory-level fragility can appear in conditional settings: for example, an image editing model may preserve the intended target while introducing changes in regions that should remain fixed. These failures are usually analyzed through solver order, score error, guidance strength, or endpoint quality. Those explanations are important, but they leave a simpler question unresolved: *what property of the path determines how well it tolerates coarse discretization or perturbation?*

Diffusion models, probability-flow ODEs, flow matching, rectified flow, stochastic interpolants, and consistency methods all specify or learn paths from a simple prior to data (Ho et al., 2020; Song et al., 2021b; Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2025; Song et al., 2023). Existing methods already make these paths easier to integrate by straightening them, changing the coupling, or changing the solver (Lee et al., 2023; Pooladian et al., 2023; Lu et al., 2022; Zhao et al., 2023). Recent work also uses trajectory curvature directly for conformal rejection or adaptive step placement (Li & Boström, 2025; Li et al., 2026). Geometric formulations place generation on Riemannian or Wasserstein spaces (Chen & Lipman, 2024; Haviv et al., 2025; Atanackovic et al., 2025; Zhang et al., 2026; Karczewski et al., 2026). These advances establish that path geometry can be useful, but first-order path summaries do not fully explain why generative trajectories respond differently to coarse integration and perturbations. We therefore ask:

What second-order geometric information do first-order path summaries miss about fragility under few-step integration and perturbations?

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

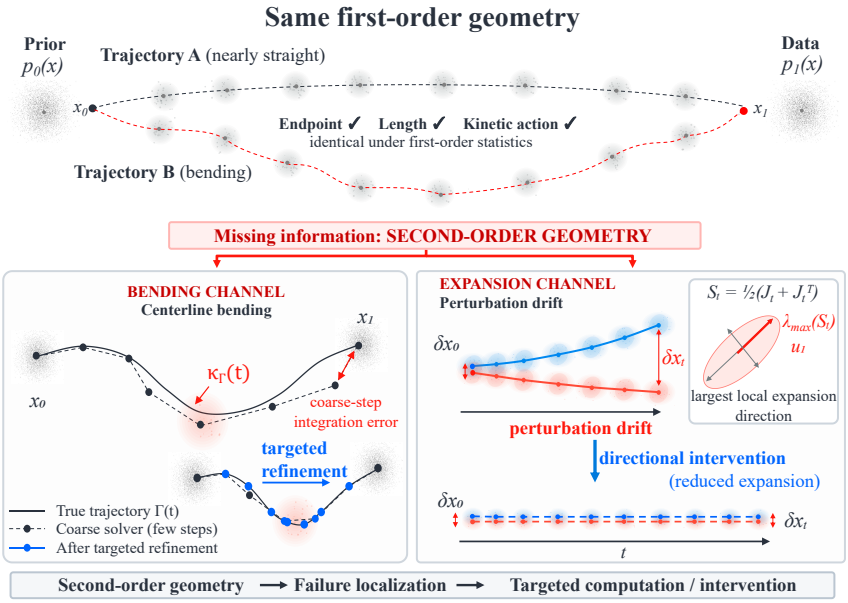


Figure 1: **The endpoint paradox and two second-order channels.** Trajectories with matched endpoints and first-order transport statistics can differ in second-order geometry. Material acceleration enters coarse-step integration error, with bending isolated under constant speed, while local expansion governs infinitesimal perturbation growth and motivates directional intervention.

Figure 1 shows the distinction we use throughout the paper. *Bending* describes direction change after speed variation is factored out, whereas *expansion* describes local separation of nearby trajectories. The theory isolates these mechanism-specific quantities and the experiments test whether they remain informative for realized coarse-step integration error and perturbation sensitivity in learned systems.

We formalize these ideas for a dynamic probability path $\Gamma = (\mu_t, v_t)$ satisfying the continuity equation. The measure curve lives in $\mathcal{P}_2(\mathbb{R}^d)$, while v_t records the implemented transport. The distinction matters because the Benamou–Brenier geometry assigns a canonical minimum-norm velocity to a measure curve (Benamou & Brenier, 2000; Ambrosio et al., 2008; Villani, 2009). Second-order Wasserstein calculus further distinguishes covariant acceleration and Jacobi fields from properties of an arbitrary realization (Otto, 2001; Lott, 2008; Gigli, 2012). A neural generator may realize another admissible velocity. Our identities are therefore stated for the implemented dynamic path and become intrinsic Wasserstein statements only for the canonical tangent. This avoids labeling an arbitrary neural-ODE acceleration as Ricci or sectional curvature of Wasserstein space.

We summarize our contributions as follows:

- An exact dynamic Frenet decomposition separates schedule-dependent speed variation from reparameterization-invariant bending. A constructive blindness theorem further shows that endpoints, path length, and kinetic action do not uniformly control bending energy, exposing second-order structure that these first-order summaries leave undetermined.
- Starting from the propagated Euler remainder, we relate material acceleration to coarse-step integration error and derive a fixed-budget allocation law through a discrete-to-continuum limit. We also prove first-order consistency of the material finite-difference estimator along exact trajectories. Controlled flows isolate the bending term, while RF++ experiments test the additional predictive value of implemented acceleration and its use for compute-matched local refinement.
- For perturbation growth, a spectral variational law identifies the locally most-expanding directions of the flow. We test this mechanism in image editing with a placebo-controlled intervention that compares leading-mode amplification against an equal-magnitude orthogonal direction, testing a mechanism distinct from centerline bending.

2 Related Work and Positioning

Generative path geometry and straightness. Trajectory-curvature objectives, multisample couplings, rectified-flow training, and straightness theory all seek easier few-step paths (Lee et al., 2023; Pooladian et al., 2023; Liu et al., 2023; Lee et al., 2024; Bansal et al., 2026); CurveFlow instead uses nonlinear paths for semantic alignment (Luo et al., 2026). FlowGuard uses velocity-jump curvature for per-sample rejection (Li & Boström, 2025), while CAT-Flow allocates steps from curvature without extra neural evaluations (Li et al., 2026). We analyze a fixed learned transport and ask what endpoints, path length, and action cannot determine, leading to a speed–bending split and a separate perturbation-expansion channel.

Adaptive numerical solvers. DPM-Solver, UniPC, and embedded Runge–Kutta methods control numerical error through solver structure or online defect estimates (Lu et al., 2022; Zhao et al., 2023; Dormand & Prince, 1980; Xiao, 2026). Our allocation law is complementary: it derives a propagated path-risk density, and the experiment tests only its localization implication under an equal added-call budget. The current offline risk cost rules out an end-to-end speedup claim.

Probability-path geometry and perturbation directions. Second-order Wasserstein calculus provides covariant derivatives and Jacobi fields (Otto, 2001; Lott, 2008; Gigli, 2012), while recent generative work places flows on Riemannian or Wasserstein spaces (Chen & Lipman, 2024; Haviv et al., 2025; Atanackovic et al., 2025; Zhang et al., 2026). Jacobian subspaces have also been used for diffusion editing (Chen et al., 2024). Our implemented-path identities become intrinsic only in the canonical tangent regime; the symmetric-Jacobian experiment instead tests the locally most-expanding direction against an equal-magnitude orthogonal placebo. A structured comparison with the closest lines of work is provided in Appendix A.

3 Second-Order Geometry of a Dynamic Probability Path

Let $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ have a smooth density and let v_t be a smooth velocity field satisfying

$$\partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0. \tag{1}$$

The measure motion and its implementation need not have the same first-order energy. Let $T_\mu \mathcal{P}_2$ be the $L^2(\mu)$-closure of gradient fields and P_μ the orthogonal projection onto it. The weighted Hodge decomposition gives $v = u + w$ with $u = P_\mu v \in T_\mu \mathcal{P}_2$ and $w \perp T_\mu \mathcal{P}_2$. In weak form $\nabla \cdot (\mu w) = 0$, so u and v generate the same infinitesimal measure evolution, while

$$\|v\|^2_{L^2(\mu)} = \|u\|^2_{L^2(\mu)} + \|w\|^2_{L^2(\mu)}\,. \tag{2}$$

Thus the canonical Wasserstein tangent is the minimum-energy representative, whereas an implemented particle flow may contain circulation invisible to the measure derivative.

For vector fields f, g, write $\langle f, g\rangle_t = \int f \cdot g \, \mathrm{d}\mu_t$ and $\|f\|_t^2 = \langle f, f\rangle_t$. Define the material derivative along the implemented transport by

$$D_t^\Gamma f = \partial_t f + D_x f \, v_t. \tag{3}$$

Using Eq. (1) gives the Reynolds identity

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle f, g\rangle_t = \left\langle D_t^\Gamma f, g\right\rangle_t + \left\langle f, D_t^\Gamma g\right\rangle_t, \tag{4}$$

so D_t^Γ is metric-compatible with the moving $L^2(\mu_t)$ inner product.

Define the dynamic acceleration $a_t = D_t^\Gamma v_t$, the transport speed $s(t) = \|v_t\|_t$, and, for $s(t) > 0$, the unit tangent $T_t = v_t / s(t)$. With $\mathrm{d}\ell = s(t)\,\mathrm{d}t$ and $D_\ell^\Gamma := s(t)^{-1} D_t^\Gamma$, define

$$\kappa_\Gamma(t) = \left\|D_\ell^\Gamma T_t\right\|_t = \frac{\left\|D_t^\Gamma T_t\right\|_t}{s(t)}. \tag{5}$$

When v_t is the canonical minimum-norm Wasserstein tangent, $s(t) = |\dot\mu_t|_{W_2}$ and the tangent projection of Eq. (3) recovers the Otto/Lott covariant derivative (Otto, 2001; Ambrosio et al., 2008; Lott, 2008; Gigli, 2012). For a smooth potential flow $v_t = \nabla\phi_t$, $a_t = \nabla(\partial_t\phi_t + \frac{1}{2}\|\nabla\phi_t\|^2)$ is already a gradient.

Proposition 1 (Canonical Wasserstein specialization)**.** *For every admissible implementation v_t, the metric derivative satisfies $|\dot{\mu}_t|_{W_2} \le s(t)$, with equality for the canonical minimum-$L^2(\mu_t)$ tangent. If that tangent is a smooth potential field, the material acceleration equals the Wasserstein covariant acceleration. Hence Theorem 1 is intrinsic in the canonical potential-flow regime and an implementation-level transport identity otherwise.*

Theorem 1 (Dynamic Frenet decomposition)**.** *Assume $s(t) > 0$. Then*

$$\|a_t\|_t^2 = |\dot{s}(t)|^2 + s(t)^4 \kappa_\Gamma(t)^2. \tag{6}$$

Moreover, κ_Γ is invariant under any C^1 increasing reparameterization of time with positive derivative, whereas $\dot{s}$ depends on the schedule.

Metric compatibility makes the two terms in Eq. (6) orthogonal. The material-acceleration norm therefore mixes schedule-induced speed variation with direction change; only the latter is the reparameterization-invariant bending term.

3.1 Why first-order summaries cannot determine bending

The split would be unimportant if first-order path summaries already controlled the bending term. They do not.

Theorem 2 (First-order blindness)**.** *Fix $m_0, m_1 \in \mathbb{R}^2$ and $L > \|m_1 - m_0\|$. For every $M > 0$ there exist two C^1, piecewise-C^2 translated-Gaussian paths $\mu_t^{(j)} = \mathcal{N}(m_j(t), \sigma^2 I)$, $j \in \{0, M\}$, such that*

$$\mu_0^{(0)} = \mu_0^{(M)}, \quad \mu_1^{(0)} = \mu_1^{(M)}, \quad \int_0^1 s_j(t)\,\mathrm{d}t = L, \quad \int_0^1 s_j(t)^2\,\mathrm{d}t = L^2, \tag{7}$$

while

$$\int_0^1 s_M(t)^4 \kappa_{\Gamma_M}(t)^2\,\mathrm{d}t - \int_0^1 s_0(t)^4 \kappa_{\Gamma_0}(t)^2\,\mathrm{d}t > M. \tag{8}$$

The paths can be chosen with constant speed $s_j(t) \equiv L$.

The construction inserts a shrinking full loop into one planar mean curve while compensating the length elsewhere. Endpoints, length, and constant-speed action stay fixed, but $\int \kappa^2\,\mathrm{d}\ell$ diverges as the loop radius goes to zero. Because translations of a fixed Gaussian satisfy $W_2(\mathcal{N}(m, \sigma^2 I), \mathcal{N}(n, \sigma^2 I)) = \|m - n\|$, the example is in probability space rather than a feature-space analogy.

Theorem 2 is an identifiability result, not a prevalence claim. The vanishing-radius loop is deliberately adversarial: it proves that the listed first-order summaries cannot determine bending without additional second-order regularity. It does not assert that trained generators typically contain such loops. This distinction matters because the theorem establishes missing information, while the experiments below ask whether a measurable version of that information is useful on learned flows.

Corollary 1 (No first-order upper certificate for bending energy)**.** *Over the translated-Gaussian class in Theorem 2, there is no finite-valued function of only the two endpoints, path length, and kinetic action that uniformly upper-bounds the bending energy $\int_0^1 s(t)^4 \kappa_\Gamma(t)^2\,\mathrm{d}t$.*

The point is not the loop itself but the missing information: any certificate based only on those first-order summaries assigns the same value to the constructed sequence while the second-order bending cost is unbounded. Additional regularity can exclude the sequence, but that regularity is extra information rather than a consequence of the matched summaries.

4 Second-Order Risk Laws and Optimal Allocation

4.1 From acceleration to coarse-step error

For an exact trajectory X_t of $\dot{X}_t = v_t(X_t)$,

$$X_{t+h} = X_t + h v_t(X_t) + \int_t^{t+h} (t + h - r) a_r(X_r)\,\mathrm{d}r. \tag{9}$$

If $\|D_x v_t\|_{\mathrm{op}} \le L(t)$ with $L \in C^1([0,1])$, define the endpoint propagation weight $G(t) = \exp(\int_t^1 L(r)\,\mathrm{d}r)$. Coupling the exact and Euler trajectories gives an exact discrete remainder bound in Appendix C, with propagation factor $G_{N,k+1} = \prod_{j=k+1}^{N-1}(1 + h_j L(t_j))$. On grids whose mesh tends to zero, $G_{N,k+1} = G(t_{k+1})(1+O(\max_j h_j))$ uniformly in k. Under the smoothness assumed above, the leading term therefore has the continuum form

$$\mathcal{E}[q] = \frac{1}{2}\int_0^1 \frac{w(t)}{q(t)}\,\mathrm{d}t, \qquad w(t) = G(t)\,\|a_t\|_t\,, \qquad \int_0^1 q(t)\,\mathrm{d}t = N, \tag{10}$$

where q is the evaluation density. The functional is not introduced independently of the solver bound; it is the continuum limit of its leading discrete term.

Proposition 2 (Discrete-to-continuum allocation limit)**.** *Let $w, p \in C^1([0,1])$, with $p(t) \ge p_{\min} > 0$ and $\int_0^1 p(t)\,\mathrm{d}t = 1$. Choose the grid so that $\int_{t_k}^{t_{k+1}} p(t)\,\mathrm{d}t = 1/N$, set $h_k = t_{k+1} - t_k$, and define the leading discrete remainder*

$$\mathcal{B}_N(p) = \frac{1}{2}\sum_{k=0}^{N-1} w(t_k) h_k^2. \tag{11}$$

Then, as $N \to \infty$,

$$\mathcal{B}_N(p) = \frac{1}{2N}\int_0^1 \frac{w(t)}{p(t)}\,\mathrm{d}t + O(N^{-2}). \tag{12}$$

Equivalently, with $q_N = Np$, the leading discrete bound converges to $\mathcal{E}[q_N]$ in Eq. (10) *up to $O(N^{-2})$.*

Theorem 3 (Bound-optimal evaluation density)**.** *Let $w \ge 0$ be measurable with $0 < \int_0^1 \sqrt{w} < \infty$. Minimize Eq.* (10) *over $q \ge 0$ with $\int_0^1 q = N$ and $q > 0$ a.e. on $\{w > 0\}$, taking $0/0 = 0$ on $\{w = q = 0\}$. The minimum is*

$$\mathcal{E}_N^* = \frac{1}{2N}\left(\int_0^1 \sqrt{w(t)}\,\mathrm{d}t\right)^2, \tag{13}$$

attained almost everywhere by

$$q^*(t) = N\frac{\sqrt{w(t)}}{\int_0^1 \sqrt{w(r)}\,\mathrm{d}r}. \tag{14}$$

If $\kappa_{\mathrm{acc}}(t) = \|a_t\|_t^2$, then $q^*(t) \propto G(t)^{1/2}\kappa_{\mathrm{acc}}(t)^{1/4}$. If $(1-\delta)w \le \widehat{w} \le (1+\delta)w$ for $\delta < 1$, then

$$\frac{\mathcal{E}[q_{\widehat{w}}]}{\mathcal{E}_N^*} \le \sqrt{\frac{1+\delta}{1-\delta}}. \tag{15}$$

The continuous law has a simple discrete counterpart that makes the connection to the experiment explicit.

Corollary 2 (Piecewise-constant discrete allocation)**.** *Partition $[0,1]$ into intervals I_i of width Δ_i and approximate w by w_i. Let $n_i \ge 0$, with $n_i > 0$ if $w_i > 0$ and $0/0 := 0$. Uniform allocation inside active intervals gives*

$$\mathcal{E} = \frac{1}{2}\sum_i \frac{w_i \Delta_i^2}{n_i}, \qquad \sum_i n_i = N, \tag{16}$$

and its continuous relaxation satisfies $n_i^* \propto \Delta_i\sqrt{w_i}$. For equal-width intervals, $n_i^* \propto \sqrt{w_i}$.

Proposition 2 and Corollary 2 separate two issues that are easy to conflate: the allocation law is forced by the leading propagated Euler remainder, while the experiment uses only a sparse integer approximation to that law. Specifically, we refine high-risk intervals by midpoint insertion rather than claim to implement Eq. (14) exactly, and we do not estimate the propagation factor $G(t)$. The empirical intervention therefore tests the acceleration-localization implication of the theory. Theorem 3 optimizes a propagated local-error envelope, not FID or a semantic metric.

4.2 SPECTRAL GROWTH OF PERTURBATIONS

Bending concerns the centerline. Perturbation growth is governed by a different object. Let $J_t = D_x v_t(X_t)$ and $S_t = (J_t + J_t^\top)/2$. An infinitesimal displacement ξ_t satisfies $\dot{\xi}_t = J_t \xi_t$.

Theorem 4 (Spectral perturbation-growth law)**.** *Along a differentiable trajectory,*

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\xi_t\right\|^2 = \xi_t^\top S_t \xi_t. \tag{17}$$

Among unit perturbations, the maximal instantaneous value of $\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\xi_t\right\|^2$ is $\lambda_1(S_t)$, attained by a leading eigenvector $e_1(t)$. If $\lambda_1(S_{t_0}) - \lambda_2(S_{t_0}) \geq \gamma > 0$, let ξ_{e_1} and ξ_η solve the variational equation with $\xi_{e_1}(t_0) = e_1(t_0)$ and $\xi_\eta(t_0) = \eta$, respectively, for any unit $\eta \perp e_1(t_0)$. Then

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\left\|\xi_{e_1}(t)\right\|^2 - \left\|\xi_\eta(t)\right\|^2\right)\Big|_{t=t_0} \geq 2\gamma. \tag{18}$$

If J_t is C^1 near t_0, then for small $h > 0$, $\left\|\xi_{e_1}(t_0+h)\right\|^2 - \left\|\xi_\eta(t_0+h)\right\|^2 \geq 2\gamma h - Ch^2$ for a finite C. For $t \geq s$, the fundamental matrix also satisfies $\left\|\Phi(t,s)\right\|_{\mathrm{op}} \leq \exp(\int_s^t \lambda_1(S_r)\,\mathrm{d}r)$.

The theorem identifies only the locally most-expanding infinitesimal direction. The intervention asks whether this local prediction survives the nonlinear frozen editor, using an equal-magnitude orthogonal perturbation as placebo.

5 OPERATIONAL ESTIMATORS

The theory uses $a_t = \partial_t v_t + D_x v_t\, v_t$ and S_t. White-box models use JVPs. The large-scale experiments use the material finite difference

$$\widehat{a}_k = \frac{v_{t_{k+1}}(X_{t_{k+1}}) - v_{t_k}(X_{t_k})}{t_{k+1} - t_k}. \tag{19}$$

The following statement gives the exact-trajectory consistency justification for the material finite difference.

Proposition 3 (Exact-trajectory consistency of the material finite difference)**.** *Let $g(t) = v_t(X_t)$ along an exact trajectory and suppose $g \in C^2$ with $\sup_t \left\|g''(t)\right\| \leq M$. For $h_k = t_{k+1} - t_k$,*

$$\left\|\widehat{a}_k - a_{t_k}(X_{t_k})\right\| \leq \frac{M}{2} h_k. \tag{20}$$

If additionally $\sup_t \left\|a_t(X_t)\right\| \leq A$, then

$$\left|\left\|\widehat{a}_k\right\|^2 - \left\|a_{t_k}(X_{t_k})\right\|^2\right| \leq AMh_k + \frac{M^2}{4}h_k^2. \tag{21}$$

Consequently, on a uniform mesh the top-q mean of sampled squared acceleration values is first-order consistent as well.

Proposition 3 covers exact trajectories; numerical state error is outside its scope and is assessed by JVP/finite-difference audits. A one-step versus two-half-step discrepancy is only a black-box truncation-risk proxy. Expansion uses Ritz/power iterations on the symmetric Jacobian. In the editing audit, 500 finite-difference estimates have median relative error 0.0722 and no non-finite values.

For RF++, the frozen sample statistic K_{top10} is the mean of the highest 10% time-point values of the squared material-acceleration norm. Earlier experiment logs used "curvature/bending" for this quantity; we call it *acceleration risk* here because Eq. (6) shows that material acceleration generally contains both speed variation and Frenet bending. Integrated acceleration remains a secondary summary. The empirical object is therefore implementation-level dynamic acceleration unless the learned field is the canonical Wasserstein tangent.

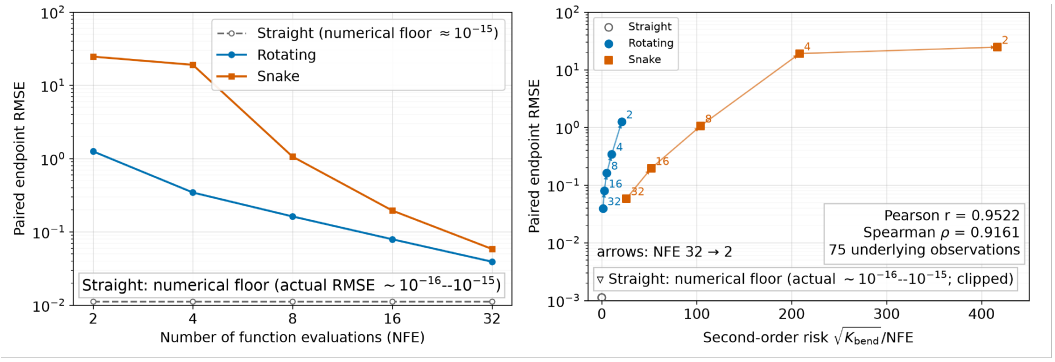


Figure 2: **Controlled bending risk tracks few-step discretization error. Left:** Euler endpoint RMSE decreases with NFE, with larger errors for higher-bending trajectories. **Right:** normalized bending risk $\sqrt{K_{\text{bend}}}/\text{NFE}$ strongly tracks paired RMSE across runs.

6 Experiments

The experiments test the preceding claims rather than optimize a single score. Controlled flows isolate pure bending. RF++ tests sample-specific prediction and compute placement. Editing tests the deformation channel and intervenes on a predicted direction. Development branches, estimator audits, and negative results are retained in the appendix rather than redefine successful endpoints.

6.1 Acceleration risk predicts few-step fragility beyond first-order metrics

Controlled flows. We integrate analytic low-, medium-, and high-bending paths with Euler at NFE $\in \{2, 4, 8, 16, 32\}$ and measure paired endpoint RMSE. Let $K_{\text{acc}} = \int_0^1 \|a_t\|_t^2 \, \mathrm{d}t$. Since these paths have constant speed, Theorem 1 makes K_{acc} exactly the weighted bending energy, isolating directional bending from speed-schedule effects. Across the canonical set, $\sqrt{K_{\text{acc}}}/\text{NFE}$ has Pearson 0.9522 and Spearman 0.9161 with paired RMSE. The broader continuous sweep remains positive (Spearman 0.7810; Appendix D.1.2) but reveals low-NFE aliasing and a high-NFE numerical floor. A true Eulerian affine branch retains Spearman 0.9869 (Appendix D.1.2). These controls support the bending mechanism but not universal linear calibration.

RF++ / DINOv2-primary evaluation. We use the public CIFAR-10 2-Rectified Flow++ model (Lee et al., 2024). Each low-NFE sample is paired with its NFE-200 trajectory from the same initial noise. The primary outcome is DINOv2-small CLS cosine deviation (Oquab et al., 2024). Same-noise pairing fixes the sample, so the comparison measures coarse integration rather than variation across generations. We use paired feature deviation rather than FID because the question is sample-level: which low-NFE trajectories depart from their high-precision references? FID is distribution-level and cannot provide that ranking.

At NFE $4/8/16$, frozen K_{top10} has Spearman $0.104/0.125/0.227$ and high-deviation AUROC $0.549/0.566/0.631$, above velocity ($0.509/0.531/0.552$) and nearly identical path-length baselines (Table 1, Figure 3). Controlling velocity and path length, acceleration-risk coefficients remain positive ($0.0666/0.0724/0.0521$) with 95% intervals excluding zero. The modest effects support incremental second-order information rather than strong standalone prediction. Integral acceleration is stronger but remains a predeclared secondary analysis (Appendix D.2.4).

6.2 Second-order risk is actionable under a matched local budget

Prediction alone would leave the work diagnostic. Starting from NFE=8, a validation $q_{0.8}$ gate selects high-risk samples; each selected sample receives two midpoint refinements. Random, uniform, velocity, and acceleration-risk strategies operate on the same selected samples and use the same added refinement-call budget. In the first 1,280-sample development run, acceleration risk beat uniform and velocity but did not significantly beat random. We therefore froze the gate and interval rule before evaluating 20,480 independent samples.

Table 1: RF++ predictive performance across NFE settings. The primary predictor is the top-10% acceleration risk K_{top10}. AUROC measures discrimination of high-DINOv2-degradation samples using validation-frozen labels, and adjusted β controls for first-order velocity and path-length risks.

NFE	Spearman	Acc. AUROC	Vel. AUROC	Path AUROC	Adjusted β_{acc} (95% CI)
4	0.1038	0.5493	0.5094	0.5095	0.0666[0.0559, 0.0773]
8	0.1252	0.5656	0.5308	0.5307	0.0724[0.0614, 0.0835]
16	0.2268	0.6310	0.5524	0.5523	0.0521[0.0419, 0.0623]
Pooled + NFE FE	-	-	-	-	0.0347[0.0288, 0.0407]

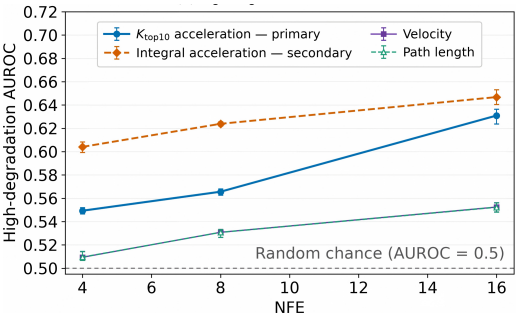


Figure 3: **High-degradation discrimination.** Acceleration risk exceeds velocity and path-length baselines across NFE.

Table 2: Compute-matched refinement under a fixed budget.

Method	Reduction ↑	95% CI
Random	0.000305	–
Uniform	0.000145	–
Velocity	0.000285	–
Acceleration risk	**0.000770**	–
Pairwise advantage		
Risk − Random	+0.000465	[0.000383, 0.000559]
Risk − Velocity	+0.000485	[0.000396, 0.000587]
Risk localization		
Above frozen $q_{0.8}$	0.003618	–
Below frozen $q_{0.8}$	0.001817	–
Above − Below	0.001801	[0.001234, 0.002442]

In confirmation, acceleration-risk allocation reduces paired DINOv2 drift by 0.000770, versus 0.000305 for random, 0.000145 for uniform, and 0.000285 for velocity. Its paired advantages over random and velocity are 0.000465 (CI $[0.000383, 0.000559]$) and 0.000485 (CI $[0.000396, 0.000587]$), with 5/5 seeds in the same direction. The base mean DINO drift falls from 0.038212 to 0.037442 (about 2.0%). Localization is stronger than the global effect: samples above the frozen $q_{0.8}$ threshold improve by 0.003618 versus 0.001817 for samples below the threshold. This is the discrete behavior expected from a where-to-compute law, not evidence that the continuous q^* is implemented exactly.

All four refinement strategies use an average of 8.425488 intervention calls per sample. Risk estimation is performed offline and requires about 23 additional calls per sample, roughly 15 more than the 8-call base sampler; these oracle calls are excluded from the matched intervention budget. The result therefore establishes allocation utility when risk is available, not wall-clock acceleration.

6.3 Expansion predicts and induces non-target editing drift

We use 1,000 first-turn MagicBrush edits (Zhang et al., 2023) with InstructPix2Pix (Brooks et al., 2023). Raw one-sided expansion ρ_+ predicts non-target LPIPS drift (Spearman 0.506). Edit-versus-empty-instruction mismatch also predicts drift (0.463), but is not stronger and adds little after multiplicity-adjusted controls (Appendix D.4.2).

We therefore use mismatch only as counterfactual localization rather than as a primary risk predictor. The archived field-change score is strongly negatively associated with drift (Spearman -0.851), while ρ_+ and this score are negatively associated (Spearman -0.408; Table 23). The post-hoc decomposition in Appendix D.4.4 shows that this archived score should not be interpreted as pure Frenet bending. We test the expansion channel more directly through a directional intervention with an equal-magnitude orthogonal placebo.

The mechanistic test perturbs the geometry directly. At each selected interval we estimate the leading expansion mode of S_t and use an equal-magnitude orthogonal direction as placebo (Fig. 4). The smallest validation-passing dose is frozen at $\alpha = 1.6$. On an untouched $n = 300$ confirmation set, leading-mode amplification relative to placebo increases non-target DINO by 0.017875, LPIPS by 0.013517, and RMSE by 0.007293, with 5/5 seeds in the same direction. Theorem 4 motivates the

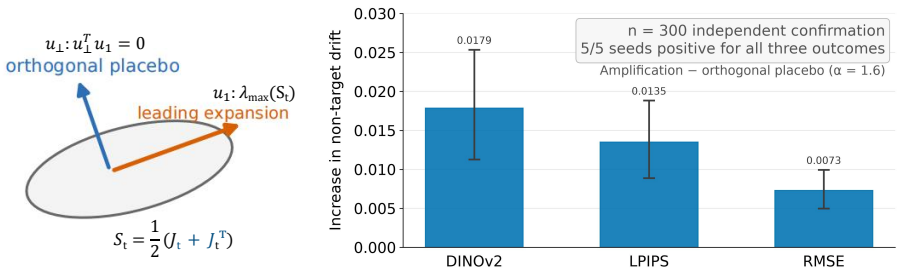


Figure 4: **Directional intervention on the deformation channel. Left:** The symmetric Jacobian $S_t = \frac{1}{2}(J_t + J_t^\top)$ defines the locally most-expanding direction through its leading eigenvector u_1. The intervention compares equal-magnitude perturbations along u_1 with an orthogonal placebo direction. **Right:** Alignment with u_1 increases mismatch-selected drift on the ndependent confirmation set across DINOv2, LPIPS, and RMSE.

selected direction, while the experiment tests the nonlinear claim: under the frozen editor and sampler, amplifying that direction produces more downstream drift than a matched orthogonal placebo.

6.4 Evidence hierarchy and negative boundaries

Tables 2 and 3 separate the two interventional stages. Each block uses a frozen primary contrast; secondary metrics and multiplicity-adjusted analyses are supporting checks rather than replacements. Negative branches bound the claims (Appendix D.5). Global acceleration-risk scheduling improves NFE-200 tracking but worsens 50k FID; the evaluated teacher–student compression setup misses its endpoint-quality gate, and a simple expansion-mode damping intervention misses its correction gate. Thus lower second-order risk is not a global quality objective.

Table 3: Leading-mode amplification relative to an equal-magnitude orthogonal placebo.

Outcome	Effect	95% CI	Seeds
DINO drift	0.017875	[0.011276, 0.025346]	5/5
LPIPS drift	0.013517	[0.008882, 0.018836]	5/5
RMSE drift	0.007293	[0.004965, 0.009937]	5/5

6.5 What the experiments jointly identify and imply

Constant-speed flows isolate bending; RF++ measures implemented acceleration and thus supports an incremental acceleration signal, not pure Frenet bending. Local refinement shows that a modest rank signal can localize fixed compute. In editing, the leading-mode placebo test isolates the leading expansion direction, while failed damping limits claims of automatic repair. Together, the evidence supports local, mechanism-specific second-order information rather than a global quality score.

These distinctions constrain how the geometry should be used. Reparameterization changes s and $\dot{s}$ but not κ_Γ, so scheduling is not path straightening. Refinement targets integration error, whereas the spectral intervention targets perturbation growth. Unlike FlowGuard rejection, CAT-Flow, or adaptive solvers (Li & Boström, 2025; Li et al., 2026; Dormand & Prince, 1980; Xiao, 2026), our analysis combines first-order non-identifiability, sample-specific allocation, and a separate perturbation mechanism. Practical deployment requires cheaper risk estimates and total-cost comparisons.

7 Conclusion

This work shows that first-order trajectory statistics do not fully characterize generative path fragility. We distinguish two second-order channels: material acceleration is linked to coarse-integration sensitivity, with bending isolated at constant speed, while expansion governs perturbation growth. Controlled flows, learned trajectories, compute allocation, and directional interventions show these quantities provide local information about where and why failures occur. They are not universal quality measures; second-order geometry offers mechanism-specific guidance for intervention, while improved local stability need not imply better global quality or correction.

REPRODUCIBILITY STATEMENT

The main text states the assumptions and scope of all theoretical results and specifies the frozen RF++ risk statistic, confirmation-set sizes, paired evaluation metrics, intervention budgets, and uncertainty estimates. Appendix A provides a structured comparison with the closest lines of work. Appendices B and C contain the complete theoretical derivations. Appendix D documents model and data preprocessing, estimator audits, the DINOv2 evaluation protocol, development–confirmation splits, the spectral intervention, protocol amendments, negative branches, and oracle-cost accounting. Appendix E states the remaining scope limitations. The supplementary package contains the code, configurations, random seeds, seed-level outputs, and frozen validation and selection rules used for the reported analyses.

AI USE STATEMENT

Generative AI tools were used strictly for language editing and LaTeX formatting to improve the readability of this manuscript. All theoretical formulations, mathematical proofs, and experimental evaluations were conducted independently by the authors. The authors have reviewed and take full responsibility for the final content of this paper.

REFERENCES

Michael Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2 edition, 2008. ISBN 978-3-7643-8722-8. doi: 10.1007/978-3-7643-8722-8.

Lazar Atanackovic, Xi (Nicole) Zhang, Brandon Amos, Mathieu Blanchette, Leo J Lee, Yoshua Bengio, Alexander Tong, and Kirill Neklyudov. Meta flow matching: Integrating vector fields on the wasserstein manifold. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 94586–94610, 2025.

Vansh Bansal, Saptarshi Roy, Alessandro Rinaldo, and Purnamrita Sarkar. On the convergence and straightness of rectified flow. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4636–4644. PMLR, 02–05 May 2026.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, Jan 2000. ISSN 0945-3245. doi: 10.1007/s002110050002.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, June 2023.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 47922–47945, 2024.

Siyi Chen, Huijie Zhang, Minzhe Guo, Yifu Lu, Peng Wang, and Qing Qu. Exploring low-dimensional subspace in diffusion models for controllable image editing. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 27340–27371. Curran Associates, Inc., 2024. doi: 10.52202/079017-0859.

J.R. Dormand and P.J. Prince. A family of embedded runge-kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, Mar 1980. ISSN 0377-0427. doi: 10.1016/0771-050x(80)90013-3.

Nicola Gigli. Second order analysis on $(\mathcal{P}_2(M), W_2)$. *Memoirs of the American Mathematical Society*, 216(1018), 2012. ISSN 0065-9266. doi: 10.1090/s0065-9266-2011-00619-2.

Doron Haviv, Aram-Alexandre Pooladian, Dana Pe'Er, and Brandon Amos. Wasserstein flow matching: Generative modeling over families of distributions. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22238–22258. PMLR, 13–19 Jul 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. volume 33, pp. 6840–6851, 2020.

Rafał Karczewski, Markus Heinonen, Alison Pouplin, Søren Hauberg, and Vikas Garg. The space-time of diffusion models: an information geometry perspective. In *International Conference on Learning Representations*, volume 2026, pp. 106128–106152, 2026.

Sangyun Lee, Beomsu Kim, and Jong Chul Ye. Minimizing trajectory curvature of ODE-based generative models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18957–18973. PMLR, 23–29 Jul 2023.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. *Advances in neural information processing systems*, 37:63082–63109, 2024.

Qinchan Li, Pedro Cisneros-Velarde, Keru Fu, Samuel Antunes Miranda, Sharan Vaswani, and Hao Zhang. CAT-Flow: Curvature-adaptive sTeps for flow matching, 2026. URL https://arxiv.org/abs/2609.01746.

Ziyun Li and Henrik Boström. FlowGuard: Guarding flow matching via conformal sampling. In Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson, and Henrik Boström (eds.), *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266 of *Proceedings of Machine Learning Research*, pp. 775–777. PMLR, 10–12 Sep 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

John Lott. Some geometric calculations on wasserstein space. *Communications in Mathematical Physics*, 277(2):423–437, 2008. ISSN 1432-0916. doi: 10.1007/s00220-007-0367-3.

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. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Yan Luo, Drake Du, Hao Huang, Yi Fang, and Mengyu Wang. CurveFlow: Curvature-guided flow matching for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 9020–9029, June 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, Jan 2001. ISSN 1532-4133. doi: 10.1081/pde-100002243.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28100–28127. PMLR, 23–29 Jul 2023.

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

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

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 23–29 Jul 2023.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 2009. ISBN 978-3-540-71050-9. doi: 10.1007/978-3-540-71050-9.

Hao Xiao. From euler to dormand-prince: ODE solvers for flow matching generative models, 2026. URL https://arxiv.org/abs/2605.00836.

Benjamin J. Zhang, Siting Liu, Wuchen Li, Markos A. Katsoulakis, and Stanley Osher. Wasserstein proximal operators describe score-based generative models and resolve memorization. *SIAM Journal on Mathematics of Data Science*, 8(3):623–648, July 2026. ISSN 2577-0187. doi: 10.1137/24m1644584.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in neural information processing systems*, 36:31428–31449, 2023.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36:49842–49869, 2023.

A Extended Comparison with Closest Work

A structured comparison with the closest lines of work is provided below. The table complements Section 2 by separating the object being measured, the intended use of that object, and the distinction made in this paper.

Research line	Main object	Main target	Distinction here
Trajectory geometry / straightness	path shape and curvature	easier few-step paths	fixed-path blindness; speed–bending split
Failure / adaptive sampling	risk score or adaptive grid	reject failures or place steps	propagated local allocation; separate expansion
Fast / adaptive solvers	local numerical defect	online step or solver control	geometry-derived risk; no solver-replacement claim
Wasserstein / Jacobian geometry	covariant or Jacobian directions	intrinsic geometry or editing	canonical specialization; placebo-tested expansion

Table 4: Positioning by research line. Prior approaches use curvature, straightness, or numerical error for training, failure detection, or adaptive integration. Our focus is the information absent from first-order path summaries and the separation between acceleration-based numerical risk and expansion-driven perturbational risk, with bending isolated under constant speed.

B Complete Proofs

B.1 Metric compatibility and the Frenet decomposition

Proof of Eq. (4)**.** Let $h_t(x) = f_t(x) \cdot g_t(x)$. By the continuity equation,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int h_t \, \mathrm{d}\mu_t = \int \partial_t h_t \, \mathrm{d}\mu_t + \int \nabla h_t \cdot v_t \, \mathrm{d}\mu_t.$$

Since $\partial_t(f \cdot g) + \nabla(f \cdot g) \cdot v = (\partial_t f + Df v) \cdot g + f \cdot (\partial_t g + Dg v)$, Eq. (4) follows.

Proof of Theorem 1. Write $v_t = s(t)T_t$. Applying D_t^Γ gives

$$a_t = \dot{s}(t)T_t + s(t)D_t^\Gamma T_t.$$

Metric compatibility and $\|T_t\|_t^2 = 1$ imply $\langle T_t, D_t^\Gamma T_t \rangle_t = 0$. Hence

$$\|a_t\|_t^2 = |\dot{s}(t)|^2 + s(t)^2 \left\|D_t^\Gamma T_t\right\|_t^2 = |\dot{s}(t)|^2 + s(t)^4 \kappa_\Gamma(t)^2.$$

Under an increasing reparameterization $t = t(\tau)$, $\widetilde{v}_\tau = t'(\tau)v_t$, $\widetilde{s} = t's$, and $D_\tau^{\widetilde{\Gamma}} T = t' D_t^\Gamma T$. Therefore $\widetilde{\kappa} = \left\|t' D_t^\Gamma T\right\| / (t's) = \kappa$, whereas the speed derivative changes.

Canonical specialization. For a fixed measure curve, $|\dot{\mu}_t|_{W_2} \le \|v_t\|_t$ for every admissible velocity, with equality for the minimum-$L^2(\mu_t)$ tangent. In the smooth Otto setting this tangent is a gradient. If $v_t = \nabla\phi_t$, then

$$\partial_t v_t + Dv_t v_t = \nabla\left(\partial_t\phi_t + \tfrac{1}{2}\|\nabla\phi_t\|^2\right),$$

so the material acceleration is tangent and agrees with the Wasserstein covariant acceleration.

B.2 First-order blindness

Proof of Theorem 2. It is enough to construct planar mean curves. By translation invariance, $W_2(\mathcal{N}(m,\sigma^2 I),\mathcal{N}(n,\sigma^2 I)) = \|m-n\|$, and the constant vector field $v_t(x) = \dot{m}(t)$ is the canonical tangent.

Let $D = \|m_1 - m_0\|$ and choose $L > D$. Start from the straight segment and attach one full circle tangent to it. Set $R = (L-D)/(2\pi)$, so the total length is L, and reparameterize by constant speed L on $[0,1]$; call this curve $m_0(t)$. For $0 < r < R$, replace the single loop by two tangent full loops of radii r and $R-r$. The length remains $D + 2\pi r + 2\pi(R-r) = L$, so the reparameterized curve m_r has the same speed, length, and action.

For a constant-speed planar curve, $\|\ddot{m}(t)\| = L^2|\kappa(\ell)|$ and $\mathrm{d}t = \mathrm{d}\ell/L$, hence

$$\int_0^1 \|\ddot{m}(t)\|^2\,\mathrm{d}t = L^3\int \kappa(\ell)^2\,\mathrm{d}\ell.$$

A full circle of radius q contributes $2\pi/q$ to $\int\kappa^2\,\mathrm{d}\ell$. Therefore

$$B(r) = 2\pi L^3\left(\frac{1}{r} + \frac{1}{R-r}\right), \qquad B_0 = \frac{2\pi L^3}{R}.$$

As $r \downarrow 0$, $B(r) - B_0 \to \infty$. Choose r so the difference exceeds M. Standard mollification of arbitrarily short junction neighborhoods gives smooth approximants with the same conclusion if desired.

Proof of Corollary 1. Suppose a finite-valued certificate F depending only on the two endpoints, path length, and kinetic action uniformly upper-bounded bending energy on this class. Theorem 2 constructs, for every M, paths with exactly the same arguments of F but with a bending-energy gap larger than M. Since F is unchanged along the construction, choosing M larger than the finite bound gives a contradiction.

C Discretization and Allocation Proofs

Lemma 1 (Propagated Euler remainder)**.** *Let X_t solve $\dot{X}_t = v_t(X_t)$ and assume $\|D_x v_t\|_{\mathrm{op}} \le L(t)$. Let $\widehat{X}_{k+1} = \widehat{X}_k + h_k v_{t_k}(\widehat{X}_k)$ with the same X_0, and define*

$$G_{N,k+1} := \prod_{j=k+1}^{N-1}\big(1 + h_j L(t_j)\big), \qquad G_{N,N} := 1.$$

Then

$$\left(\mathbb{E}\left\|X_1 - \widehat{X}_N\right\|^2\right)^{1/2} \le \sum_{k=0}^{N-1} G_{N,k+1}\int_{t_k}^{t_{k+1}} (t_{k+1}-r)\,\|a_r\|_r\,\mathrm{d}r.$$

If, in addition, $L \in C^1([0,1])$ and the grid is generated by a density $p \in C^1([0,1])$ with $p \ge p_{\min} > 0$ as in Proposition 2, then, uniformly in k,

$$G_{N,k+1} = G(t_{k+1})\big(1 + O(N^{-1})\big), \qquad G(t) := \exp\left(\int_t^1 L(s)\,\mathrm{d}s\right).$$

If also $A(t) := \|a_t\|_t \in C^1([0,1])$, then the propagated bound has leading term

$$\left(\mathbb{E}\left\|X_1 - \widehat{X}_N\right\|^2\right)^{1/2} \le \frac{1}{2}\sum_{k=0}^{N-1} G(t_k)A(t_k)h_k^2 + O(N^{-2}).$$

Proof. Along the exact path, $\mathrm{d}[v_t(X_t)]/\,\mathrm{d}t = a_t(X_t)$, so integrating twice gives Eq. (9). If $e_k = (\mathbb{E}\left\|X_{t_k} - \widehat{X}_k\right\|^2)^{1/2}$, Minkowski's inequality and Lipschitzness give

$$e_{k+1} \le \big(1 + h_k L(t_k)\big) e_k + r_k, \qquad r_k \le \int_{t_k}^{t_{k+1}} (t_{k+1} - r)\, \|a_r\|_r \,\mathrm{d}r.$$

Iterating this recursion gives the stated discrete bound with the product $G_{N,k+1}$; no comparison between a finite Riemann sum and a continuous integral is needed. For the equal-p grids of Proposition 2, $\max_j h_j = O(N^{-1})$. Hence, uniformly in k,

$$\begin{aligned}\log G_{N,k+1} &= \sum_{j=k+1}^{N-1} \log\big(1 + h_j L(t_j)\big) \\ &= \sum_{j=k+1}^{N-1} h_j L(t_j) + O\left(\sum_{j=k+1}^{N-1} h_j^2\right) \\ &= \int_{t_{k+1}}^{1} L(s)\,\mathrm{d}s + O(N^{-1}),\end{aligned}$$

where the last step is the Riemann-sum error for $L \in C^1$. Exponentiating yields $G_{N,k+1} = G(t_{k+1})(1 + O(N^{-1}))$. If $A \in C^1$, then

$$\int_{t_k}^{t_{k+1}} (t_{k+1} - r) A(r)\,\mathrm{d}r = \frac{1}{2} A(t_k) h_k^2 + O(h_k^3)$$

uniformly in k. Replacing $G(t_{k+1})$ by $G(t_k)$ changes each term by $O(h_k^3)$, so summing over k gives the $O(N^{-2})$ remainder. The same coupling also bounds the endpoint W_2 distance.

Proof of Proposition 2. Let $F(t) = \int_0^t p(r)\,\mathrm{d}r$ and $T = F^{-1}$. Since $p \in C^1$ and $p \ge p_{\min} > 0$, $T \in C^2$. The equal-mass grid is $t_k = T(k/N)$. Taylor expansion gives

$$h_k = T((k+1)/N) - T(k/N) = \frac{1}{N p(t_k)} + O(N^{-2}),$$

uniformly in k. Hence

$$\mathcal{B}_N(p) = \frac{1}{2N^2} \sum_{k=0}^{N-1} \frac{w(t_k)}{p(t_k)^2} + O(N^{-2}).$$

In the $y = F(t)$ coordinate, the Riemann sum satisfies

$$\frac{1}{N} \sum_{k=0}^{N-1} \frac{w(T(k/N))}{p(T(k/N))^2} = \int_0^1 \frac{w(T(y))}{p(T(y))^2}\,\mathrm{d}y + O(N^{-1}) = \int_0^1 \frac{w(t)}{p(t)}\,\mathrm{d}t + O(N^{-1}),$$

which proves Eq. (12).

Proof of Theorem 3. For any admissible q, interpret the integrand as zero on $\{w = q = 0\}$. Since $q > 0$ almost everywhere on $\{w > 0\}$,

$$\left(\int_0^1 \sqrt{w(t)}\,\mathrm{d}t\right)^2 = \left(\int_0^1 \sqrt{\frac{w(t)}{q(t)}}\sqrt{q(t)}\,\mathrm{d}t\right)^2 \le \left(\int_0^1 \frac{w(t)}{q(t)}\,\mathrm{d}t\right)\left(\int_0^1 q(t)\,\mathrm{d}t\right).$$

Thus $\mathcal{E}[q] \ge (2N)^{-1}(\int \sqrt{w})^2$. Equality in Cauchy–Schwarz requires $q \propto \sqrt{w}$, giving Eq. (14). For the plug-in bound, write $S = \int \sqrt{w}$ and $\widehat{S} = \int \sqrt{\widehat{w}}$. Then

$$\mathcal{E}[q_{\widehat{w}}] = \frac{\widehat{S}}{2N} \int_0^1 \frac{w(t)}{\sqrt{\widehat{w}(t)}}\,\mathrm{d}t.$$

The multiplicative assumption gives $\widehat{S} \le \sqrt{1+\delta}\,S$ and $w/\sqrt{\widehat{w}} \le \sqrt{w}/\sqrt{1-\delta}$, which yields Eq. (15).

Proof of Corollary 2. For active intervals, piecewise-constant $q_i = n_i/\Delta_i$ turns Eq. (10) into $\frac{1}{2}\sum_i w_i\Delta_i^2/n_i$; intervals with $w_i = n_i = 0$ contribute zero by convention. Applying the same Cauchy–Schwarz argument to the active finite sum gives $n_i \propto \Delta_i\sqrt{w_i}$.

C.1 Spectral perturbation growth

Proof of Theorem 4. The variational equation is $\dot{\xi} = J_t\xi$. Therefore

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2 = \xi^\top J_t\xi = \xi^\top S_t\xi,$$

because the skew-symmetric part contributes zero. Rayleigh–Ritz gives the maximum over unit vectors as $\lambda_1(S_t)$. If e_1 is a leading eigenvector and $\eta \perp e_1$ is unit, then $\eta^\top S_{t_0}\eta \le \lambda_2(S_{t_0})$, proving Eq. (18). If J_t is C^1, the second derivatives of the two squared norms are bounded near t_0, so Taylor's theorem yields the $2\gamma h - Ch^2$ bound. Finally, the Euclidean logarithmic norm of J_t is $\lambda_1(S_t)$, which gives the fundamental-matrix bound.

C.2 Finite-difference consistency

Proof of Proposition 3. Along the exact trajectory, $g'(t) = \frac{\mathrm{d}}{\mathrm{d}t}v_t(X_t) = a_t(X_t)$. Taylor's theorem with integral remainder gives

$$\widehat{a}_k - g'(t_k) = \frac{1}{h_k}\int_{t_k}^{t_{k+1}} (t_{k+1} - r)g''(r)\,\mathrm{d}r,$$

so $\|\widehat{a}_k - a_{t_k}(X_{t_k})\| \le Mh_k/2$. Writing $e_k = \widehat{a}_k - a_{t_k}(X_{t_k})$ and using $\|a_{t_k}(X_{t_k})\| \le A$,

$$\left|\, \|\widehat{a}_k\|^2 - \|a_{t_k}(X_{t_k})\|^2 \,\right| \le 2A\,\|e_k\| + \|e_k\|^2 \le AMh_k + \frac{M^2h_k^2}{4}.$$

For sampled scalar risks, the mean of the largest fixed fraction of coordinates is 1-Lipschitz under the ℓ_∞ norm: perturbing every coordinate by at most ε changes the top-fraction mean by at most ε. The final statement follows on a uniform mesh.

D Experimental Details and Additional Results

We report protocols, additional results, and limitations for the four main experiments, separating original analyses, confirmations, and post-hoc reanalyses. Numerical-reference tracking, real-data quality, and intrinsic geometry are distinct outcomes. Unless specified otherwise, intervals are 95%.

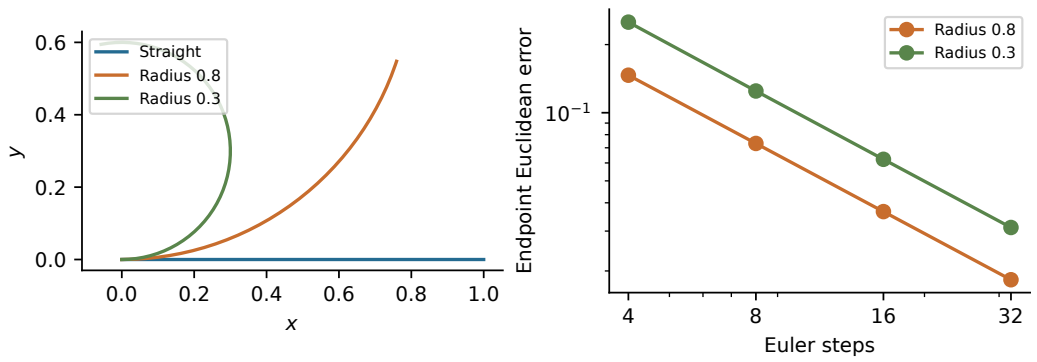
D.1 Controlled flows

D.1.1 Constant-speed control

We translate the same reference distribution along a straight path and two circular arcs with radii 0.8 and 0.3. All paths have unit speed, length, and action, but different endpoints. The smaller-radius arc has greater normal-acceleration energy and larger Euler endpoint errors at every tested NFE (Table 5; Figure 5).

Table 5: Estimated normal energy B and Euclidean endpoint error for the unit-speed controls.

Path	B	$N = 4$	$N = 8$	$N = 16$	$N = 32$
Straight	0.000000	0.000000	0.000000	0.000000	0.000000
Arc, $R = 0.8$	1.562498	0.146473	0.073162	0.036574	0.018288
Arc, $R = 0.3$	11.111029	0.251298	0.124728	0.062257	0.031117



(a) Unit-speed translation paths

(b) Archived discretization error

Figure 5: Unit-speed paths and Euler endpoint errors. The straight path's zero error is omitted from the logarithmic panel.

D.1.2 Historical analytic suite and robustness sweeps

The archived canonical three-path analysis gives Pearson/Spearman $0.9522/0.9161$ between $\sqrt{K_{\mathrm{acc}}}/\mathrm{NFE}$ and paired RMSE. The broader continuous sweep remains positive (Spearman 0.7810) but shows low-NFE aliasing and a high-NFE numerical floor. A separate state-dependent Eulerian affine branch retains Spearman 0.9869 for paired RMSE. These are descriptive robustness checks and do not replace the constant-speed control above.

D.1.3 Reparameterization check

We traverse the same curve with uniform, front-loaded, and back-loaded schedules. Curvature agrees after arc-length alignment, with maximum discrepancy 2.445×10^{-6}, while speed and acceleration-based temporal summaries change (Table 6; Figure 6). This distinguishes geometric curvature from clock-dependent operational risk; no image-quality outcomes are evaluated.

Table 6: Time-grid means for three schedules on the same curve with fixed endpoints; these are not invariant geometric integrals.

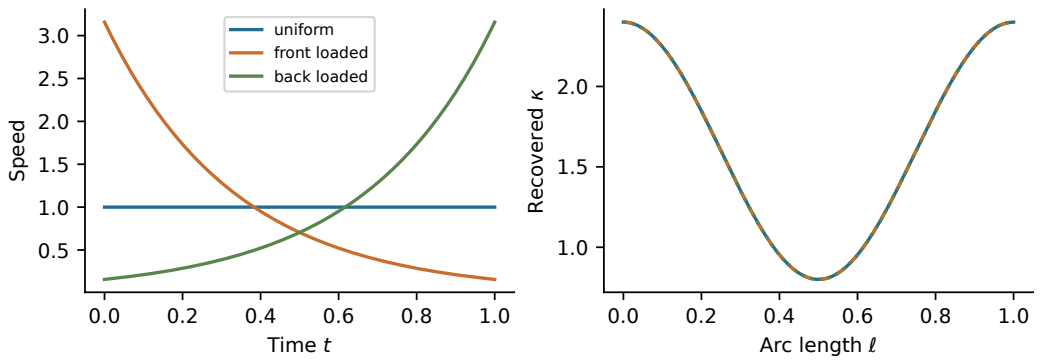
Schedule	$\langle \|a\|^2 \rangle_t$	$\langle \dot{s}^2 \rangle_t$	$\langle B \rangle_t$
uniform	2.880720	0.000000	2.880720
front loaded	44.816232	14.922198	29.894034
back loaded	44.816232	14.922198	29.894034

D.2 RF++ DINOv2-primary evaluation

D.2.1 Model, pairing, and outcomes

We evaluate the fixed CIFAR-10 RF++ configuration-G checkpoint using FP32 descending Euler sampling. NFE 4, 8, and 16 outputs share initial noise with 200-step references. Each seed contains 10,000 samples: seed 0 defines the 80th-percentile high-drift threshold and selects the reanalysis's first-order comparator; seeds 1–4 are held out. Risks retain $t \in [0.02, 0.98]$.

Primary drift is one minus DINOv2 ViT-S/14 CLS cosine similarity. Pixel RMSE and VGG LPIPS provide separate discrepancy measures. The numerical reference is not image-quality ground truth; real-data FID and KID assess distributional quality.



(a) Different schedules (b) Matched geometric locations

Figure 6: Speed profiles and arc-length-aligned curvature under three time schedules.

D.2.2 Operational scores and interpretation

R_{acc} averages the largest 10% of retained squared-acceleration values, summed over pixel coordinates without dimension normalization. R_{bend} subtracts the top-10% mean of the squared time derivative of cohort RMS speed, then clips the result at zero. $R_{\mathrm{bend,loc}}$ instead subtracts this time-aligned speed-change term and clips at each time before top-10% aggregation. Cohorts comprise all samples within each seed/NFE setting.

These population speed corrections are not samplewise normal projections. Their predictive performance therefore does not isolate an intrinsic curvature contribution.

D.2.3 Implemented Euler probe and consistency

For $v_0 = v(t, x)$ and $F(r) = v(t + r, x + rv_0)$, the actual estimator uses

$$\widehat{a}^{c}_{\epsilon} = \frac{F(\epsilon) - F(-\epsilon)}{2\epsilon}, \qquad \widehat{a}^{+}_{\delta} = \frac{F(\delta) - F(0)}{\delta}, \qquad \widehat{a}^{-}_{\delta} = \frac{F(0) - F(-\delta)}{\delta}. \tag{22}$$

Here $\epsilon = 10^{-3}$ is distinct from the sampler step. The central branch requires admissible probes for every batch query. Otherwise each query uses a forward step $\delta = \epsilon$ when admissible, or a backward step $\delta = t - \max(t - \epsilon, t_{\min}) > 0$. These are frozen-direction Euler displacements, not exact trajectory states.

Consistency and calibration. Since $F'(0) = a(t, x)$, cancellation of even Taylor terms gives $\|\widehat{a}^{c}_{\epsilon} - a\| \le M_3\epsilon^2/6$ when $F \in C^3$ and $\sup \|F'''\| \le M_3$. The one-sided bound is $M_2\delta/2$ for $F \in C^2$ and $\sup \|F''\| \le M_2$. An acceleration error E with $\|a\| \le A_0$ gives squared-norm error at most $2A_0E + E^2$. The top-fraction mean, a maximum of fixed-size subset means, is 1-Lipschitz in the maximum norm; the same uniform error bounds risk on a fixed grid.

These are pointwise truncation bounds for exact field evaluations. An L_a-Lipschitz acceleration field adds state error $L_a\|x_h - X_t\|$. Finite-precision cancellation and direction errors can prevent improvement as ϵ shrinks; continuous-time tail convergence also needs grid refinement and profile regularity. An exact-trajectory difference proposition alone would not justify this implementation.

The FD/AD audit compares squared norms on 1,000 paths and 32 times per seed. Relative L_p error is $\|q_{\mathrm{FD}} - q_{\mathrm{AD}}\|_p / \|q_{\mathrm{AD}}\|_p$. Spearman 0.923–0.924 coexists with relative L_2 error 0.771–0.789: ranking agrees more than scale. No NaNs were recorded. Only $\epsilon = 10^{-3}$ is archived, so no multi-ϵ convergence claim is made. The editing spatial-Jacobian audit concerns another derivative.

Table 7: RF++ squared-acceleration audit: finite differences versus automatic differentiation at $\epsilon = 10^{-3}$, using 1,000 paths and 32 query times per seed. Rank agreement does not imply scale calibration.

Seed	Pearson	Spearman	Relative L_1	Relative L_2
0	0.931	0.924	0.761	0.774
1	0.669	0.924	0.765	0.789
2	0.948	0.924	0.763	0.781
3	0.861	0.922	0.755	0.771
4	0.941	0.923	0.757	0.776

D.2.4 Original primary analysis and integral-acceleration secondary analysis

The original DINO analysis reuses saved outcomes. Its primary score is $K_{\mathrm{top10}} = T_{0.1}(\|\widehat{a}\|^2)$; secondary scores are the retained maximum $K_{\max}$ and trapezoidal integral K_{int}, using positive elapsed time between retained nodes. These summarize acceleration, not intrinsic curvature. The original velocity baseline is the saved velocity-energy integral divided by 0.96; drift energy is its exact alias, and the time-only predictor is constant within NFE.

Tables 8–10 provide the original main-text comparison. Macro metrics average four test-seed values; their intervals use 10,000 resamples of those values. Pooled metrics use 40,000 rows per NFE, without asserted pooled intervals. The saved `test_pooled` row contains both point estimates, but its interval columns belong to the macro means. Test-positive proportions are 0.194175, 0.194075, and 0.194750 at NFE 4, 8, and 16. These seeds vary generation noise for one checkpoint, not model training.

Table 8: Original DINOv2 analysis: four-seed macro means with 95% seed-bootstrap intervals. K_{top10} is primary; maximum and integral are secondary. Undefined constant-predictor entries are retained.

NFE	Predictor	Macro Spearman [95% CI]	Macro AUROC [95% CI]
4	K_{top10}	0.104 $[0.098, 0.110]$	0.549 $[0.547, 0.552]$
4	$K_{\max}$	0.104 $[0.098, 0.110]$	0.549 $[0.547, 0.552]$
4	K_{int}	0.206 $[0.196, 0.215]$	0.604 $[0.599, 0.608]$
4	Velocity energy	0.041 $[0.036, 0.044]$	0.509 $[0.507, 0.514]$
4	Path length	0.041 $[0.036, 0.044]$	0.510 $[0.507, 0.514]$
4	Drift energy (alias)	0.041 $[0.036, 0.044]$	0.509 $[0.507, 0.514]$
4	Time-only (constant)	Undefined	Not reported
8	K_{top10}	0.125 $[0.120, 0.132]$	0.566 $[0.563, 0.568]$
8	$K_{\max}$	0.125 $[0.120, 0.132]$	0.566 $[0.563, 0.568]$
8	K_{int}	0.227 $[0.218, 0.235]$	0.624 $[0.622, 0.626]$
8	Velocity energy	0.078 $[0.072, 0.085]$	0.531 $[0.527, 0.533]$
8	Path length	0.078 $[0.071, 0.085]$	0.531 $[0.526, 0.533]$
8	Drift energy (alias)	0.078 $[0.072, 0.085]$	0.531 $[0.527, 0.533]$
8	Time-only (constant)	Undefined	Not reported
16	K_{top10}	0.227 $[0.221, 0.232]$	0.631 $[0.624, 0.636]$
16	$K_{\max}$	0.169 $[0.162, 0.175]$	0.599 $[0.591, 0.604]$
16	K_{int}	0.253 $[0.248, 0.258]$	0.647 $[0.640, 0.653]$
16	Velocity energy	0.114 $[0.107, 0.120]$	0.552 $[0.548, 0.556]$
16	Path length	0.113 $[0.107, 0.120]$	0.552 $[0.548, 0.556]$
16	Drift energy (alias)	0.114 $[0.107, 0.120]$	0.552 $[0.548, 0.556]$
16	Time-only (constant)	Undefined	Not reported

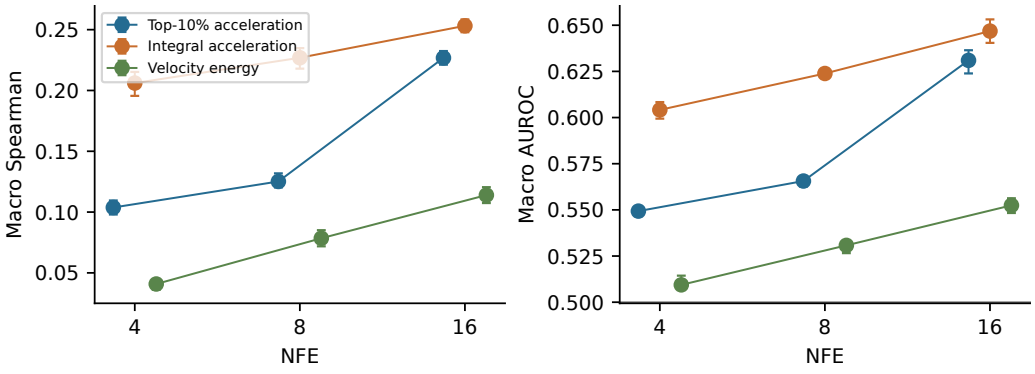
Integral-acceleration AUROC is $0.6041/0.6239/0.6468$, versus $0.5493/0.5656/0.6310$ for the primary score. This supports cumulative acceleration's predictive value in this cohort, not universal superiority. Original regressions control for velocity energy, path length, and residual motion of successive denoised-state estimates; the pooled fit adds NFE effects. Their intervals are classical OLS Student-t intervals, not HC3 or seed-cluster intervals.

Table 9: Original RF++ aggregation: pooled Spearman/AUROC over 40,000 test samples per NFE, and four-seed macro AUPRC with 95% seed-bootstrap intervals. No pooled intervals are asserted.

NFE	Score	Pooled Spearman	Pooled AUROC	Macro AUPRC [95% CI]
4	K_{top10}	0.10372	0.54930	0.258 [0.252, 0.264]
4	K_{int}	0.20605	0.60409	0.309 [0.302, 0.316]
8	K_{top10}	0.12518	0.56561	0.272 [0.268, 0.275]
8	K_{int}	0.22694	0.62387	0.321 [0.317, 0.324]
16	K_{top10}	0.22687	0.63094	0.332 [0.322, 0.340]
16	K_{int}	0.25324	0.64680	0.343 [0.333, 0.353]

Table 10: Original standardized K_{top10} coefficients with classical OLS 95% intervals. Controls include velocity energy, path length, and residual motion; the pooled fit adds NFE effects.

Fit	n	Std. β [classical 95% CI]
4	40000	0.0666 [0.0559, 0.0773]
8	40000	0.0724 [0.0614, 0.0835]
16	40000	0.0521 [0.0419, 0.0623]
Pooled + NFE effects	120000	0.0347 [0.0288, 0.0407]



(a) Original four-seed analysis

(b) Primary and secondary scores

Figure 7: Original RF++ primary and secondary scores, with four-seed macro 95% bootstrap intervals. Integral acceleration has higher point estimates at each NFE but remains secondary; the comparison does not isolate curvature.

D.2.5 LATER ROBUST-STATISTICS REANALYSIS

The later replay changes quantities, predictors, and inference; its intervals cannot be attached to the original K_{top10} table. At each NFE, all predictors share outcomes, validation-defined labels, and paired resamples. The 2,000 bootstrap replicates resample within seed-by-outcome strata for AUROC and within seed with fixed ranks for Spearman. They condition on four held-out generation seeds.

RMS velocity, path length, and action form the first-order block. Near-collinearity is handled by PCA of standardized controls, retaining singular values above 10^{-4} of the largest (rank two). Outcomes and added scores are standardized. HC3 is sample-level, not seed-cluster. Leave-one-test-seed-out (LOSO) prediction refits standardization, PCA, and regression on three seeds before testing the fourth. RMS velocity is the validation-selected AUROC comparator at every NFE.

Table 11: RF++ reanalysis on 40,000 test samples per NFE: 95% paired-bootstrap intervals, using seed/outcome strata for AUROC and within-seed fixed ranks for Spearman.

NFE	Predictor	Spearman [95% CI]	AUROC [95% CI]
4	RMS velocity	0.041 $[0.031, 0.051]$	0.509 $[0.502, 0.516]$
4	Path length	0.041 $[0.031, 0.051]$	0.509 $[0.502, 0.516]$
4	Action	0.041 $[0.031, 0.051]$	0.509 $[0.502, 0.516]$
4	R_{acc}	0.170 $[0.160, 0.179]$	0.596 $[0.589, 0.604]$
4	R_{bend}	0.170 $[0.160, 0.179]$	0.596 $[0.589, 0.604]$
4	$R_{\mathrm{bend,loc}}$	0.169 $[0.160, 0.179]$	0.596 $[0.589, 0.604]$
8	RMS velocity	0.078 $[0.069, 0.088]$	0.531 $[0.523, 0.538]$
8	Path length	0.078 $[0.069, 0.088]$	0.531 $[0.523, 0.538]$
8	Action	0.078 $[0.069, 0.088]$	0.531 $[0.523, 0.538]$
8	R_{acc}	0.174 $[0.165, 0.184]$	0.608 $[0.601, 0.615]$
8	R_{bend}	0.174 $[0.165, 0.184]$	0.608 $[0.601, 0.615]$
8	$R_{\mathrm{bend,loc}}$	0.174 $[0.165, 0.184]$	0.608 $[0.601, 0.615]$
16	RMS velocity	0.114 $[0.104, 0.124]$	0.552 $[0.545, 0.560]$
16	Path length	0.114 $[0.104, 0.124]$	0.552 $[0.545, 0.560]$
16	Action	0.114 $[0.104, 0.124]$	0.552 $[0.545, 0.560]$
16	R_{acc}	0.271 $[0.262, 0.280]$	0.659 $[0.652, 0.666]$
16	R_{bend}	0.271 $[0.262, 0.280]$	0.659 $[0.652, 0.666]$
16	$R_{\mathrm{bend,loc}}$	0.271 $[0.262, 0.280]$	0.659 $[0.652, 0.666]$

Table 12: Incremental $R_{\mathrm{bend,loc}}$ results: paired-bootstrap AUROC gain over validation-selected RMS velocity, sample-level HC3 coefficient intervals, and LOSO out-of-fold ΔR^2.

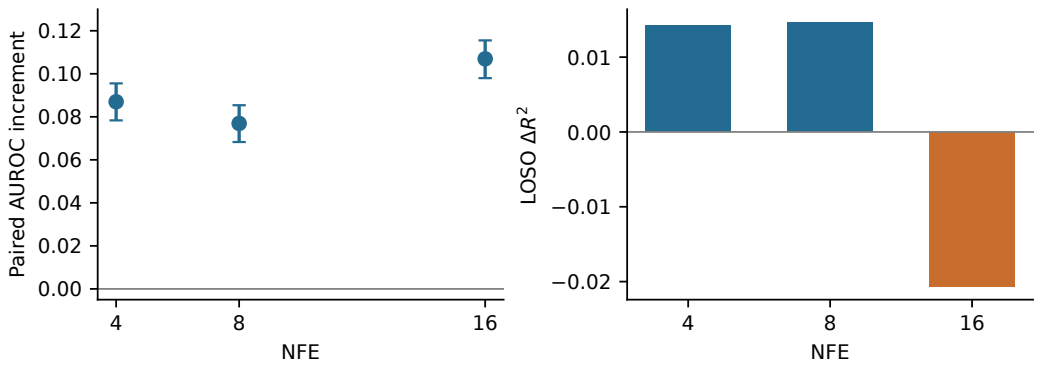
NFE	ΔAUROC [95% CI]	Std. β [HC3 95% CI]	LOSO ΔR^2
4	0.087 $[0.078, 0.096]$	0.121 $[0.071, 0.172]$	0.0142
8	0.077 $[0.068, 0.085]$	0.136 $[0.069, 0.204]$	0.0146
16	0.107 $[0.098, 0.116]$	0.077 $[-0.256, 0.410]$	-0.0207

Discrimination is moderate; conditional effects are not uniformly stable. At NFE 16, the HC3 interval includes zero and LOSO ΔR^2 is negative. Nearly identical AUROCs for R_{acc}, R_{bend}, and $R_{\mathrm{bend,loc}}$ prevent attributing the gain specifically to normal curvature.

D.2.6 DENSE REFERENCE TRAJECTORIES AND ENDPOINT ABLATIONS

Low-NFE population-moment analysis gives pooled DINO Spearman 0.943 for the curvature integral, but NFE-centered association is 0.375 with seed-cluster interval $[-0.945, 0.927]$. Pooling NFE therefore does not establish an independent geometric effect.

The dense analysis uses 64 trajectories per seed, five seeds, a 200-step reference interpolated to 1,000 time points, and independently computed 50/100-step trajectories for mesh sensitivity. Geometry is estimated once per seed from population moments and reused across outcome NFE settings. Thus each within-NFE correlation has five geometry/outcome pairs; evaluating 320 individual outcomes does not create 320 independent geometry scores. The integral is fixed-clock $\int \widehat{\kappa}_\Gamma^2 dt$, not arc-length



(a) High-drift discrimination

(b) Cross-seed prediction

Figure 8: RF++ reanalysis: (a) paired AUROC gains with 95% bootstrap intervals; (b) LOSO ΔR^2 point estimates. AUROC gains are positive, but the NFE-16 LOSO increment is negative.

invariant. Nested fits have 960 outcome rows, operational-risk controls, NFE effects, and five clusters; exact tests enumerate 120 seed assignments.

Table 13: Dense-reference geometry: five seed-level scores per NFE, evaluated against 320 outcomes. AUROC intervals resample seeds. Trimming retains $[0.02, 0.98]$; outcome rows do not add independent geometry observations.

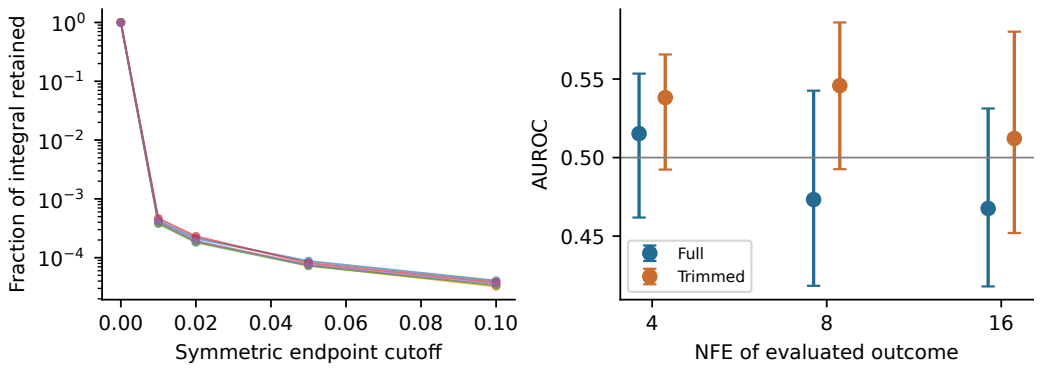
NFE	Interval	Seed-level Spearman	AUROC [95% CI]
4	Full	0.000	$0.515\ [0.462, 0.553]$
4	Trimmed	0.900	$0.538\ [0.492, 0.566]$
8	Full	-0.700	$0.473\ [0.418, 0.543]$
8	Trimmed	0.600	$0.546\ [0.493, 0.586]$
16	Full	-0.700	$0.468\ [0.418, 0.531]$
16	Trimmed	-0.100	$0.512\ [0.452, 0.580]$

Table 14: Adding full/trimmed curvature integrals to operational-risk controls and NFE effects: 960 outcome rows, five seed clusters. Bootstrap intervals are descriptive; permutation tests enumerate all 120 seed assignments.

Added geometry	Std. β [95% CI]	Exact permutation p
Full integral	$-0.023\ [-0.084, 0.028]$	0.433
Trimmed integral	$0.026\ [0.007, 0.087]$	0.367

The first grid cell contains about 99.67% of the integral; trimming to $[0.02, 0.98]$ removes about 99.98%. Full-grid relative L_2 curvature discrepancies against the 200-step reference are 0.281% and 0.442% at 50/100 steps, without monotone convergence. Trimmed discrepancies decrease from 6.345% to 3.950%, improving in every seed; the reference remains numerical, not exact.

Trimmed NFE-4 Spearman is 0.9 (nominal $p = 0.0374$), based on five pairs and several unadjusted comparisons; it does not persist at NFE 8/16. The trimmed regression's positive descriptive bootstrap interval coexists with exact permutation $p = 0.367$. Independent curvature prediction is therefore not robustly established, nor ruled out given endpoint sensitivity and few clusters. Historical/replayed outcomes are close but not bitwise identical; this analysis uses its own within-run pairings.



(a) Endpoint concentration

(b) Seed-level geometry scores

Figure 9: Dense-reference geometry: (a) retained fixed-clock curvature integral, one line per seed; (b) AUROC with seed-cluster 95% intervals, all including 0.5. Endpoint trimming removes most energy; geometry remains seed-level despite individual outcome evaluation.

D.3 Local refinement

D.3.1 Development and common-gate confirmation

Development uses 1,280 samples, policy-specific sample gates, and matched aggregate extra calls. Confirmation uses fresh IDs 30,000–34,095 for each seed 0–4 (20,480 samples) with a frozen validation-quantile gate. The primary setting refines two intervals of an 8-step trajectory (nominal fraction 0.20), evaluated against the paired 200-step DINO reference. Only seven intervals remain after trimming to $[0.02, 0.98]$; two therefore represent $2/7$ of eligible intervals.

All confirmation policies treat the same 4,357 acceleration-high samples. Only interval selection changes: acceleration risk, random without replacement, uniform bin-midpoint indices, or velocity magnitude. Risk/velocity rankings use stable descending order. Each selected interval becomes two half-steps, adding one call; untreated endpoints remain unchanged. The stages differ in both sample size and sample gating and are not pooled.

Table 15: Common-gate refinement confirmation ($n = 20,480$; 4,357 treated). Baseline-minus-refined improvements and paired 95% intervals are scaled by 10^3. Each treated sample receives two extra rollout calls; probe costs are excluded.

Interval policy	DINO improvement	LPIPS improvement	RMSE improvement
Acceleration risk	0.770 [0.647, 0.906]	0.275 [0.237, 0.318]	0.336 [0.317, 0.356]
Random	0.305 [0.218, 0.412]	0.110 [0.084, 0.143]	0.136 [0.124, 0.149]
Uniform	0.145 [0.119, 0.171]	0.054 [0.046, 0.062]	0.085 [0.080, 0.090]
Velocity	0.285 [0.201, 0.394]	0.096 [0.071, 0.128]	0.131 [0.118, 0.145]

Table 16: Acceleration-guided advantage at NFE 8, nominal fraction 0.20 ($\times 10^3$; positive favors acceleration). Development uses policy-specific gates; confirmation uses a common gate. The stages are not pooled.

Stage	Comparator	n	Paired advantage $\times 10^3$
Development	Random	1280	0.120 [−0.269, 0.447]
Development	Uniform	1280	0.420 [0.167, 0.682]
Development	Velocity	1280	0.405 [0.160, 0.647]
Confirmation	Random	20480	0.465 [0.383, 0.559]
Confirmation	Uniform	20480	0.625 [0.511, 0.758]
Confirmation	Velocity	20480	0.485 [0.396, 0.587]

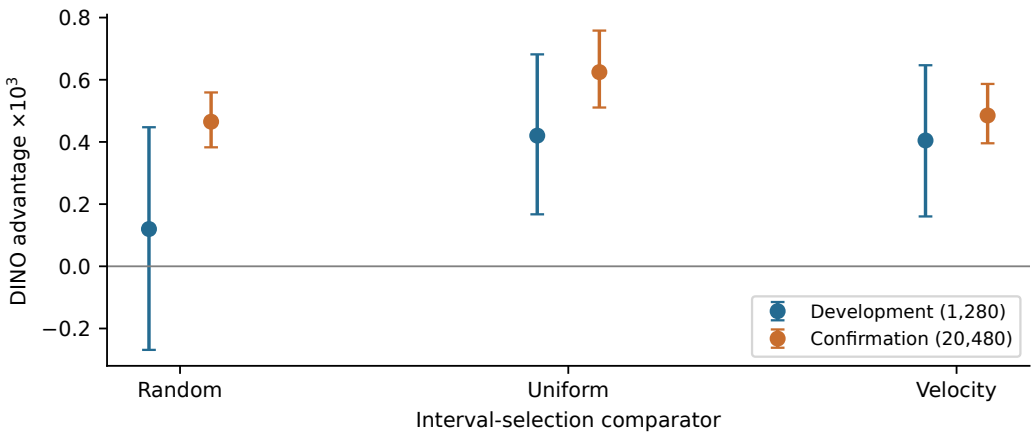


Figure 10: Refinement advantages with paired 95% intervals. Development's random-control contrast includes zero; confirmation contrasts are positive. Policy-specific development gates differ from the common confirmation gate. Risk-estimation calls are excluded.

Acceleration-guided mean DINO drift decreases from 0.038212 to 0.037442 (2.01%). Mean rollout calls are $8 + 2(4357/20480) = 8.425488$. Improvements mean baseline minus refined discrepancy; positive paired advantages favor acceleration. Tables 15 and 16 scale effects and intervals by 10^3. Intervals use 10,000 paired sample resamples conditional on the evaluated seeds, not a population of trained checkpoints.

D.3.2 Risk-stratified effects

Ungated refinement yields potential DINO improvements of 0.003618 for 4,357 high-risk samples and 0.001817 for 16,123 low-risk samples. The mean difference is 0.001801 (95% interval [0.001234, 0.002442]). These are potential outcomes under treatment in both strata; the primary policy leaves low-risk samples unchanged. Archived policy-specific-gating sensitivity is separate from the common-gate comparison.

D.3.3 Probe cost and total-budget comparison

The 8-step scout plus 15 incremental material probes costs 23 calls total. The 8.425488 rollout budget excludes risk discovery and supports offline localization, not end-to-end speedup. Local-defect diagnostics do not provide a compute-matched comparison with a step-doubling adaptive solver.

A separate 1,280-sample, five-seed experiment budgets 32/48/64 total calls. Uniform sampling spends all calls on rollout. The fixed-state difference proxy uses 8 scout plus 8 probe calls and approximates $\partial_t v$; the material-acceleration oracle uses 8 plus 15. Adaptive methods restart from paired initial noise using the remaining budget. The 200-step reference is excluded equally from all budgets. Intervals use 2,000 hierarchical resamples (seed, then samples); only five clusters are available, and FID was not measured.

Both adaptive policies worsen DINO, LPIPS, and RMSE in all five seed strata; this does not imply failure of all adaptive solvers. Checkpoint/configuration hashes are available, but the current material-probe source differs from the frozen budget-experiment manifest. We report archived measurements without claiming an exact source-replay match.

D.4 Editing experiments

D.4.1 Data, masks, metrics, and prediction

We use 1,000 first-turn MagicBrush edits (Zhang et al., 2023) from public training-source records, split by source image into disjoint 600/200/200 train/validation/test sets (seed 20260809). The orig-

Table 17: Total-budget comparison on 1,280 paired samples and five seeds. Prepass includes scout/probe calls; lower DINO drift is better. Intervals are hierarchical-bootstrap 95% intervals. FID was not measured.

Budget	Method	Prepass	Final	DINO drift [95% CI]
32	Uniform	0	32	0.0115 [0.0081, 0.0155]
32	Fixed-state FD	16	16	0.0372 [0.0314, 0.0430]
32	Material-acc. oracle	23	9	0.0377 [0.0317, 0.0439]
48	Uniform	0	48	0.0081 [0.0053, 0.0116]
48	Fixed-state FD	16	32	0.0373 [0.0315, 0.0437]
48	Material-acc. oracle	23	25	0.0351 [0.0297, 0.0412]
64	Uniform	0	64	0.0057 [0.0034, 0.0085]
64	Fixed-state FD	16	48	0.0371 [0.0314, 0.0434]
64	Material-acc. oracle	23	41	0.0347 [0.0292, 0.0404]

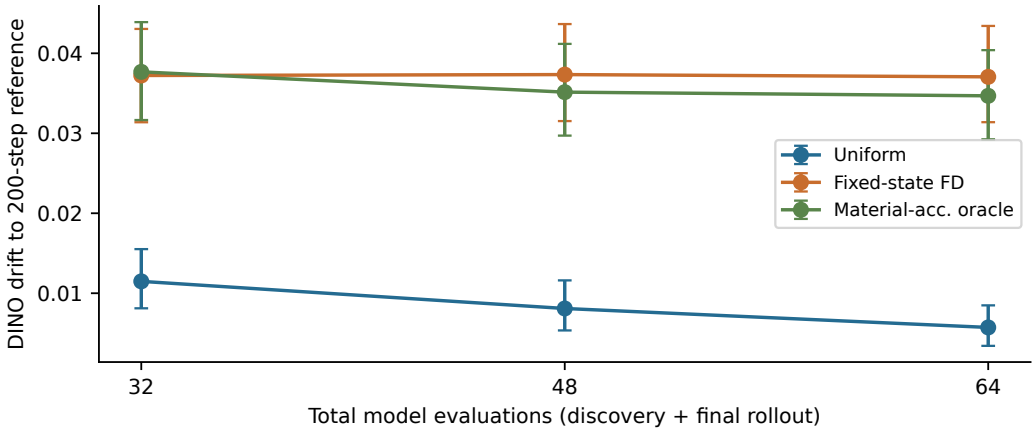


Figure 11: Total-budget comparison with hierarchical-bootstrap 95% intervals. Uniform sampling has lower reference drift than both tested adaptive policies at every budget, limiting cost-inclusive efficiency claims.

inal InstructPix2Pix checkpoint (Brooks et al., 2023), without MagicBrush fine-tuning, operates on 256×256 images and $4 \times 32 \times 32$ latents. Generation uses FP32, 32 DDIM steps (Song et al., 2021a), $\eta = 0$, text guidance 7.5, and image guidance 1.5. Each edit receives one deterministic seed from 0–4; test-stratum counts are 43/40/36/42/39, not repeated evaluations of every edit.

Masks and outcomes. RGBA alpha below 0.5 defines the target mask, dilated by 8 pixels; its complement is non-target. Six examples with non-target area below 20% remain included. Sensitivities use dilation 0/4/8/16. Non-target LPIPS is mask-weighted spatial VGG LPIPS (v0.1); DINO uses normalized patch embeddings and masked patch discrepancies, with CLS secondary. Nonlocal patch context precludes strict spatial isolation. Editing uses bicubic resizing to 224 without an explicit antialias option, so magnitudes are not calibrated to the RF++ pipeline. Prediction's RGB metric is mean absolute change; feedback uses RMSE. Target LPIPS compares to the supplied target image, and target success is its negative exponential. Backend audits record no fallback encoder.

Expansion estimator and cost. With diffusion length $N = 1000$, the guided reverse-progress field is

$$v_s(x) = \tfrac{1}{2} N \beta_t \left[x - \frac{\epsilon_{\text{guided}}(x, t)}{\sqrt{1 - \bar{\alpha}_t}} \right]. \tag{23}$$

It is evaluated along DDIM states; DDIM increments need not be Euler steps in this field. Branchwise autodifferentiation supplies $J_s = D_x v_s$ and transpose products. Ten Lanczos steps estimate the leading Ritz value of $H_s = (J_s + J_s^\top)/2$; clipping at zero gives $\widehat{\rho}_s^+$. Risk $R_\rho = T_{0.1}(\widehat{\rho}^+)$ averages four of 32 values. Ritz values are approximate. A separate 500-query spatial-Jacobian audit has median finite-difference relative error 0.0722 and no nonfinite values. Profiling costs 1,920 branchwise derivative products per edit ($32 \times 10 \times 3 \times 2$); recorded median generation/profiling times are 4.530/435.722 seconds on two RTX 4090s. This is not a demonstrated cheap online predictor.

Table 18: Non-target LPIPS prediction on 200 test edits (24 high-drift positives). Spearman has 95% edit-bootstrap intervals; AUROC/AUPRC retain the recorded score direction. Negative association does not imply no information.

Predictor	Spearman [95% CI]	AUROC	AUPRC
Expansion ρ^+	0.506 $[0.387, 0.613]$	0.735	0.237
Image complexity	0.160 $[0.033, 0.289]$	0.624	0.173
Mask fraction	0.148 $[0.008, 0.281]$	0.596	0.171
Dilated-mask fraction	0.114 $[-0.022, 0.249]$	0.576	0.161
Instruction length	-0.101 $[-0.233, 0.039]$	0.510	0.139
Instruction embedding norm	-0.081 $[-0.220, 0.054]$	0.346	0.090
Latent path length	-0.590 $[-0.698, -0.464]$	0.309	0.123
Velocity energy	-0.781 $[-0.835, -0.711]$	0.240	0.077
Legacy field-change energy	-0.851 $[-0.880, -0.810]$	0.192	0.075
Target-success score	-0.544 $[-0.648, -0.426]$	0.150	0.071
Contraction magnitude ρ^-	-0.828 $[-0.863, -0.778]$	0.110	0.070
Positive edit–control mismatch	0.463 $[0.340, 0.572]$	0.718	0.254

Prediction and adjustment. On 200 test edits, expansion–LPIPS Spearman is 0.506 with interval $[0.387, 0.613]$ and AUROC 0.735; 24 edits exceed the validation threshold. Spearman intervals use 10,000 edit resamples: expansion/mismatch retain their respective primary-analysis intervals, and other scores use the baseline table. AUROC/AUPRC retain recorded score directions. A negative association or AUROC below 0.5 does not mean no information; score reversal was not selected on test outcomes.

On 600 training edits, the standardized expansion coefficient is 0.0936 (95% interval [0.0437, 0.1434]); partial $R^2 = 0.0228$. Controls include masks, latent motion, image/instruction descriptors, category, and target success. Some are post-generation variables, so adjustment is associational. Held-out partial Spearman is about 0.451 (primary controls) or 0.345 (expanded); mask sensitivities give raw Spearman about 0.506–0.508.

D.4.2 EDIT–CONTROL MISMATCH

The matched no-edit run uses the same source and initial noise with an empty instruction; it is not an identity map. Mismatch is $R_{\text{mismatch}} = T_{0.1}([\widehat{\rho}^{+}_{\text{edit}} - \widehat{\rho}^{+}_{\text{ctrl}}]_{+})$. Its test LPIPS Spearman is 0.463 ($[0.340, 0.572]$), AUROC 0.718. The training-fitted mismatch coefficient is 0.2134 with unadjusted interval $[0.0315, 0.3953]$, partial $R^2 = 0.0090$, and Holm $p = 0.259$. Held-out $\Delta R^2 = 0.0020$ has interval $[-0.0135, 0.0167]$. Paired advantages over raw expansion are -0.0429 for Spearman ($[-0.0936, 0.0078]$) and -0.0168 for AUROC ($[-0.0727, 0.0353]$). Mismatch defines the intervention plan but is not a robustly superior predictor; its unadjusted interval does not override the corrected test.

D.4.3 DOSE SELECTION AND EQUAL-GAIN STATE-DEPENDENT FEEDBACK

The plan selects four interior times with largest positive mismatch, fixing a unit leading Ritz vector u_k, an orthogonal unit vector p_k, and paired control state x_k^{ctrl}. Amplification (A) and orthogonal feedback (P) use

$$\begin{aligned} \gamma_k &= \min\{\alpha[\widehat{\rho}^{+}_{\text{edit},k} - \widehat{\rho}^{+}_{\text{ctrl},k}]_{+}, 40\}, \\ \Delta v_{A,k}(x) = \gamma_k u_k u_k^{\top}(x - x_k^{\text{ctrl}}), &\qquad \Delta v_{P,k}(x) = -\gamma_k p_k u_k^{\top}(x - x_k^{\text{ctrl}}). \end{aligned} \tag{24}$$

Norms agree at the same state, but the arms evolve at different states; equal gain therefore need not yield equal realized magnitude. Orthogonal feedback is not geometrically neutral: its added symmetric Jacobian is $-\gamma_k(p_k u_k^{\top} + u_k p_k^{\top})/2$, with plane eigenvalues $\pm\gamma_k/2$.

Selection and confirmation. After unsuccessful bidirectional/lower-dose stages, 50 validation edits compare $\alpha = 1.2, 1.6, 2.4$. The smallest passing dose is 1.6. Gates require positive DINO, LPIPS, and RMSE contrasts, at least four positive DINO seed-stratum directions, target-LPIPS noninferiority margin 0.02, and orthogonal-control DINO equivalence margin ± 0.003. The 300 confirmation edits come from the original predictive training split, excluding earlier intervention edits; they are new to intervention, not to every prior analysis. The recorded lock reduces 400 planned edits to a deterministic 300-edit prefix after ten outputs existed, reportedly before outcome inspection. Final seed counts are 50/53/68/66/63.

Table 19: Dose selection on 50 validation edits. Entries show amplification-minus-orthogonal DINO drift with 95% intervals; $\alpha = 1.6$ is the smallest dose passing all gates.

α	DINO contrast [95% CI]	All gates	Decision
1.2	0.0061 [0.0022, 0.0115]	No	RMSE gate failed
1.6	0.0087 [0.0034, 0.0158]	Yes	Selected
2.4	0.0159 [0.0058, 0.0297]	Yes	Larger passing dose

Table 20: Equal-gain feedback confirmation ($\alpha = 1.6$, 300 edits): paired means and 95% intervals. A: amplification; P: orthogonal feedback; 0: baseline. Positive drift means greater damage; seed strata partition the edits.

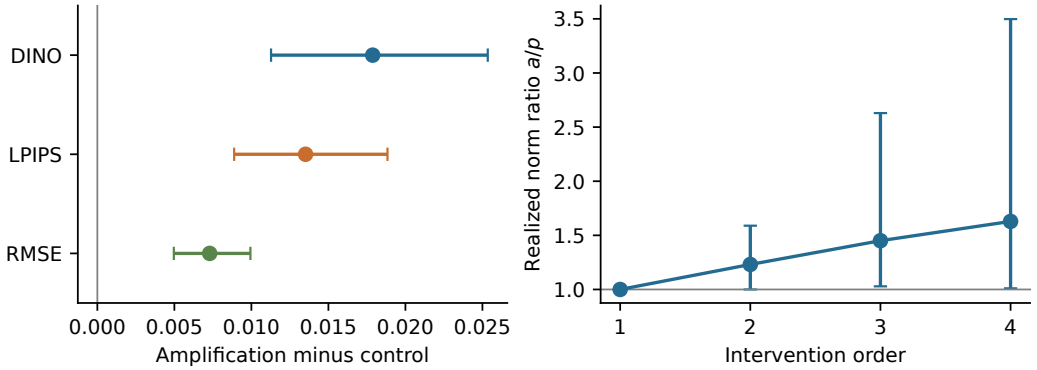
Contrast	n	Mean [95% CI]	Positive strata
DINO, $A - P$	300	0.01788 [0.01128, 0.02535]	5/5
LPIPS, $A - P$	300	0.01352 [0.00888, 0.01884]	5/5
RMSE, $A - P$	300	0.00729 [0.00497, 0.00994]	5/5
Target LPIPS, $A - 0$	300	0.00380 [0.00093, 0.00695]	3/5
DINO, $P - 0$	300	0.00039 [0.00005, 0.00074]	5/5

Intervals use 10,000 paired edit resamples. Amplification increases non-target DINO, LPIPS, and RMSE relative to orthogonal feedback in every seed stratum. Target LPIPS worsens slightly versus baseline, but its upper interval stays below 0.02; control DINO satisfies the stated equivalence margin, which does not mean zero effect.

Across 1,200 paired interventions, median amplification/control norm ratios are 1.000, 1.231, 1.450, and 1.629. Ratios exclude zero-control pairs; discrepancy counts use all 300 pairs per time. For

Table 21: Realized feedback norms in 1,200 paired interventions. Ratios use nonzero-control pairs; discrepancy counts use all 300 pairs at each intervention. The relative-discrepancy definition is given in the text.

Order	Nonzero P	Median a/p	Both zero	Discrepancy $> 1\%$
1	250	1.000	50	0/300
2	299	1.231	1	241/300
3	297	1.450	3	284/300
4	297	1.629	3	279/300



(a) Paired effects (95% CI)

(b) Median and interquartile range

Figure 12: Feedback confirmation: (a) paired effects with 95% bootstrap intervals; (b) norm-ratio medians and IQRs among nonzero-control pairs. Later realized magnitudes differ despite equal gain; this is a policy comparison, not an equal-dose direction-only effect.

realized norms a, p, discrepancy exceeds 1% when $|a - p|/\max(a, p) > 0.01$, with both-zero pairs assigned zero. Figure 12(b) shows IQRs, not confidence intervals. Saved-plan reproduction matches Ritz values within 3.82×10^{-5} and orthogonality within 1.08×10^{-8}, but relative Ritz residuals have median 2.83×10^{-4}, 95th percentile 0.0269, and maximum about 1.066. Reproduction does not make every direction an exact eigenvector.

These are effects of frozen feedback policies. They do not identify a direction-only effect at equal realized dose or an infinitesimal equal-norm perturbation under an unchanged field. Failed damping tests also preclude claiming automatic editing repair.

D.4.4 Post-hoc geometric decomposition

The legacy "bending curvature" score differences stored field velocities without removing tangential change. We reanalyze all 1,000 saved trajectories and evaluate on the original 200 test edits, without new model/eigensolver calls. This is post-hoc, not independent confirmation.

There are 32 equally spaced registered progress nodes $\tau_j \in [0, 1]$; the last stores the actual final latent at diffusion time zero. For field values w_j, use $a_j^w = (w_j - w_{j-1})/(\tau_j - \tau_{j-1})$, duplicating the first interval at the first node. For states $z_j \in \mathbb{R}^d$, compute $u_j = D_\tau z_j$ and $b_j = D_\tau u_j$ by central interior and second-order one-sided endpoint differences. With $P_u^\perp = I - uu^\top/\|u\|^2$,

$$a_j^{w,\perp} = P_{w_j}^\perp a_j^w, \qquad B_{z,j} = \frac{\|P_{u_j}^\perp b_j\|^2}{d}, \qquad \widehat{\kappa}_{z,j}^2 = \frac{\|P_{u_j}^\perp b_j\|^2}{\|u_j\|^4}. \tag{25}$$

Field-change projection is algebraic, not necessarily the DDIM curve's normal acceleration. All speeds are nonzero. Energies are normalized per latent dimension; curvature uses raw coordinates. The arc integral uses speed-weighted quadrature; other scores use top-10% temporal means. Interior sensitivity removes two nodes per end. Strides two/four coarsen the same path, not independent integrations or a convergence test.

Table 22: Post-hoc geometry–LPIPS Spearman correlations on 200 test edits, with unadjusted 95% stratified-bootstrap intervals. Scores use temporal top-10% means except the arc integral; these are not independent confirmations.

Quantity	Spearman [95% CI]
Legacy total field-change energy	-0.851 $[-0.880, -0.811]$
Projected field-change normal energy	-0.822 $[-0.855, -0.780]$
Projected field-change tangent energy	-0.841 $[-0.873, -0.799]$
Recorded-curve normal energy	-0.787 $[-0.830, -0.729]$
Recorded-curve squared curvature	0.636 $[0.532, 0.723]$
Arc-weighted squared-curvature integral	0.761 $[0.683, 0.817]$
Interior normal energy	-0.321 $[-0.468, -0.172]$
Interior squared curvature	0.725 $[0.638, 0.792]$
Squared curvature, stride 2	0.579 $[0.467, 0.671]$
Squared curvature, stride 4	0.492 $[0.377, 0.589]$
Expansion ρ^+	0.506 $[0.387, 0.612]$

Table 23: Expansion–geometry Spearman correlations on the same 200 test edits. The legacy point estimate reproduces -0.408; intervals belong to the post-hoc reanalysis. Pairwise association does not identify independent causal channels.

Second score	Spearman with R_ρ [95% CI]
Legacy total field-change energy	-0.408 $[-0.526, -0.278]$
Projected field-change normal energy	-0.366 $[-0.488, -0.235]$
Recorded-curve normal energy	-0.419 $[-0.536, -0.291]$
Recorded-curve squared curvature	0.271 $[0.128, 0.402]$
Interior squared curvature	0.374 $[0.235, 0.489]$

Intervals use 2,000 paired edit resamples within fixed seed strata, reranking each resample. They are unadjusted post-hoc intervals, distinct from the original analysis. Expansion correlates -0.4082 with legacy field-change energy, which correlates -0.8508 with LPIPS; the latter point estimate is reproduced. Expansion instead correlates 0.2713 with state-curve squared curvature. Normal energy and speed-normalized curvature have different signed LPIPS associations, also sensitive to endpoint stencils. These results do not identify independent causal channels or a general inverse expansion–curvature relation.

Maximum relative legacy-score reproduction error is 2.36×10^{-7}; projection-identity residual is below 10^{-15}. These verify arithmetic, not continuous-time accuracy. Median relative state-derivative/field discrepancy is 0.534, reinforcing their distinction. Thus the old negative "bending" correlation does not establish negative pure-curvature association. Per-edit latent curvature also differs from the population quantity in Section D.2.6.

D.5 Negative results and limitations

Global scheduling and real-data quality. The global risk schedule improves DINO tracking at NFE 8/16 but not archived official FID at any tested NFE. Tracking averages five equal-size seed groups (10,000 samples each); negative differences favor risk scheduling, without constructing pooled intervals from seed intervals. FID is computed once on 50,000 pooled images, not averaged across seed FIDs; no intervals support significance claims for its small differences. The 200-step reference has FID about 2.348 versus 2.260 for uniform NFE 8, so closer tracking need not improve real-data quality.

Smaller mixing weights yield no established FID advantage. Distribution- and quality-aware searches fail nontrivial-candidate gates and revert to zero strength or standard sampling. Local-defect, classwise, and error-budget diagnostics do not convert these failures into successful comparisons.

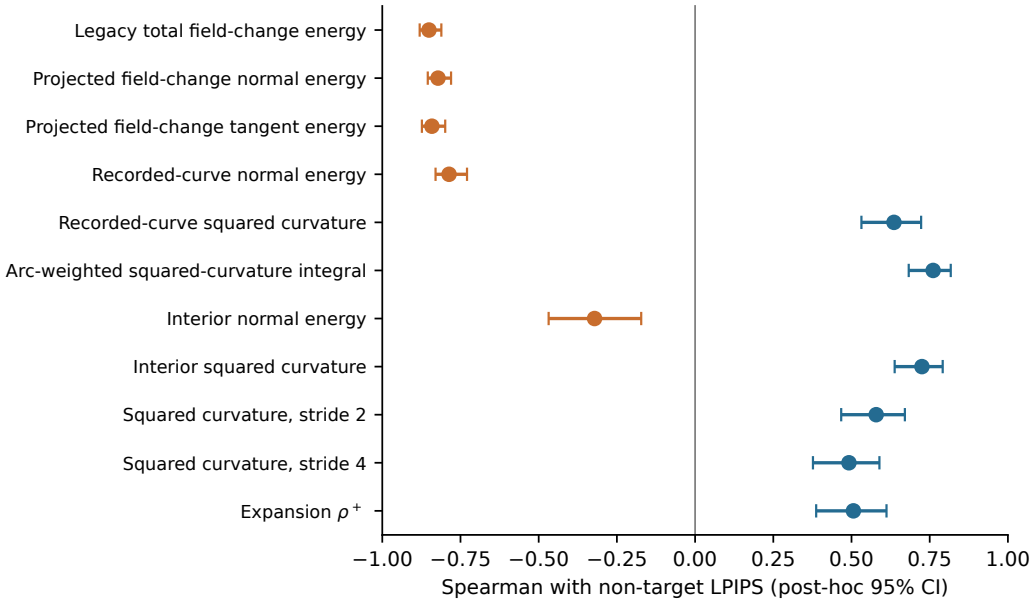


Figure 13: Post-hoc geometry–LPIPS associations with unadjusted 95% intervals, conditional on test edits and seed strata. Normal energy and squared curvature have different signs; these are not independent causal effects or intrinsic distribution-level curvature estimates.

Table 24: Global-schedule DINO discrepancy to the 200-step reference, averaged over five equal-size seeds (10,000 samples each). Negative differences improve tracking; no pooled intervals are supplied. FID is reported separately.

NFE	Uniform	Risk schedule	Risk − uniform	Improved seeds
4	0.062662	0.062752	0.000091	1/5
8	0.039115	0.033941	-0.005174	5/5
16	0.023072	0.017244	-0.005828	5/5

Table 25: Global schedule ($\alpha = 0.25$): official FID on 50,000 pooled images, not mean seed FID. Positive differences favor uniform sampling. No intervals were archived; significance is not established.

NFE	Uniform	Risk schedule	Risk − uniform
4	2.31661	2.33544	0.01883
8	2.25998	2.27682	0.01684
16	2.26616	2.28687	0.02071
32	2.28723	2.31195	0.02472

Distillation and endpoint matching. A bank of 300,000 teacher trajectories (seeds 1000–1004) undergoes five sequential matching attempts, not independent training-seed replications. All pass class-histogram matching but fail the FID/KID gates: the best maximum pairwise FID gap is 0.7083 versus tolerance 0.05. KID intervals use 10,000 resamples of 100 paired subset contrasts (1,000 examples/subset); the gate requires every pairwise interval to contain zero. This screening criterion does not prove distributional equivalence. Table 26 retains decision-audit gaps and Inception/DINO PCA dimensions; no valid matched-teacher student comparison follows.

Table 26: Five sequential endpoint-matching attempts. PCA dimensions are Inception/DINO; FID gaps are maxima over group pairs. High-minus-low KID contrasts and 95% intervals are scaled by 10^3. All attempts fail quality matching.

Attempt	PCA dims.	Max. FID gap	KID contrast $\times 10^3$	Gate
1	32/16	0.8420	$0.866\ [0.744, 0.990]$	Failed
2	64/16	0.8018	$0.709\ [0.574, 0.843]$	Failed
3	96/24	0.7977	$0.815\ [0.684, 0.947]$	Failed
4	128/32	0.7336	$0.810\ [0.686, 0.941]$	Failed
5	192/32	0.7083	$0.638\ [0.498, 0.774]$	Failed

Conditional distillation also fails matching: whitened Inception PCA-64 FID gap is about 0.480 versus tolerance 0.082, with KID failure. This is distinct from official full-feature FID. Exploratory students use one training seed. At NFE 16, high-minus-low DINO contrasts are -0.0327 for group-specific students and -0.0387 for the shared student, opposite to the hypothesized ordering. Unmatched endpoints and limited replication prevent attributing distillation difficulty causally to geometry.

Damping and replication limits. Early damping-minus-base LPIPS is -0.00553 with interval $[-0.01141, 0.00027]$; the separate $\alpha = 0.8$ amplification-minus-control DINO contrast is 0.00374 with interval $[-0.00371, 0.01133]$. Neither passes all criteria. Later amplification confirmation answers a narrower feedback-policy question and does not validate these failed tests.

D.6 Reproducibility

D.6.1 Evaluation units and source traceability

Table 27 separates development, confirmation, and reused trajectory analyses. RF++ seeds vary noise for a fixed checkpoint; editing seed strata partition edits. Neither substitutes for independent model training. Exploratory intervals are unadjusted unless stated; interval overlap is not a paired-comparison test. Resampling units and conditioning are specified per experiment.

D.6.2 Checkpoints, environments, and reproduction

Checkpoints are `cifar-configG.pth` and `instruct-pix2pix-00-22000`. The recorded editing environment uses Python 3.8.5, PyTorch 1.12.1, CUDA 11.3, torchvision 0.13.1, and transformers 4.33.3 with 1 RTX4090; these versions are not asserted for every later analysis.

Source tables, scripts, configurations, and SHA-256 manifests trace each result. Intervals retain archived draws; the feedback summary recomputes 1,200 application pairs, and editing decomposition includes trajectory hashes and analysis code. The directory inventory includes pilots, debugging, empty placeholders, and derived artifacts, without counting them as independent confirmations. Together with the limits above, this supports operational prediction, offline refinement allocation, and specified feedback-policy comparisons, not universal curvature–quality or cost-inclusive efficiency claims.

E Limitations and Scope

Our theoretical results characterize second-order geometric mechanisms under controlled assumptions, while empirical studies examine representative generative trajectories and editing dynamics.

Table 27: Evaluation populations, analysis roles, and dependence. RF++ seeds vary noise for one checkpoint; editing strata partition the edits.

Analysis	Evaluation population	Role and dependence
Constant-speed / reparameterization	Analytic curves; no trained model	Deterministic checks of specific geometric statements
RF++ original / later prediction	10,000 samples per seed; seeds 0–4; NFE 4, 8, 16	Seed 0 for validation; four test seeds, 40,000 outcomes per NFE
Dense-reference geometry	64 paired trajectories per seed; five seeds	Five seed-level geometry observations; shared across evaluated NFE
Local refinement, development	1,280 samples; five seeds	Strategy development; not pooled with confirmation
Local refinement, confirmation	20,480 fresh samples; five seeds	Frozen common sample gate and interval policies
Total-budget ablation	1,280 paired samples; five seeds	Separate test of total model-call efficiency
Editing prediction / mismatch	1,000 edits; 600/200/200 split	Source-image-disjoint training, validation, and test
Feedback confirmation	300 edits; five seed strata	Unused in earlier interventions; drawn from the predictive training split
Editing geometry reanalysis	1,000 stored trajectories; 200 test edits	Post-hoc reuse of the original predictive cohort

The proposed allocation principle targets trajectory-level integration fragility rather than directly optimizing distributional quality metrics. Similarly, the spectral intervention identifies a local expansion mechanism under the evaluated dynamics rather than a universal correction rule. Extending these analyses to broader generators, online risk estimation, and additional intervention strategies are promising directions for future work.