STRAIGHTNESS AND INTRINSIC GEOMETRY IN OT FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimal transport (OT) flow matching promotes straight ambient trajectories, but exact ambient fitting need not recover intrinsic transport velocities on curved supports. We quantify this mismatch and examine whether tangent-projected straightness can detect it. For short circle arcs and full-dimensional product spherical bands, we prove shared ambient and intrinsic optimality and derive exact midpoint velocity-gap laws. The gap determines the leading excess kinetic action of the projected path, despite matching endpoint laws. A structured product-torus extension accommodates correlated source densities. Exact calibrations show that the diagnostic can vanish despite a positive intrinsic gap, while nonlinear representations can produce a positive score on a flat path. Separate CIFAR-10 experiments evaluate training behavior and sample selection without intrinsic ground truth. Across two scoring representations and three quality metrics, lower-score retention performs worse on average than matched random retention; reversing DINOv2 rankings yields metric-dependent gains. Projector ablations preserve this selection pattern. An exact finite-pool MMD identity explains why favorable bucket association need not improve the retained distribution.

1 INTRODUCTION

Flow matching learns a velocity field from conditional probability paths (Lipman et al., 2023; Albergo et al., 2025). Ambient optimal transport (OT) can pair endpoints for linear interpolation (Pooladian et al., 2023; Tong et al., 2024). Minimizing intrinsic kinetic action is a path-design objective in dynamic OT and geometric flow matching (Benamou & Brenier, 2000; Shaul et al., 2023; Chen & Lipman, 2024; Kapuśniak et al., 2024). We show that projecting ambient motion onto the manifold need not recover the intrinsic optimum, even under a shared optimal coupling and exact ambient fitting. This isolates interpolation from differences in the coupling or learned-field approximation. Establishing a field discrepancy also requires controlling conditional averaging, which can erase a pairwise difference.

The starting comparator is ambient straightness: Liu et al. (2023, Eq. (3)) compare instantaneous velocity with endpoint displacement. We ask whether applying such a comparison in a representation and projecting it onto local data directions certifies intrinsic fidelity. We propose this feature endpoint-chord probe here; prior ambient straightness results do not claim that interpretation. A prefix-only version uses a one-step endpoint forecast. Its possible value for sample selection is a separate question from its validity as an intrinsic measurement.

What gap remains, and what does it imply? For short circle arcs and full-dimensional product spherical bands, we prove shared ambient and intrinsic optimality, coincident midpoint laws, and pair identification. These steps turn a paired calculation into an exact conditional-field law. The gap also determines the leading excess kinetic action of the actual projected path. A product-torus extension permits correlated densities and motion in several coordinates; a general local expansion identifies the directional extrinsic coefficient without asserting common optimality.

Does tangent-projected straightness measure that gap? Exact calibrations show that a zero chord residual remains zero under any linear projection even with positive intrinsic discrepancy; nonlinear features can instead yield a positive score on a flat path. These constructions delimit a comparator family, rather than all geometric diagnostics.

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.

Can the ranking still help select samples? Failure as a geometric measurement does not settle this question. Matched-subset experiments show that lower-score retention loses at the primary setting across two feature representations and three quality measures; reversed DINO rankings give metric-dependent gains. An exact finite-pool MMD identity explains why favorable bucket association need not improve the retained distribution. The resulting evaluation procedure separates a known geometric target, calibration of a proposed observable, and a direct test of its decision value.

2 Related work

Transport and ambient straightness. Dynamic OT and displacement interpolation (Benamou & Brenier, 2000; Ambrosio et al., 2008; Villani, 2009) motivate ambient straight paths in rectified flow and OT-CFM (Liu et al., 2023; Pooladian et al., 2023; Tong et al., 2024; Hertrich et al., 2025). Related coupling constructions include kinetic-optimal, semidiscrete, and discrete-state transport (Shaul et al., 2023; Mousavi-Hosseini et al., 2026; Haxholli et al., 2026). Bansal et al. (2026) connect acceleration to discretization error and show that endpoint-displacement residuals can miss oscillations. We instead compare ambient and intrinsic targets even when the ambient path is exactly straight.

Manifold geometry and approximation error. Riemannian and metric flow matching specify geometric interpolants (Chen & Lipman, 2024; Kapuśniak et al., 2024). Zaghen et al. (2026) compare endpoint and velocity losses through intrinsic curvature and Jacobi fields. Our coefficient is extrinsic and can be positive on an intrinsically flat product torus. Learning/solver decompositions (Baheri, 2026), endpoint bounds (Benton et al., 2024; Zhou & Liu, 2025), statistical rates (Fukumizu et al., 2025; Kunkel & Trabs, 2025; Kumar et al., 2026), and manifold total-variation bounds (Guan et al., 2026) concern approximation to a target flow. We identify a difference between two target fields under verified shared optimality.

Manifold estimation and representations. Reach (Federer, 1959) supports reconstruction, estimation, and projection regularity (Niyogi et al., 2008; Genovese et al., 2012; Aamari et al., 2019; Boissonnat et al., 2019; Leobacher & Steinicke, 2021). Local approximation depends on smoothness and sampling (Aamari & Levrard, 2019). Local PCA (Singer & Wu, 2012) motivates our neighborhoods; intrinsic-dimension estimation (Levina & Bickel, 2004; Ansuini et al., 2019) is a separate problem. Our variance threshold supplies neither a dimension guarantee nor an intrinsic interpretation. The calibrations test the latter. See Loaiza-Ganem et al. (2024) for the broader generative-modeling context.

3 Objects being compared

Let $\mathcal{M} \subset \mathbb{R}^D$ be a closed C^2 embedded manifold of positive reach τ, and let π denote nearest-point projection in its reach tube. Write P_p for the orthogonal projector onto $T_p\mathcal{M}$. For an endpoint coupling ζ on $\mathcal{M} \times \mathcal{M}$, put $c_t = (1-t)x_0 + tx_1$ and $\Delta = x_1 - x_0$. All projected quantities below are restricted to pairs and times with $\operatorname{dist}(c_t, \mathcal{M}) < \tau$; an arbitrary coupling need not satisfy this condition. Set $p_t = \pi(c_t)$ and let γ_t be a constant-speed minimizing intrinsic geodesic, excluding cut-locus ambiguity where needed. The three velocities are

$$a_t = P_{p_t}\Delta, \qquad b_t = \frac{d}{dt}\pi(c_t) = D\pi(c_t)\Delta, \qquad h_t = P_{p_t}\dot{\gamma}_t. \tag{1}$$

We compare the tangent component of the ambient target with the intrinsic target expressed in the same tangent space. This asks whether fitting the ambient target also recovers the specified intrinsic motion. The operation defining h_t is ambient reprojection from $T_{\gamma_t}\mathcal{M}$ onto $T_{p_t}\mathcal{M}$, not parallel transport. Our pairwise discrepancy is $G_t = \|a_t - h_t\|$. In contrast, b_t describes the actual projected-path dynamics. A positive G_t alone implies neither an incorrect endpoint law nor numerical sampling error. Section 4.1 supplies the coupling and midpoint-law conditions for a corresponding field comparison.

Conditional fields and paired targets. For a prescribed coupling, the ambient regression field is $u_\zeta(t,x) = \mathbb{E}[\Delta \mid c_t = x]$. The *paired-target regression residual* $\|P_{p_t}(v_\theta(t,c_t) - \Delta)\|$ compares a

model with an individual target. Its expected square separates into field approximation and conditional target variance (Appendix C). We distinguish these residuals from the geometric comparison in Eq. (1).

Mean norm, RMS, and integration. We distinguish mean norms $\mathbb{E}[G_t]$ from RMS norms $(\mathbb{E}[G_t^2])^{1/2}$, and normalized discrepancies from their unnormalized counterparts. The exact sphere integral covers $[0, 1]$; empirical integrals use their stated evaluation grids. The learned common-OT controls report midpoint L^2 norms.

4 Exact geometry and the role of reach

Projection regularity. Using the classical differential of nearest-point projection (Leobacher & Steinicke, 2021, Theorem C) and the reach curvature bound, we obtain $\left\|D\pi(x) - P_{\pi(x)}\right\|_{\mathrm{op}} \leq \rho/(\tau - \rho)$ when $\mathrm{dist}(x, \mathcal{M}) \leq \rho < \tau$. The bound is sharp on inward sphere offsets (Proposition 4, Appendix B.1). It controls the projection operation; the transport-field comparison also requires the coupling and base point.

Directional curvature at short distances. On a C^5 embedded manifold, let m and ξ be the midpoint and unit tangent of a short geodesic of length d. With second fundamental form II and shape operator S, Proposition 6 gives

$$G_{1/2} = \frac{d^3}{24} K_m(\xi) + O(d^5), \quad K_m(\xi) = \left\|S_{\mathrm{II}_m(\xi,\xi)}\xi\right\|, \quad \|\mathrm{II}_m(\xi,\xi)\|^2 \leq K_m(\xi) \leq \tau^{-2}. \tag{2}$$

Thus the coefficient is extrinsic and directional; reach supplies an upper bound. On a sphere, $K_m(\xi) = r^{-2}$. Even an intrinsically flat product torus can have a nonzero gap. Appendix B.3 proves the expansion and its weighted sixth-moment form, and checks thirteen cases across five geometry families. This same-pair result does not establish a shared optimal coupling.

The sphere gives a closed-form paired reference (Theorem 5, Appendix B.2). Its tangent-component discrepancy and projected-path velocity differ even on the same great circle. A field claim additionally needs compatible couplings and identifiable conditional targets, which we establish next.

4.1 A shared transport with an identified field gap

A flow-matching field averages conditional targets, so a same-pair calculation need not survive conditioning. Ambient and intrinsic OT can also choose different pairs. The following class resolves both issues: it has a shared optimal coupling, equal projected and intrinsic midpoint laws, and a midpoint that identifies its endpoint pair.

Theorem 1 (Common-OT midpoint field law)**.** *Let two absolutely continuous probability laws on $\mathbb{S}^1(r)$ be supported in a common unwrapped angular interval of length less than $\pi/2$. Their increasing angular transport T induces a coupling ζ optimal for both squared ambient and squared intrinsic distance. For a source angle α, put $\delta = T(\alpha) - \alpha$, $d = r|\delta|$, and $p = \pi(c_{1/2})$. Then $p = \gamma_{1/2}$, and the ambient and intrinsic displacement fields u and w satisfy*

$$B_{1/2} := \left\|P_p u(1/2, c_{1/2}) - w(1/2, p)\right\|_{L^2(\zeta)} = r\left(\mathbb{E}_\zeta[(\delta - 2\sin(\delta/2))^2]\right)^{1/2}. \tag{3}$$

If $|\delta| \leq \varepsilon < \pi/2$, then

$$\left(1 - \frac{\varepsilon^2}{80}\right) \frac{\sqrt{\mathbb{E}d^6}}{24r^2} \leq B_{1/2} \leq \frac{\sqrt{\mathbb{E}d^6}}{24r^2}. \tag{4}$$

Both angular transport costs are strictly convex functions of the displacement on this interval, giving the same increasing coupling. The midpoint angle is $(\alpha + T(\alpha))/2$, which identifies each pair. The two tangent velocity components are $2r\sin(\delta/2)$ and $r\delta$; elementary sine bounds yield Eq. (4). Appendix B.5 proves these claims, including the equality of the projected and intrinsic midpoint laws.

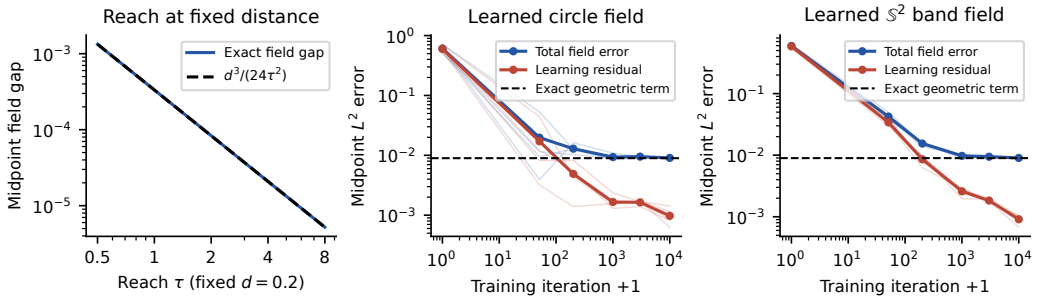

Figure 1: Verified common-OT field laws. Left: varying reach at fixed intrinsic displacement. Middle and right: learned circle and full-dimensional sphere-band fields, with thin lines for five individual seeds and thick lines for their means. The geometric term is analytic, not fitted to learning curves.

Theorem 2 (Full-dimensional spherical bands). *On $\mathbb{S}^m(r)$, $m \geq 2$, write $X(a,\omega) = r(\sin a\,\omega, \cos a)$, where $\omega \in \mathbb{S}^{m-1}$. Let $\mu_i = X_\#(\nu_i \otimes \lambda)$, where $\nu_i \ll da$ are supported on a common interval $I \subset (0,\pi)$ of length less than $\pi/2$, and the common λ is absolutely continuous on $\mathbb{S}^{m-1}$. These laws are absolutely continuous with respect to full surface area. If T is the increasing transport from ν_0 to ν_1, the map $X(a,\omega) \mapsto X(T(a),\omega)$ is optimal for both squared costs. With $\delta = T(a) - a$, Eqs.* (3)–(4) *hold for its midpoint fields.*

The two sphere costs dominate the corresponding polar-angle costs; keeping ω fixed attains both bounds. The midpoint coordinates $((a + T(a))/2, \omega)$ identify the pair. Appendix B.6 gives the proof. Independence and a common directional law are substantive restrictions; no globally positive density is required.

For a learned field, let $D = \|P_p v_\theta - w(p)\|_{L^2}$ and $e = \|P_p(v_\theta - u)\|_{L^2}$. Then $|D - B_{1/2}| \leq e$: exact ambient learning approaches a positive intrinsic gap. Appendix B.8 extends shared optimality and midpoint identification to coordinatewise monotone transports of correlated product-torus densities, allowing several coordinates to move. This remains a structured class, not a general coupled-manifold OT theorem.

4.2 A CONSEQUENCE FOR ACTUAL PROJECTED-PATH ACTION

The preceding gap compares $P_p u$ with the intrinsic target. To relate it to actual dynamics, put $\bar{\mu}_t = (\pi \circ c_t)_\#\zeta$ and define $\mathcal{A}_{\mathrm{proj}} = \int_0^1 \int \|b_t(p)\|^2 \, d\bar{\mu}_t(p)\, dt$, without a factor of one half.

Corollary 3 (Projected-path action). *Under either Theorem 1 or 2, the projected point identifies its pair for every $0 < t < 1$, and $b_t(p_t) = D\pi(c_t)u(t,c_t)$ is its conditional velocity. Then*

$$\begin{aligned} \mathcal{A}_{\mathrm{proj}} - W^2_{2,\mathcal{M}}(\mu_0,\mu_1) &= r^2\mathbb{E}[1 - \cos\delta + \delta\tan(\delta/2) - \delta^2] \\ &= \frac{16}{5}B^2_{1/2} + O(r^2\mathbb{E}[\delta^8]). \end{aligned} \tag{5}$$

The excess is positive for nonidentity transport; the expansion holds for uniformly short angles.

Appendix B.7 proves all-time identification and integrates the projected speed. Both paths follow the same geodesic arc with different time parametrizations: correct endpoints do not imply minimum intrinsic action. Figure 3 separates the velocity magnitudes and illustrates the short-distance action scaling. This consequence applies when intrinsic action is the evaluation target; it is not an image-quality or endpoint-error bound. Global reach alone supplies no lower law outside these classes (Appendix B.9).

5 DEFINING AND CALIBRATING THE FEATURE PROBE

What the probe compares. The ambient functional $\int_0^1 \mathbb{E}\,\|\dot{x}_t - (x_1 - x_0)\|^2\,dt$ measures straightness (Liu et al., 2023). We study a feature-space, tangent-projected norm version and an endpoint-free forecast, without assuming an intrinsic interpretation. For a feature map F, an estimated tangent projector $\widehat{P}_t$, and step ϵ, define

$$U_t^\epsilon = \frac{F(x_t + \epsilon v_\theta(t, x_t)) - F(x_t)}{\epsilon}, \qquad \widehat{G}_t(x_0, x_1) = \left\| \widehat{P}_t\{U_t^\epsilon - [F(x_1) - F(x_0)]\} \right\|. \quad (6)$$

For images, $F = \Phi \circ C$, with $C(x) = (\mathrm{clamp}(x, -1, 1) + 1)/2$ and a fixed InceptionV3 encoder (Szegedy et al., 2016), later crossed with DINOv2. Centered PCA on neighboring real training features retains 95% of variance, and $\epsilon = 0.01$. These choices define the operational score; they do not supply an intrinsic endpoint geodesic.

Information available during generation. If x_1 is a completed generated image, Eq. (6) is retrospective at earlier times. We therefore also define the endpoint-free forecast

$$\widetilde{x}_1(t) = x_t + (1 - t)v_\theta(t, x_t), \qquad S_T = \mathrm{Trapz}_{t \le T} \left\| \widehat{P}_t\{U_t^\epsilon - [F(\widetilde{x}_1(t)) - F(x_0)]\} \right\|. \quad (7)$$

Its inputs are available from the prefix and fixed training bank. This makes prospective ranking possible, without establishing that the ranking is useful.

Two calibration failures. Even for $v_\theta = \Delta$ and a perfectly straight ambient path, $DF(c_t)\Delta$ need not equal $F(x_1) - F(x_0)$. On the linear interpolant c_t, if F is C^2 along the relevant segments, with $\|DF\| \le L$ and $\left\|D^2F\right\| \le H$, any orthogonal projector satisfies

$$\widehat{G}_t \le L\,\|v_\theta - \Delta\| + \frac{H}{2}\,\|\Delta\|^2\,\{t^2 + (1 - t)^2\} + \frac{\epsilon H}{2}\,\|v_\theta\|^2. \quad (8)$$

Appendix B.11 proves this identity-based bound. In the flat example $x_t = t$, $v = 1$, $F_\alpha(x) = (x, \alpha x^2)$, the true ambient discrepancy is zero, while projection onto the exact feature tangent gives

$$\widehat{G}_t = \frac{2\alpha^2 |t(2t - 1)|}{\sqrt{1 + 4\alpha^2 t^2}} \quad (9)$$

in the derivative limit (Figure 2). Conversely, affine features give zero retrospective and forecast residuals for an exact straight ambient field although $B_{1/2} > 0$ in a nonidentity common-OT construction. Any linear projection and zero-preserving aggregation retain this zero, so this comparator family cannot certify intrinsic error by a bound $B_{1/2} \le f(S)$ with $f(0) = 0$ (Corollary 12). With identity features and the exact tangent projector, learned-field midpoint RMS scores are e and $e/2$, while $|D - B_{1/2}| \le e$ (Proposition 11). Better tangent estimation alone cannot repair this cancellation. The nonlinear example does not identify every image signal as a representation artifact.

From calibration to utility. A score can fail to identify intrinsic discrepancy yet still rank useful samples. We therefore test selection directly against equal-sized random subsets in Section 6.3, and distinguish bucket association from distributional gain in Section 6.4. The smooth calibrations do not cover image clipping, and the image study supplies no intrinsic ground truth. It tests measurement behavior and selection utility separately.

6 EXPERIMENTS

The known-manifold controls test the field law and probe against exact intrinsic targets. The image experiments ask whether the score tracks a paired-target residual during learning or helps select samples when intrinsic ground truth is unavailable. Error bars and $\pm$ values denote sample standard deviation across training seeds; checkpoints within a run are dependent. The supplement provides per-seed values and protocols.

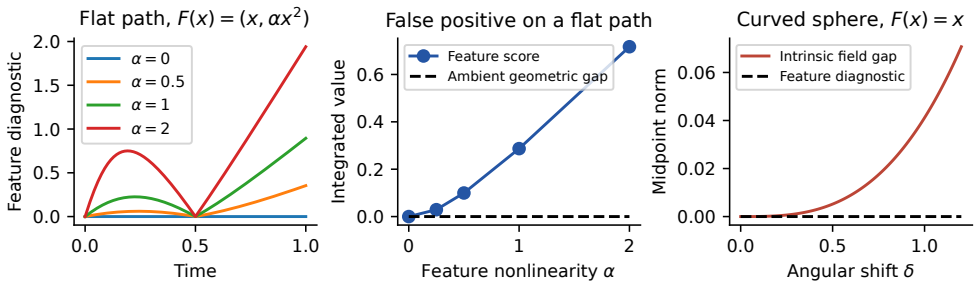

Figure 2: Complementary calibration failures without tangent-estimation error. Left and middle: a flat path acquires a positive diagnostic under a nonlinear feature map. Right: identity features on a unit sphere give a zero diagnostic despite the positive common-OT midpoint field gap $\delta - 2\sin(\delta/2)$. The right panel uses constant angular shifts on short arcs or spherical bands.

6.1 Controlled optimal-transport fields

We check seven increasing transports, including a nonlinear map, negative and mixed-sign displacements, and the identity. Independent ambient and intrinsic assignment solvers agree for circle arcs and product grids on $\mathbb{S}^2, \mathbb{S}^3, \mathbb{S}^5$; direct vector evaluation matches the field formulas within 6.1×10^{-17}. A radius sweep fixes intrinsic pair distance at $d = 0.2$, verifying reach dependence without changing transported distance.

An additional sweep varies radius, mean transported distance, and prescribed residuals independently in 24 sphere-band transport configurations. Across 288 grid combinations, the retrospective and instantaneous forecast midpoint RMS scores with identity features equal e and $e/2$, while D depends on residual direction. At $r = 0.5$ with constant displacement $d = 0.4$, score 0.004 accompanies $D = 0.006582$ toward the intrinsic target or 0.014582 away from it. Feature and tangent-projector controls further alter the score at fixed intrinsic discrepancy on the original sphere (Appendix B.12). These prescribed perturbations are separate from the learned controls below.

We train five MLPs on each of two ambient flow-matching tasks for 10,000 iterations. The circle source angle is uniform on $[-0.15, 0.15]$; the $\mathbb{S}^2$ source has independent uniform polar angle on $[0.8, 1.1]$ and uniform azimuth. Both shift the polar/arc angle by 0.6, preserving direction. The exact geometric term is 0.008960. The circle's final total error is 0.009056 ± 0.000096 with learning residual 0.000976 ± 0.000315; the sphere band's is 0.009018 ± 0.000047 with residual 0.000911 ± 0.000158. Measurements at all recorded iterations satisfy $|D - B_{1/2}| \le e$. Figure 1 shows approach to the predicted value. Appendix I gives settings and per-seed results.

For the exact target paths, excess action is 0.074% of the intrinsic optimum at the controls' shift $\delta = 0.6$, and 1.29% at $\delta = 1.2$ (Figure 3).

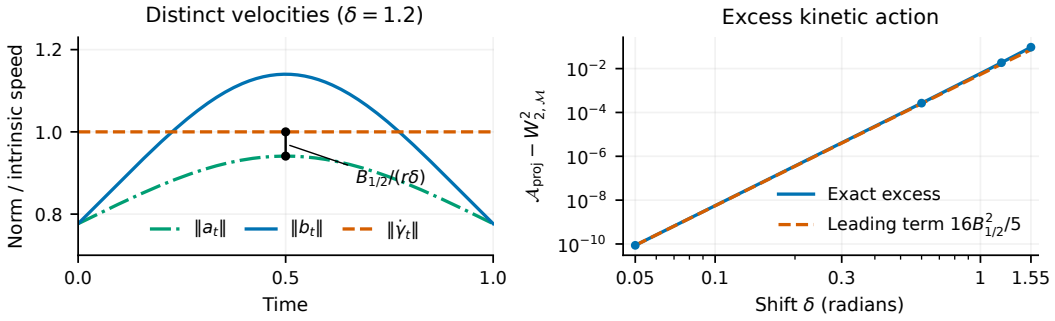

Figure 3: Common-OT unit-circle shifts from source angles uniform on $[-0.005, 0.005]$. Left: velocity norms divided by intrinsic speed at $\delta = 1.2$; tangent spaces coincide at the midpoint. Right: exact excess action and its leading term. All curves are analytic.

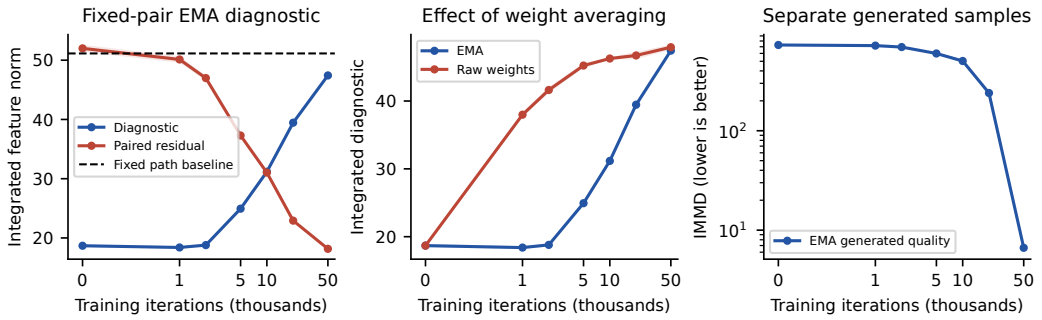

Figure 4: Three fresh early-training trajectories. Shading gives sample SD. Diagnostic and residual norms do not add; the fixed baseline is independent of model weights. Raw weights respond earlier than EMA. Generated quality uses separate ODE samples. The iteration axis is linear through 1k and logarithmic thereafter; IMMD uses a log scale.

6.2 Dense early-training diagnostics and generated quality

Does a smaller diagnostic indicate a smaller paired-target residual during learning? We train three fresh CIFAR-10 (Krizhevsky, 2009) UNets through 50k iterations and record $0, 1, 2, 5, 10, 20, 50$k checkpoints. Each uses the same 512 noise–real OT pairs, ten times in $[0.05, 0.95]$, and cached $k = 300$ PCA projectors. These are *linear training interpolants*, not generated paths or held-out generalization probes. Raw and EMA weights are evaluated separately, with EMA primary. The 400k cosine learning-rate horizon and 2k warmup are preserved; these fresh runs are distinct from the fixed 400k models used below. Appendix G details the training and execution protocols.

For each fixed pair, the diagnostic vector is $r_t + \beta_t$: r_t compares the learned and paired-target feature lifts, while β_t compares the paired-target lift with the feature endpoint chord. We record their norms separately and verify that β_t is constant across checkpoints. Their scalar norms need not add, and neither component is identified as intrinsic manifold error. The EMA parameter average retains coefficient 0.9999^n on its initialization after n iterations, motivating the raw-weight comparison.

Separately, each EMA checkpoint generates 2,048 samples using 50 Euler steps and common initial noise within a seed. Quality uses unbiased Gaussian-kernel MMD squared in Inception features, scaled by 1,000, with bandwidth 10 and 10,000 matched real test features. We call it IMMD to distinguish it from CLIP-based CMMD (Jayasumana et al., 2024); it is a specified evaluation geometry, not a universal quality measure. Crossed representations and a separate FID encoder (Heusel et al., 2017; Chong & Forsyth, 2020) are used in the selection study. All real/generated comparisons use identical encoding, verified in Appendix H.1.

The diagnostic increases while the paired-target residual and generated IMMD decrease (Figure 4). By 1k iterations the raw diagnostic rises from 18.68 ± 0.24 to 37.98 ± 0.33, while its residual falls from 52.00 ± 0.60 to 27.04 ± 0.11. EMA delays this response: its diagnostic is 18.39 ± 0.20 at 1k and 47.43 ± 0.09 at 50k. By 50k its residual falls to 18.16 ± 0.02 and generated IMMD falls from 729.09 ± 3.08 to 6.64 ± 0.17. A small initial score therefore indicates neither a small residual nor better generated quality. The fixed baseline remains 51.14 and has an endpoint rise independently of model weights (Appendix G); the image curves do not identify intrinsic geometric error.

A post-hoc reconstruction from every released per-pair norm quantifies cancellation between r_t and β_t. Relative to their summed time-integrated squared norms, the negative cross term cancels $96.41 \pm 0.09\%$ at initialization and $24.25 \pm 1.82\%$ for raw weights at 50k. These percentages concern integrated mean squared norms, not a decomposition of Figure 4's mean-norm curves. They quantify opposition and magnitude balance without explaining generated-quality changes (Appendix G.1).

Score	Half	Inc. $10^3\mathrm{MMD}^2$	DINO $10^3\mathrm{MMD}^2$	FID
Random	Matched	6.14 ± 0.45	18.84 ± 0.28	38.97 ± 0.35
Inc. forecast	Lower	10.53 ± 0.88	27.33 ± 0.26	45.62 ± 0.93
Inc. forecast	Upper	9.67 ± 0.58	16.49 ± 1.18	40.89 ± 0.29
Inc. retro.	Lower	9.27 ± 1.15	23.51 ± 0.85	42.52 ± 1.42
Inc. retro.	Upper	12.78 ± 0.68	18.22 ± 1.50	42.54 ± 0.36
DINO forecast	Lower	12.20 ± 0.59	32.73 ± 0.43	48.61 ± 0.37
DINO forecast	Upper	6.66 ± 0.90	12.29 ± 0.79	37.55 ± 0.57
DINO retro.	Lower	10.74 ± 0.82	31.74 ± 0.76	46.68 ± 0.19
DINO retro.	Upper	6.31 ± 0.82	12.09 ± 0.90	37.37 ± 0.41

Table 1: Same 6,144 candidates; 1,024 retained per seed at $T = 0.5, k = 300$. Mean $\pm$ sample SD; lower metric values are better. Both score directions are shown. MMD uses raw Inception features ($\sigma = 10$) or unit DINO features ($\sigma = 1$). FID uses a separate TensorFlow-ported encoder; all metrics use 10,000 test references.

6.3 Prospective filtering and matched baselines

Does ranking improve the retained distribution relative to an equally sized random subset? Before generating the discovery pool, we fix three 400k CIFAR-10 checkpoints, 2,048 trajectories per checkpoint, 50-step float32 Euler integration, primary horizon $T = 0.5$, and 50% retention. Scores integrate the actual times $\{0, 0.10, 0.26, 0.50, 0.74\}$ up to the horizon. Secondary horizons are 0.26 and 0.74; $k = 100$ supplements primary $k = 300$. Exact search over 50,000 training features determines centered PCA neighborhoods. Completed endpoints enter only quality evaluation and the retrospective comparator.

Baselines are pixel velocity, feature velocity, normal feature velocity, nearest-reference distance, and a fixed random split. The predeclared retained direction is the lower score; both halves and every method are reported. Each half contains 1,024 samples, compared with 10,000 real test images. Balanced-label tests use 9,999 permutations, two-sided tails, and the plus-one correction; they test bucket association, not superiority over random retention.

Both Inception lower-score halves lose to matched random retention in every seed. The forecast's high-minus-low IMMD gap is -0.86 ± 0.46, and per-seed permutation tests do not confirm the predicted positive association. The retrospective gap is positive (3.51 ± 0.57), yet its lower half also loses to random: bucket association alone is insufficient. Section 6.4 explains this distinction. Appendix E reports every baseline, secondary setting, permutation result, and computational cost.

Representation, metric, and direction. Crossing Inception and DINOv2 (Oquab et al., 2024) scoring/quality features and adding FID leaves all four lower-score probe selectors worse than random under all three primary metrics in every seed (Table 1). Reversing DINO rankings improves DINO MMD and FID, while Inception MMD worsens on average. This extension was designed after the original Inception results, with settings frozen before DINO outcomes were measured.

A further 6,144-candidate evaluation, with independent noise and settings fixed before generation, repeats the mean pattern. The forecast FID gain occurs in two of three confirmation pools (Appendix F.2). These pools reuse the same three trained models and real references, so they test noise robustness rather than transfer to new models or data.

A fixed DINO 20-neighbor classifier with 97.0% test accuracy shows greater predicted-class imbalance in the lower halves. A post-hoc discovery-pool analysis compares selection with random subsets having identical predicted-class counts. The DINO forecast upper half retains a 3.56 ± 0.27 DINO-MMD gain under this comparison, versus -0.03 ± 0.74 in Inception MMD. Thus the metric-specific DINO benefit persists beyond changes in class counts. Appendix F.1 gives the exact conditional expectation and all results; the generated labels are predictions, not ground-truth classes.

Is the selection pattern specific to PCA? We hold the feature chord residual fixed and compare its PCA-projected norm with no projection and ten random orthogonal frames at the same pointwise ranks. This post-hoc analysis reuses both previously inspected pools. Without PCA, model-mean rank correlations with the original score are 0.938–0.979, with 88.4–94.4% half-membership overlap. Mean lower-half losses and metric-dependent upper-DINO benefits persist. PCA changes effect sizes: on the confirmation pool, upper DINO forecast FID gains are 1.27 with PCA, 0.77 without projection, and 0.71 averaged over random frames. The main selection pattern therefore does not require PCA orientation; differences do not establish intrinsic validity. Appendix F.3 reports all metrics and both directions.

Separating measurement choices. A separate post-hoc audit freezes feature states, forecasts and PCA projectors for 256 trajectories per model. Replacing $\epsilon = 0.01$ by an automatic-differentiation JVP preserves 96.9% and 98.7% of the within-audit lower halves for Inception and DINO. At the same clipped feature input, omitting only the derivative mask lowers agreement to 78.4% and 66.7%. Ranking thus depends on this measurement operation even when finite-step ranking is stable. Appendix F.4 reports every variant. This audit measures ranking sensitivity, not sample quality or intrinsic error.

6.4 An exact criterion for selection gains

A favorable difference between quality buckets need not improve the retained distribution. Here the reference changes from the particular random half in Table 1 to the exact expectation over equally sized random subsets of each fixed pool. For an equal split L, H into subsets of size m, let $\widehat{M}$ be diagonal-excluding MMD squared and write $\Gamma = \widehat{M}(H, Q) - \widehat{M}(L, Q)$. The exact partition identity is

$$\widehat{M}(L \cup H, Q) - \widehat{M}(L, Q) = \frac{1}{2}\left[\Gamma - \frac{m}{2m-1}\widehat{M}(L, H)\right]. \tag{10}$$

The full-pool statistic equals the expected statistic of a uniform random subset of size m. Retaining L therefore beats expected uniform retention *if and only if* its bucket gap exceeds $m/(2m-1)\widehat{M}(L, H)$. The criterion applies to any ranking at this retention fraction. For the Inception retrospective score, the mean gap is 3.51, but the required gap is 9.42, giving a gain of -2.95. This accounts for the loss despite its positive bucket association.

Additional cross-fitted MMD-witness and quartile analyses do not yield monotone quality rankings (Appendix J, Figure 17). Together with predicted-class conditioning, they separate ranking association, retained-distribution quality, and the representation in which a gain occurs. The identity above is retrospective accounting using completed quality features, not a prospective estimator of selection gain.

7 Implications and scope

The common-OT results isolate a discrepancy caused by interpolation, despite a shared optimal coupling, exact ambient fitting, and matching midpoint laws. Equation (5) relates it to excess intrinsic kinetic action. The field formulas and controlled transports provide exact reference cases for calibrating geometric diagnostics. The affine construction shows why a tangent projector cannot recover information absent from the chord residual. For selection, the matched comparisons and Eq. (10) give a distinct criterion: a favorable bucket gap must exceed the exact finite-pool threshold.

The common-midpoint field laws and all-time action results concern the stated circle, product-band, and coordinatewise torus classes; arbitrary coupled-manifold OT remains open. The general local expansion is a same-pair result. Gaussian-to-image experiments supply no intrinsic ground truth. Selection reuses three fixed CIFAR-10 models; three separate fresh runs provide Inception-only training curves. Predicted classes do not establish full semantic coverage, and reversed DINO gains are metric-dependent. Forecast retention requires 76 imagewise velocity evaluations per output versus 50 for direct generation, before feature costs (Appendix E); an equal-cost sampling benefit remains open.

AI USE STATEMENT

Generative AI tools, including OpenAI Codex, assisted with literature search, mathematical claim and proof development, designing and implementing synthetic numerical and image experiments, interpreting results, and manuscript drafting and editing. All authors independently verified all mathematical claims and proofs in this paper. The authors take responsibility for the final content.

ETHICS STATEMENT

This work studies transport geometry and diagnostic limitations using synthetic controls and existing image benchmarks. Feature representations and predicted class labels can reflect biases in their training data. Aggregate image-quality gains do not establish fairness, safety, or reliable decisions about individual images. We report adverse selection outcomes and limits of the tested diagnostics; these diagnostics have not been validated for decisions about individuals.

REPRODUCIBILITY STATEMENT

Appendices give definitions, proofs, data splits, time grids, normalizations, and per-seed results. The supplement contains analytic checks, code and protocols, small controlled-model weights, machine-readable results, and figure/table scripts. Generated features replay the original-pool Inception/DINO quality analyses. Compact kernel sums and covariance spectra replay confirmation MMD and FID arithmetic; re-estimating their kernels or covariances requires the omitted features. Per-pair arrays replay early-training diagnostic aggregates. Large generative checkpoints, full image captures, and omitted feature arrays are identified by hashes. Regeneration requires the specified data, models, and encoders; compact replay does not. The README separates immutable measurement records from scratch evaluation outputs.

REFERENCES

Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177–204, 2019. doi: 10.1214/18-AOS1685. URL https://doi.org/10.1214/18-AOS1685.

Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13(1):1359–1399, 2019. doi: 10.1214/19-EJS1551. URL https://doi.org/10.1214/19-EJS1551.

Michael S. 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. URL https://jmlr.org/papers/v26/23-1605.html.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser Basel, 2 edition, 2008. doi: 10.1007/978-3-7643-8722-8. URL https://doi.org/10.1007/978-3-7643-8722-8.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/cfcce0621b49c983991ead4c3d4d3b6b-Abstract.html.

Ali Baheri. Geometry-Grounded Flow Matching on Compact Manifolds. In *Proceedings of GRaM: the Second Edition of the Workshop on Geometry-grounded Representation Learning and Generative Modeling*, volume 326 of *Proceedings of Machine Learning Research*, pp. 32–44. PMLR, 2026. URL https://proceedings.mlr.press/v326/baheri26a.html.

Vansh Bansal, Saptarshi Roy, Alessandro Rinaldo, and Purnamrita Sarkar. On the Convergence and Straightness of Rectified Flow. In *Proceedings of The 29th International Conference*

on Artificial Intelligence and Statistics, volume 300 of *Proceedings of Machine Learning Research*, pp. 4636–4644. PMLR, 2026. URL https://proceedings.mlr.press/v300/bansal26a.html.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000. doi: 10.1007/s002110050002. URL https://doi.org/10.1007/s002110050002.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error Bounds for Flow Matching Methods. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=uqQPyWFDhY.

Jean-Daniel Boissonnat, André Lieutier, and Mathijs Wintraecken. The reach, metric distortion, geodesic convexity and the variation of tangent spaces. *Journal of Applied and Computational Topology*, 3(1-2):29–58, 2019. doi: 10.1007/s41468-019-00029-8. URL https://doi.org/10.1007/s41468-019-00029-8.

Ricky T. Q. Chen and Yaron Lipman. Flow Matching on General Geometries. In *International Conference on Learning Representations*, pp. 47922–47945, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d1f9936d3be6997ffffab692977eebe6-Abstract-Conference.html.

Min Jin Chong and David Forsyth. Effectively Unbiased FID and Inception Score and Where to Find Them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6070–6079, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Chong_Effectively_Unbiased_FID_and_Inception_Score_and_Where_to_Find_CVPR_2020_paper.html.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A Downsampled Variant of ImageNet as an Alternative to the CIFAR datasets. *arXiv preprint arXiv:1707.08819*, 2017. URL https://arxiv.org/abs/1707.08819.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959. doi: 10.1090/S0002-9947-1959-0110078-1. URL https://doi.org/10.1090/S0002-9947-1959-0110078-1.

Kenji Fukumizu, Taiji Suzuki, Noboru Isobe, Kazusato Oko, and Masanori Koyama. Flow matching achieves almost minimax optimal convergence. In *International Conference on Learning Representations*, pp. 27608–27640, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/44a427b77e9727fcec560e2f8d6925e0-Abstract-Conference.html.

Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax Manifold Estimation. *Journal of Machine Learning Research*, 13(43):1263–1291, 2012. URL https://jmlr.org/papers/v13/genovese12a.html.

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

Yunrui Guan, Krishnakumar Balasubramanian, and Shiqian Ma. Total Variation Rates for Riemannian Flow Matching. *arXiv preprint arXiv:2602.05174*, 2026. URL https://arxiv.org/abs/2602.05174.

Etrit Haxholli, Yeti Z. Gurbuz, Oğul Can, and Eli Waxman. Minibatch Optimal Transport and Perplexity Bound Estimation in Discrete Flow Matching. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=A8rmJlSET9.

Johannes Hertrich, Antonin Chambolle, and Julie Delon. On the Relation between Rectified Flows and Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 172578–172610. Curran Associates, Inc., 2025. doi: 10.52202/085713-5743. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/fb9f946a34d5b5dfc582f9e7b5277bbf-Abstract-Conference.html.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a Better Evaluation Metric for Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9307–9315, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Jayasumana_Rethinking_FID_Towards_a_Better_Evaluation_Metric_for_Image_Generation_CVPR_2024_paper.html.

Kacper Kapuśniak, Peter Potaptchik, Teodora Reu, Leo Zhang, Alexander Tong, Michael Bronstein, Avishek Joey Bose, and Francesco Di Giovanni. Metric Flow Matching for Smooth Interpolations on the Data Manifold. In *Advances in Neural Information Processing Systems*, volume 37, pp. 135011–135042. Curran Associates, Inc., 2024. doi: 10.52202/079017-4291. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f381114cf5aba4e45552869863deaaa7-Abstract-Conference.html.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009. URL https://cave.cs.toronto.edu/kriz/learning-features-2009-TR.pdf.

Shivam Kumar, Yixin Wang, and Lizhen Lin. Flow Matching is Adaptive to Manifold Structures. *arXiv preprint arXiv:2602.22486*, 2026. URL https://arxiv.org/abs/2602.22486.

Lea Kunkel and Mathias Trabs. On the minimax optimality of Flow Matching through the connection to kernel density estimation. *arXiv preprint arXiv:2504.13336*, 2025. URL https://arxiv.org/abs/2504.13336.

Gunther Leobacher and Alexander Steinicke. Existence, uniqueness and regularity of the projection onto differentiable manifolds. *Annals of Global Analysis and Geometry*, 60(3):559–587, 2021. doi: 10.1007/s10455-021-09788-z. URL https://doi.org/10.1007/s10455-021-09788-z.

Elizaveta Levina and Peter J. Bickel. Maximum Likelihood Estimation of Intrinsic Dimension. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/hash/74934548253bcab8490ebd74afed7031-Abstract.html.

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. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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. URL https://openreview.net/forum?id=XVjTT1nw5z.

Gabriel Loaiza-Ganem, Brendan Leigh Ross, Rasa Hosseinzadeh, Anthony L. Caterini, and Jesse C. Cresswell. Deep Generative Models through the Lens of the Manifold Hypothesis: A Survey and New Connections. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a90WpmSi0I.

Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/5acdc9ca5d99ae66afdfe1eea0e3b26b-Abstract.html.

Alireza Mousavi-Hosseini, Stephen Zhang, Michal Klein, and Marco Cuturi. Flow Matching with Semidiscrete Couplings. In *International Conference on Learning Representations*, pp. 95449–95486, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9a263e235f6d1521d13a8531c7974951-Abstract-Conference.html.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the Homology of Submanifolds with High Confidence from Random Samples. *Discrete & Computational Geometry*, 39(1-3): 419–441, 2008. doi: 10.1007/s00454-008-9053-2. URL https://doi.org/10.1007/s00454-008-9053-2.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

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 *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28100–28127. PMLR, 2023. URL https://proceedings.mlr.press/v202/pooladian23a.html.

Neta Shaul, Ricky T. Q. Chen, Maximilian Nickel, Matthew Le, and Yaron Lipman. On Kinetic Optimal Probability Paths for Generative Models. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30883–30907. PMLR, 2023. URL https://proceedings.mlr.press/v202/shaul23a.html.

Amit Singer and Hau-tieng Wu. Vector diffusion maps and the connection Laplacian. *Communications on Pure and Applied Mathematics*, 65(8):1067–1144, 2012. doi: 10.1002/cpa.21395. URL https://doi.org/10.1002/cpa.21395.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW.

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

Olga Zaghen, Floor Eijkelboom, Alison Pouplin, Cong Liu, Max Welling, Jan-Willem van de Meent, and Erik J. Bekkers. Riemannian Variational Flow Matching for Material and Protein Design. In *International Conference on Learning Representations*, pp. 34243–34286, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/39c5871aa13be86ab978cba7069cbcec-Abstract-Conference.html.

Zhengyu Zhou and Weiwei Liu. An Error Analysis of Flow Matching for Deep Generative Modeling. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78903–78932. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhou25l.html.

A NOTATION AND SCOPE

Symbol	Meaning
$\mathcal{M}, \tau$	Embedded manifold and its reach
$\pi(x), P_p$	Nearest-point projection and tangent orthogonal projector
$\zeta, (x_0, x_1), \Delta$	Endpoint coupling, sampled pair, and $x_1 - x_0$
c_t, p_t, γ_t	Ambient chord, projected chord, and same-pair intrinsic geodesic
a_t, b_t, h_t	Tangent chord component, projected-path velocity, reprojected geodesic velocity
$G_t, g(t, \theta)$	Raw pairwise discrepancy and intrinsic-speed-normalized sphere discrepancy
u_ζ, v_θ	Conditional regression field and learned field
$B_{1/2}, D, e$	Exact common-OT midpoint field gap, learned total gap, and projected learning residual
$F, \widehat{P}_t, U_t^\epsilon$	Image feature map, local PCA projector, and finite-difference feature velocity
$\widehat{G}_t, S_T$	Endpoint-conditioned feature diagnostic and endpoint-free prefix score

The use of "geodesic" in the controlled results refers to the shortest Riemannian path for a specified pair. A distribution-level Wasserstein statement additionally requires optimality of the endpoint coupling. Shared ambient/intrinsic optimality is proved for the circle arcs, spherical-band product laws, coordinatewise product-torus transports, and ellipsoid rings. It is not presumed for the preserved sphere sweep. Image scores are empirical feature diagnostics and are not substituted for geometric quantities in any theorem.

B PROOFS AND ANALYTIC CHECKS

B.1 REACH-CONTROLLED PROJECTION DEFECT

Proposition 4 (Reach-controlled projection defect)**.** *Suppose* $\operatorname{dist}(x, \mathcal{M}) \le \rho < \tau$. *Then*

$$\left\|D\pi(x) - P_{\pi(x)}\right\|_{\mathrm{op}} \le \frac{\rho}{\tau - \rho}, \quad \left\|D\pi(x)v - P_{\pi(x)}v\right\| \le \frac{\rho}{\tau - \rho}\|v\|. \tag{11}$$

The coefficient is attained on an inward normal offset of a sphere by a nonzero tangent vector.

The standard reach inequality bounds the second fundamental form: for a unit normal ν and a unit tangent u, $|\langle \nu, \mathrm{II}(u,u)\rangle| \le 1/\tau$. One way to see this is to expand the inequality

$$|\langle \nu, q - p\rangle| \le \frac{\|q-p\|^2}{2\tau}, \qquad p, q \in \mathcal{M},$$

along a smooth manifold curve through p with initial velocity u (Federer, 1959; Aamari et al., 2019). If $x = p + n$ is in the tube, the shape operator S_n therefore satisfies $\|S_n\|_{\mathrm{op}} \le \|n\|/\tau \le \rho/\tau$. With the sign convention matching Leobacher & Steinicke (2021), the projection differential is

$$D\pi(x) = (I - S_n)^{-1} P_p.$$

Here I acts on $T_p\mathcal{M}$. It follows that

$$\begin{aligned}\|D\pi(x) - P_p\|_{\mathrm{op}} &= \left\|\{(I - S_n)^{-1} - I\}P_p\right\|_{\mathrm{op}} \\ &\le \frac{\|S_n\|_{\mathrm{op}}}{1 - \|S_n\|_{\mathrm{op}}} \le \frac{\rho}{\tau - \rho}.\end{aligned}$$

For sharpness, take a sphere of radius $r = \tau$, $x = (r - \rho)e_1$, and $v = e_2$. Radial projection gives $D\pi(x)v = r/(r-\rho)v$ and $P_{re_1}v = v$, so equality holds. This argument compares two projection operations; no statistical estimation or optimal-coupling assumption enters it.

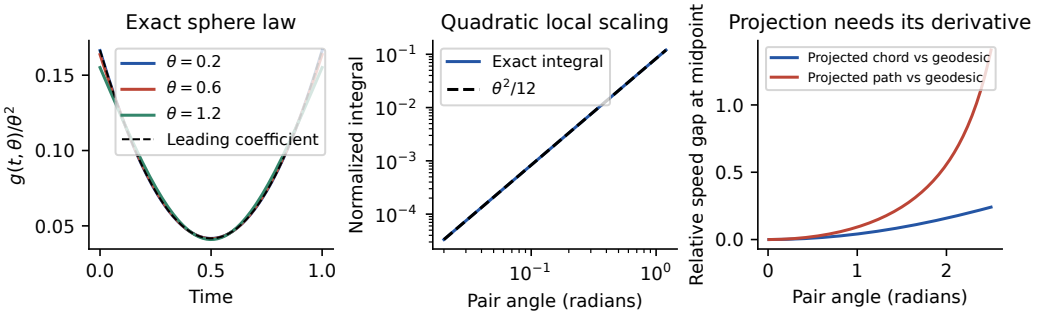

Figure 5: Exact sphere identities. Left: the normalized time profile approaches the quadratic coefficient. Middle: integration gives $\theta^2/12$ for short pairs. Right: tangent-projecting a chord velocity and differentiating the projected chord give different midpoint speed errors, despite traversing the same great circle.

B.2 Exact sphere law

Theorem 5 (Round-sphere comparison)**.** *Let $x_0, x_1 \in \mathbb{S}^m(r)$ have intrinsic distance $d = r\theta$, $0 < \theta < \pi$. Put*

$$q_t = \sqrt{1 - 2t(1-t)(1-\cos\theta)}, \quad \phi_t = \operatorname{atan2}(t\sin\theta, 1 - t + t\cos\theta). \tag{12}$$

For the unit tangent e_t along the connecting great circle at p_t, the velocities in Eq. (1) *satisfy*

$$a_t = \frac{r\sin\theta}{q_t} e_t, \qquad b_t = \frac{r\sin\theta}{q_t^2} e_t, \qquad h_t = r\theta\cos(\phi_t - t\theta)e_t. \tag{13}$$

Consequently, the intrinsic-speed-normalized discrepancy is exactly

$$g(t,\theta) = \frac{G_t}{r\theta} = \left|\frac{\sin\theta}{\theta q_t} - \cos(\phi_t - t\theta)\right|, \qquad \int_0^1 g(t,\theta)\,dt = \frac{\theta^2}{12} + O(\theta^4). \tag{14}$$

The remainder is uniform over time for uniformly short pairs. Thus, for couplings supported on $d/r \le \varepsilon \to 0$,

$$\mathbb{E}_\zeta \int_0^1 g(t, d/r)dt = \frac{\mathbb{E}_\zeta[d^2]}{12r^2} + O\left(\frac{\mathbb{E}_\zeta[d^4]}{r^4}\right). \tag{15}$$

The leading normalized time profile is $\theta^2[1/6 - t(1-t)/2]$. Here reach equals r, while transported-distance moments depend on the endpoint laws. Changing concentration on a fixed sphere changes neither reach nor diameter. The empirical normalization by $\|h_t\|$ has the same leading law. By rotational symmetry, write $x_0 = re_1$ and $x_1 = r(\cos\theta\, e_1 + \sin\theta\, e_2)$. Set

$$A = 1 - t + t\cos\theta, \quad B = t\sin\theta, \quad q = \sqrt{A^2 + B^2}, \quad \phi = \operatorname{atan2}(B, A).$$

Then $p_t = r(\cos\phi\, e_1 + \sin\phi\, e_2)$ and the unit forward tangent is $e_t = -\sin\phi\, e_1 + \cos\phi\, e_2$. Direct calculation gives

$$\begin{aligned}\langle x_1 - x_0, e_t\rangle &= r\{(1-\cos\theta)\sin\phi + \sin\theta\cos\phi\} \\ &= r\frac{(1-\cos\theta)B + \sin\theta A}{q} = \frac{r\sin\theta}{q}.\end{aligned}$$

Both vectors lie in the plane spanned by the endpoints, establishing the formula for a_t. On the sphere, $D\pi(c_t) = P_{p_t}/q$, yielding b_t. The constant-speed geodesic is

$$\gamma_t = r(\cos(t\theta)e_1 + \sin(t\theta)e_2), \qquad \dot\gamma_t = r\theta(-\sin(t\theta)e_1 + \cos(t\theta)e_2),$$

so $\langle\dot\gamma_t, e_t\rangle = r\theta\cos(\phi - t\theta)$. Subtracting the two scalar coefficients proves the exact $g(t,\theta)$ formula.

For $\theta \to 0$, Taylor expansion uniformly on $t \in [0,1]$ gives

$$q^{-1} = 1 + \tfrac{1}{2}t(1-t)\theta^2 + O(\theta^4),$$
$$\frac{\sin\theta}{\theta} = 1 - \tfrac{1}{6}\theta^2 + O(\theta^4),$$
$$\phi - t\theta = \tfrac{1}{6}t(1-t)(2t-1)\theta^3 + O(\theta^5).$$

Hence $\cos(\phi - t\theta) = 1 + O(\theta^6)$ and

$$g(t,\theta) = \theta^2\left(\frac{1}{6} - \frac{t(1-t)}{2}\right) + O(\theta^4).$$

The leading coefficient is at least $1/24$, so the absolute value causes no sign ambiguity for sufficiently small θ. Its integral is $1/12$. The expressions extend smoothly to $\theta = 0$ and are analytic on a compact neighborhood of $[0,1] \times \{0\}$; the remainder is therefore uniform for sufficiently short pairs. Averaging the remainder yields $O(\mathbb{E}[\theta^4])$, rather than $O((\mathbb{E}[\theta])^4)$, and proves Eq. (15).

The released empirical normalization is $\|a_t - h_t\| \,/\, \|h_t\|$. Its exact scalar form is

$$\bar{g}(t,\theta) = \frac{g(t,\theta)}{\cos(\phi - t\theta)}.$$

For $0 < \theta < \pi$, the cosine is positive: on either half of the interval, the deviation of ϕ from $t\theta$ is smaller than $\theta/2 < \pi/2$. This normalization shares the same leading quadratic law but is not the same exact finite-angle quantity.

Projected path speed. At the midpoint, the tangent-projected chord speed is $2r\sin(\theta/2)$, whereas the projected-path speed is $2r\tan(\theta/2)$. The intrinsic speed is $r\theta$. Their normalized midpoint gaps are respectively $1 - 2\sin(\theta/2)/\theta$ and $2\tan(\theta/2)/\theta - 1$. This accounts for the two distinct curves in Figure 5; the difference is a chain-rule effect, not a different geometric path.

B.3 Local directional curvature beyond spheres

The sphere coefficient is a special case of a local expansion. This uses ordinary Taylor expansion and the Weingarten identity; it does not assume an optimal coupling or identify an empirical image score with a geometric quantity.

Proposition 6 (Local midpoint curvature term)**.** *Let $\mathcal{M} \subset \mathbb{R}^D$ be a C^5 embedded manifold of positive reach. Let γ be a unit-speed geodesic with $m = \gamma(0)$ and $\xi = \dot{\gamma}(0)$. For sufficiently small $d > 0$, set*

$$x_- = \gamma(-d/2), \quad x_+ = \gamma(d/2), \quad c_d = (x_- + x_+)/2, \quad p_d = \pi(c_d).$$

Define the shape operator by $\langle S_n v, w\rangle = \langle n, \mathrm{II}_m(v,w)\rangle$. Then, as vectors in the ambient embedding,

$$P_{p_d}(x_+ - x_- - d\xi) = -\frac{d^3}{24} S_{\mathrm{II}_m(\xi,\xi)}\xi + O(d^5). \tag{16}$$

Consequently, with $K_m(\xi) = \left\|S_{\mathrm{II}_m(\xi,\xi)}\xi\right\|$,

$$G_{1/2} = \frac{d^3}{24}K_m(\xi) + O(d^5), \qquad \|\mathrm{II}_m(\xi,\xi)\|^2 \le K_m(\xi) \le \tau^{-2}. \tag{17}$$

Remainders are uniform on compact families with bounded geodesic derivatives through order five and a common projection neighborhood. Their constants are not asserted to depend on reach alone.

Here the unit-speed geodesic is reparameterized to the unit time interval, so its midpoint velocity is $d\xi$. Thus the left side of Eq. (16) is exactly the projected same-pair difference defined in Section 3. The proposition does not compare fields under arbitrary ambient and intrinsic transport laws.

Proof. Symmetric Taylor expansions give

$$x_+ - x_- = d\xi + \frac{d^3}{24}\gamma'''(0) + O(d^5), \qquad c_d = m + \frac{d^2}{8}\mathrm{II}_m(\xi,\xi) + O(d^4).$$

The truncated point $m + (d^2/8)\mathrm{II}_m(\xi,\xi)$ is a normal offset whose nearest point is exactly m for sufficiently small d. The actual chord midpoint differs from this truncated point by $O(d^4)$. Local Lipschitz continuity of π therefore yields $p_d - m = O(d^4)$, and smoothness of the tangent projector gives $P_{p_d} - P_m = O(d^4)$. Along a geodesic, $\gamma'' = \mathrm{II}(\dot\gamma,\dot\gamma)$ is normal. Differentiating its orthogonality to any parallel tangent field V gives

$$\langle \gamma'''(0), V(0)\rangle = -\langle \mathrm{II}_m(\xi,\xi), \mathrm{II}_m(\xi, V(0))\rangle.$$

Hence $P_m\gamma'''(0) = -S_{\mathrm{II}_m(\xi,\xi)}\xi$. Projecting the first Taylor expansion proves Eq. (16); replacing P_{p_d} by P_m contributes only $O(d^7)$. The reverse triangle inequality proves the scalar expansion even if its leading coefficient vanishes.

Since $\|\xi\| = 1$,

$$\langle S_{\mathrm{II}(\xi,\xi)}\xi, \xi\rangle = \|\mathrm{II}(\xi,\xi)\|^2 \le K_m(\xi).$$

The reach bounds in Appendix B.1 give $\|\mathrm{II}(\xi,\xi)\| \le \tau^{-1}$ and $\|S_n\|_{\mathrm{op}} \le \|n\|/\tau$, establishing the upper bound.

Moment form and scope. For any pair law ζ supported on $0 \le d \le d_0$ in a family with a uniform remainder bound Cd^5, Minkowski's inequality gives

$$\left| (\mathbb{E}_\zeta G_{1/2}^2)^{1/2} - \frac{1}{24}\sqrt{\mathbb{E}_\zeta[d^6 K_m(\xi)^2]} \right| \le C\sqrt{\mathbb{E}_\zeta[d^{10}]}. \tag{18}$$

The bounded-distance and uniform-geometry assumptions ensure the required moments are finite. On a round sphere, $K_m(\xi) = r^{-2}$, recovering the leading sixth-moment term of the common-OT theorem. On a general manifold, its coefficient depends on the transported directions and local second fundamental form. A nonzero third-order lower term requires nonzero directional normal curvature; reach alone supplies no such lower bound. The expansion remains a local same-pair statement. The common-OT theorems separately verify the conditions needed for an Eulerian midpoint-field interpretation.

Independent numerical checks. The script `verify_local_curvature.py` compares actual endpoint differences and nearest-point tangent projections with Eq. (16). Thirteen cases cover round spheres, ellipses, helices, product tori in $\mathbb{R}^4$, and ring tori in $\mathbb{R}^3$. Ellipse arc-length parameters and ring-torus geodesics are independently integrated with DOP853; other paths are explicit. Pair distances are $0.4, 0.2, 0.1, 0.05, 0.025$. Across the final three distances, vector-remainder slopes range from 4.9998 to 5.0035. At the shortest distance, the largest relative coefficient error among nonzero cases is below 0.102%. A ring-torus asymptotic direction has a vanishing third-order coefficient and a measured fifth-order gap. The product torus also illustrates why this is an *extrinsic* directional coefficient: its intrinsic metric is flat, yet its nonzero embedding curvature produces a gap. These numerical cases check the expansion; the proof establishes its stated scope.

B.4 Conditional-field bridge

Proposition 7 (Conditional-field bridge)**.** *Fix a time and a coupling ζ for which c_t lies in the reach tube almost surely. Suppose a specified tangent field w_t satisfies $\dot\gamma_t = w_t(\gamma_t)$ almost surely, is defined at both p_t and γ_t, and is L_t-Lipschitz in ambient distance on these points. For square-integrable quantities,*

$$\|P_{p_t}u_\zeta(t,c_t) - w_t(p_t)\|_{L^2(\zeta)} \le (\mathbb{E}_\zeta G_t^2)^{1/2} + L_t(\mathbb{E}_\zeta \|\gamma_t - p_t\|^2)^{1/2}. \tag{19}$$

Write $P = P_{p_t}$ and condition on c_t. Because p_t is a function of c_t and $w_t(p_t)$ is tangent, $Pw_t(p_t) = w_t(p_t)$. Under the proposition's deterministic-field assumption,

$$\begin{aligned} Pu_\zeta(t,c_t) - w_t(p_t) = \mathbb{E}[P(\Delta - \dot\gamma_t) \mid c_t] \\ + \mathbb{E}[P\{w_t(\gamma_t) - w_t(p_t)\} \mid c_t]. \end{aligned}$$

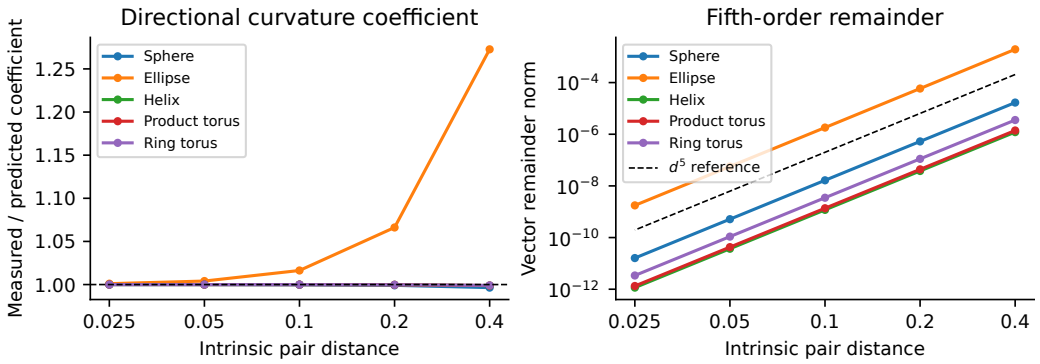

Figure 6: Local midpoint expansion on five geometry families. Left: the measured coefficient approaches the directional shape-operator prediction. Right: vector errors decay at fifth order in intrinsic pair distance. Each plotted curve is one specified case; all thirteen cases, including a vanishing leading coefficient, are released.

Conditional Jensen bounds the first term's L^2 norm by $(\mathbb{E}G_t^2)^{1/2}$. Contractivity and the Lipschitz assumption bound the second by $L_t(\mathbb{E}\,\|\gamma_t - p_t\|^2)^{1/2}$. The triangle inequality yields Eq. (19). On the sphere, the two positions have angular difference $\phi_t - t\theta$, so their Euclidean distance is $2r\sin(|\phi_t - t\theta|/2)$. The expansion in Appendix B.2 establishes the stated uniform order. The argument neither changes the endpoint coupling nor supplies a regular extension of an almost-everywhere intrinsic field; these are assumptions to verify in each application.

B.5 Common optimal coupling and midpoint field law

Choose an unwrapped interval I of length $\ell < \pi/2$ containing both angular supports. The two costs in angular coordinates are

$$c_E(\alpha,\beta) = 2r^2\{1-\cos(\beta-\alpha)\}, \qquad c_M(\alpha,\beta) = r^2(\beta-\alpha)^2.$$

There is no ambiguity from a wrapped shortest path. On $I \times I$, their mixed derivatives are $-2r^2\cos(\beta-\alpha) < 0$ and $-2r^2 < 0$. Thus uncrossing any two ordered source angles and oppositely ordered target angles strictly decreases either cost: integrate the mixed derivative over the rectangle determined by the four angles. The increasing rearrangement is optimal for each cost, by the standard one-dimensional rearrangement argument (Villani, 2009). Absolute continuity of the source identifies the same transport map $T = F_1^{-1} \circ F_0$ almost everywhere. In this theorem it is with respect to arc length on the circle; the next result separately treats full-dimensional spherical-band laws.

Write $e(\beta) = (-\sin\beta, \cos\beta)$, $\delta = T(\alpha) - \alpha$, and $\beta = (\alpha + T(\alpha))/2$. Then

$$c_{1/2} = r\cos(\delta/2)(\cos\beta, \sin\beta), \quad p = r(\cos\beta, \sin\beta) = \gamma_{1/2}.$$

Because $|\delta| < \pi/2$, the radial factor is positive and the nearest-point projection is unique. The map $\alpha \mapsto \beta$ is strictly increasing: T is nondecreasing and the identity map is strictly increasing. Hence both $c_{1/2}$ and p identify their generating pair. The ambient conditional velocity is therefore $u(1/2, c_{1/2}) = x_1 - x_0$, and the intrinsic displacement field is $w(1/2, p) = r\delta e(\beta)$ almost everywhere under its own midpoint law. Moreover,

$$P_p u(1/2, c_{1/2}) = 2r\sin(\delta/2)e(\beta).$$

The projected ambient midpoint law and the intrinsic midpoint law coincide. Their fields are compared at the same intrinsic base point, without assigning an arbitrary extension outside the support. Taking the squared $L^2(\zeta)$ norm proves Eq. (3).

For $0 \le z \le \pi/2$, alternating Taylor bounds for $\sin(z/2)$ give

$$\frac{z^3}{24}\left(1 - \frac{z^2}{80}\right) \le z - 2\sin(z/2) \le \frac{z^3}{24}.$$

Apply this inequality to $z = |\delta| \le \varepsilon$, take the L^2 norm, and substitute $d = r|\delta|$ to obtain Eq. (4). The result also covers zero displacements. In particular, for uniformly short pairs,

$$B_{1/2} = \frac{\sqrt{\mathbb{E}d^6}}{24\tau^2}\{1 + O(\varepsilon^2)\}, \qquad \tau = r.$$

No learning-error assumption, local-PCA approximation, or unknown intrinsic-field extension is used in this geometric value. The learned-field comparison $|D - B_{1/2}| \le e$ is the reverse triangle inequality for the explicitly defined projected L^2 quantities. It does not say that a finite-error learner must have discrepancy at least $B_{1/2}$: its residual can partially cancel the geometric difference.

B.6 Full-dimensional spherical-band extension

In the coordinates of Theorem 2, surface area has element $r^m \sin^{m-1} a \, da \, d\sigma_{m-1}(\omega)$. Since $I \subset (0, \pi)$, product measures $\nu_i \otimes \lambda$ with the stated absolute continuity push forward to surface-absolutely-continuous probability laws. They need not have positive density outside their bands or equal the uniform surface law inside a band.

For arbitrary directions ω, ω', $\sin a \sin b > 0$ gives

$$\begin{aligned}\|X(a,\omega) - X(b,\omega')\|^2 &= 2r^2\{1 - \cos(b-a)\} + 2r^2 \sin a \sin b(1 - \langle \omega, \omega' \rangle) \\ &\ge 2r^2\{1 - \cos(b-a)\}.\end{aligned}$$

Likewise, the inner product is at most $r^2 \cos(a-b)$, so the spherical distance is at least $r|a-b|$. Equality in both bounds holds for $\omega' = \omega$, using the shortest constant-direction meridian. Every coupling of the full laws induces a coupling of the polar marginals, so its expected cost is bounded below by the corresponding one-dimensional OT value. By the preceding strict-Monge argument the same increasing T minimizes both polar costs. The product assumption with common λ makes $(a, \omega) \mapsto (T(a), \omega)$ feasible for the full laws and it attains both lower bounds. This proves common optimality without positing it as a separate assumption.

For $\delta = T(a) - a$ and $\eta = (a + T(a))/2$,

$$c_{1/2} = \cos(\delta/2) X(\eta, \omega), \qquad p = \pi(c_{1/2}) = X(\eta, \omega) = \gamma_{1/2}.$$

The projected and intrinsic midpoint laws therefore coincide. The strictly increasing map $a \mapsto \eta$ is injective, and $\eta \in (0, \pi)$ identifies ω from the nonzero first m coordinates of p. Thus either midpoint identifies its pair, and both displacement fields are determined under the appropriate law. For the unit meridian tangent $e = (\cos\eta\, \omega, -\sin\eta)$ their velocities are $P_p u = 2r \sin(\delta/2) e$ and $w = r\delta e$. The exact L^2 formula and the same sine bounds now give Eqs. (3)–(4). The proof permits an arbitrary absolutely continuous common directional law; rotational symmetry is not required, but its independence from the polar angle is.

B.7 Actual projected-path action and all-time identification

We prove Corollary 3. Work first in the common unwrapped circle interval from Theorem 1. For two endpoint angles α, β, the projected angle is

$$f_t(\alpha, \beta) = \arg\{(1-t)e^{i\alpha} + te^{i\beta}\}, \qquad Q_t^2 = 1 - 2t(1-t)(1 - \cos(\beta - \alpha)).$$

The argument has an unambiguous branch in this interval. Direct differentiation gives

$$\partial_\alpha f_t = \frac{(1-t)((1-t) + t\cos(\beta-\alpha))}{Q_t^2}, \qquad \partial_\beta f_t = \frac{t(t + (1-t)\cos(\beta-\alpha))}{Q_t^2}. \tag{20}$$

Both derivatives are strictly positive for $0 < t < 1$ because $|\beta - \alpha| < \pi/2$. As T is increasing, $\alpha \mapsto f_t(\alpha, T(\alpha))$ is strictly increasing. This argument uses coordinatewise monotonicity, not differentiability of T. Thus p_t identifies α and its paired endpoint at every interior time. The chord point c_t also identifies the pair because its projection does. Consequently $u(t, c_t) = \Delta$ and the conditional velocity of $\bar{\mu}_t = (\pi \circ c_t)_\# \zeta$ is

$$b_t(p_t) = \mathbb{E}[D\pi(c_t)\Delta \mid p_t] = D\pi(c_t)\Delta = D\pi(c_t)u(t, c_t).$$

The paths are smooth in time with uniformly bounded speed under the short-interval condition. Their pushforward laws and conditional velocity satisfy the continuity equation, and their Eulerian action equals the expected paired action. For the spherical-band class, projection preserves ω; the polar angle obeys the same formula and stays in $(0, \pi)$. Hence $(f_t(a, T(a)), \omega)$ identifies the full pair, proving the same field conclusion. Equality with the intrinsic intermediate law is neither asserted nor needed away from the midpoint.

For a signed displacement δ, $\dot{f}_t = \sin\delta/Q_t^2$. Put $h = \delta/2$, $A = \cos h$, and $B = \sin h$. By symmetry and $v = 2t - 1$,

$$\int_0^1 \|b_t(p_t)\|^2 \, dt = 4r^2A^2B^2 \int_0^1 \frac{dv}{(A^2 + B^2v^2)^2}$$
$$= 2r^2B^2 + 2r^2hB/A = r^2\{1 - \cos\delta + \delta\tan(\delta/2)\}.$$

The formula extends continuously to $\delta = 0$ and is even in δ. Common optimality gives $W^2_{2,\mathcal{M}}(\mu_0, \mu_1) = r^2\mathbb{E}[\delta^2]$, attained by the constant-speed intrinsic displacement path. For each pair, Cauchy–Schwarz gives $\int_0^1 \dot{f}_t^2 dt \geq (\int_0^1 \dot{f}_t dt)^2 = \delta^2$. For $\delta \neq 0$, Q_t varies with time, so the speed is nonconstant and the inequality is strict. Averaging proves the exact identity and positivity in Eq. (5).

Finally,

$$1 - \cos\delta + \delta\tan(\delta/2) - \delta^2 = \delta^6/180 + \delta^8/2520 + O(\delta^{10}),$$
$$(\delta - 2\sin(\delta/2))^2 = \delta^6/576 - \delta^8/23040 + O(\delta^{10}).$$

Uniform Taylor remainders on a fixed compact subinterval of $(-\pi/2, \pi/2)$ yield

$$\mathcal{A}_{\mathrm{proj}} - W^2_{2,\mathcal{M}} = \frac{\mathbb{E}[d^6]}{180r^4} + O(\mathbb{E}[d^8]/r^6) = \frac{16}{5}B^2_{1/2} + O(r^2\mathbb{E}[\delta^8]).$$

The next coefficient in the last difference is $3r^2\mathbb{E}[\delta^8]/5600$. This is a consequence for intrinsic action, not a lower bound on endpoint or numerical sampling error. Normalized linear interpolation's nonuniform speed is familiar; the role here is to connect its action to the identified conditional-field gap under the proved shared coupling. At $r = 1, \delta = 0.6$, for example, the relative action excess is below 0.08%, so short-angle effects should not be presented as necessarily large.

The script `verify_intrinsic_action.py` evaluates the Cartesian derivative of radial projection and performs separate 80-digit numerical quadrature in 18 radius/signed-angle cases. It checks the exact integral and three decreasing-angle coefficient limits. The numerical checks support the implementation of the formulas; the arguments above establish the identities for the continuous laws.

B.8 Several moving coordinates with correlated source densities

Theorem 8 (Coordinatewise common OT on a product torus)**.** *Let $\mathcal{M} = \prod_{j=1}^k \mathbb{S}^1(r_j) \subset \mathbb{R}^{2k}$, $k \geq 2$. In angular coordinates, let μ_0 be absolutely continuous, and let $T(\alpha) = (T_1(\alpha_1), \ldots, T_k(\alpha_k))$, where each coordinate map is C^1 with positive derivative on its source interval. Set $\mu_1 = T_\# \mu_0$. Suppose the source and target supports in coordinate j lie in a common unwrapped interval I_j of length less than $\pi/2$, and the paired chords stay within the reach tube. No independence of the source coordinates is required. Then $(\mathrm{id}, T)_\#\mu_0$ is optimal for both squared ambient and squared intrinsic distance. Its projected and intrinsic midpoint laws coincide and identify the endpoint pair. For $\delta_j = T_j(\alpha_j) - \alpha_j$ and $d_j = r_j|\delta_j|$,*

$$B^2_{1/2} = \mathbb{E}\sum_{j=1}^k r_j^2(\delta_j - 2\sin(\delta_j/2))^2 = \frac{1}{576}\mathbb{E}\sum_{j=1}^k \frac{d_j^6}{r_j^4} + O\left(\mathbb{E}\sum_{j=1}^k \frac{d_j^8}{r_j^6}\right) \tag{21}$$

for uniformly short coordinate displacements, where $B_{1/2}$ is defined by the same tangent-component field comparison as in Eq. (3)*.*

Proof. The two costs split into sums of coordinate costs,

$$c_E(\alpha, \beta) = \sum_j 2r_j^2(1 - \cos(\beta_j - \alpha_j)), \qquad c_{\mathcal{M}}(\alpha, \beta) = \sum_j r_j^2(\beta_j - \alpha_j)^2.$$

The second identity follows from the product metric and the unambiguous shorter coordinate arcs. For every coupling of the joint laws, each coordinate marginal is a feasible coupling of the corresponding marginal laws. Its total cost is therefore at least the sum of the marginal optimal costs. Each increasing T_j attains both marginal minima by the short-interval argument of Theorem 1. The joint map T is feasible by the definition of μ_1 and simultaneously attains these lower bounds, proving common optimality even when μ_0 is correlated.

Nearest projection onto the embedded product torus acts radially in each coordinate plane; all chord blocks are nonzero. At the midpoint, coordinate j has angle $(\alpha_j + T_j(\alpha_j))/2$, which is both its intrinsic midpoint and an injective function of α_j. The full midpoint thus identifies the endpoint pair. Conditional averaging introduces no variance, and the orthogonal coordinate blocks have respective tangent velocity components $2r_j \sin(\delta_j/2)$ and $r_j\delta_j$. Summing their squared differences and averaging gives the exact formula. The uniform sine expansion gives Eq. (21). □

The reach is $\min_j r_j$. A sufficient tube condition is $\sum_j r_j^2(1-\cos(\delta_j/2))^2 < (\min_j r_j)^2$ for each pair, since the maximum chord distance from the product torus occurs at the midpoint. It holds for uniformly small displacements with fixed radii and dimension. Coordinatewise all-time identification also follows from Eq. (20). Hence the projected-action identity adds across coordinates:

$$\mathcal{A}_{\mathrm{proj}} - W_{2,\mathcal{M}}^2 = \mathbb{E}\sum_j r_j^2[1-\cos\delta_j + \delta_j\tan(\delta_j/2) - \delta_j^2].$$

The construction permits several moving coordinates and nonproduct densities, but its map and costs are separable. Source and target dependence structures cannot be chosen arbitrarily and independently. It establishes no unrestricted coupled OT result. The intrinsically flat product torus can have a positive gap because its embedding is curved; a single global reach does not replace the directional weighted coefficient above.

Numerical checks use $k = 2, 3, 4$, two seeds, and 96 samples per case. Source angles are $0.22\tanh(\sqrt{0.65}z_0 + \sqrt{0.35}z_j)$ with independent standard normal inputs. Coordinate maps have positive derivative and alternating signed shifts. Their common support intervals are shorter than $\pi/2$. Independent linear assignments under both costs recover the prescribed pairing, and direct Cartesian tangent-velocity calculations agree with Eq. (21) within 6.3×10^{-19} squared-gap error. These finite examples check the formulas rather than proving joint optimality; the marginal lower-bound argument does that.

B.9 Within-tube ellipsoid construction

Theorem 9 (A within-tube ellipsoid example)**.** *For $a \geq 1$, let $\mathcal{M}_a = \{(x,y,z) : x^2+y^2+z^2/a^2 = 1\}$, with reach $\tau_a = 1/a$ and ambient diameter $2a$. Set $s_a = 1 - 1/(4a)$, $c_a = \sqrt{1-s_a^2}$, and $\theta_a = \arcsin s_a$. Couple the two uniform latitude rings at heights $\pm ac_a$ by equal azimuth. This coupling is optimal for both ambient squared distance and squared intrinsic distance. Every chord remains within distance $\tau_a/4$ of $\mathcal{M}_a$. At $t = 1/2$, $p_t = \gamma_t$ and the projected-chord versus intrinsic-velocity gap is*

$$G_a = \int_{\theta_a}^{\pi-\theta_a} \sqrt{\cos^2 u + a^2\sin^2 u}\, du - 2ac_a \sim \frac{1}{3\,2^{3/2}} a^{-5/2}. \tag{22}$$

Hence $G_a \to 0$ while $\mathrm{diam}_E(\mathcal{M}_a)/\tau_a = 2a^2 \to \infty$.

Parameterize $\mathcal{M}_a$ by

$$X(u,\varphi) = (\sin u\cos\varphi, \sin u\sin\varphi, a\cos u), \qquad ds^2 = (\cos^2 u + a^2\sin^2 u)du^2 + \sin^2 u\, d\varphi^2.$$

Its reach is $1/a$. For completeness, write $A = \mathrm{diag}(1,1,a^{-2})$. For $p, q \in \mathcal{M}_a$, the equality $p^TAp = q^TAq = 1$ gives

$$\left|\left\langle \frac{Ap}{\|Ap\|}, q-p\right\rangle\right| = \frac{(q-p)^TA(q-p)}{2\,\|Ap\|} \leq \frac{a}{2}\,\|q-p\|^2,$$

because $\|A\|_{\mathrm{op}} = 1$ and $\min_{p\in\mathcal{M}_a}\|Ap\| = 1/a$. The reach characterization by tangent-distance quotients (Aamari et al., 2019) implies reach at least $1/a$. The principal curvature at a pole is a, so reach is at most $1/a$. The ambient diameter is $2a$.

Optimality of the same-azimuth coupling. The endpoint rings are $X(\theta_a, \varphi)$ and $X(\pi - \theta_a, \varphi)$, with φ uniform. The squared ambient distance between arbitrary ring endpoints is

$$4a^2c_a^2 + 2s_a^2\{1 - \cos(\varphi_1 - \varphi_0)\},$$

which is minimized by matching azimuths. Any intrinsic curve between the rings has length at least

$$L_a = \int_{\theta_a}^{\pi-\theta_a} \sqrt{\cos^2 u + a^2 \sin^2 u}\, du,$$

since its line element dominates $\sqrt{\cos^2 u + a^2 \sin^2 u}|du|$. The monotone constant-azimuth meridian attains this bound. Equal-azimuth matching preserves the uniform marginal on each ring and attains the pointwise minimum under either cost, proving optimality for both squared-distance transport problems.

Tube confinement and the midpoint. A chord point has cylindrical radius s_a and height $z \in [-ac_a, ac_a]$. The same-height surface point has radius $\sqrt{1 - z^2/a^2} \in [s_a, 1]$. Thus

$$\operatorname{dist}(c_t, \mathcal{M}_a) \le 1 - s_a = \frac{1}{4a} = \frac{\tau_a}{4} < \frac{\tau_a}{2}$$

for all t, giving unique nearest projection throughout the chord. At $z = 0$, the nearest point is the equator point with the same azimuth. Indeed, after matching azimuth, squared distance to a surface point of cylindrical radius $r \in [0, 1]$ is

$$(r - s_a)^2 + a^2(1 - r^2),$$

which is strictly decreasing in r for $a \ge 1$, and hence minimized at $r = 1$. Symmetry of the meridian speed about $u = \pi/2$ puts the intrinsic constant-speed midpoint at this same point. The vertical chord velocity is tangent there and has magnitude $2ac_a$; the intrinsic velocity is vertical with magnitude L_a. The midpoint projected laws also coincide (both uniform on the equator), so this particular midpoint comparison admits a common-base-point field interpretation in addition to its pairwise interpretation.

Asymptotic size and explicit bounds. Rationalize the length difference rather than subtract nearly equal quadratures:

$$G_a = 2\int_{\theta_a}^{\pi/2} \frac{\cos^2 u}{\sqrt{\cos^2 u + a^2 \sin^2 u} + a \sin u}\, du.$$

For $C_a = \log\cot(\theta_a/2) - c_a = \operatorname{atanh}(c_a) - c_a > 0$,

$$\frac{2C_a}{a\{1 + \sqrt{1 + \cot^2\theta_a/a^2}\}} \le G_a \le \frac{C_a}{a}. \tag{23}$$

This follows from $\int_{\theta_a}^{\pi/2} \cos^2 u / \sin u\, du = C_a$ and the maximum of $\cot u$ on the integration interval. Since $c_a \sim (2a)^{-1/2}$ and $\operatorname{atanh}(c_a) - c_a \sim c_a^3/3$, both bounds have the asymptotic coefficient in Eq. (22). The verification script checks the two bounds and independent quadratures for nine aspect ratios. There is no shooting solver, learned model, fitted exponent, or statistical uncertainty in this example.

The measures vary with a and are one-dimensional ring measures on a two-dimensional surface. This scope is essential: the construction is not a refutation of an existential large-bias example or a claim for densities bounded below on the entire surface. It establishes that tube confinement and shared OT optimality alone do not impose a lower growth law in the global diameter-to-reach ratio.

B.10 Full-dimensional ellipsoid bands

The exact ring formula has a common-midpoint interpretation, but its singularity is not needed to refute a measure-free lower law for the same-pair discrepancy.

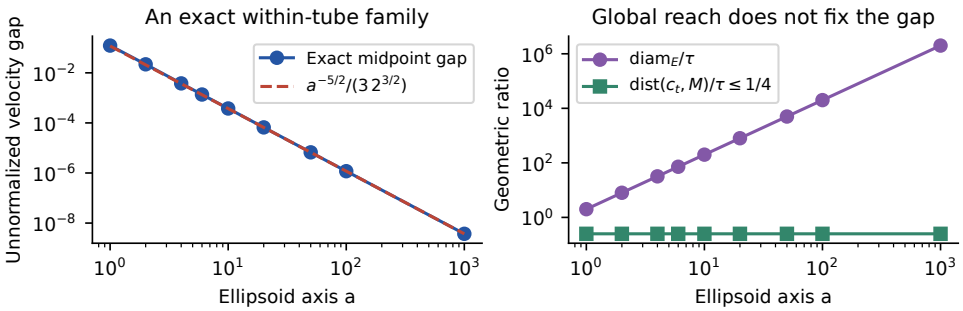

Figure 7: Exact ring transports on prolate ellipsoids. The velocity gap decreases while the global geometric ratio increases. All chords remain inside a fixed fraction of the reach tube. The family changes the transported rings with a and has no learned model or numerical geodesic solver.

Corollary 10 (Smooth-density same-pair counterexample)**.** *For $a \ge 4$, let $h_a = \arccos(1-1/(4a))$ and choose a polar law $\nu_a \ll du$ supported on $[\pi/2 - h_a - 0.1, \pi/2 - h_a + 0.1]$. With independent uniform azimuth, transport $X(u,\varphi)$ to $X(u+2h_a,\varphi)$ on $\mathcal{M}_a$. The source and target are surface-absolutely-continuous laws, and this coupling is optimal for both squared costs. With $C = 38400/12691 < 4$, every coupled chord satisfies*

$$\mathrm{dist}(c_t, \mathcal{M}_a) \le \frac{C}{8a} < \frac{\tau_a}{2}, \qquad G_{1/2} = \left\| P_p\{\Delta - \dot\gamma_{1/2}\} \right\| \le \frac{C}{4a} < \tau_a.$$

Consequently both its mean and L^2 same-pair discrepancy vanish while $\mathrm{diam}_E(\mathcal{M}_a)/\tau_a = 2a^2 \to \infty$. Every coupled intrinsic distance is at least $(7/8)\sqrt{2a}$. The two surface densities may be chosen globally smooth, with support in these bands.

Proof. First $h_a < 0.4$: $\cos(0.4) \le 1 - 0.4^2/2 + 0.4^4/24 < 15/16 \le \cos h_a$. All relevant meridians therefore lie within polar distance 0.5 of the equator, where $\sin u \ge \cos(0.5) \ge 7/8$ and $|\cos u| \le 0.5$. Ambient squared distance dominates the same-azimuth meridian cost, since its additional azimuth term is $2\sin u \sin v\{1 - \cos(\varphi' - \varphi)\} \ge 0$. The meridian cost has mixed derivative

$$-2\{\cos u \cos v + a^2 \sin u \sin v\} \le -2(49a^2/64 - 1/4) < 0.$$

Let $s(u)$ be meridian arc length. Every surface curve between polar angles u, v has length at least $|s(v) - s(u)|$, attained by a constant-azimuth meridian; the squared lower cost has mixed derivative $-2s'(u)s'(v) < 0$. Thus the increasing shift $T(u) = u + 2h_a$ is optimal for both polar problems. The common independent azimuth law makes this coupling feasible on the whole surface, where it attains both lower bounds.

The curvature of a meridian and its coupled arc length satisfy

$$\kappa(u) = \frac{a}{(\cos^2 u + a^2 \sin^2 u)^{3/2}} \le \frac{512}{343a^2}, \qquad L \le 2ah_a.$$

Also $1 - \cos h_a \ge h_a^2/2 - h_a^4/24 \ge 37h_a^2/75$, giving $4ah_a^2 \le 75/37$. Conversely, $1 - \cos h_a \le h_a^2/2$ gives $h_a \ge 1/\sqrt{2a}$ and $L \ge (7/8)2ah_a \ge (7/8)\sqrt{2a}$. The constant-speed meridian has acceleration bounded by

$$\|\ddot\gamma\|_\infty \le \kappa_{\max} L^2 \le \frac{38400}{12691a} = \frac{C}{a}.$$

The twice-integrated endpoint interpolation bound yields

$$\|c_t - \gamma_t\| \le \tfrac{1}{2}t(1-t)\,\|\ddot\gamma\|_\infty \le \frac{C}{8a}.$$

Since $\gamma_t \in \mathcal{M}_a$, this implies the tube bound and unique nearest projection. Furthermore, $\Delta = \int_0^1 \dot\gamma_s\, ds$, so

$$\left\|\Delta - \dot\gamma_{1/2}\right\| \le \|\ddot\gamma\|_\infty \int_0^1 |s - 1/2|\, ds \le \frac{C}{4a}.$$

Contractivity of P_p proves the uniform same-pair bound and hence both moment bounds. Surface area has element $\sin u\, s'(u)\, du\, d\varphi$, which is positive in the bands. Choosing a C^∞ polar bump and translating it gives smooth surface densities that extend by zero outside the bands. Global positivity is not claimed. □

In general $p \neq \gamma_{1/2}$ for an off-center band pair: both velocities in $G_{1/2}$ are projected with P_p. This corollary does not assert equality of midpoint laws or define the intrinsic field at unsupported points. Its tube fraction is $\tau_a/2$, not the sharper $\tau_a/4$ of the ring construction.

Numerical checks use nine aspect ratios from 4 to 64. A Cartesian ambient assignment on 17×8 polar/azimuth grids agrees with the increasing same-azimuth map. A separate assignment verifies the polar intrinsic lower bound; full intrinsic optimality follows from the proved bound and its attainment, not a numerical ellipsoid distance matrix. Meridian lengths use independent quadrature. For interior chord points, nearest projection solves $q = (I + \lambda \operatorname{diag}(1, 1, a^{-2}))^{-1} c$ with $-1 < \lambda < 0$; the positive-definite Lagrangian Hessian certifies the global nearest point. The maximum surface-constraint residual is 2.7×10^{-15}. Across sampled paths the maximum distance/reach ratio is 0.25125, below the proved 0.37823 bound. Sixty-four-point Gauss quadrature evaluates mean and L^2 gaps for both uniform polar laws and normalized bumps proportional to $\exp\{-1/(1-z^2)\}$, $|z| < 1$, where $z = (u - \pi/2 + h_a)/0.1$.

a	diam/τ	Sampled dist.$/\tau$	Uniform L^2	Smooth L^2	Bound
4	32	0.2511	0.00381	0.0038	0.189
6	72	0.2512	0.00137	0.00137	0.126
8	128	0.2512	0.000664	0.000662	0.0946
12	288	0.2512	0.00024	0.000239	0.063
16	512	0.2512	0.000117	0.000116	0.0473
24	1152	0.2512	4.22×10^{-5}	4.21×10^{-5}	0.0315
32	2048	0.2512	2.05×10^{-5}	2.05×10^{-5}	0.0236
48	4608	0.2512	7.44×10^{-6}	7.42×10^{-6}	0.0158
64	8192	0.2512	3.62×10^{-6}	3.61×10^{-6}	0.0118

Table 2: Full-dimensional ellipsoid band checks. Both L^2 columns use same-pair projected velocities, with uniform or smooth polar densities. The proved upper bound is $C/(4a)$ and holds uniformly over each band; sampled tube maxima are not substitutes for the analytic bound.

B.11 Feature-map bound and complementary calibrations

Fix a linear ambient path $c_t = x_0 + t\Delta$. Assume F is C^2 on the chord and the finite-difference segment at c_t, with derivative bound L and bilinear Hessian bound H. For $v = v_\theta(t, c_t)$,

$$U_t^\epsilon - [F(x_1) - F(x_0)] = [U_t^\epsilon - DF(c_t)v] + DF(c_t)(v - \Delta) + \int_0^1 [DF(c_t) - DF(c_s)]\Delta\, ds.$$

Taylor's theorem bounds the first term by $\epsilon H \|v\|^2 /2$. The second is at most $L\|v - \Delta\|$. The last is at most

$$H\|\Delta\|^2 \int_0^1 |t - s|\, ds = \frac{H}{2}\|\Delta\|^2 \{t^2 + (1-t)^2\}.$$

Orthogonal projection is contractive, which proves Eq. (8). No manifold curvature is needed for this error source. A nonlinear representation changes the geometry of the embedded feature curve, which is precisely why feature and ambient conclusions must be kept distinct.

For $F_\alpha(x) = (x, \alpha x^2)$, the feature velocity is $(1, 2\alpha t)$, the endpoint difference is $(1, \alpha)$, and a unit tangent is $e_t = (1, 2\alpha t)/\sqrt{1 + 4\alpha^2 t^2}$. Their projected difference has norm

$$|\langle e_t, (0, \alpha(2t - 1))\rangle| = \frac{2\alpha^2 |t(2t-1)|}{\sqrt{1 + 4\alpha^2 t^2}}.$$

The finite-difference implementation replaces $2t - 1$ by $2t - 1 + \epsilon$. The accompanying script records both versions, confirming convergence to the analytic expression. No claim of differentiability through image clipping is used.

Proposition 11 (Zero feature score with positive intrinsic field error)**.** *Consider either common-OT class in Theorems 1–2, with* $\Pr(d > 0) > 0$*. Evaluate the diagnostics on its paired ambient paths* c_t*. Use affine features* $F(x) = Ax + b$*, without clipping, and any orthogonal feature projector. At the midpoint, the exact ambient field* u *gives zero retrospective and forecast diagnostic for every* $\epsilon > 0$*, yet its intrinsic field discrepancy* $B_{1/2}$ *in Eq.* (3) *is positive. In particular, this occurs for the injective identity feature map.*

For identity features and the oracle projector P_p*, an arbitrary learned field has*

$$\left\|\widehat{G}^{\mathrm{retro}}_{1/2}\right\|_{L^2} = e, \qquad \left\|\widehat{G}^{\mathrm{forecast}}_{1/2}\right\|_{L^2} = \tfrac{1}{2}e, \qquad |D - B_{1/2}| \le e, \tag{24}$$

where $e = \|P_p(v - u)\|_{L^2}$ *and* $D = \|P_p v - w(p)\|_{L^2}$*. Thus, as* $e \to 0$*, both diagnostics vanish while* $D \to B_{1/2} > 0$*.*

Proof. For an affine feature map, $U^\epsilon_t = Av$ exactly. On a paired linear path, the retrospective vector before projection is $A(v - \Delta)$. The forecast endpoint satisfies

$$F(c_t + (1 - t)v) - F(x_0) = A\{t\Delta + (1 - t)v\},$$

so its diagnostic vector is $tA(v - \Delta)$. At the common-OT midpoint, the point identifies its endpoint pair, hence $u = \Delta$. Both vectors vanish for $v = u$, before any projector is applied. Meanwhile $\delta - 2\sin(\delta/2)$ is nonzero for each nonzero displacement in these short-interval classes, giving $B_{1/2} > 0$. Taking $A = I$, the exact projector, and midpoint L^2 norms gives the first two equalities in Eq. (24); the reverse triangle inequality gives the last bound. □

For the unit-radius, constant-shift $\delta = 0.6$ transports used in the learned controls, $B_{1/2} = 0.0089595867$ despite both exact-field feature scores being zero. This connects the geometric controls to the operational comparator: with identity features, the diagnostic tracks the projected ambient learning residual rather than the remaining intrinsic discrepancy. It does not claim that nonlinear image encoders behave affinely. Cartesian checks cover 48 circle/band, radius, and signed-shift settings, including the identity transport, three finite-difference steps, identity and injective affine maps, and perturbed fields. The tests verify both zero-score identities and the factors in Eq. (24); the continuous proof establishes the transport claim.

Corollary 12 (A zero-preserving comparator family)**.** *On an exact straight paired ambient path* c_t*, let* F *be affine and let* L_t *be any linear map, possibly depending on the path or data. Form a score from* $L_t\{DF(c_t)\Delta - [F(x_1) - F(x_0)]\}$ *by transformations and aggregation that send identically zero input to zero. Then the score is zero. This family admits no uniform intrinsic-gap certificate* $B_{1/2} \le f(S)$ *with* $f(0) = 0$ *over the nonidentity common-OT constructions of Theorems 1–2.*

Proof. Affineness makes the residual zero before applying L_t. The preceding proposition supplies a common-OT construction with positive $B_{1/2}$ and exact ambient field. A bound with $f(0) = 0$ would make this positive gap at most zero. □

The family includes tangent projectors, weights, norm powers, and defined normalizations that preserve zero. The forecast residual also vanishes on the exact straight field because its forecast is the paired endpoint. This algebraic extension is elementary; the substantive comparison requires the proved common coupling and conditional-field gap. Diagnostics using intrinsic targets or curvature information outside this residual are not excluded, and prior ambient straightness measures are not claimed to have an intrinsic interpretation.

B.12 Independent variation of geometry, residuals, and measurement

This additional calibration uses prescribed fields rather than new model training. Its complete parameter grid and figure slices were declared before computation, after the earlier results had been

Prescribed field	B	e	D	Score
Exact ambient	0.010582	0.000000	0.010582	0.000000
Toward intrinsic	0.010582	0.004000	0.006582	0.004000
Away from intrinsic	0.010582	0.004000	0.014582	0.004000
Azimuthal	0.010582	0.004000	0.011312	0.004000

Table 3: Prescribed fields on the same sphere-band transport, $r = 0.5$, $\bar{d} = 0.4$, $q = 0$. "Score" is the retrospective midpoint RMS diagnostic with identity features and the exact tangent projector. The forecast score is half this value. No field in this table is a newly trained model.

inspected. On $\mathbb{S}^2(r)$, the source has independent uniform polar angle $a \in [0.8, 1.1]$ and uniform azimuth. The transport preserves azimuth and maps

$$T(a) = a + \frac{\bar{d}}{r}\left(1 + q\frac{a - 0.95}{0.15}\right), \qquad r \in \{0.5, 1, 2, 4\}, \quad \bar{d} \in \{0.1, 0.2, 0.4\}, \quad q \in \{0, 0.25\}.$$

Here $\bar{d}$ is mean intrinsic distance and q controls displacement variation. Every map is increasing, and every source/target support lies in the common interval $[0.8, 2.1]$ of length $1.3 < \pi/2$. Thus all 24 geometries satisfy Theorem 2. A 64×64 midpoint quadrature grid evaluates polar/azimuth averages. Doubling polar resolution changes $B_{1/2}$ by at most 2.44×10^{-5} relatively.

At the common midpoint, add a tangent residual of magnitude $e \in \{0, 0.001, 0.004, 0.016\}$ to the exact ambient field. Its direction is toward the intrinsic target, away from it, or azimuthal. The 288 grid combinations contain 240 distinct prescribed fields, since zero residual repeats across direction labels. We evaluate Cartesian endpoint differences, exact tangent projectors, identity features, and finite-difference step 0.01. Across the geometries, $B_{1/2}$ ranges from 2.60×10^{-6} to 0.0121535, while exact-field diagnostic values remain below 1.46×10^{-14}. For every prescribed residual, the retrospective and midpoint forecast RMS scores equal e and $e/2$ to the same numerical precision. The target discrepancy D also depends on the residual direction (Table 3). For constant displacement, its three values are $|B_{1/2} - e|$, $B_{1/2} + e$, and $\sqrt{B_{1/2}^2 + e^2}$. Consequently, equal scores need not imply equal intrinsic discrepancies, and decreasing the residual score need not decrease D.

To vary measurement while fixing physical geometry, set $r = 1$, $\bar{d} = 0.4$, $q = 0.25$, and use a 12×24 grid with

$$F_\alpha(x) = (x, \alpha x_z^3/r^2), \qquad \alpha \in \{0, 0.25, 0.5, 1\}.$$

This smooth injective map changes the representation; the target remains intrinsic motion on the original sphere. We evaluate exact feature tangents at both $F(p)$ and the nearest point on $F(\mathbb{S}^2(r))$ to $F(c_{1/2})$, and centered local PCA with known rank two or a 95%-variance cutoff. The fixed reference bank contains 8,192 Fibonacci points covering the whole sphere, with neighborhoods of 32, 128, or 512 points. The variance rule selects rank two throughout this grid, so it does not test dimension-selection failures. Tangent residuals $0, 0.004$, normal residuals $0, 0.02$, and derivative/finite-difference lifts with $\epsilon = 0.01, 0.001$ give 384 measurement cases. All are released.

With no residual, an exact derivative lift, and the exact nearest-feature tangent, the score changes from zero at $\alpha = 0$ to 0.0043804 at $\alpha = 1$, while the original-sphere discrepancy remains 0.0030583. This isolates representation dependence without finite-difference or tangent-estimation error. With identity features and a purely normal residual of 0.02, the exact projector gives zero score up to roundoff; estimated rank-two projectors give 1.06×10^{-4} at $k = 32$ and 4.20×10^{-5} at $k = 512$. These are controlled measurement effects, not estimates of image-manifold error.

Independent ambient and intrinsic assignment matrices recover the declared coupling on all 24 transport grids. Central differences of slerp geodesics and radial nearest-point projection agree with the separately evaluated velocities within 7.8×10^{-11}. Feature Jacobians and tangent spaces are checked by independent central differences; PCA covariance eigenspaces are checked against direct SVD. The nearest-feature point is obtained by scalar root finding, with a positive Hessian bound and an independent dense objective check. These numerical checks supplement the common-OT theorem; they do not extend it to unrestricted couplings. The supplement includes the declared design, all outputs, and the numerical and figure-rendering sources.

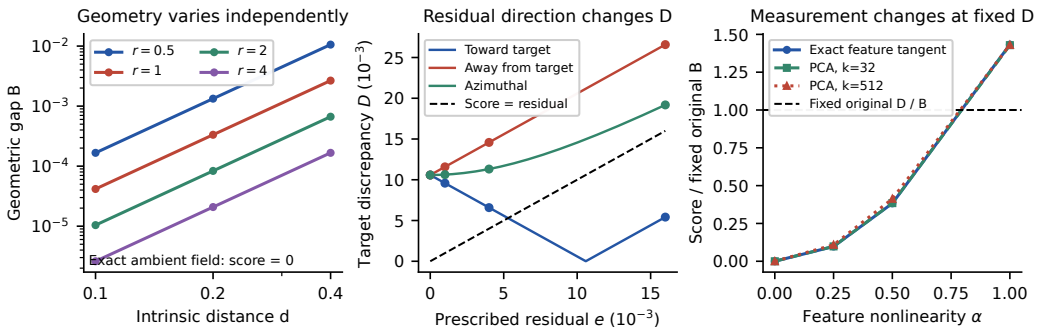

Figure 8: Separating geometry, prescribed residuals, and measurement. Left: radius and distance vary while exact-field identity scores remain zero. Middle: residual direction changes D at the same score; markers are computed grid cases and lines are the exact response formulas. Right: feature nonlinearity changes the score at fixed original-sphere discrepancy, with exact and estimated tangents. The dashed ratio one marks the physical target. No statistical error bars apply to these deterministic calculations.

C Conditional fields and what an additive bound establishes

Paired-target variance. Fix a coupling and a deterministic field v and projector $P(c_t)$. Let $u(c_t) = \mathbb{E}[\Delta \mid c_t]$. The conditional mean of $P(c_t)(\Delta - u(c_t))$ vanishes. Squaring and taking expectations yields

$$\mathbb{E}\,\|P(v-\Delta)\|^2 = \mathbb{E}\,\|P(v-u)\|^2 + \mathbb{E}\,\|P(\Delta-u)\|^2 . \tag{25}$$

The last term need not vanish even if $v = u$. If the field or projector was estimated from a random dataset, the identity applies conditionally on the fixed fitted objects with the corresponding conditional target field. Mean norms do not obey this variance equality; the experiment's raw mean-norm residual is not being decomposed additively by Eq. (25). For individual vectors one only has triangle and reverse-triangle bounds.

A specified tangent comparator. Let $A_t = \{\mathrm{dist}(c_t, \mathcal{M}) \le \rho\}$, with $\rho < \tau$, and define the unnormalized restricted norm

$$\|f\|_{2,A_t} = (\mathbb{E}[\|f\|^2 \mathbf{1}_{A_t}])^{1/2}.$$

It is not the normalized conditional norm. Suppose an explicitly specified tangent field $w_t(p)$ is defined at every projected point under consideration, and $\left\|\widehat{P}_t - P_{p_t}\right\|_{\mathrm{op}} \le \eta_t$ on A_t. Set

$$\epsilon_t = \|v_\theta - u_\zeta\|_{L^2}, \qquad R_t = \|v_\theta\|_{2,A_t}, \qquad B_t = \|P_{p_t} u_\zeta - w_t(p_t)\|_{2,A_t}.$$

Then the triangle inequality gives

$$\left\|\widehat{P}_t v_\theta - w_t(p_t)\right\|_{2,A_t} \le \eta_t R_t + \epsilon_t + B_t. \tag{26}$$

If the intended object is the derivative of a projected learned trajectory, Proposition 4 introduces a further projection-defect bound $\rho R_t/(\tau-\rho)$ when comparing $D\pi(c_t)v_\theta$ with $P_{p_t}v_\theta$. If $\Pr(A_t) > 0$ and a normalized conditional norm is desired, the corresponding restricted quantities must be divided by $\sqrt{\Pr(A_t)}$.

Equation (26) organizes errors, but does not itself identify B_t. Taking w_t to be an intrinsic Wasserstein field requires that a suitable version is defined on the projected law, not merely almost everywhere under a different intrinsic interpolation law. Coupling changes, base-point changes, and any extension regularity must be analyzed separately. Theorems 1 and 2 resolve these issues at the midpoint for their specified classes, yielding the exact $B_{1/2}$. No general finite-sample rate for ϵ_t follows from entropic OT cost estimation (Mena & Niles-Weed, 2019), and no PCA rate for η_t follows from choosing $k = 300$. These remain separate estimation questions.

D EXPERIMENTAL PROTOCOLS AND FULL PRESERVED RESULTS

D.1 INTERPRETATION OF THE PRESERVED SPHERE SWEEP

The source is uniform on $\mathbb{S}^2(1)$. Targets use four tetrahedral centers, with Gaussian tangent noise of variance $1/\kappa$ added to each center followed by radial normalization; this is a normalized tangent-Gaussian mixture, not an exact von Mises–Fisher sampler. We use five training seeds, 2,000 optimizer updates, training assignments of size 256, evaluation assignments of size 500, and $\kappa \in \{5, 10, 20, 50, 100, 200\}$. Each assignment minimizes squared ambient distance. Architecture and full settings are in Appendix D.

As κ increases, the mean paired distance increases, the raw geometric integral increases from approximately 0.0017 to 0.0213, and the raw regression residual decreases from 0.1400 to 0.0931 (Figure 9). The opposite trends distinguish two measured components, but do not establish statistical independence or invariant optimization difficulty. The speed normalization amplifies the residual's decrease. A separate update-budget sweep holds the evaluation pairs fixed; the geometric component is then constant by definition because it contains no model evaluation.

The analytic formula reproduces all 30 stored normalized geometric integrals to maximum absolute error 1.3×10^{-8}. Comparing the leading law on the *same ten-point grid* gives maximum relative error below 0.37%. Independent double-precision vector calculations over 2,000 randomly chosen angles, radii, and times agree with the exact formula to 1.3×10^{-15}. These checks validate the stated sphere proxy, not equality of ambient and intrinsic optimal couplings in general.

D.2 CONTROLLED DISTRIBUTION AND TRAINING

The four centers are the unit-normalized vectors $(1, 1, 1)$, $(1, -1, -1)$, $(-1, 1, -1)$, and $(-1, -1, 1)$. For a center m, draw $z \sim \mathcal{N}(0, I_3/\kappa)$, replace it by $(I - mm^T)z$, and return $(m + z)/\|m + z\|$. Evaluation uses balanced center counts. The velocity network is an MLP with three hidden layers of width 128, SiLU activations, and concatenated scalar time input. Adam uses learning rate 10^{-3}. Each optimizer update uses a fresh 256-pair minibatch and an ambient squared-distance OT assignment. There are 2,000 updates in the concentration sweep and $\{250, 500, 1000, 2000, 4000\}$ updates in the budget sweep, with five seeds numbered 0–4.

Evaluation uses 500 fresh source and target points coupled in one assignment, with a fixed evaluation seed offset of 99,999 and ten equally spaced times in $[0.05, 0.95]$. Ground-truth sphere tangent projectors and same-pair great circles are used; no local-PCA tangent estimation enters these results. The integration is the trapezoid rule applied to the mean per-pair norm. The normalized version divides each pair's norm by $\|P_{p_t}\dot{\gamma}_t\|$ before averaging.

κ	Paired distance	Raw G	Raw R	Norm. G	Norm. R
5	0.248 ± 0.025	0.0017 ± 0.0004	0.1400 ± 0.0121	0.0051 ± 0.0009	0.8220 ± 0.1528
10	0.350 ± 0.025	0.0042 ± 0.0007	0.1308 ± 0.0128	0.0096 ± 0.0012	0.4566 ± 0.0516
20	0.439 ± 0.023	0.0079 ± 0.0009	0.1205 ± 0.0089	0.0149 ± 0.0013	0.3209 ± 0.0423
50	0.529 ± 0.021	0.0137 ± 0.0011	0.1104 ± 0.0062	0.0215 ± 0.0014	0.2369 ± 0.0243
100	0.577 ± 0.020	0.0178 ± 0.0012	0.1004 ± 0.0055	0.0257 ± 0.0014	0.1952 ± 0.0220
200	0.612 ± 0.019	0.0213 ± 0.0012	0.0931 ± 0.0052	0.0289 ± 0.0014	0.1656 ± 0.0165

Table 4: Concentration sweep: mean $\pm$ sample SD over five seeds. G is the same-pair geometric proxy and R the paired-target regression residual. Normalized and raw quantities are separate measurements.

D.3 IMAGE MODELS, FEATURES, AND EVALUATION GEOMETRY

The CIFAR-10 (Krizhevsky, 2009) experiments use the canonical 50,000 training and 10,000 test images. The backbone is the archived `diffusers.UNet2DModel`: image size 32, three input/output channels, two layers per block, widths $(128, 256, 256, 256)$, down blocks (Down, AttnDown, Down, Down), and up blocks (Up, Up, AttnUp, Up). Evaluation prefers the stored exponential-moving-average weights. Images are generated by 50 explicit Euler steps of size 0.02. The checkpoint and code hashes in the result manifest identify the evaluated assets.

Updates	Raw residual R	Raw geometry G	Normalized residual
250	0.1732 ± 0.0157	0.0079 ± 0.0009	0.3999 ± 0.0318
500	0.1648 ± 0.0144	0.0079 ± 0.0009	0.3928 ± 0.0144
1000	0.1371 ± 0.0101	0.0079 ± 0.0009	0.3556 ± 0.0267
2000	0.1205 ± 0.0089	0.0079 ± 0.0009	0.3209 ± 0.0423
4000	0.1144 ± 0.0047	0.0079 ± 0.0009	0.2967 ± 0.0316

Table 5: Budget sweep at $\kappa = 20$. The geometric quantity is fixed by the evaluation pairs and is therefore invariant to training budget by construction.

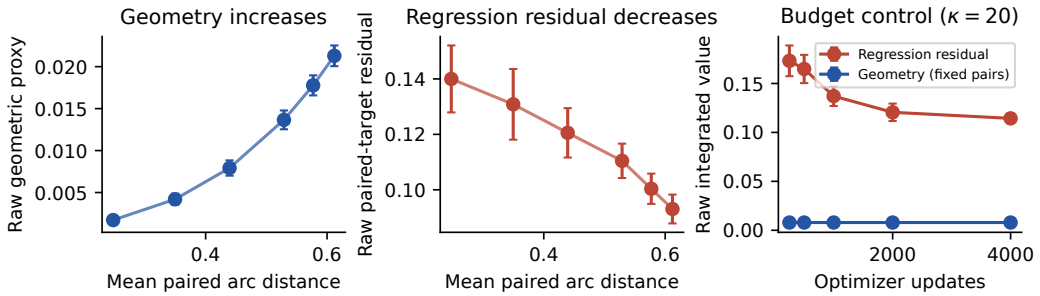

Figure 9: Five-seed controlled results; bars show sample SD. The geometric pairwise proxy increases with paired distance, while the raw regression residual decreases. Fixing evaluation pairs makes geometry independent of the optimization budget by construction. All plotted integrals use $[0.05, 0.95]$.

The three evaluated CIFAR-10 checkpoint configurations specify batch size 128, AdamW with learning rate 2×10^{-4} and zero weight decay, 2,000 warmup iterations followed by cosine decay, gradient clipping at 1, EMA decay 0.9999, and 400,000 training iterations. Training uses exact minibatch EMD with uniform marginals and squared pixel distance; the stored configuration enables mixed precision during training. The older 10k controls are independent short-budget runs with a 10k cosine horizon and 500 warmup iterations; they are not saved prefixes of these 400k runs.

The preserved ImageNet pilot uses the downsampled 64×64 training-shard format of Chrabaszcz et al. (2017), with 1,281,167 RGB images. Its backbone has 12 DiT layers (Peebles & Xie, 2023), six attention heads of dimension 64, and 2×2 image patches. The loader ignores dataset labels, applies random horizontal flips, and maps pixels to $[-1, 1]$; the DiT receives a constant conditioning label. Both stored configurations specify 200,000 iterations, batch size 64, learning rate 2×10^{-4}, 2,000 warmup iterations, exact minibatch EMD, mixed precision, and EMA decay 0.9999. Diagnostic endpoints come from augmented training images; the reference-bank builder encodes unaugmented training images. The separate validation-feature cache belongs only to the excluded historical quality evaluation. These are preserved diagnostic observations, not newly replicated training runs.

The feature extractor is torchvision InceptionV3 with `IMAGENET1K_V1` weights, the classifier replaced by the identity, and the auxiliary head removed. Inputs are resized and center-cropped to 299×299, with ImageNet channel normalization, following the preserved wrapper. The pretrained factory also enables its internal `transform_input`; together these transformations produce the model's expected input scale. The new evaluation uses the same wrapper for generated images, training-reference images, and test-reference images, with PyTorch 2.8.0, torchvision 0.23.0, diffusers 0.30.0, float32 arithmetic, and no automatic mixed precision or TF32 matrix multiplication. This specifies one torchvision feature pipeline; it is not a claim of equivalence to every external FID feature implementation.

For preserved checkpoint diagnostics, a Gaussian source is paired with real training images in minibatches and evaluated at the linear interpolants. The model velocity is lifted with a finite image-space perturbation of $\epsilon = 0.01$, including clipping. Real training targets follow the training dataset transformations, including random horizontal flipping. Local PCA centers the selected neighbor features and retains the minimum rank explaining at least 95% variance. The older implementation uses a

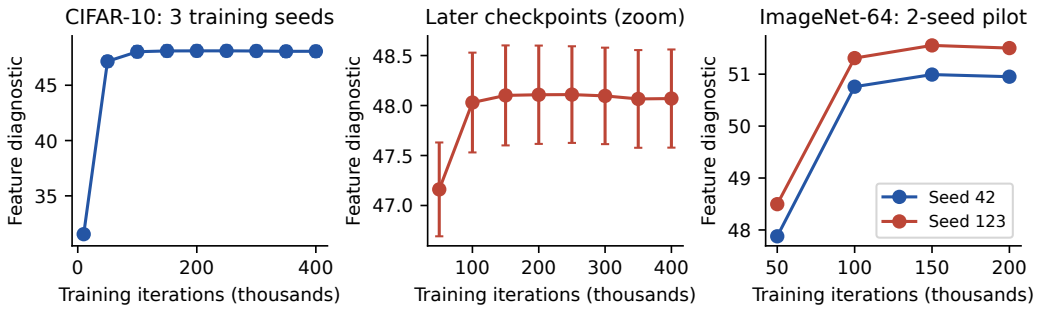

Figure 10: Preserved feature diagnostics. Separate 10k-budget CIFAR-10 controls differ from 50k–400k runs; the later-run series nearly saturates; bars show sample SD across three seeds. ImageNet-64 gives two descriptive checkpoint series. Older quality values are excluded because their real-reference feature pipeline could not be validated.

FAISS IVF index with 100 lists and 10 probes, so these neighborhoods are approximate. Queries are the state features, not oracle projections onto an unknown manifold. The prospective experiment uses exact feature-space search and therefore has a separately stated neighborhood protocol.

Iterations	Feature diagnostic
10k	31.53 ± 0.19
50k	47.16 ± 0.47
100k	48.03 ± 0.50
150k	48.10 ± 0.50
200k	48.11 ± 0.49
250k	48.11 ± 0.48
300k	48.10 ± 0.48
350k	48.07 ± 0.49
400k	48.07 ± 0.49

Table 6: CIFAR-10 preserved diagnostic results: three training seeds, 500 diagnostic pairs, and $k = 300$. Scores average norms and integrate over $[0.05, 0.95]$. The incompatible quality-reference cache is not used for these values.

Iterations	Score (42)	Score (123)
50k	47.88	48.50
100k	50.76	51.31
150k	50.99	51.55
200k	50.95	51.50

Table 7: ImageNet-64 DiT pilot, with 300 diagnostic pairs and $k = 200$ for each of two seeds. These are diagnostic values without a validated quality correlation; checkpoints from a training run are not independent replicates.

D.4 The quality statistic

For features $X = \{x_i\}_{i=1}^n$, $Y = \{y_j\}_{j=1}^m$, define $k(x, y) = \exp(-\|x - y\|^2 / (2\sigma^2))$, with $\sigma = 10$. We use

$$\mathrm{IMMD}(X, Y) = 1000 \left[\frac{\sum_{i \neq j} k(x_i, x_j)}{n(n-1)} + \frac{\sum_{i \neq j} k(y_i, y_j)}{m(m-1)} - \frac{2 \sum_{i,j} k(x_i, y_j)}{nm} \right]. \quad (27)$$

For independent samples this is 1,000 times the unbiased MMD-squared estimator for the specified feature kernel (Gretton et al., 2012, Section 2.2), so negative finite-sample estimates are allowed rather than clipped to zero. For score-selected subsets we retain the same diagonal-excluding statistic

Method	Prefix-only	Low IMMD ↓	High IMMD	High−low
Random split	Yes	6.14 ± 0.45	6.49 ± 0.25	0.35 ± 0.40
Pixel velocity magnitude	Yes	8.41 ± 0.29	6.36 ± 0.03	-2.05 ± 0.29
Feature velocity magnitude	Yes	9.88 ± 0.88	6.98 ± 0.52	-2.89 ± 0.50
Normal feature velocity	Yes	9.73 ± 0.63	6.82 ± 0.71	-2.91 ± 0.11
Nearest-reference distance	Yes	6.81 ± 0.16	6.45 ± 0.38	-0.36 ± 0.45
Forecast deviation	Yes	10.53 ± 0.88	9.67 ± 0.58	-0.86 ± 0.46
Final-endpoint deviation	No	9.27 ± 1.15	12.78 ± 0.68	3.51 ± 0.57

Table 8: Original Inception primary comparison, $T = 0.5, k = 300$. Errors are sample SD across three training seeds. A positive high-minus-low bucket gap does not establish improvement over an equally sized random subset.

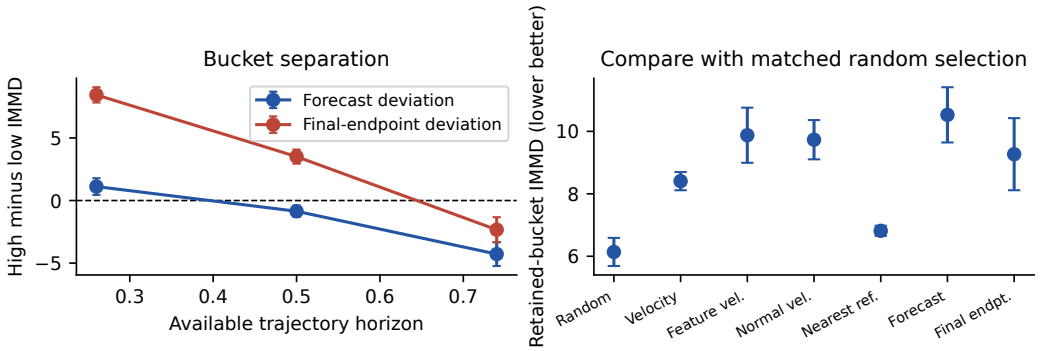

Figure 11: Consistent-feature evaluation. Left: high-minus-low bucket differences at all prespecified horizons ($k = 300$). Right: IMMD of the retained half at the primary horizon. A positive bucket difference does not establish improvement over matched random selection. Bars show sample SD across three seeds.

without claiming population unbiasedness after selection. The image-evaluation design is inspired by Jayasumana et al. (2024); its numerical values are not CLIP-CMMD or FID values. A selected subset can change the generated distribution even when it looks individually better, which makes comparison with a same-sized random subset essential.

D.5 Retrospective endpoint-conditioned filtering

The preserved filtering experiment uses three training seeds and 5,000 completed trajectories per seed. The score uses the completed endpoint's features, although its state/velocity inputs come from the first half of the trajectory. The nominal integration labels are $\{0, .05, .15, .25, .35, .45, .5\}$; nearest Euler states actually correspond to $\{0, .06, .14, .24, .34, .44, .5\}$. Each bucket contains 2,500 generated samples and was compared with 5,000 cached real features. These quality conclusions are withdrawn because of the reference-feature incompatibility in Appendix H.1.

Independently of that incompatibility, the final-endpoint dependency prevents an early-stopping or saved-compute conclusion. Zero exceedances in a 500-permutation Monte Carlo test should be reported with the plus-one estimate $1/501$, not as probability zero. The new experiment in Section 6.3 corrects feature compatibility, information availability, time labels, sample-matched baselines, and permutation reporting. Stored older aggregates remain in the audit files with an explicit invalid-reference status, rather than being reused as quality evidence.

E Prospective evaluation: detailed results and availability

The sampling, score, baseline, horizon, and retention protocol was fixed before generating the new trajectories. A subsequent implementation audit found the incompatible real-reference cache; it was replaced uniformly with freshly encoded test features and all quality statistics were recalculated.

The generated samples and rankings were unchanged. The three training seeds are 42, 123, and 7; fresh sampling uses the respective seed plus 20,260,925. Noise is generated on CPU before batching to keep the sample set independent of evaluation batch size. The primary horizon is 0.5, the retained fraction is 0.5, and all 2,048 candidates for a training seed are ranked together using only quantities available at the stated horizon.

For each selected Euler step, the capture phase records the state feature, a finite-difference feature velocity, the single-step endpoint-forecast feature, and pixel velocity magnitude before completing that trajectory. Final-image features are stored separately. The scoring stage constructs the PCA neighborhood from the state alone; final features enter only the retrospective comparator and the downstream quality calculation. An initial Gram-PCA/direct-SVD check gives maximum norm error 1.7×10^{-5} on its checked inputs; Appendix F.4 reports additional rank-threshold sensitivity. Perturbing final-image features and all post-horizon inputs leaves the primary forecast score unchanged exactly, while changing the retrospective comparator. The matched real-reference check agrees to maximum feature-coordinate error 5.3×10^{-6} on 32 evenly spaced canonical test images.

Baseline definitions. At each recorded time, the pixel-magnitude baseline is $\|v_\theta\|$, the feature-magnitude baseline is $\|U_t^\epsilon\|$, the normal-velocity baseline is $\left\|(I - \widehat{P}_t)U_t^\epsilon\right\|$, and the neighborhood baseline is the distance from $F(x_t)$ to its nearest real training feature. The same trapezoid rule on actual Euler times produces each integrated score. The retained direction is always the lower score, including when the empirical association has the opposite sign; signs are not selected after observing outcomes. The retrospective method uses the final endpoint in Eq. (6), and the random baseline uses a fixed independent pseudorandom ordering. No method is tuned on test IMMD.

Permutation calculation. Let there be $N = 2b$ generated features and m reference features. Write $K_{ij} = k(x_i, x_j)$ with zero diagonal, $c_i = m^{-1} \sum_j k(x_i, y_j)$, and let $s_i = 1$ identify the high bucket and $s_i = -1$ the low bucket. For equal bucket sizes,

$$\mathrm{IMMD}_{\mathrm{high}} - \mathrm{IMMD}_{\mathrm{low}} = \sum_i s_i \, 1000 \left[\frac{\sum_j K_{ij}}{b(b-1)} - \frac{2c_i}{b} \right]. \tag{28}$$

The real–real term cancels. This identity permits exact evaluation of each balanced-label permutation after computing kernel row sums once. The implementation verifies agreement with direct bucket U-statistics. For $B = 9999$ permutations, we report $(1 + \#\{|T^{(b)}| \geq |T_{\mathrm{obs}}|\})/(B+1)$. The permutation null is exchangeability of bucket labels conditional on the generated feature set; it is not a null of equal predictive quality between competing methods.

Computation and practical scope. The new runs use a single RTX 3080 GPU with 10 GB memory and batch size 32. The 50,000-image reference bank is constructed once. Per-seed records separate model-forward time, feature-extraction time, and post-generation diagnostic processing. The latter currently includes both k values and the retrospective comparator and therefore should not be presented as the cost of only one deployable score. Likewise, completing all candidates for evaluation does not constitute a deployed early-rejection benchmark. Equal-output-count runtime, calibration of an online threshold, and semantic coverage would be needed before claiming a practical sampling advantage.

Even ideal deployment requires an equal-output-count comparison. If N candidates yield $M = \rho N$ outputs, each accepted trajectory takes S velocity evaluations and each rejected candidate takes J, then the imagewise evaluation count per output is

$$S + J(\rho^{-1} - 1). \tag{29}$$

For our 50-step Euler implementation, the forecast score at $T = 0.5$ needs the velocity at step 25, after evaluations at steps 0 through 24: $J = 26$, including that boundary evaluation. Accepted paths reuse it and complete the remaining 24 steps. Thus $\rho = 0.5$ requires 76 imagewise velocity evaluations per retained output, versus 50 for direct random generation, before any feature or PCA cost. Relative to completing all N candidates the model count decreases, but that baseline produces twice as many completed images. This arithmetic is not a wall-clock benchmark; it specifies the compute comparison needed even when a reversed score improves a quality metric.

Seed	Method	Low	High	Gap	p	Var. ratio
42	Random split	6.151	6.751	0.600	0.3722	0.990
123	Random split	5.681	6.256	0.576	0.3733	0.999
7	Random split	6.582	6.470	-0.112	0.8569	1.004
42	Pixel velocity magnitude	8.515	6.328	-2.187	0.0016	0.992
123	Pixel velocity magnitude	8.075	6.364	-1.711	0.0078	0.993
7	Pixel velocity magnitude	8.627	6.384	-2.242	0.0005	0.990
42	Feature velocity magnitude	9.709	6.611	-3.097	0.0001	0.987
123	Feature velocity magnitude	9.089	6.770	-2.319	0.0002	0.991
7	Feature velocity magnitude	10.831	7.573	-3.258	0.0001	0.993
42	Normal feature velocity	9.570	6.552	-3.019	0.0001	0.989
123	Normal feature velocity	9.201	6.292	-2.909	0.0001	0.987
7	Normal feature velocity	10.428	7.629	-2.798	0.0001	0.994
42	Nearest-reference distance	7.000	6.283	-0.717	0.2856	0.990
123	Nearest-reference distance	6.703	6.192	-0.510	0.4254	0.983
7	Nearest-reference distance	6.742	6.886	0.144	0.8373	0.991
42	Forecast deviation	10.334	9.249	-1.085	0.1080	0.983
123	Forecast deviation	9.758	9.433	-0.325	0.6184	0.987
7	Forecast deviation	11.495	10.334	-1.161	0.0810	0.988
42	Final-endpoint deviation	8.949	12.856	3.907	0.0001	0.935
123	Final-endpoint deviation	8.308	12.066	3.758	0.0001	0.938
7	Final-endpoint deviation	10.550	13.411	2.861	0.0001	0.936

Table 9: New per-seed primary results. p denotes a two-sided within-method balanced-label permutation test. Variance ratio is the retained feature covariance trace divided by the full candidate-set trace; it is not a semantic coverage metric. Multiple method/horizon comparisons are exploratory, and these unadjusted p-values are not evidence of superiority over the random baseline.

k	Horizon	Method	Low IMMD	High−low
300	0.26	Forecast deviation	8.80 ± 0.17	1.12 ± 0.67
300	0.26	Final-endpoint deviation	7.79 ± 1.09	8.44 ± 0.61
300	0.5	Forecast deviation	10.53 ± 0.88	-0.86 ± 0.46
300	0.5	Final-endpoint deviation	9.27 ± 1.15	3.51 ± 0.57
300	0.74	Forecast deviation	12.34 ± 1.25	-4.28 ± 0.95
300	0.74	Final-endpoint deviation	11.44 ± 1.10	-2.33 ± 1.00
100	0.26	Forecast deviation	8.34 ± 0.23	1.40 ± 0.82
100	0.26	Final-endpoint deviation	6.28 ± 0.86	8.88 ± 0.44
100	0.5	Forecast deviation	9.91 ± 0.82	-0.37 ± 0.53
100	0.5	Final-endpoint deviation	8.33 ± 1.05	4.27 ± 0.51
100	0.74	Forecast deviation	11.54 ± 1.08	-3.61 ± 0.95
100	0.74	Final-endpoint deviation	10.65 ± 1.11	-1.51 ± 0.85

Table 10: All prespecified horizon and neighbor-count settings for forecast and retrospective deviation. Errors are sample SD over three training seeds. The primary setting was $T = 0.5, k = 300$; this table is a sensitivity analysis, not a selection of the best test setting.

Seed	Model	Features	Capture wall	All diagnostics
42	154.2	44.7	199.6	168.7
123	153.6	44.5	199.2	171.1
7	153.8	44.5	199.1	171.8

Table 11: Measured seconds for 2,048 candidates per seed. Model time covers all 50 Euler steps; features include all five recorded times, endpoint forecasts, and final images. Capture wall includes these operations and transfer overhead. "All diagnostics" includes both neighbor counts at all five times and both prospective and retrospective scores. It excludes the separate quality-kernel and permutation calculations.

F FEATURE-BACKBONE, METRIC, AND COMPOSITION CHECKS

This extension was designed after inspecting the original Inception results. Its settings and numerical sources were frozen before computing DINO outcomes; it is not an externally preregistered experiment. It keeps the same three trained models, all 6,144 candidate identities, the original score definitions, and the original random splits. Both lower and upper halves are reported. The added FID calculation was separately declared before extracting FID features or observing FID results. No quality outcome chooses the score direction, horizon, neighborhood, bandwidth, or reference set within either declared extension.

Frozen feature extraction and candidate identity. We use the official DINOv2 ViT-S/14 backbone (Oquab et al., 2024), pinned to source commit `7764ea0f912e53c92e82eb78a2a1631e92725fc8`. Both real and generated float tensors in $[0, 1]$ undergo antialiased bicubic resize to 256, center crop to 224, and ImageNet channel normalization. The 384-dimensional normalized CLS token returned by the backbone is used directly for Euclidean neighborhood search and centered PCA; there is no additional unit normalization in the diagnostic. All 50,000 training images form its reference bank, with the same $k \in \{300, 100\}$ and 95% variance rule as before. The unchanged scoring routine processes the same state, finite-difference velocity, and endpoint-forecast inputs. Thus only the representation and its induced neighborhoods change.

Regeneration uses the same full CPU initial-noise tensors, row order, EMA checkpoints, batch size 32, and 50 float32 Euler steps, without autocast or TF32. Every regenerated Inception state, lifted velocity, forecast, initial and final feature, and pixel-velocity norm matches the corresponding saved value exactly in all three seeds. The original run did not save raw pixels; this is complete agreement of its recorded features, not retrospective proof of pixel-bitwise identity. The new run saves final pixels for the additional FID evaluation. Protocols bind source files, model weights, banks, captures, and score arrays by SHA256.

Crossed representations and bandwidths. Each scoring backbone is evaluated under both quality backbones using all 10,000 canonical test images. Inception quality uses raw 2,048-dimensional features with Gaussian bandwidths $5, 10, 20$; 10 remains primary. DINO quality alone uses unit-normalized endpoint features and bandwidths $0.5, 1, 2$; 1 is primary. Both use the diagonal-excluding $10^3\mathrm{MMD}^2$ definition, float64 kernels and sums, and no clipping of negative estimates. Scores retain their original scales because candidates are ranked separately within each representation. We evaluate 2 scoring backbones $\times 2$ quality backbones $\times 3$ bandwidths $\times 2$ neighborhoods $\times 3$ horizons $\times 7$ methods: 504 comparisons per seed, each with both halves. All values are released; the main table fixes the primary settings.

At those primary settings, every lower forecast/retrospective half is worse than its matched random half in each seed, for both quality backbones and all three respective bandwidths. This does not extend to every sensitivity setting. For example, across the 18 bandwidth/neighborhood/horizon combinations, the Inception retrospective lower half improves mean Inception MMD once, and the DINO retrospective lower half improves mean DINO MMD three times. Upper-score choices improve more often: the DINO forecast upper half improves mean DINO MMD in all 18 combinations, but mean Inception MMD in eight. These are descriptive sensitivity counts, not multiplicity-adjusted discoveries or evidence for selecting the most favorable setting.

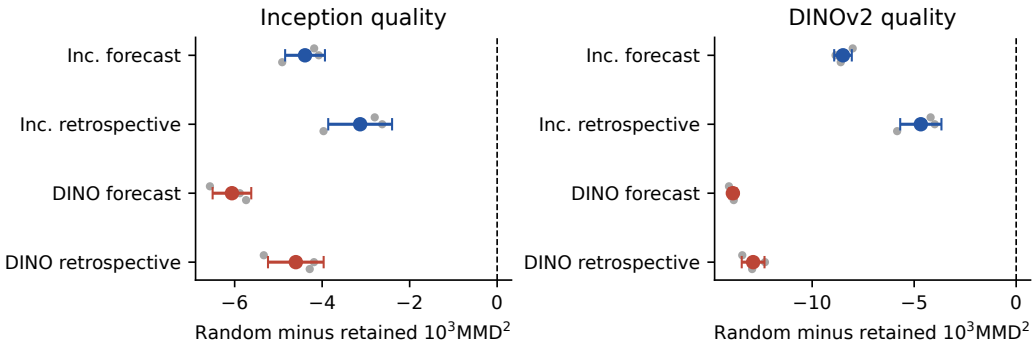

Figure 12: Primary lower-score retention under both scoring and quality representations. Positive gain would beat the identical fixed random half. Bars give sample SD over the same three trained models. The upper-score results are separately reported in Table 1; the lower-direction failure does not imply failure of both directions.

Score features	Selection	Inc. 10^3MMD^2	DINO 10^3MMD^2
—	Random half	6.14 ± 0.45	18.84 ± 0.28
Inception	Forecast	10.53 ± 0.88	27.33 ± 0.26
Inception	Retrospective	9.27 ± 1.15	23.51 ± 0.85
Inception	Pixel velocity	8.41 ± 0.29	21.93 ± 1.02
Inception	Feature velocity	9.88 ± 0.88	23.68 ± 1.11
Inception	Normal velocity	9.73 ± 0.63	23.59 ± 0.63
Inception	Nearest reference	6.81 ± 0.16	20.03 ± 0.99
DINOv2	Forecast	12.20 ± 0.59	32.73 ± 0.43
DINOv2	Retrospective	10.74 ± 0.82	31.74 ± 0.76
DINOv2	Pixel velocity	8.41 ± 0.29	21.93 ± 1.02
DINOv2	Feature velocity	11.59 ± 0.40	27.51 ± 0.50
DINOv2	Normal velocity	11.38 ± 0.44	27.57 ± 0.75
DINOv2	Nearest reference	7.56 ± 0.70	19.53 ± 0.46

Table 12: All lower-score baselines at $T = 0.5, k = 300$, evaluated under both primary feature kernels. Mean $\pm$ sample SD across three seeds. Both halves at every declared setting are included in the machine-readable results.

A separate FID implementation. We use `pytorch-fid 0.3.0`, its TensorFlow-ported Inception weights, and its 2,048-dimensional final pooling block. This encoder differs from the torchvision ImageNet classifier used to score trajectories. Generated pixels are clipped to $[-1, 1]$, mapped to $[0, 255]$, rounded to nearest integer with ties to even, and cast to uint8. Canonical real uint8 pixels are unchanged. Both are divided by 255 before the implementation's default bilinear resize to 299×299 and internal $[-1, 1]$ normalization. Feature inference uses float32 and batch size 64. We use the usual mean/covariance FID (Heusel et al., 2017), with float64 sample covariances normalized by $n - 1$. The reference is the same 10,000-image test set, not a published 50,000-training-image benchmark. Finite-sample FID is biased (Chong & Forsyth, 2020); equal subset sizes control one comparison difference without eliminating this bias.

The covariance cross term is evaluated through the nonzero spectrum of

$$\frac{1}{n-1} H X \widehat{\Sigma}_Q X^\top H, \qquad H = I - \frac{\mathbf{1}\mathbf{1}^\top}{n}, \tag{30}$$

where rows of X are selected generated features. This symmetric matrix has the same nonzero eigenvalues as the covariance product used in FID. We enforce its known centering nullity and check tiny negative eigenvalues before rounding them to zero. For one 1,024-image random half per seed, the result agrees with the public covariance-square-root implementation within 2.59×10^{-5} FID units. SciPy 1.18 removed an older return-signature option; a scoped adapter restores that interface without changing its matrix-square-root calculation. The code, pinned weights, library-source hashes, and adapter are recorded in the protocol. The supplement includes the eigenvalues

Score features	Selection	Lower-half FID	Upper-half FID
Inception	Forecast	45.62 ± 0.93	40.89 ± 0.29
Inception	Retrospective	42.52 ± 1.42	42.54 ± 0.36
Inception	Pixel velocity	41.55 ± 0.43	39.96 ± 0.26
Inception	Feature velocity	43.18 ± 0.83	40.38 ± 0.40
Inception	Normal velocity	43.10 ± 0.76	40.26 ± 0.49
Inception	Nearest reference	39.78 ± 0.54	39.43 ± 0.49
Inception	Random	38.97 ± 0.35	39.60 ± 0.30
DINOv2	Forecast	48.61 ± 0.37	37.55 ± 0.57
DINOv2	Retrospective	46.68 ± 0.19	37.37 ± 0.41
DINOv2	Pixel velocity	41.55 ± 0.43	39.96 ± 0.26
DINOv2	Feature velocity	47.03 ± 0.66	40.91 ± 0.89
DINOv2	Normal velocity	46.76 ± 0.82	40.26 ± 0.77
DINOv2	Nearest reference	40.52 ± 0.08	39.31 ± 0.25
DINOv2	Random	38.97 ± 0.35	39.60 ± 0.30

Table 13: All FID comparisons on the original candidate pool. Each half has 1,024 images, compared with 10,000 real test images; mean $\pm$ sample SD. Full-pool FID uses twice as many generated images and is released only as context, not as an expected random-half value.

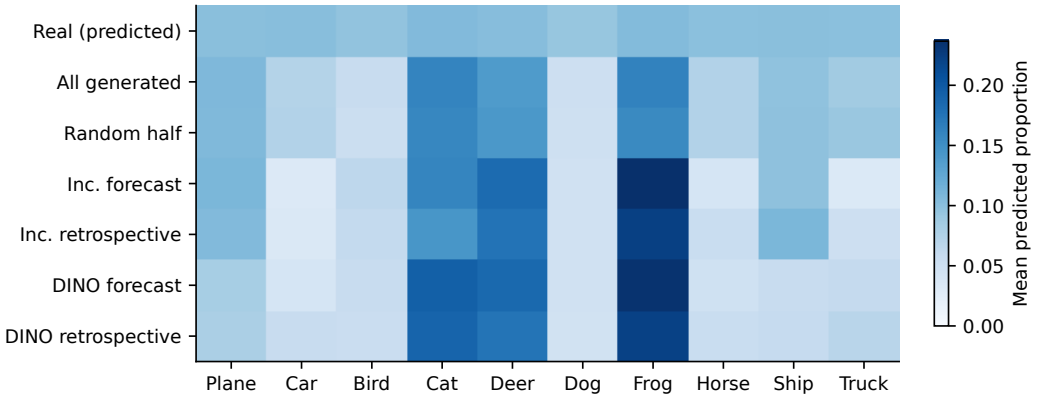

Figure 13: Mean predicted-class proportions for the fixed DINO 20-neighbor classifier. Selected rows use lower-score halves at the primary setting. These are classifier-based composition proxies, not ground-truth semantic coverage.

and scalar terms for arithmetic replay; full covariance re-estimation requires the hash-identified feature arrays, which are not embedded.

Predicted-class composition. A fixed, unweighted cosine 20-neighbor vote on the 50,000 labeled DINO training features predicts CIFAR-10 classes. Vote ties choose the lowest class index. It achieves 97.0% accuracy on all 10,000 test images; the full confusion matrix and votes are included. The same fixed predictor supplies generated-image labels, without tuning on generated quality. We report entropy and total variation from the *predicted* test-class distribution, always using actual equal-sized subsets. The MMD identity for expected random retention does not apply to these nonlinear composition statistics.

The random half's predicted-class TV is 0.155 ± 0.032, compared with 0.278 ± 0.027 for the Inception forecast lower half and 0.299 ± 0.013 for the DINO forecast lower half. Their class histograms become more imbalanced (Figure 13). Conversely, the DINO forecast upper half has TV 0.106 ± 0.011. The classifier shares DINO features with one quality measure, so these are not independent semantic and DINO-quality validations. Its predictions are imperfect, especially for generated images outside the training distribution; they neither establish true generated class labels nor prove that composition causes the metric changes.

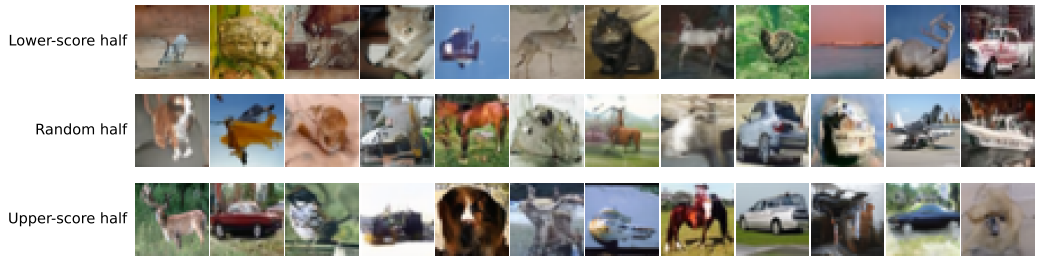

Figure 14: An illustrative view of the DINO forecast subsets at $T = 0.5, k = 300$, using the first declared model seed (42). A fixed independent draw selects twelve images uniformly within each 1,024-image subset; it does not select score extremes or visually preferred examples. The row memberships and draw seed are released. This small post-hoc illustration is not a human quality evaluation.

Score features	Selection	Entropy (nats)	Predicted-class TV
—	Random half	2.23 ± 0.02	0.15 ± 0.03
Inception	Forecast	2.08 ± 0.04	0.28 ± 0.03
Inception	Retrospective	2.13 ± 0.03	0.25 ± 0.02
Inception	Pixel velocity	2.21 ± 0.01	0.18 ± 0.00
Inception	Feature velocity	2.14 ± 0.04	0.24 ± 0.03
Inception	Normal velocity	2.15 ± 0.04	0.23 ± 0.03
Inception	Nearest reference	2.23 ± 0.01	0.14 ± 0.02
DINOv2	Forecast	2.08 ± 0.02	0.30 ± 0.01
DINOv2	Retrospective	2.12 ± 0.01	0.27 ± 0.01
DINOv2	Pixel velocity	2.21 ± 0.01	0.18 ± 0.00
DINOv2	Feature velocity	2.09 ± 0.01	0.29 ± 0.01
DINOv2	Normal velocity	2.10 ± 0.01	0.28 ± 0.01
DINOv2	Nearest reference	2.22 ± 0.01	0.16 ± 0.01

Table 14: Predicted-class composition for all lower-score selectors; mean $\pm$ sample SD. Entropy uses natural logarithms. TV measures discrepancy from the predicted test distribution.

Replay checks. All 117 original Inception subset hashes checked across methods and settings are unchanged. Independent CPU float64 quality replay differs from the original primary values by at most 3.56×10^{-6} units. Every equal-half MMD partition identity holds within 4.45×10^{-13}. The supplement contains DINO final features, all scores, kernel summaries, votes, per-seed measurements, and the complete sensitivity grid. Re-estimating the real-reference kernels requires the identified feature banks; replay of the included summary inputs does not.

F.1 Conditioning on predicted-class counts

This further post-hoc analysis was designed after inspecting both discovery and confirmation outcomes. It uses only the original three discovery pools, preserving $k = 300$, $T = 0.5$, both score backbones, all six nonrandom scores and the original random split, both halves, and the two primary MMD kernels. It does not refit the class predictor or optimize subsets. The exact comparison asks whether the selected examples improve MMD beyond uniform selection with the *same predicted-class counts*. This is a conditional finite-pool comparison, not a causal intervention on true semantic classes.

Let C_c be the predicted-class strata of the full candidate pool P, with sizes N_c. A selected subset S of size m has counts m_c. Independently within each stratum, draw a uniform m_c-element subset without replacement and denote their union by S^*. Write $K_{ij} = k(x_i, x_j)$ for $i \neq j$, $K_{ii} = 0$,

		Inception MMD gain			DINO MMD gain		
Score / features	Half	Total	Class	Within	Total	Class	Within
Forecast / Inc.	low	-4.21 ± 0.67	-2.39 ± 0.28	-1.82 ± 0.42	-8.17 ± 1.00	-6.71 ± 0.54	-1.46 ± 0.96
Forecast / Inc.	high	-3.36 ± 0.50	-0.61 ± 0.48	-2.75 ± 0.11	2.67 ± 0.46	3.34 ± 0.28	-0.67 ± 0.73
Retro. / Inc.	low	-2.95 ± 0.95	-1.38 ± 0.39	-1.57 ± 0.58	-4.35 ± 1.39	-4.17 ± 0.79	-0.18 ± 0.87
Retro. / Inc.	high	-6.46 ± 0.43	-0.26 ± 0.30	-6.20 ± 0.37	0.94 ± 0.68	2.39 ± 0.30	-1.45 ± 0.68
Forecast / DINO	low	-5.88 ± 0.39	-2.78 ± 0.45	-3.11 ± 0.20	-13.56 ± 0.43	-6.76 ± 0.55	-6.81 ± 0.97
Forecast / DINO	high	-0.35 ± 1.05	-0.32 ± 0.32	-0.03 ± 0.74	6.87 ± 0.38	3.31 ± 0.29	3.56 ± 0.27
Retro. / DINO	low	-4.42 ± 0.61	-1.93 ± 0.35	-2.49 ± 0.38	-12.57 ± 0.38	-4.90 ± 0.45	-7.68 ± 0.80
Retro. / DINO	high	0.00 ± 0.95	0.05 ± 0.24	-0.04 ± 0.72	7.07 ± 0.43	2.77 ± 0.17	4.30 ± 0.46

Table 15: Discovery-pool MMD gains relative to expected uniform retention. "Class" is the composition-associated term and "Within" is the residual gain at identical predicted-class counts in Eq. (32). Positive gains are favorable. Mean ± sample SD across three models; the two terms add per model, while their SDs do not. All 168 metric rows, including other scores and random halves, are released.

$q_i = |Q|^{-1} \sum_{y \in Q} k(x_i, y)$, and let κ_Q be the diagonal-excluding real–real kernel mean. Define

$$B_{cd} = \sum_{i \in C_c, j \in C_d} K_{ij}, \qquad q_c = \sum_{i \in C_c} q_i, \qquad p_{cd} = \begin{cases} \dfrac{m_c(m_c-1)}{N_c(N_c-1)}, & c = d, \\ \dfrac{m_c m_d}{N_c N_d}, & c \neq d. \end{cases}$$

An empty stratum has $m_c = 0$ and zero inclusion weights; the diagonal weight is zero for $N_c < 2$. Linearity of expectation and the point/pair inclusion probabilities give

$$M_{\text{cond}} := \mathbb{E}[\widehat{M}(S^*, Q) \mid (m_c)_c] = 10^3 \left\{ \frac{\sum_{c,d} p_{cd} B_{cd}}{m(m-1)} + \kappa_Q - \frac{2}{m} \sum_c \frac{m_c}{N_c} q_c \right\}. \tag{31}$$

The c, d sum uses ordered blocks; the denominator is the total subset's $m(m-1)$. No Monte Carlo sampling is needed. Since $\widehat{M}(P, Q)$ equals the expected statistic of a uniform random subset of size m, the total gain decomposes exactly as

$$\widehat{M}(P,Q) - \widehat{M}(S,Q) = \underbrace{\widehat{M}(P,Q) - M_{\text{cond}}}_{\text{composition-associated}} + \underbrace{M_{\text{cond}} - \widehat{M}(S,Q)}_{\text{residual at fixed class counts}}. \tag{32}$$

Both terms can be negative. The residual changes which examples are selected within the fixed quotas, including their *cross-class* kernel interactions; it is not a sum of classwise MMD values. These terms are descriptive differences, not causal effects or percentages explained. The reference here is *expected* random retention, distinct from the particular fixed random half used in Table 1.

The DINO forecast upper half improves expected DINO MMD by 6.87 ± 0.38: 3.31 ± 0.29 is composition-associated and 3.56 ± 0.27 remains at fixed class counts. All three residual gains are positive. Under Inception MMD, the corresponding residual is -0.03 ± 0.74, positive in one seed and negative in two. For the DINO retrospective upper half, the residual DINO gain is 4.30 ± 0.46, but its residual Inception gain is -0.04 ± 0.72. Thus class counts alone do not account for the DINO gains, while conditioning does not yield a metric-independent improvement. All four primary lower-score selectors retain negative mean residual gains under both kernels (Table 15).

The class predictions share DINO features with one evaluation kernel and may mislabel generated samples; conditioning does not remove this dependence. No claim about true class balance, human realism, or causal mediation follows. FID is excluded from this linear-expectation calculation, and these results are not an additional confirmation on new samples. Exhaustive enumeration checks 40 toy count configurations covering 114 subsets, including empty/singleton strata; uniform and unique-label reductions also pass. All 168 selected metrics reproduce the frozen discovery records. The supplement includes class-block sums, cross sums, counts, membership hashes, and the public features needed to recompute the kernels.

F.2 Independent-noise confirmation

After inspecting the complete discovery results above, we freeze a second evaluation on new initial noise: 2,048 candidates for each of the same three 400k-step models, for another 6,144 trajecto-

Score	Half	Inc. $10^3\mathrm{MMD}^2$	DINO $10^3\mathrm{MMD}^2$	FID
Random	Matched	6.41 ± 0.30	18.68 ± 1.44	39.74 ± 0.94
Inc. forecast	Lower	10.21 ± 0.60	26.53 ± 0.72	46.26 ± 0.86
Inc. forecast	Upper	9.80 ± 1.16	16.52 ± 1.27	41.09 ± 0.61
Inc. retro.	Lower	8.20 ± 0.43	21.89 ± 0.58	42.13 ± 0.68
Inc. retro.	Upper	13.73 ± 0.61	18.71 ± 1.44	43.60 ± 0.36
DINO forecast	Lower	12.03 ± 1.21	31.50 ± 2.46	49.46 ± 1.72
DINO forecast	Upper	7.33 ± 1.02	13.04 ± 0.24	38.47 ± 0.85
DINO retro.	Lower	11.02 ± 1.22	31.86 ± 3.09	48.24 ± 1.84
DINO retro.	Upper	6.85 ± 0.76	12.22 ± 0.44	38.04 ± 0.32

Table 16: Independent-noise confirmation: 6,144 new candidates from the same three trained models, with 1,024 retained per model. Mean $\pm$ sample SD. Both directions and all three metrics were fixed before new generation; all additional baselines are included in the released records.

ries. The CPU noise seed is the training seed plus 40,260,925, disjoint from the original seeds and initial-tensor hashes. A new matched random split uses the training seed plus 314,159. Generation, encoders, preprocessing, score definitions, $k = 300$, $T = 0.5$, both score directions, all six nonrandom methods, and the primary MMD bandwidths and FID definition are fixed before this generation. No confirmation outcome chooses a new setting. The same real references are reused, but all generated–real and generated–generated kernels are recomputed; only real-only kernel means and reference features are reused. This is a test over new candidate noise conditional on the same trained models and test references, not an independent-model or independent-dataset replication.

Every lower forecast/retrospective half again loses to matched random retention under all three metrics in every model. The reversed DINO forecast yields Inception MMD 7.33 ± 1.02 versus random 6.41 ± 0.30, DINO MMD 13.04 ± 0.24 versus 18.68 ± 1.44, and FID 38.47 ± 0.85 versus 39.74 ± 0.94. The direction of the three mean contrasts repeats the discovery pattern. In seed order 42, 123, and 7, the FID gains are $0.835, -0.018, 2.996$, with mean 1.27 ± 1.55 sample SD; thus this benefit does not occur in every confirmation pool. The corresponding Inception MMD gains are $-0.638, -2.332, 0.214$, while DINO MMD improves in all three. We do not interpret three reused models as establishing a general sampling policy.

The released score arrays reproduce every fixed subset. Compact generated-kernel sums, generated–real means, and real-only means replay MMD arithmetic; FID eigenvalues and scalar terms replay its covariance formula. Re-estimating kernels and covariances requires the omitted, hash-identified capture/feature arrays. Independent CPU reconstruction from these features reproduces random and DINO-forecast half metrics within 2.21×10^{-13} MMD units and 2.0×10^{-12} FID units. This checks full kernels and selected covariance terms, not a new image encoding or PCA run. The public covariance-square-root comparison agrees within 2.60×10^{-5} FID units on all three random halves. These records preserve both the initial discovery and its subsequent confirmation without merging them into additional independent training seeds.

F.3 Isolating projection on the same feature chord residual

The original baselines vary the residual as well as its projection. We therefore add a bounded component analysis that changes only the projector. Its protocol was recorded before measuring the new comparisons, after the original discovery and confirmation results had been inspected. Both pools are reused; this is a post-hoc analysis, not another independent confirmation experiment.

For each model, representation and residual definition, we retain the same 2,048 candidates, four times $\{0, 0.10, 0.26, 0.50\}$, trapezoidal aggregation, and 1,024-image halves. The residual vectors are

$$e_{t,i}^{\mathrm{forecast}} = U_{t,i}^{\epsilon} - [F(\widetilde{x}_{1,i}(t)) - F(x_{0,i})], \qquad e_{t,i}^{\mathrm{retro}} = U_{t,i}^{\epsilon} - [F(x_{1,i}) - F(x_{0,i})].$$

We preserve the captured float32 subtractions, clipping and $\epsilon = 0.01$ lifts. We compare the frozen PCA norm $\left\| \widehat{P}_{t,i} e_{t,i} \right\|$, the unprojected norm $\|e_{t,i}\|$, and ten random-projector norms. For repetition b, draw a Gaussian matrix and take its thin QR factor $Q^{(b)}$; with the frozen local rank $r_{t,i}$, use

$$P_{t,i}^{(b)} = Q_{[:,1:r_{t,i}]}^{(b)} Q_{[:,1:r_{t,i}]}^{(b)\top}.$$

Each frame is shared across models, pools, times and residual definitions within a representation. The seeds are $2026092600 + b$ for Inception and $2026093600 + b$ for DINO, $b = 0, \ldots, 9$. This isolates directional alignment conditional on the observed rank; it does not remove all PCA information. There is no dimension/rank rescaling. Ranks range from 175 to 199 for Inception and 142 to 158 for DINO. Random and unprojected norms use float64 arithmetic on the same captured residuals; frozen PCA norms remain unchanged.

Stable sorting defines both halves. Quality uses the identical endpoint features and 10,000 real references from the corresponding original analyses: raw Inception Gaussian MMD ($\sigma = 10$), unit-normalized DINO MMD ($\sigma = 1$), and separately encoded FID. Both MMD quantities are diagonal-excluding squared statistics multiplied by 1,000. Gains subtract selected-subset quality from the same fixed random-half quality, so positive values are favorable. FID always compares equal-sized subsets. Recomputing the unchanged PCA selections reproduces their memberships exactly and their reported MMD/FID values within 2.3×10^{-13} and 4.0×10^{-12}, respectively.

Table 17 shows the confirmation-pool rank comparison and FID gains; Tables 18–21 give all crossed quality results. Random-frame results average the ten frames within each trained model before computing the mean and sample SD across three models. The supplement separately records each frame, each model, and within-model frame SD. Frames and the two noise pools are not additional independent training replications.

Score	Residual	Projection	ρ	Overlap	ΔFID L	ΔFID H
Inception	Forecast	PCA	1.000	100.0%	-6.52 ± 0.78	-1.34 ± 1.03
Inception	Forecast	None	0.943	89.6%	-6.57 ± 1.00	-1.56 ± 1.05
Inception	Forecast	Random rank	0.917	87.2%	-6.01 ± 0.93	-1.39 ± 1.01
Inception	Retro.	PCA	1.000	100.0%	-2.38 ± 1.24	-3.86 ± 0.58
Inception	Retro.	None	0.944	89.9%	-3.69 ± 1.30	-3.27 ± 0.84
Inception	Retro.	Random rank	0.912	86.7%	-3.51 ± 1.29	-3.15 ± 0.85
DINO	Forecast	PCA	1.000	100.0%	-9.72 ± 1.54	1.27 ± 1.55
DINO	Forecast	None	0.979	93.4%	-12.13 ± 1.92	0.77 ± 1.34
DINO	Forecast	Random rank	0.950	90.6%	-11.62 ± 1.85	0.71 ± 1.27
DINO	Retro.	PCA	1.000	100.0%	-8.49 ± 1.49	1.70 ± 1.18
DINO	Retro.	None	0.939	88.8%	-13.83 ± 1.97	1.27 ± 1.41
DINO	Retro.	Random rank	0.896	85.6%	-12.54 ± 2.20	1.34 ± 1.35

Table 17: Post-hoc projector comparison on the existing confirmation pools. ρ is model-mean Spearman correlation with PCA; overlap is the retained-half intersection divided by 1,024. L/H denote lower/upper halves. "Random rank" means random orientation at the frozen pointwise PCA rank. Positive FID gains beat the matched random half. Random-frame means precede the three-model SD.

Across both pools and residual definitions, removing PCA leaves model-mean Spearman correlations 0.938–0.979 and half overlaps 88.4–94.4%. Mean lower-half losses persist under all three metrics. Upper DINO selection retains mean DINO-MMD and FID benefits while worsening Inception MMD. These patterns also occur after averaging the rank-matched random orientations. Thus the qualitative selection findings do not require local PCA orientation. Projection can still change magnitudes and the ranking of methods: on confirmation DINO forecasts, upper-half FID gains are 1.27 (PCA), 0.77 (none), and 0.71 (random), whereas DINO-MMD gains are 5.63, 6.22, and 5.95. This metric dependence does not support a universal quality advantage for any projector or an intrinsic interpretation of its score.

The evaluator and frozen protocol are `evaluate_projector_ablation.py` and `results/projector_ablation/protocol.json`. Compact replay reconstructs score ordering, overlap, correlations, kernel-sum arithmetic, covariance-spectrum arithmetic and tables. Regenerating residual vectors, projections, kernels or covariances requires the separately identified large captures and encoded features. No additional model training, image generation, or feature encoding was used.

Score	Residual	Projection	ΔIMMD	ΔDINO MMD	ΔFID
Inception	Forecast	PCA	-4.39 ± 0.46	-8.49 ± 0.43	-6.65 ± 0.63
Inception	Forecast	None	-4.73 ± 0.38	-7.97 ± 0.30	-6.26 ± 0.40
Inception	Forecast	Random rank	-4.61 ± 0.42	-7.85 ± 0.37	-6.14 ± 0.40
Inception	Retro.	PCA	-3.13 ± 0.73	-4.67 ± 1.01	-3.55 ± 1.06
Inception	Retro.	None	-3.90 ± 0.34	-5.55 ± 0.37	-4.20 ± 0.47
Inception	Retro.	Random rank	-3.69 ± 0.45	-5.36 ± 0.52	-4.11 ± 0.54
DINO	Forecast	PCA	-6.06 ± 0.44	-13.89 ± 0.17	-9.65 ± 0.48
DINO	Forecast	None	-7.20 ± 0.53	-16.12 ± 0.61	-11.45 ± 0.14
DINO	Forecast	Random rank	-7.09 ± 0.66	-15.64 ± 0.40	-11.33 ± 0.36
DINO	Retro.	PCA	-4.60 ± 0.64	-12.89 ± 0.56	-7.72 ± 0.49
DINO	Retro.	None	-7.70 ± 0.75	-21.41 ± 1.05	-13.29 ± 0.32
DINO	Retro.	Random rank	-6.86 ± 0.63	-19.63 ± 0.79	-11.84 ± 0.39

Table 18: Discovery pools, lower-half gains over matched random retention. Positive is favorable. Random projections preserve the original pointwise ranks. Mean ± sample SD across three models, with random frames first averaged within each model.

Score	Residual	Projection	ΔIMMD	ΔDINO MMD	ΔFID
Inception	Forecast	PCA	-3.53 ± 0.38	2.35 ± 0.94	-1.93 ± 0.25
Inception	Forecast	None	-2.78 ± 0.27	2.22 ± 0.77	-2.24 ± 0.17
Inception	Forecast	Random rank	-2.42 ± 0.28	2.37 ± 0.69	-1.92 ± 0.11
Inception	Retro.	PCA	-6.64 ± 0.23	0.62 ± 1.22	-3.58 ± 0.57
Inception	Retro.	None	-4.56 ± 0.04	1.15 ± 0.90	-3.58 ± 0.21
Inception	Retro.	Random rank	-4.48 ± 0.15	1.16 ± 0.85	-3.34 ± 0.12
DINO	Forecast	PCA	-0.52 ± 1.08	6.55 ± 0.53	1.42 ± 0.74
DINO	Forecast	None	-1.00 ± 0.89	6.89 ± 0.47	0.99 ± 0.64
DINO	Forecast	Random rank	-1.00 ± 0.89	6.60 ± 0.40	0.86 ± 0.50
DINO	Retro.	PCA	-0.17 ± 0.97	6.75 ± 0.64	1.60 ± 0.53
DINO	Retro.	None	-1.10 ± 1.15	9.09 ± 0.60	1.24 ± 0.93
DINO	Retro.	Random rank	-0.71 ± 0.87	8.83 ± 0.50	1.32 ± 0.72

Table 19: Discovery pools, upper-half gains. Definitions and aggregation match Table 18; both score directions are reported.

Score	Residual	Projection	ΔIMMD	ΔDINO MMD	ΔFID
Inception	Forecast	PCA	-3.80 ± 0.44	-7.85 ± 0.78	-6.52 ± 0.78
Inception	Forecast	None	-4.33 ± 0.37	-7.71 ± 0.95	-6.57 ± 1.00
Inception	Forecast	Random rank	-3.93 ± 0.36	-7.01 ± 0.71	-6.01 ± 0.93
Inception	Retro.	PCA	-1.79 ± 0.39	-3.22 ± 0.96	-2.38 ± 1.24
Inception	Retro.	None	-2.88 ± 0.66	-4.71 ± 1.30	-3.69 ± 1.30
Inception	Retro.	Random rank	-2.67 ± 0.56	-4.39 ± 0.98	-3.51 ± 1.29
DINO	Forecast	PCA	-5.61 ± 1.11	-12.82 ± 1.53	-9.72 ± 1.54
DINO	Forecast	None	-7.02 ± 1.32	-15.80 ± 1.91	-12.13 ± 1.92
DINO	Forecast	Random rank	-6.75 ± 1.18	-15.04 ± 1.79	-11.62 ± 1.85
DINO	Retro.	PCA	-4.61 ± 1.08	-13.18 ± 2.23	-8.49 ± 1.49
DINO	Retro.	None	-7.48 ± 1.28	-21.23 ± 1.85	-13.83 ± 1.97
DINO	Retro.	Random rank	-6.72 ± 1.35	-19.48 ± 2.37	-12.54 ± 2.20

Table 20: Existing confirmation pools, lower-half gains. The models and real references are reused; this new component analysis follows inspection of the pools.

Score	Residual	Projection	ΔIMMD	ΔDINO MMD	ΔFID
Inception	Forecast	PCA	-3.39 ± 1.43	2.16 ± 0.29	-1.34 ± 1.03
Inception	Forecast	None	-2.51 ± 1.14	2.29 ± 0.43	-1.56 ± 1.05
Inception	Forecast	Random rank	-2.30 ± 1.10	2.19 ± 0.52	-1.39 ± 1.01
Inception	Retro.	PCA	-7.31 ± 0.83	-0.03 ± 0.71	-3.86 ± 0.58
Inception	Retro.	None	-4.96 ± 0.70	0.89 ± 0.86	-3.27 ± 0.84
Inception	Retro.	Random rank	-4.75 ± 0.77	0.80 ± 0.71	-3.15 ± 0.85
DINO	Forecast	PCA	-0.92 ± 1.30	5.63 ± 1.29	1.27 ± 1.55
DINO	Forecast	None	-1.40 ± 1.28	6.22 ± 1.03	0.77 ± 1.34
DINO	Forecast	Random rank	-1.38 ± 1.12	5.95 ± 1.12	0.71 ± 1.27
DINO	Retro.	PCA	-0.44 ± 1.01	6.46 ± 1.22	1.70 ± 1.18
DINO	Retro.	None	-1.42 ± 1.26	8.53 ± 1.17	1.27 ± 1.41
DINO	Retro.	Random rank	-1.02 ± 1.02	8.24 ± 1.33	1.34 ± 1.35

Table 21: Existing confirmation pools, upper-half gains. Positive values favor selection over its matched random half. Variability is across three models, not across individual samples or projector frames.

Backbone	Lift	Score ratio	Spearman	Agreement (%)
Inception	$\epsilon = 0.02$	1.0001 ± 0.0001	0.9951 ± 0.0007	96.35 ± 1.63
Inception	$\epsilon = 0.01$	1.0000 ± 0.0000	1.0000 ± 0.0000	100.00 ± 0.00
Inception	$\epsilon = 0.005$	1.0002 ± 0.0004	0.9975 ± 0.0002	97.66 ± 0.00
Inception	$\epsilon = 0.0025$	1.0004 ± 0.0004	0.9948 ± 0.0006	96.61 ± 0.45
Inception	AD JVP, clipped mask	1.0008 ± 0.0006	0.9915 ± 0.0005	96.88 ± 0.00
Inception	AD JVP, no mask	1.0704 ± 0.0044	0.7544 ± 0.0579	78.39 ± 2.39
DINOv2	$\epsilon = 0.02$	1.0001 ± 0.0005	0.9984 ± 0.0001	98.44 ± 0.00
DINOv2	$\epsilon = 0.01$	1.0000 ± 0.0000	1.0000 ± 0.0000	100.00 ± 0.00
DINOv2	$\epsilon = 0.005$	0.9998 ± 0.0001	0.9992 ± 0.0001	99.22 ± 0.78
DINOv2	$\epsilon = 0.0025$	0.9997 ± 0.0001	0.9985 ± 0.0001	99.22 ± 0.00
DINOv2	AD JVP, clipped mask	0.9996 ± 0.0002	0.9974 ± 0.0002	98.70 ± 0.45
DINOv2	AD JVP, no mask	1.5347 ± 0.0040	0.4586 ± 0.0324	66.67 ± 1.97

Table 22: Measurement sensitivity on the fixed audit subsets. Ratios, Spearman correlations, and lower-half membership agreement are relative to the original $\epsilon = 0.01$ scores within each audited model pool. Mean $\pm$ sample SD across three reused models; every declared variant is shown. These are stability measures, not quality gains.

F.4 Finite differences and the clipping derivative mask

This bounded post-hoc audit was designed after inspecting all discovery/confirmation outcomes and the subsequent review. It reuses every eighth candidate index, starting at zero, from each original 2,048-candidate discovery pool: 256 trajectories for each of the same three models. Both scoring backbones and all four prefix times $\{0, 0.10, 0.26, 0.50\}$ are retained. Candidate indices, steps, interventions, and summaries were fixed before these measurements. No new training, generation, or quality evaluation is performed.

The raw captured image state x and velocity v permit changes to the feature lift while holding the generator fixed. Let $y = C(x)$ and $M_x = \mathrm{diag}(\mathbf{1}_{|x_i| \le 1})$. We compare the original forward differences at $\epsilon \in \{0.02, 0.01, 0.005, 0.0025\}$ with two automatic-differentiation Jacobian–vector products:

$$J_{\mathrm{clip}} = D\Phi(y)[M_x v/2], \qquad J_{\mathrm{unmasked}} = D\Phi(y)[v/2].$$

Both use the same clipped base input y, encoder weights, preprocessing, forecast comparator, initial features, neighbors, and numerical PCA projector. Only the derivative mask differs between the two JVPs. The implementation uses its standard derivatives at nonsmooth operations, float32 arithmetic, math attention for differentiation, and disabled TF32. The unmasked JVP is an artificial local measurement intervention; it is not evaluation of an unclipped image pipeline and is not proposed as a replacement encoder.

Before intervention, re-encoding at $\epsilon = 0.01$ reproduces the saved state features and lifts, and recomputed neighborhood ranks and scores reproduce the frozen measurements within the declared tolerances. The JVP forward values are separately checked against the stored features. Scores integrate the same four prefix times and select 128 of the 256 audited candidates. Thus membership agreement is conditional on this audit subset, not on the original full-pool retention experiment.

Automatic differentiation with the clipping mask retains mean lower-half agreements of $96.88 \pm 0.00\%$ for Inception and $98.70 \pm 0.45\%$ for DINO. Removing only the derivative mask at the same input yields $78.39 \pm 2.39\%$ and $66.67 \pm 1.97\%$ agreement with the original ranking. Corresponding Spearman correlations change from 0.9915 ± 0.0005 to 0.7544 ± 0.0579 and from 0.9974 ± 0.0002 to 0.4586 ± 0.0324. These observations separate sensitivity to the derivative mask from the small-step comparison. They do not establish that clipping causes the adverse quality results or that the unmasked ranking improves quality.

The released arrays also report clipping occupancy, removed pixel-velocity energy, step-dependent mask crossings, neighborhood ranks, and the query's distance from its neighborhood mean relative to neighborhood RMS radius. Occupancy alone is not used as evidence of a causal contribution. Figure 15 shows the projected lift comparisons and ranking stability for every model separately.

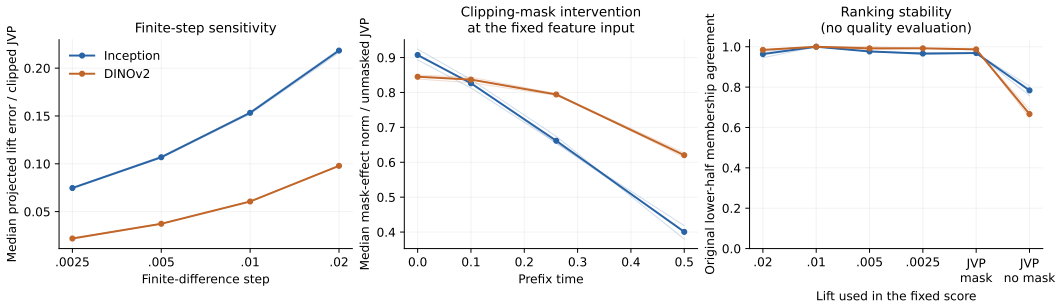

Figure 15: Fixed-input measurement interventions. Left: median projected finite-difference lift error relative to the clipped JVP. Middle: the change from removing the derivative mask, normalized by the unmasked projected lift. Right: within-audit lower-half agreement with the original finite-step scores. Thin lines show each reused model; thick lines average their summaries. No panel measures selection utility or intrinsic geometry.

PCA numerical sensitivity. The original validation required float32 Gram PCA and float64 direct SVD to select the same minimum rank. One of 24 predeclared checked points failed: float64 selects rank 183, with cumulative variance 0.9500010188, while float32 selects 184. The latter's ratio at rank 183 is 0.9499996901. A recorded amendment preserves this failed check and freezes the original numerical projectors for all interventions, while adding fixed-rank SVD comparisons and threshold sensitivity measurements. Across these 24 points, the maximum fixed-rank score difference is 1.68×10^{-6} relatively; using each arithmetic's threshold rank changes the checked score by at most 0.0350%. At the changed rank, projector operator distance is approximately one, so this small observed score change is not a uniform guarantee over arbitrary velocity vectors. The original quality measurements are retained, and no universal minimum-rank agreement is claimed. Protocols, amendment, numerical checks, compact arrays, and summary replay are included in the supplement.

G Fresh early-training protocol

The new dense trajectories use seeds 42, 123, and 7, with snapshots at iterations $0, 1000, 2000, 5000, 10000, 20000, 50000$. Each run stops at 50k while preserving the archived 400k cosine horizon and 2k warmup; it does not compress the schedule into a 50k budget. The UNet architecture, batch size 128, AdamW learning rate 2×10^{-4}, zero weight decay, gradient clipping at 1, exact uniform minibatch OT objective, and EMA decay 0.9999 match the archived configuration. The fresh iterator independently shuffles the 50,000 CIFAR-10 training images without replacement, drops the last incomplete batch, and applies horizontal flips. It has a separate CPU random generator initialized with the training seed plus 710,003. These are new replications with recorded random states, not recovered prefixes of the archived runs.

Training uses float16 autocast and a GradScaler with float32 parameters and optimizer state. Logged training iterations include any AMP-skipped optimizer updates; the actual optimizer-step count is recorded separately. EMA updates follow each training iteration. The resumable `latest.pt` checkpoint preserves the model, EMA, optimizer, scaler, global random states, data-generator state, permutation, and cursor. Evaluated snapshots retain the weights and configuration. cuDNN benchmarking is enabled, so resumption state does not imply bitwise deterministic training across independent runs. Protocol, source, dataset, and evaluated-checkpoint hashes guard provenance.

An execution amendment, declared before the second and third seeds started and before any new diagnostic or generated-quality results, assigns seed 42 to an RTX 3080 on Windows and seeds 123 and 7 to separate RTX 4090 GPUs on Linux. Source bytes, data hashes, and numerical package versions are matched; Python patch versions differ (3.13.2 versus 3.13.12). A second execution-only amendment, also declared before any new diagnostic or generated-quality measurement, assigns all 21 checkpoint evaluations to the same Linux/RTX 4090 environment after training releases those GPUs. Its two workers share one prepared diagnostic cache and must reproduce 32 canonical test features within the declared tolerance before evaluation. It changes no evaluator source or numer-

ical setting. Both original protocols, amendments, per-worker environments, and result hashes are retained. With one training run per seed, training hardware and seed effects are not separately identifiable; no hardware comparison is claimed.

Fixed diagnostic probes. A separate seed 20,260,926 determines 512 Gaussian inputs and shuffled/flipped real training targets. Exact pixel-distance OT pairs them in batches of 64, matching the diagnostic pairing convention rather than the training batch size. These same pairs probe every seed and checkpoint at ten times $0.05, 0.15, \ldots, 0.95$. They are fixed *training* probes, not held-out generalization tests. State features, endpoint feature chords, exact $k = 300$ neighborhoods in the matched 50,000-image training bank, and centered 95%-variance PCA bases are computed once and reused. Raw and EMA weights are evaluated separately using float32 inference, batch size 32, and no autocast or TF32. The primary curve uses EMA. The EMA parameter average assigns coefficient 0.9999^n to its initialization after n iterations (about 0.905 at 1k and 0.368 at 10k), making the raw-weight comparison necessary for interpreting early changes.

An operational decomposition. For a fixed pair, let $\Delta = x_1 - x_0$, $c_t = (1-t)x_0 + tx_1$, and define the finite-difference pair lift $U_t^{\epsilon,\text{pair}} = \{F(c_t + \epsilon\Delta) - F(c_t)\}/\epsilon$. With the same cached projector and $\epsilon = 0.01$, write

$$r_t = \widehat{P}_t(U_t^{\epsilon} - U_t^{\epsilon,\text{pair}}), \qquad \beta_t = \widehat{P}_t\{U_t^{\epsilon,\text{pair}} - [F(x_1) - F(x_0)]\}.$$

The diagnostic is exactly $\|r_t + \beta_t\|$. We separately record $\|r_t\|$, the paired-target feature residual, and $\|\beta_t\|$, a fixed representation/path baseline. These are vectors before taking norms: the reported scalar curves are not an additive decomposition. The baseline does not depend on the learned field, and is checked for equality across all evaluations. The residual includes paired-target variation and nonlinear representation effects, so it is not identified as pure optimization error. Neither component measures intrinsic manifold error. All curves average per-pair norms and then integrate by the trapezoidal rule over $[0.05, 0.95]$. Pixel velocity, pixel paired-target residual, and lifted feature-velocity norms are retained as additional diagnostics.

Separate generated quality. Each EMA checkpoint generates 2,048 fresh samples with 50-step float32 Euler integration. A CPU noise generator uses the training seed plus 20,260,925; its initial tensor is identical across the seven checkpoints within that seed. Quality is measured against the same matched 10,000 real test features using the specified Gaussian IMMD. These generated ODE paths are distinct from the fixed linear-interpolant diagnostic probes. Raw-weight diagnostic correlations, when reported, also use this EMA-generated quality series; no raw-weight generation experiment is implied. Checkpoints within each seed are dependent, so correlations are descriptive and do not carry independent-checkpoint significance tests. Adjacent changes span unequal iteration intervals and are neither transition rates nor formal change-point estimates.

Completed trajectories and interpretation. All three runs reached 50k iterations, and all 21 checkpoints were evaluated, generating 43,008 images for the quality curves. Table 23 gives seed means and sample SD; Table 24 gives every checkpoint. The raw diagnostic's largest observed adjacent change occurs between initialization and 1k in every seed; the EMA curve's occurs between 10k and 20k. These are comparisons of the recorded intervals, not estimated change points. The raw paired-target feature residual falls from 52.00 ± 0.60 initially to 27.04 ± 0.11 at 1k and 18.17 ± 0.04 at 50k. Its pixel-space counterpart falls from 56.30 ± 0.77 to 20.42 ± 0.12 at 1k. Consequently, the rising raw diagnostic is not a rising paired-target residual. Parameter averaging delays the score response under otherwise identical probes.

Descriptive Spearman correlations with the EMA-generated IMMD sequence are $-0.893, -0.964, -0.964$ for the EMA diagnostic and $-0.964, -1.000, -1.000$ for the raw diagnostic, in seed order 42, 123, and 7. These sequences largely track training progress and do not establish a within-checkpoint quality ranking. The latter is tested separately in Section 6.4.

Figure 16 separates time-resolved diagnostic and residual profiles for both weight choices. The fixed baseline itself increases from 20.28 at $t = 0.05$ to 198.24 at $t = 0.95$, with integrated value 51.14. It therefore already contains a strong endpoint rise without a learned field. This does not assign that component to intrinsic curvature: the baseline contains feature-map, clipping, finite-difference, and path effects. The norms describe vectors that can cancel; no scalar additive decomposition or strictly positive learned-error floor is inferred.

Iteration	EMA score	Raw score	EMA residual	Baseline	IMMD
0	18.68 ± 0.24	18.68 ± 0.24	52.00 ± 0.60	51.14 ± 0.00	729.09 ± 3.08
1,000	18.39 ± 0.20	37.98 ± 0.33	50.12 ± 0.55	51.14 ± 0.00	718.71 ± 4.45
2,000	18.79 ± 0.21	41.62 ± 0.46	46.98 ± 0.45	51.14 ± 0.00	695.02 ± 9.61
5,000	24.94 ± 0.26	45.23 ± 0.36	37.29 ± 0.20	51.14 ± 0.00	599.14 ± 16.93
10,000	31.17 ± 0.22	46.26 ± 0.17	31.03 ± 0.29	51.14 ± 0.00	503.38 ± 19.91
20,000	39.45 ± 0.01	46.71 ± 0.55	22.92 ± 0.06	51.14 ± 0.00	239.79 ± 1.72
50,000	47.43 ± 0.09	47.93 ± 0.36	18.16 ± 0.02	51.14 ± 0.00	6.64 ± 0.17

Table 23: Fresh early-training results; mean $\pm$ sample SD across three seeds. Residual denotes the paired-target feature residual of EMA weights. Baseline is fixed across all checkpoints. IMMD uses separate EMA-generated images and a matched test bank.

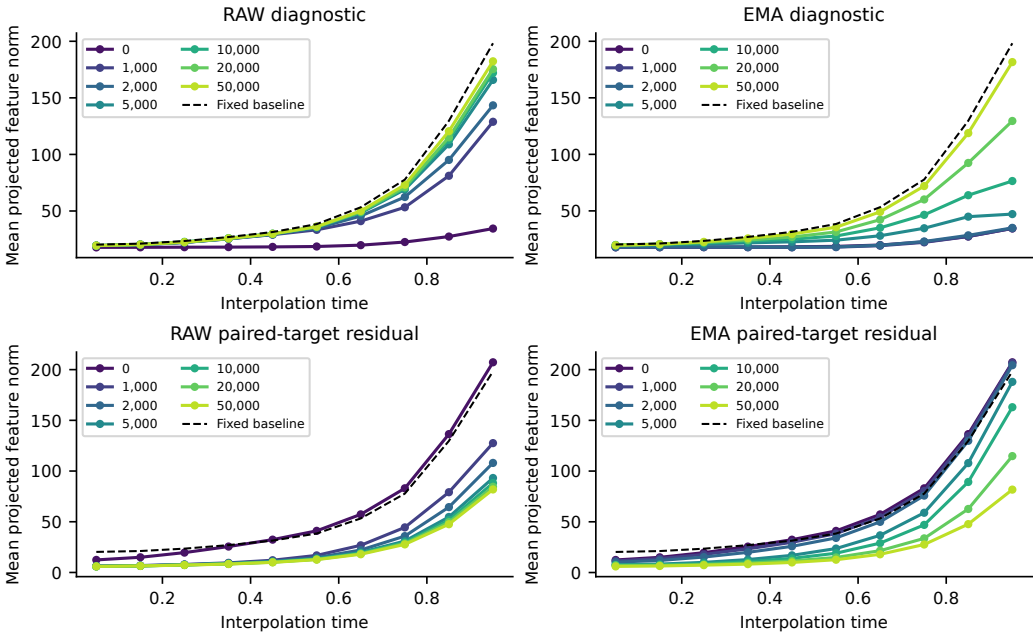

Figure 16: Three-seed mean profiles along the fixed linear training interpolants. Colors denote training iterations. Top: the total feature diagnostic. Bottom: paired-target feature residual. The dashed representation/path baseline is identical in every run; it is included for comparison and is not added to the scalar residual curve.

Execution and numerical checks. Both server workers reproduce the 32 reference-feature checks with relative error 1.53×10^{-6} and maximum coordinate error 8.11×10^{-6}. Stored per-pair norms replay the reported diagnostic aggregates exactly; the fixed baseline has zero recorded change, and all norm triangle/reverse-triangle checks pass. The maximum recorded float32 coordinate discrepancy in the vector decomposition is 5.04×10^{-4}. Independent CPU generated-kernel calculations differ from the GPU IMMD values by at most 1.21×10^{-4} units. All seven checkpoints within each seed use identical initial-quality-noise hashes. Per-checkpoint execution records bind the checkpoint, frozen evaluator, shared cache, worker environment, and result files; the supplement includes compact kernel sums for arithmetic replay.

Seed	Iteration	EMA score	Raw score	EMA residual	Pixel residual	IMMD
42	0	18.950	18.950	52.612	57.183	729.900
42	1,000	18.543	37.612	50.719	52.887	720.295
42	2,000	18.787	42.045	47.499	45.715	696.760
42	5,000	25.105	45.633	37.364	27.398	601.349
42	10,000	31.385	46.286	30.802	22.814	500.175
42	20,000	39.467	46.077	22.977	19.902	238.696
42	50,000	47.327	47.549	18.190	17.779	6.451
123	0	18.617	18.617	51.417	55.898	731.676
123	1,000	18.477	38.241	49.648	51.874	722.151
123	2,000	19.006	41.141	46.756	45.328	703.636
123	5,000	24.640	44.967	37.441	27.947	614.861
123	10,000	30.950	46.419	31.359	23.151	524.698
123	20,000	39.454	47.039	22.861	19.898	238.900
123	50,000	47.513	47.991	18.155	17.774	6.679
7	0	18.482	18.482	51.965	55.805	725.682
7	1,000	18.163	38.101	50.005	51.616	713.690
7	2,000	18.583	41.687	46.693	44.732	684.657
7	5,000	25.073	45.085	37.065	27.333	581.219
7	10,000	31.175	46.088	30.935	22.721	485.263
7	20,000	39.441	47.025	22.928	19.882	241.769
7	50,000	47.444	48.251	18.148	17.772	6.794

Table 24: All 21 fresh checkpoints. The feature and pixel residual columns both use EMA weights; the raw column reports only the raw-weight diagnostic. IMMD is computed from 2,048 generated samples per row.

G.1 Algebraic cancellation across training

The small initial diagnostic can be examined using the recorded vector identity $r_t + \beta_t$. A post-hoc protocol includes all 512 pairs at all ten times, all seven checkpoints, all three seeds, and both weight choices. Let $\langle\cdot\rangle$ average over pairs and integrate with the existing nonnegative trapezoidal weights on $[0.05, 0.95]$. Define

$$T = \langle \|r_t + \beta_t\|^2 \rangle, \quad R = \langle \|r_t\|^2 \rangle, \quad A = \langle \|\beta_t\|^2 \rangle,$$
$$X = \frac{T - R - A}{2} = \langle r_t^T \beta_t \rangle, \qquad C = \frac{-2X}{R + A} = 1 - \frac{T}{R + A}.$$

These are integrals of mean *squared* norms, distinct from the integrals of mean norms plotted in Figure 4. When $R + A > 0$, Cauchy–Schwarz gives $-1 \le C \le 1$. Positive C measures cancellation relative to the sum of the two squared-norm integrals. It combines vector opposition and relative magnitude; it is not itself a cosine or a fraction of the original mean-norm curve explained.

Reconstructing X from the frozen per-pair norms gives $C = 0.96409 \pm 0.00088$ initially. For raw weights it becomes 0.66597 ± 0.00862 at 1k and 0.24247 ± 0.01823 at 50k. EMA retains 0.96440 ± 0.00075 at 1k and reaches 0.25790 ± 0.00421 at 50k. Thus a small initial total can coexist with large opposing components, and their cancellation weakens during these runs. This is an algebraic description of the operational diagnostic, not a causal account of training, a curvature measurement, or an explanation of generated-quality changes.

The supplement releases every seed/checkpoint value, the pre-calculation analysis declaration, input hashes, and replay code. Direct inner products of 1,000 independently constructed vector pairs verify the norm-based reconstruction, including aligned, opposed, and zero-component cases. This check validates the algebra; it does not create new experimental replications. The original per-pair norm and decomposition checks remain unchanged.

H Reproduction and validation checks

The source package separates preserved observations from new calculations. Figures 2, 5, and 7 are generated by `verify_geometry.py`; raw numerical outputs include exact sphere identities, evaluation-pair moments, finite-angle errors, and the rationalized ellipsoid quadrature. `evaluate_prefix.py` records the new image protocol, builds the reference bank, captures prefix inputs, and computes all selection baselines and quality statistics. `build_results.py` recomputes tables and figures from per-seed records, using sample rather than population SD. Figure 3 is rendered from the verified action formulas by `render_intrinsic_action.py`.

For the original prefix evaluation, the supplementary material includes the final generated features, all time-resolved scores, and the real-reference cross-kernel and self-kernel means. Its quality-replay mode recomputes generated-sample kernels, bucket assignments, IMMD, permutation tests, and retained-feature variance without a generative checkpoint. Rebuilding the real-reference summaries additionally requires the canonical CIFAR-10 test images and the specified encoder. Recomputing all trajectory scores from images requires the original generative checkpoints and training-feature bank.

Meaningful checks include direct-vector versus scalar sphere identities, an independent finite-difference check of the projection differential, comparison of the analytic sphere proxy with all 30 preserved seed/concentration measurements, agreement of two ellipsoid quadratures and their finite bounds, and fixed-rank agreement of Gram-based PCA with direct SVD. The rank-threshold discrepancy is reported separately in Appendix F.4. The analytic feature calibration is compared with its finite-difference implementation. Bibliographic records are checked against primary publisher or conference metadata and the corresponding paper or independent publisher record; published venue information is used in place of preprint-only citations when verified. These checks support the stated calculations and definitions, not an unrestricted claim that any automated audit proves a research result correct.

Seed	Old gap (invalid)	Matched gap	Random IMMD	Forecast IMMD
42	3.517	-1.085	6.151	10.334
123	3.783	-0.325	5.681	9.758
7	4.668	-1.161	6.582	11.495

Table 25: Same generated samples and forecast rankings, different quality-reference features. "Old gap" uses the incompatible cache and is shown only to quantify the evaluation error. "Matched gap" and the retained IMMD columns use consistently re-encoded references. Gap means high-minus-low at $T = 0.5, k = 300$.

H.1 Reference-feature compatibility audit

The PCA training-reference bank passes a separate CPU re-encoding check at 32 evenly spaced canonical training indices, with relative feature error 2.76×10^{-6} and maximum coordinate error 1.60×10^{-5}. This is an independent-backend compatibility check, not exhaustive validation of all 50,000 rows.

The archived image evaluations loaded real quality-reference features separately from the encoder used for generated images. Re-encoding 32 evenly spaced canonical CIFAR-10 test images produced relative feature error 0.643 against that cache. A freshly generated test bank, checked with a different batch size at those same indices, agreed to relative error 1.1×10^{-6}. The full 10,000-row feature means differ by 0.217 relative to the matched mean's norm, which is invariant to row order. Searching each of the 32 matched queries against every old row still gives nearest-neighbor discrepancies of 0.397–0.707 relative to each query's norm. Thus a row permutation alone cannot explain the mismatch. The supplement includes the two mean vectors for arithmetic replay; recomputing the searches requires the full identified banks. A limited eight-image ImageNet check also failed to reproduce its cached reference. We could not establish the exact original preprocessing or reference-image selection. Consequently, we exclude the archived IMMD values, checkpoint correlations, and quality-filtering conclusions; the source preserves them only as invalid-reference audit evidence. All new results use 10,000 real test images encoded with the same wrapper as generated samples.

This consistency requirement affects rankings, not merely a metric's offset. Let $\mu_{F,P}$ be the kernel mean embedding of features obtained from image law P using pipeline F, and suppose the real law Q is instead encoded by H. Define $\delta = \mu_{F,Q} - \mu_{H,Q}$. Then

$$\mathrm{MMD}^2(F_{\#}P, H_{\#}Q) = \mathrm{MMD}^2(F_{\#}P, F_{\#}Q) + 2\langle \mu_{F,P} - \mu_{F,Q}, \delta \rangle + \|\delta\|^2 . \tag{33}$$

The identity is an expansion of a squared Hilbert-space norm. Although the final term is fixed across methods, the cross term depends on the selected distribution P. Thus holding an incompatible real cache fixed does not make method rankings comparable to those of a consistent-feature evaluation.

For the new generated samples, using the old cache gives a forecast high-minus-low IMMD difference of 3.99 ± 0.60. Replacing only the real-reference features changes it to -0.86 ± 0.46 (Table 25). Sample identities, score directions, neighborhoods, and checkpoints are unchanged. This controlled recalculation isolates an evaluation dependence; it does not identify the undocumented cache's exact preprocessing or reference-image selection, or imply that every feature discrepancy reverses rankings. The same implementation correction was applied to all methods and all prespecified horizons.

I Controlled common-OT field validation

The analytic checks use 257 sorted angles sampled uniformly on $[-0.25, 0.25]$, with NumPy seed 20,260,926. The seven increasing target maps are $T(\alpha) = \alpha + 0.6$, $0.7\alpha + 0.5$, $1.3\alpha + 0.45$, $0.8\alpha + 0.7\alpha^3 + 0.5$, $\alpha - 0.6$, 1.3α, and α. Every combined source/target support lies within an interval shorter than $\pi/2$. Squared ambient costs are computed from Cartesian endpoints; intrinsic costs use squared angular distances. Independent linear-assignment solves are checked against increasing matching, and direct Cartesian projection and velocity calculations against Eq. (3). The identity map and mixed-sign transport check zero and signed displacements. The numerical samples validate the calculation; the theorem concerns the stated continuous laws.

For the reach sweep, $d = 0.2$ is fixed and $r = \tau$ varies over 25 logarithmically spaced values from 0.5 to 8. The exact midpoint error is $d - 2r\sin(d/(2r))$, with leading term $d^3/(24r^2)$. This varies radius while holding intrinsic transported distance fixed, rather than varying concentration at constant geometry.

The learned circle experiment uses uniform source angles in $[-0.15, 0.15]$ and $T(\alpha) = \alpha + 0.6$. Source angles and $t \sim \mathrm{Uniform}[0, 1]$ are sampled independently in batches of 256. A 3-input, 2-output MLP has three 64-wide hidden layers and SiLU activations. Adam minimizes ambient velocity MSE for 10,000 iterations, with cosine learning-rate decay from 10^{-3} to 10^{-5} and no EMA. Five CPU float32 runs use seeds 0–4. A fixed 4,096-angle grid measures the projected L^2 learning residual and total field discrepancy at the midpoint, at iterations $0, 50, 200, 1000, 3000, 10000$. The geometric value is $0.6 - 2\sin(0.3) = 0.0089595866773$. The protocol, measurements at all recorded iterations, and final weights accompany the code.

Seed	Total field error	Learning residual	Exact geometry
0	0.0092157	0.0014285	0.0089596
1	0.0090665	0.0008160	0.0089596
2	0.0089887	0.0008515	0.0089596
3	0.0089775	0.0006297	0.0089596
4	0.0090298	0.0011551	0.0089596

Table 26: Final controlled learning results. The geometric value is identical across seeds; the learning residual and total field discrepancy are separately measured projected L^2 norms.

Full-dimensional sphere bands. An independent script checks the product extension on $\mathbb{S}^2, \mathbb{S}^3, \mathbb{S}^5$. Each of seven monotone polar transports is tested on a 16×16 product grid of polar angles and directions, including negative, mixed-sign, and zero displacements. Cartesian ambient costs and arccosine intrinsic costs dominate their polar counterparts; independent assignment solvers return the same pairing. Direct projected velocities match the formula within 6.1×10^{-17} across all 21 cases. These finite samples check the calculations; the theorem establishes optimality for continuous laws.

The learned $\mathbb{S}^2(1)$ source samples polar angle uniformly on $[0.8, 1.1]$ and azimuth independently and uniformly on $[0, 2\pi)$. Its surface-area density is $1/(0.3 \cdot 2\pi \sin a)$ within this band. It is a full-dimensional surface-absolutely-continuous distribution, not a uniform surface-area sample. The target shifts polar angle by 0.6 while preserving azimuth. A 4-input, 3-output MLP with three 128-wide SiLU hidden layers uses the same batch size, optimizer, learning-rate schedule, 10,000-iteration budget, and five seeds as the circle control. A fixed 64×64 product midpoint quadrature evaluates the entire tangent projection, including any learned azimuthal error, rather than only its meridian component. Learning uses float32; evaluation geometry uses float64. The exact geometric term remains 0.0089595866773. Final projected learning residual is 0.000911 ± 0.000158, including azimuthal RMS error 0.000302 ± 0.000027. The total field error is 0.009018 ± 0.000047. Finite residuals can partially cancel the geometric difference, so no monotonic convergence or lower bound by the limiting geometric value is asserted.

Seed	Total field L^2	Learning L^2	Azimuth error L^2
0	0.009016	0.001047	0.000293
1	0.009003	0.000677	0.000272
2	0.009002	0.000819	0.000310
3	0.009096	0.001019	0.000343
4	0.008970	0.000991	0.000293

Table 27: Learned full-dimensional $\mathbb{S}^2$ band controls. The total and learning errors include both tangent directions. Final saved weights reproduce the measurements.

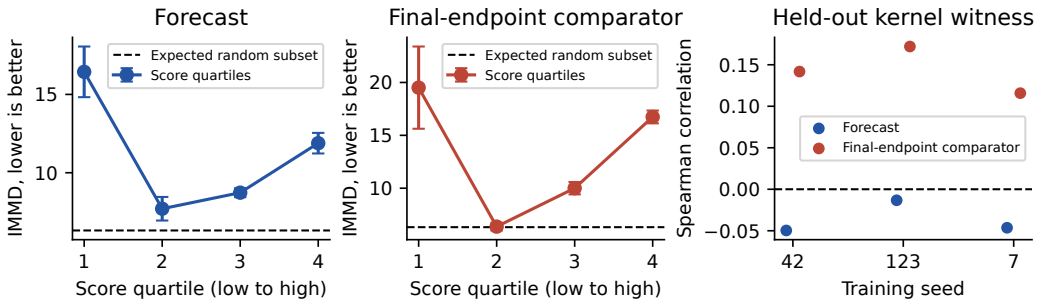

Figure 17: Fixed 400k checkpoints, with 512 samples per quartile. Bars show sample SD across three training seeds. The dashed reference is the exact expected diagonal-excluding IMMD of an equally sized uniform random subset of each fixed pool. Right: score–witness correlations for each seed.

J FIXED-CHECKPOINT SAMPLE ANALYSIS

This is a post-hoc analysis of the previously generated three 400k checkpoints and 2,048 samples per checkpoint. The primary horizon $T = 0.5$, neighborhood $k = 300$, feature pipeline, Gaussian kernel bandwidth 10, and score directions remain fixed. It does not select a new horizon or reverse a score after inspecting quality. All six nonrandom scores are retained.

Cross-fitted witness. A fixed permutation (training seed plus 20,260,926) partitions each checkpoint's generated features into four folds of 512. For feature z_i in fold $f(i)$, define

$$\widehat{w}_i = \frac{1}{|I_{-f(i)}|} \sum_{j \in I_{-f(i)}} k(z_i, z_j) - \frac{1}{10000} \sum_{a=1}^{10000} k(z_i, y_a), \qquad k(z, z') = \exp(-\|z - z'\|^2 / 200).$$

This estimates the standard, unnormalized generated-minus-real MMD witness (Gretton et al., 2012, Section 2.3). Neither the evaluated sample nor its fold enters its generated reference. The CIFAR-10 test-reference features use the same encoder pipeline. The witness is a signed distribution contrast in a specified feature/kernel geometry, not a per-image IMMD or a human realism label.

Writing $w(z) = \langle \mu_P - \mu_Q, k(z, \cdot) \rangle$, the derivative of $\mathrm{MMD}^2((1-\epsilon)P + \epsilon\delta_z, Q)$ at zero is $2\{w(z) - \mathbb{E}_P w\}$. This is a distributional influence calculation. It does not assert that a point with a larger witness is intrinsically a worse image and does not replace the direct subset-quality comparison.

Method	Seed 42	Seed 123	Seed 7
Forecast	-0.050	-0.013	-0.046
Final-endpoint comparator	+0.142	+0.172	+0.116
Pixel velocity	-0.090	-0.058	-0.103
Feature velocity	-0.083	-0.097	-0.111
Normal velocity	-0.089	-0.102	-0.115
Nearest reference	-0.028	-0.026	+0.013

Table 28: Spearman correlations with the four-fold witness at each fixed checkpoint. Training seeds, rather than the pooled 6,144 samples, are the model replicates.

Quartiles and uncertainty. Each score is sorted stably into four groups of 512. We compute the same diagonal-excluding Gaussian MMD statistic against all 10,000 real features for every group. For a uniform random subset of fixed size $m \geq 2$, its expected statistic equals the full pool's statistic: every ordered pair and every individual point has the same selection probability. This is the exact matched random-retention reference conditional on the pool. We also record 512 random subsets per seed. Their central 95% range describes random-subset variation, not uncertainty across newly trained models. Figure 17 uses sample standard deviations across the three training seeds.

The per-seed correlation intervals use 1,000 paired bootstrap resamples of the fixed score/witness vectors, recomputing ranks each time. These conditional resampling summaries do not account for relearning the models, re-estimating shared kernel references, or repeating cross-fitting. We do not use them to claim model-level significance. All methods are retained.

Seed	Method	ρ	Conditional interval	Q1	Q2	Q3	Q4
42	Forecast	-0.050	[-0.092,-0.005]	15.66	7.70	8.46	11.95
42	Final-endpoint comparator	+0.142	[+0.099,+0.179]	17.50	6.00	10.27	16.77
42	Pixel velocity	-0.090	[-0.132,-0.048]	11.57	6.37	6.61	6.67
42	Feature velocity	-0.083	[-0.122,-0.040]	12.96	8.55	6.02	8.17
42	Normal velocity	-0.089	[-0.134,-0.048]	12.36	8.78	5.99	7.96
42	Nearest reference	-0.028	[-0.073,+0.014]	8.30	6.03	5.88	6.73
123	Forecast	-0.013	[-0.053,+0.027]	15.36	6.95	8.69	11.20
123	Final-endpoint comparator	+0.172	[+0.128,+0.216]	17.04	6.38	9.33	16.12
123	Pixel velocity	-0.058	[-0.104,-0.018]	10.71	6.65	5.49	8.22
123	Feature velocity	-0.097	[-0.138,-0.056]	12.89	7.09	6.56	7.68
123	Normal velocity	-0.102	[-0.147,-0.060]	12.29	7.77	5.76	7.64
123	Nearest reference	-0.026	[-0.071,+0.019]	7.53	6.37	6.45	6.03
7	Forecast	-0.046	[-0.088,-0.005]	18.30	8.46	9.03	12.51
7	Final-endpoint comparator	+0.116	[+0.076,+0.161]	23.99	6.73	10.41	17.33
7	Pixel velocity	-0.103	[-0.147,-0.062]	11.56	6.68	6.20	7.32
7	Feature velocity	-0.111	[-0.154,-0.070]	14.81	8.37	7.56	8.41
7	Normal velocity	-0.115	[-0.155,-0.072]	14.59	7.58	8.37	7.80
7	Nearest reference	+0.013	[-0.029,+0.058]	6.20	7.44	6.24	7.54

Table 29: Per-seed witness correlations, conditional bootstrap ranges, and quartile IMMD. Q1 contains the lowest scores. All quartiles contain 512 samples; lower IMMD is better.

Exact finite-pool selection identity. Let A_L, A_H be within-group generated kernel averages excluding diagonals, and C the cross-group kernel average for disjoint groups of size m. For their union the within-generated average is

$$\frac{m-1}{2(2m-1)}(A_L + A_H) + \frac{m}{2m-1}C.$$

The real-reference self term is common and its cross term is the average of the two group cross terms. Substituting $\widehat{M}(L,H) = A_L + A_H - 2C$ gives

$$\widehat{M}(L \cup H, Q) = \frac{\widehat{M}(L,Q) + \widehat{M}(H,Q)}{2} - \frac{m}{2(2m-1)}\widehat{M}(L,H),$$

proving Eq. (10). This elementary identity neither introduces a new metric nor assumes independence of score-selected groups. The finite between-group statistic can be negative; the exact comparison remains valid. We symmetrize kernels to remove roundoff asymmetry and verify the identity for every method and seed. Primary half-bucket values agree with the previous GPU results to within 4.7×10^{-6} IMMD units.

Method	Observed gap	Required gap	Gain over random
Forecast	-0.86	+7.57	-4.21
Final-endpoint comparator	+3.51	+9.42	-2.95
Pixel velocity	-2.05	+2.13	-2.09
Feature velocity	-2.89	+4.23	-3.56
Normal velocity	-2.91	+3.93	-3.42
Nearest reference	-0.36	+0.64	-0.50

Table 30: Finite-pool selection accounting, averaged across three seeds at $m = 1024$. The required gap is $m\widehat{M}(L,H)/(2m-1)$. A positive gain over expected random retention requires the observed gap to exceed this quantity, not merely zero.