NORMAL FRÉCHET DRIFT: MEASURING POST-TRAINING DEPARTURE FROM PRETRAINED REFERENCE MANIFOLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training can improve a generator's reward while moving its outputs away from pretrained reference geometry. We distinguish this *normal drift* from *tangential drift*, which redistributes probability along the same reference manifold. We introduce *Normal Fréchet Drift* (NFD), the expected squared distance to a fixed reference set in a shared feature space. Its ideal value is zero for distributions confined to that set. To estimate NFD from finite pretrained samples, we construct bounded Multi-Patch reconstructions: local plane patches centered at sampled features, with finite radii that control extrapolation. We separate evaluation sampling variance from reconstruction bias and bound the bias through local coverage, curvature, tangent, and anchor errors. Controlled geometries validate the distinction between normal and tangential drift and characterize reconstruction resolution. On Stable Diffusion 3.5 Medium, FlowDPO exhibits the largest reconstruction-relative departure, with stable ordering across prompt splits, encoders, and reference refits. A FLUX.2 comparison extends NFD to distillation. In blinded evaluation on 120 selected image pairs, lower-NFD images receive 84.4% of overall-quality judgments, counting ties as half a win. NFD quantifies post-training departure from pretrained reference geometry.

1 INTRODUCTION

Preference optimization adapts pretrained generators to human preferences and task rewards (Wallace et al., 2024; Lee et al., 2025; Eyring et al., 2024; Rafailov et al., 2023). Yet text-to-image (T2I) optimization can increase reward at the cost of diversity or visual plausibility (Rafailov et al., 2024; Jena et al., 2025). This raises a geometric question: *does post-training move generated outputs away from pretrained reference geometry?*

We distinguish *normal drift*, where outputs leave a pretrained reference manifold, from *tangential drift*, which redistributes probability along it (Han et al., 2026) (Fig. 1). A normal-drift diagnostic should vanish for on-manifold motion and quantify departure. FID can change under tangential drift (Heusel et al., 2017), while finite-radius precision/recall measures neighborhood membership rather than normal displacement (Kynkäänniemi et al., 2019).

Finite reference samples create two difficulties: an on-manifold output can lie far from the observed samples, while an off-manifold output can lie on an extrapolated local plane. Measuring drift therefore requires controlling reference-reconstruction error.

We introduce *Normal Fréchet Drift* (NFD), the expected squared distance to a fixed reference set. Its ideal form uses a reference manifold; real-model evaluation uses a *bounded Multi-Patch* (MP) reconstruction from pretrained samples. MP centers local planes at sampled anchors and restricts their radii (Fig. 2). Our theory connects ideal and reconstructed NFD through local coverage, curvature, tangent, and anchor errors.

Our contributions are:

1. **A geometric estimand.** NFD measures normal departure and remains zero under tangential redistribution. A transport characterization formalizes its distinction from distribution matching.

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.

2. **Finite-sample estimation and theory.** Bounded MP controls structural false zeros. We separate evaluation sampling variance from reconstruction bias through an error decomposition and bound the bias using local geometry.
3. **Controlled and real-model validation.** Known geometries test accuracy and resolution. SD3.5-M and FLUX.2 comparisons establish reconstruction-relative effects, with encoder and reference-refit checks and selected-pair perceptual associations.

2 Related Work

NFD connects preference optimization, generative-model evaluation, and local geometry.

Preference optimization and drift. DPO (Rafailov et al., 2023) and diffusion adaptations (Wallace et al., 2024) optimize preferences without an explicit reward model, while multi-reward and inference-time methods broaden the alignment toolbox (Lee et al., 2025; Eyring et al., 2024). NFD uses the distinction between normal and tangential drift of Han et al. (2026) to assess normal drift across post-trained checkpoints.

Generative-model evaluation. FID and KID compare marginal feature distributions (Heusel et al., 2017; Bińkowski et al., 2018). MMD is a kernel two-sample discrepancy (Gretton et al., 2012). Improved precision and recall, together with density and coverage, separate fidelity from diversity with finite-radius neighborhoods (Kynkäänniemi et al., 2019; Naeem et al., 2020), with probabilistic and frontier variants improving reliability (Park et al., 2023; Cheema & Urner, 2023). NFD instead isolates distance to reference geometry, allowing probability redistribution along it. Representation and finite-sample effects in feature metrics (Stein et al., 2023; Jayasumana et al., 2024) motivate our encoder and reference-refit comparisons.

Geometry in learned representations. Image and diffusion representations exhibit measurable local dimension and curvature (Kaufman & Azencot, 2023; Stanczuk et al., 2024), and diffusion theory can adapt to low-dimensional structure (Tang & Yang, 2024; Li & Yan, 2024; Potaptchik et al., 2025). Local-PCA tangent estimation has non-asymptotic guarantees under manifold assumptions (Aamari & Levrard, 2019). We use this connection to analyze bounded local reconstructions and empirically audit their geometric assumptions (App. K).

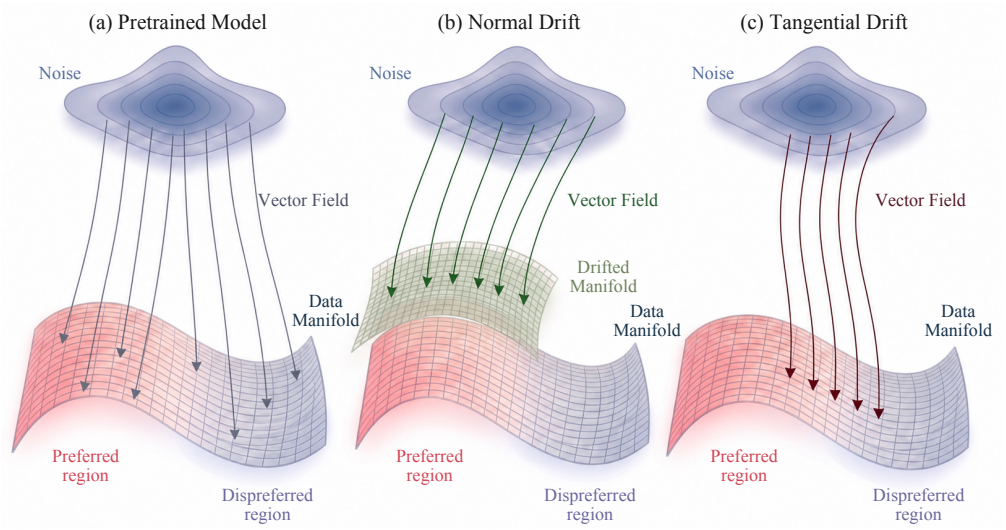

Figure 1: **Pretrained generation and two forms of drift.** (a) Pretrained trajectories reach the reference manifold. (b) Normal departure yields positive ideal NFD. (c) Tangential redistribution has zero ideal NFD. "Data Manifold" denotes the ideal reference geometry. Adapted from Han et al. (2026); formal distinctions appear in App. B.2.

3 NORMAL FRÉCHET DRIFT

We distinguish the ideal reference manifold $\mathcal{M}_c$ from its finite reconstruction $\widehat{\mathcal{M}}_c$. For condition c and noise ξ, let $Z_\theta = \psi(G_\theta(\xi, c)) \in \mathbb{R}^d$, where $\psi = \Pi \circ \phi$ combines a frozen encoder ϕ with a projection Π fitted to pretrained samples. Write $P_{\theta,c}$ for the feature distribution and $P_{0,c}$ for its pretrained counterpart. All checkpoints share ψ and the reference geometry.

3.1 DEFINITION AND INTERPRETATION OF NFD

NFD is defined for a closed reference set. A smooth manifold additionally provides local Euclidean coordinates and tangent spaces for the interpretation of normal drift and tangential drift. App. A specifies the assumptions.

Definition 1 (Normal Fréchet Drift)**.** For condition c, let $\mathcal{M}_c \subset \mathbb{R}^d$ be a nonempty closed reference set in the frozen feature space. The ideal NFD of checkpoint θ is

$$\mathrm{NFD}(\theta \mid c) = \mathbb{E}_{Z\sim P_{\theta,c}} d^2_{\mathcal{M}_c}(Z), \qquad d_{\mathcal{M}_c}(z) = \inf_{x\in\mathcal{M}_c} \|z - x\|_2.$$

The name emphasizes its squared-transport interpretation in Prop. 1; unlike FID, NFD does not fit Gaussians to the two distributions.

For the manifold interpretation we take $\mathcal{M}_c$ to be the ideal reference manifold. Distances are computed per output without pairing with pretrained samples.

Optimal transport interpretation. For a distribution P and reference set A, write $\mathsf{D}(P; A) = \mathbb{E}_{Z\sim P} d^2_A(Z)$. In particular,

$$\mathrm{NFD}(\theta \mid c) = \mathsf{D}(P_{\theta,c}; \mathcal{M}_c).$$

The reference $\mathcal{M}_c$ stays fixed across checkpoints. Let W_2 be the 2-Wasserstein distance and $\mathcal{P}_2(\mathbb{R}^d)$ the distributions with finite second moment (App. B.1).

Proposition 1 (Optimal transport characterization of NFD)**.** *Let $A \subset \mathbb{R}^d$ be nonempty and closed, and $P \in \mathcal{P}_2(\mathbb{R}^d)$. Then*

$$\mathsf{D}(P; A) = \inf_{\substack{Q\in\mathcal{P}_2(\mathbb{R}^d)\\ Q(A)=1}} W_2^2(P, Q). \tag{1}$$

Moreover, $\mathsf{D}(P; A) = 0$ if and only if $P(A) = 1$.

Applying Eq. 1 to $P_{\theta,c}$ and $\mathcal{M}_c$ identifies ideal NFD as the minimum *squared* W_2 distance to any distribution on $\mathcal{M}_c$ (see App. B.1 for the proof). The destination need not match $P_{0,c}$. Thus mode reweighting or removal on $\mathcal{M}_c$ leaves ideal NFD zero, while distribution distances or recall can change (see App. B.2).

Comparison with FID. In a common feature space, let $\mathcal{G}(P)$ be the Gaussian distribution with the same mean and covariance as P. FID is the squared 2-Wasserstein distance between two such Gaussian fits (Heusel et al., 2017, Eq. (6)):

$$\mathrm{FID}(P_{\theta,c}, P_{0,c}) = W_2^2(\mathcal{G}(P_{\theta,c}), \mathcal{G}(P_{0,c}))\,. \tag{2}$$

FID therefore compares two fixed Gaussian fits, whereas NFD keeps the original output distribution and minimizes over all destination distributions on $\mathcal{M}_c$. Translating a distribution on a line by length δ produces tangential drift when the translation is along the line and normal drift when it is perpendicular to the line, giving ideal NFD 0 and δ^2, respectively. With finite second moments, both translations give FID δ^2 relative to the original distribution. App. B.2.1 gives the derivation.

3.2 REFERENCE RECONSTRUCTION FOR NFD

Bounded Multi-Patch (MP) builds a finite reference set from pretrained features: a union of local plane patches centered at observed features and restricted to finite radii. We fit the patches and select reconstruction parameters using pretrained samples, then freeze the resulting geometry before evaluating post-trained outputs (Fig. 2). This pretrained-only protocol is shared by all subsequent comparisons.

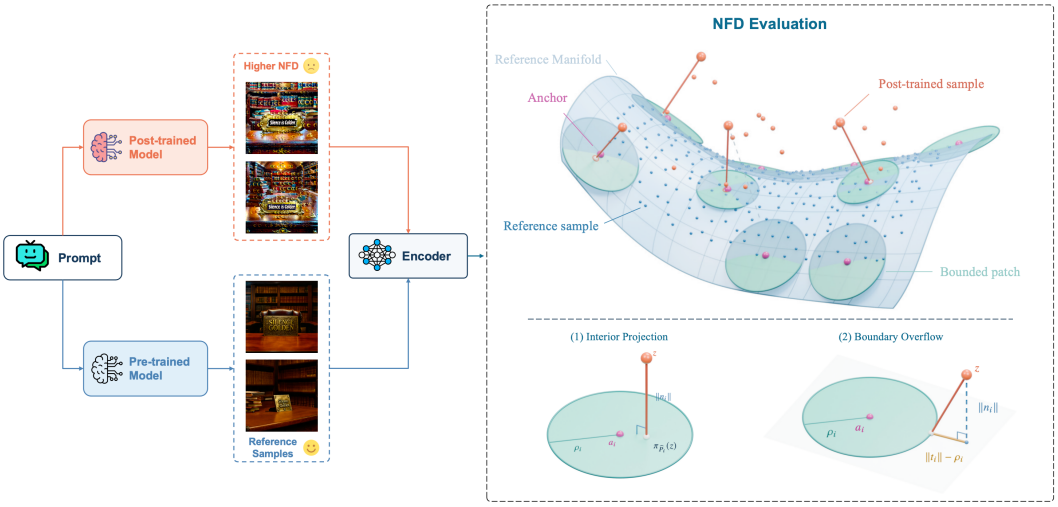

Figure 2: **Reference construction and NFD evaluation.** Pretrained features determine bounded patches in the shared feature space. NFD averages each output's squared distance to the nearest patch. The lower panels show an interior projection (perpendicular residual) and an exterior projection (additional radial excess). Algs. 1–2 give pseudocode.

Unlike nearest-neighbor distance, local patches interpolate between finite references and reduce spacing residuals; anchoring and bounded radii are then needed to avoid the false zeros introduced by unconstrained local planes.

Constructing the patches. For each condition (omitting c), split pretrained features into a fit bank $B_{\text{fit}} = \{y_n\}_{n=1}^{N}$ for constructing patches and disjoint calibration banks for selecting reconstruction parameters. Select M anchors a_i by farthest-point sampling (FPS): each new anchor is the fit sample farthest from the selected set. At each anchor, PCA on its K nearest fit samples gives m orthonormal directions $U_i \in \mathbb{R}^{d\times m}$. Set the radius to $\rho_i = \lambda_\rho h_i$, where h_i is the farthest-neighbor distance and λ_ρ is chosen using calibration data. The patches and their union are

$$\widehat{\mathcal{P}}_i = \{a_i + U_i t : \|t\|_2 \le \rho_i\}, \qquad \widehat{\mathcal{M}} = \bigcup_{i=1}^{M} \widehat{\mathcal{P}}_i. \tag{3}$$

Sampled anchors avoid off-manifold centers: averaging points from two separated regions can place a query-centered plane in the gap and give it zero residual. The neighborhood mean is therefore used only for PCA directions, not as the patch center.

Finite radii control extrapolation. A point on the fitted plane but outside its patch incurs positive distance instead of the zero assigned by an unbounded plane.

Handling unreliable fits. We check that each neighborhood has enough samples, the last retained PCA eigenvalue is sufficiently larger than the first discarded one, and the fitted plane is stable across two neighborhood sizes. If a check fails, we set $\rho_i = 0$, reducing the patch to its anchor point. App. C.5 specifies the checks and thresholds.

Computing the distance. For an output feature z, each patch distance combines perpendicular displacement and radial excess. We take the minimum over patches:

Proposition 2 (Exact patch distance). *For $r_i = z - a_i$, $t_i = U_i^\top r_i$, and $n_i = (I - U_i U_i^\top) r_i$,*

$$d^2_{\widehat{\mathcal{P}}_i}(z) = \|n_i\|_2^2 + (\|t_i\|_2 - \rho_i)_+^2, \qquad \widehat{d}^2_{\text{MP}}(z) = \min_i d^2_{\widehat{\mathcal{P}}_i}(z). \tag{4}$$

Here $(u)_+ = \max\{u, 0\}$. An interior projection contributes only the perpendicular residual; an exterior projection adds radial excess to the boundary. App. B.3 derives the formula and Fig. 2 illustrates both cases.

3.3 NFD Estimation

Estimating NFD. With the representation and reconstruction fixed, we estimate NFD by averaging squared distances over generated outputs. Let $\mathcal{R}$ denote the pretrained reference samples and their fit and calibration assignments. At a fixed condition c, the reconstruction population value and its estimate from J outputs are

$$\mathrm{NFD}_{\mathrm{rec}}(\theta \mid c, \mathcal{R}) = \mathsf{D}(P_{\theta,c}; \widehat{\mathcal{M}}_c) \quad \text{and} \quad \widehat{\mathrm{NFD}}_{J,\mathrm{rec}}(\theta \mid c, \mathcal{R}) = \frac{1}{J}\sum_{j=1}^{J} d^2_{\widehat{\mathcal{M}}_c}(Z_{\theta,j}) \tag{5}$$

where $Z_{\theta,j} \sim P_{\theta,c}$. For conditions $c \sim \nu$, average the population value over c and use condition-specific distances for evaluation:

$$\mathrm{NFD}_{\mathrm{rec}}(\theta \mid \mathcal{R}) = \mathbb{E}_{c\sim\nu}\,\mathrm{NFD}_{\mathrm{rec}}(\theta \mid c, \mathcal{R}) \quad \text{and} \quad \widehat{\mathrm{NFD}}_{J,\mathrm{rec}}(\theta \mid \mathcal{R}) = \frac{1}{J}\sum_{j=1}^{J} d^2_{\widehat{\mathcal{M}}_{c_j}}(Z_{\theta,j}) \tag{6}$$

where $c_j \sim \nu$ and $Z_{\theta,j} \sim P_{\theta,c_j}$. Sec. 3.4 bounds the estimation error.

Comparing checkpoints with Base. Our primary comparison measures change relative to held-out pretrained (Base) outputs, accounting for their finite-reconstruction residual. We pair each checkpoint ℓ with Base at the same condition c and seed s:

$$\Delta_\ell(c, s \mid \mathcal{R}) = d^2_{\widehat{\mathcal{M}}_c}(Z_\ell(c,s)) - d^2_{\widehat{\mathcal{M}}_c}(Z_{\mathrm{Base}}(c,s)). \tag{7}$$

Positive Δ_ℓ indicates greater departure than Base; a negative Δ_ℓ means the checkpoint output is closer to the frozen reconstruction than its paired Base output. Pairing controls condition and seed variation; reference refits assess sensitivity to reconstructed geometry (App. I.3).

3.4 Error Analysis of NFD Estimation

Fix the representation, condition c, and reference data $\mathcal{R}$. For Eq. 5, we bound reconstruction bias through local geometry and combine it with finite-evaluation sampling variance.

Overestimation and underestimation. Gaps between patches can overestimate distance, while patches extending off the manifold can underestimate it. We control both directions of $e(z) = d_{\widehat{\mathcal{M}}_c}(z) - d_{\mathcal{M}_c}(z)$.

Four local error components. *Curvature* bends the manifold away from flat patches. *Tangent estimation* tilts fitted directions. *Coverage* gaps increase distance to the patch union. *Anchor deviation* shifts patches off the manifold. The last term vanishes for exact on-manifold anchors. Each component contributes to the two-sided distance-error bound.

NFD estimation error. For each component k, let Γ_k be the root expected sum of its squared upper- and lower-error bounds under $P_{\theta,c}$ (Eq. 26).

Theorem 3 (NFD estimation error)**.** *Let $e(z) = d_{\widehat{\mathcal{M}}_c}(z) - d_{\mathcal{M}_c}(z)$ denote the signed distance error and assume the local geometric conditions in App. B.5 hold,* $\mathrm{NFD}(\theta \mid c) < \infty$*, and all four errors Γ_k are finite. Then*

$$\Gamma_{P_{\theta,c}} := \Gamma_{\mathrm{cov}} + \Gamma_{\mathrm{curv}} + \Gamma_{\tan} + \Gamma_{\mathrm{anchor}} \quad \textit{and} \quad \left(\mathbb{E}_{Z\sim P_{\theta,c}}[e(Z)^2]\right)^{1/2} \le \Gamma_{P_{\theta,c}} \tag{8}$$

If, conditional on $\mathcal{R}$, the J evaluation outputs are independent draws from $P_{\theta,c}$ and $d^2_{\widehat{\mathcal{M}}_c}(Z)$ has finite variance, then, with the fixed arguments of the NFD terms omitted,

$$\mathbb{E}\Big[\big(\widehat{\mathrm{NFD}}_{J,\mathrm{rec}} - \mathrm{NFD}\big)^2 \,\Big|\, \mathcal{R}\Big] \le \underbrace{\frac{1}{J}\,\mathrm{Var}_{P_{\theta,c}}\big[d^2_{\widehat{\mathcal{M}}_c}(Z)\big]}_{\text{sampling variance}} + \underbrace{\big(2\sqrt{\mathrm{NFD}}\,\Gamma_{P_{\theta,c}} + \Gamma^2_{P_{\theta,c}}\big)^2}_{\text{squared bias bound}}. \tag{9}$$

Apps. B.5 and B.6 give the definitions and proofs.

Reducing estimation error. Increasing J reduces variance, while improving reconstruction balances coverage against local accuracy. Denser anchors shorten patch coverage distances. Larger radii close gaps but increase curvature and tangent error. Neighborhood size controls PCA accuracy, and anchor noise shifts patches. App. B.5 analyzes these tradeoffs and Fig. 13 tests these components.

4 Experiments

Our experiments assess (1) the validity of NFD estimation under different geometries (Sec. 4.1), (2) the NFD estimation of different post-training methods (Sec. 4.2), (3) the robustness of NFD estimation (Sec. 4.3) and (4) the relation of NFD to existing metrics and human judgments (Sec. 4.4). Below, raw NFD denotes the empirical reconstruction score $\widehat{\mathrm{NFD}}_{J,\mathrm{rec}}$, RMS-NFD denotes its square root, and Δ_ℓ denotes the paired checkpoint-minus-Base effect.

4.1 Controlled Geometric Experiments

In this subsection, we use known geometries to test three properties: (1) separating normal drift from tangential drift, (2) avoiding structural false zeros, and (3) distinguishing sampling error from reconstruction bias. Apps. D–G give the protocols.

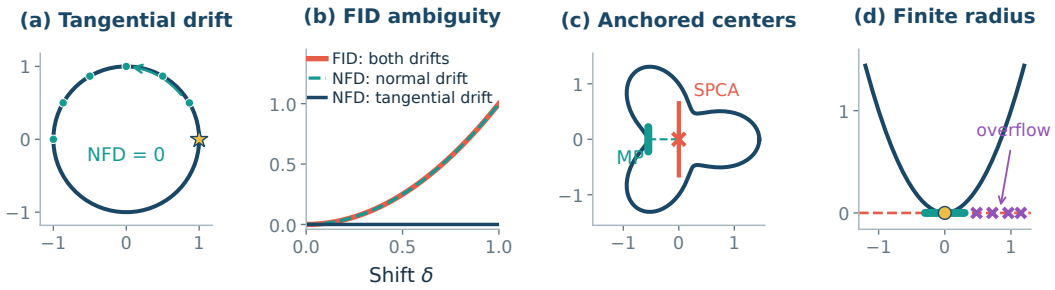

Figure 3: **Geometry behind drift separation and false zeros.** (a) Tangential drift follows the circle and leaves ideal NFD zero. (b) Equal normal and tangential drifts have identical FID but different ideal NFD. (c) A mean-centered SPCA fit passes through the flower's off-manifold center. An anchored patch avoids it. (d) Finite radii penalize points beyond the local tangent patch. These panels show analytic constructions. Apps. D and F.9 report quantitative comparisons.

Normal drift versus tangential drift. NFD distinguishes the two drift types in settings where FID alone cannot. Tangential drift around the unit circle leaves ideal NFD zero (Fig. 3a). MP's maximum squared distance is 5.17×10^{-9}, while a normal projection fixed at the starting point reaches 4 (Fig. 9b). Translating a line-supported distribution by length δ in the tangential or normal direction yields FID δ^2 in both cases but ideal NFD 0 and δ^2, respectively (Fig. 3b). Under controlled normal drift, estimates correlate 0.997 with truth (Tab. 4).

Avoiding false-zero NFD estimates. Anchoring and finite radii address two structural false-zero mechanisms. Mean-centered fitting can place a patch through the flower's off-manifold center (Fig. 3c); an unbounded tangent can assign zero distance to distant off-manifold points (Fig. 3d). Across five fixed toy cases, mean false-zero rates are 0% for bounded MP and NN, 20% for SPCA, and 40% for unbounded patches. App. F.9 gives the comparison.

Sampling error and reconstruction bias. More evaluation samples reduce sampling variance but leave a reconstruction-bias floor (Fig. 4a). Across 12 combinations of reference and evaluation sizes, the decomposition predicts observed MSE to within 2.31 Monte Carlo standard errors (App. F.2). Relative NFD error also decreases as $\delta/\epsilon_{\mathcal{M}}$ increases on three spheres, a flower, and a saddle (Fig. 4b), where $\delta = \sqrt{\mathrm{NFD}_{\mathrm{true}}}$ and $\epsilon_{\mathcal{M}}$ is the measured Hausdorff reconstruction error.

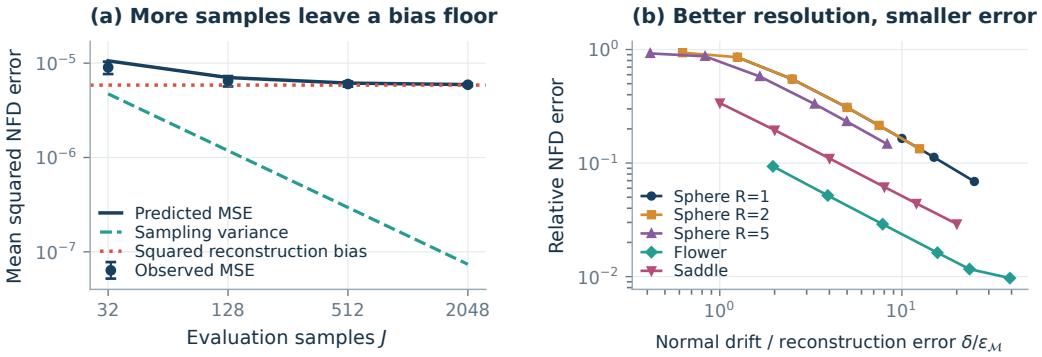

Figure 4: **NFD accuracy depends on sampling variance and reconstruction bias.** (a) For a fixed MP reconstruction ($N = 4096$ reference samples), increasing evaluation size J reduces sampling variance, while MSE approaches the squared reconstruction bias. Points show empirical MSE over 200 repeats per J. Error bars indicate ± 2 Monte Carlo standard errors. (b) Relative NFD error decreases with $\delta/\epsilon_{\mathcal{M}}$ across spheres ($R = 1, 2, 5$), a flower, and a saddle. Here $\delta = \sqrt{\mathrm{NFD}_{\mathrm{true}}}$ is the RMS normal drift and $\epsilon_{\mathcal{M}}$ is the measured reconstruction error. Apps. F.2 and G.1 give the protocols.

4.2 NFD Estimates Across Post-Training Methods

FlowDPO shows the largest SD3.5-M normal drift, with lower scores for all three mitigation variants. FLUX.2 also shows the normal drift after distillation.

Models and protocol. We compare SD3.5-M (Esser et al., 2024) and five post-trained checkpoints (Han et al., 2026) (Tab. 1). The primary set has 24 prompts and four paired seeds per method; 12 disjoint prompts provide confirmation. Each prompt uses 64 Base fit, 16 calibration, and 16 cross-fit references. The primary representation is SigLIP2 (Tschannen et al., 2025) with Base-fit PCA to $d = 16$; checks use VAE-GAP, DINOv2 (Oquab et al., 2024), and DINOv3 (Siméoni et al., 2025). Intervals use 10,000 hierarchical bootstrap draws over prompts and paired seeds (Efron & Tibshirani, 1994). See Sec. 3.2 and App. I.1 for the protocol.

Table 1: **Primary reconstruction-relative T2I results.** Brackets are 95% hierarchical bootstrap CIs over prompts and paired seeds ($B = 10{,}000$). Values are rounded to three decimals.

Method	Raw NFD [95% CI]	Δ vs. Base [95% CI]
Base SD3.5-M	0.028 [0.023, 0.033]	0
RFT	0.031 [0.026, 0.037]	0.004 [−0.002, 0.009]
FlowDPO	0.105 [0.078, 0.134]	0.077 [0.050, 0.108]
FlowDPO+KL	0.064 [0.050, 0.081]	0.036 [0.022, 0.054]
FlowDPO+RFT	0.051 [0.039, 0.063]	0.023 [0.012, 0.035]
ThermoDPO-weighted	0.034 [0.027, 0.042]	0.006 [0.000, 0.013]

SD3.5-M: NFD across post-training methods. FlowDPO+KL, FlowDPO+RFT, and ThermoDPO-weighted all have lower NFD than FlowDPO. ThermoDPO-weighted gives the largest mean reduction, 0.071 (95% CI [0.045, 0.100]; Fig. 5a). The paired 95% confidence intervals for all three reductions lie above zero (App. I.1).

FLUX.2: NFD before and after distillation. The official distilled FLUX.2 [klein] 4B model has higher NFD than its Base counterpart (Black Forest Labs, 2026). Across 128 pairs matched by prompt and seed, the mean NFD increase is 0.00255 (95% CI [0.00065, 0.00490]). Mean NFD is higher after distillation on 14 of 16 non-OCR prompts (App. J).

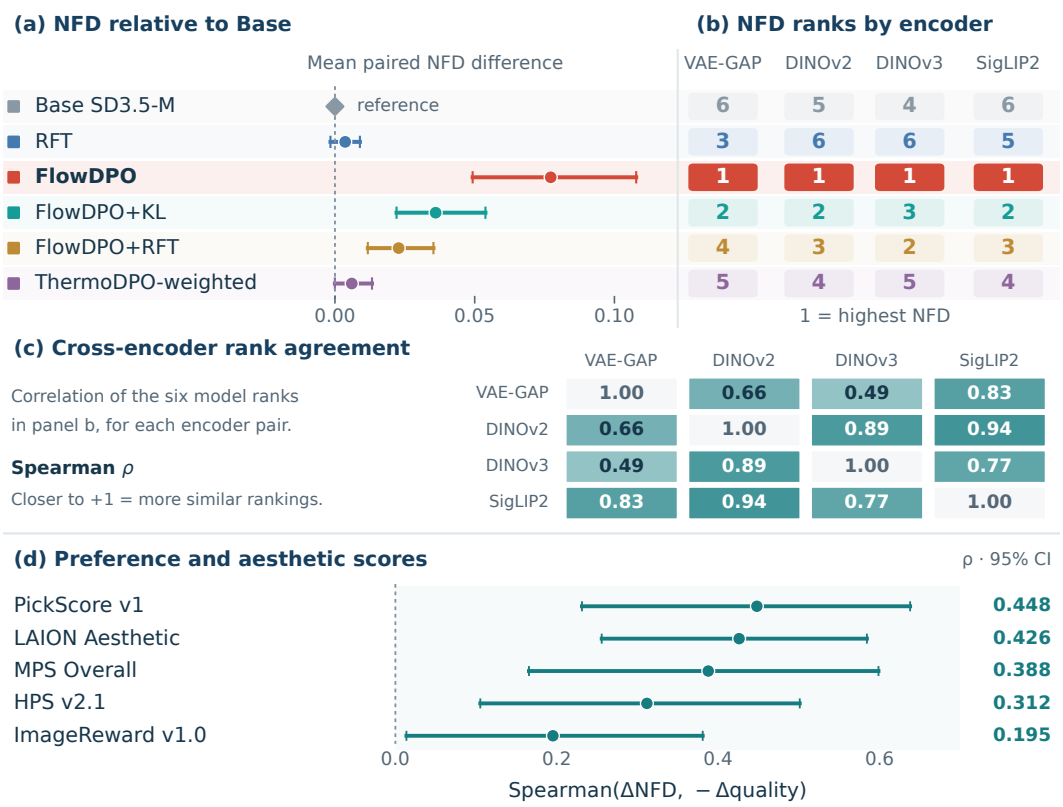

Figure 5: **NFD comparisons across methods, encoders, and external evaluators.** (a) Paired NFD differences from Base for SD3.5-M checkpoints. (b) NFD ranks under four frozen encoders (1 = highest NFD). (c) Spearman correlations between the six-method rankings in (b) (1 indicates identical rankings). (d) Spearman correlations of Base-relative NFD with negated Base-relative scores from five preference and aesthetic evaluators across 216 prompt–method cells. Positive values indicate higher NFD with lower evaluator scores. Task-specific EasyOCR results are reported in App. I.8. Error bars in (a,d) show 95% confidence intervals.

4.3 Robustness of NFD Estimation

For SD3.5-M, FlowDPO retains higher NFD than Base and the three mitigation variants across the tested encoders, prompt sets, estimators, and reference refits.

Feature encoders. FlowDPO has the highest NFD under all four encoders (Fig. 5b). The six-method rankings are positively correlated across encoders (Spearman $\rho = 0.49$–0.94; Fig. 5c).

Prompt sets. The disjoint 12-prompt confirmation set yields a similar method ranking to the 24-prompt primary set (Spearman $\rho = 0.829$). All three mitigation variants again have lower NFD than FlowDPO, with paired 95% confidence intervals for the reductions above zero (App. I.1).

Distance estimators. NN, SPCA, and MP all estimate higher NFD for FlowDPO than for Base on both prompt sets. They also agree that all three mitigation variants have lower NFD than FlowDPO. App. I.4 reports the estimator comparisons and computational costs.

Reference refits. FlowDPO retains the highest NFD in all 200 refits. Each refit repartitions each prompt's 96 Base references and rebuilds the reference geometry, keeping evaluation images and hyperparameters fixed. The FlowDPO-minus-Base difference is positive in every refit (mean 0.076; central 95% range $[0.068, 0.084]$). Across methods, the standard deviation of NFD over refits is 3.1–4.3% of its mean (App. I.3).

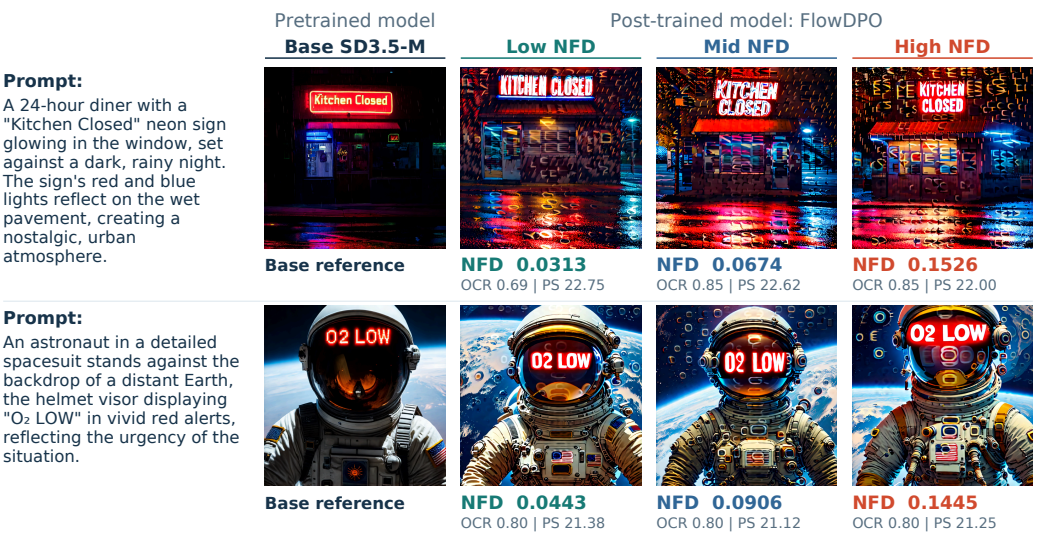

Figure 6: **Pretrained references and increasing post-training drift.** Each row shows a Base reference and three FlowDPO outputs from different seeds, ordered by squared MP distance. OCR: EasyOCR; PS: PickScore. Selection protocol: App. L.

4.4 Relation to Existing Metrics and Perceptual Quality

In our SD3.5-M experiments, NFD agrees with existing metrics and correlates with perceptual quality.

Existing metrics. Standard metrics also detect FlowDPO's departure from Base (App. I.5). Across 864 images, NFD and NN squared distance have Spearman $\rho = 0.923$, but disagree on 12.3% of method comparisons matched by prompt and seed. See App. I.6 for the sample-level comparison.

Automated quality scores. Higher NFD accompanies lower scores from all five preference and aesthetic evaluators. Across 216 prompt–method combinations, changes are measured relative to Base within each prompt. NFD increases correlate with PickScore decreases (Kirstain et al., 2023) (Spearman $\rho = 0.448$, 95% CI $[0.231, 0.638]$; Fig. 5d). App. I.8 additionally reports task-specific EasyOCR results.

Blinded human evaluation. In blinded evaluation, 84.4% of overall-quality judgments favor lower-NFD images on 120 selected pairs (rater-cluster 95% CI $[78.5\%, 90.0\%]$; ties count as half a win). Thirteen raters provided 195 judgments (App. I.9). Fig. 6 illustrates outputs ordered by NFD. The selection protocol and expanded galleries appear in App. L.

5 Discussion and Conclusion

NFD measures normal drift from pretrained geometry. Bounded MP enables estimation from finite samples, while theory separates sampling variance from reconstruction bias. SD3.5-M and FLUX.2 experiments demonstrate its use in post-training comparisons.

Real-model comparisons measure departure from the frozen pretrained reconstruction $\widehat{\mathcal{M}}_c$; controlled geometries connect this score to ideal normal drift. In SD3.5-M, NFD identifies the largest departure for FlowDPO and lower departure for its mitigation variants. The FLUX.2 comparison extends this measurement to distillation. These results establish a geometric measure of post-training change distinct from distribution matching.

AI USE STATEMENT

Generative AI tools were used to assist with code development and debugging, as well as to improve the language, clarity, and formatting of the manuscript. All AI-assisted content was reviewed and verified by the authors.

ETHICS STATEMENT

This work studies diagnostic measurements for already-trained image generators and includes a low-risk anonymous pairwise image-rating study. Participants affirmed voluntary consent in the interface, could stop at any time, and were represented only by random rater codes; no names, emails, IP addresses, or sensitive attributes were collected. The realised cohort comprised 13 raters and 195 judgments; the manuscript reports its sample size and pair coverage in full. The real-image study otherwise uses prompts and model checkpoints supplied for research evaluation; only aggregate statistics and selected generated examples are reported. Representation bias and downstream safety require separate evaluation. Sec. 5 consolidates the interpretation of NFD.

REPRODUCIBILITY STATEMENT

Apps. A–N give complete proofs, executable pseudocode, model and checkpoint identities, fixed prompt and seed splits, hyperparameters, statistical procedures, ablations, sensitivity analyses, and compute measurements. The supplementary materials include experiment configurations, machine-readable results, plotting scripts, and replication materials. See App. F.1 for the geometric validation protocol and App. F.2 for the sampling tests.

REFERENCES

Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177–204, 2019.

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

Black Forest Labs. FLUX.2 [klein]. Official model card and implementation, 2026. URL https://huggingface.co/black-forest-labs/FLUX.2-klein-base-4B.

Fahad Cheema and Ruth Urner. Precision-recall curves using information divergence frontiers. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, 2023. URL https://proceedings.mlr.press/v206/cheema23a.html.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12691–12718, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. ReNO: Enhancing one-step text-to-image models through reward-based noise optimization. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e31bdea0a93741c2157eea705dd219eb-Abstract-Conference.html.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3): 418–491, 1959.

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

Yansen Han, Shengyi Liao, Yuanxing Zhang, Pengfei Wan, and Tao Lin. Manifold drift in flow preference optimization: A root cause of reward hacking, 2026. URL https://arxiv.org/abs/2608.20011.

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, 2017.

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.

Rohit Jena, Ali Taghibakhshi, Sahil Jain, Gerald Shen, Nima Tajbakhsh, and Arash Vahdat. Elucidating optimal reward-diversity tradeoffs in text-to-image diffusion models. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 232–242, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Jena_Elucidating_Optimal_Reward-Diversity_Tradeoffs_in_Text-to-Image_Diffusion_Models_WACV_2025_paper.html.

Ilya Kaufman and Omri Azencot. Data representations' study of latent image manifolds. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15928–15945, 2023. URL https://proceedings.mlr.press/v202/kaufman23a.html.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/73aacd8b3b05b4b503d58310b523553c-Abstract-Conference.html.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, 2 edition, 2013. doi: 10.1007/978-1-4419-9982-5. URL https://doi.org/10.1007/978-1-4419-9982-5.

Kyungmin Lee, Xiahong Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. Calibrated multi-preference optimization for aligning diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18465–18475, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Lee_Calibrated_Multi-Preference_Optimization_for_Aligning_Diffusion_Models_CVPR_2025_paper.html.

Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e46c303cfd3b93bd32ecf6070318efc3-Abstract-Conference.html.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185, 2020. URL https://proceedings.mlr.press/v119/naeem20a.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, 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.

Jisung Park, Kyungmin Kim, Jaehoon Kim, Joon-Young Kim, and Gunhee Kim. Probabilistic precision and recall towards reliable evaluation of generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Park_Probabilistic_Precision_and_Recall_Towards_Reliable_Evaluation_of_Generative_Models_ICCV_2023_paper.html.

Peter Potaptchik, Iskander Azangulov, and George Deligiannidis. Linear convergence of diffusion models under the manifold hypothesis. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4668–4685, 2025. URL https://proceedings.mlr.press/v291/potaptchik25a.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e45caa3d5273d105b8d045e748636957-Abstract-Conference.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46412–46440, 2024. URL https://proceedings.mlr.press/v235/stanczuk24a.html.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Renjie Liu, Anthony L. Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0bc795afae289ed465a65a3b4b1f4eb7-Abstract-Conference.html.

Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1648–1656, 2024. URL https://proceedings.mlr.press/v238/tang24a.html.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wallace_Diffusion_Model_Alignment_Using_Direct_Preference_Optimization_CVPR_2024_paper.html.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023. URL https://arxiv.org/abs/2306.09341.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8018–8027, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_Learning_Multi-Dimensional_Human_Preference_for_Text-to-Image_Generation_CVPR_2024_paper.html.

CONTENTS

Abstract 1

1 Introduction 1

2 Related Work 2

3 Normal Fréchet Drift 3
3.1 Definition and Interpretation of NFD . . . 3
3.2 Reference Reconstruction for NFD . . . 3
3.3 NFD Estimation . . . 5
3.4 Error Analysis of NFD Estimation . . . 5

4 Experiments 6
4.1 Controlled Geometric Experiments . . . 6
4.2 NFD Estimates Across Post-Training Methods . . . 7
4.3 Robustness of NFD Estimation . . . 8
4.4 Relation to Existing Metrics and Perceptual Quality . . . 9

5 Discussion and Conclusion 9

AI use statement 10

Ethics statement 10

Reproducibility statement 10

References 10

A Notation and geometric preliminaries 16
A.1 Notation . . . 16
A.2 Geometric preliminaries . . . 17

B Theoretical results and interpretations 18
B.1 Proof of proposition 1 . . . 18
B.2 Comparing NFD with distribution and coverage metrics . . . 19
B.3 Proof of proposition 2 . . . 21
B.4 Sampling error and reconstruction bias . . . 21
B.5 Local reconstruction error bound . . . 22
B.6 Proof of Theorem 3 . . . 25

C Estimator implementations and pseudocode 26
C.1 Inputs, split discipline, and outputs . . . 26
C.2 Nearest neighbor . . . 26
C.3 Query-centered Single-PCA . . . 26
C.4 Bounded Multi-Patch construction . . . 27
C.5 Quality gate . . . 27
C.6 Aggregation, uncertainty, and resolution fields . . . 28
C.7 Computational complexity . . . 28

D Controlled geometry: experiments 1–7 29
D.1 Long-range tangential drift . . . 29
D.2 Averaging across manifold regions . . . 29
D.3 Finite patches block infinite-tangent shortcuts . . . 30

E Analytic separation from distribution and coverage metrics 33
E.1 Analytic separation map . . . 33

F Estimator theory validation and stress tests 33
F.1 Direct validation of the NFD error decomposition . . . 33
F.2 Evaluation sampling and the reconstruction-bias floor . . . 35
F.3 Patch-union Hausdorff convergence . . . 36
F.4 Tangent estimation convergence . . . 36
F.5 Anchor coverage and reference size . . . 37
F.6 Curvature stress with oracle-geometry selection . . . 37

F.7 Neighborhood, radius, anchor-count, and dimension ablations . . . 37
F.8 Separation between manifold regions . . . 38
F.9 Unified estimator comparison and quality-gate validation . . . 39

G Detection resolution and pretrained-only selection **41**
G.1 Normal-drift resolution sweep . . . 41
G.2 Oracle-free hyperparameter selection . . . 41
G.3 Practical resolution proxy . . . 42

H Optimization experiments **43**
H.1 Toy preference optimization . . . 43

I Real T2I protocol and complete quantitative results **45**
I.1 Checkpoints and sampling . . . 45
I.2 All primary effects . . . 45
I.3 Reference-refit uncertainty . . . 45
I.4 Estimator ablation and timing . . . 46
I.5 Standard-metric comparison . . . 47
I.6 Sample-level comparison with NN . . . 47
I.7 Absolute-resolution diagnostics . . . 47
I.8 Independent evaluators . . . 47
I.9 Blinded human evaluation . . . 48

J Independent FLUX.2 distillation validation **48**
J.1 Purpose and protocol . . . 48
J.2 Resolution and standard-metric controls . . . 49

K Empirical audit of local reference geometry **50**
K.1 Protocol and nulls . . . 50
K.2 Local structure and reconstruction . . . 50

L Visual interpretation **51**
L.1 Selection protocol . . . 51

M Reporting and interpretation **57**
M.1 Interpretation . . . 57

N Reproducibility details **57**
N.1 Replication materials . . . 57
N.2 Software and hardware . . . 57
N.3 Statistics and leakage control . . . 57

A NOTATION AND GEOMETRIC PRELIMINARIES

A.1 NOTATION

Table 2: Symbols used throughout the paper.

Symbol	Meaning
c, ν	condition (prompt) and its evaluation distribution
ϕ, Π, ψ	frozen encoder, pretrained-only projection/normalization, and composed representation $\psi = \Pi \circ \phi$
$Z_\theta, P_{\theta,c}, P_{0,c}$	generated feature, its probability distribution for fixed c, and the pretrained feature distribution
$\mathcal{M}_c$	ideal closed reference set; geometric bounds impose the stated manifold assumptions
$\mathcal{R}$	pretrained reference samples and their assignment to fitting and calibration banks (including cross-fit checks)
$\widehat{\mathcal{M}}_c$	frozen finite reconstruction fitted and calibrated using $\mathcal{R}$
$\mathsf{D}(P; A)$	expected squared distance $\mathbb{E}_{Z\sim P}[d_A^2(Z)]$ to A
$\mathcal{P}_2(\mathbb{R}^d), \mathrm{Cpl}(P, Q)$	distributions with finite second moment, and joint distributions with marginals P, Q
$W_2, \mathcal{G}(P)$	2-Wasserstein distance, and the Gaussian with the same mean and covariance as P
$\mathrm{NFD}(\theta \mid c)$	ideal value $\mathsf{D}(P_{\theta,c}; \mathcal{M}_c)$ at condition c
$\mathrm{NFD}_{\mathrm{rec}}(\theta \mid c, \mathcal{R})$	population value $\mathsf{D}(P_{\theta,c}; \widehat{\mathcal{M}}_c)$ for the frozen reconstruction at condition c
$\mathrm{NFD}_{\mathrm{rec}}(\theta \mid \mathcal{R})$	average $\mathbb{E}_{c\sim\nu}\,\mathrm{NFD}_{\mathrm{rec}}(\theta \mid c, \mathcal{R})$ over evaluation conditions
m, d, D	local tangent rank, retained feature dimension, and raw encoder dimension, respectively
N, M, K, J	fit-reference size, anchor count, local neighbors, evaluation samples
a_i, U_i	sampled anchor and fitted tangent basis $U_i \in \mathbb{R}^{d\times m}$
$P_T(x)$	orthogonal projector onto $T_x\mathcal{M}$, for $x \in \mathcal{M}$
h_i, ρ_i, ρ	local-PCA bandwidth (Kth-neighbor distance), patch radius, and common patch radius
λ_ρ	patch-to-bandwidth ratio ρ_i/h_i
δ_A	covering radius $\sup_{x\in\mathcal{M}} \min_i \|x - a_i\|$
$\epsilon_{\mathcal{M}}$	upper bound on Hausdorff reconstruction error $d_H(\mathcal{M}, \widehat{\mathcal{M}})$
$\widehat{\mathrm{NFD}}_{J,\mathrm{rec}}(\theta \mid c, \mathcal{R})$	sample mean of J squared distances at fixed condition c, defined in Eq. 5
$\widehat{\mathrm{NFD}}_{J,\mathrm{rec}}(\theta \mid \mathcal{R})$	sample mean across conditions $c_j \sim \nu$, defined in Eq. 6
$B_{\mathcal{R}}(\theta \mid c)$	reconstruction bias $\mathrm{NFD}_{\mathrm{rec}}(\theta \mid c, \mathcal{R}) - \mathrm{NFD}(\theta \mid c)$
$v_{\mathcal{R}}(\theta \mid c)$	variance $\mathrm{Var}_{P_{\theta,c}}[d^2_{\widehat{\mathcal{M}}_c}(Z)]$ of one squared distance with the reference frozen
$e(z)$	signed distance error $d_{\widehat{\mathcal{M}}_c}(z) - d_{\mathcal{M}_c}(z)$ at fixed c and $\mathcal{R}$
$\bar{a}_i, \epsilon_{A,i}, \epsilon_{T,i}$	projected anchor, its displacement from the observed anchor, and local tangent-projector error
$\Gamma_k, \Gamma_{P_{\theta,c}}$	component norms under $P_{\theta,c}$ and their sum, defined in Eq. 26
$\widehat{d}^2_{\mathrm{NN}}, \widehat{d}^2_{\mathrm{SPCA}}, \widehat{d}^2_{\mathrm{MP}}$	NN, query Single-PCA, and bounded Multi-Patch squared-distance estimators
BHR	boundary-hit rate: fraction whose selected patch has $\|U_i^\top(z - a_i)\| > \rho_i$
$\Delta_\ell(c, s \mid \mathcal{R})$	paired method-ℓ-minus-Base raw NFD for prompt c and seed s

Model-specific results retain the full notation $P_{\theta,c}$, $\mathcal{M}_c$, and $\widehat{\mathcal{M}}_c$. In general transport statements, P is an arbitrary distribution and A an arbitrary reference set; unsubscripted $\mathcal{M}$ denotes a generic manifold in geometric statements and examples. The arguments $(\theta \mid c, \mathcal{R})$ specify a fixed-condition reconstruction value, while $(\theta \mid \mathcal{R})$ indicates averaging over conditions. Reconstruction-dependent statements condition on the pretrained reference data $\mathcal{R}$.

A.2 Geometric preliminaries

NFD is defined for any nonempty closed reference set. Smooth manifold structure supports its local normal interpretation and the geometric reconstruction bounds. The assumptions for each error bound are stated with the corresponding result in App. B.

Tangent and normal spaces. An m-dimensional C^2 embedded manifold $\mathcal{M} \subset \mathbb{R}^d$ is a smooth set with m local coordinates (Lee, 2013, Chs. 1, 5). At $x \in \mathcal{M}$, the tangent space $T_x\mathcal{M}$ consists of velocities of smooth curves on $\mathcal{M}$ through x; the normal space $N_x\mathcal{M} = (T_x\mathcal{M})^\perp$ contains directions perpendicular to it. We write $P_T(x)$ for the tangent projector. The affine plane $x + T_x\mathcal{M}$ approximates the manifold locally, but a finite step along it can leave a curved manifold. Tangential drift refers to redistribution on $\mathcal{M}$, which leaves ideal NFD zero.

Reach and normal distance. For a nonempty closed set, reach is the supremum of radii $r > 0$ for which every point z with $d_{\mathcal{M}}(z) < r$ has a unique nearest point $\pi_{\mathcal{M}}(z)$ (Federer, 1959). For a closed C^2 manifold without boundary and with positive reach, if $x \in \mathcal{M}$, $n \in N_x\mathcal{M}$, and $\|n\| < \mathrm{reach}(\mathcal{M})$, then

$$\pi_{\mathcal{M}}(x + n) = x, \qquad d_{\mathcal{M}}(x + n) = \|n\|.$$

Within this neighborhood, distance to the manifold equals the length of the normal displacement from the nearest point.

Reference manifold, support, and reconstruction. The ideal manifold $\mathcal{M}_c$, the support of $P_{0,c}$, and the finite reconstruction $\widehat{\mathcal{M}}_c$ are distinct objects. An everywhere-positive feature density has support $\mathbb{R}^d$, so distance to that support is zero everywhere. The MP reconstruction $\widehat{\mathcal{M}}_c$ is a finite union of bounded planar patches and need not itself be a smooth manifold. Its distance remains well defined because $\widehat{\mathcal{M}}_c$ is nonempty and closed.

At fixed c and reference data $\mathcal{R}$, ideal NFD measures expected squared distance to $\mathcal{M}_c$, reconstruction NFD measures expected squared distance to $\widehat{\mathcal{M}}_c$, and the empirical estimate averages the latter distances over J outputs. More evaluation outputs improve this sample average; approximating ideal NFD also requires control of reconstruction error. App. B.4 makes this distinction explicit.

B THEORETICAL RESULTS AND INTERPRETATIONS

B.1 PROOF OF PROPOSITION 1

The 2-Wasserstein distance. The 2-Wasserstein distance measures the minimum root-mean-square displacement needed to transport one distribution to another (Villani, 2009). For $P, Q \in \mathcal{P}_2(\mathbb{R}^d)$, the distributions with finite second moment, define

$$W_2(P,Q) = \left(\inf_{\gamma \in \mathrm{Cpl}(P,Q)} \int \|z-y\|_2^2 \, \mathrm{d}\gamma(z,y) \right)^{1/2}, \tag{10}$$

where $\mathrm{Cpl}(P,Q)$ contains all joint distributions with marginals P and Q. Each joint distribution specifies how mass is transported; both marginals are fixed. The squared distance $W_2^2(P,Q)$ is the minimum average squared displacement.

Proof of Prop. 1. Write $C_A(P) = \inf_{Q \in \mathcal{P}_2(\mathbb{R}^d):\, Q(A)=1} W_2^2(P,Q)$. Fix $a_0 \in A$. Since $d_A(z) \le \|z - a_0\|$ and P has finite second moment, $\mathsf{D}(P;A)$ is finite.

For the lower bound, consider any admissible Q and any coupling γ of P and Q. Its second coordinate lies in A almost surely, so

$$\int \|z-a\|^2 \, \gamma(dz,da) \ge \int d_A^2(z) \, P(dz) = \mathsf{D}(P;A). \tag{11}$$

Taking the infimum over couplings and then over Q gives $C_A(P) \ge \mathsf{D}(P;A)$.

For the upper bound, we first construct a countable dense sequence in A. For each integer $n \ge 1$, partition $\mathbb{R}^d$ into a grid of half-open cubes of side length $1/n$, and choose one point from A in each cube that intersects A. There are countably many cubes at each scale and countably many scales, so the set of all selected points is countable. To see that it is dense in A, fix $a \in A$ and $\varepsilon > 0$, and choose n such that $\sqrt{d}/n < \varepsilon$. The cube containing a contains a selected point at distance at most $\sqrt{d}/n < \varepsilon$ from a. Enumerate the selected points as $(a_j)_{j \ge 1}$, allowing repetitions if there are only finitely many distinct points.

Using this sequence, for each integer $k \ge 1$ and $z \in \mathbb{R}^d$, define

$$j_k(z) = \min\{j \ge 1 : \ \|z - a_j\|^2 \le d_A^2(z) + 1/k\}, \qquad \pi_k(z) = a_{j_k(z)}.$$

Density gives $\inf_{j \ge 1} \|z - a_j\|^2 = d_A^2(z)$, so the positive tolerance $1/k$ ensures that the index set is nonempty for every z. The map is Borel measurable because d_A is continuous and the first-index selection uses countably many Borel sets. Moreover,

$$\|\pi_k(z)\| = \|z + \pi_k(z) - z\| \le \|z\| + d_A(z) + k^{-1/2} \le 2\|z\| + \|a_0\| + k^{-1/2}.$$

Let $Z \sim P$ and $Y_k = \pi_k(Z)$. The pushforward $Q_k = (\pi_k)_{\#}P$ is the probability distribution of Y_k: for every Borel set B,

$$Q_k(B) = \Pr(\pi_k(Z) \in B) = P\big(\{z \in \mathbb{R}^d : \pi_k(z) \in B\}\big).$$

Every member of the sequence (a_j) belongs to A, so $\pi_k(z) = a_{j_k(z)} \in A$ for every input $z \in \mathbb{R}^d$. The tolerance $1/k$ controls the distance approximation; the selected point still lies in A exactly. Taking $B = A$ therefore gives

$$Q_k(A) = P\big(\{z \in \mathbb{R}^d : \pi_k(z) \in A\}\big) = P(\mathbb{R}^d) = 1.$$

Thus all probability mass of Q_k lies in A, even if P assigns mass outside A.

The second moment of Q_k is finite as well. Applying $(a+b)^2 \le 2a^2 + 2b^2$ to the preceding norm bound gives

$$\int \|y\|^2 \, Q_k(\mathrm{d}y) = \mathbb{E}\|Y_k\|^2 \le 8\mathbb{E}\|Z\|^2 + 2\big(\|a_0\| + k^{-1/2}\big)^2 < \infty,$$

because $P \in \mathcal{P}_2(\mathbb{R}^d)$, a_0 is fixed, and $k \ge 1$. Thus Q_k is an admissible destination in the definition of $C_A(P)$. The joint law $(\mathrm{id}, \pi_k)_{\#}P$ of (Z, Y_k) has marginals P and Q_k, so it is a valid coupling. Its cost gives

$$C_A(P) \le W_2^2(P,Q_k) \le \mathbb{E}\|Z - Y_k\|^2 \le \mathsf{D}(P;A) + 1/k. \tag{12}$$

Letting $k \to \infty$ proves $C_A(P) = \mathsf{D}(P; A)$.

Finally, the nonnegative variable $d_A^2(Z)$ has zero expectation if and only if it is zero almost surely. Since A is closed, $d_A(z) = 0$ if and only if $z \in A$. Hence $\mathsf{D}(P; A) = 0$ if and only if $P(A) = 1$. □

Relation to the pretrained distribution. If $P_{0,c} \in \mathcal{P}_2(\mathbb{R}^d)$ satisfies $P_{0,c}(\mathcal{M}_c) = 1$, it is an admissible destination in Eq. 1. Therefore, for $P_{\theta,c} \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathrm{NFD}(\theta \mid c) \le W_2^2(P_{\theta,c}, P_{0,c}).$$

This compares NFD with transport to the *original* pretrained distribution. Replacing both distributions by Gaussian fits does not preserve this bound, so it does not imply $\mathrm{NFD}(\theta \mid c) \le \mathrm{FID}(P_{\theta,c}, P_{0,c})$.

B.2 Comparing NFD with distribution and coverage metrics

Ideal NFD measures a geometric quantity distinct from distribution matching and coverage. Equal FID values can correspond to different NFD values. Zero NFD allows both changes in probabilities on the manifold and loss of reference modes. Conversely, some divergence and overlap scores can remain unchanged as normal distance increases. The comparisons below establish these distinctions for population distributions.

B.2.1 Comparing FID and NFD

Equal-length translations along and perpendicular to a line have the same FID but different ideal NFD, as stated in Sec. 3.

Setup. Fix a condition c and let $Z_0 \sim P_{0,c}$ lie on the reference line $\mathcal{M}_c = \{(x, 0) : x \in \mathbb{R}\}$ almost surely, with $\mathbb{E}\|Z_0\|^2 < \infty$. Consider translated outputs $Z_\theta = Z_0 + u$ with distribution $P_{\theta,c}$. For the same length $\delta > 0$, take $u = (\delta, 0)$ for tangential drift or $u = (0, \delta)$ for normal drift.

FID under translation. Translation shifts the mean by u and preserves the covariance $\Sigma = \mathrm{Cov}(Z_0)$. Substituting these moments into the FID formula (Heusel et al., 2017, Eq. (6)) gives

$$\begin{aligned}\mathrm{FID}(P_{\theta,c}, P_{0,c}) &= \|u\|^2 + \mathrm{tr}\Big(2\Sigma - 2(\Sigma^{1/2}\Sigma\Sigma^{1/2})^{1/2}\Big) \\ &= \|u\|^2 = \delta^2.\end{aligned}$$

The covariance term vanishes because Σ is positive semidefinite, so $(\Sigma^{1/2}\Sigma\Sigma^{1/2})^{1/2} = (\Sigma^2)^{1/2} = \Sigma$. This identity also holds for singular covariances, as occurs for a distribution on a line.

NFD in the two directions. Distance to the reference line is $d_{\mathcal{M}_c}((z_1, z_2)) = |z_2|$. Tangential translation keeps every output on the line, whereas normal translation places every output at distance δ. Thus

$$\mathrm{NFD}(\theta \mid c) = \mathbb{E}\, d^2_{\mathcal{M}_c}(Z_\theta) = \begin{cases} 0, & u = (\delta, 0) \quad \text{(tangential drift)}, \\ \delta^2, & u = (0, \delta) \quad \text{(normal drift)}. \end{cases}$$

Neither FID nor NFD bounds the other. There is no general ordering between NFD and FID, even when $P_{0,c}(\mathcal{M}_c) = 1$. The tangential translation above gives NFD 0 and FID $\delta^2 > 0$. For the reverse ordering, let $P_{0,c}$ be uniform on the unit circle $\mathcal{M}_c \subset \mathbb{R}^2$ and let $P_{\theta,c} = \mathcal{N}(0, \frac{1}{2}I_2)$. Both distributions have mean zero and covariance $\frac{1}{2}I_2$, so their Gaussian fits coincide and FID is zero. The Gaussian output lies off the circle almost surely, giving strictly positive NFD.

B.2.2 OTHER DISTRIBUTION AND COVERAGE METRICS

Proposition 4 (NFD, distribution matching, and mode coverage). *Fix a condition c. Let $\mathcal{M}_c \subset \mathbb{R}^d$ be nonempty and closed, and let $P_{0,c}, P_{\theta,c} \in \mathcal{P}_2(\mathbb{R}^d)$ be the pretrained and generated feature distributions, with $P_{0,c}(\mathcal{M}_c) = 1$. As in Sec. 3, $\mathrm{NFD}(\theta \mid c) = \mathbb{E}_{Z_\theta \sim P_{\theta,c}} d^2_{\mathcal{M}_c}(Z_\theta)$.*

(i) ***Zero NFD allows distribution mismatch.*** *If $P_{\theta,c}(\mathcal{M}_c) = 1$ but $P_{\theta,c} \neq P_{0,c}$, then*

$$\mathrm{NFD}(\theta \mid c) = 0, \qquad W_2(P_{\theta,c}, P_{0,c}) > 0.$$

Also, $\mathrm{MMD}_k(P_{\theta,c}, P_{0,c}) > 0$ for any bounded characteristic kernel k.

(ii) ***Zero NFD allows mode dropping.*** *Let $P_{0,c}$ be uniform on the unit circle $\mathcal{M}_c \subset \mathbb{R}^2$, and let $P_{\theta,c}$ be uniform on a closed arc $B \subset \mathcal{M}_c$ of length $2\pi\alpha$, where $0 < \alpha < 1$. Then*

$$\mathrm{NFD}(\theta \mid c) = 0, \qquad P_{0,c}(\operatorname{supp} P_{\theta,c}) = \alpha.$$

(iii) ***Divergences can stay fixed as NFD increases.*** *Use the reference line and translated distribution from App. B.2.1, with $u = (0, \delta)$ and $\delta > 0$. Then*

$$\mathrm{NFD}(\theta \mid c) = \delta^2, \quad D_{\mathrm{KL}}(P_{\theta,c} \| P_{0,c}) = \infty,$$
$$\mathrm{TV}(P_{\theta,c}, P_{0,c}) = 1, \qquad \mathrm{JS}(P_{\theta,c}, P_{0,c}) = \log 2.$$

In (iii), total variation is normalized to $[0, 1]$ and logarithms are natural.

Proof. (i) Changing probabilities on the manifold leaves NFD zero. The assumption $P_{\theta,c}(\mathcal{M}_c) = 1$ means that every generated feature lies on $\mathcal{M}_c$ almost surely. Its distance to $\mathcal{M}_c$ is therefore zero, giving $\mathrm{NFD}(\theta \mid c) = 0$. However, W_2 is zero only when the two distributions agree (Villani, 2009). The same is true for MMD with a characteristic kernel (Gretton et al., 2012). Both the W_2 and MMD are consequently positive when $P_{\theta,c} \neq P_{0,c}$.

This distinction also holds for strictly convex f-divergences under an additional condition. Suppose $P_{\theta,c} \ll P_{0,c}$, meaning that any set of zero pretrained probability also has zero generated probability. The density ratio $w = \mathrm{d}P_{\theta,c}/\mathrm{d}P_{0,c}$ then satisfies $\mathbb{E}_{P_{0,c}} w = 1$. Since the distributions differ, w is not $P_{0,c}$-almost surely constant. For strictly convex $f : [0, \infty) \to \mathbb{R}$ with $f(1) = 0$, strict Jensen inequality gives

$$D_f(P_{\theta,c} \| P_{0,c}) = \mathbb{E}_{P_{0,c}} f(w) > f(\mathbb{E}_{P_{0,c}} w) = f(1) = 0,$$

with $+\infty$ also allowed for the left-hand side.

(ii) Staying on the manifold does not require covering all of it. In the circle example, every generated feature lies on $B \subset \mathcal{M}_c$, so NFD is zero. The generated support is exactly B. Because $P_{0,c}$ is uniform on the full circle, the reference probability retained on this support equals the fraction of the circumference covered:

$$P_{0,c}(\operatorname{supp} P_{\theta,c}) = P_{0,c}(B) = \frac{2\pi\alpha}{2\pi} = \alpha.$$

Thus the retained reference mass can be arbitrarily small while NFD remains zero.

(iii) Disjoint supports can have different separation distances. Every generated feature has second coordinate δ, so $d_{\mathcal{M}_c}(Z_\theta) = \delta$ and $\mathrm{NFD}(\theta \mid c) = \delta^2$. The reference and generated distributions lie on disjoint parallel lines. In particular,

$$P_{0,c}(\mathcal{M}_c) = 1, \qquad P_{\theta,c}(\mathcal{M}_c) = 0.$$

The generated distribution therefore assigns probability one outside $\mathcal{M}_c$, where the reference probability is zero, giving $D_{\mathrm{KL}}(P_{\theta,c} \| P_{0,c}) = \infty$.

For total variation, the event $B = \mathcal{M}_c$ attains the largest possible probability difference:

$$1 \geq \mathrm{TV}(P_{\theta,c}, P_{0,c}) = \sup_{B \text{ Borel}} \left|P_{\theta,c}(B) - P_{0,c}(B)\right| \\ \geq \left|P_{\theta,c}(\mathcal{M}_c) - P_{0,c}(\mathcal{M}_c)\right| = |0 - 1| = 1.$$

For JS, let $H = (P_{0,c} + P_{\theta,c})/2$. On each line, H assigns half the probability mass of the corresponding distribution. Hence each distribution has density ratio 2 relative to H on its own line, and both KL terms below equal $\log 2$:

$$\mathrm{JS}(P_{\theta,c}, P_{0,c}) = \tfrac{1}{2} D_{\mathrm{KL}}(P_{\theta,c} \| H) + \tfrac{1}{2} D_{\mathrm{KL}}(P_{0,c} \| H) = \log 2.$$

These divergence values are the same for every $\delta > 0$, whereas NFD grows as δ^2. □

Equal precision and recall can correspond to different NFD values. Precision and recall based on a fixed radius can also stay constant as normal distance changes. Use the same perpendicular translation, now assuming $\operatorname{supp} P_{0,c} = \mathcal{M}_c$, and fix $r > 0$. Define precision as the generated mass within distance r of the reference support, and recall as the reference mass within distance r of the generated support:

$$\mathrm{Prec}_r(P_{\theta,c}, P_{0,c}) = P_{\theta,c}\big(\{z : d_{\operatorname{supp} P_{0,c}}(z) \leq r\}\big), \\ \mathrm{Rec}_r(P_{\theta,c}, P_{0,c}) = P_{0,c}\big(\{z : d_{\operatorname{supp} P_{\theta,c}}(z) \leq r\}\big).$$

The two supports are parallel lines separated by δ, so every point on either line has distance δ to the other line. Both scores therefore equal 1 when $\delta \leq r$ and 0 when $\delta > r$. Two different shifts within either range give the same precision and recall, although their NFD values differ.

The exact $0/1$ result above assumes known supports and a single fixed radius r. In practice, nearest-neighbor precision/recall and Density/Coverage build neighborhoods from finite samples (Kynkäänniemi et al., 2019; Naeem et al., 2020), so their scores need not follow this rule. We compare these metrics with NFD in App. E (Tab. 5).

B.3 Proof of proposition 2

proof of Prop. 2. Fix anchor a, a column-orthonormal U, and $\mathcal{P} = \{a + Ut : \|t\| \leq \rho\}$. Let $r = z - a$, $t_0 = U^\top r$, and $n = (I - UU^\top)r$. Since $r = Ut_0 + n$ and $U^\top n = 0$,

$$\|z - (a + Ut)\|^2 = \|n\|^2 + \|t_0 - t\|^2. \tag{13}$$

The second term is minimized by Euclidean projection of t_0 onto the closed radius-ρ ball: $t = t_0$ if $\|t_0\| \leq \rho$, and $t = \rho t_0 / \|t_0\|$ otherwise. The minimum is $\|n\|^2 + (\|t_0\| - \rho)_+^2$, proving Eq. 4. □

B.4 Sampling error and reconstruction bias

For independent evaluation outputs and a frozen reference, mean squared NFD estimation error equals sampling variance plus squared reconstruction bias. We prove this identity, then bound the bias through distance error.

Fix the representation, condition c, and reference data $\mathcal{R}$. The ideal and reconstruction population values are

$$\mathrm{NFD}(\theta \mid c) = \mathsf{D}(P_{\theta,c}; \mathcal{M}_c) = \mathbb{E}_{Z \sim P_{\theta,c}}[d^2_{\mathcal{M}_c}(Z)], \\ \mathrm{NFD}_{\mathrm{rec}}(\theta \mid c, \mathcal{R}) = \mathsf{D}(P_{\theta,c}; \widehat{\mathcal{M}}_c) = \mathbb{E}_{Z \sim P_{\theta,c}}[d^2_{\widehat{\mathcal{M}}_c}(Z)].$$

Both depend on θ through $P_{\theta,c}$. The estimate $\widehat{\mathrm{NFD}}_{J,\mathrm{rec}}(\theta \mid c, \mathcal{R})$ in Eq. 5 averages J outputs from this distribution.

Proposition 5 (Sampling error and reconstruction bias). *Conditional on $\mathcal{R}$, let $Z_{\theta,1}, \dots, Z_{\theta,J}$ be independent draws from $P_{\theta,c}$. Define*

$$v_{\mathcal{R}}(\theta \mid c) = \mathrm{Var}_{P_{\theta,c}}[d^2_{\widehat{\mathcal{M}}_c}(Z)], \quad B_{\mathcal{R}}(\theta \mid c) = \mathrm{NFD}_{\mathrm{rec}}(\theta \mid c, \mathcal{R}) - \mathrm{NFD}(\theta \mid c).$$

If $\mathrm{NFD}(\theta \mid c) < \infty$ *and* $v_{\mathcal{R}}(\theta \mid c) < \infty$, *then*

$$\begin{aligned} &\widehat{\mathrm{NFD}}_{J,\mathrm{rec}}(\theta \mid c, \mathcal{R}) - \mathrm{NFD}(\theta \mid c) \\ &\quad = \underbrace{\widehat{\mathrm{NFD}}_{J,\mathrm{rec}}(\theta \mid c, \mathcal{R}) - \mathrm{NFD}_{\mathrm{rec}}(\theta \mid c, \mathcal{R})}_{\textit{sampling error}} + \underbrace{B_{\mathcal{R}}(\theta \mid c)}_{\textit{reconstruction bias}} \end{aligned} \tag{14}$$

$$\mathbb{E}_{\mathrm{eval}}\Big[\big(\widehat{\mathrm{NFD}}_{J,\mathrm{rec}}(\theta \mid c, \mathcal{R}) - \mathrm{NFD}(\theta \mid c)\big)^2 \mid \mathcal{R}\Big] = \frac{v_{\mathcal{R}}(\theta \mid c)}{J} + B_{\mathcal{R}}(\theta \mid c)^2 \tag{15}$$

Proof. The sample mean $\widehat{\mathrm{NFD}}_{J,\mathrm{rec}}(\theta \mid c, \mathcal{R}) = J^{-1}\sum_j d^2_{\widehat{\mathcal{M}}_c}(Z_{\theta,j})$ has conditional mean $\mathrm{NFD}_{\mathrm{rec}}(\theta \mid c, \mathcal{R})$ and variance $v_{\mathcal{R}}(\theta \mid c)/J$. Adding and subtracting this mean proves Eq. 14. Write $S_J = \widehat{\mathrm{NFD}}_{J,\mathrm{rec}}(\theta \mid c, \mathcal{R}) - \mathrm{NFD}_{\mathrm{rec}}(\theta \mid c, \mathcal{R})$ for the sampling error. Since $\mathbb{E}_{\mathrm{eval}}[S_J \mid \mathcal{R}] = 0$,

$$\begin{aligned} &\mathbb{E}_{\mathrm{eval}}[(S_J + B_{\mathcal{R}}(\theta \mid c))^2 \mid \mathcal{R}] \\ &\quad = \mathbb{E}_{\mathrm{eval}}[S_J^2 \mid \mathcal{R}] + 2B_{\mathcal{R}}(\theta \mid c)\mathbb{E}_{\mathrm{eval}}[S_J \mid \mathcal{R}] + B_{\mathcal{R}}(\theta \mid c)^2 \\ &\quad = v_{\mathcal{R}}(\theta \mid c)/J + B_{\mathcal{R}}(\theta \mid c)^2, \end{aligned}$$

which proves Eq. 15. □

Increasing the number of evaluation outputs J reduces sampling variance but leaves the fixed reconstruction bias unchanged.

From distance error to NFD bias. A bound on root mean squared distance error also bounds the reconstruction bias. Let $e(z) = d_{\widehat{\mathcal{M}}_c}(z) - d_{\mathcal{M}_c}(z)$ and assume $d_{\mathcal{M}_c}, d_{\widehat{\mathcal{M}}_c} \in L^2(P_{\theta,c})$. Expanding the squared distances gives

$$\begin{aligned} B_{\mathcal{R}}(\theta \mid c) &= \mathbb{E}_{P_{\theta,c}}[(d_{\mathcal{M}_c}(Z) + e(Z))^2 - d_{\mathcal{M}_c}(Z)^2] \\ &= 2\mathbb{E}_{P_{\theta,c}}[d_{\mathcal{M}_c}(Z)e(Z)] + \mathbb{E}_{P_{\theta,c}}[e(Z)^2]. \end{aligned} \tag{16}$$

For any finite $\Gamma \geq \|e\|_{L^2(P_{\theta,c})}$, Cauchy–Schwarz therefore yields

$$\begin{aligned} |B_{\mathcal{R}}(\theta \mid c)| &\leq 2\big|\mathbb{E}_{P_{\theta,c}}[d_{\mathcal{M}_c}e]\big| + \mathbb{E}_{P_{\theta,c}}[e^2] \\ &\leq 2 \cdot \|d_{\mathcal{M}_c}\|_{L^2(P_{\theta,c})} \cdot \|e\|_{L^2(P_{\theta,c})} + \|e\|^2_{L^2(P_{\theta,c})} \\ &\leq 2\sqrt{\mathrm{NFD}(\theta \mid c)}\Gamma + \Gamma^2, \end{aligned} \tag{17}$$

where $\|d_{\mathcal{M}_c}\|_{L^2(P_{\theta,c})} = \sqrt{\mathrm{NFD}(\theta \mid c)}$. The moment assumptions make all terms integrable.

The cross term in Eq. 16 can be negative, so reconstruction can raise or lower NFD. If ideal NFD is zero, however, $d_{\mathcal{M}_c}(Z) = 0$ almost surely and $B_{\mathcal{R}}(\theta \mid c) = \mathbb{E}_{P_{\theta,c}}[e(Z)^2] \geq 0$. The next subsection constructs Γ from the four local errors. App. B.6 then combines this bias bound with sampling variance.

B.5 Local reconstruction error bound

We bound the signed distance error

$$e(z) = d_{\widehat{\mathcal{M}}_c}(z) - d_{\mathcal{M}_c}(z).$$

Positive values mean overestimation, while negative values mean underestimation. We first control each direction for a single output, then average over outputs to obtain the geometric part of Thm. 3. Fix the representation, condition c, and reference data $\mathcal{R}$. Let $\mathcal{M}_c$ be a nonempty closed C^2 embedded manifold and $\widehat{\mathcal{M}}_c$ the finite MP union in Eq. 3, allowing $\rho_i = 0$.

Overestimation: from the true nearest point to a patch. Let $x(z)$ be a nearest point to z on $\mathcal{M}_c$. Going through $x(z)$ gives an upper bound on the distance to the reconstruction:

$$\begin{aligned} d_{\widehat{\mathcal{M}}_c}(z) &\le \|z - x(z)\| + d_{\widehat{\mathcal{M}}_c}(x(z)) \\ &= d_{\mathcal{M}_c}(z) + d_{\widehat{\mathcal{M}}_c}(x(z)). \end{aligned}$$

Subtracting the true distance yields, for any patch i,

$$e(z) \le d_{\widehat{\mathcal{M}}_c}(x(z)) \le d_{\widehat{\mathcal{P}}_i}(x(z)).$$

Thus, to control overestimation, we only need to bound how far the true nearest point lies from a patch. The following local model identifies four contributions to this distance.

For each anchor a_i, choose a nearest manifold point $\bar{a}_i$ and write $T_i = T_{\bar{a}_i}\mathcal{M}_c$. Within a local radius $H_i > 0$, describe the manifold by

$$\bar{a}_i + u + g_i(u) \in \mathcal{M}_c, \qquad u \in T_i, \quad \|u\| \le H_i, \quad g_i(u) \in T_i^{\perp}, \quad \|g_i(u)\| \le \beta_i \|u\|^2. \tag{18}$$

Here u moves along the true tangent plane and $g_i(u)$ accounts for bending. Since $g_i(0) = Dg_i(0) = 0$, the local quadratic bound controls the departure from the tangent plane. The anchor and tangent errors are

$$\epsilon_{A,i} = \|a_i - \bar{a}_i\|, \qquad \epsilon_{T,i} = \|U_i U_i^{\top} - P_T(\bar{a}_i)\|_{\mathrm{op}}. \tag{19}$$

Consider a patch whose local graph contains $x = x(z)$, and put $v = x - \bar{a}_i$. Then $v = u + g_i(u)$ with $u = P_T(\bar{a}_i)v$ and $\|u\| \le \|v\|$.

Curvature. Projecting x onto the true tangent plane removes the bending term, whose length satisfies

$$\|(I - P_T(\bar{a}_i))v\| = \|g_i(u)\| \le \beta_i \|u\|^2 \le \beta_i \|v\|^2. \tag{20}$$

This contribution is quadratic in the local length; for flat local geometry, one can take $\beta_i = 0$.

Tangent estimation. Replacing the true tangent projector by the fitted one adds at most $\epsilon_{T,i}\|v\|$:

$$\begin{aligned} \|(I - U_i U_i^{\top})v\| &\le \|(I - P_T(\bar{a}_i))v\| + \|(P_T(\bar{a}_i) - U_i U_i^{\top})v\| \\ &\le \beta_i \|v\|^2 + \epsilon_{T,i}\|v\|. \end{aligned} \tag{21}$$

This contribution is the direction error times the local length. More PCA samples alone do not guarantee smaller direction error, since neighborhood geometry also affects the fitted directions.

Coverage. The projection onto the fitted plane may exceed the patch radius. Clipping it to the radius-ρ_i ball centered at $\bar{a}_i$ requires an additional distance $(\|U_i^{\top} v\| - \rho_i)_+$. This term vanishes when the projection is covered.

Anchor deviation. Moving the patch center from $\bar{a}_i$ to the observed anchor a_i adds at most $\epsilon_{A,i}$. Combining this shift with the exact patch-distance formula in Eq. 4 gives

$$\begin{aligned} d_{\widehat{\mathcal{P}}_i}(x) &\le \|a_i - \bar{a}_i\| + d_{\widehat{\mathcal{P}}_i}(x + a_i - \bar{a}_i) \\ &= \epsilon_{A,i} + \sqrt{\|(I - U_i U_i^{\top})v\|^2 + (\|U_i^{\top} v\| - \rho_i)_+^2} \\ &\le \epsilon_{A,i} + \|(I - U_i U_i^{\top})v\| + (\|U_i^{\top} v\| - \rho_i)_+. \end{aligned} \tag{22}$$

The anchor term vanishes for on-manifold anchors; changing the PCA neighborhood size does not move the sampled anchor.

Together, the four contributions yield

$$e(z) \le d_{\widehat{\mathcal{P}}_i}(x) \le \beta_i \|v\|^2 + \epsilon_{T,i}\|v\| + \epsilon_{A,i} + (\|U_i^{\top} v\| - \rho_i)_+. \tag{23}$$

Among patches whose local graphs contain $x(z)$, choose $i_+(z)$ to minimize the four-term sum in Eq. 23, and define

$$\ell_+(z) = \|x(z) - \bar{a}_{i_+(z)}\|.$$

This comparison patch is used only to construct the bound.

Underestimation: from the scoring point to the manifold. Let $i_-(z)$ be the patch closest to z, selected by Eq. 4, and let $\hat{x}(z)$ be its nearest point to z. Thus $\|z - \hat{x}(z)\| = d_{\widehat{\mathcal{M}}_c}(z)$. Going through the scoring point gives

$$\begin{aligned} d_{\mathcal{M}_c}(z) &\le \|z - \hat{x}(z)\| + d_{\mathcal{M}_c}(\hat{x}(z)) \\ &= d_{\widehat{\mathcal{M}}_c}(z) + d_{\mathcal{M}_c}(\hat{x}(z)), \end{aligned}$$

and hence

$$-e(z) \le d_{\mathcal{M}_c}(\hat{x}(z)).$$

To control underestimation, we therefore need to bound how far the actual scoring point lies from the true manifold. The scoring patch i_- can differ from the comparison patch i_+.

Define $\ell_-(z) = \|\hat{x}(z) - a_{i_-(z)}\|$. Write $i = i_-(z)$, $v = \hat{x}(z) - a_i$, and $u = P_T(\bar{a}_i)v$. If $\ell_- \le H_i$, then $\|u\| \le \ell_- \le H_i$, so $y = \bar{a}_i + u + g_i(u)$ lies on $\mathcal{M}_c$. Since

$$\hat{x}(z) - y = (a_i - \bar{a}_i) + (v - u) - g_i(u),$$

the three corrections are anchor displacement, tangent error, and curvature:

$$\begin{aligned} -e(z) \le d_{\mathcal{M}_c}(\hat{x}(z)) &\le \|\hat{x}(z) - y\| \\ &\le \|a_i - \bar{a}_i\| + \|v - u\| + \|g_i(u)\| \\ &\le \epsilon_{A,i} + \epsilon_{T,i}\ell_- + \beta_i \ell_-^2. \end{aligned} \tag{24}$$

For the last line, $v = U_i U_i^\top v$ gives $v - u = (U_i U_i^\top - P_T(\bar{a}_i))v$, and Eq. 18 controls $g_i(u)$. No coverage correction is needed: $\hat{x}(z)$ already lies inside its patch, so there is no radius truncation.

Four component bounds. Tab. 3 summarizes the terms derived above, denoting them by $q_k^+(z)$ and $q_k^-(z)$ for each component k.

Table 3: **Four components of the distance-error bound.** For a single output z, $q_k^+(z)$ and $q_k^-(z)$ are nonnegative terms used to bound distance overestimation and underestimation, respectively. Under the conditions of Thm. 6, their sums satisfy $-\sum_k q_k^-(z) \le e(z) \le \sum_k q_k^+(z)$, where $e(z) = d_{\widehat{\mathcal{M}}_c}(z) - d_{\mathcal{M}_c}(z)$. The index $k \in \{\mathrm{cov}, \mathrm{curv}, \mathrm{tan}, \mathrm{anchor}\}$ labels the rows in order. $i_\pm$ and $\ell_\pm$ depend on z.

Component	Overestimation $q_k^+(z)$	Underestimation $q_k^-(z)$
Coverage	$(\|U_{i_+}^\top(x(z) - \bar{a}_{i_+})\| - \rho_{i_+})_+$	0
Curvature	$\beta_{i_+}\ell_+^2$	$\beta_{i_-}\ell_-^2$
Tangent estimation	$\epsilon_{T,i_+}\ell_+$	$\epsilon_{T,i_-}\ell_-$
Anchor deviation	ϵ_{A,i_+}	ϵ_{A,i_-}

Wherever both local arguments apply, these terms give

$$-\sum_k q_k^-(z) \le e(z) \le \sum_k q_k^+(z). \tag{25}$$

Use measurable nearest-point and minimizing-index selections wherever the local conditions hold, breaking finite index ties by the smallest index.

Average distance error and NFD bias. For each component k, average its squared bounds over outputs and take the square root:

$$\Gamma_k = \left\{\mathbb{E}_{Z \sim P_{\theta,c}}[(q_k^+(Z))^2 + (q_k^-(Z))^2]\right\}^{1/2}. \tag{26}$$

All expectations condition on $\mathcal{R}$. The next theorem combines the four component norms to bound average distance error and NFD bias.

Theorem 6 (Local MP error bound). *Let $e(z) = d_{\widehat{\mathcal{M}}_c}(z) - d_{\mathcal{M}_c}(z)$. Assume $\mathrm{NFD}(\theta \mid c) < \infty$, all four component norms in Eq. 26 are finite, and the following hold $P_{\theta,c}$-almost surely:*

(i) At least one local graph in Eq. 18 contains $x(z)$.

(ii) The scoring point stays within the valid local radius: $\ell_-(z) \le H_{i_-(z)}$.

Then

$$\begin{aligned} \Gamma_{P_{\theta,c}} &:= \Gamma_{\mathrm{cov}} + \Gamma_{\mathrm{curv}} + \Gamma_{\tan} + \Gamma_{\mathrm{anchor}}, \\ \|e\|_{L^2(P_{\theta,c})} &\le \Gamma_{P_{\theta,c}}, \\ |B_{\mathcal{R}}(\theta \mid c)| &\le 2\sqrt{\mathrm{NFD}(\theta \mid c)}\,\Gamma_{P_{\theta,c}} + \Gamma^2_{P_{\theta,c}}. \end{aligned} \tag{27}$$

Proof. Nearest points exist because $\mathcal{M}_c$ is nonempty and closed and $\widehat{\mathcal{M}}_c$ is a finite union of compact patches. Conditions (i)–(ii) ensure that both local arguments above apply almost surely. The two-sided bound in Eq. 25 and the triangle inequality for two-dimensional vectors give

$$|e(z)| \le \sqrt{\left(\sum_k q_k^+(z)\right)^2 + \left(\sum_k q_k^-(z)\right)^2} \le \sum_k \sqrt{(q_k^+(z))^2 + (q_k^-(z))^2}.$$

Taking the $L^2(P_{\theta,c})$ norm and applying its triangle inequality yields

$$\|e\|_{L^2(P_{\theta,c})} \le \sum_k \left\{\mathbb{E}_{P_{\theta,c}}[(q_k^+)^2 + (q_k^-)^2]\right\}^{1/2} = \sum_k \Gamma_k = \Gamma_{P_{\theta,c}}.$$

Because $d_{\widehat{\mathcal{M}}_c} \le d_{\mathcal{M}_c} + |e|$, the moment assumptions also give $d_{\widehat{\mathcal{M}}_c} \in L^2(P_{\theta,c})$. Substituting this distance-error bound into Eq. 17 proves the NFD bias bound. □

The theorem requires the two stated local conditions in addition to the empirical quality checks. Condition (ii) holds whenever $\rho_{i_-} \le H_{i_-}$. Zero-radius patches are allowed: then $\ell_- = 0$, and the table and proof still apply.

Larger radii reduce coverage gaps but permit scoring points farther from their anchors, which can increase curvature and tangent bounds. Reconstruction therefore requires balancing these components; Fig. 13 tests them individually.

B.6 Proof of Theorem 3

To complete Thm. 3, combine the local reconstruction bound with the sampling-variance identity from App. B.4.

Proof of Thm. 3. Under the main theorem's local and moment assumptions, Thm. 6 gives Eq. 8 and the bound on $|B_{\mathcal{R}}(\theta \mid c)|$. With conditionally iid evaluation outputs and finite $v_{\mathcal{R}}(\theta \mid c)$, substituting that bias bound into Eq. 15 yields

$$\begin{aligned} &\mathbb{E}_{\mathrm{eval}}\left[\left(\widehat{\mathrm{NFD}}_{J,\mathrm{rec}}(\theta \mid c, \mathcal{R}) - \mathrm{NFD}(\theta \mid c)\right)^2 \mid \mathcal{R}\right] \\ &\quad = \frac{v_{\mathcal{R}}(\theta \mid c)}{J} + B_{\mathcal{R}}(\theta \mid c)^2 \\ &\quad \le \frac{v_{\mathcal{R}}(\theta \mid c)}{J} + \left(2\sqrt{\mathrm{NFD}(\theta \mid c)}\,\Gamma_{P_{\theta,c}} + \Gamma^2_{P_{\theta,c}}\right)^2. \end{aligned}$$

This is Eq. 9. □

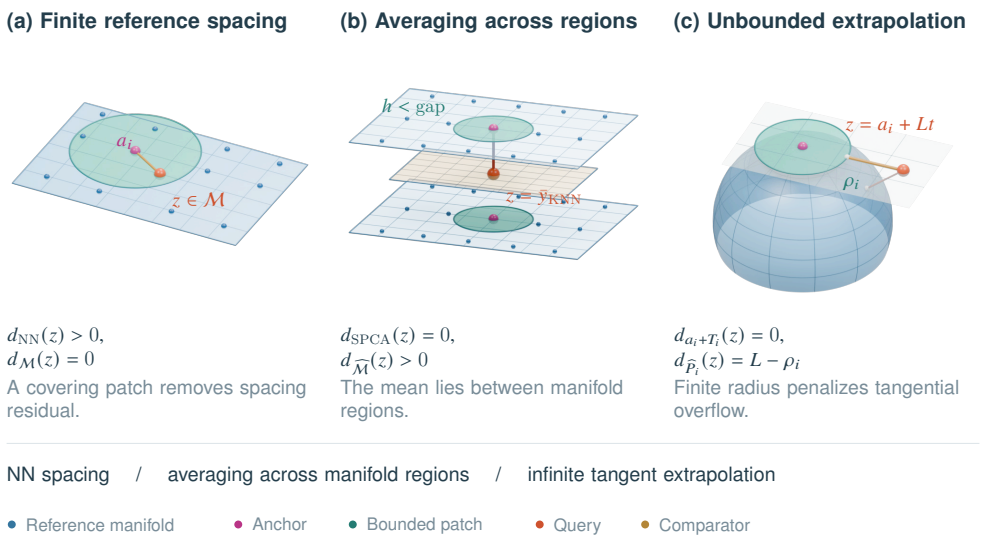

Figure 7: **Three structural estimator failures.** Nearest-neighbor distance contains finite-reference spacing even for an on-manifold query; query-centered PCA can place the mean of separated manifold regions at an off-manifold query; and an unbounded tangent plane assigns zero distance arbitrarily far from its observed neighborhood. Anchor-centered bounded patches address these mechanisms without using target samples to fit geometry.

C Estimator implementations and pseudocode

C.1 Inputs, split discipline, and outputs

Every estimator receives a fit reference matrix $Y \in \mathbb{R}^{N\times d}$ and query matrix $Z \in \mathbb{R}^{J\times d}$. Calibration and cross-fit references are never passed to the fine-tuned ranking stage during selection. All implementations return a vector of squared distances before aggregation, permitting paired statistics, prompt bootstraps, tail summaries, and boundary-hit diagnostics.

C.2 Nearest neighbor

$$\widehat{d}^2_{\text{NN}}(z) = \min_{1\le n\le N} \|z - y_n\|^2. \tag{28}$$

The implementation batches the pairwise squared-distance matrix and takes a minimum along the reference axis. With direct search, online complexity is $O(JNd)$ and memory is controlled by chunks along the query and reference axes. NN is the zero-radius special case of Multi-Patch. It cannot create a false zero from an erroneous tangent, but a point on the ideal reference set generally has positive distance because that set is represented by a finite bank.

C.3 Query-centered Single-PCA

For each query, find its K nearest references, compute their mean $\bar{y}_z$, and form the centered matrix X_z. If U_z contains the top m right singular vectors of X_z, then

$$\widehat{d}^2_{\text{SPCA}}(z) = \|(z - \bar{y}_z) - (z - \bar{y}_z)U_z U_z^\top\|^2. \tag{29}$$

The query-specific fit costs $O(JKd\min\{K, d\})$ without amortization. The method is useful as a flexible local baseline, but its query-dependent center and unbounded affine plane are precisely the two mechanisms targeted by our adversarial tests.

Algorithm 1 Anchor-centered bounded Multi-Patch NFD

Require: fit features B_{fit}; queries Z; anchors M; neighbors K; dimension m; radius rule $\rho_i = \lambda_\rho h_i$

choose $\{a_i\}_{i=1}^{M}$ by FPS on B_{fit}
for $i = 1, \dots, M$ **do**
  $\mathcal{N}_i \leftarrow K\text{NN}(a_i; B_{\text{fit}})$; $U_i \leftarrow$ top-m eigenvectors of centered $\mathcal{N}_i$
  compute eigengap and two-scale projector instability; if unreliable set $\rho_i \leftarrow 0$
for query $z \in Z$ **do**
  $q_i \leftarrow \|(I - U_i U_i^\top)(z - a_i)\|^2 + \big(\|U_i^\top (z - a_i)\| - \rho_i\big)_+^2$
  $\widehat{d}^2(z) \leftarrow \min_i q_i$
return $|Z|^{-1} \sum_{z \in Z} \widehat{d}^2(z)$ and diagnostics

Algorithm 2 Unified estimator, calibration, and paired Base-relative NFD workflow

Require: prompt set $\mathcal{C}$; pretrained generator G_0; checkpoints $\{G_\ell\}$; frozen representation $\psi = \Pi \circ \phi$; paired seeds $\mathcal{S}$

for prompt $c \in \mathcal{C}$ **do**
  generate disjoint pretrained banks $B_{\text{fit}}(c), B_{\text{cal}}(c), B_{\text{cross}}(c)$
  fit candidate NN, SPCA, or MP geometry using $B_{\text{fit}}(c)$ only
  compute calibration residual, coverage, eigengap, and cross-scale and cross-fit stability
  select hyperparameters without reading checkpoint outputs; freeze $\widehat{\mathcal{M}}_c$
  for seed $s \in \mathcal{S}$ **do**
    generate $x_\ell(c, s) = G_\ell(c, s)$ for all ℓ with identical sampler settings
    encode $z_\ell(c, s) = \psi(x_\ell(c, s))$ and compute $d_\ell^2(c, s) = d^2_{\widehat{\mathcal{M}}_c}(z_\ell(c, s))$
    compute $\Delta_\ell(c, s) = d_\ell^2(c, s) - d^2_{\text{Base}}(c, s)$
aggregate raw NFD; bootstrap prompts then seeds for paired intervals
report encoder, estimator, calibration, BHR, cross-fit disagreement, and resolution ratio

C.4 Bounded Multi-Patch construction

See Fig. 2 for the workflow and Alg. 1 for pseudocode. The construction is:

1. Select M observed references by greedy FPS. The first anchor is selected with a fixed seed; each later anchor maximizes its distance to the selected set.
2. At anchor a_i, retrieve K nearest fit references. Let h_i be the farthest selected-neighbor distance.
3. Subtract the local mean only for covariance estimation. Compute the top m singular directions U_i and eigengap $\lambda_m - \lambda_{m+1}$.
4. Set the geometric center to a_i and nominal radius $\rho_i = \lambda_\rho h_i$.
5. Compute a second tangent projector at a neighboring scale. If the local count, eigengap, or projector-stability rule fails, replace the disk by its anchor point ($\rho_i = 0$); do not delete it.

The dense query kernel broadcasts residuals $R_{ji} = z_j - a_i$. Tangent coordinates are $T_{ji} = U_i^\top R_{ji}$, normal energy is $\|R_{ji}\|^2 - \|T_{ji}\|^2$ (clamped at zero only against floating-point roundoff), and radial excess is $(\|T_{ji}\| - \rho_i)_+^2$. Taking the minimum over anchors implements Eq. 4 exactly.

C.5 Quality gate

The quality gate is a safeguard against *false low distance*. For every anchor it records (i) neighborhood size; (ii) normalized eigengap; (iii) projector instability $\|\widehat{P}_h - \widehat{P}_{h/2}\|_{\text{op}}$; and optionally (iv) stress tests for region separation and coverage. Here $\widehat{P}_h = U_h U_h^\top$ is the fitted projector at neighborhood scale h. Thresholds are selected using pretrained samples and fixed before target evaluation. A gate failure sets $\rho_i = 0$. This point fallback preserves the anchor as an upper-bounding discrete candidate and avoids holes that deleting the patch would create. The corresponding false-zero, coverage, and fallback results are reported in App. F.9.

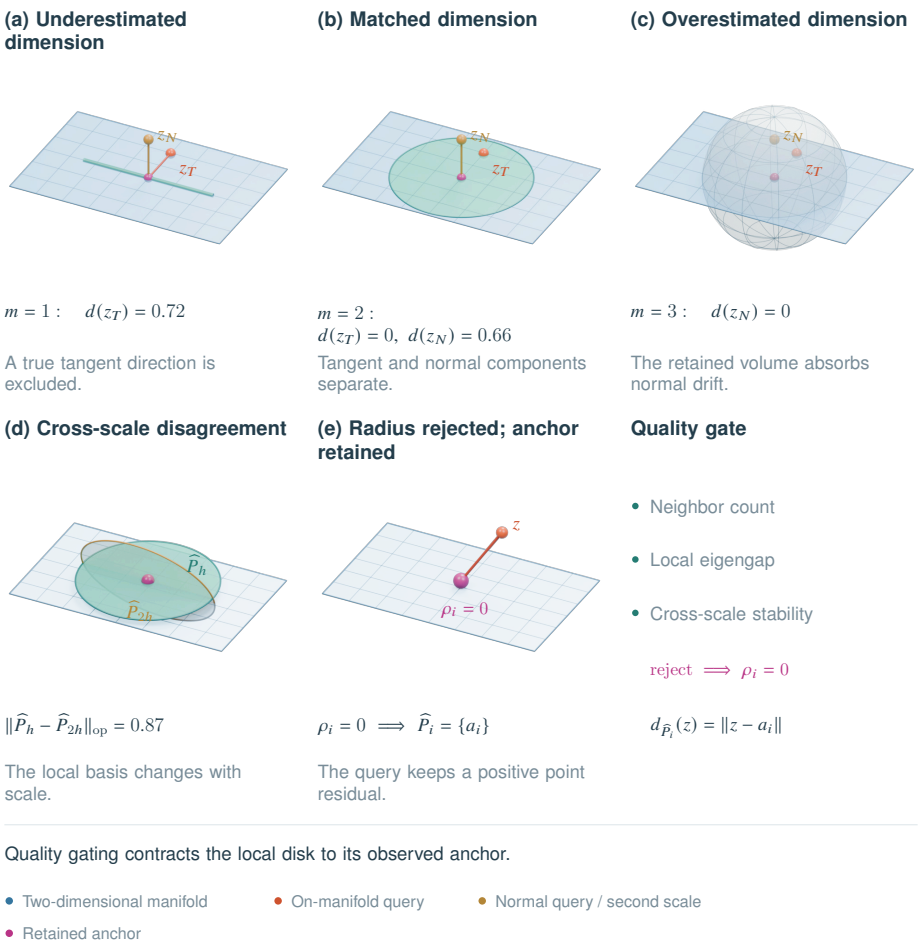

Figure 8: **Retained dimension and pretrained-only quality control.** Underestimating m excludes a genuine tangent direction, whereas overestimating it can absorb normal drift. Cross-scale projector disagreement detects unstable local bases. When a patch fails the gate, setting $\rho_i = 0$ contracts it to its observed anchor, retaining a positive point residual instead of deleting coverage.

C.6 Aggregation, uncertainty, and resolution fields

Raw NFD is the arithmetic mean of squared query distances and RMS-NFD its square root. Real comparisons first form $\Delta_\ell(c, s)$ at the paired unit. The primary hierarchical bootstrap samples prompts with replacement and paired seeds within prompt. All methods remain paired in a replicate. Unless otherwise specified in the experiment protocol, real NFD intervals in this paper use $B = 10{,}000$ percentile replicates from the same implementation and a fixed seed. The resolution record includes held-out calibration RMS, coverage gap, tangent instability, curvature proxy, cross-fit disagreement, a conservative safe resolution ϵ_{safe}, and the ratio RMS-NFD/ϵ_{safe}. Resolution diagnostics assess the scale of drift relative to reconstruction uncertainty (App. G).

C.7 Computational complexity

Ignoring search-index acceleration, MP construction costs $O(MNd)$ for FPS distance updates plus local neighbor search and M small SVDs. Online scoring costs $O(JMdm)$ and is fully batched. The

fit is amortized across every checkpoint evaluated against the same pretrained reconstruction. See Tab. 9 for measured construction and query costs under the real $d = 16$, $M \leq 48$ protocol.

D CONTROLLED GEOMETRY: EXPERIMENTS 1–7

The seven tests establish three properties: sensitivity to normal drift, invariance to tangential drift, and avoidance of false zeros from local-plane fitting (Tab. 4). All use double precision on one RTX 4090 and fixed seeds. Analytic geometry and dense meshes supply evaluation truth, not estimator fit references. Fig. 12 illustrates the underlying geometries.

Table 4: Controlled experiments and their main findings. Detailed settings are given below and in the replication materials. Errors are squared-distance errors unless identified as distance errors.

Exp.	Geometry	Question	Main finding
1	Line	Separate normal drift from tangential drift	Normal MAE 2.17×10^{-19} and tangential MP maximum 2.86×10^{-4}
2	Line distributions	Distinguish equal-FID shifts	FID differs by at most 2.36×10^{-16}, while population NFD separates the shifts
3	Unit circle	Preserve long-range tangential drift	MP maximum 5.17×10^{-9}, versus 4 for the fixed-base normal chord
4	Parallel lines	Avoid mixing separated regions	SPCA gives a false zero, while bounded MP recovers the true value a^2
5	Three-petal flower	Resolve medial-axis ambiguity	Center distance error 5.97×10^{-4} and no MP false zeros
6	Circle, flower, parabola	Limit tangent extrapolation	Bounded-patch distance MAE is 5.1–6.3× lower than for the unbounded plane
7	Circle, flower, sphere	Estimate NFD under normal drift	Pooled estimated-versus-truth correlation 0.997 and MAE 0.002091

The line tests use displacement magnitudes $a \in \{0.02, 0.05, 0.1, 0.2, 0.5, 1\}$. Exp. 1 evaluates $J = 4096$ queries with nested banks of $N = 200, 500, 1000, 5000$, $K = 32$, and $\lambda_\rho = 0.75$; error falls $15.76\times$ between the smallest and largest banks. Exp. 2 uses 8192 paired queries, $N = 5000$, and $M = 1024$ on the same displacement grid.

For controlled normal drift (Exp. 7), $\delta \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.3\}$, with up to $N = 12000$ and $M = 600$. Correlations with truth are 1.000, 0.999999, and 0.999566 for the circle, flower, and sphere. Flower locations are filtered through the dense oracle so each perturbation remains in a numerically unique normal tube.

D.1 LONG-RANGE TANGENTIAL DRIFT

Bounded MP stays near zero under long-range tangential drift, whereas a fixed-base normal projection does not. On the unit circle, $z_\varphi = (\cos\varphi, \sin\varphi)$ moves from $x_0 = (1, 0)$ through eight angles between 0 and π (Fig. 9). True squared distance remains zero. The squared normal component of the chord relative to the *starting* tangent is $(\cos\varphi - 1)^2$, reaching 4 at the antipode. With $N = 4096$ and $M = 256$, bounded MP has maximum squared distance 5.17×10^{-9}. The contrast shows why a local normal decomposition cannot be used as a fixed-base global metric. Tangent and normal-space definitions are given in App. A.2.

D.2 AVERAGING ACROSS MANIFOLD REGIONS

Observed anchors avoid the false zeros caused by averaging separated manifold regions. For parallel lines $y = \pm a$, symmetry places a query-centered neighborhood mean at the off-manifold midpoint, giving SPCA zero residual despite true squared distance a^2. An anchor-centered patch on either reference line recovers this separation. Exp. 4 tests $a = 0.05, 0.1, 0.2, 0.5, 1$ with 2001 points per line and retains all radius ties.

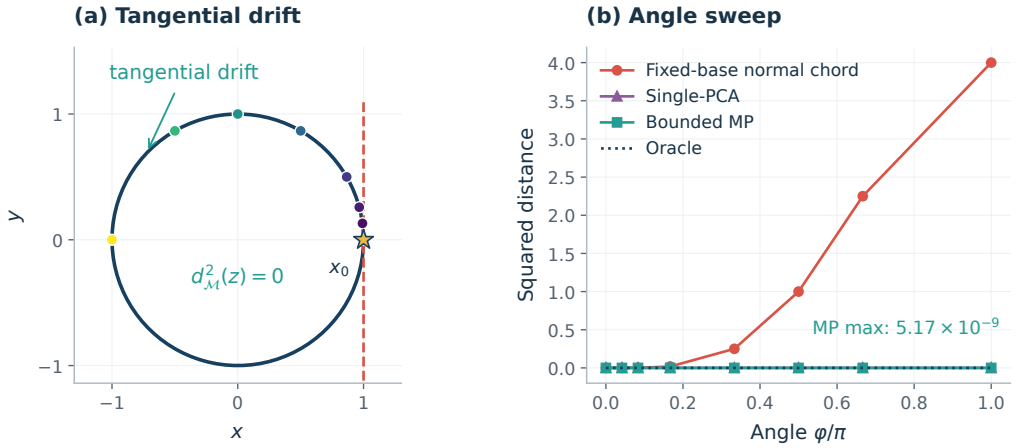

Figure 9: **Long-range tangential drift (Exp. 3).** (a) Queries move around the unit circle; the star is the starting point and the dashed line its tangent. (b) All eight tested angles are shown. The fixed-base normal-chord surrogate grows although true NFD remains zero; finite-reference estimators stay near zero. The plotted quantities are squared distances, not arc lengths.

Distance at the flower center. The flower extends this mechanism to one connected smooth curve. For $r(\theta) = 1 + \beta\cos(3\theta)$ with $0 < \beta < 1$, the distance from the origin to the curve point at angle θ is $r(\theta)$. Hence

$$d_{\mathcal{M}}(0) = \min_{\theta} r(\theta) = 1 - \beta, \tag{30}$$

attained at $\theta \in \{\pi/3, \pi, 5\pi/3\}$. At the center, these three nearest points are equally valid; the distance is well defined even though the nearest-point map is not unique. For $\beta = 0.45$, $d_{\mathcal{M}}(0) = 0.55$ and $d^2_{\mathcal{M}}(0) = 0.3025$. These are distinct from the mean squared distance over a query group.

The three nearest-point vectors sum to zero. A threefold-symmetric query-centered neighborhood therefore has mean at the off-manifold query and again produces an exact SPCA zero; sampled anchors do not use the mean of these separated manifold regions as a patch center.

The bounded patch estimates center distance 0.549403, with error 5.97×10^{-4} (Fig. 3c). Exp. 5 uses $N = 12000$, $M = 600$, $K = 96$, and a 131072-point oracle. Queries comprise the center, six symmetry axes with 60 radii, 720 on-curve points, and 360 locations at three normal offsets. A false zero means true distance at least 0.05 and estimated distance at most 0.01; bounded MP has none. Fig. 10 reports all four regimes, including the dense NN baseline, which is also accurate at the center.

D.3 Finite patches block infinite-tangent shortcuts

Bounding the tangent patch reduces distance MAE by factors 5.27, 6.30, and 5.11 on the circle, flower, and parabola, respectively. Exp. 6 holds one observed anchor and its local PCA direction fixed, then places queries at $a + L\widehat{u}$ for four lengths per geometry (Fig. 11). Every query lies on the unbounded fitted tangent, which therefore gives zero distance even when the query is off the true curve. Once $L > \rho$, the bounded distance is $L - \rho$ and its square is $(L - \rho)^2$, exactly the tangential overflow term in Eq. 4. Coverage and radius selection are studied in App. F.

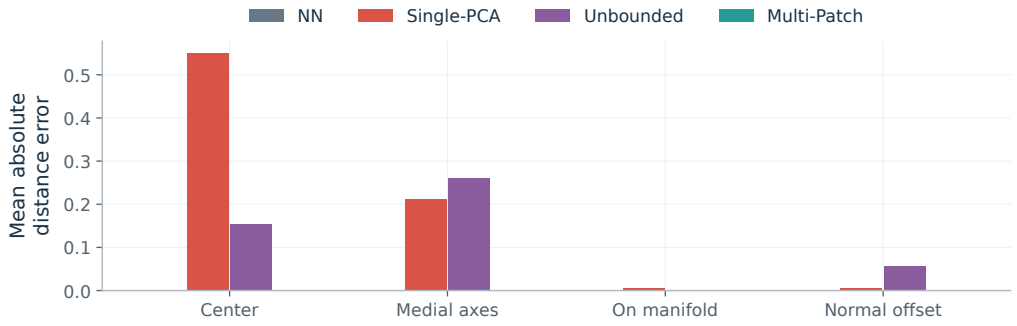

Figure 10: **Flower medial-axis errors (Exp. 5).** Mean absolute distance error across the center, medial axes, on-manifold points, and points undergoing normal drift, including NN and unbounded-patch controls. NFD itself averages squared distances.

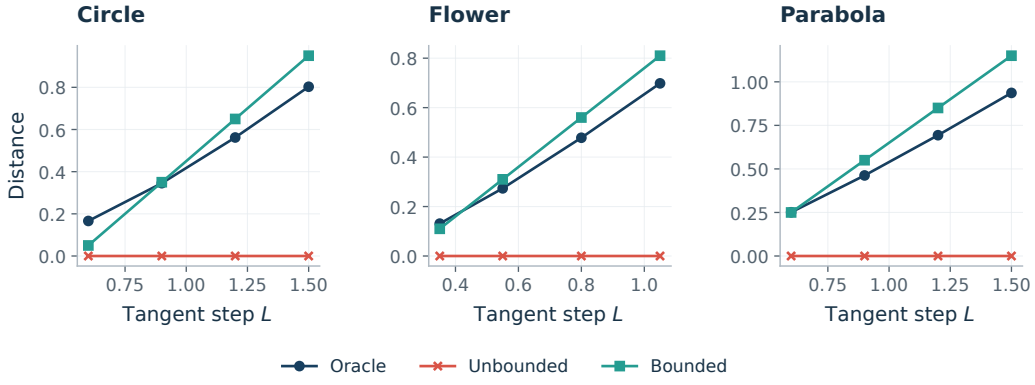

Figure 11: **Bounded versus unbounded tangent extrapolation (Exp. 6).** Complete distance sweeps on three geometries with one fixed local patch each. Unbounded distance stays zero; bounding penalizes queries beyond the segment endpoint and follows the oracle more closely. The full MP estimator minimizes squared distance over all patches.

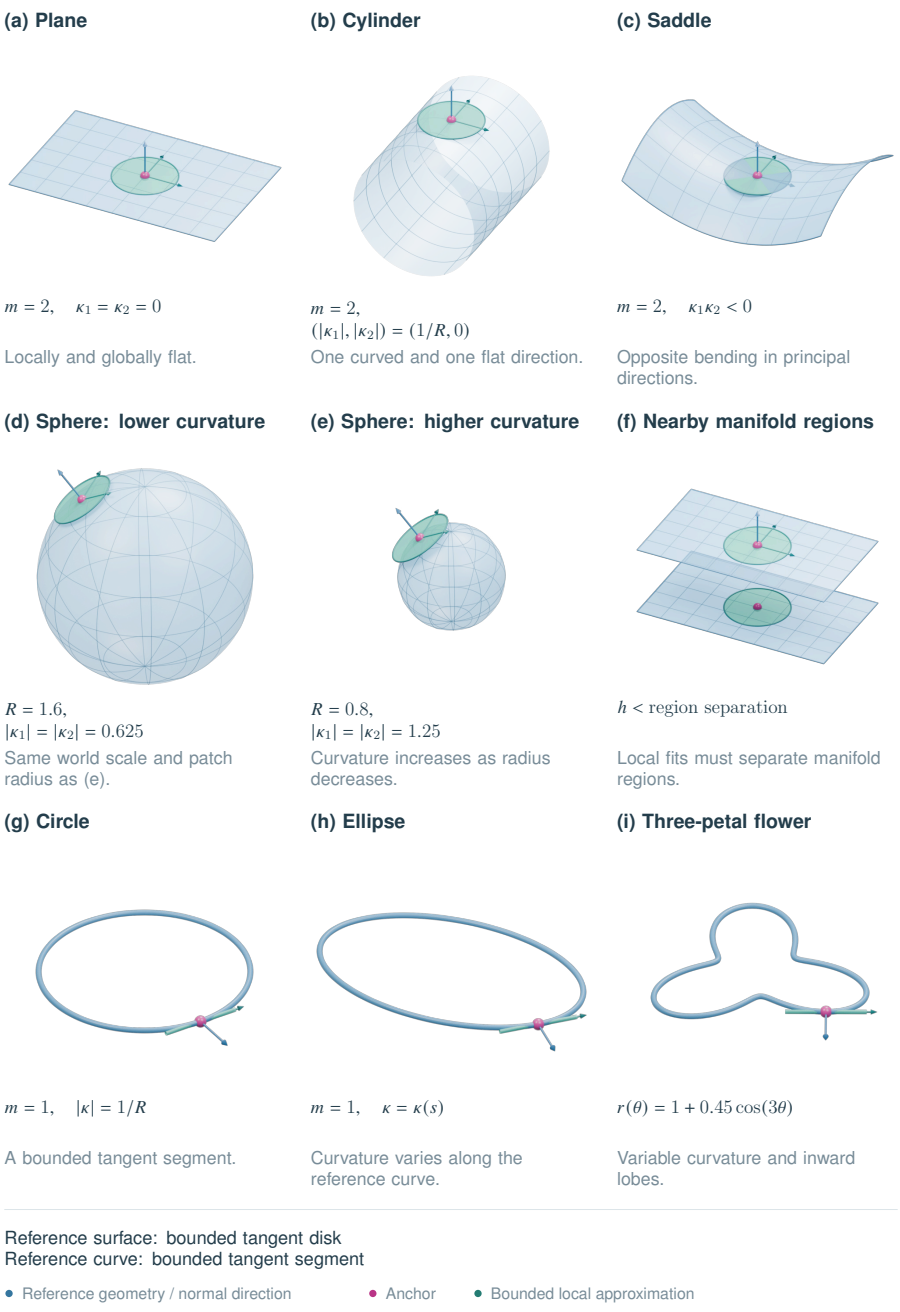

Figure 12: **Atlas of local manifold geometry.** Plane, cylinder, saddle, low- and high-curvature spheres, nearby manifold regions, circle, ellipse, and flower illustrate how intrinsic dimension, curvature, reach, and separation between manifold regions affect a bounded local tangent approximation. Every panel is schematic; see Apps. D–F for the experimental configurations.

E ANALYTIC SEPARATION FROM DISTRIBUTION AND COVERAGE METRICS

Distributional change does not uniquely identify normal departure. The constructions below separate these quantities on a known circle; App. I.5 tests their empirical association on generated images.

E.1 ANALYTIC SEPARATION MAP

Tangential redistribution changes distributional and coverage metrics while leaving ideal NFD zero, whereas a normal shift produces positive NFD (Tab. 5). The circle construction uses 2,048 reference and 2,048 generated points. The reference circle is known analytically. Empirical distribution metrics are computed in the same two-dimensional coordinates with $k = 5$ for precision, recall, density, and coverage, a median-bandwidth RBF kernel, and 256 fixed sliced-Wasserstein directions. Small negative unbiased MMD estimates are sampling noise and are not clipped.

Table 5: Circle separation map. NFD measures normal drift; the other columns can also respond to tangential drift or finite-radius overlap.

Scenario	NFD	MMD^2	SW_2^2	Prec.	Recall	Coverage
Identical	0	−0.00035	0	1.000	1.000	1.000
Tangential drift: reweight	0	0.050	0.116	1.000	1.000	1.000
Tangential drift: mode drop	0	0.127	0.291	1.000	0.502	0.503
Normal drift $\delta = 0.04$	0.00160	0.00004	0.00080	0	0	0

F ESTIMATOR THEORY VALIDATION AND STRESS TESTS

The validation follows the theory: local reconstruction errors determine bias, evaluation size controls sampling variance, and reference geometry controls the remaining error. We then stress the geometric design choices and compare estimators under the same toy cases.

F.1 DIRECT VALIDATION OF THE NFD ERROR DECOMPOSITION

CPU/float64 tests confirm the decomposition and local bound in Eq. 9 to numerical tolerance on queries satisfying the chart conditions. We first isolate the geometric error sources, then test their interaction in fitted reconstructions and compare local with uniform bounds. Evaluation-sampling error is tested separately in App. F.2.

Oracle geometry and integration. For a line, $d_{\mathcal{M}}(z) = |z_2|$; for a radius-$R$ circle, $d_{\mathcal{M}}(z) = |\|z\| - R|$. Main circle queries have independent uniform angle and outward displacement $\delta \sim \mathrm{Unif}[0, 0.2R]$, giving ideal NFD $(0.2R)^2/3$. We use fixed seeds for numerical integration and Monte Carlo experiments. We approximate population moments using a scrambled Sobol grid of 2^{15} points and compare against its 2^{14}-point prefix. Across the 60 reconstructions, the largest changes are 4.79×10^{-7} in bias and 1.24×10^{-5} in the bias bound.

For circle charts we use $H_i = R/2$ and $\beta_i = 1/R$. Indeed, for $|u| \le R/2$, the graph remainder is $R - \sqrt{R^2 - u^2} = u^2/(R + \sqrt{R^2 - u^2}) \le u^2/R$. A chord distance at most $R/2$ is a sufficient upper-chart membership check, and the selected patch length is checked against H_i in the lower direction. For a line, $\beta_i = 0$ and the chart is unrestricted. No geometric constants are fitted to NFD errors.

Isolating the components. Exact covered line patches have zero bias and zero component bounds to numerical precision. To isolate the individual error sources, line queries use $z_1 \sim \mathrm{Unif}[-1, 1]$ and $z_2 \sim \mathrm{Unif}[0, 0.2]$. Four anchors at $-0.75, -0.25, 0.25, 0.75$ isolate coverage as their common radius varies over $0, 0.25, 0.5, 1$. A single radius-2 patch isolates tangent rotations of $0, 0.05, 0.1, 0.2$ radians, or normal anchor shifts of $0, 0.02, 0.05, 0.1$. To vary curvature while holding physical scales fixed, a radius-0.45 tangent patch at $(R, 0)$ scores circle queries with arc coordinate uniform on $[-0.4, 0.4]$ and independent outward displacement uniform on $[0, 0.2]$, for $R = 1, 2, 5$. Fig. 13 reports absolute bias and its bound; the raw results also retain signed bias and both terms of Eq. 16.

Zero-radius patches pass the local check and expose the coverage term. The largest algebraic identity residual is 2.23×10^{-16}; on chart-valid queries, the largest local-inequality violation is 1.39×10^{-17}, both below the 10^{-10} numerical tolerance (proof: App. B.5).

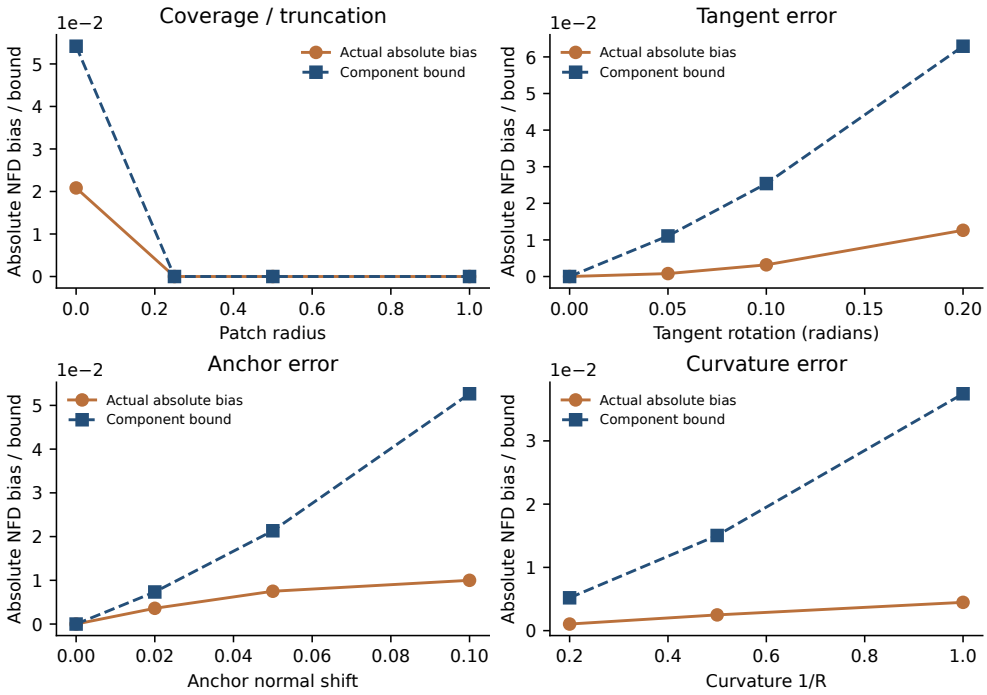

Figure 13: **Local error sources in isolation.** Coverage, tangent direction, anchor position, and curvature each affect NFD bias. Analytic local-geometry bounds track the four error sources.

Fitted MP and the radius tradeoff. Reference density exposes the predicted coverage–local-accuracy tradeoff at fixed anchor count and neighborhood size (Tab. 6). We fit 20 independent uniform reference banks per $N \in \{256, 1024, 4096\}$ with FPS initialized at the first bank sample, $M = 32$, centered local PCA on $K = 16$ neighbors including the anchor, and $\rho_i = h_i$. This is a fixed-parameter MP ablation: no oracle tuning or additional quality rejection is used. The separate zero-radius test exercises the estimator's fallback geometry. Every fitted case satisfies the local conditions on the evaluation grid.

Table 6: Median results across 20 reference fits per N. Component columns have distance units; bias and its bound have squared-distance units. The anchor term is zero up to floating-point precision.

N	Absolute bias	Bias bound	Γ_{cov}	Γ_{curv}	$\Gamma_{\tan}$
256	2.97×10^{-3}	1.21×10^{-2}	0	3.98×10^{-2}	3.95×10^{-3}
1024	3.21×10^{-4}	6.91×10^{-3}	2.13×10^{-2}	4.97×10^{-3}	3.96×10^{-4}
4096	2.38×10^{-3}	1.44×10^{-2}	4.66×10^{-2}	4.32×10^{-3}	6.71×10^{-5}

At fixed K, denser banks shrink h_i and the patch radii. Curvature and tangent errors decrease, but the fixed number of patches leaves more of the intervening manifold uncovered, explaining the nonmonotone bias.

An unused remote patch. An unused remote patch weakens the global Hausdorff bound without changing NFD or its local bound. We use one fixed $N = 1024$ reconstruction with query angles restricted to $[3\pi/4, 5\pi/4]$ and append a zero-radius patch at $(L, 0)$ for $L = 4, 8, 16, 32$. It is never selected for scoring or for the upper local envelope. The baseline manifold-to-patch distance is at most 2, and its patch-to-manifold distance is bounded by $\max_i \rho_i < 1$. The added point is at distance

$L - 1 > 2$ from the unit circle, so the union's Hausdorff error is exactly $L - 1$. The global NFD bound therefore increases from 9.69 to 968.16, while the local bound remains 7.20×10^{-3} and actual absolute bias remains 4.12×10^{-4}. Scores are unchanged to machine precision.

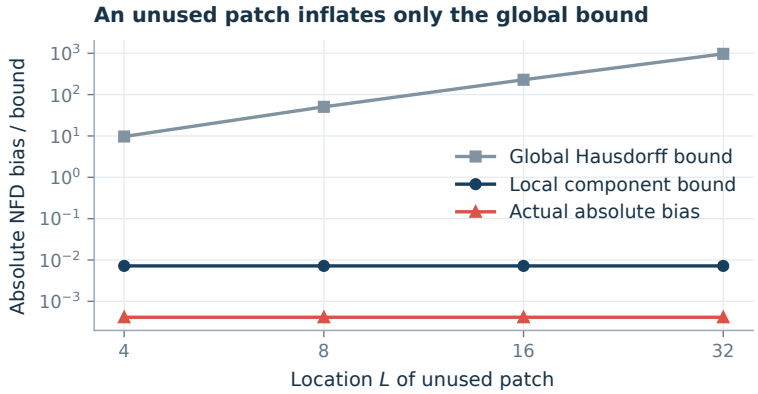

Figure 14: **A remote unused patch inflates the global bound without changing NFD.** Adding a point patch at $(L, 0)$ away from the evaluated circle region leaves actual bias and the local component bound unchanged. The global Hausdorff bound grows with L; the local bound remains conservative. All values use the same reference reconstruction and query grid.

F.2 Evaluation sampling and the reconstruction-bias floor

These two fixed-reconstruction tests distinguish decreasing sampling variance from persistent reconstruction bias.

Standard-error scaling. With reconstruction fixed ($N = 10000$, $M = 512$) and $J \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$, 30 independent evaluation repeats yield slope -0.513 in $\log \mathrm{sd}(\widehat{\mathrm{NFD}})$ versus $\log J$, close to the $-1/2$ prediction.

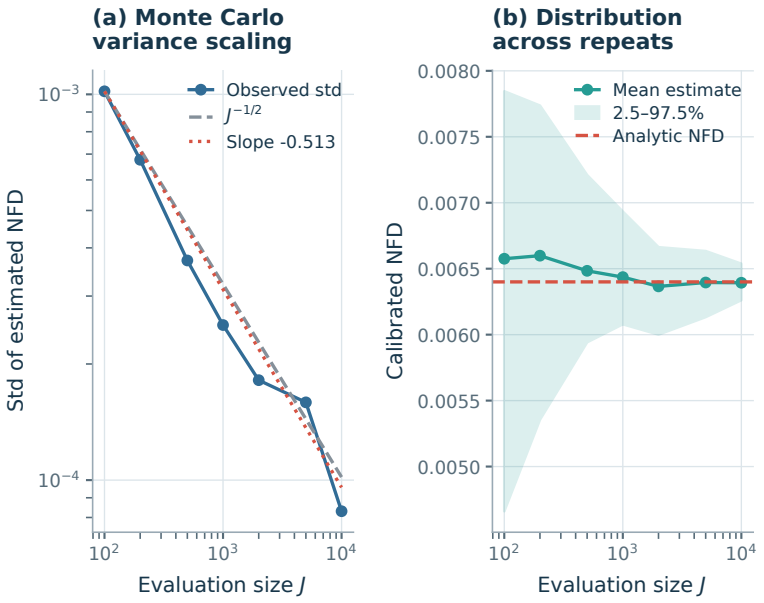

Figure 15: Finite-evaluation standard error follows the predicted $J^{-1/2}$ law.

MSE decomposition and the bias floor. Increasing evaluation size reduces MSE toward the reconstruction-bias floor rather than toward zero. We hold one fitted reconstruction fixed for each $N \in \{256, 1024, 4096\}$ from App. F.1. For each $J \in \{32, 128, 512, 2048\}$ we draw 200 independent evaluation replicates (2,400 in total). We compare empirical squared error against the analytic ideal NFD with the prediction in Eq. 15, using integrated variance and bias. The empirical MSE differs from the prediction by at most 2.31 estimated Monte Carlo standard errors across the 12 settings. For $N = 4096$, the predicted bias floor is 5.85×10^{-6} and observed MSE at $J = 2048$ is 5.90×10^{-6}. Fig. 4a displays this floor; the complete replicate data retain the finite Monte Carlo discrepancies.

F.3 Patch-union Hausdorff convergence

Patch-union Hausdorff error decreases on all five tested manifolds (Fig. 16). Circle, ellipse, flower, sphere, and saddle references use $N \in \{200, 500, 1000, 2000, 5000, 10000, 20000\}$. Curves take $M = \min(N, \lceil 8\sqrt{N} \rceil)$ and surfaces $M = \min(N, \lceil 4\sqrt{N} \rceil)$; $\rho = 1.10\delta_A$. A dense 8192-point true mesh and patch meshes evaluate both directed distances. Final Hausdorff errors include 8.58×10^{-4} (circle), 1.40×10^{-3} (flower), and 4.21×10^{-2} (sphere). We report finite-grid convergence slopes.

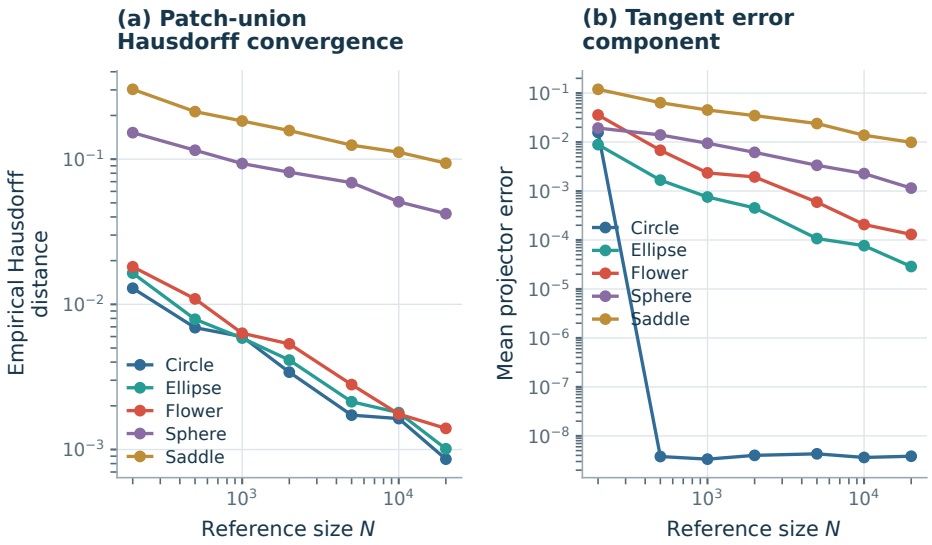

Figure 16: Patch-union Hausdorff convergence on five analytic manifolds. Both directions of the discretized Hausdorff distance are included.

F.4 Tangent estimation convergence

Local tangent-projector error decreases with reference size on both the circle and sphere. These synthetic bandwidth studies use a common Euclidean radius h: the PCA neighborhood of a_i contains fit references within distance h, with a nearest-neighbor replacement of size $\min\{N, \max(2m+3, 8)\}$ if that neighborhood is too sparse. The nominal patch radius remains $\rho = \lambda_\rho h$. Here h_0 is a prescribed baseline bandwidth, and h/h_0 is its dimensionless multiplier. This radius-based parameterization differs from the fixed-K construction in App. C.4, where the realized bandwidth is the anchor-specific h_i.

For the tangent-convergence test,

$$h_0 = \begin{cases} 2.2R(\log N/N)^{1/3}, & \text{circle}, \\ 2.5R(\log N/N)^{1/4}, & \text{sphere}. \end{cases} \tag{31}$$

For circle and sphere radii $0.5, 1, 2, 5$, we add isotropic noise $\sigma = 0.01R$, use $M = 192$, scan $N = 200$ to 10000, and evaluate $h/h_0 \in \{0.5, 0.75, 1, 1.5, 2\}$. Projector operator error, Frobenius error, local count, and eigengap are reported. At the selected scale, circle and sphere operator errors have slopes -0.432 ($R^2 = 0.997$) and -0.610 ($R^2 = 0.992$).

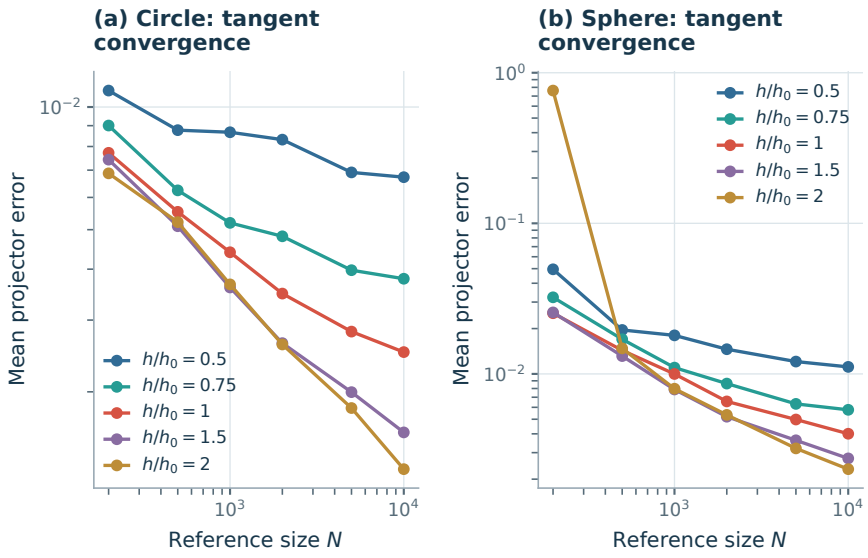

Figure 17: Local tangent-projector error versus reference size.

F.5 Anchor coverage and reference size

FPS improves worst-case coverage over random anchors and yields a better tradeoff between error and runtime. The anchor-coverage study isolates anchor selection by using analytic tangents on flower and sphere, $N = 5000$, $M = 16$ to 5000, $J = 4096$, and $\rho = 1.10$ times empirical coverage. FPS implements greedy k-center selection.

Larger reference banks give small relative errors for both drift levels in the separate reference-size study. It uses a unit circle with $N = 128$ to 10000, $J = 6000$, calibration size 4000, four anchor policies, and 300 bootstrap replicates. Queries are $z = (1 + \xi)(\cos\vartheta, \sin\vartheta)$ with independent $\vartheta \sim \mathrm{Unif}[0, 2\pi)$ and $\xi \sim \mathcal{N}(0, \sigma^2)$. We call $\sigma = 0.05$ the *low-drift generator* and $\sigma = 0.12$ the *high-drift generator*; their ideal NFD values are 0.0025 and 0.0144 to the reported precision. At $N = 5000$, raw relative errors are 1.8% for low drift and 2.6% for high drift. At $N = 10000$, they are 3.3% and 2.7%, respectively. Tail behavior is stable, and boundary hits are tracked.

F.6 Curvature stress with oracle-geometry selection

The curvature study covers plane, cylinder, spheres of radius $0.5, 1, 2, 5$, saddle, and flower. All queries use normal displacement 0.04 with analytic truth 0.0016, $N = 4000$, and held-out $J = 2000$. An oracle-geometry score (true Hausdorff error, true tangent error, and coverage penalty) selects among $h/h_0 = 0.5, 0.75, 1, 1.5$, $\rho/h = 0.1, 0.25, 0.5, 0.75$, and $M \in \{192, 384, 768\}$, with $M = 1536$ additionally evaluated for spheres. MP attains $\leq 35\%$ error in 75% of selected reconstructions, $\leq 60\%$ in all cases, and outperforms SPCA in 75%.

This study sets $h_0 = 2.2s(\log N/N)^{1/(m+2)}$, with $s = R$ for spheres and cylinders and $s = 1$ for plane, saddle, and flower; $m = 1$ for the flower and $m = 2$ for the surfaces. Thus h/h_0 rescales a geometry-specific bandwidth, while ρ/h independently controls the common patch extent.

F.7 Neighborhood, radius, anchor-count, and dimension ablations

An interior patch radius and sufficient anchors give the best coverage–error tradeoff in these ablations. The bandwidth, radius, and anchor-count tests use the unit circle with $N = 4000$, $h_0 = 2.2(\log N/N)^{1/3}$, and reference noise $\sigma = 0.006$. They scan $h/h_0 = 0.5, 0.75, 1, 1.5, 2$, $\rho/h = 0, 0.25, 0.5, 0.75, 1$, and $M = 16$ to 4000. A separate rank ablation uses a unit 2-sphere embedded in $D = 16$, $N = 3000$, $M = 128$, ambient noise 0.003, fixed $h = 2.5(\log N/N)^{1/4}$, and $m = 1, 2, 3, 4, 8$.

Both use Gaussian normal-offset queries with standard deviations 0.04 (low drift) and 0.10 (high drift), with $J = 3000$ per generator; these scales differ from the reference-size study above. For this ablation, the plotted error in Fig. 18 compares $\max\{\widehat{\mathrm{NFD}}_{J,\mathrm{rec}} - \bar{q}_{\mathrm{clean}}, 0\}$ with the low-drift bank's mean squared normal offset, where $\bar{q}_{\mathrm{clean}}$ is the mean squared reconstruction distance of clean reference-manifold queries. The ablations place the optimum near $\rho/h = 0.25$ and the stable bandwidth around $h/h_0 = 0.5$–1. Increasing M improves coverage until saturation, and eigengap-based selection recovers $m = 2$.

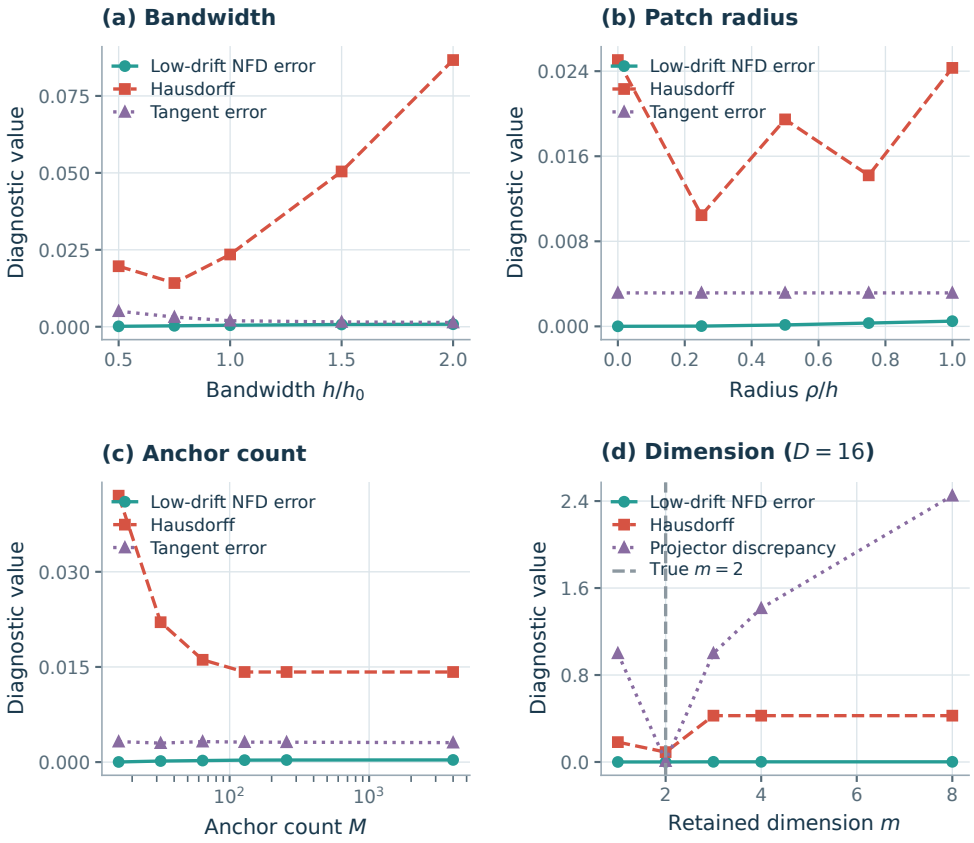

Figure 18: Joint h, ρ, M, and m ablations. The intended operating region is interior rather than a degenerate solution with zero or infinite radius.

F.8 Separation between manifold regions

Denser references and smaller neighborhoods reduce error and tangent mixing when manifold regions are close. We scan $g \in \{1, 0.5, 0.2, 0.1, 0.05\}$ on two curved branches and a three-petal flower, with reference-noise standard deviation $\sigma = 0.0025$. For branches $y = \pm(g/2 + 0.28x^2)$, g is their minimum Euclidean separation. For the flower, g is twice the minimum radial distance from the center, a controlled separation proxy rather than the minimum distance between all pairs of manifold points.

The baseline uses $(N, M, h) = (4096, 160, 0.24)$. The *recovery with denser references and smaller neighborhoods* uses $(N, M) = (12288, 240)$ and $h = \min\{0.24, \max(0.35g, 0.035)\}$, retaining $\rho/h = 0.5$. Thresholds are calibrated from pretrained references without query truth. The diagnostic scale comparison $h + 2\sigma < g$ motivates resolving the regions. In hard $g \leq 0.1$ cases, recovery lowers absolute error from 1.42×10^{-3} to 5.81×10^{-4} and tangent mixing from 0.062 to 0.014.

F.9 Unified estimator comparison and quality-gate validation

Bounded MP achieves zero structural false zeros and perfect case-level rank agreement with the oracle, with median error comparable to dense NN. Across five toy cases, median absolute errors are 4.48×10^{-4} (NN), 9.66×10^{-4} (SPCA), 1.59×10^{-3} (anchor-unbounded), and 5.14×10^{-4} (bounded MP). Structural false-zero rates, averaged equally over the five toy cases, are 0, 0.20, 0.40, and 0, respectively. These summaries are computed over the five controlled cases. NN's low toy error is expected because references are dense and queries lie near densely sampled reference manifolds; its finite-spacing term becomes apparent in on-manifold and lower-density studies.

Quality-gate intervention. On the toy validation grid, the combined gate reduced false-zero rate from 0.122 to 0.045, kept 0.996 coverage, and used fallback for 0.119 of patches. This is a separate gate-validation grid from the unified-estimator comparison above. The point-fallback rule and its pretrained-only diagnostics are defined in App. C.5 and illustrated in Fig. 8.

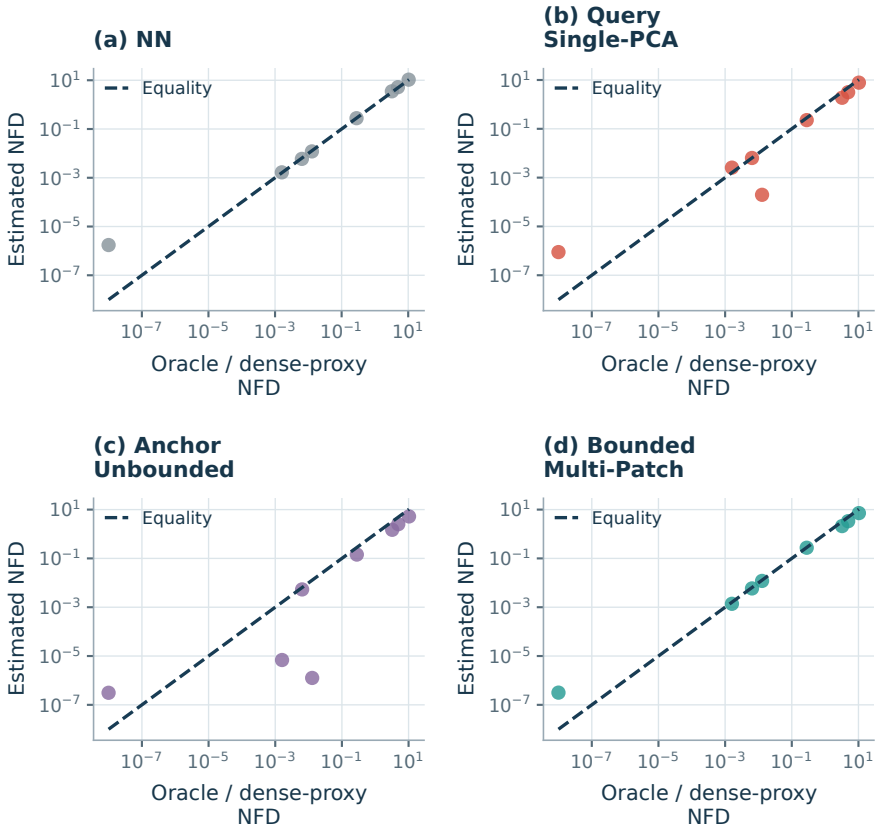

Figure 19: Oracle values or dense approximations versus four estimators. Bounded MP stays near equality without the SPCA and unbounded-patch false-zero outliers.

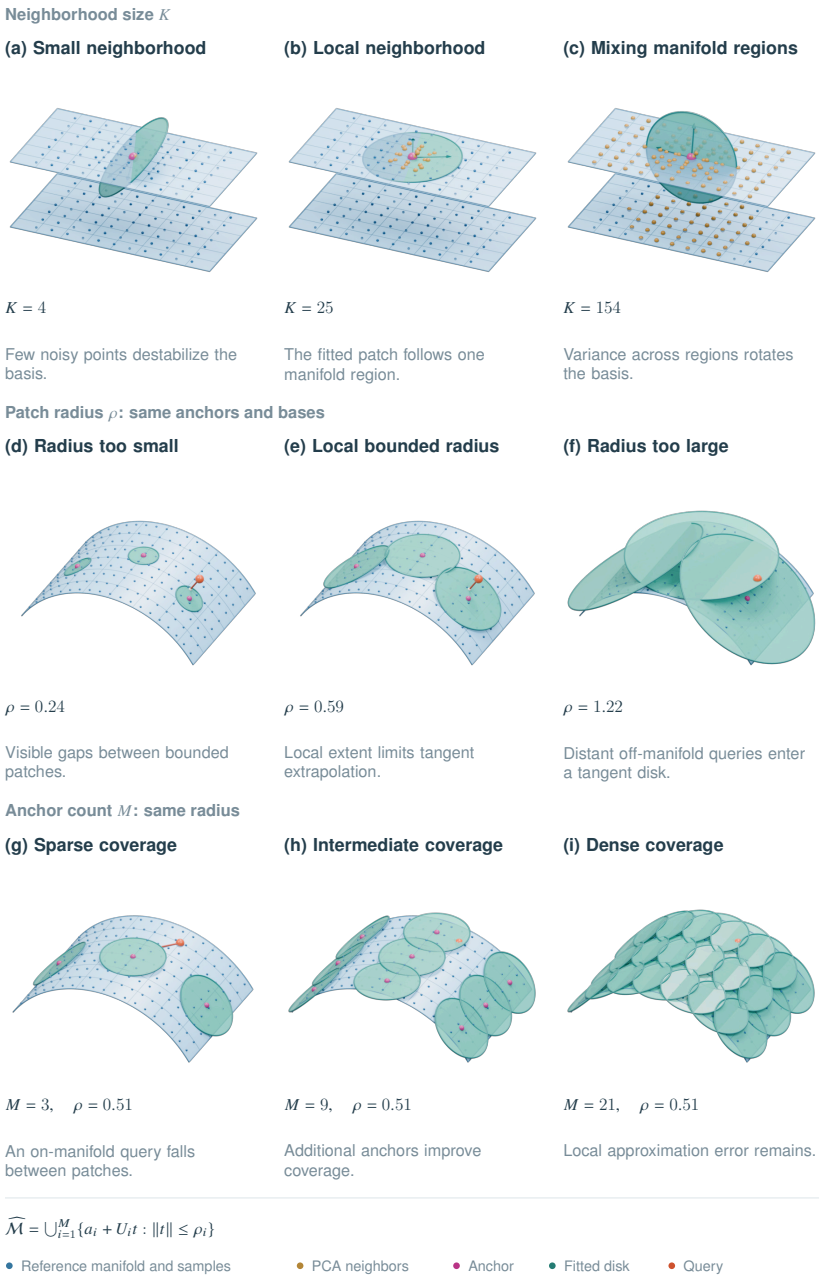

Figure 20: **Distinct roles of neighborhood size, patch radius, and anchor count.** Increasing K increases the induced local-PCA scale h_i, trading variance against curvature and mixing across manifold regions; ρ controls patch extent and tangent extrapolation; and M controls coverage at fixed radius. The radius-based quantitative ablations in Fig. 18 vary h directly, as defined in App. F.4.

G Detection resolution and pretrained-only selection

Reconstruction error limits quantitative drift resolution. We first measure this relationship using known geometry, then evaluate pretrained-only parameter selection and an observable resolution proxy. Oracle-geometry selection (App. F.6) serves as the benchmark.

G.1 Normal-drift resolution sweep

Resolution bound. The following bound motivates the drift-to-reconstruction ratio used in this experiment. Suppose $d_H(\mathcal{M}_c, \widehat{\mathcal{M}}_c) \le \epsilon_{\mathcal{M}} < \infty$ and $0 < \delta = \sqrt{\mathrm{NFD}(\theta \mid c)} < \infty$. Then

$$\frac{|B_{\mathcal{R}}(\theta \mid c)|}{\mathrm{NFD}(\theta \mid c)} \le 2\frac{\epsilon_{\mathcal{M}}}{\delta} + \left(\frac{\epsilon_{\mathcal{M}}}{\delta}\right)^2. \tag{32}$$

The Hausdorff bound and triangle inequality give $d_{\widehat{\mathcal{M}}_c}(z) \le d_{\mathcal{M}_c}(z) + \epsilon_{\mathcal{M}}$ and the reverse inequality with the sets exchanged. Thus $|e(z)| \le \epsilon_{\mathcal{M}}$. Apply Eq. 17 with $\Gamma = \epsilon_{\mathcal{M}}$ and divide by $\mathrm{NFD}(\theta \mid c) = \delta^2$. Small reconstruction error relative to δ is therefore sufficient for small relative bias. For a constant normal displacement, δ equals its magnitude.

Experimental results. Relative NFD error decreases as drift becomes larger than reconstruction error on every tested geometry. Spheres of radii 1, 2, and 5, a flower, and a saddle use drift magnitudes

$$\delta \in \{0.01, 0.02, 0.04, 0.08, 0.12, 0.20\}, \qquad \mathrm{NFD}_{\mathrm{true}} = \delta^2. \tag{33}$$

For each combination of geometry and reconstruction we compute the true Hausdorff error $\epsilon_{\mathcal{M}}$ from dense meshes and regress the raw relative NFD error against $\delta/\epsilon_{\mathcal{M}}$. Global Spearman is -0.834, the log–log slope is -0.952, and every within-case correlation is negative. Fifteen conditions lie above the fitted empirical operating point $c = 4.999015$. The fitted c is an empirical operating point for this grid; Eq. 32 supplies the sufficient bias bound.

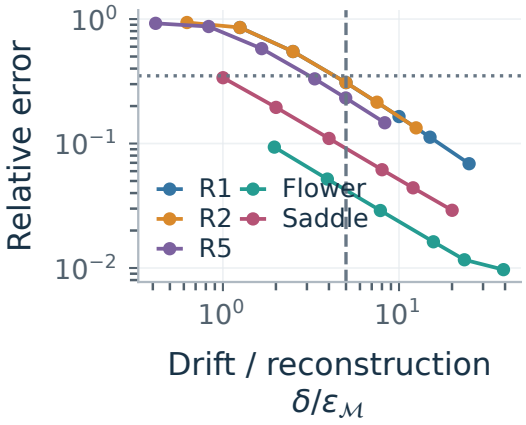

Figure 21: Relative NFD error against the drift-to-reconstruction ratio across the five geometries. Dotted lines mark 35% error and the empirical ratio 5.00.

G.2 Oracle-free hyperparameter selection

Pretrained-only selection achieves $\le 35\%$ NFD error in 62.5% of held-out cases (Fig. 22). Its score uses only fit and calibration samples, combining clean residual, coverage, tangent stability, eigengap, and fallback frequency.

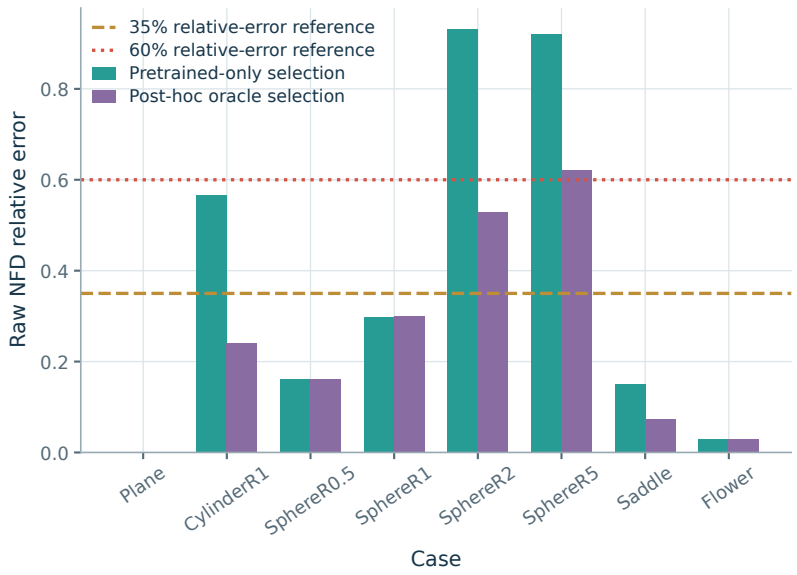

Figure 22: Held-out NFD accuracy after pretrained-only selection compared with oracle selection.

G.3 Practical resolution proxy

To estimate resolution from observable pretrained-sample diagnostics, we test

$$\widehat{\epsilon}^{\mathrm{prac}}_{\mathcal{M}} = \max\{E_{\mathrm{cal}}, E_{\mathrm{cov}}, E_{\tan}, E_{\mathrm{curv}}, E_{\mathrm{cross}}\}, \qquad \epsilon_{\mathrm{safe}} = s_{\mathrm{toy}}\widehat{\epsilon}^{\mathrm{prac}}_{\mathcal{M}}, \tag{34}$$

where E_{cal} is held-out clean RMS residual, E_{cov} the RMS patch-radius exceedance, $E_{\tan}$ two-scale projector disagreement times local scale, E_{curv} a second-order cross-scale proxy, and E_{cross} the 95th percentile disagreement between reconstructions from disjoint fit splits. A leave-one-manifold-out toy calibration gives $s_{\mathrm{toy}} = 1.721557$.

The practical proxy tracks true resolution (Spearman 0.972; safety coverage 92.2%). On this toy validation grid, it operates conservatively, achieving 100% precision and zero false-safe cases, with 29.7% recall.

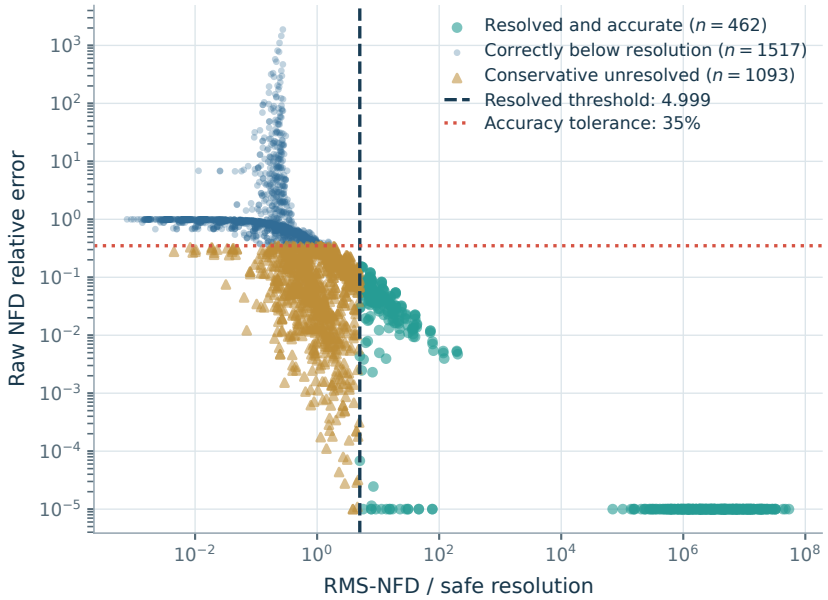

Figure 23: Practical resolution separates resolved and unresolved cases on the toy validation grid, with zero false-safe cases.

H Optimization experiments

The optimization studies connect geometry to post-training behavior: endpoint rankings test sensitivity to departure, binary OnManifold comparisons test retained severity information, and the gradient bridge tests the local quadratic prediction.

H.1 Toy preference optimization

MP preserves the oracle ordering of all six methods, with FlowDPO showing the largest departure (Tab. 7). A common toy generator and preference objective produce endpoints for Base, RFT, FlowDPO, FlowDPO+KL, FlowDPO+RFT, and ThermoDPO-weighted. Each method is evaluated with $J = 10{,}000$ using both an oracle based on analytic geometry and dense numerical evaluation and a separately fitted Multi-Patch reconstruction. The comparison uses one training seed.

Table 7: Toy optimization endpoints.

Method	Oracle NFD	Multi-Patch NFD
Base	0.008047	0.005816
RFT	0.006374	0.004475
FlowDPO	0.417	0.304
FlowDPO+KL	0.006993	0.005001
FlowDPO+RFT	0.008225	0.006169
ThermoDPO-weighted	0.006056	0.004445

The binary *OnManifold indicator* is $\mathbf{1}\{|x| \le 5,\ |y| \le 1,\ |z - f(x,y)| \le .15\}$ for the surface f below. Its threshold is on vertical residual, not Euclidean point-to-surface distance. Oracle and MP method ranks have Spearman 1.0. FlowDPO has 0% OnManifold endpoints, while ThermoDPO-weighted has 92.37%. This is a controlled instance of the qualitative mechanism in Han et al. (2026).

Fig. 24 contrasts preference concentration with geometric departure: RFT concentrates near the preferred bumps, whereas FlowDPO also moves above the reference surface

$$\mathcal{M} = \{(x, y, f(x, y)) : x \in [-5, 5],\ y \in [-1, 1]\},$$
$$f(x, y) = \left[1.2e^{-(x-3)^2} + 1.2e^{-(x+3)^2} - 0.9e^{-x^2}\right](1 - 0.15y^2).$$

The display uses 2,600 uniformly sampled stored (x, y, z) endpoints per method with a fixed plotting seed. All 10,000 contribute to Tab. 7. Colors encode Euclidean point-to-surface distance. Shared axes retain the full ranges, including the high-z FlowDPO tail.

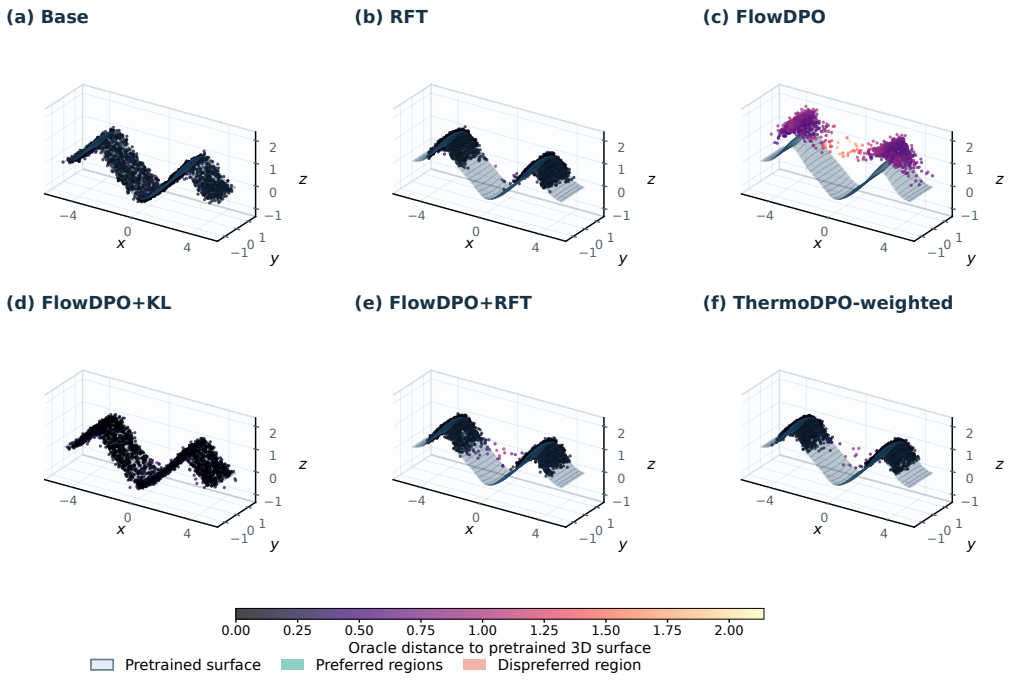

Figure 24: **Three-dimensional endpoint geometry after toy preference optimization.** Six methods share axes, viewing angle, and a common Oracle-distance color scale. The mesh shows the reference surface, with teal preferred x regions and a pale red dispreferred region. Points show the stored (x, y, z) endpoints, colored by three-dimensional point-to-surface distance. FlowDPO combines preferred-region concentration with substantial departure. The accompanying quantitative comparison uses mean *squared* distance and retains all endpoints.

Continuous NFD versus binary OnManifold. NFD retains severity information that the binary OnManifold label discards. Pooling all toy methods, MP NFD has Spearman 0.694 with $1 -$ OnManifold and area under the curve 0.955 for detecting thresholded off-manifold endpoints. Its Spearman correlation with continuous tube-threshold excess is 0.717.

Gradient bridge. Oracle NFD grows quadratically along the tested gradient direction, while MP eventually reaches its reconstruction floor. We construct a same-noise finite-difference direction from a FlowDPO loss gradient, project Base endpoints onto the oracle manifold, restrict to a 0.01 interior margin (9,642 of 10,000 endpoints), and calibrate a base step $\alpha_0 = 0.0110569$ to endpoint RMS 0.06. Across shrinking fractions of α_0, oracle NFD has local log–log slope 1.995. At the smallest step, NFD $/\alpha^2$ differs from the directly computed normal endpoint energy by 0.081%. MP follows the curve until its 2.83×10^{-6} reconstruction floor (Fig. 25).

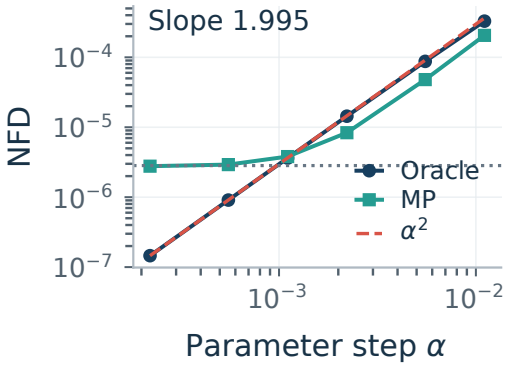

Figure 25: Oracle NFD scales quadratically along the tested gradient direction, while MP reaches its reconstruction floor.

I Real T2I protocol and complete quantitative results

FlowDPO has the largest departure across encoders and reference refits. This appendix gives the sampling protocol and complete effect estimates, followed by estimator costs, metric comparisons, resolution diagnostics, and external evaluations.

I.1 Checkpoints and sampling

All checkpoints share the SD3.5-M base and are evaluated as supplied, without weight modification. Sampling uses the Flow-Factory DPM2 sampler, 40 steps, guidance 4.5, maximum sequence length 128, and 512×512 output. Each method uses the same prompt and random seed. Checkpoint identifiers are Base (pretrained), RFT step 1010, FlowDPO step 620, FlowDPO+KL step 490, FlowDPO+RFT step 340, and ThermoDPO-weighted step 890.

The primary and confirmatory sets contain 24 and 12 disjoint prompts, respectively. Per prompt, 64 Base features fit geometry, 16 calibrate it, and 16 supply an independent cross-fit check. Candidate grids are $m \in \{2, 4, 8\}$, $M \in \{16, 32, 48\}$, $K \in \{8, 16, 24\}$, and $\lambda_\rho \in \{0.25, 0.5, 0.75, 1\}$. Projectors, prompt lists, candidate scores, and thresholds are frozen without reading fine-tuned NFD or reward outputs.

I.2 All primary effects

FlowDPO has greater departure than each mitigation variant, with all three paired intervals excluding zero. Tab. 1 reports the primary raw NFD and paired effects for $J = 96$ outputs per method. Corresponding RMS-NFD values are 0.166, 0.177, 0.324, 0.252, 0.225, and 0.184 for Base, RFT, FlowDPO, FlowDPO+KL, FlowDPO+RFT, and ThermoDPO-weighted. All intervals use the $B = 10{,}000$ hierarchical bootstrap over prompts and paired seeds.

FlowDPO-minus-mitigation mean differences and 95% intervals are 0.041 $[0.014, 0.070]$ for KL, 0.054 $[0.030, 0.083]$ for RFT, and 0.071 $[0.045, 0.100]$ for ThermoDPO-weighted. Across four encoders, median pairwise rank Spearman is 0.800, FlowDPO is rank one under every encoder, and every mitigation lies below FlowDPO.

I.3 Reference-refit uncertainty

FlowDPO remains the highest-scoring method in all 200 reference refits. Unlike the preceding bootstrap intervals, these refits isolate reference-sampling uncertainty. Each prompt's 96 Base

references are randomly repartitioned with fixed seeds into 64 fit, 16 calibration, and 16 cross-fit references. FPS anchors and PCA bases are refitted, while prompt-specific hyperparameters and target images remain fixed.

Table 8: Variation across 200 complete reference refits. Intervals are empirical 2.5–97.5% refit quantiles, not bootstrap intervals over prompts and paired seeds.

Method	Primary	Refit mean	SD	CV	Refit interval
Base	0.02772	0.02767	0.00086	0.031	[0.02612, 0.02921]
RFT	0.03138	0.03188	0.00117	0.037	[0.02967, 0.03408]
FlowDPO	0.10492	0.10368	0.00450	0.043	[0.09578, 0.11241]
FlowDPO+KL	0.06374	0.06313	0.00221	0.035	[0.05878, 0.06725]
FlowDPO+RFT	0.05051	0.04851	0.00177	0.036	[0.04500, 0.05165]
ThermoDPO-weighted	0.03374	0.03271	0.00129	0.040	[0.03048, 0.03530]

Its paired contrast against Base has mean 0.076, SD 0.004509, empirical interval $[0.068, 0.084]$, and is positive in every refit. Rank Spearman against the primary method ordering has median 1.000 and 2.5–97.5% range $[0.943, 1.000]$. These refits support stability at the evaluated reference-bank size.

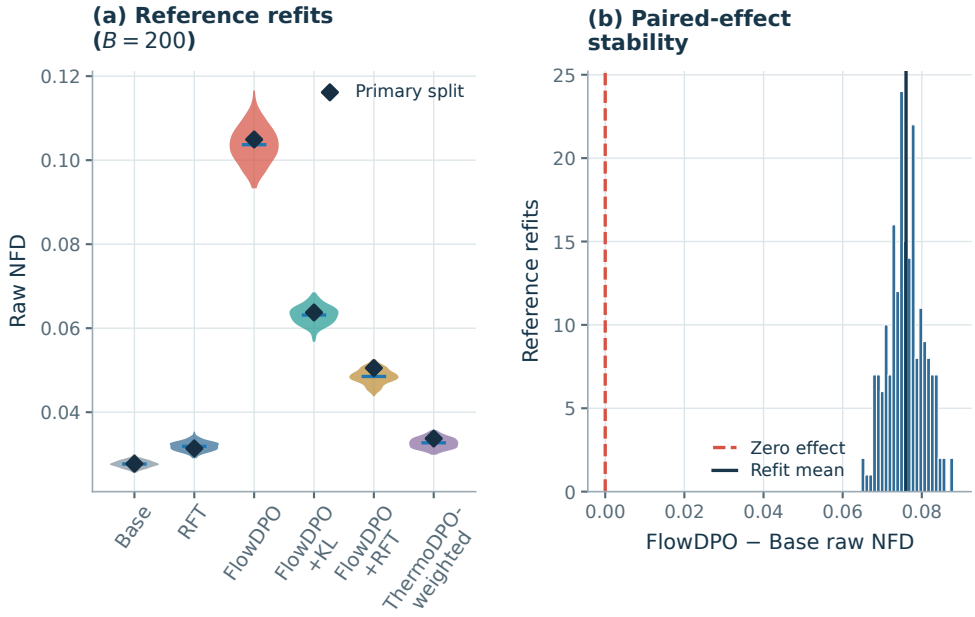

Figure 26: Reference-refit uncertainty is small relative to the FlowDPO effect. Horizontal intervals vary only the pretrained split and fitted geometry.

I.4 Estimator ablation and timing

NN, SPCA, and MP agree on the positive FlowDPO-minus-Base effect and all three mitigation comparisons. They use identical frozen features and paired samples. SPCA selects $K = 16, m = 8$ using calibration RMS plus 0.25 times tangent instability, from pretrained calibration data only. The estimator-ablation analysis reports FlowDPO-minus-Base means (95% CIs) of NN 0.110 $[0.068, 0.160]$ and SPCA 0.067 $[0.044, 0.092]$. For MP, we use the primary contrast and interval in Tab. 1 throughout. The NN and SPCA intervals remain those of the estimator-ablation analysis. Original-to-confirmatory rank Spearman is 0.829, 0.943, and 0.829.

The reported "one evaluation" time refers to scoring already-extracted features, not image generation or encoder inference. MP takes 8.05 ms online for all 864 queries and 52.05 ms including a cold patch build.

Table 9: Measured estimator efficiency on one RTX 4090; 864 queries, median of 20 timed repeats after five warmups.

Estimator	Build (s)	Online 864 (s)	ms/query	Peak GPU MB
NN	0.00426	0.00500	0.00579	9.23
Query SPCA	0.00121	0.01925	0.02228	9.35
Bounded MP	0.04400	0.00805	0.00931	9.83

I.5 Standard-metric comparison

Standard metrics also detect the large FlowDPO shift, despite measuring different properties. NFD uses the frozen SigLIP2-d16 representation and prompt-conditioned reconstructions described in App. I.1. NN squared distance, RBF-MMD, sliced W_2^2, symmetric Gaussian KL, and precision, recall, density, and coverage use the same feature space, with 384 Base cross-reference features and 96 evaluation images per method. FID and KID use Inception features. Tab. 10 reports the unnormalized values.

Table 10: Standard-metric controls on the six SD3.5-M methods.

Method	NFD	FID	MMD^2	SW_2^2	Prec.	Recall	Density	Coverage
Base	0.028	115.78	-0.00435	0.00046	0.979	1.000	1.033	0.742
RFT	0.031	129.32	0.00544	0.00210	0.906	0.997	0.760	0.492
FlowDPO	0.105	186.53	0.047	0.00844	0.396	0.857	0.198	0.143
FlowDPO+KL	0.064	148.94	0.021	0.00431	0.615	0.992	0.350	0.266
FlowDPO+RFT	0.051	153.48	0.00945	0.00311	0.760	0.990	0.496	0.352
ThermoDPO-weighted	0.034	128.38	0.00649	0.00240	0.917	0.997	0.835	0.565

Spearman correlations with NFD are 0.886 for FID and KID, 1.000 for MMD^2, sliced W_2^2, symmetric Gaussian KL, and mean NN squared distance, 0.943 for $1-$precision and $1-$coverage, and 0.928 for $1-$recall. These metrics share sensitivity to the large FlowDPO change. The analytic map in App. E establishes non-equivalence; see App. I.6 for sample-level discordance analyses.

I.6 Sample-level comparison with NN

Across 864 images, NFD and NN squared distance have Spearman 0.923 but disagree on 12.31% of method orderings matched by prompt and seed. These sample-level differences accompany the analytic non-equivalence established in App. E.

I.7 Absolute-resolution diagnostics

Prompt-conditioned banks reduce the conservative resolution floor by 70.9% relative to a global bank, with geometry plateauing around $N = 128$. Under the conservative absolute-resolution criterion, we report real-model comparisons relative to the fitted reconstruction.

I.8 Independent evaluators

Larger Base-relative NFD covaries with lower scores from all five preference and aesthetic evaluators, while EasyOCR shows a different pattern (Tab. 11). The six evaluators are PickScore v1, LAION Aesthetic, MPS Overall, HPS v2.1, ImageReward v1.0, and EasyOCR. We use PickScore, MPS, HPS, and ImageReward as their respective short names elsewhere; the model versions are unchanged.

The complete study contains 36 prompts, six methods, and four paired seeds: 864 images, 216 prompt–method cells, and 2,592 estimator scores. For each evaluator, we center method quality and NFD against Base within prompt, then correlate Δ NFD with negative Δquality. Prompt-level bootstrap keeps all method cells in a prompt together.

The first five are separately trained preference and aesthetic channels (Kirstain et al., 2023; Wu et al., 2023; Xu et al., 2023; Zhang et al., 2024). EasyOCR measures the text-rendering task optimized by several checkpoints. A separate blinded study below evaluates visible task correctness, overall quality, and artifact severity without exposing method or NFD labels.

Table 11: External validation for bounded MP across 216 prompt–method cells.

Evaluator	Spearman	95% CI low	95% CI high
PickScore v1	0.448	0.231	0.638
LAION Aesthetic	0.426	0.256	0.584
MPS Overall	0.388	0.166	0.599
HPS v2.1	0.312	0.105	0.501
ImageReward v1.0	0.195	0.014	0.381
EasyOCR	-0.144	-0.334	0.056

I.9 Blinded human evaluation

Lower-NFD images receive 84.4% of overall-quality judgments on the selected pairs, counting ties as half a win (Tab. 12). The protocol records task correctness and overall visual quality, 0–4 artifact severity for each image, and 1–5 confidence; it specifies tie coding, incomplete-assignment exclusion, and pair-cluster bootstrap. The 120 pairs comprise 40 FlowDPO–Base, 40 FlowDPO–FlowDPO+KL, and 40 FlowDPO–ThermoDPO-weighted contrasts. Left and right placement is blinded and nearly balanced (FlowDPO left in 58 pairs and right in 62). The interface explicitly separates task correctness, including text rendering from naturalness and artifacts, records affirmative consent, permits withdrawal, and stores only anonymous rater codes—no names, emails, or IP addresses.

Thirteen raters completed 15 assignments each, providing 195 judgments across all 120 pairs: 45 pairs were rated once and 75 twice. Two incomplete assignments were excluded. We report 10,000-draw bootstrap intervals clustered by pair and by rater, with participants as the independent units in the latter.

Table 12: Blinded human evaluation. Win-equivalent rates encode ties as 0.5. The first two rows favor FlowDPO; the last two favor the lower-NFD side.

Endpoint	Rate	Pair-cluster 95% CI	Rater-cluster 95% CI
FlowDPO task correctness	0.182	$[0.126, 0.245]$	$[0.131, 0.233]$
FlowDPO overall quality	0.059	$[0.030, 0.093]$	$[0.036, 0.082]$
Lower-NFD task correctness	0.721	$[0.646, 0.792]$	$[0.667, 0.785]$
Lower-NFD overall quality	0.844	$[0.781, 0.902]$	$[0.785, 0.900]$

Among the 75 twice-rated pairs, agreement is 0.787 for task correctness and 0.893 for quality (Cohen $\kappa = 0.627$ and 0.807). Mean confidence is 4.61 out of 5. Pair-mean left-minus-right NFD and artifact-severity differences have Spearman $\rho = 0.778$ (10,000-permutation $p < 10^{-4}$). Overall left-choice rates are 0.500 for task and 0.505 for quality, arguing against a material position bias.

J Independent FLUX.2 distillation validation

J.1 Purpose and protocol

The distilled FLUX.2 model has a positive Base-relative NFD effect, extending the comparison beyond SD3.5-M OCR preference optimization. We compare FLUX.2 [klein] Base 4B with the official step- and guidance-distilled 4B checkpoint (Black Forest Labs, 2026). This comparison tests NFD across a second model family and a different post-training operation.

For every prompt the projection and the choices of m, M, K, and ρ are frozen before evaluation features are read. The same prompt and seed are used for the Base and distilled evaluation pair. Across the 128 pairs, distilled-minus-Base NFD is 0.002549 with 95% interval $[0.000647, 0.004902]$ and bootstrap direction probability 0.997. The sample-level positive fraction is 0.797, and 14 of 16 prompt means are positive. Base has raw NFD 0.004153 $[0.002834, 0.006051]$; distilled has 0.006702 $[0.004933, 0.008834]$.

Table 13: FLUX.2 evaluation protocol. Selection sees only Base fit, calibration, and cross-fit features.

Component	Configuration
Prompts	16 diverse non-OCR English prompts
Base banks / prompt	64 fit, 16 calibration, 16 cross
Evaluation / prompt	8 Base and 8 distilled, paired by seed
Total images	1,792 at 512×512
Generation	Base: 50 steps, CFG 4; distilled: 4 steps, guidance 1
Representation	frozen SigLIP2; Base-fit-only global PCA, $d = 16$
Geometry	prompt-specific bounded MP; pretrained-only selection
Statistics	shared prompt–seed bootstrap, $B = 10{,}000$
Implementation	Diffusers v0.37.1

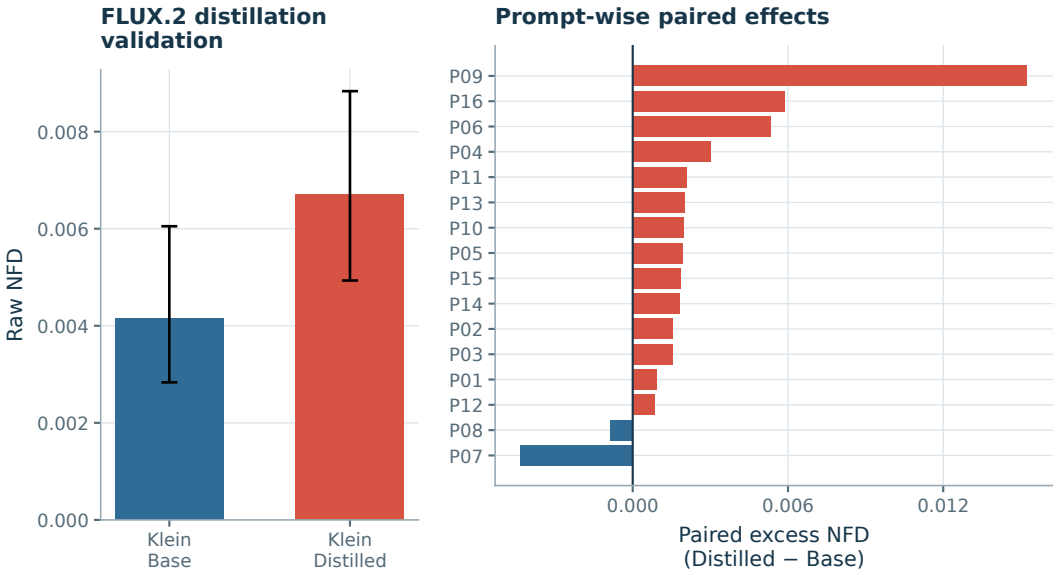

Figure 27: Independent FLUX.2 result. Raw reconstruction NFD values (left) and prompt-wise paired distilled-minus-Base effects (right).

J.2 Resolution and standard-metric controls

We interpret the paired FLUX.2 effect relative to the frozen reconstruction. Median held-out Base calibration RMS is 0.060 and median cross-fit disagreement 0.027. Median ratios of RMS-NFD to safe resolution are 0.202 (Base) and 0.248 (distilled). Mean boundary-hit rates are 0.086 and 0.180.

Table 14: Pooled FLUX.2 standard metrics. Negative unbiased KID and MMD estimates are retained rather than clipped.

Model	NFD	FID	KID	MMD^2	SW_2^2	Prec.	Recall	Coverage
Base	0.004153	0.00222	-0.001692	-0.003181	0.000193	0.898	0.975	0.463
Distilled	0.006702	0.014	-0.000933	-0.001781	0.000815	0.812	0.649	0.179

The positive paired effect extends the operational comparison to a second architecture and a distillation operation.

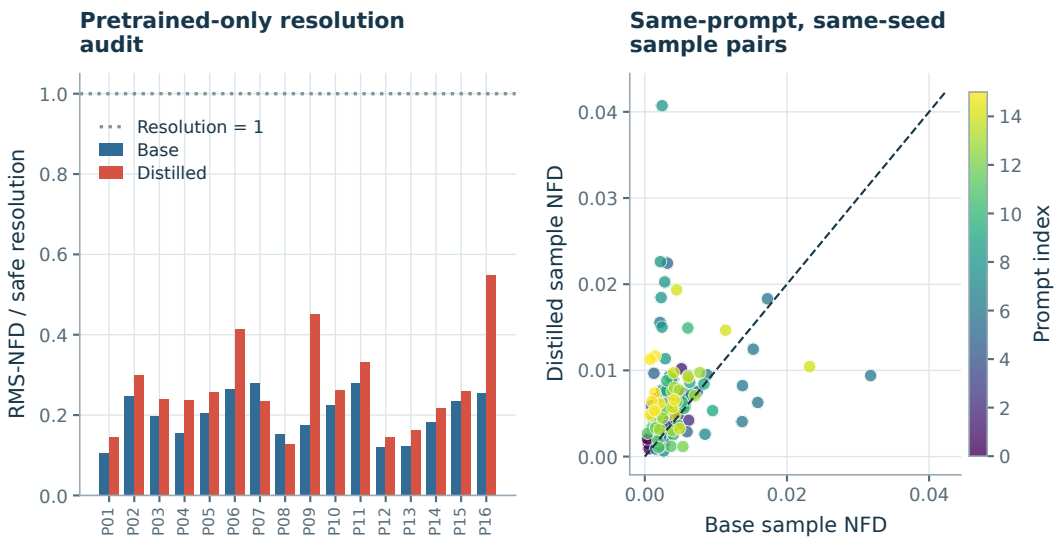

Figure 28: Pretrained-only resolution diagnostics (left) and same-seed sample pairs (right). The dashed scatter diagonal marks equal Base and distilled scores.

K Empirical audit of local reference geometry

We examine local tangent structure, neighborhood-mass scaling, and held-out reconstruction in pretrained representations (Fig. 29).

K.1 Protocol and nulls

The audit reuses Base samples from 24 prompts and two frozen representations, SigLIP2-d16 and DINOv2-d16. Each prompt provides 128 fit, 32 held-out, and 32 test features. Neighborhood sizes are $K = 16, 24, 32, 48, 64$, with $K = 32$ primary. We use 5,000 prompt-block bootstrap replicates. Covariance-matched Gaussian and coordinate-shuffled samples provide null comparisons. Bounded MP is compared with global PCA and NN on held-out points.

We evaluate six diagnostics: intrinsic dimension, spectral decay, normal residual scaling, tangent smoothness, neighborhood-mass scaling, and reconstruction with null specificity.

K.2 Local structure and reconstruction

Local tangents are smoother than far tangents: the mean local-minus-far Frobenius distance is -0.335 with interval $[-0.365, -0.302]$. Median neighborhood-mass dimension is 5.830, with valid fraction above 0.96. MP improves held-out error over global PCA for every prompt (mean difference -0.003157), and DINOv2 reproduces the broad pattern.

These local results coexist with similar dimension and spectral decay under the Gaussian null and a larger relative MP advantage under coordinate shuffling. The residual-scaling exponent is 0.698 $[0.668, 0.804]$, rather than the quadratic exponent of 2. The audit supports local reconstruction structure, without establishing a thin-manifold model.

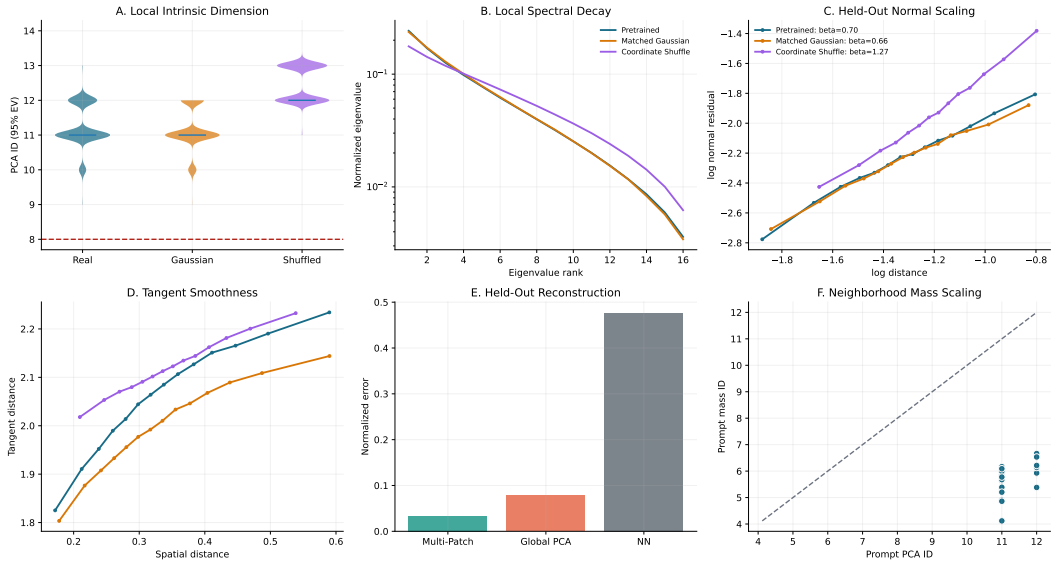

Figure 29: Local-reference-geometry audit: dimension, spectral decay, normal residual scaling, tangent smoothness, reconstruction, and neighborhood-mass scaling, with Gaussian and shuffled null comparisons.

L VISUAL INTERPRETATION

The galleries compare within-prompt NFD variation, reward-matched outputs, and nearest Base references. Two representative prompts illustrate each view; an all-prompt overview and balance audits document the six-reward matching.

Compact labels use OCR for EasyOCR, Pick or PS for PickScore v1, HPS for HPS v2.1, IR for ImageReward v1.0, AES for LAION Aesthetic, and MPS for MPS Overall (App. I.8).

L.1 SELECTION PROTOCOL

The visual bank contains 256 FlowDPO samples for each of ten prompts (2,560 images). Generation, features, patches, quantiles, OCR bins, reward calipers, and tie-breaking are fixed before inspection, with no manual sample replacement. Four views are retained:

1. **NFD quantiles:** six images at each of Q5, Q25, Q50, Q75, and Q95 per prompt (300 images).
2. **Matched task:** three images at five NFD levels within a fixed OCR bin (150 images).
3. **Nearest references:** two queries per prompt with five nearest Base neighbors each (100 comparisons).
4. **Six-reward matching:** twelve NFD-ordered images per prompt, jointly caliper-matched on EasyOCR, PickScore, HPS, ImageReward, LAION Aesthetic, and MPS.

Across prompts the median within-prompt Spearman between RMS-NFD and negative PickScore is 0.319. In matched OCR bins it is 0.282. The selection audits below retain both the matched contrasts and the wider association structure.

Compact main-text illustration. Fig. 6 illustrates two prompts selected after reviewing the completed gallery; within-prompt image selection and ordering follow the fixed protocol above. FlowDPO columns use positions 1, 6, and 12 from each preselected twelve-image six-reward sequence, ordered by squared MP distance. Base is the rank-one fit reference retrieved for Q5; no Base NFD is assigned. Seeds differ and pixels are unmodified. Reward calipers apply to complete low- and high-NFD groups, not the displayed subset.



Figure 30: Expanded NFD-quantile gallery for prompt ID 05. Each row is a fixed quantile and contains six distinct images; all English labels and selections are automatic.



Figure 31: Queries and nearest Base references for the same prompt ID 05. Each query is accompanied by five automatically retrieved references in the NFD representation.



Figure 32: Expanded NFD-quantile gallery for prompt ID 10. Each row is a fixed quantile and contains six distinct images; all English labels and selections are automatic.

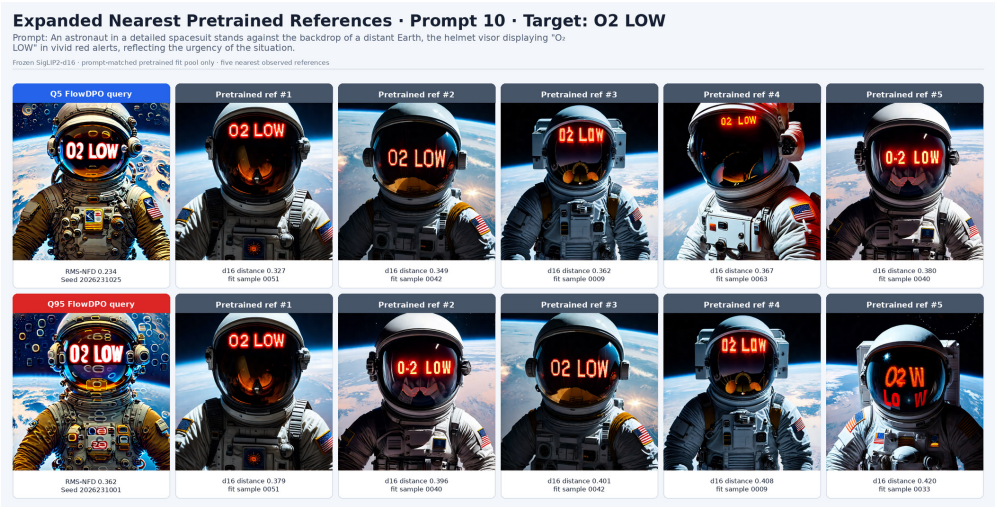

Figure 33: Queries and nearest Base references for the same prompt ID 10. Each query is accompanied by five automatically retrieved references in the NFD representation.

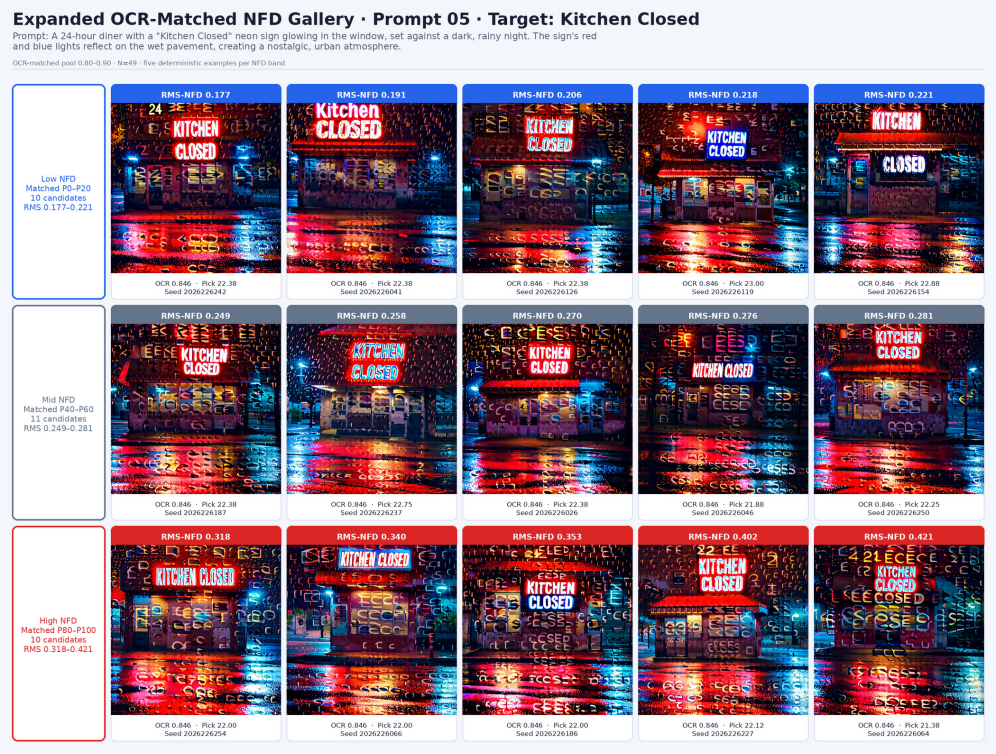

Figure 34: OCR-caliper-matched NFD gallery for prompt ID 05. Images occupy the same predefined OCR bin and span five NFD levels.

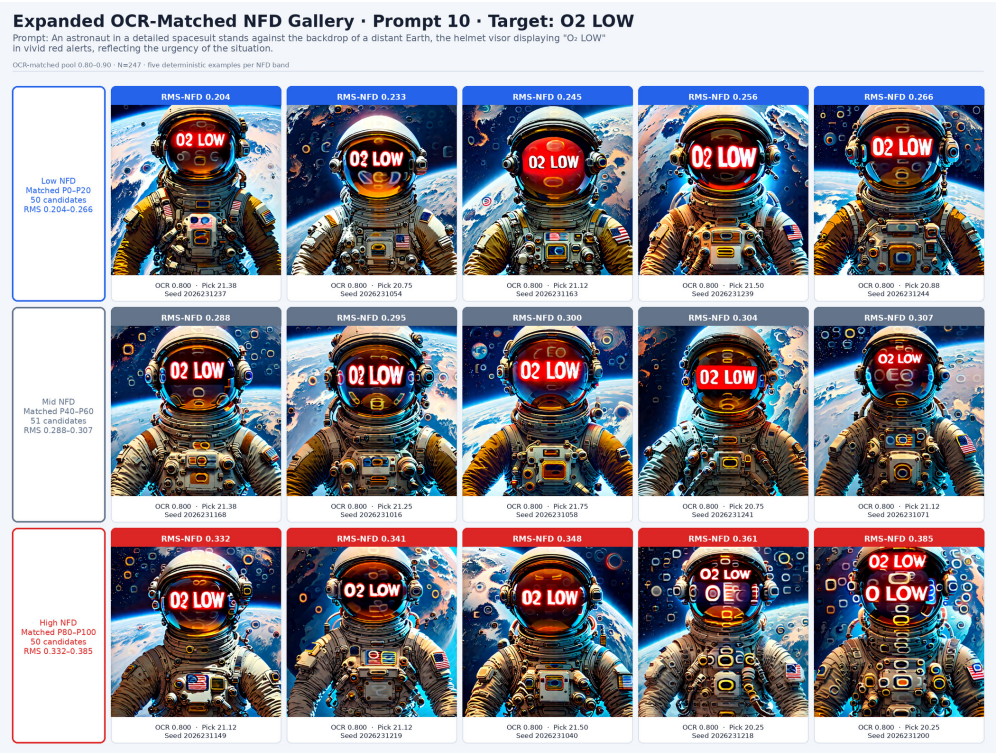

Figure 35: OCR-caliper-matched NFD gallery for prompt ID 10. Selection uses the same fixed OCR bins and NFD quantiles across prompts.



Figure 36: All-prompt six-reward caliper match. Within each prompt, images are ordered by NFD. Low- and high-NFD groups differ by at most 0.315 robust units on any measured reward while retaining at least 82.6% of the full Q05–Q95 NFD span.

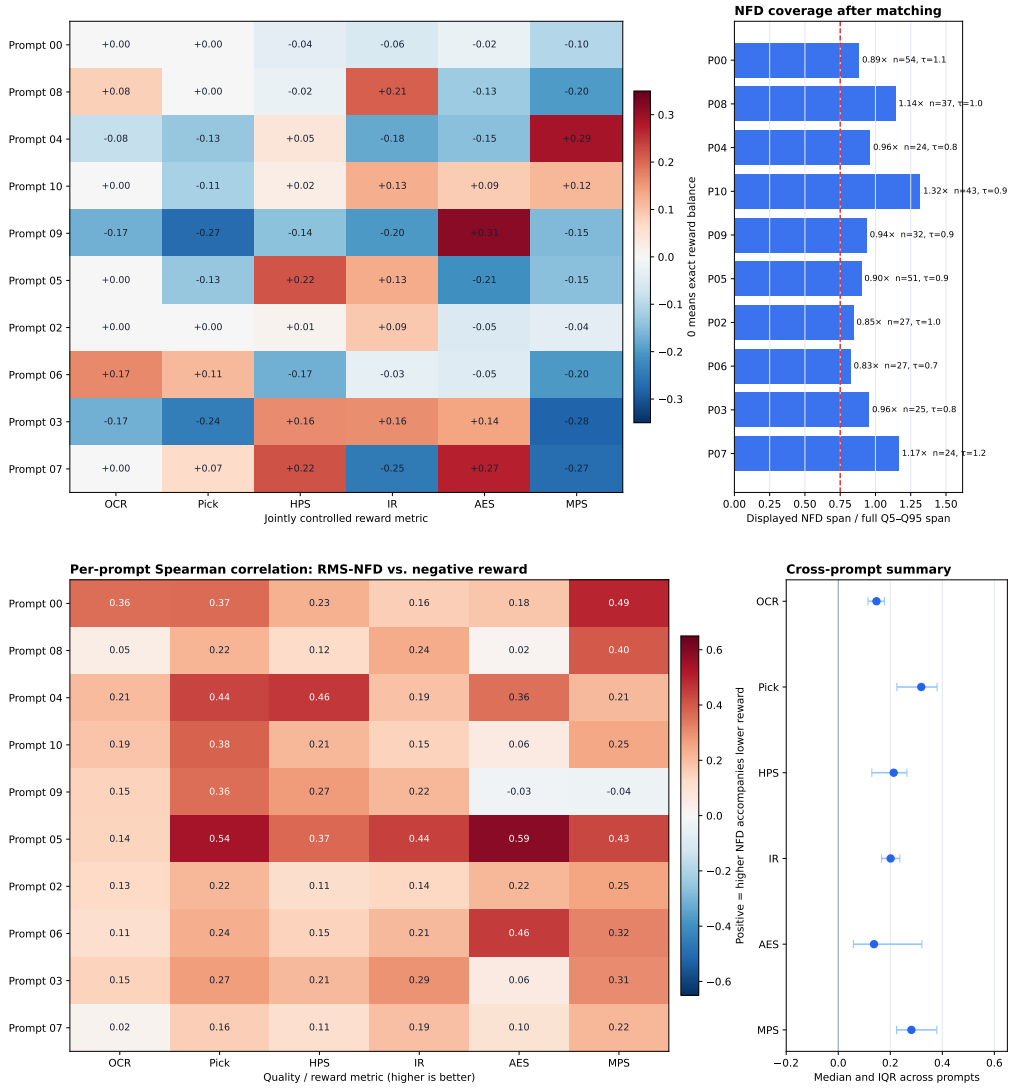

Figure 37: Selection audits for the six-reward gallery: reward balance (top) and the wider sample-level association between rewards and NFD (bottom). Calipers bound the selected groups' measured reward differences.

M REPORTING AND INTERPRETATION

For reproducible NFD evaluation, report the frozen representation, prompt-conditioned reference construction, pretrained-only parameter selection, raw NFD for Base and each checkpoint, and paired Base-relative effects with confidence intervals. Reconstruction diagnostics and reference-refit variation describe uncertainty in the fitted geometry; prompt and seed resampling describes evaluation uncertainty (Apps. C and I.3).

M.1 INTERPRETATION

Real-model NFD measures geometric departure relative to a frozen pretrained reconstruction in a specified representation. Apps. G and K characterize reconstruction resolution and local geometry.

N REPRODUCIBILITY DETAILS

N.1 REPLICATION MATERIALS

The supplementary materials include estimator code, generation and evaluation procedures, experiment configurations, numerical results, and plotting scripts. Prompt-wise galleries accompany the visual analyses. Protocols and statistical details are given with each experiment in Apps. D–L.

N.2 SOFTWARE AND HARDWARE

Error-decomposition tests use CPU float64 with Python, NumPy, and SciPy. Earlier geometry tests use Python 3.11.9 and PyTorch 2.7.1+cu126 in float64 on RTX 4090 GPUs, with NumPy, SciPy, and scikit-learn. Both generation studies use eight RTX 4090 GPUs: SD3.5-M uses the supplied implementation and LoRAs, and FLUX.2 uses official checkpoints with Diffusers v0.37.1. Encoders and evaluators use published weights. Estimator timings use one RTX 4090, synchronization, five warmups, and 20 repeats (App. I.4).

N.3 STATISTICS AND LEAKAGE CONTROL

Human-study sampling, exclusions, pair coverage, and clustered intervals are reported in App. I.9.

Table 1 reports the primary values and paired contrasts. Real-model intervals use $10{,}000$ hierarchical bootstrap draws over prompts and paired seeds. Reference-refit intervals vary only the pretrained split and fitted geometry, with target images fixed (App. I.3). The manifold audit uses $5{,}000$ prompt-block draws, and human judgments use both pair and rater clustering (Apps. K and I.9).

Fit, calibration, cross-fit, and evaluation images are disjoint. Geometry selection uses only pretrained data before post-trained outputs or held-out rewards are inspected. Confirmatory and FLUX.2 prompts are fixed before generation. Visual selection follows the rules in App. L without manual image replacement; the two main-text prompts are chosen after gallery inspection.