Local Manifold Identification with Latent Linear Models and OT Flows

Sherman Khoo[*]
sherman.khoo@bristol.ac.uk
University of Bristol

Song Liu[*]
song.liu@bristol.ac.uk
University of Bristol

Mark Beaumont
M.Beaumont@bristol.ac.uk
University of Bristol

Abstract

The manifold hypothesis states that high-dimensional data often concentrate near low-dimensional structures. Identifying these local manifold structures allows us to separate signals from ambient noise. In this paper, we study the problem of identifying local manifold structure from data. Given a query point y_0 from a dataset, our goal is to recover the geometry of the data manifold in a neighborhood of y_0 using a *pretrained* optimal transport flow from the reference to the data distribution. First, we prove that the Brenier optimal transport map preserves manifold structure: the preimage of an m-dimensional data manifold is itself an m-dimensional manifold in the reference space. Second, motivated by this result, we propose a latent variable model that maps a linear model through the transport flow. We prove that the linear approximation error is significantly reduced by the optimal transport map, leading to a tight fit of the non-linear data manifold. Third, noting the intractability of the resulting likelihood, we deploy denoising Fisher score estimation — a recent development from simulation-based inference that learns the Fisher score over parameter–observation pairs — to perform likelihood-based inference effectively. Experiments on both synthetic and real-world datasets demonstrate the effectiveness of the proposed method.

1 Introduction

The manifold hypothesis—that high-dimensional data such as images inherently lie on a low-dimensional manifold—is the foundation of many machine learning algorithms [11, 34, 26]. Recovering this geometric structure allows us to extract signals from the data distribution while removing noise, thereby enabling downstream applications. While directly estimating the global manifold is difficult due to the curse of dimensionality [29, 31], accurately identifying its *local structure*, i.e., the tangent space at a query point, remains highly tractable and is sufficient for many applications. The challenge is to estimate this local structure accurately and efficiently [2, 19].

The study of local manifold recovery has a long history, with the classic tool being local PCA [18]. This method collects the k nearest neighbors of a query point y_0, centers them, and takes the top m eigenvectors of the resulting covariance matrix as an estimate of the tangent space of an m-dimensional data manifold. While local PCA is simple, fast, and comes with finite-sample tangent-space recovery guarantees [38, 44, 2], it is also limited by a fundamental bias-variance trade-off. A linear approximation is only accurate when the neighborhood radius is small relative to the manifold's curvature [44]; however, a small radius captures few samples, resulting in a noisy

[*]equal contribution

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

covariance estimate, especially in high dimensions. Consequently, practitioners are forced to trade off linearization bias against estimator variance.

Recent advances in generative modeling shed new light on this problem. Diffusion models [16, 40] and flow matching [24, 43] learn smooth, invertible maps from a simple isotropic reference distribution to the data distribution. Crucially, a growing body of work now shows that these maps implicitly *encode the manifold geometry of the data*: as the flow compresses the reference distribution onto a thin neighborhood of the data manifold $\mathcal{M}$, its Jacobian contracts along the normal directions of $\mathcal{M}$ much more strongly than its tangent directions [41, 26, 12]. These discoveries naturally raise the question: can we recover the local manifold structure from a *pretrained generative flow*? Among the broader class of generative models, we focus on Brenier Optimal Transport (OT) flow and its practical approximation, OT-Conditional Flow Matching (OT-CFM).

In this work, we demonstrate that the *reference space* of OT flow contains a latent preimage that perfectly mirrors the geometry of the data manifold $\mathcal{M}$. By fitting a local linear model on this latent structure and pushing it forward through the pretrained OT map, we generate a curved, high-fidelity local approximation of $\mathcal{M}$ in the data space. Crucially, since the OT flow strongly contracts along normal directions [12], this pushforward actively suppresses linearization errors in the reference space. As a result, our model tightly tracks $\mathcal{M}$ across extended neighborhoods where standard local PCA breaks down. To accurately model real-world observations, we further incorporate structured ambient noise orthogonal to the local tangent space, capturing the measurement errors that push data points off the underlying signal manifold.

While pushing this model through a non-linear OT map yields a highly accurate geometric approximation, it also makes evaluating the exact likelihood intractable. However, the model remains *trivial to simulate from*: given the parameters, we sample latent variables and push them through the flow to generate observations. This easy-simulation, hard-likelihood asymmetry is the hallmark of simulation-based inference (SBI) [9]. We adopt denoising score matching [45], which has been applied to estimate the gradient of the log-likelihood function (Fisher score) from joint samples in SBI [37, 21]. Leveraging the estimated score, likelihood-based inference reduces to highly efficient gradient ascent, bypassing the need for exact likelihood evaluation.

To summarize, our core contributions are as follows:

- **Geometric formalization of the OT latent space.** We show that the preimage of an m-dimensional data manifold under the OT map is itself an m-dimensional manifold. This formally defines "the latent twin" of a data manifold, with tangent spaces explicitly linked by the inverse Hessian of the Brenier potential (Proposition 1).
- **OT-based local manifold model.** We propose a latent linear model that approximates the preimage of the data manifold in the reference space. Pushing this subspace forward through a pretrained OT-CFM flow produces a naturally curved local approximation of the data manifold (Definition 1).
- **Rigorous approximation error bounds.** We prove our latent approximation achieves an asymptotically tighter leading-order error bound than the standard tangent-plane approximation used by local PCA. We demonstrate that this error reduction is directly proportional to the spectral anisotropy of the OT Jacobian (Corollary 1).
- **Scalable model inference algorithm.** To bypass the model's intractable likelihood, we adopt a simulation-based inference approach. By estimating the Fisher score from simulated observation-parameter pairs, we perform likelihood-based inference efficiently via gradient ascent on the learned score.

2 Background

2.1 Optimal Transport

Given two probability measures P_0 and P_1 on $\mathbb{R}^d$, the *Monge optimal transport problem* with quadratic cost seeks a map $G : \mathbb{R}^d \to \mathbb{R}^d$ that pushes P_0 forward to P_1 (denoted $G_{\#}P_0 = P_1$) while minimizing the transport cost:

$$\inf_{G:\, G_{\#}P_0 = P_1} \int \|z - G(z)\|^2 \, \mathrm{d}P_0(z).$$

Brenier's theorem [6] establishes that when P_0 has a density with respect to Lebesgue measure, the unique minimizer takes the form $G = \nabla\varphi$ for a convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$. The map $G = \nabla\varphi$ is called the *Brenier map*, and φ is the *Brenier potential*. The convexity of φ implies that the Jacobian $J_G = \nabla^2\varphi$ is positive semidefinite.

Computing the exact Brenier map over large datasets is intractable. Earlier approaches approximate the transport map through entropy-regularized linear programming via the Sinkhorn algorithm [10], parameterization of the Brenier potential φ with input convex neural networks [28, 3], or numerical solution of the Monge–Ampère PDE [15]. More recent works focus on training straight and continuous flows that closely approximate the OT map [25, 43].

2.2 OT Flow Matching

Flow Matching (FM) [24] learns a continuous-time flow that transports a reference distribution $P_0 = \mathcal{N}(0, I_d)$ to the data distribution P_1. The flow is defined by the ODE

$$\dot{z} = v_\theta(z, t), \quad z(0) = z_0 \sim P_0, \tag{1}$$

and the transport map is the time-1 map $G_\theta(z_0) := z(1)$. In general, the transport map of FM does not approximate the OT flow as the trajectory of the ODE may be strongly curved. Liu et al. [25] propose to train the flow repeatedly, to straighten the flow trajectories, approximating the OT map.

Optimal Transport Conditional Flow Matching (OT-CFM) [43] improves upon standard FM by coupling the source and target samples via mini-batch optimal transport. The velocity field v_θ is learned by minimizing the conditional flow matching objective:

$$\mathcal{L}_{\text{OT-CFM}}(\theta) = \mathbb{E}_{t,(z_0,x_1)\sim\pi^*} \left\| v_\theta(z_t, t) - (x_1 - z_0) \right\|^2, \tag{2}$$

where $t \sim \text{Uniform}(0, 1)$, $z_t = (1 - t)z_0 + t\, x_1$ is the linear interpolant, and π^* is the optimal transport coupling within each mini-batch, obtained by solving the assignment problem $\pi^* = \arg\min_\pi \sum_{ij} \pi_{ij} \|z_{0,i} - x_{1,j}\|^2$.

The OT coupling reduces the variance of the velocity field targets, encouraging straighter transport paths and closely approximating the true OT map. In our paper, we train G_θ using OT-CFM to approximate the true OT map in experiments.

2.3 Score Estimation and Likelihood-Free Inference

Classical latent-variable models such as PPCA [42] and GPLVM [22] admit tractable marginal likelihoods because of their Gaussian structure. In contrast, many modern implicit, latent variable models — including the model proposed in this paper — are easy to sample from, but have intractable likelihoods, making direct parameter inference difficult. Specifically, for a parameter ψ, one can simulate observations $Y \sim p(\cdot \mid \psi)$, but cannot evaluate the likelihood $p(Y \mid \psi)$ in closed form. This is the setting of simulation-based inference (SBI) [9].

Recent score-based SBI methods avoid explicit conditional density estimation over high-dimensional observations. Instead, they leverage the technique of denoising score matching (DSM) [45] to learn the parameter-space score function, from simulated parameter-observation pairs.

Neural posterior score estimation [37, 13] learns the posterior score across a diffusion noise schedule, and generates posterior samples via the reverse-time dynamics. Similarly, the *likelihood score* $s(\psi; Y) = \nabla_\psi \log p(Y \mid \psi)$ can also be directly targeted [21, 17]. The posterior and likelihood scores are connected by Bayes' rule, and if the prior score is known, the likelihood score can be recovered from the posterior score. Once either the posterior score or likelihood score is estimated, the MLE/MAP can be obtained through stochastic gradient ascent [35].

3 Manifold Latent Variable Models

Problem Setting: Let $\{Y_i\}_{i=1}^N \subset \mathbb{R}^d$ be observations sampled from a distribution P_1 supported on or near an m-dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$, where $m \ll d$. Given a query point $y_0 \in \mathbb{R}^d$ on, or near $\mathcal{M}$, the goal is to identify the local tangent space $T_{y_0}\mathcal{M}$ and normal space $N_{y_0}\mathcal{M}$ at y_0.

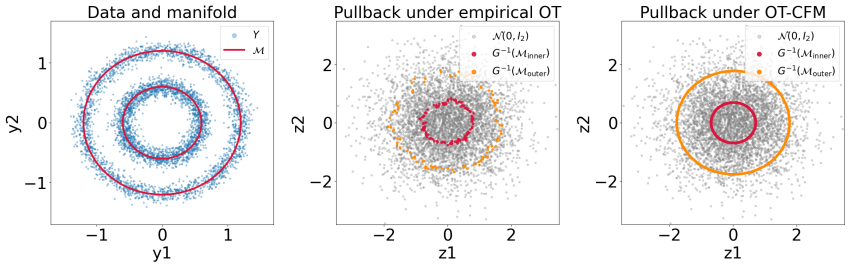


Figure 1: Illustration of $G^{-1}(\mathcal{M})$ for a noisy double-ring target $\mathcal{M}$ with reference $\mathcal{N}(0, I_2)$. Both empirical OT (center) and OT-CFM (right) reveal clear manifold structure in the preimage of $\mathcal{M}$ — specifically, two disjoint rings of the same dimension as $\mathcal{M}$, consistent with Proposition 1. Empirical OT (Hungarian on 5000×5000 samples) yields only a discrete coupling rather than a continuous map, so we plot the preimage using data points lying close to $\mathcal{M}$, producing a noisy point cloud. The continuous OT-CFM map, in contrast, can be inverted at any point of $\mathcal{M}$ via reverse-time ODE integration, yielding clean preimage circles. Gray points are reference samples.

3.1 Optimal Transport Pullback Preserves Manifolds

Suppose G is a Brenier OT map that transports a d-dimensional reference distribution to this data distribution. We show that the preimage of $\mathcal{M}$ under G is also an m-dimensional manifold $\mathcal{M}_0$. In other words, the pullback of the data manifold $\mathcal{M}$ is also an m-dimensional manifold.

Proposition 1 (Manifold preservation under OT pullback)**.** *Let $P_0 = \mathcal{N}(0, I_d)$ and let P_1 be a probability measure on $\mathbb{R}^d$ with density f_1 satisfying $0 < \lambda \le f_1 \le \Lambda < \infty$ on compact sets (e.g., P_1 is a Gaussian-smoothed distribution supported near a smooth m-dimensional submanifold $\mathcal{M}$). Let $G = \nabla\varphi$ be the Brenier optimal transport map from P_0 to P_1. Then:*

1. *$\mathcal{M}_0 := G^{-1}(\mathcal{M})$ is a smooth m-dimensional submanifold of $\mathbb{R}^d$.*
2. *For each $y \in \mathcal{M}$ with $z = G^{-1}(y)$,*

$$T_z\mathcal{M}_0 = [\nabla^2\varphi(z)]^{-1} \cdot T_y\mathcal{M}. \tag{3}$$

The proof of this result can be found in Appendix A. Proposition 1 shows that a manifold in the data space will also have a "twin" manifold structure in the reference space, with the exact same dimension, whose tangent space is precisely defined under (3). See Figure 1 for an illustration.

3.2 OT Latent Variable Model

In this paper, instead of directly modeling the manifold $\mathcal{M}$ in the data space, we propose to model its latent twin in the reference space. We specify the latent manifold to be an affine subspace, which is then pushed through the OT map G, to produce a curved approximation of the data manifold $\mathcal{M}$. Now we present our generative model for an observation Y near a query point y_0:

Definition 1 (Manifold Latent Variable Model)**.** *Given a query point $y_0 \in \mathbb{R}^d$ and a pretrained OT flow $G_\theta : \mathbb{R}^d \to \mathbb{R}^d$, define $w_0 := G_\theta^{-1}(y_0)$. The generative model for observations $Y \in \mathbb{R}^d$ near y_0 is:*

$$\boxed{Y = G_\theta(w_0 + Wx) + \Pi_\perp(x, W)\,\varepsilon, \qquad x \sim \mathcal{N}(0, I_m), \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_d),} \tag{4}$$

where the components are defined as follows:

1. ***Signal direction.*** *$W \in \mathbb{R}^{d\times m}$ is the parameter to be estimated. Its columns span the direction space of the m-dimensional affine subspace $\{w_0 + Wx : x \in \mathbb{R}^m\}$ in the reference space.*
2. ***Signal manifold.*** *The image $\mathcal{M}_W := \{G_\theta(w_0 + Wx) : x \in \mathbb{R}^m\}$ is a curved m-dimensional submanifold in data space that locally approximates $\mathcal{M}$ near y_0.*
3. ***Local tangent space.*** *At latent coordinate x, the tangent space of $\mathcal{M}_W$ at the signal point $G_\theta(w_0 + Wx)$ is spanned by the columns of*

$$T(x, W) = J_{G_\theta}(w_0 + Wx) \cdot W \in \mathbb{R}^{d\times m}, \tag{5}$$

where $J_{G_\theta}(z) = \partial G_\theta / \partial z$ is the Jacobian of the OT map.

4. ***Normal noise.*** *Let $Q(x, W) \in \mathbb{R}^{d\times m}$ be the orthonormal basis of* $\mathrm{col}(T(x, W))$ *obtained via QR decomposition. The noise is projected onto the $(d - m)$-dimensional normal subspace via*

$$\Pi_\perp(x, W) = I_d - Q(x, W)\, Q(x, W)^\top. \tag{6}$$

Note that another tempting choice of latent model is

$$Y = G_\theta(w_0 + Wx + \epsilon), \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 I_d),$$

where isotropic noise is added to the linear model and then pushed forward through the OT map. This model has a tractable likelihood: applying the change of variables rule gives

$$p(Y \mid W) = p_Z\big(G_\theta^{-1}(Y); W\big)\, \big|\det J_{G_\theta^{-1}}(Y)\big|,$$

where $Z := w_0 + WX + \epsilon \sim \mathcal{N}(w_0, WW^\top + \sigma^2 I)$ is Gaussian. Since the Jacobian determinant does not depend on W, MLE reduces to fitting PCA to the pullback local samples $\{G_\theta^{-1}(Y_{i,\mathrm{local}})\}$. We refer to this method as *Pullback PCA*.

Although computationally attractive, this latter model does not reflect the manifold data-generation assumption. In many applications, data points are sampled from a manifold $\mathcal{M}$ (via the latent coordinate x) and then perturbed by measurement errors or environmental fluctuations that displace them *away from* $\mathcal{M}$. Pushing reference-space error through the nonlinear G_θ produces a different noise structure that does not match how noise enters observations in practice.

This also motivates our choice of structured noise instead of the isotropic Gaussian noise $\mathcal{N}(0, \sigma^2 I_d)$ used in classical latent variable models such as PPCA [42]. The projection $\Pi_\perp$ ensures that all on-manifold variation is attributed to the signal x and all off-manifold variation is attributed to the noise ε. This separation makes the model more interpretable — W controls the signal subspace and σ controls the off-manifold noise level — and ensures that the likelihood faithfully represents the assumed data-generating mechanism. We will benchmark our method against Pullback PCA.

Although Proposition 1 shows the preimage of $\mathcal{M}$ is also a manifold, it does not state the local manifold structure in the reference space can be approximated by a linear model. The advantage of pushing the linear model through an OT map remains unclear. In the next section, we show why fitting a local linear model in the reference space is much less restrictive than fitting a local linear model directly in the data space.

3.3 Local Linear Approximation in Reference or Data Space?

A classic approach to identifying the local manifold structure around y_0 is local PCA, which fits a linear model $y_0 + Wx$ in the data space using samples in the neighborhood of y_0; the tangent space at y_0 is then estimated as the column span of W. Our model (4), by contrast, fits a linear model $w_0 + Wx$ in the *reference space* to the preimage $G^{-1}(\mathcal{M})$ and pushes it through G_θ to obtain a curved approximation of the data manifold. The question is: which is better?

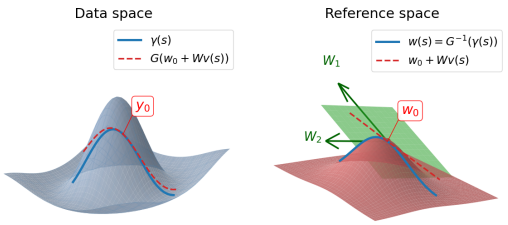


Figure 2: Geometric objects in Definition 2. $\gamma(s)$ is a curve on $\mathcal{M}$. $w(s)$ is its preimage on $\mathcal{M}_0$. The red dashed line on the right is $w_0 + Wv(s)$, the orthogonal projection of $w(s)$ onto $w_0 + T_{w_0}\mathcal{M}_0$. Its pushforward through G is the red dashed line on the left. $v(s)$ represents the coordinates on the green plane.

Intuitively, fitting a linear model in the reference space yields a better approximation. In generative modeling, the reference distribution is typically full-dimensional and isotropic (e.g., Gaussian), while the target distribution concentrates on a thin neighborhood of the data manifold $\mathcal{M}$. The OT map G_θ achieves this concentration by strongly compressing distances in directions perpendicular to $\mathcal{M}$ — the spectral gap of $\nabla^2\varphi$ formalizes this [12]. Consequently, when we approximate the manifold by a linear model in the reference space, any approximation error in the normal direction is "compressed" by the OT map.

The following theoretical discussion makes this intuition precise.

Definition 2 (Local geometry at y_0)**.** *Let $y_0 \in \mathcal{M}$ be the query point in the data space, and let $w_0 = G^{-1}(y_0) \in \mathcal{M}_0$ be its latent preimage. We denote the maximum extrinsic curvatures of $\mathcal{M}$ at y_0 and $\mathcal{M}_0$ at w_0 by κ and κ_0, respectively.*

Let $T \in \mathbb{R}^{d\times m}$ and $W \in \mathbb{R}^{d\times m}$ be matrices whose orthonormal columns span $T_{y_0}\mathcal{M}$ and $T_{w_0}\mathcal{M}_0$, respectively. For any smooth curve $\gamma : (-\epsilon, \epsilon) \to \mathcal{M}$ with $\gamma(0) = y_0$, parameterized by arc length, define the following latent and ambient coordinate trajectories:

$$w(s) := G^{-1}(\gamma(s)) \in \mathcal{M}_0, \quad v(s) := W^\top\big(w(s) - w_0\big) \in \mathbb{R}^m, \quad u(s) := T^\top\big(\gamma(s) - y_0\big) \in \mathbb{R}^m.$$

In plain words, $\gamma(s)$ is a smooth curve on $\mathcal{M}$, while $w(s)$ is its preimage on $\mathcal{M}_0$. $w_0 + Wv(s)$ is the orthogonal projection of $w(s)$ onto the affine tangent plane $w_0 + T_{w_0}\mathcal{M}_0$, thus $v(s)$ can be seen as the coordinates on this tangent plane. Similarly, $u(s)$ acts as coordinates on the ambient tangent plane. We plot these objects in Figure 2.

We now impose a local contraction assumption on the OT map G.

Assumption 1 (Anisotropic Contraction)**.** *Let $w_0 = G^{-1}(y_0) \in \mathcal{M}_0$ and $J := J_G(w_0)$. There exist $\rho \in (0, 1)$ and constants $0 < c \le C < \infty$, independent of ρ, such that*

$$\begin{aligned} c\|v\| \le \|Jv\| \le C\|v\|, \ \forall v \in T_{w_0}\mathcal{M}_0 \\ \|Jn\| \le \rho\|n\|, \ \forall n \in N_{w_0}\mathcal{M}_0 \end{aligned} \tag{7}$$

Essentially, we assume that the Jacobian contracts the normal direction of $\mathcal{M}_0$ by at most, a factor of ρ, while along the tangent direction, we assume that it is bounded above and below, independent of ρ. Note that this assumption imposes a constraint purely on the magnitude of the tangent and normal spaces, and does not require further alignment with the eigenvectors of the Jacobian.

In the manifold-concentration regime studied by Fukumizu et al. [12], ρ is connected to the off-manifold noise scale of P_1 and the assumption holds when the data is concentrated near $\mathcal{M}$. We provide further validation of the spectral gap in Appendix G.1 and Figure 8.

Corollary 1 (Local linear approximation through the OT flow)**.** *Suppose Proposition 1 and Assumption 1 hold. With the notation of Definition 2, the **latent tangent-plane approximation** incurs the approximation error*

$$\big\|\gamma(s) - G\big(w_0 + Wv(s)\big)\big\| = O\big(\rho\,\kappa_0\,s^2\big) + O(s^3), \tag{8}$$

*while the standard **ambient tangent-plane approximation** incurs the approximation error*

$$\big\|\gamma(s) - \big(y_0 + Tu(s)\big)\big\| = O(\kappa\,s^2) + O(s^3), \tag{9}$$

as $s \to 0$. Hence, whenever $\rho\kappa_0 \ll \kappa$, our model yields an asymptotically smaller linearization error.

The proof of the corollary can be found in Appendix B.

The corollary formalizes the intuition that the OT flow "does the curving for us". Initially, the pullback manifold $\mathcal{M}_0$ in the reference space is approximated linearly with curvature error $O(\kappa_0 s^2)$. When the pullback manifold is mapped back to data space via G_θ, the OT map's normal contraction shrinks it by a further factor of ρ, where ρ is the normal contraction factor in Assumption 1. The tangent-plane approximation in data space enjoys no such contraction, so its linear approximation error remains at $O(\kappa s^2)$. This is the theoretical justification for our latent variable model (4): by parameterizing the signal as a linear subspace in reference space and pushing it through G_θ, we obtain a local approximation that is tighter than data-space tangent approximation such as local PCA whenever $\rho\kappa_0 \ll \kappa$.

The corollary predicts that the reference space linear approximation is valid over a neighborhood of radius $O(\rho^{-1/2})$ times larger than the tangent-plane approximation. By Fukumizu et al. [12], ρ scales with the off-manifold noise scale, and is small when the data is distributed tightly around the manifold. The flow-based approximation therefore enjoys a substantially larger usable neighborhood, allowing the method to exploit more local data along the curved manifold for estimating W. We validate this prediction in Figure 3.

Note that, although Corollary 1 considers the tangent plane of $\mathcal{M}_0$ at w_0, the normal contraction behavior suggests that alternate latent affine planes through w_0 could also provide an accurate local approximation to the manifold $\mathcal{M}$. We provide an illustration of this in Appendix C, comparing our fitted latent model with alternative affine planes.

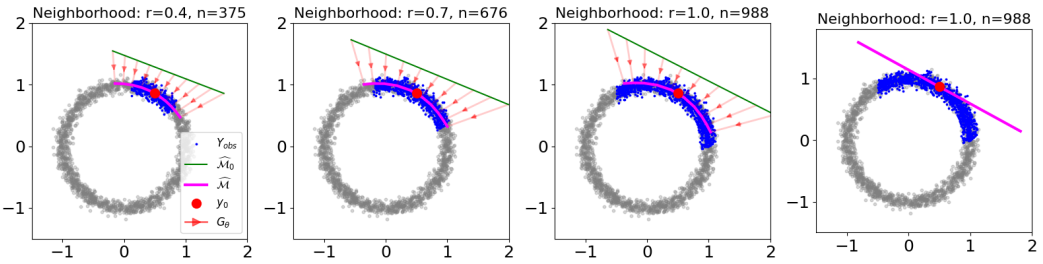


Figure 3: **Effective neighborhood of linear approximation in reference and data space.** Across panels, the local neighborhood used for fitting (blue, $\|y - y_0\| < r$) is enlarged from $r = 0.4$ to $r = 1.0$. The first three panels show a linear latent model $\widehat{\mathcal{M}}_0 = w_0 + \widehat{W}x$ (green) fitted by Algorithm 1 and pushed forward through the OT map G_θ (red arrows) to give the estimated manifold $\widehat{\mathcal{M}}$ (magenta). The estimated manifold tracks the true circular manifold across the full neighborhood at every radius, *including the largest one*, despite the latent model being strictly linear. This robustness is due to the OT map's normal contraction: linearization error along the direction perpendicular to the manifold is strongly suppressed by G_θ, so a flat latent approximation produces an accurate curved data-space estimate even over wide neighborhoods. **Right:** local PCA fit in the data space at $r = 1.0$ (same neighborhood as the third panel), shown for comparison. Without an OT map to absorb normal error, the linear fit is rigid and fails to track the true manifold on both ends of the local neighborhood. See Figure 7 for a more detailed comparison and Figure 4 for a quantitative comparison.

4 Likelihood-Free Inference

To estimate the parameter W in our latent variable model (4), one might resort to likelihood-based inference. However, the marginal likelihood, i.e., $p(Y \mid W) = \int p(Y \mid x, W)\, p(x)\mathrm{d}x$ is analytically intractable because both the signal $G_\theta(w_0 + Wx)$ and the noise projection $\Pi_\perp(x, W)$ depend on the latent variable x through the OT map G_θ. This prevents the application of Gaussian identities. However, the model is *trivial to simulate from*: sample x and ε, compute $G_\theta(w_0 + Wx)$ by integrating the flow, obtain $T(x, W)$ via Jacobian-vector product (JVP), then Q via QR decomposition, and finally add the projected noise $\Pi_\perp(x, W)\varepsilon$. The intractability of likelihood and the simplicity of simulation motivate a *simulation-based likelihood inference*.

We adopt DSM [45] to learn the score of the posterior distribution $\nabla_W \log p(W \mid Y)$ from simulated parameter–observation pairs. Place a Gaussian prior on the parameters: $W_{i,j} \sim \mathcal{N}(0, \sigma^2_{\text{prior}})$. We generate training data by sampling W from this prior and simulating Y from model (4), producing pairs (Y, W) from the joint distribution $p(Y, W) = p(Y \mid W)p(W)$. We then corrupt the parameters by adding Gaussian noise:

$$\widetilde{W}_{i,j} = W_{i,j} + \sigma_w \eta_{i,j}, \qquad \eta_{i,j} \sim \mathcal{N}(0, 1) \tag{10}$$

where $\sigma_w > 0$ controls the corruption level. A score model $s_\phi(Y, \widetilde{W})$ is trained to predict the denoising direction by minimizing

$$\mathcal{L}_{\text{DSM}}(\phi) = \mathbb{E}_{W,Y,\eta} \left\| s_\phi(Y, \widetilde{W}) + \frac{\eta}{\sigma_w} \right\|^2. \tag{11}$$

By the classical result of Vincent [45], the minimizer of (11) satisfies

$$s_{\phi^*}(Y, \widetilde{W}) = \nabla_{\widetilde{W}} \log p_{\sigma_w}(\widetilde{W} \mid Y), \tag{12}$$

the score of the noise-smoothed posterior. As $\sigma_w \to 0$, this converges to the true posterior score $\nabla_W \log p(W \mid Y)$. The likelihood score can then be approximated by subtracting the known prior score via Bayes' rule:

$$\nabla_W \log p(Y \mid W) \approx \underbrace{s_{\phi^*}(Y, W)}_{\approx \nabla_W \log p(W|Y)} \underbrace{- \nabla_W \log p(W)}_{= -W/\sigma^2_{\text{prior}}}. \tag{13}$$

Note, instead of MLE, one may also perform maximum a posteriori (MAP) on W, then the subtraction of the prior score in (13) is optional. Once $\nabla_W \log p(Y \mid W)$ is estimated, a simple gradient ascent algorithm can be used to approximate the MLE with $\widehat{W}$. We provide more details about the likelihood-free inference procedure in Appendix D as well as the full inference algorithm in Algorithm 1.

5 Related Work

Local manifold estimation aims to recover the local geometry of a target manifold $\mathcal{M}$ around a query point, using local observations. Local PCA estimates tangent spaces from the principal components of these local observations [18, 44], providing a first-order estimation with neighborhood selection balancing between the curvature-induced bias and sampling variability. Higher-order approaches include local polynomial estimators [2] and manifold moving least squares (MMLS) [39], which construct local polynomial approximations. Higher-order methods are able to model non-linear geometry, using only local observations. However, our setting additionally incorporates information from a broader data distribution through a pretrained generative map, allowing us to obtain a better local estimation.

Generative models provide a complementary source of geometric information, either through explicit manifold modeling, or implicitly from pretrained models. For example, manifold-learning flows learn manifold embeddings and related densities directly [5, 8], while pretrained diffusion models have been used for intrinsic dimension estimation [41, 19] and tangent-space projections [20].

Our approach combines a pretrained generative model with local statistical fitting. For each query point, we use a fixed, pretrained generative map, and directly estimate an affine subspace in its reference space from local observations. The image of this subspace under the generative map defines a nonlinear local manifold, thus allowing us to leverage the geometric information from the broader data distribution through the pretrained generative model. Furthermore, under conditions on the normal contraction property of the generative map, we are able to show how this approach reduces the approximation error over tangent-space approximation in data space.

6 Experiments

We evaluate our approach against two primary baselines: *local PCA* [18] and *pullback PCA* (Section 3.2). Comparing against local PCA validates the asymptotic error reduction established in Corollary 1. Meanwhile, the comparison with pullback PCA isolates the importance of properly modeling structured noise in the ambient space.

The two primary baselines and our method estimate the linear basis $\widehat{W}$ from the same local training observations around the query y_0. As a further comparison, we include *Jacobian-SVD*, which uses the leading right singular vectors of $J_{G_\theta}(w_0)$ to define a latent affine model and pushes this model through the flow. Its basis is obtained without fitting to local observations, but leverages the same pretrained OT-flow as our method.

The estimated manifolds are then parameterized directly in data space for local PCA, and through the latent affine models for the other methods:

- **Data space:** $\widehat{\mathcal{M}} = \{y_0 + \widehat{W}x \mid x \in \mathbb{R}^m\}$.
- **Latent space:** $\widehat{\mathcal{M}} = \{G_\theta(w_0 + \widehat{W}x) \mid x \in \mathbb{R}^m\}$.

For our synthetic data experiments and ablation studies, we define the true data manifold $\mathcal{M}$ as a 1D ring of radius 1 for simplicity. We quantify the performance of each method by computing the mean and maximum orthogonal distances from the estimated manifold $\widehat{\mathcal{M}}$ to the true manifold $\mathcal{M}$ in the local neighborhood. In the real-world image data experiments, we evaluate the quality of $\widehat{\mathcal{M}}$ through a held-out reconstruction error, which measures how closely $\widehat{\mathcal{M}}$ reproduces the neighbors of y_0 in a held-out dataset. Further details on experimental settings can be found in Appendices F and H.

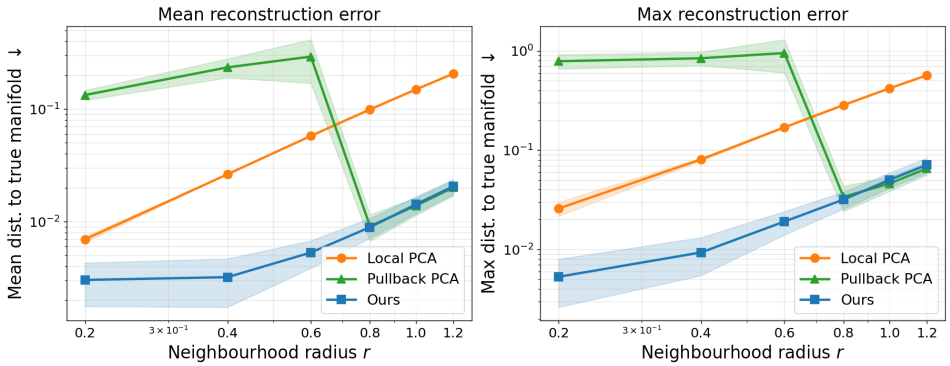


Figure 4: Reconstruction error vs. neighborhood radius r on a noisy unit circle embedded in a 2D space. **Left**: mean error; **right**: max error. Error bars indicate the standard deviation over 12 query points $\times$ 10 seeds.

6.1 Manifold Estimation Error with Growing Neighborhood

Corollary 1 states that the estimated local manifold of all methods would incur larger linearization error as the neighborhood increases. We test this claim quantitatively on the same synthetic dataset used in the visualization in Figure 3. It can be seen in Figure 4, for all methods, as the neighborhood radius (r) increases, the manifold reconstruction error increases. Moreover, Corollary 1 predicts that our method will have a linearization error that is lower than the local PCA, up to a constant that is determined by the spectral gap defined in Assumption 1. Indeed, we can see the reconstruction error of our method also increases, but remains lower than the local PCA up to a constant. We can also observe that when the neighborhood radius is small, pullback PCA does not perform well. However, its error drops suddenly as the neighborhood increases. This is because when the neighborhood is small, the off-manifold noise overwhelms the local manifold structure. After the inverse OT, the signal, i.e., the preimage manifold structure is unidentifiable from the pullback samples. However, when the neighborhood is large, the local structure survives the inverse OT map, thus the pullback PCA can recover the correct tangent space of the preimage. In comparison, our model works robustly in both small and large neighborhood radius settings. In Appendix G, we further demonstrate our model's robustness to varying off-manifold noise levels (Figure 9) and flow training quality (Figure 8), consistently outperforming local PCA and pullback PCA.

6.2 Local Manifold Recovery on Real-world Images

We now move on to high-dimensional image datasets, which exhibit different notions of locality. For the MNIST digits dataset [1], the locality is visual similarity within the same digit class. For the Frey face dataset [36], since images are collected from a single video, locality is represented by temporal contiguity. Finally, for the COIL-20 dataset [30], which contains various images captured from different viewpoints, we use this viewpoint angle as a notion of locality. For each dataset, we train an OT-CFM model, and estimate a local manifold around a small neighborhood of a query or reference image y_0. We provide further details of this performance metric and chosen local observations in Appendix H.

Figure 5 presents the quantitative results for the MNIST and Frey Faces datasets. Across both datasets, our proposed method consistently achieves a lower held-out reconstruction error than the baselines, particularly maintaining its advantage as the latent dimension m increases (right column). Notably, the purely local Jacobian-SVD baseline performs poorly across all settings, demonstrating that simply extracting the singular directions of OT flow's Jacobian at w_0 is insufficient at recovering the local geometry. We also observe a natural geometric degradation in the MNIST radius sweep (top-left): while our method provides a substantial reduction in error at small to moderate radii, the error eventually converges toward that of local PCA at the largest radius ($r \approx 0.8$), showing the inherent limits of fitting a linear latent model over a highly nonlinear image neighborhoods. Further qualitative visualizations of these local manifold approximations, alongside further experimental results and details are provided in Appendix H.

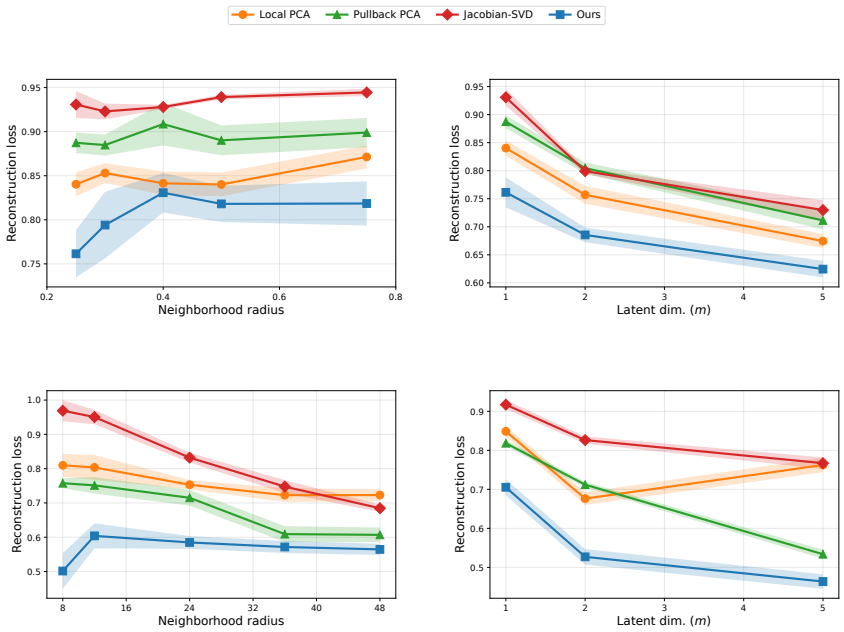


Figure 5: Held-out reconstruction loss with increasing neighborhood size and latent dimension on the MNIST and Frey Faces datasets, in the top and bottom rows, respectively. (Lower is better, error bars indicate the 95% confidence interval)

7 Conclusion and Limitations

We propose a novel latent variable model for identifying local manifold structure by modeling the preimage of the manifold in the reference space, and pushing it through the OT map to get a curved manifold in data space. Although experiments on toy and real-world datasets have been promising, our approach has several limitations that present exciting avenues for future work. First, the validity of Assumption 1 needs further empirical verification. The middle panel of Figure 8 shows that it holds approximately for well-trained OT-CFM flows on low-dimensional toy targets. Whether the assumption holds for more general high-dimensional datasets is an open empirical question. Second, our generative model assumes the intrinsic dimension m of the data manifold is known a priori. Automatically determining the intrinsic dimension within our framework is a direction for future work. Third, despite the effectiveness of our method on real-world image datasets, estimating a high-dimensional W remains a challenging problem. How to leverage inductive bias to effectively compute the Fisher score is an interesting future direction.

Acknowledgments and Disclosure of Funding

The authors thank the three anonymous reviewers for their constructive feedback. SK was supported by the EPSRC Centre for Doctoral Training in Computational Statistics and Data Science, grant number EP/S023569/1.

References

[1] MNIST Database of Handwritten Digits, 1998.

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

[3] Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

[4] Silvere Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

[5] Johann Brehmer and Kyle Cranmer. Flows for simultaneous manifold learning and density estimation. *Advances in Neural Information Processing Systems*, 33:442–453, 2020.

[6] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

[7] Luis A Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

[8] Anthony L Caterini, Gabriel Loaiza-Ganem, Geoff Pleiss, and John P Cunningham. Rectangular flows for manifold learning. *Advances in Neural Information Processing Systems*, 34:30228–30241, 2021.

[9] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[10] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[11] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[12] Kenji Fukumizu, Wei Huang, Han Bao, Shuntuo Xu, and Nisha Chandramoorthy. Flow matching from viewpoint of proximal operators. *arXiv preprint arXiv:2602.12683*, 2026.

[13] Tomas Geffner, George Papamakarios, and Andriy Mnih. Compositional score modeling for simulation-based inference. In *International Conference on Machine Learning*, pages 11098–11116. PMLR, 2023.

[14] David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, pages 2404–2414. PMLR, 2019.

[15] Xianfeng Gu, Feng Luo, Jian Sun, and Shing-Tung Yau. Variational principles for Minkowski type problems, discrete optimal transport, and discrete Monge–Ampère equations. *Asian Journal of Mathematics*, 20(2):383–398, 2016.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[17] Haoyu Jiang, Yuexi Wang, and Yun Yang. Likelihood-free inference via structured score matching. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 1567–1575. PMLR, 02–05 May 2026.

[18] Nandakishore Kambhatla and Todd K. Leen. Dimension reduction by local principal component analysis. *Neural Computation*, 9(7):1493–1516, 1997. doi: 10.1162/neco.1997.9.7.1493.

[19] Hamidreza Kamkari, Brendan Leigh Ross, Rasa Hosseinzadeh, Jesse C Cresswell, and Gabriel Loaiza-Ganem. A geometric view of data complexity: Efficient local intrinsic dimension estimation with diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pages 38307–38354, 2024.

[20] Andrey Kharitenko, Zebang Shen, Riccardo De Santi, Niao He, and Florian Dorfler. Landing with the score: Riemannian optimization through denoising. In *International Conference on Learning Representations*, volume 2026, pages 129064–129104, 2026.

[21] Sherman Khoo, Yakun Wang, Song Liu, and Mark Beaumont. Direct Fisher score estimation for likelihood maximization. In *Advances in Neural Information Processing Systems*, 2025.

[22] Neil D Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems*, volume 16, 2004.

[23] John M Lee. *Introduction to Riemannian Manifolds*. Springer, 2nd edition, 2018.

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

[25] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

[26] Gabriel Loaiza-Ganem, Brendan Leigh Ross, Rasa Hosseinzadeh, Anthony L. Caterini, and Jesse C. Cresswell. Deep generative models through the lens of the manifold hypothesis: A survey and new connections. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a90WpmSi0I.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[28] Ashok Vardhan Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

[29] Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

[30] S.A. Nene, S.K. Nayar, and H. Murase. Columbia Object Image Library (COIL-20). In *Technical Report, Department of Computer Science, Columbia University CUCS-005-96*, Feb 1996.

[31] Arkadas Ozakin and Alexander G Gray. Submanifold density estimation. In *Advances in Neural Information Processing Systems*, volume 22, 2009.

[32] George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[33] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *International Conference on Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.

[34] Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

[35] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[36] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[37] Louis Sharrock, Jack Simons, Song Liu, and Mark Beaumont. Sequential neural score estimation: Likelihood-free inference with conditional score based diffusion models. In *International Conference on Machine Learning*, pages 44565–44602. PMLR, 2024.

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

[39] Barak Sober and David Levin. Manifold approximation by moving least-squares projection (MMLS). *Constructive Approximation*, 52(3):433–478, 2020.

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

[41] 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*, pages 46412–46440. PMLR, 2024. URL `https://proceedings.mlr.press/v235/stanczuk24a.html`.

[42] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B*, 61(3):611–622, 1999.

[43] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

[44] Hemant Tyagi, Elif Vural, and Pascal Frossard. Tangent space estimation for smooth embeddings of Riemannian manifolds. *Information and Inference: A Journal of the IMA*, 2(1):69–114, 2013. doi: 10.1093/imaiai/iat003.

[45] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

A Proof of Proposition 1

Proof. First, for any Brenier transport map, the vector field is the gradient field of a convex function φ.

By Caffarelli's regularity theorem [7], with the densities f_0 and f_1 lower-bounded by a constant $\epsilon > 0$, $\nabla\varphi$ is smooth, $\nabla^2\varphi \succ 0$, and its eigenvalues are also lower-bounded by a constant $\lambda_{\min} > 0$. The Inverse Function Theorem states that the inverse transport $G^{-1} = (\nabla\varphi)^{-1}$ exists and is also smooth.

Since $\mathcal{M}$ is a smooth manifold, for an arbitrary point $y \in \mathcal{M}$, there exists an open neighborhood $U \subset \mathbb{R}^d$ and a smooth function $\psi : U \to \mathbb{R}^{d-m}$, such that $U \cap \mathcal{M} = \psi^{-1}(0)$.

Define $V := G^{-1}(U)$. Then we can rewrite the local preimage $V \cap \mathcal{M}_0$ as $\{z \in V : G(z) \in \mathcal{M}\} = \{z \in V : \psi(G(z)) = 0\} = (\psi \circ G)^{-1}(0)$.

The composition $\psi \circ G$ is a map from V to $\mathbb{R}^{d-m}$. We know that J_ψ has rank $d - m$ by definition, and $J_G = \nabla^2\varphi$ is invertible. Therefore, $J_{\psi\circ G} = J_\psi J_G$ has full row rank $d - m$ everywhere in V, which means that 0 is a regular value of $\psi \circ G$.

Applying the Preimage Theorem to $\psi \circ G$, we get that the level set $\{z \in V : (\psi \circ G)(z) = 0\}$ is a manifold with dimension $d - (d - m) = m$. Because this holds locally in a neighborhood V for every $z \in \mathcal{M}_0$, the entire preimage $\mathcal{M}_0$ is an m-dimensional manifold. This proves our first claim.

Furthermore, the Preimage Theorem states that the tangent space is given by the kernel of the Jacobian:

$$T_z\mathcal{M}_0 = \ker(J_{\psi\circ G}) = \ker(J_\psi J_G) = J_G^{-1}\ker(J_\psi) = J_G^{-1}T_y(\mathcal{M}),$$

where the third equality is due to the fact that $J_G = \nabla^2\varphi$ is an invertible linear operator (guaranteed by Caffarelli's theorem), and the fourth equality follows from the definition of ψ, which dictates that $T_y\mathcal{M} = \ker(J_\psi)$. This concludes the proof. □

B Proof of Corollary 1

Proof. Our proof consists of three parts: firstly, we derive a linear approximation error in the reference space, then analyze the error after pushing through G_θ, and finally, derive and compare with the linear approximation error in the data space. For clarity, we refer to the pretrained OT map G_θ as G.

Step 1: Linear approximation in the reference space. The pullback curve $w(s) = G^{-1}(\gamma(s))$ lies on the smooth m-dimensional submanifold $\mathcal{M}_0$. Using the curve $v(s) = W^\intercal(w(s) - w_0)$ from Definition 2, we have:

$$w(s) - w_0 - Wv(s) = (I_d - WW^\intercal)(w(s) - w_0)$$

Note that since $w'(0) \in T_{w_0}\mathcal{M}_0$, we can perform a Taylor expansion to obtain:

$$w(s) = w_0 + Wv(s) + \frac{1}{2}N_0 s^2 + O(|s|^3) \tag{14}$$

where $N_0 = (I_d - WW^\intercal)\ddot{w}_0 \in N_{w_0}\mathcal{M}_0$ is the normal component of the acceleration $\ddot{w}(0)$.

Because γ is parameterized by the arc length $Jw'(0) = \gamma'(0)$, from Assumption 1, this implies:

$$c\|w'(0)\| \le \|Jw'(0)\| = 1 \le C\|w'(0)\|$$

In particular, note that $\|w'(0)\| \le 1/c$

By the second-fundamental-form characterization of submanifold curvature (e.g., Lee [23, Chapter 8]),

$$\|N_0\| \le \kappa_0\|w'(0)\|^2 \le \frac{\kappa_0}{c^2}$$

Step 2: Pushing the approximation through G. Rearranging (14), and performing a Taylor expansion of the map G around the point $w(s)$ gives:

$$\begin{aligned} G(w_0 + Wv(s)) &= G(w(s) - \frac{1}{2}N_0 s^2 + O(|s|^3)) \\ &= G(w(s)) - \frac{1}{2}J_G(w(s))N_0 s^2 + O(|s|^3) \end{aligned}$$

Assuming standard Caffarelli regularity for the Brenier potential ($\varphi \in C^3$), the Jacobian J_G is locally Lipschitz. Since $w(s) = w_0 + O(|s|)$, we have $J_G(w(s)) = J_G(w(0)) + O(|s|)$. Thus, we can obtain:

$$\gamma(s) - G(w_0 + Wv(s)) = \frac{1}{2}J_G(w_0)N_0 s^2 + O(|s|^3)$$

By definition, $N_0 \in N_{w_0}\mathcal{M}_0$. By the normal-contraction bound in Assumption 1, this gives:

$$\|J_G(w_0)N_0\| \leq \rho\|N_0\| \leq \frac{\rho\kappa_0}{c^2}$$

Therefore, we can obtain:

$$\begin{aligned} \|\gamma(s) - G(w_0 + Wv(s))\| &\leq \frac{\rho\kappa_0}{2c^2}s^2 + O(|s|^3) \\ &= O(\rho\kappa_0 s^2) + O(|s|^3) \end{aligned}$$

This establishes (8).

Step 3: Comparison with the ambient tangent plane. Applying the identical second-fundamental-form expansion directly to γ on the ambient manifold $\mathcal{M}$ gives the standard tangent-plane approximation error: $\|\gamma(s) - (y_0 + Tu(s))\| = O(\kappa s^2) + O(|s|^3)$, which proves (9). Comparing the two bounds, the flow-based latent approximation achieves an asymptotically smaller leading-order error whenever $\rho\kappa_0 \ll \kappa$. □

C Local manifold approximation with alternative affine planes

In this section, we illustrate how different affine planes in latent space can give good approximations to the data manifold $\mathcal{M}$, after passing through the pretrained OT map G_θ. With respect to Figure 6, we compare our fitted latent model (under the same settings as in Appendix F) with two alternate affine planes, obtained via rotating the direction about the latent anchor $w_0 = G_\theta^{-1}(y_0)$. Although the latent affine planes have different orientations, the resulting manifold approximation after G_θ provides a good approximation of the manifold, with the approximation quality deteriorating as the neighborhood increases.

Further insight to this behavior can be seen from the grey segment in the figures. In the data space, each point on the single grey segment has the same nearest point p on the clean manifold. The resulting inverse of this grey segment, which we define as the fibre associated with p, intersects the three affine planes at different points. However, consistent with the normal contraction property discussed in Section 3.3, the differences in the latent points along the fibre are contracted, allowing different affine planes in the latent space to give good approximations to the same local manifold in data-space.

D Likelihood-Free Inference

D.1 Intractability of the Likelihood

To estimate the parameter W, we seek to perform maximum likelihood or maximum a posteriori estimation. However, the likelihood $p(Y \mid W)$ of model (4) is not available in closed form. Marginalizing over the latent variable x yields the integral

$$p(Y \mid W) = \int \underbrace{\mathcal{N}(x \mid 0, I_m)}_{p(x)} \underbrace{\mathcal{N}\big(Y \mid G_\theta(w_0 + Wx),\, \sigma^2\Pi_\perp(x, W)\big)}_{p(Y|x,W)} \,\mathrm{d}x, \tag{15}$$

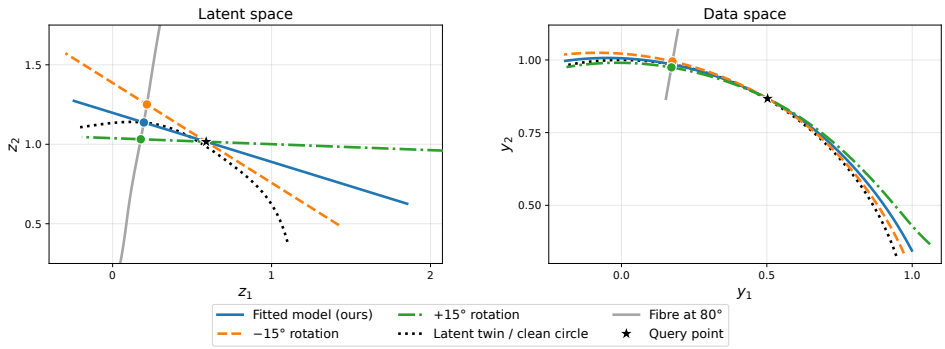


Figure 6: **Alternative affine lines under the same fixed flow.** Left: the fitted line, its $\pm 15°$ rotations, and the latent circle manifold (black dotted). Right: the pushforward fits and the circle manifold. The colored dots select different points on the highlighted fiber and decode to different normal displacements from the same clean point at $80°$. Stars mark the latent anchor and the query point.

where the conditional $p(Y \mid x, W)$ is a degenerate Gaussian whose covariance $\sigma^2 \Pi_\perp(x, W)$ varies with x through the Jacobian of the nonlinear flow G_θ. Since both the mean $G_\theta(w_0 + Wx)$ and the covariance $\Pi_\perp(x, W)$ depend nonlinearly on x, standard Gaussian integration identities do not apply, and the integral has no obvious closed-form expression.

D.2 DSM Training Procedure

First, we sample pairs of $\{W\}$ and for each W, we sample a paired Y from model (4) using the pretrained flow G_θ, to form a paired set $\mathcal{D}_{\text{DSM}} := \{(Y, W)\}$. We then start the training loop: at each iteration, sample a batch from $\mathcal{D}_{\text{DSM}}$, corrupt the parameters as (10), and update ϕ by gradient descent on (11). Crucially, after simulating the training pairs $\mathcal{D}_{\text{DSM}}$, the training procedure does not have to interact with G_θ nor marginalize over x, making the score gradient update highly efficient and entirely integration-free.

D.3 Gradient Ascent for MLE

Given a set of local observations $\{Y_i\}_{i=1}^n$ in a neighborhood of the query point y_0, we estimate the signal direction W by gradient ascent on the log-likelihood. At each step, the posterior score is obtained by querying the trained score network s_ϕ at the current parameter estimate and averaging over the local observations. The likelihood score is then recovered by (13), and the parameters are updated iteratively:

$$W^{(k+1)} = W^{(k)} + \gamma \left[\frac{1}{n} \sum_{i=1}^{n} s_\phi\left(Y_i,\, W^{(k)}\right) + \frac{W^{(k)}}{\sigma^2_{\text{prior}}} \right], \tag{16}$$

where $\gamma > 0$ is the learning rate. Upon convergence, the estimated tangent space at y_0 is read off as $\hat{T} = \text{col}(J_{G_\theta}(w_0) \cdot \widehat{W})$, and the orthonormal tangent basis is obtained via QR decomposition. The full inference procedure is summarized in Algorithm 1.

A key advantage of the amortized approach is the computational cost at inference time. Each gradient step in (16) requires only forward passes through the score network s_ϕ — a small neural network evaluated at n input pairs (Y_i, W). The main expensive operations, i.e., simulating (Y, W) pairs and integrating the flow, are performed once before training the score network. The Jacobian–vector product $J_{G_\theta}(w_0) \cdot \widehat{W}$ in the final step is computed only once at convergence to extract the tangent basis in data space.

E Denoising Score Matching in SBI

Classical neural SBI methods learn the posterior [32, 14] or the likelihood [33] directly via normalizing flows. More recent work targets the *score function* — the gradient of the log-density — rather

Algorithm 1 Local Manifold Identification via Likelihood Free Inference

Input: Trained score network s_ϕ, pretrained flow G_θ, query point y_0, local observations $\{Y_i\}_{i=1}^n$, learning rate γ, prior std σ_{prior}

Output: Estimated tangent basis $\hat{Q} \in \mathbb{R}^{d\times m}$

$w_0 \leftarrow G_\theta^{-1}(y_0)$

Initialize $W^{(0)} \in \mathbb{R}^{d\times m}$ randomly

for $k = 0, 1, 2, \ldots$ until convergence **do**

  $\hat{g}_{\text{post}} \leftarrow \frac{1}{n}\sum_{i=1}^n s_\phi(Y_i, W^{(k)})$ ▷ Average posterior score

  $\hat{g}_{\text{lik}} \leftarrow \hat{g}_{\text{post}} + W^{(k)}/\sigma^2_{\text{prior}}$ ▷ Subtract prior score $\nabla \log p(W)$, optional

  $W^{(k+1)} \leftarrow W^{(k)} + \gamma\, \hat{g}_{\text{lik}}$

end for

$\hat{Q}, _ \leftarrow \text{QR}(J_{G_\theta}(w_0) \cdot \widehat{W})$ ▷ Tangent basis in data space

return $\hat{Q}$

than the density itself. The score can be learned via a least-squares training procedure known as *denoising score matching* (DSM; Vincent [45]). Given samples (θ, x) from the joint distribution, DSM corrupts the parameters as $\tilde{\theta} = \theta + \sigma\eta$ with $\eta \sim \mathcal{N}(0, I)$ and trains a network $s_\phi(x, \tilde{\theta})$ by minimizing

$$\mathcal{L}_{\text{DSM}}(\phi) = \mathbb{E}_{\theta,x,\eta}\left\| s_\phi(x, \tilde{\theta}) + \frac{\eta}{\sigma} \right\|^2. \tag{17}$$

The minimizer is attained at $s_{\phi^*}(x, \tilde{\theta}) = \nabla_{\tilde{\theta}} \log p_\sigma(\tilde{\theta} \mid x)$, the score of the noise-smoothed posterior. As $\sigma \to 0$, this converges to the score of the true posterior $\nabla_\theta \log p(\theta \mid x)$.

F Experimental Settings

F.1 Evaluation Metric on Toy Dataset

To quantitatively assess the geometric fidelity of the estimated local manifolds, we measure the distance from a discretized grid of points on the predicted curves to the true underlying manifold (a circle of radius R in our toy experiments). To ensure a fair comparison, the predicted curves for all methods are explicitly rescaled so that they span the exact same arc length in the ambient data space. We evaluate performance using two key metrics:

- **Mean Reconstruction Error:** The average orthogonal distance between the points on the predicted curve and the true manifold. This metric captures the overall geometric alignment and average fidelity of the local approximation across the evaluated neighborhood.
- **Max Reconstruction Error:** The maximum orthogonal distance from any point on the predicted curve to the true manifold. This metric isolates the worst-case deviation, typically highlighting the extreme linearization error where the predicted curve diverges most sharply from the true manifold at the outer edges of the local neighborhood.

G Additional Experimental Results

In this section, we add some additional figures and tables to supplement our experimental results in Section 6. All experiments are run on the same 1D ring dataset used in Figure 3 with two ambient dimensions.

First, we show an illustration of local PCA vs. our linear model on this dataset in Figure 7. It can be seen that our method traces the true manifold nicely, while the local PCA makes large approximation error at the both ends.

Second, we show how the quality of the pretrained flow affects the recovered manifold. In Figure 8, the leftmost plot shows that as we train the OT-CFM model with more iterations, the manifold reconstruction error reduce significantly. However, after 2000 training epochs, the flow model may have over-fitted, leading to an increased reconstruction error.

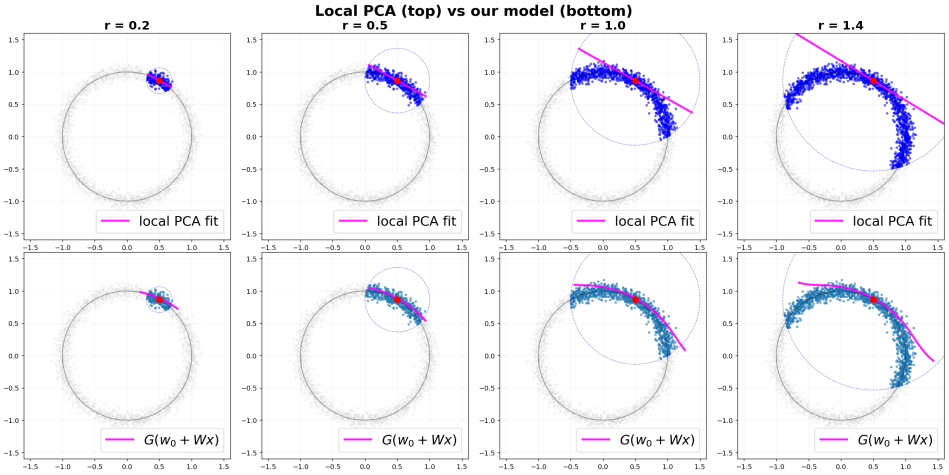


Figure 7: Comparison of Local PCA and our methods on 1D circle dataset. Our method has a significantly lower approximation error comparing to Local PCA (particularly at the two ends of the local arc).

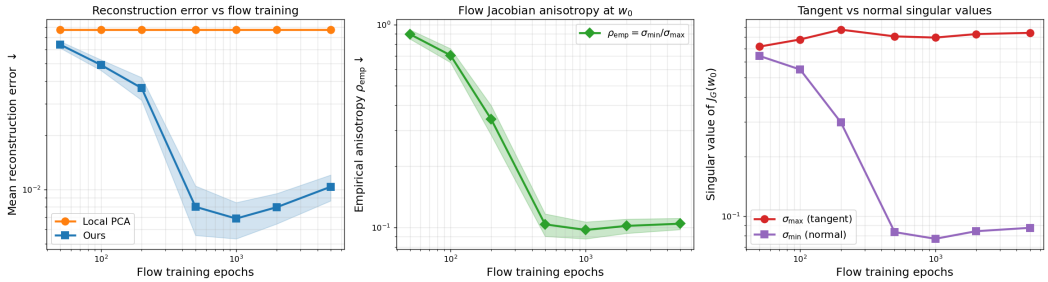


Figure 8: Ablation study, how the quality of the pretrained flow affects the recovered manifold, repeated over 8 query points spread over the manifold and 10 independently trained flows, giving 80 evaluations per configuration. Shaded bands are one standard deviation

The center and the right plot show the spectral gap of tangent vs. normal space of G_θ increase as the training goes on. This is a direct validation of our Assumption 1.

Third, we compare the impact of the data noise on our model. We change the off-manifold noise to 0.05, 0.1 and 0.2 and plot the manifold reconstruction error. It can be seen our method remains the most robust in all three settings with the smallest reconstruction error.

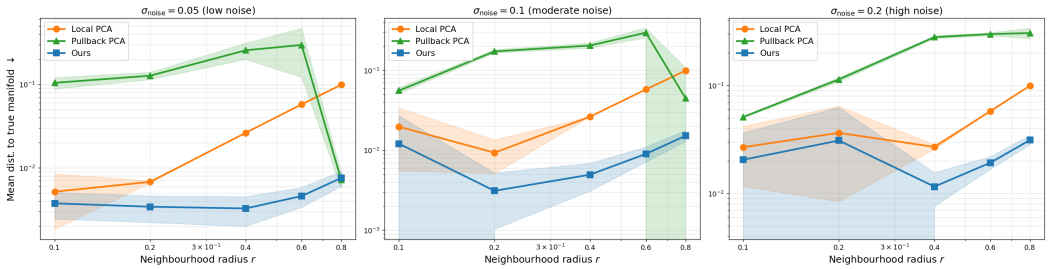


Figure 9: Ablation study, how the dataset noise affects the recovered manifold, repeated over 8 query points spread over the manifold and 10 independently trained flows, giving 80 evaluations per configuration. Shaded bands are one standard deviation

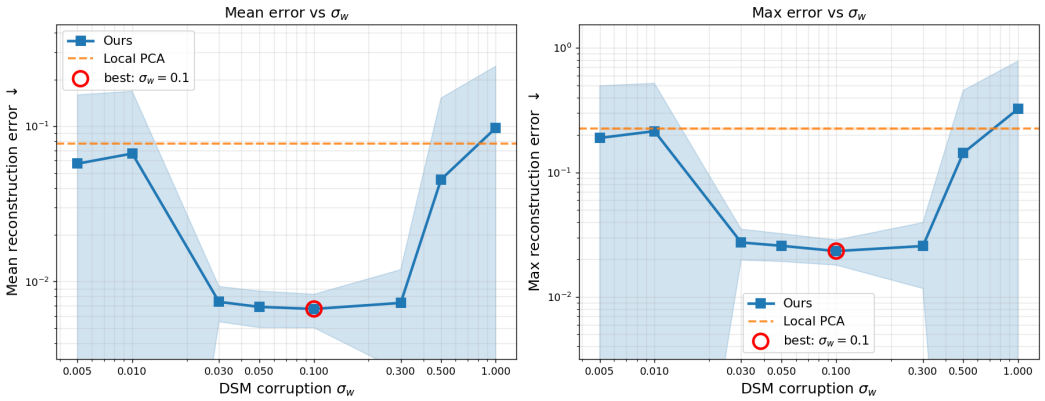


Figure 10: Ablation study, how the corruption level σ_w of the DSM affects the recovered manifold, repeated over 8 query points spread over the manifold and 10 independently trained flows, giving 80 evaluations per configuration. Shaded bands are one standard deviation

Table 1: Angular error in degrees (mean $\pm$ std; lower is better) for varying r, averaged over 12 queries and 10 runs. Best result per row in bold.

r	Local PCA	Pullback PCA	Jacobian-SVD	Ours
0.2	1.8 ± 1.2	74.7 ± 11.2	$\mathbf{0.6 \pm 0.4}$	1.2 ± 0.9
0.4	$\mathbf{0.6 \pm 0.4}$	70.9 ± 13.5	$\mathbf{0.6 \pm 0.4}$	0.7 ± 0.6
0.6	$\mathbf{0.4 \pm 0.3}$	49.2 ± 20.0	0.6 ± 0.4	0.6 ± 0.5
0.8	$\mathbf{0.4 \pm 0.3}$	0.6 ± 0.5	0.6 ± 0.4	0.6 ± 0.5

Finally, we compare the impact of the σ_w, i.e., noise we add on to W when performing denoising score matching. It can be seen that there exists a global sweet spot, in which we attain the lowest reconstruction error. However, even a suboptimal choice, e.g., $\sigma_w = 0.01$, our method still achieves a lower manifold reconstruction error comparing to local PCA.

In addition to the reconstruction error, we report tangent space recovery, using the angle between the estimated and true tangent vector. With reference to Table 1, we see that our method is competitive against both the local PCA and Jacobian-SVD, while the pullback PCA performs poorly for small radius.

In order to provide further understanding of the performance of our method against more modern manifold estimation baselines, in addition to the local and pullback PCA baselines, we also include a local quadratic model, and the method of Stanczuk et al. [41] (which we call Score SVD) and Brehmer and Cranmer [5] (M-flow). With reference to Table 2, we find that our method, along with the local quadratic model, is able to show good reconstruction performance relative to other baseline methods.

G.1 Validation of anisotropic contraction

In this section, we perform additional experiments to test the anisotropic contraction property in Assumption 1, across a variety of different manifold configurations with differing curvatures and intrinsic/ambient dimensions.

At each query point, the Jacobian of the estimated OT map G_θ is evaluated, and the singular values $s_1, \ldots, s_d$ are obtained. The ratio s_{m+1}/s_m serves as a measure of the spectral gap between the leading m directions and the other directions, as well as an empirical counterpart to ρ/c in Assumption 1. Furthermore, since the ground-truth manifold is known, we also directly evaluate tangent-space alignment.

With reference to Table 3, we observe a clear spectral gap in all 10 manifolds, for both constant and non-constant curvature manifolds, and also strong tangent-space alignment.

Table 2: Mean distance to the true manifold (lower is better) for varying r, averaged over 12 queries and 10 runs. Best result per row in bold, second best underlined.

r	Local PCA	Local quadratic	Pullback PCA	Score SVD [41]	M-flow [5]	Ours
0.2	0.0069	<u>0.0056</u>	0.1332	0.0074	0.199	**0.0030**
0.4	0.0263	<u>0.0051</u>	0.2346	0.0266	0.718	**0.0032**
0.6	0.0577	**0.0051**	0.2924	0.0578	1.268	<u>0.0053</u>
0.8	0.0991	**0.0070**	0.0091	0.0992	2.082	<u>0.0089</u>
1.0	0.1491	**0.0134**	<u>0.0138</u>	0.1490	2.974	0.0142
1.2	0.2061	0.0337	**0.0200**	0.2060	3.970	<u>0.0206</u>

Table 3: Empirical validation of Jacobian spectral anisotropy and tangent-space alignment across ten synthetic manifold configurations. m/d denotes intrinsic/ambient dimension, and κ denotes curvature at the query points (const. indicates constant curvature). The spectral-gap ratio s_{m+1}/s_m is reported as mean (maximum) across queries. Min. tangent alignment denotes the minimum cosine alignment between the leading singular subspace and the ground-truth tangent space across query points (1 indicates perfect alignment). † denotes the setting used in the main text.

Setting	m/d	κ	s_{m+1}/s_m	Min. tangent alignment
Ring $S^{1\dagger}$	1/2	const.	0.105 (0.111)	0.9999
Ring S^1	1/5	const.	0.107 (0.116)	0.9998
Sphere S^2	2/3	const.	0.114 (0.119)	0.9999
Sphere S^3	3/4	const.	0.102 (0.104)	0.9999
Clifford torus T^2	2/4	const.	0.144 (0.153)	0.9998
Sphere S^5	5/6	const.	0.139 (0.148)	0.9999
Wavy ring $r = 1 + 0.3\sin 3\theta$	1/2	0.61–4.08	0.213 (0.296)	0.982
Ellipsoid, axes $(1.5, 1.0, 0.6)$	2/3	0.64–4.17	0.173 (0.319)	0.9994
Ellipsoid, axes $(1.5, 1.1, 0.8, 0.6)$	3/4	0.99–4.17	0.162 (0.345)	0.9996
Ellipsoid, axes 0.7–1.4	10/11	1.18–2.71	0.315 (0.492)	0.9986

We also further investigated the anisotropic contraction property on the MNIST dataset, where $d = 784$. Since ground-truth tangent spaces are unavailable, we examine the singular-value spectrum of the estimated OT flow Jacobian. Specifically, the 90%-energy rank r_{90} is defined as the smallest integer k satisfying

$$\frac{\sum_{i=1}^{k} s_i^2}{\sum_{i=1}^{d} s_i^2} \geq 0.9,$$

where $s_1 \geq \cdots \geq s_d$ are the singular values of the Jacobian.

We observe empirically an average r_{90} of 7.4, which is roughly consistent with previously reported estimates of MNIST intrinsic dimension (approximately 7–13) in the literature [34].

This provides further empirical evidence consistent with Assumption 1 on high-dimensional data.

H Real-world dataset experiments

We describe the full experimental setting used in Section 6.2. We first provide the settings that are shared across all three image datasets used (MNIST, Frey Faces, COIL-20), before specifying dataset-specific settings.

OT Flow Pre-training For each dataset we train an OT-CFM model [43] to convergence, and use the frozen model for subsequent local manifold learning. The reference distribution is $P_0 = N(0, I_d)$ in pixel space, and the velocity network $v_\theta(z, t)$ is parameterized by either a ConvNet or U-Net. Training was done with the AdamW optimizer [27] with a fixed step size of 2×10^{-4}.

Table 4: Configuration of pre-trained OT-CFM flows used in all experiments.

Dataset	Image shape	Architecture	Hidden dim.	Blocks	Batch	Epochs	Params
MNIST	$1 \times 28 \times 28$	ConvNet	192	10	128	250	7.2M
Frey Faces	$1 \times 28 \times 28$	U-Net	64	1	128	300	3.0M
COIL-20	$1 \times 128 \times 128$	U-Net	96	2	16	800	11.4M

Evaluation Protocol For each reference/query image, we generate a local neighborhood around the reference image, representing the local observations used for the local manifold estimation. This local observation dataset is split into a train/test split, and the test dataset is chosen such that it is interspersed equally across the local geometry.

Given the local manifold approximation $\hat{\mathcal{M}}$ for each method, we follow the same evaluation pipeline. Firstly, the manifold approximation is rendered on a single, shared coordinate grid in image space. Then, for each held-out, test image, we locate the nearest rendered image point and calculate a reconstruction error, which is finally pooled across all test images.

We evaluate each fitted manifold on a shared coordinate grid $\{\alpha_k\}_{k=1}^{K} \subset \mathbb{R}^m$, based on the pullback PCA scores of the local observations. We then render $c(\alpha) = G_\theta(w_0 + W\alpha)$ (or for local PCA, $c(\alpha) = y_0 + W\alpha$). We then define, for each held-out image y_i, $\hat{k}_i = \arg\min_k \|y_i - c(\alpha_k)\|_2^2$, with the corresponding estimated image $\hat{y}_i = c(\alpha_{\hat{k}_i})$. Our corresponding reconstruction loss is then the total squared distance to the nearest rendered chart point, normalized to the total squared distance to the anchor image, evaluated over the held-out observations.

$$L_{\text{recon}} = \frac{\sum_{i \in \mathcal{T}} \|y_i - \hat{y}_i\|_2^2}{\sum_{i \in \mathcal{T}} \|y_i - y_0\|_2^2}$$

where $\mathcal{T}$ indexes the held-out observations.

All shaded bands represent a 95% confidence interval over the local observations and simulation randomness.

Baseline Methods For the PCA methods, which are done either on the data-space neighborhood points or on the reference space inverted neighborhood points, we obtain the manifold direction W using the first m principal components. For the Jacobian-SVD, we take the leading m singular vector of the Jacobian $J_{G_\theta}(w_0)$.

For our proposed method, note that in the likelihood-free estimation of W, as opposed to the approach discussed in Appendix D.2, which uses a pretrained score network, allowing for amortization over the parameters W, we use a non-amortized approach which estimates a new likelihood-score for each parameter iterate $W^{(k)}$ in Algorithm 1. This approach is proposed as Fisher score matching MLE in Khoo et al. [21]. As opposed to a full neural network score parameterization, we use a linear score model with a handcrafted image feature map. We furthermore constrain W to have orthonormal columns. The MNIST and Frey Faces experiment uses a standard Adam optimizer update following by projection onto this constraint, while the COIL-20 uses Riemannian SGD [4].

All experiments were ran with a Nvidia RTX 4090 GPU, with 24 GB of RAM.

H.1 MNIST dataset

We use the MNIST dataset [1], with the reference image being the digit "3", as seen in the top left image of Figure 11. We calculate the radius to this image using lower-order image moments as a measure of distance. For our proposed method which leverages Fisher score matching [21] to obtain the MLE, we use a proposal distribution for W which is a Gaussian distribution centered at the current iterate, with a normalized proposal scale of 0.5, and with 1024 proposals per iteration. We ran the projected Adam with a total of 40 steps with a learning rate of 1×10^{-2}.

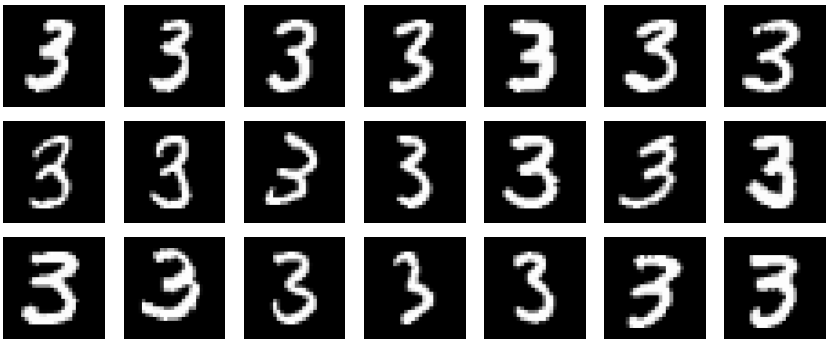


Figure 11: Local observations for MNIST dataset ($r = 0.25$). Reference image is in the top left corner.

As we can see in Figures 12 and 13, our proposed method is able to provide a better approximation to the local manifold, varying smoothly across grid points and staying on-manifold.

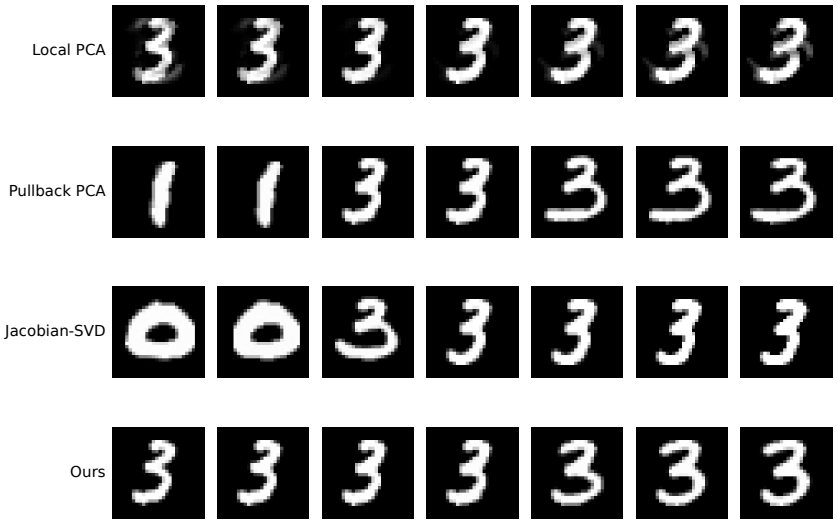


Figure 12: MNIST manifold approximation across grid points.

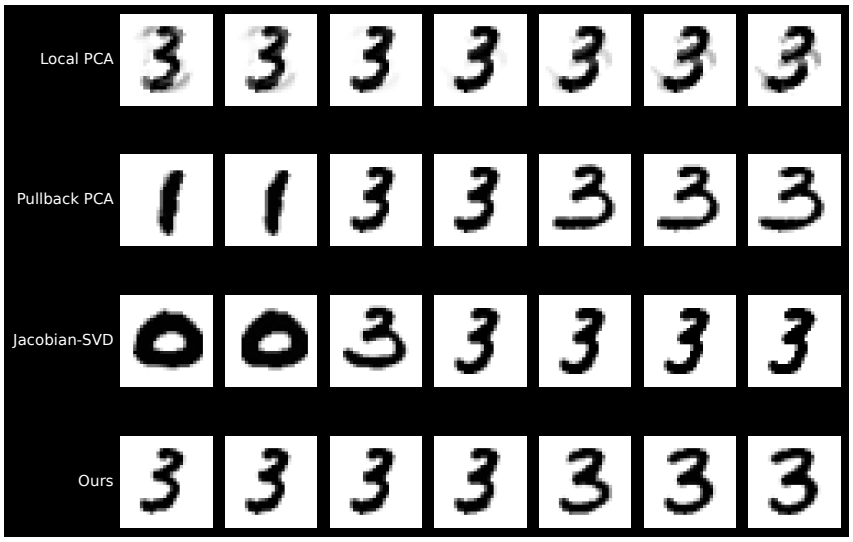


Figure 13: MNIST manifold approximation across grid points. (Inverted colors)

Similarly, Figure 14 reflects that our proposed method is able to provide better reconstruction error over a range of latent dimension and neighborhood radius. This is further reflected in the tangent direction of the various baseline methods in data-space, from Figure 15. As we can see, our proposed method is able to better capture the tangent direction representing the variation in the digits within the local neighborhood.

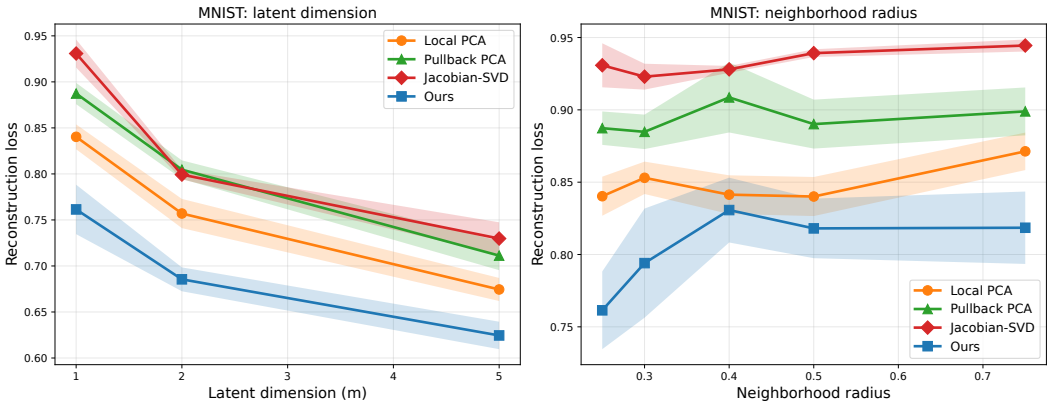


Figure 14: MNIST reconstruction error across latent dimension and neighborhood radius. Shaded bands represent 95% confidence intervals

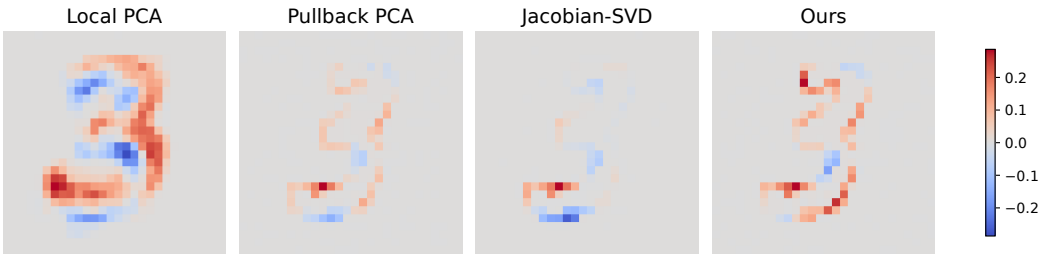


Figure 15: MNIST tangent direction.

In order to investigate the robustness of our evaluation to different choices of grid points, we rescale the grid points used for evaluation using a scaling factor, and provide the ablation in Figure 16. Generally, our proposed method is able to stay competitive and provide better local manifold approximation across a range of chosen grid points.

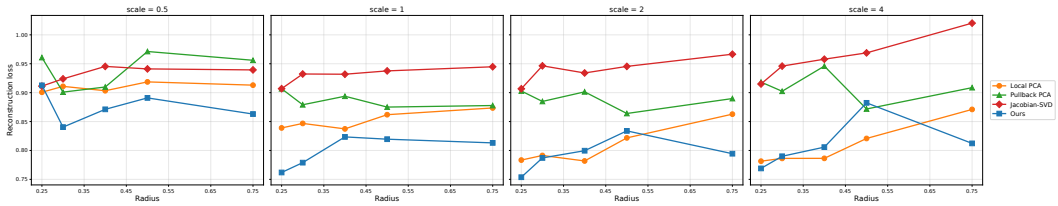


Figure 16: MNIST grid points rescaling.

H.1.1 Rotated MNIST Dataset

To provide a controlled experiment where the ground-truth tangent is known, we constructed a rotated-MNIST benchmark, which has a known one-dimensional tangent. In this setting, we train our OT flow on a collection of rotated versions of a single MNIST image.

Furthermore, we perform an ablation with fixed, data-space isotropic Gaussian corruption $\sigma \in \{0, 0.1, 0.3, 0.5\}$. As opposed to normally projected noise, this is a setting where we have model misspecification, as the noise is not correctly modeled.

We include a comparison with two modern manifold estimation methods: Manifold Moving Least Squares (MMLS) [39], which is a local manifold estimation approach, as well as M-flow [5], which is global.

With respect to Table 6, under the fixed setting with increasing neighborhood size, Jacobian-SVD performs best overall in this controlled single-transformation setting, where the dominant flow direction can directly represent the rotation of the manifold, as opposed to the other image dataset experiments which we have provided, which involve more heterogeneous local variation. Our method is still able to show superior performance compared to other baseline approaches, although MMLS performs better in larger neighborhood sizes. With respect to Table 5, in the purely noiseless setting, local PCA and MMLS give the best performance because of the OT-CFM estimation error. However, for moderate noise levels, our method is able to perform relatively better than local PCA, pullback PCA and MMLS. Across all settings, we find that M-flow has higher mean error than our method in these experiments, which may be due to the method not being specific to local observations.

H.2 Frey Faces dataset

We use the Frey Face dataset [36], which consists of 2000 video frames. Each frame is assigned a temporal index t, and since the video is of an individual's face with different expressions and poses, the temporal index t encodes a smooth manifold. Using a fixed anchor image at t_0, we

Table 5: Tangent-space angular error (degrees; lower is better) versus injected noise level σ, with neighborhood size $k = 256$. Entries are mean $\pm$ sample SD over 3 queries $\times$ 5 seeds; best mean per column in bold.

Method	$\sigma = 0$	$\sigma = 0.1$	$\sigma = 0.3$	$\sigma = 0.5$
Local PCA	0.8 ± 0.2	26.7 ± 0.6	35.2 ± 0.5	88.0 ± 2.1
Pullback PCA	3.1 ± 2.2	80.5 ± 5.5	84.8 ± 4.5	70.6 ± 9.0
MMLS	$\mathbf{0.2 \pm 0.0}$	27.0 ± 0.4	34.8 ± 0.7	61.3 ± 2.4
Jacobian-SVD	11.9 ± 5.7	$\mathbf{8.6 \pm 1.6}$	$\mathbf{16.1 \pm 0.4}$	$\mathbf{54.8 \pm 11.2}$
Ours	5.9 ± 3.6	12.3 ± 2.1	22.0 ± 2.5	64.8 ± 15.6
M-flow	31.1 ± 10.5	40.7 ± 13.8	61.0 ± 16.8	67.2 ± 16.0

Table 6: Tangent-space angular error (degrees; lower is better) versus neighborhood size k, with noise level $\sigma = 0.1$. Setup as in Table 5.

Method	$k = 64$	$k = 128$	$k = 256$	$k = 512$	$k = 1024$
Local PCA	51.5 ± 2.8	39.3 ± 1.6	26.7 ± 0.6	15.8 ± 0.3	8.8 ± 0.1
Pullback PCA	81.6 ± 4.4	80.6 ± 6.3	80.5 ± 5.5	77.7 ± 7.4	74.8 ± 10.8
MMLS	59.8 ± 3.7	41.2 ± 1.6	27.0 ± 0.4	16.0 ± 0.3	$\mathbf{8.7 \pm 0.1}$
Jacobian-SVD	$\mathbf{7.7 \pm 0.8}$	$\mathbf{7.6 \pm 1.0}$	$\mathbf{8.6 \pm 1.6}$	$\mathbf{10.5 \pm 4.2}$	10.8 ± 4.0
Ours	11.5 ± 2.5	11.2 ± 2.4	12.3 ± 2.1	13.4 ± 2.8	12.2 ± 3.3
M-flow	40.7 ± 13.8 (independent of k)				

define a temporal radius r_t around the anchor image, and select the local neighborhood based on $[t_0 - r_t, t_0 + r_t]$. The local observations are shown in Figure 17.

We use a proposal distribution for W which is a Gaussian distribution centered at the current iterate, with a normalized proposal scale of 0.1, and with 2048 proposals per iteration, and ran the projected Adam with a total of 80 steps with a learning rate of 6×10^{-3}.



Figure 17: Local observations for Frey face dataset ($r = 12$)

From Figure 19, we can see that across neighborhood radius and latent dimensions, our proposed method provides a better reconstruction loss compared to baseline methods. From Figure 18, we

can see that the local PCA fails to provide a good representation of the local manifold, and the Jacobian-SVD approach seems to go off-manifold. On the other hand, our proposed method varies smoothly, and matches the variation of the local observation.



Figure 18: Frey face manifold approximation across grid points

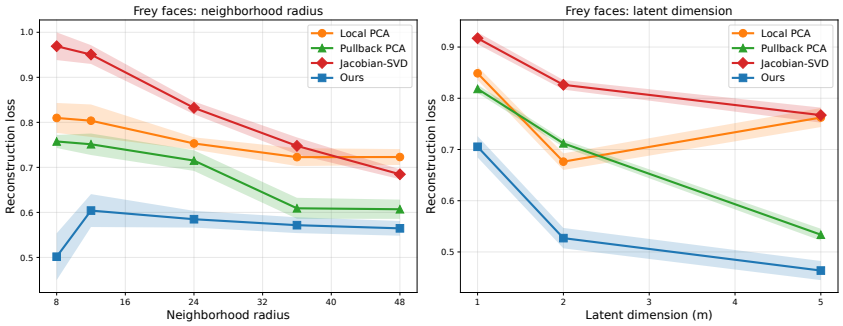


Figure 19: Frey face reconstruction error across latent dimension and neighborhood radius. Shaded bands represent 95% confidence intervals

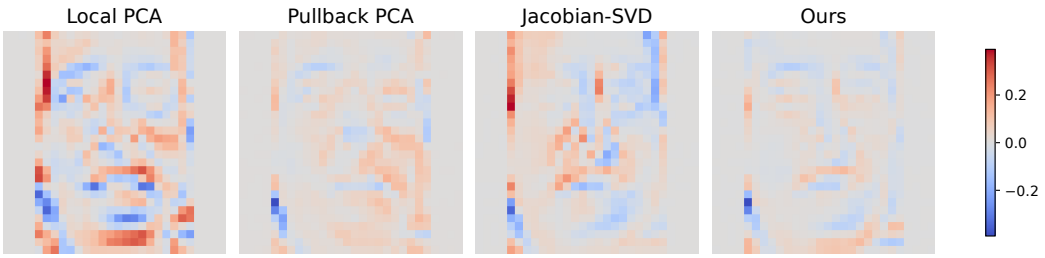


Figure 20: Frey face tangent direction

H.3 COIL-20 dataset

We use the COIL-20 [30], which consists of 1440 images which are constructed from 20 objects horizontally rotated across 72 viewpoints. The angular radius thus serves as a natural measure of locality. The local observations are provided in Figure 21.

We use a proposal distribution for W which is a Gaussian distribution centered at the current iterate, with a normalized proposal scale of 0.05, and with 512 proposals per iteration, and ran the Riemannian SGD with a total of 10 steps with a learning rate of 4×10^{-3}.

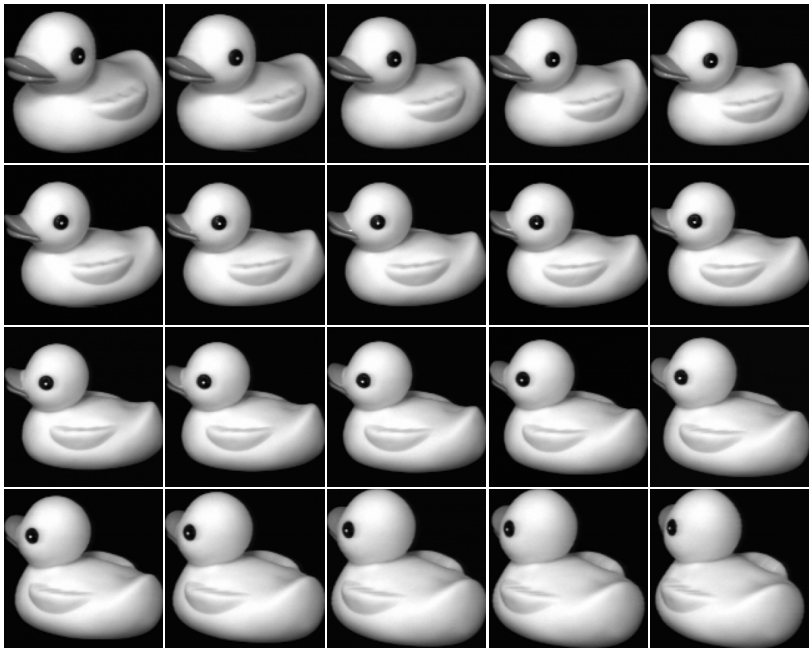


Figure 21: Local observations for COIL-20 dataset

Figures 22 and 24 provide a qualitative assessment of the local manifold approximation of our method, and show that our proposed method is able to obtain a smoother manifold compared to existing baselines. Similarly, this is shown in the quantitative benchmark in Figure 23 as well.

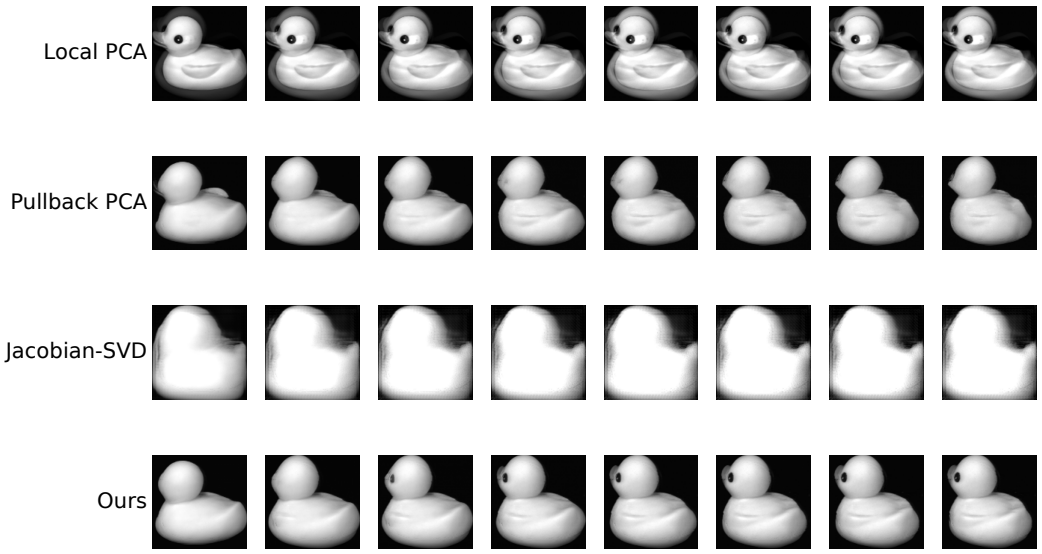


Figure 22: COIL-20 manifold approximation across grid points

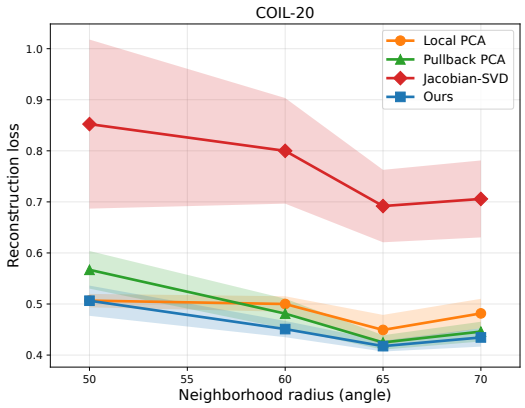


Figure 23: COIL-20 reconstruction error across neighborhood radius. Shaded bands represent 95% confidence intervals

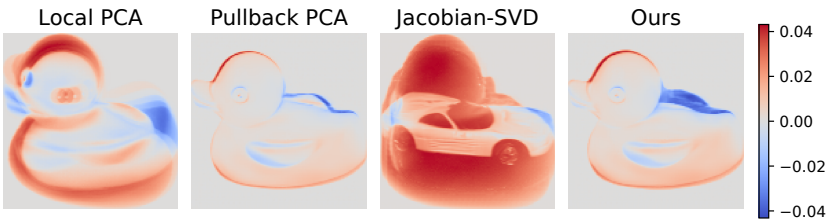


Figure 24: COIL-20 tangent direction

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the conclusion, Section 7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the stand-alone assumption (Assumption 1), as well as proofs in Appendices A and B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide full experimental detail in Section F and the main algorithm in Algorithm 1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide sufficient instructions for running main experiments in the supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide experimental settings in Section F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide standard deviations or confidence interval as error bars in all quantitative plots.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The baseline methods (local PCA and pullback PCA) are not particularly computationally demanding. However, for large scale experiments, we describe our experimental platform in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and our research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The proposed algorithm and research are mostly theoretical in nature.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our research poses no security risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use public datasets and their original papers are cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We do not use crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We do not use crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core development of this method does not involve LLM.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.