Compatible Likelihoods for Flow Matching on Manifolds

Lucas Ng[1]
ln373@cantab.ac.uk

Georgios Batzolis[2]

Mark Girolami[1,2]

Abstract

On the simplest non-Euclidean manifold, the 1D flat torus, we exhibit a distribution where a recent extension of Riemannian flow matching is globally optimized by a velocity field that pushes samples away from the data distribution. We trace this failure mode to a property we call *compatibility*; a likelihood-based flow matching objective is compatible if its population minimizer recovers the marginal velocity field induced by the data and coupling. We derive a general condition for compatibility, apply it to existing methods, and bound the worst-case error by a residual term that vanishes under compatibility. On compact manifolds, we show that existing compatible objectives put density on infeasible velocities. To tackle both issues, we propose an exponential family over endpoints with the conditional velocity as sufficient statistic, which is both compatible and supported only on feasible velocities. In practice, and with low overhead, this reduces to a reweighting of the flow matching loss. On spaces of constant curvature, this treats errors in flow speed and direction differently. Compared to prior art, our method finds the correct velocities on the test case, outperforms in general on high-dimensional product tori, and, for two of four standard geospatial benchmarks on the sphere, achieves substantial NLL improvements. When evaluated on a large dataset of human poses, our method surpasses prior work according to sample-based metrics.

1 Introduction

Modern generative modelling owes much of its success to transport-based models, which learn to deform a simple base distribution into the target data distribution. Among them are diffusion models [Sohl-Dickstein et al., 2015, Ho et al., 2020] and normalizing flows [Tabak and Turner, 2013, Rezende and Mohamed, 2015, Papamakarios et al., 2021]. In particular, conditional flow matching models [Lipman et al., 2023, Albergo and Vanden-Eijnden, 2023, Liu et al., 2023] are effective at generating images [Hu et al., 2024], video [Jin et al., 2025], and heterogeneous tabular data [Guzmán-Cordero et al., 2025]. As flow matching models learn the transport map in terms of a time-dependent vector field, they belong to the class of continuous normalizing flows [Chen et al., 2018]. To overcome the scalability challenges afflicting other models in this class, flow matching decomposes the intractable marginal probability path into simpler conditional paths that couple the base and data distributions. The idea is to learn the marginal velocity field by regression onto many examples of the simple fields generating these simple paths.

Many datasets are inherently non-Euclidean and more naturally modelled as distributions on curved manifolds of nontrivial topology. The dominant generalization of flow matching to manifolds is the Riemannian flow matching (RFM) model of Chen and Lipman [2024]. Recently, Eijkelboom et al. [2024] have observed that the Euclidean flow matching objective is equivalent to maximizing the likelihood of a Gaussian model $q_t^\theta(x_1 \mid x_t)$ of the true endpoint posterior $p_t(x_1 \mid x_t)$. Cast in this light, flow matching generalizes to data supported on arbitrary, non-Euclidean spaces through an

Affiliations. [1] University of Cambridge; [2] Devotion AI Labs

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

Table 1: Comparison of our method against prior work.

Method	Network output	Compatible?	Support-aware on compact manifolds?
RFM	velocity	✓	✗
RG-VFM	endpoint	✗	✓
Ours	velocity	✓	✓

appropriate choice of parametric family q_t^θ. Applying this to Riemannian manifolds, Zaghen et al. [2026] propose Riemannian Gaussian variational flow matching (RG-VFM), which maximizes the likelihood of Pennec [2006]'s Riemannian Gaussian distribution over conditional endpoints.

By delineating two properties, we characterize loss functions for flow matching in terms of their likelihood model, whether that is explicit (as in RG-VFM) or implicit (as in RFM). First, we say a loss function is *compatible* if its population optimum is at the true marginal flow induced by the data and coupling. This is a minimal correctness check for an objective to be of general utility. Second, we say it is *support-aware* if the likelihood model puts mass only on velocities that are achievable. As depicted in table 1, these are independent properties in the sense that RFM is compatible but not support-aware on compact manifolds, while RG-VFM is support-aware but not compatible.

Our theoretical contributions are:

- We derive a general compatibility condition for likelihood-based losses, satisfied by RFM but not RG-VFM. We show this analytically on the simplest non-Euclidean manifold, the flat torus $\mathbb{T}^1$. This general condition is of particular relevance to the manifold setting, but applies to *any* flow matching problem in which the conditional paths are nonlinear.
- We show that the squared error in the local flow scales with the product of distributional fit and a residual term. For a compatible model, the residual is zero, which eliminates this source of error even when the likelihood model is misspecified.
- We view RFM as maximizing likelihood over velocities in the tangent space under a Gaussian model of fixed covariance. This reveals how, despite being compatible, it omits global structure from its training signal. One example is compact manifolds, where the maximum flow speed is limited. Here, the RFM likelihood puts density on infeasible velocities.

This motivates our proposed method, which is both compatible and support-aware. We use an exponential family in endpoint space with the conditional velocity as sufficient statistic. This enforces compatibility regardless of the underlying geometry, inducing a geometrically preconditioned regression loss (with the RFM pipeline otherwise unchanged). Experimentally, we first compare methods on a simple distribution on $\mathbb{T}^1$ where the ideal behaviour of all models is analytic. Faced with rapid spatiotemporal variation in incompatible models, where gradients are uncontrolled except by the network Lipschitz constant, ODE solvers struggle. We then evaluate on high-dimensional tori and four geospatial datasets, following Chen and Lipman [2024] and De Bortoli et al. [2022], as well as on a dataset of human poses in $\mathrm{SO}(3)^{19}$ with nearly 12 million training samples.

2 Preliminaries

Riemannian flow matching (RFM). Chen and Lipman [2024] apply flow matching to general manifolds, as long as we have a suitable notion of distance to hand. If we have the exponential and logarithmic maps, a natural way to construct conditional velocity fields is to follow the minimum-length geodesic at constant speed from a base sample $x_0 \sim p_0$ at $t = 0$ to a data sample $x_1 \sim p_1$ at $t = 1$. Then $u_t^{\text{cond}}(x_t \mid x_1) = \frac{1}{1-t} \log_{x_t}(x_1)$, where x_t is the intermediate point[1] along the geodesic at time t. Together with the coupling, these conditional velocity fields $u_t^{\text{cond}}(x_t \mid x_1)$ uniquely determine conditional probability paths $p_t(x_t \mid x_1)$ via the continuity equation,

$$\partial_t p_t(\cdot \mid x_1) + \operatorname{div}_g \left(p_t(\cdot \mid x_1) u_t^{\text{cond}}(\cdot \mid x_1) \right) = 0,$$

under the boundary conditions $\mathbb{E}_{x_1}[p_t(x_t \mid x_1, t = 0)] = p_0$ and $\mathbb{E}_{x_1}[p_t(x_t \mid x_1, t = 1)] = p_1$. The converse holds modulo gauge freedom in the velocity fields, as different velocity fields can induce

[1] This is computable without clamping the $1 - t$ denominator as it is $\frac{\mathrm{d}}{\mathrm{d}t} x_t$ with $x_t = \exp_{x_0}(t \log_{x_0}(x_1))$. Considering nonuniqueness of $\log_{x_t}(x_1)$ is nonessential as the set of coupled pairs (x_0, x_1) with multiple geodesics has measure zero.

the same probability path. Learning the marginal velocity field produces a transport map from the base distribution into the data distribution. To achieve this, the RFM loss is

$$\mathcal{L}_{\mathrm{RFM}}(\theta) = \mathbb{E}_{t,x_0,x_1} \| v_t^\theta(x_t) - u_t^{\mathrm{cond}}(x_t \mid x_1) \|_g^2.$$

Variational flow matching (VFM). In the Euclidean setting, Eijkelboom et al. [2024] reformulate flow matching as likelihood maximization for a virtual parametric model $q_t^\theta(x_1 \mid x_t)$ of the endpoint posterior $p_t(x_1 \mid x_t)$. The objective is to minimize the KL divergence or, equivalently, to maximize a likelihood in endpoint space,

$$\mathcal{L}_{\mathrm{VFM}}(\theta) = \mathbb{E}_{t,x_t}\left[\mathrm{KL}(p_t(x_1 \mid x_t) \| q_t^\theta(x_1 \mid x_t))\right] = -\,\mathbb{E}_{t,x_0,x_1}[\log q_t^\theta(x_1 \mid x_t)] + \text{const.}$$

This gives a unified framework in which, by choosing an appropriate likelihood model $q_t^\theta(x_1 \mid x_t)$, flow matching generalizes to arbitrary data supports like the space of graphs. Eijkelboom et al. [2024] show further that the squared norm loss of Euclidean flow matching corresponds to implicitly modelling the endpoint posterior as a Gaussian of fixed covariance.

Riemannian Gaussian variational flow matching (RG-VFM). Zaghen et al. [2026] extend this endpoint-likelihood view to manifolds. They choose the family of virtual models to be the Riemannian Gaussians of Pennec [2006] with mean parameter μ_t^θ and constant variance $\sigma^2 = 1$,

$$q_t^\theta(x_1 \mid x_t) \propto \exp\left(-\frac{1}{2\sigma^2}\,\mathrm{dist}_g(x_1, \mu_t^\theta(x_t))^2\right).$$

By restricting to homogeneous manifolds, where all points are geometrically equivalent under the symmetries of the manifold, Zaghen et al. [2026] propose the RG-VFM loss,

$$\mathcal{L}_{\text{RG-VFM}}(\theta) = \mathbb{E}_{t,x_0,x_1}\left[\mathrm{dist}_g(x_1, \mu_t^\theta(x_t))^2\right].$$

This exploits the fact that the partition function of the Riemannian Gaussian on a homogeneous manifold is independent of $\mu_t^\theta(x_t)$. Hence, it can be dropped from the loss without affecting the optimization. Because μ_t^θ is an endpoint, RG-VFM predicts the velocity as $v_t^\theta(x_t) = \frac{1}{1-t}\log_{x_t}(\mu_t^\theta(x_t))$.

3 Compatibility of training objectives

We call a training objective *compatible* if every population minimizer, after applying the prescribed decoder from network output to velocity, produces the true marginal velocity,

$$u_t(x_t) = \mathbb{E}_{p_t(x_1 \mid x_t)}\left[u_t^{\mathrm{cond}}(x_t \mid x_1)\right]. \tag{1}$$

This is the minimal requirement for correctness as it guarantees that transport along integral curves of $u_t(x_t)$ transforms the base distribution p_0 into the data p_1. Statistical decision theory would see this as an application of *Fisher consistency*; we use an alternative name here to avoid confusion with Consistency Models [Song et al., 2023, Cheng et al., 2025b]. Since $\arg\min_v \mathbb{E}\|v - U\|_g^2 = \mathbb{E}[U]$, RFM satisfies this. By contrast, the RG-VFM objective has global minima at any and all Fréchet means of $p_t(x_1 \mid x_t)$. Converting such an endpoint on the manifold into a velocity tangent vector generally does not produce the true $u_t(x_t)$, except in Euclidean space.

The likelihood perspective of Eijkelboom et al. [2024] is gaining traction, proliferating to a diversity of settings. As well as Zaghen et al. [2026], the VFM lineage now includes the work of Guzmán-Cordero et al. [2025], Matişan et al. [2026], and Carrasco-Pollo et al. [2026]. Alongside RG-VFM, future work is at risk of designing new losses for flow matching which *seem* natural, without checking for compatibility. Without an explicit check, we risk building on incompatible models. Possible mechanisms are that (1) the parametrization makes this analytically non-obvious, or (2) the added level of indirection obscures the connection to the careful construction of conditional flow matching.

In this section, we seek a general framework for compatibility of flow matching losses, both those that exist and those that *will* exist. Under the likelihood lens, we study the behaviour of these losses in the manifold setting. As our primary interest is the performance of such a likelihood model at best fit, we proceed under the idealized assumption of infinite data and network capacity. This allows us to reason beyond a binary condition of YES or NO and quantify how far from the true marginal flow a likelihood model predicts in the worst case, even under optimal training.

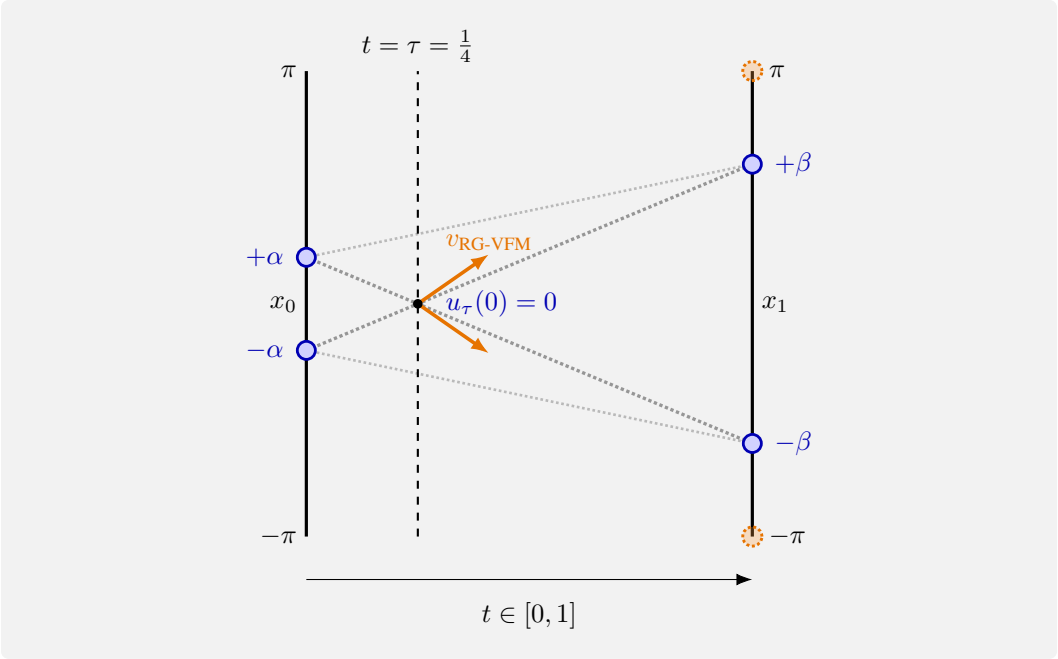


Figure 1: Example on $\mathbb{T}^1$ with $\tau = \frac{1}{4}$, $\alpha = \frac{\pi}{5}$, $\beta = \frac{3\pi}{5}$. Dotted lines show conditional paths; their tangents are conditional velocities. Two cross-pairs meet at $(x, t) = (0, \tau)$, where the true marginal velocity is zero. We show both possible branches of the optimal RG-VFM prediction in orange.

3.1 A motivating example: incompatibility on a manifold without curvature

Zaghen et al. [2026, §4.2] write that RG-VFM implicitly captures the full geometric structure of the manifold while RFM lacks curvature information, resulting in weaker, less precise supervision. Therefore, we begin by comparing both methods on the flat torus $\mathbb{T}^1 \cong \mathbb{R}/2\pi\mathbb{Z}$ where the geometric structure is both conceptually simple and without curvature. Consider the base and data distributions

$$\begin{aligned} p_0 &= \tfrac{1}{2}\delta_{-\alpha} + \tfrac{1}{2}\delta_{+\alpha}, \\ p_1 &= \tfrac{1}{2}\delta_{-\beta} + \tfrac{1}{2}\delta_{+\beta} \end{aligned} \tag{2}$$

under independent coupling $p_0 \otimes p_1$. With some $\tau \in (0, \frac{1}{2})$, let $\beta \in (\frac{\pi}{2}, \pi(1-\tau))$ and $\alpha = \frac{\tau\beta}{1-\tau}$. As illustrated in fig. 1, the two cross-pair geodesics then meet at $x = 0$ at time $t = \tau$, carrying equal and opposite velocities. At $(x, t) = (0, \tau)$, their conditional expectation (the true marginal velocity) is zero. At the same $(x, t) = (0, \tau)$, RG-VFM minimizes the Fréchet function,

$$\mathcal{L}[\mu_\tau^\theta(0)] = \frac{1}{2}\,\mathrm{dist}_g(-\beta, \mu_\tau^\theta(0))^2 + \frac{1}{2}\,\mathrm{dist}_g(\beta, \mu_\tau^\theta(0))^2. \tag{3}$$

This has local minima at $\mu_\tau^\theta(0) = 0$ and $\mu_\tau^\theta(0) = \pi$; the loss is locally quadratic around each. Because $\beta > \pi/2$, the shorter arc between $x_1 = -\beta$ and $x_1 = \beta$ is the one through π. Under the RG-VFM loss, the unique and globally optimal endpoint prediction is $\mu_\tau^\theta(0) = \pi$. Decoding this endpoint prediction into a velocity prediction at $(0, \tau)$ gives a velocity of magnitude $\frac{\pi}{1-\tau}$. This is substantially different to the zero velocity we expect. Hence, despite the endpoint prediction being globally optimal under the loss, we have incompatible flow on a flat manifold.

Although this construction is intuitive and analytically tractable, a natural question is whether this behaviour is a consequence of the singular distributions and not the objective itself; we analyse a smoothed version in appendix A. Next, we formalize compatibility for all flow matching objectives viewed as likelihood maximization, agnostic of the base, coupling, or data.

3.2 The condition for compatibility of likelihood-based objectives

Think of training as fitting a model distribution $q^\theta(z) = q_t^\theta(z \mid x)$ at each (x, t) to a data posterior $p(z) = p_t(z \mid x)$, where z is whichever variable carries the likelihood, for example an endpoint x_1,

a conditional velocity u_t^{cond}, or something else.[2] The marginal velocity field is the expectation under p of a feature $\psi_{x,t}(z)$ mapping z to the space of conditional velocities, $u_t(x) = \mathbb{E}_p[\psi_{x,t}(z)]$. For a model in endpoint space (like RG-VFM), $\psi_{x,t}(z) = u_t^{\text{cond}}(x \mid z)$, while for a model in velocity space (like RFM), $\psi_{x,t}(z) = z$. The question is whether likelihood optimization can distinguish between data posteriors that differ in $u_t(x)$.

Suppose $\theta = \theta^*$ is optimal for p under the loss $\mathcal{L}_p(\theta) = \mathbb{E}_p[-\log q^\theta(z)]$. We wish to know if this optimum could give an incorrect velocity prediction. Consider an alternative $p' = p + h \geq 0$ for some (signed, square-integrable, mean-zero) h. Then

$$\nabla_\theta \mathcal{L}_{p+h}(\theta^*) = \nabla_\theta \mathcal{L}_p(\theta^*) + \left[\nabla_\theta \int -\log q^\theta(z) h(z)\,\mathrm{d}z\right]_{\theta=\theta^*} = -\int s^*(z) h(z)\,\mathrm{d}z, \tag{4}$$

where $s^*(z) = \nabla_\theta \log q^\theta(z)|_{\theta=\theta^*}$ is the model score at optimum. This perturbation h is invisible to the likelihood at first order if this integral vanishes. Meanwhile, it shifts the marginal velocity by

$$\mathbb{E}_{p+h}[\psi_{x,t}(z)] - \mathbb{E}_p[\psi_{x,t}(z)] = \int \psi_{x,t}(z)\, h(z)\,\mathrm{d}z. \tag{5}$$

For a correct prediction, we need both eq. (4) and eq. (5) to agree that a perturbation actually occurs. That is, the following logical implication must hold for all h:

$$\int s^*(z)\, h(z)\,\mathrm{d}z = 0 \quad \Longrightarrow \quad \int \psi_{x,t}(z)\, h(z)\,\mathrm{d}z = 0. \tag{6}$$

Equivalently, this is a check for the existence of a z-independent linear map $A_{x,t}^\theta \colon \mathbb{R}^m \to T_x\mathcal{M}$ and vector $b_{x,t}^\theta \in T_x\mathcal{M}$ such that we have (almost everywhere),

$$\psi_{x,t}(z) = A_{x,t}^\theta s^\theta(z) + b_{x,t}^\theta, \qquad \theta = \theta^*. \tag{7}$$

Since θ^* is unknown before training, a practical objective must satisfy eq. (7) at every θ in the model class, not just at the optimum. For RG-VFM, whose network output is $\mu = \mu_t^\theta(x_t)$, the target feature is $\psi_{x,t}(z) = \frac{1}{1-t}\log_x(z)$, and the score with respect to the network output is $s(z) = \frac{1}{\sigma^2}\log_\mu(z)$. As there is no global affine relationship between $\log_\mu(z)$ and $\log_x(z)$ on non-Euclidean manifolds, RG-VFM does not satisfy compatibility, obstructed by both curvature and topology.

3.3 Measuring the flow error

A natural way to quantify the departure from compatibility is via the best approximation at the likelihood optimum, so as to satisfy eq. (7) as much as possible. Specifically, we solve the least-squares problem

$$(A_{x,t}^*,\, b_{x,t}^*) \in \arg\min_{A,b} \mathbb{E}_{q^{\theta^*}} \|\psi_{x,t}(z) - As^*(z) - b\|_g^2, \tag{8}$$

denoting the residual term as $r_{x,t}(z) = \psi_{x,t}(z) - A_{x,t}^* s^*(z) - b_{x,t}^*$. We obtain the following bound, which is locally sharp in the sense of Cauchy–Schwarz:[3]

Theorem 1. *Assume measurable $\psi_{x,t} \in L^2(q^{\theta^*}; T_x\mathcal{M})$, $s^* \in L^2(q^{\theta^*}; \mathbb{R}^m)$. Let p be absolutely continuous with respect to q^{θ^*}. Assume the first-order optimality condition $\mathbb{E}_p[s^*(z)] = 0$ (requiring θ^* to be interior) and enough regularity for the score identity $\mathbb{E}_{q^{\theta^*}}[s^*(z)] = 0$. Then*

$$\left\|\mathbb{E}_p[\psi_{x,t}(z)] - \mathbb{E}_{q^{\theta^*}}[\psi_{x,t}(z)]\right\|_g \leq \sqrt{\chi^2(p\|q^{\theta^*})}\sqrt{\left\|\mathbb{E}_{q^{\theta^*}}[r_{x,t} \otimes r_{x,t}]\right\|_{\text{op},g}}, \tag{9}$$

where $\chi^2(p\|q^{\theta^*}) < \infty$ is the chi-squared divergence between p and q^{θ^*}.

Proof. See appendix A.1. □

[2] Any likelihood over endpoints can be reparameterized as a likelihood over conditional velocities via the map $x_1 \mapsto u_t^{\text{cond}}(x_t \mid x_1)$, but not vice versa; we discuss the implications of this further in section 3.4.

[3] The bound is locally sharp with χ^2. All f-divergences with $f''(1) > 0$ are locally equivalent so KL has the same rate.

This elucidates why strong performance does not need the likelihood to be an accurate approximation to the density. A compatible likelihood model has zero residual $r_{x,t}$ and zero flow error at the population optimum, regardless of the residual distributional misfit $\chi^2(p\|q^{\theta^*})$. This gives a proof of the compatibility condition alternative to the exposition in section 3.2. Adding more degrees of freedom to the likelihood model, aiming for better density estimation, seems a natural way to reduce flow error. However, if the model is incompatible, the residual $r_{x,t}$ remains nonzero. Unconstrained by first-order conditions, a flow error generically persists. That is, we should favour likelihood families that possess compatibility over those that are more expressive but do not.

Moreover, the velocity implied by the likelihood model, $\mathbb{E}_{q^{\theta^*}}[\psi_{x,t}(z)]$, is not necessarily the prediction the model uses in practice, $v_t^\theta(x_t)$. To avoid computing the expectation online, methods like RG-VFM use a decoder. This takes the parameters of the likelihood at the network head and maps them to a velocity prediction. Via the triangle inequality, this mismatch between the decoder and the likelihood moments leads to an additional error term:

$$\|v_t^\theta(x_t) - u_t(x_t)\|_g = \|v_t^\theta(x_t) - \mathbb{E}_{q^{\theta^*}}[\psi_{x,t}] + \mathbb{E}_{q^{\theta^*}}[\psi_{x,t}] - u_t(x_t)\|_g \tag{10}$$

$$\leq \underbrace{\|v_t^\theta - \mathbb{E}_{q^{\theta^*}}[\psi_{x,t}]\|_g}_{\text{likelihood-decoder mismatch}} + \underbrace{\sqrt{\chi^2(p\|q^{\theta^*})}\sqrt{\left\|\mathbb{E}_{q^{\theta^*}}[r_{x,t} \otimes r_{x,t}]\right\|_{\mathrm{op},g}}}_{\text{likelihood misfit } \times \text{ incompatibility bound}}. \tag{11}$$

This reiterates that, in settings with nonlinear conditional paths, like on a manifold, a family of likelihoods needs more than geometric naturalness. In addition to compatibility, we would like to parametrize the likelihood family without this decoder error.

3.4 Riemannian flow matching does likelihood maximization implicitly

Here, we express the RFM loss in maximum-likelihood form in velocity space as

$$\mathcal{L}_{\mathrm{RFM}}(\theta) = \mathbb{E}_{t,x_0,x_1}\left[-\log q_t^\theta(u_t^{\mathrm{cond}} \mid x_t)\right] + \mathrm{const}, \tag{12}$$

where $q_t^\theta(u_t^{\mathrm{cond}} \mid x_t)$ is the Gaussian $\mathcal{N}(u_t^{\mathrm{cond}}; v_t^\theta(x_t), \frac{1}{2}g^{-1})$ over the full tangent space. The score with respect to the network output is $s(u) = 2(u - v_t^\theta)$, which is affine in the feature $\psi_{x,t}(u) = u$. Hence, RFM satisfies the compatibility condition. However, this Gaussian is not support-aware on compact manifolds. On a compact manifold $\mathcal{M}$, conditional velocities are bounded as

$$\|u_t^{\mathrm{cond}}(x_t \mid x_1)\|_g \leq \frac{\mathrm{diameter}(\mathcal{M})}{1-t}. \tag{13}$$

The map $x_1 \mapsto u_t^{\mathrm{cond}}(x_t \mid x_1)$ is not surjective. That is, many tangent vectors do not correspond to any endpoint on the manifold. Despite this, the Gaussian puts density on velocities outside the feasible set, regardless of the training distributions. Two adjacent velocities, straddling the boundary of the feasible set so that one is feasible and the other is not, have near equal loss. This discards prior knowledge of the global manifold structure and suggests an improvement to the method: using this information to improve training efficiency and optima achieved.

4 Method

These theoretical developments point to a likelihood family whose score is affine in the conditional velocity. In addition, we want to support the likelihood on manifold endpoints only, so as to avoid inefficient density assignment. As in Zaghen et al. [2026], defining the likelihood in endpoint space (at least initially) is a natural way to achieve this. Keeping velocity prediction $v_t^\theta(x)\colon \mathcal{M} \to T_x\mathcal{M}$ as the network output, we propose an exponential family with conditional velocity as the sufficient statistic,

$$q_t^\theta(x_1 \mid x_t) \propto c_{x,t}(x_1)\exp\left(\langle \eta, u_t^{\mathrm{cond}}(x_t \mid x_1)\rangle_g\right). \tag{14}$$

This is an exponential family with natural parameter $\eta \in T_{x_t}\mathcal{M}$, carrier $c_{x,t}(x_1)\colon \mathcal{M} \to \mathbb{R}^+$. This uniquely maximizes entropy relative to $c_{x,t}\,\mathrm{dVol}_g(x_1)$ amongst endpoint distributions of a given mean velocity. Per the standard property of exponential families, let the mean map be

$$m_{x,t}(\eta) = \nabla_\eta A_{x,t}(\eta) = \mathbb{E}_{q_t^\theta}[u_t^{\mathrm{cond}}(x_t \mid x_1)], \tag{15}$$

where $A_{x,t}(\eta)$ is the log-partition function,

$$A_{x,t}(\eta) = \log \int_{\mathcal{M}} c_{x,t}(x_1) \exp(\langle \eta, u_t^{\text{cond}}(x_t \mid x_1) \rangle_g) \, \mathrm{dVol}_g(x_1). \tag{16}$$

To tie the family to the network, we set $\eta = m_{x,t}^{-1}(v_t^\theta(x_t))$. This makes the prediction of the model match the expectation of the likelihood. We discuss the domain of this in appendix C. The score with respect to η is $u_t^{\text{cond}}(x_t \mid x_1) - v_t^\theta(x_t)$ and since the target feature is $\psi_{x,t}(x_1) = u_t^{\text{cond}}(x_t \mid x_1)$, the compatibility condition of eq. (7) is satisfied. Hence, the model fulfils both key desiderata.

Training loss. Differentiating the negative log-likelihood $\ell_{\text{NLL}}(\theta)$ with respect to θ gives

$$\nabla_\theta \ell_{\text{NLL}}(\theta) = (\nabla_\theta v_t^\theta(x_t))^\top V_{x,t}(v_t^\theta(x_t))^{-1}(v_t^\theta(x_t) - u_t^{\text{cond}}(x_t \mid x_1)), \tag{17}$$

where $V_{x,t}(v) = \nabla_\eta^2 A_{x,t}(\eta)\big|_{\eta = m_{x,t}^{-1}(v)}$ is the covariance of u_t^{cond} at the natural parameter η matching the predicted velocity v. This is the *Fisher information* of the model. Detaching $V_{x,t}$ from the backward pass yields a surrogate loss with gradients identical to those of the likelihood objective,[4]

$$\mathcal{L}(\theta) = \frac{1}{2} \mathbb{E}_{t,x_0,x_1} \left[\langle v_t^\theta(x_t) - u_t^{\text{cond}}, \, \mathrm{sg}(V_{x,t}(v_t^\theta(x_t)))^{-1}(v_t^\theta(x_t) - u_t^{\text{cond}}) \rangle_g \right], \tag{18}$$

where $\mathrm{sg}(\cdot)$ is the stop-gradient operator. Our method naturally reduces to a reweighting of the RFM objective, and crucially without opening up the entries of the preconditioning matrix as a new design space. In other settings, using the inverse Fisher information to adapt dynamics to the intrinsic geometry is an idea that has seen performance gains [Girolami and Calderhead, 2011].

Choice of carrier. Although our construction addresses both key desiderata, the choice of carrier $c_{x,t}$ is apparently open. However, this design axis mostly collapses under several structural constraints. We seek a centred, isotropic carrier that factorizes on product manifolds, localizes around x_t as $t \to 1$, and recovers ordinary flow matching in Euclidean space. Among radial carriers, product factorization imposes a Gaussian form; terminal locality sets the decay; and matching the Euclidean covariance sets its scale (appendix A). This leads to the following choice, reminiscent of the Riemannian Gaussian in Zaghen et al. [2026]:

$$c_{x,t}(x_1) = \exp\left(-\frac{1}{2}\frac{\mathrm{dist}_g(x_t, x_1)^2}{(1-t)^2}\right) = \exp\left(-\frac{1}{2}\|u_t^{\text{cond}}(x_t \mid x_1)\|_g^2\right). \tag{19}$$

This is also normalizable on common noncompact manifolds, including hyperbolic space, which requires the carrier to have superlinear decay. As $t \to 1$, the loss tends to that of RFM. In appendix B, we study sensitivity to the carrier. The resulting velocity-space likelihood is Gaussian in Euclidean space and a truncated Gaussian on flat tori; on spheres, the loss upweights errors in speed relative to those in direction (fig. 7). This reflects the intuition that geodesic convergence on manifolds of positive curvature means errors in magnitude are more significant than those in direction.

Practical computation. At first, it seems we must compute the mean map $m_{x,t}$ and covariance $V_{x,t}$ per training step, online and via numerical differentiation. The insight is that the more symmetries a manifold has, the fewer scalar functions $V_{x,t}$ depends on. By differentiating under the integral sign, we get a formula for each scalar function. On many manifolds, we pretabulate these via quadrature in low dimensions, often just one. Different datasets on a given manifold can reuse the same lookup table. On the flat torus $\mathbb{T}^d$, the covariance $V_{x,t}$ is diagonal with each entry an independent function of η_i. All factors share a single 1D lookup table. On spaces of constant curvature, isotropy cuts $V_{x,t}$ down to two scalar functions of the magnitude. On product manifolds, $V_{x,t}$ is block diagonal with one block per factor. This lets us compose tables and highlights that on many manifolds, the table dimension stays low as the manifold dimension grows. Queries to the lookup tables take $\mathcal{O}(1)$ time. In appendix F, we show that, when compiled, the overhead per iteration is negligible.

5 Experiments

In this section, we evaluate our theoretical predictions and method across four settings. We defer implementation details to appendix C and evaluation and training protocols to appendix E.

[4] $V_{x,t}$ is self-adjoint under g.

Table 2: $\mathbb{T}^1$ analytic test case. Mean±std across seeds for distributional metrics. 20 000 samples per seed. Median $[Q_1, Q_3]$ for NFE and wall time of dopri5 with tolerance 10^{-5}. RG-VFM seeds split by $\|v_\tau^\theta(0)\|_g$ into two clusters (see fig. 2): 'lower' (< 0.1) and 'upper' (> 4.0).

Method (# seeds)	**Error** ↓ $\|v_\tau^\theta(0)\|_g$	W_1 ↓	W_2 ↓	**Energy dist.** ↓ via $\sqrt{\text{dist}_g}$	**MMD²** ↓ Laplace	**NFE** ↓ dopri5	**Time** ↓ dopri5, (s)
RFM (30)	0.070±0.040	0.45±0.20	0.88±0.30	0.214±0.064	0.055±0.028	497 [426, 578]	1.92 [1.65, 2.22]
RG-VFM (30)	2.080±2.078	0.54±0.29	0.83±0.32	0.311±0.191	0.085±0.079	36 839 [33 796, 60 574]	141.4 [130.3, 233.9]
upper cluster (15)	4.123±0.042	0.65±0.30	1.00±0.31	0.370±0.212	0.107±0.089	55 436 [32 228, 70 115]	212.9 [124.0, 271.9]
lower cluster (15)	**0.037**±0.019	**0.42**±0.24	**0.66**±0.25	0.251±0.151	0.063±0.063	36 638 [35 411, 41 675]	140.4 [136.8, 161.3]
Our method (30)	0.060±0.041	0.43±0.18	0.84±0.31	**0.212**±0.055	**0.054**±0.024	**443** [388, 582]	**1.71** [1.49, 2.24]

Analytic test case. We test the predictions of section 3.1 under the parameters illustrated in fig. 1 with 30 seeds. Figure 2 shows the velocity error at $(x, t) = (0, \frac{1}{4})$. Both RFM and our method achieve velocities close to the ideal value of zero, while RG-VFM is bimodal and splits into two clusters of 15 seeds each. As the magnitude is bounded per section 3.4, the upper cluster has its centre just below the value of $\frac{4\pi}{3}$ from section 3.1. The lower cluster corresponds to the population local minimum of RG-VFM at zero velocity, which is a stable attractor for some seeds but not globally optimal.

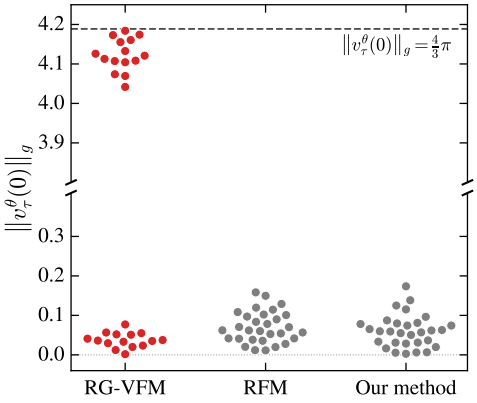


Figure 2: Swarmplot of velocity errors $\|v^\theta_{t=\frac{1}{4}}(0)\|_g$, 30 seeds. The upper RG-VFM cluster has better training loss but poorer generative performance; see table 2.

Table 2 shows distributional metrics and, for dopri5, NFE and wall times. Ranking by W_1 agrees with ranking by velocity error at $(0, \frac{1}{4})$, supporting the idea that learning the marginal velocity is important for generative performance. RG-VFM seeds in the upper cluster are better by training loss than those in the lower cluster, but perform worse on generative metrics and require more ODE steps. Ranking by kernel methods favours our method and RFM over both RG-VFM clusters. A potential cause of this discrepancy is that optimal transport is more sensitive to the ratio of samples generated near to each Dirac and more affected by the $(0, \frac{1}{4})$ flow branch. RG-VFM has the highest inference NFEs and wall times by far. This is consistent with the rapid spatiotemporal variation we observe in the learned RG-VFM field in the absence of compatibility (see also fig. 4).

High-dimensional product tori. Following the setup of De Bortoli et al. [2022] and Chen and Lipman [2024], we evaluate on product tori $\mathbb{T}^d$ with uniform base and wrapped Gaussian data distribution of scale $\sigma = 0.2$. This setup is of particular interest because the scale $0.2 \ll 2\pi$ is so small that the optimal NLL of a continuous normalizing flow is analytic. Recall that $\mathbb{E}_{p_1}[-\log p_1^\theta(x_1)] = \mathcal{H}[p_1] + \mathrm{KL}(p_1 \parallel p_1^\theta)$ so the differential entropy of the true data gives a lower bound. As the density is isotropic, the per-dimension NLL lower bound is $\mathcal{H}[p_1]/d \approx \frac{1}{2}\log(2\pi e\sigma^2) \approx -0.1905$, exact to 4 decimal places for all dimensions here. We see in table 3 that NLL is uninformative for RG-VFM and several values reported by Chen and Lipman [2024] are below the entropic lower bound, which may be due to use of Rademacher vectors to estimate divergence. To mitigate this, we use exact divergence via `jacrev`. In appendix B, we run the checkerboard experiment of Zaghen et al. [2026], testing many solver settings to sanity-check that the NLL results of RG-VFM are not an artefact of our implementation.

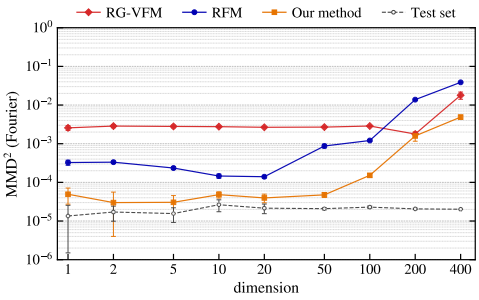


Figure 3: Fourier-kernel MMD² ↓ on flat tori of varying dimension. Mean±std, 5 seeds. 20 000 samples per seed.

In fig. 3, the MMD² gives an interesting comparison. We see that before dimension 200, there is a consistent ordering of methods with our method best, RFM second, and RG-VFM last. Beyond this,

Table 3: High-dimensional flat tori: NLL/dim↓. Mean±std, 5 seeds for our method, RFM (Reproduced), and RG-VFM. 20 000 samples/seed. Target is a wrapped Gaussian ($\sigma = 0.2$). Values shaded grey are below the entropic bound. The best of the *un*-shaded are in **bold**. RFM (Paper results) and Riemannian Score-Based [De Bortoli et al., 2022] values are extracted from fig. 5 of Chen and Lipman [2024] using WebPlotDigitizer [Rohatgi, 2026] with a uniform uncertainty estimate of 0.01.

d	**Our method**	**RFM** Paper results	**RFM** Reproduced	**RG-VFM**	**Riemannian Score-Based**
1	**−0.1904**±0.001	−0.2051±0.010	−0.1715±0.006	−0.2171±0.001	−0.1482±0.010
2	**−0.1900**±0.005	−0.1495±0.010	−0.1658±0.004	−0.2142±0.003	−0.1412±0.010
5	**−0.1879**±0.004	−0.1851±0.010	−0.1681±0.003	−0.2157±0.002	−0.1519±0.010
10	−0.1827±0.004	**−0.1838**±0.010	−0.1678±0.002	−0.2211±0.002	−0.1404±0.010
20	**−0.1871**±0.001	−0.2105±0.010	−0.1673±0.001	−0.2935±0.011	−0.1230±0.010
50	**−0.1847**±0.001	−0.2111±0.010	−0.1687±0.001	−0.3524±0.010	−0.0840±0.010
100	**−0.1758**±0.003	−0.2065±0.010	−0.1553±0.001	−0.4091±0.008	0.0355±0.010
200	−0.1643±0.002	**−0.1701**±0.010	−0.0993±0.001	−0.3368±0.017	0.0745±0.010
400	−0.1374±0.002	**−0.1432**±0.010	−0.0232±0.002	−1.1979±0.040	0.1423±0.010

Table 4: Earth and climate science datasets on the sphere: test NLL↓. Mean±std, 5 seeds.

Dataset size (train + val + test)	**Volcano** 827	**Earthquake** 6 120	**Flood** 4 875	**Fire** 12 809
Riemannian CNF [Mathieu and Nickel, 2020]	−0.97±0.15	0.19±0.04	0.90±0.03	−0.66±0.05
Moser Flow [Rozen et al., 2021]	−2.02±0.42	−0.09±0.02	0.62±0.04	−1.03±0.03
CNF Matching [Ben-Hamu et al., 2022]	−2.38±0.17	−0.38±0.01	**0.25**±0.02	−1.40±0.02
Riemannian Score-Based [De Bortoli et al., 2022]	−4.92±0.25	−0.19±0.07	0.45±0.17	−1.33±0.06
Riemannian Diffusion Model [Huang et al., 2022]	−6.61±0.97	**−0.40**±0.05	0.43±0.07	−1.38±0.05
RFM (Paper results)	−7.93±1.67	−0.28±0.08	0.42±0.05	−1.86±0.11
RFM (Reproduced)	−6.96±1.71	−0.21±0.05	0.48±0.04	−1.77±0.06
Our method	**−9.69**±1.94	−0.28±0.04	0.47±0.04	**−2.11**±0.11

our method maintains its advantage while RFM and RG-VFM swap places. This is consistent with the idea of section 3.4 that awareness of geometric constraints is important for scalability.

Earth and climate science datasets on the sphere. Following Chen and Lipman [2024], we apply our method to datasets compiled from NGDC/WDS [2020a, 2020b], Brakenridge [2017], and EOSDIS [2020] by Mathieu and Nickel [2020]. In table 4, we see uniform improvement relative to our reproduction of the baseline. While our method, like RFM, struggles on the earthquake and flood datasets, it improves over prior work on the volcano and fire datasets, which are the smallest and largest datasets respectively and have highly concentrated regions requiring high fidelity [see plots in De Bortoli et al., 2022]. The models outperforming our method and RFM (Reproduced) do not require implicitly specifying a particular transport path. This allows them to optimize the path. The trade-off is lack of scalability above low manifold dimension.

High-dimensional human pose data. We further evaluate on a large-scale dataset of human poses from AMASS [Mahmood et al., 2019], embedded in $\mathrm{SO}(3)^{19}$, following Dünkel et al. [2024] in the partition of the dataset we use. We train RFM, RG-VFM, and our method on 11 823 971 samples from this partition. We also compare with the released checkpoint of HuProSO3 [Dünkel et al., 2024], which was trained on the full partition, including our test samples. We see in table 5 that our method outperforms according to both sample-based metrics. This supports the use of our method, especially in high-dimensional applications where sample quality is of particular importance.

6 Limitations

Similar to RFM [Chen and Lipman, 2024] and RG-VFM [Zaghen et al., 2026], our method requires specialization per manifold. Both RFM and RG-VFM have the advantage of only requiring well-studied manifold objects. Tables of metric tensors and logarithmic maps are readily available; our

Table 5: Human pose on $\mathrm{SO}(3)^{19}$. Mean±std, 5 seeds, except HuProSO3, which uses the released checkpoint `id_0900`. 20 000 samples per seed.

Method	**NLL** ↓ w.r.t. Haar	**Energy distance** ↓ chordal, $\times 10^3$	**Gaussian MMD²** ↓ chordal, $\times 10^3$
HuProSO3	-137.78	56.76	9.746
RG-VFM	-150.58 ± 0.41	18.60 ± 1.72	3.601 ± 0.308
RFM	$\mathbf{-198.11} \pm 0.81$	2.16 ± 0.20	0.379 ± 0.035
Our method	-197.41 ± 0.44	$\mathbf{1.68} \pm 0.04$	$\mathbf{0.293} \pm 0.008$

method brings the added work of devising an efficient way to compute and invert the mean map and pretabulate the covariance components. This generally involves identifying a symmetry that reduces the integration to low-dimensional quadrature, differentiating under the integral sign, and crafting a numerically stable implementation. We have not implemented our method on general geometries as extensively as Chen and Lipman [2024] have for RFM. For manifolds without closed-form geodesics, even RFM is costly and complex. In this setting, probability paths are defined via diffusive or biharmonic premetrics, so our method would need to perform quadrature over these or estimate the mean and variance via Monte Carlo, which may be prohibitive. We discuss these potential extensions in appendix C. As in RG-VFM, we target data with known symmetries and geometric structure rather than arbitrary 3D meshes, but it may be possible to extend our method to these settings. Unlike RG-VFM, we do not require homogeneity, although it is useful in practice.

7 Related Work

Guzmán-Cordero et al. [2025] previously suggested, based on the ideas of Eijkelboom et al. [2024], that likelihoods over endpoints should be exponential families with a parameter of interest as the sufficient statistic. Their work targets heterogeneous tabular data, relying on the identity $u_t(x_t) = \mathbb{E}[u_t^{\mathrm{cond}}(x_t \mid x_1)] = u_t^{\mathrm{cond}}(x_t \mid \mathbb{E}[x_1])$ via linearity of conditional paths to estimate the marginal field from the endpoint prediction, similar to RG-VFM. As this work establishes, the nonlinearity of conditional paths on manifolds means that such constructions do not generally possess compatibility.

Prior intrinsic generative models include Riemannian CNFs [Mathieu and Nickel, 2020], Moser Flow [Rozen et al., 2021], CNF Matching [Ben-Hamu et al., 2022], and score-based [De Bortoli et al., 2022] and ELBO-based diffusion [Huang et al., 2022]. However, as Chen and Lipman [2024] discuss, these require simulated random walks [Huang et al., 2022, De Bortoli et al., 2022], approximations to the heat kernel [De Bortoli et al., 2022], heavy divergence computation [Huang et al., 2022] or Hutchinson estimation [Mathieu and Nickel, 2020]. Earlier simulation-free methods [Rozen et al., 2021, Ben-Hamu et al., 2022] scale poorly to high dimensions and general geometries.

With the exception of RG-VFM, recent manifold extensions of flow matching mostly specialize RFM to particular geometries or applications [as in Cheng et al., 2024, Miller et al., 2024, Cheng et al., 2025a], or propose an alternative method for a single manifold class [as in Collas et al., 2025]. An orthogonal line of work focuses on accelerated inference. Consistency models [Song et al., 2023] distil one-step maps from intermediate states to data, and Riemannian consistency models [Cheng et al., 2025b] port this idea to manifolds. Similarly, flow maps [Boffi et al., 2025] use self-distillation for few-step inference, and Davis et al. [2026] propose Generalized Flow Maps for manifolds.

8 Conclusion

At population optima, likelihood maximization generically fails to recover the true marginal velocity. Our compatibility condition characterizes when it succeeds. Our residual bound shows that compatibility is enough to avoid this source of error even if the likelihood is misspecified. On compact manifolds, prior work satisfying the condition assigns Gaussian likelihood to infeasible velocities. We propose an exponential family in endpoint space that preconditions the loss while being both compatible and support-aware. Empirically, our method recovers the expected velocity in the analytic test case, improves NLL on two spherical benchmarks, and improves sample-based metrics on high-dimensional tori and human poses. More broadly, in choosing a likelihood for flow matching, we must consider not only the intrinsic geometry, but also whether optimizing yields the right flow.

Acknowledgments and Disclosure of Funding

This research is supported through the EPSRC ROSEHIPS grant [EP/W005816/1].

References

Advanced Computing Center for the Arts and Design. ACCAD MoCap Dataset. URL https://accad.osu.edu/research/motion-lab/mocap-system-and-data.

Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*, June 2015. URL https://doi.org/10.1109/CVPR.2015.7298751.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Heli Ben-Hamu, Samuel Cohen, Joey Bose, Brandon Amos, Maximillian Nickel, Aditya Grover, Ricky T. Q. Chen, and Yaron Lipman. Matching normalizing flows and probability paths on manifolds. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1749–1763. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ben-hamu22a.html.

Nicholas Matthew Boffi, Michael Samuel Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=cqDH0e6ak2.

Clément Bonet, Lucas Drumetz, and Nicolas Courty. Sliced-Wasserstein distances and flows on Cartan-Hadamard manifolds. *Journal of Machine Learning Research*, 26(32):1–76, 2025. URL https://jmlr.org/papers/v26/24-0359.html.

G. R. Brakenridge. Global active archive of large flood events. Dartmouth Flood Observatory, University of Colorado, 2017. URL http://floodobservatory.colorado.edu/Archives/index.html. Accessed 2026-04-13. License: CC-BY 4.0.

Carnegie Mellon University. CMU MoCap Dataset. URL http://mocap.cs.cmu.edu.

Jorge Carrasco-Pollo, Floor Eijkelboom, and Jan-Willem van de Meent. Pawsterior: Variational flow matching for structured simulation-based inference. In *ICLR 2026 Workshop on Geometry-grounded Representation Learning and Generative Modeling*, 2026. URL https://openreview.net/forum?id=QeiyEvdyVq.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=g7ohDlTITL. Accompanying codebase released under CC BY-NC 4.0.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Austin Henry Cheng, Alston Lo, Kin Long Kelvin Lee, Santiago Miret, and Alan Aspuru-Guzik. Stiefel flow matching for moment-constrained structure elucidation. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=84WmbzikPP.

Chaoran Cheng, Jiahan Li, Jian Peng, and Ge Liu. Categorical flow matching on statistical manifolds. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=5fybcQZ0g4.

Chaoran Cheng, Yusong Wang, Yuxin Chen, Xiangxin Zhou, Nanning Zheng, and Ge Liu. Riemannian consistency model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=SAlCQdk5lx.

Antoine Collas, Ce Ju, Nicolas Salvy, and Bertrand Thirion. Riemannian flow matching for brain connectivity matrices via pullback geometry. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=NY3LzmUXl7.

Oscar Davis, Michael Samuel Albergo, Nicholas Matthew Boffi, Michael M. Bronstein, and Joey Bose. Generalised flow maps for few-step generative modelling on Riemannian manifolds. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1YHF7B8Yjk.

Valentin De Bortoli, Emile Mathieu, Michael John Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=oDRQGo8I7P.

Olaf Dünkel, Tim Salzmann, and Florian Pfaff. Normalizing flows on the product space of SO(3) manifolds for probabilistic human pose modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2024. URL https://doi.org/10.1109/CVPR52733.2024.00222.

Floor Eijkelboom, Grigory Bartosh, Christian A. Naesseth, Max Welling, and Jan-Willem van de Meent. Variational flow matching for graph generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=UahrHR5HQh.

EOSDIS. Active fire data. Land, Atmosphere Near Real-time Capability for EOS (LANCE) system operated by NASA's Earth Science Data and Information System (ESDIS), 2020. URL https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/active-fire-data. Accessed 2026-04-13. License: CC0 1.0 (NASA EOSDIS).

Eyes JAPAN Co. Ltd. Eyes Japan MoCap Dataset. URL http://mocapdata.com.

Aasa Feragen, Francois Lauze, and Soren Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3032–3042, Los Alamitos, CA, USA, Jun 2015. IEEE Computer Society. URL https://doi.org/10.1109/CVPR.2015.7298922.

Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Bharath K. Sriperumbudur. Characteristic kernels on groups and semigroups. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/hash/d07e70efcfab08731a97e7b91be644de-Abstract.html.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2):123–214, 03 2011. ISSN 1369-7412. URL https://doi.org/10.1111/j.1467-9868.2010.00765.x.

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

Andrés Guzmán-Cordero, Floor Eijkelboom, and Jan-Willem van de Meent. Exponential family variational flow matching for tabular data generation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=kjtvCSkSsy.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Ludovic Hoyet, Kenneth Ryall, Rachel McDonnell, and Carol O'Sullivan. Sleight of hand: Perception of finger motion from reduced marker sets. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 79–86, 2012. URL https://doi.org/10.1145/2159616.2159629.

Vincent Tao Hu, Wei Zhang, Meng Tang, Pascal Mettes, Deli Zhao, and Cees Snoek. Latent space editing in transformer-based flow matching. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3):2247–2255, Mar. 2024. URL https://doi.org/10.1609/aaai.v38i3.27998.

Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron Courville. Riemannian diffusion models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=ecevn9kPm4.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=66NzcRQuOq.

Simon Lyngby Kokkendorff. *Geometry and Combinatorics*. Ph.D. thesis, Department of Mathematics, Technical University of Denmark, 2002. URL https://orbit.dtu.dk/en/publications/geometry-and-combinatorics. See p. 88, Theorem 15: compact Riemannian manifolds of negative type and dimension at least 2 are simply connected.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f0935e4cd5920aa6c7c996a5ee53a70f-Abstract.html.

Franziska Krebs, Andre Meixner, Isabel Patzer, and Tamim Asfour. The KIT bimanual manipulation dataset. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, pages 499–506, 2021. URL https://doi.org/10.1109/HUMANOIDS47582.2021.9555788.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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

Russell Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305, September 2013. ISSN 0091-1798. URL https://doi.org/10.1214/12-AOP803.

Russell Lyons. Strong negative type in spheres. *Pacific Journal of Mathematics*, 307(2):383–390, September 2020. ISSN 0030-8730. URL https://doi.org/10.2140/pjm.2020.307.383.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5441–5450, 2019. URL https://doi.org/10.1109/ICCV.2019.00554.

Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The KIT whole-body human motion database. In *International Conference on Advanced Robotics (ICAR)*, pages 329–336, 2015. URL https://doi.org/10.1109/ICAR.2015.7251476.

Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. Unifying representations and large-scale whole-body motion databases for studying human motion. *IEEE Transactions on Robotics*, 32(4):796–809, 2016. URL https://doi.org/10.1109/TRO.2016.2572685.

Emile Mathieu and Maximilian Nickel. Riemannian continuous normalizing flows. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 2503–2515, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546. URL https://proceedings.neurips.cc/paper/2020/hash/1aa3d9c6ce672447e1e5d0f1b5207e85-Abstract.html.

Răzvan-Andrei Matişan, Vincent Tao Hu, Grigory Bartosh, Björn Ommer, Cees G. M. Snoek, Max Welling, Jan-Willem van de Meent, Mohammad Mahdi Derakhshani, and Floor Eijkelboom. Purrception: Variational flow matching for vector-quantized image generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=SA8xDYrUYB.

Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M Wood. FlowMM: Generating materials with Riemannian flow matching. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=W4pB7VbzZI.

National Geophysical Data Center / World Data Service (NGDC/WDS). NCEI/WDS global significant earthquake database. NOAA National Centers for Environmental Information, 2020a. URL https://doi.org/10.7289/V5TD9V7K. Accessed 2026-04-13.

National Geophysical Data Center / World Data Service (NGDC/WDS). NCEI/WDS global significant volcanic eruptions database. NOAA National Centers for Environmental Information, 2020b. URL https://doi.org/10.7289/V5JW8BSH. Accessed 2026-04-13.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. URL https://jmlr.org/papers/v22/19-1028.html.

Xavier Pennec. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, July 2006. ISSN 1573-7683. URL https://doi.org/10.1007/s10851-006-6228-4.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. URL https://doi.org/10.1137/0330046.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions, 2018. URL https://openreview.net/forum?id=SkBYYyZRZ.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/rezende15.html.

Ankit Rohatgi. WebPlotDigitizer, 2026. URL https://automeris.io. Version 5.2.

Noam Rozen, Aditya Grover, Maximilian Nickel, and Yaron Lipman. Moser flow: Divergence-based generative modeling on manifolds. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=qGvMv3undNJ.

Raif M. Rustamov and Subhabrata Majumdar. Intrinsic sliced Wasserstein distances for comparing collections of probability distributions on manifolds and graphs. In *Proceedings of the 40th International Conference on Machine Learning*, pages 29388–29415, 2023. URL https://proceedings.mlr.press/v202/rustamov23a.html.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32211–32252. PMLR, 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11 (50):1517–1561, 2010. URL https://jmlr.org/papers/v11/sriperumbudur10a.html.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. URL https://doi.org/10.1016/j.jspi.2013.03.018.

E. G. Tabak and Cristina V. Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013. URL https://doi.org/10.1002/cpa.21423.

Nikolaus F. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2(5):2–2, September 2002. URL https://doi.org/10.1167/2.5.2.

Matt Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total Capture: 3D human pose estimation fusing video and inertial sensors. In *2017 British Machine Vision Conference (BMVC)*, 2017. URL https://doi.org/10.5244/C.31.14.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009. URL https://doi.org/10.1007/978-3-540-71050-9.

Olga Zaghen, Floor Eijkelboom, Alison Pouplin, Cong Liu, Max Welling, Jan-Willem van de Meent, and Erik J Bekkers. Riemannian variational flow matching for material and protein design. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=NlnDselrtl.

Appendix

A Proofs and theoretical derivations 15

B Supplementary experimental results 19

C Implementation details 21

D Specialization to common geometries 22

E Experimental methodology and evaluation 26

F Computational performance 28

Organization of the appendix. We prove and derive the main theoretical results of interest in appendix A; we share more experiments and ablations in appendix B; we describe implementation details and training algorithms in appendix C; we specialize the method to some common manifolds of interest in appendix D; we outline the experimental protocols and hyperparameters in appendix E; and we report measurements of computational performance in appendix F.

A Proofs and theoretical derivations

A.1 Compatibility bound

Theorem 1. *Assume measurable $\psi_{x,t} \in L^2(q^{\theta^*}; T_x\mathcal{M})$, $s^* \in L^2(q^{\theta^*}; \mathbb{R}^m)$. Let p be absolutely continuous with respect to q^{θ^*}. Assume the first-order optimality condition $\mathbb{E}_p[s^*(z)] = 0$ (requiring θ^* to be interior) and enough regularity for the score identity $\mathbb{E}_{q^{\theta^*}}[s^*(z)] = 0$. Then*

$$\left\|\mathbb{E}_p[\psi_{x,t}(z)] - \mathbb{E}_{q^{\theta^*}}[\psi_{x,t}(z)]\right\|_g \leq \sqrt{\chi^2(p\|q^{\theta^*})}\sqrt{\left\|\mathbb{E}_{q^{\theta^*}}[r_{x,t} \otimes r_{x,t}]\right\|_{\mathrm{op},g}}, \tag{9}$$

where $\chi^2(p\|q^{\theta^*}) < \infty$ is the chi-squared divergence between p and q^{θ^*}.

Proof. The residual $r_{x,t}$ is defined as that of the least-squares problem

$$(A^*_{x,t}, b^*_{x,t}) \in \arg\min_{A,b} \mathbb{E}_{q^{\theta^*}}\left[\|\psi_{x,t}(z) - As^*(z) - b\|_g^2\right].$$

In particular,

$$\psi_{x,t} = A^*_{x,t}s^* + b^*_{x,t} + r_{x,t}.$$

Now taking expectations under each distribution p and q^{θ^*},

$$\mathbb{E}_p[\psi_{x,t}] = A^*_{x,t}\underbrace{\mathbb{E}_p[s^*]}_{0} + b^*_{x,t} + \mathbb{E}_p[r_{x,t}],$$

$$\mathbb{E}_{q^{\theta^*}}[\psi_{x,t}] = A^*_{x,t}\underbrace{\mathbb{E}_{q^{\theta^*}}[s^*]}_{0} + b^*_{x,t} + \mathbb{E}_{q^{\theta^*}}[r_{x,t}].$$

In the subtraction of these expressions, the $b^*_{x,t}$ terms cancel and we are left with

$$\begin{aligned}\mathbb{E}_p[\psi_{x,t}] - \mathbb{E}_{q^{\theta^*}}[\psi_{x,t}] &= \mathbb{E}_p[r_{x,t}] - \mathbb{E}_{q^{\theta^*}}[r_{x,t}] \\ &= \mathbb{E}_{q^{\theta^*}}\left[\left(\frac{\mathrm{d}p}{\mathrm{d}q^{\theta^*}} - 1\right) r_{x,t}\right].\end{aligned}$$

For any unit vector $w \in T_x\mathcal{M}$, Cauchy–Schwarz yields

$$\begin{aligned}\left|\left\langle w,\, \mathbb{E}_p[\psi_{x,t}] - \mathbb{E}_{q^{\theta^*}}[\psi_{x,t}]\right\rangle_g\right| &\leq \sqrt{\chi^2(p\|q^{\theta^*})}\sqrt{\mathbb{E}_{q^{\theta^*}}\left[\langle w, r_{x,t}\rangle_g^2\right]} \\ &= \sqrt{\chi^2(p\|q^{\theta^*})}\sqrt{\left\langle w,\, \mathbb{E}_{q^{\theta^*}}[r_{x,t} \otimes r_{x,t}]w\right\rangle_g} \\ &\leq \sqrt{\chi^2(p\|q^{\theta^*})}\sqrt{\left\|\mathbb{E}_{q^{\theta^*}}[r_{x,t} \otimes r_{x,t}]\right\|_{\mathrm{op},g}}.\end{aligned}$$

Taking supremum over unit vectors w concludes the proof. □

A.2 Motivating example with smooth data on $\mathbb{T}^1$

Here, we check whether the example of section 3.1 survives after smoothing the mixture of Diracs into that of wrapped Gaussians,

$$p_0 = \frac{1}{2}\mathcal{WN}(-\alpha, \sigma_0^2) + \frac{1}{2}\mathcal{WN}(+\alpha, \sigma_0^2), \tag{20}$$

$$p_1 = \frac{1}{2}\mathcal{WN}(-\beta, \sigma_1^2) + \frac{1}{2}\mathcal{WN}(+\beta, \sigma_1^2). \tag{21}$$

As before, let $\tau \in (0, \frac{1}{2})$, $\beta \in (\frac{\pi}{2}, \pi(1-\tau))$, and $\alpha = \frac{\tau\beta}{1-\tau}$. Geodesics joining the component centres $(+\alpha, -\beta)$ and $(-\alpha, +\beta)$ meet at $x_\tau = 0$. For a coupled pair with geodesic velocity $w = \log_{x_0}(x_1) \in [-\pi, \pi]$ at $x_t = x$, the endpoints satisfy $x_0 = x - tw$ and $x_1 = x + (1-t)w$ (modulo 2π). As the change of variables has unit Jacobian, under independent coupling the conditional density satisfies

$$p_t(w \mid x) \propto p_0(x - tw)p_1(x + (1-t)w). \tag{22}$$

Both mixtures satisfy $p_i(x) = p_i(-x)$. Hence, $p_t(w \mid 0) = p_t(-w \mid 0)$. For all $\sigma_0, \sigma_1 > 0$ and $0 \le t < 1$,

$$u_t(0) = \int_{-\pi}^{\pi} w\, p_t(w \mid 0)\, \mathrm{d}w = 0. \tag{23}$$

Local behaviour of the marginal velocity. The derivative $u'_\tau(0)$ determines the direction of the true marginal field near the crossing. Let $\tau = \frac{1}{4}$, $\alpha = \frac{1}{3}\beta$, and $\sigma_0 = \sigma_1 = \sigma$. For small σ, the two crossing component pairs dominate $p_{\frac{1}{4}}(w \mid x)$ near $x = 0$. Via the corresponding Gaussian components on $\mathbb{R}$, we calculate the conditional mean of w for each crossing pair. For pairs centred at $(+\alpha, -\beta)$ and $(-\alpha, +\beta)$, we have, via completion of the square,

$$\mathcal{N}\left(x - \frac{1}{4}w;\, \pm\alpha, \sigma^2\right)\mathcal{N}\left(x + \frac{3}{4}w;\, \mp\beta, \sigma^2\right) \propto \exp\left(-\frac{(x - \frac{1}{4}w \mp \alpha)^2 + (x + \frac{3}{4}w \pm \beta)^2}{2\sigma^2}\right) \tag{24}$$

$$= \exp\left(-\frac{5}{16\sigma^2}\left(w + \frac{4x}{5} \pm (\alpha + \beta)\right)^2 - \frac{4x^2}{5\sigma^2}\right). \tag{25}$$

The two Gaussian products have the same dependence on x (after integration over $w \in \mathbb{R}$), so they also have equal posterior weights. Averaging over the two crossing pairs cancels the terms in $\alpha + \beta$,

$$\mathbb{E}\left[w \mid x_{\frac{1}{4}} = x,\ \text{crossing pairs}\right] = \frac{1}{2}\left(-(\alpha+\beta) - \frac{4}{5}x + (\alpha+\beta) - \frac{4}{5}x\right) \tag{26}$$

$$= -\frac{4}{5}x. \tag{27}$$

For pairs whose endpoints are of the same sign, the mean of $x_{\frac{1}{4}}$ is $\pm\frac{1}{2}\beta$; for crossing pairs, it is zero. For all four cases, the variance is $\frac{5}{8}\sigma^2$. For each pair whose endpoints are of the same sign, the distribution of $x_{\frac{1}{4}}$ has density at zero equal to $e^{-\beta^2/(5\sigma^2)}$ times that for a crossing pair. To obtain the derivative for the wrapped distributions, we must also account for periodic copies and the restriction $|w| \le \pi$. We have

$$u_t(x) = \frac{\int_{-\pi}^{\pi} w\, p_0(x - tw)p_1(x + (1-t)w)\, \mathrm{d}w}{\int_{-\pi}^{\pi} p_0(x - tw)p_1(x + (1-t)w)\, \mathrm{d}w}. \tag{28}$$

Since $u_t(0) = 0$, differentiating gives

$$u'_t(0) = \frac{\int_{-\pi}^{\pi} w\, [p'_0(-tw)p_1((1-t)w) + p_0(-tw)p'_1((1-t)w)]\, \mathrm{d}w}{\int_{-\pi}^{\pi} p_0(-tw)p_1((1-t)w)\, \mathrm{d}w} \tag{29}$$

$$= \mathbb{E}_{p_t(\cdot|0)}\left[w\left(\frac{p'_0(-tw)}{p_0(-tw)} + \frac{p'_1((1-t)w)}{p_1((1-t)w)}\right)\right]. \tag{30}$$

At $x_{\frac{1}{4}} = 0$, the endpoints satisfy $x_0 = -\frac{1}{4}w$ and $x_1 = \frac{3}{4}w$. Both endpoints coincide with their Gaussian means for only the crossing pairs, at $w = \pm\frac{4}{3}\beta \in (-\pi, \pi)$. For the other pairs, no value of w puts both endpoints at their Gaussian means. The crossing peaks lie some fixed distance from the integration boundaries. Hence, other Gaussian terms and the tails, added when extending the crossing integrals to $\mathbb{R}$, change the derivative of the numerator and the value of the denominator by at most $C\sigma^{-K}e^{-c/\sigma^2}$ for some $C, c > 0$ and finite K. The crossing terms contribute a denominator of order σ^{-1}, so

$$u'_{\frac{1}{4}}(0) = -\frac{4}{5} + \mathcal{O}(e^{-c'/\sigma^2}) \tag{31}$$

for some $c' > 0$. Reflection means $u_{\frac{1}{4}}(x)$ is odd, implying that

$$u_{\frac{1}{4}}(x) = \left(-\frac{4}{5} + \mathcal{O}(e^{-c'/\sigma^2})\right) x + \mathcal{O}_\sigma(x^3). \tag{32}$$

That is to say, for sufficiently narrow components, the true marginal velocity points toward zero on both sides of the crossing.

Global optimum of the RG-VFM loss for smooth data. RG-VFM minimizes the Fréchet function of the conditional endpoint distribution. At $x_\tau = 0$, the endpoint corresponding to w is $x_1 = (1-\tau)w$, so

$$F(\mu) = \int_{-\pi}^{\pi} \operatorname{dist}_g((1-\tau)w, \mu)^2 p_\tau(w \mid 0)\, \mathrm{d}w. \tag{33}$$

By symmetry, $F(\pi+\delta) = F(\pi-\delta)$, meaning $\mu = \pi$ is always a stationary point. Let $Y = (1-\tau)|w|$ be the magnitude of the endpoint, where $p_\tau(w \mid 0)$ induces the density ρ of Y. Again by symmetry, it will suffice to compare π with $\mu = \pi - \delta$ for $0 \le \delta \le \pi$. For symmetric endpoints $\pm y$, the Fréchet function changes by

$$\frac{1}{2}\operatorname{dist}_g(\pi-\delta, y)^2 + \frac{1}{2}\operatorname{dist}_g(\pi-\delta, -y)^2 - (\pi-y)^2 = \begin{cases} \delta^2, & \delta \le y, \\ \delta^2 - 2\pi(\delta - y), & \delta > y \end{cases} \tag{34}$$

$$= \delta^2 - 2\pi[\delta - y]_+. \tag{35}$$

Now, averaging over Y,

$$F(\pi-\delta) - F(\pi) = \delta^2 - 2\pi\int_0^\delta (\delta - y)\rho(y)\, \mathrm{d}y. \tag{36}$$

Only endpoint magnitudes $Y < \delta$ contribute to the second term, which offsets the quadratic cost of moving away from π. Choose $b \in (\pi/2, \beta)$ and let $M = \sup_{0 \le y \le b} \rho(y)$. If $M < 1/\pi$ and $b\,\mathbb{P}(Y \ge b) > \pi/2$, then, for $0 < \delta \le b$,

$$F(\pi-\delta) - F(\pi) \ge \delta^2 - 2\pi M \int_0^\delta (\delta - y)\, \mathrm{d}y \tag{37}$$

$$= \underbrace{(1 - \pi M)}_{>0}\, \delta^2. \tag{38}$$

For $b \le \delta \le \pi$,

$$F(\pi-\delta) - F(\pi) = \delta^2 - 2\pi\,\mathbb{E}\left[[\delta - Y]_+\right] \tag{39}$$

$$= \delta^2 - 2\pi\delta + 2\pi\,\mathbb{E}[\min(\delta, Y)] \tag{40}$$

$$\ge (\pi-\delta)^2 + \underbrace{2\pi b\,\mathbb{P}(Y \ge b) - \pi^2}_{>0}. \tag{41}$$

For the wrapped Gaussian mixtures, Y concentrates at β as $\sigma_0, \sigma_1 \to 0$, so $M \to 0$ and $\mathbb{P}(Y \ge b) \to 1$. The strict inequalities hold for sufficiently small component variances, persisting for small perturbations (within the family of symmetric mixtures). Hence $\mu = \pi$ is the unique global optimum for RG-VFM on an open region within the parameterized family of symmetric smooth mixtures. At $\mu = \pi$, either minimum-length geodesic from zero has length π. Because of this, the smoothed case has the same branch ambiguity as for the mixture of Diracs. While symmetry gives $u_\tau(0) = 0$, RG-VFM decodes this endpoint into a velocity of magnitude $\frac{\pi}{1-\tau}$. Again, the velocity error at the population optimum is

$$\left\| v_\tau^{\theta^*}(0) - u_\tau(0) \right\|_g = \frac{\pi}{1-\tau}. \tag{42}$$

A.3 Characterization of the Gaussian carrier

As the theory of compatibility does not select a universal carrier, we impose a small number of heuristic desiderata. We desire for the carrier to be centred, isotropic, factorizable on product manifolds, give ordinary CFM in the Euclidean case, linearize to RFM in the limit $t \to 1$, and be normalizable on manifolds of interest. For a radially isotropic carrier, coordinates on product manifolds enter only through the joint norm $\|u_t^{\text{cond}}\|_g$. Following the same derivation as for Herschel–Maxwell, we write such a carrier as $f(\|u\|_g^2)$ and use factorization to obtain

$$\frac{f(a+b)}{f(0)} = \frac{f(a)}{f(0)}\frac{f(b)}{f(0)}, \tag{43}$$

where $a, b \geq 0$. Hence, by continuity, $f(a) = f(0)e^{-\beta a}$. Geodesic exponentials provide a natural setting in which to vary the tailedness of the carrier in ablation (see appendix B.3). Terminal locality excludes the constant and growing choices. This means that $\beta > 0$, and equating the Euclidean covariance to I selects $\beta = \frac{1}{2}$ uniquely. The requirements narrow to the present choice of Gaussian carrier.

A.4 Limiting behaviour and relationship to Riemannian flow matching

Manifolds with diffeomorphic exponential map. On these *non*compact manifolds, a special choice of carrier recovers Riemannian flow matching. We see that the Jacobian term introduces an artificial linearization of the likelihood model over endpoints, collapsing the variance to the identity. Let the carrier be

$$c_{x,t}(x_1) = \exp\left(-\frac{1}{2}\|u_t^{\text{cond}}\|_g^2\right)\left|\det Du_t^{\text{cond}}\right|. \tag{44}$$

Here, $\left|\det Du_t^{\text{cond}}\right|$ is the Jacobian determinant with respect to the Riemannian volume measures. The exponential family over endpoints

$$q_t(\mathrm{d}x_1|x_t;\eta) = c_{x,t}(x_1)\exp\left(\langle\eta, u_t^{\text{cond}}\rangle_g - A_{x,t}(\eta)\right)\,\mathrm{dVol}_g(x_1) \tag{45}$$

pushes forward under u_t^{cond} to a density on $T_x\mathcal{M}$,

$$q_u(u;\eta) \propto \exp\left(-\frac{1}{2}\|u\|_g^2 + \langle\eta,u\rangle_g\right)\overbrace{\left|\det Du_t^{\text{cond}}\right|}^{\text{carrier}}\underbrace{\left|\det Du_t^{\text{cond}}\right|^{-1}}_{\text{change of variables}} \tag{46}$$

$$= \exp\left(-\frac{1}{2}\|u-\eta\|_g^2 + \frac{1}{2}\|\eta\|_g^2\right). \tag{47}$$

Thus the pushforward is $\mathcal{N}(\eta, g^{-1})$, and

$$A_{x,t}(\eta) = \log\int_{T_x\mathcal{M}}\exp\left(-\frac{1}{2}\|u\|_g^2 + \langle\eta,u\rangle_g\right)\,\mathrm{dVol}_g(u) \tag{48}$$

$$= \frac{1}{2}\|\eta\|_g^2 + \frac{d}{2}\log(2\pi). \tag{49}$$

In consequence, $m_{x,t}(\eta) = \eta$, $V_{x,t} = I$, and the likelihood-based loss is equivalent to RFM up to the factor $\frac{1}{2}$. On flat manifolds like product tori, this carrier is equivalent to our choice up to a constant factor, but the pushforward for product tori is truncated, as in appendix D.1.

Asymptotic convergence as $t \to 1$. Under our choice of carrier, as $t \to 1$, the endpoint distribution concentrates around x_t, with typical distances of order $1-t$. In this limit, the geometry becomes locally Euclidean by linearization of the exponential map, and the covariance operator $V_{x,t}(v_t^\theta(x_t))$ converges to the identity, so the sample loss approaches $\frac{1}{2}\|v_t^\theta(x_t) - u_t^{\text{cond}}\|_g^2$, the RFM sample loss up to the factor $\frac{1}{2}$. Examples of this limit can be found in appendix D.

B Supplementary experimental results

B.1 Profiles of the vector fields for the analytic test case on $\mathbb{T}^1$

In the absence of compatibility, first-order conditions for likelihood maximization do not fully determine the learned marginal velocity in the sense we discuss in section 3.2. Here, we study the learned field corresponding to the analytic example in section 3.1, in particular its behaviour with respect to ODE solvers. The analytic predictions are visible in $v_t^\theta(x{=}0)$ along $t \in [0, 1]$, as shown in fig. 4. Here, there is a branch ambiguity in that the RG-VFM decoder has the arbitrary choice of two shortest paths to π, which approach from opposing directions. The upper cluster for RG-VFM (the globally optimal case, defined as in fig. 2) exhibits what is nearly a discontinuity at $t = \tau = \frac{1}{4}$, with the one-sided limits matching the $\pm\frac{4\pi}{3}$ values predicted in section 3.1. The lower cluster is smooth but converges to the local minimum at zero velocity, which is not globally optimal. RFM and our method are smooth throughout. We attribute the high NFE with dopri5 reported in table 2 to the high local Lipschitz constant of the field in the upper cluster at $t \approx \frac{1}{4}$. The kink at $t = 0.9$ in both RG-VFM clusters is due to the clipping of $\frac{1}{1-t}$ to a maximum of 10. This is necessary because the pole at $t = 1$ is not automatically separable in the same way as it is for RFM and our method, as the network output is not guaranteed to match the rate necessary for pole cancellation. We take the clamp bounds from the implementation provided by Zaghen et al. [2026].

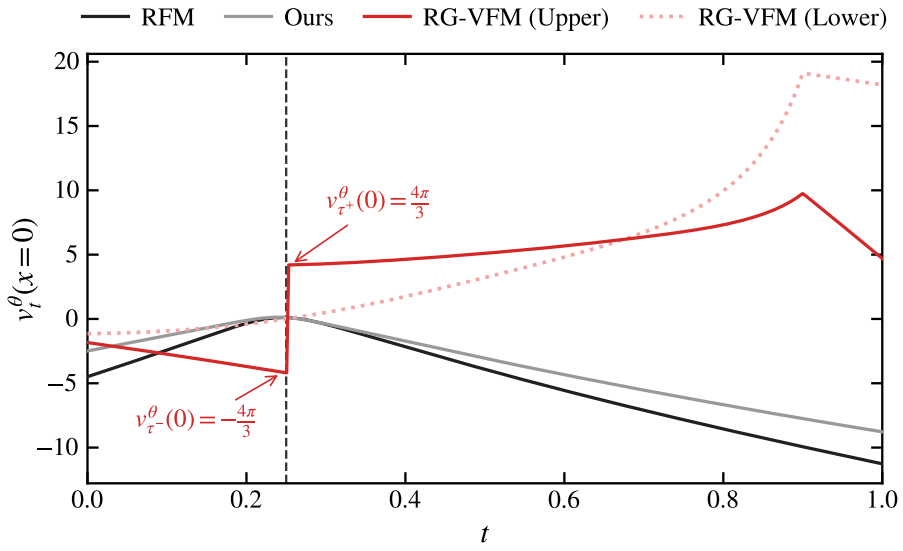


Figure 4: $v_t^\theta(x{=}0)$ as a function of t. One representative seed for RFM, our method, and each RG-VFM cluster. The vertical dashed line marks $\tau = \frac{1}{4}$. RG-VFM (Upper) jumps between the two globally minimizing branches at $\pm\frac{4\pi}{3}$ predicted in section 3.1. It takes many small steps for the ODE solver to resolve the stiff field.

B.2 Evaluating the NLL for RG-VFM

To check that the uniform violations of the entropic bound we find in the RG-VFM NLLs in table 3 are not an artefact of our implementation, we reproduce the checkerboard task of Zaghen et al. [2026] on $\mathbb{S}^2$. We use the published codebase on the same task (intrinsic variant, 3 000 epochs at the published hyperparameters) and study the resulting NLL of the continuous normalizing flow under a wide range of solvers. We compute the CNF likelihood via change of variables with exact divergence (per-coordinate `torch.autograd.grad`), reporting values in bits. Here, we observe values that are too large rather than too small. We cannot predict which way the NLL will diverge, but we hypothesize this to be a function of the data and solver.

Increasing the number of steps. Figure 5 shows a sweep over the number N of Euler steps. RFM stabilizes at approximately 2.51 bits (below the entropic upper bound $\log_2(4\pi) \approx 3.65$ bits for the

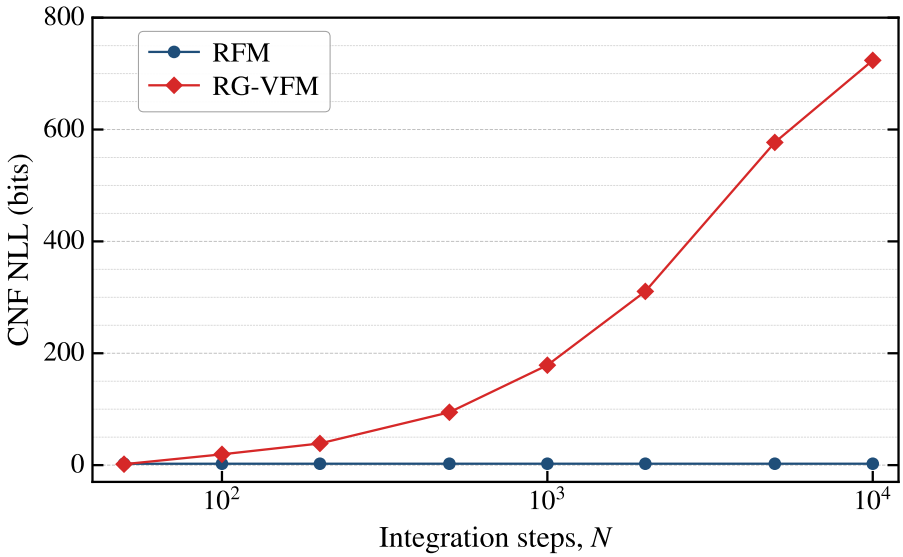


Figure 5: NLL of the continuous normalizing flow on the $\mathbb{S}^2$ checkerboard task of Zaghen et al. [2026] as a function of the number N of Euler integration steps (log x, linear y). 1 000 samples. RFM converges to approximately 2.5 bits; RG-VFM appears to diverge monotonically. Exact divergence.

Table 6: Varying the tolerance of dopri5 on the $\mathbb{S}^2$ checkerboard. 1 000 samples. We allow 300 s for each call to the solver.

	RFM		RG-VFM	
Tolerance	**NLL** ↓ (bits)	**NFE** ↓	**NLL** ↓ (bits)	**NFE** ↓
10^{-2}	2.48	32	16.6	452
10^{-3}	2.47	50	130.0	2 318
10^{-4}	2.47	134	312.3	10 328
10^{-5}	2.47	494	533.5	34 694
10^{-6}	2.47	2 216	timeout	—
10^{-7}	2.47	9 902	timeout	—

uniform distribution on $\mathbb{S}^2$) by $N = 200$ and stays there. RG-VFM rises from 1.7 bits at $N = 50$ to 723.7 bits at $N = 10\,000$. Repeating the sweep with Heun (2nd order) and RK4 (4th order) gives qualitatively identical behaviour, exceeding 50 bits by $N = 500$ for both. This eliminates the possibility of this phenomenon being an artefact of discretization to first order.

Reducing the tolerance of dopri5. Table 6 exhibits the same divergence under dopri5 at tightening tolerances. RFM stays around 2.47 bits while RG-VFM rises from 16.6 bits to 533.5 bits as tolerance drops from 10^{-2} to 10^{-5}. At tolerance 10^{-5}, RG-VFM requires $\sim 70\times$ as many function evaluations as RFM. RG-VFM at tolerances $\leq 10^{-6}$ exceeds the 300 s time limit; RFM completes in under 2 s at the tolerances we test.

Specialized solvers for stiff ODEs. After increasing the time limit to 900 s per call, we also test SciPy's stiff integrators LSODA, BDF and Radau via `torchdiffeq`. We use tolerances $\{10^{-2}, 10^{-3}, 10^{-5}, 10^{-7}\}$, resulting in 12 runs per model. For RFM, 10 runs complete, taking between 0.2 s and 399 s. Their NLLs are within 0.04 bits of 2.47 bits. LSODA and Radau time out at tolerance 10^{-7}. For RG-VFM, all 12 runs reach the time limit.

Table 7: Carrier ablation on $\mathbb{S}^2$. Mean±std [median], 15 seeds. 20 000 samples per seed.

	KL ↓ $\times 10^3$	**Energy distance** ↓ $\times 10^6$	**Laplace MMD^2** ↓ $\times 10^6$
RFM	0.1765 ± 0.1447 [0.1607]	0.945 ± 5.453 [−0.336]	2.026 ± 7.359 [−0.187]
$p = 0$	0.5485 ± 0.2773 [0.4488]	9.811 ± 7.990 [5.828]	15.173 ± 11.696 [9.467]
$p = 1$	0.1398 ± 0.1339 [0.1070]	1.389 ± 4.884 [−0.306]	2.681 ± 6.189 [−0.098]
$p = 2$	$\mathbf{0.1350} \pm 0.1091$ **[0.1037]**	$\mathbf{0.584} \pm 5.086$ **[−1.432]**	$\mathbf{1.465} \pm 6.405$ [−1.383]
$p = 3$	0.1779 ± 0.1667 [0.1215]	0.713 ± 4.996 [−0.819]	1.649 ± 6.420 [−1.138]
$p = 4$	0.1617 ± 0.1567 [0.1078]	0.876 ± 5.740 [−1.237]	1.872 ± 7.662 **[−1.641]**

B.3 Carrier ablation

Here, as a natural setting to compare our choice of carrier against potential alternatives, we consider the family of geodesic exponential carriers

$$c_p = \begin{cases} \exp\left[-\left(\dfrac{\|u_t^{\text{cond}}\|_g}{a_p} \right)^p \right], & p \in \{1, 2, 3, 4\}, \\ 1, & p = 0, \end{cases} \tag{50}$$

where, for $p > 0$,

$$a_p = \sqrt{2} \sqrt{\frac{\Gamma(2/p)}{\Gamma(4/p)}}. \tag{51}$$

We choose a_p so that, in two-dimensional Euclidean space at $\eta = 0$, the covariance is $V = I$. This is for the same reason that we omit the factor of $\frac{1}{2}$ in the loss—it allows us to compare based on shape and not global rescaling. A global rescaling would largely correspond to just a factor change in the learning rate. We choose c_0 as a degenerate carrier that does not localize the endpoint distribution around x_t as $t \to 1$. Clearly, this is inefficient density assignment in the $t = 1$ limit, and more so than RFM, whose Gaussian tails decay outside the feasible velocity support.

Unlike c_0 and c_2, the other carriers c_1, c_3, and c_4 do not factor coordinate-wise on product manifolds. For c_2, factorization follows from the additive separability of the squared product norm. Hence, we conduct the ablation on $\mathbb{S}^2$, where pretabulating works for all carriers via quadrature in 1D. Specifically, we compare under the wrapped Gaussian target (a Gaussian in the tangent space pushed forward onto the manifold) described in section 5 for the $\mathbb{T}^d$ experiments. We use RFM as a baseline. As well as the target, the uniform base, coupling, and manifold geometry determine the posterior.

In table 7, many negative medians occur for energy distance and MMD as the finite sample estimators exhibit skew although the population mean must be non-negative. KL is computed from the difference of the test/NLL and the wrapped Gaussian entropy estimated from the same test samples (i.e., acting as a Monte Carlo control variate). The lowest variance comparison is by KL. Ranked by KL, c_0 is clearly worst. Except for $p = 0$, the other powers do not show a clear difference, consistent with the carrier being a robust secondary choice after discarding alternatives filtered out by desiderata such as normalizability.

C Implementation details

C.1 Network output and domain of the mean map

For a full, regular exponential family with positive carrier and bounded sufficient statistic, under minimality $A_{x,t}$ is strictly convex. Hence, $m_{x,t}(\eta) = \nabla_\eta A_{x,t}$ is one-to-one and maps the natural-parameter space onto the relative interior of the convex hull of feasible conditional velocities. Thus, $m_{x,t}^{-1}(v)$ exists when v lies in that interior. At finite interior parameters, $V_{x,t} = \nabla_\eta^2 A_{x,t}$ is positive definite under minimality. If v approaches the boundary, some eigenvalues of the covariance can tend to zero. That is, $V_{x,t}^{-1}$ does have the potential to become ill-conditioned or diverge in the limit. Our notion of support awareness relies on the conditional velocity field being well defined. This is the same prerequisite already required by RFM for the regression target, and by RG-VFM in the decoder from endpoints to velocities—both assume that the set of coupled pairs with non-unique

probability paths has measure zero. That is to say, we do not introduce branch dependence at cut loci as an additional assumption or limitation of our method. Our endpoint likelihood induces density only on velocities produced by the conditional map from endpoints to velocities, relative to a fixed choice of branch.

In implementation, the network predicts an unconstrained velocity $v_t^\theta(x_t) \in T_{x_t}\mathcal{M}$. We evaluate the surrogate loss where the mean map is invertible. Outside this region, we use a barrier-style extension, which we obtain by clamping the argument of the mean parameter prior to inversion of the table. This has no effect on the population minimizer, provided it lies in the interior. This is because the effective loss we implement in practice is larger outside that interior and the gradients of the penalty all point inwards. On tori we clamp per coordinate while on spheres we clamp the norm. Across the selected checkpoints in our reported experiments, the fraction of predictions for which the barrier remained active was zero. Notwithstanding potentially large gradients as the mean parameter approaches the boundary of its domain, we do not observe any stability issues in practice and did not require gradient clipping, warmup, or other special techniques to train our method. We scale our loss to be on the same scale as the RFM loss by omitting the factor of $\frac{1}{2}$.

C.2 Training algorithms

In fig. 6, we compare our method and RFM on simple geometries in algorithmic terms.

Algorithm 1 Riemannian Flow Matching

Require: base distribution $p_0(x_0)$, target distribution $p_1(x_1)$

Initialize parameters θ of $v_t^\theta(x)$
while not converged **do**
  sample time $t \sim \mathcal{U}(0,1)$
  sample training example $x_1 \sim p_1$
  sample noise $x_0 \sim p_0$
  $x_t \leftarrow \exp_{x_0}\big(t \log_{x_0}(x_1)\big)$
  $\Delta \leftarrow v_t^\theta(x_t) - u_t^{\text{cond}}(x_t \mid x_1)$
  $\ell(\theta) \leftarrow \|\Delta\|_g^2$
  $\theta \leftarrow \text{optimizer_step}(\ell(\theta))$
end while

Algorithm 2 Our method

Require: base distribution $p_0(x_0)$, target distribution $p_1(x_1)$

Pretabulate components of $V_{x,t}(v)$
Initialize parameters θ of $v_t^\theta(x)$
while not converged **do**
  sample time $t \sim \mathcal{U}(0,1)$
  sample training example $x_1 \sim p_1$
  sample noise $x_0 \sim p_0$
  $x_t \leftarrow \exp_{x_0}\big(t \log_{x_0}(x_1)\big)$
  $\Delta \leftarrow v_t^\theta(x_t) - u_t^{\text{cond}}(x_t \mid x_1)$
  $\ell(\theta) \leftarrow \langle \Delta, \text{sg}(V_{x,t}(v_t^\theta(x_t)))^{-1}\,\Delta\rangle_g$
  $\theta \leftarrow \text{optimizer_step}(\ell(\theta))$
end while

Figure 6: Side-by-side training procedures for RFM and our method on simple geometries. The figure highlights the differing steps in our method in colour. We omit the factor of $\frac{1}{2}$ from our loss to keep it on the same scale as RFM.

C.3 Implementation on less symmetric manifolds

In principle, the usual steps of our method work on less symmetric manifolds—compute the mean and covariance of the conditional velocity under the endpoint family, numerically invert the mean map, and cache the resulting functions. Instead of quadrature, Monte Carlo could be feasible; the structure of the exponential family lets us express these quantities as expectations. Computing them offline and interpolating seems the most plausible route. Neural or piecewise optimization is plausibly most useful in either inverting the mean map, or selecting good choices of nodes and weights for quadrature. Generally, if there is a particular high-dimensional manifold supporting several datasets of interest, it would be prudent to learn the function approximations once, reusing them across datasets.

D Specialization to common geometries

The surrogate loss of our method is preconditioned by the covariance $V_{x,t}(v)$ of the conditional velocity. For the purposes of this section, let $v = v_t^\theta(x_t)$ and $\eta = m_{x,t}^{-1}(v)$ so that $V_{x,t}(v) =$

$\nabla^2_\eta A_{x,t}(\eta)$. It will also be particularly convenient to adopt the definition

$$\Delta = v - u_t^{\text{cond}}(x_t \mid x_1). \tag{52}$$

The range of geodesic velocities that are feasible and the Riemannian volume measure determine the log-partition function. In each case considered here, we find that symmetries of the manifold reduce the covariance to a small number of scalar functions. This makes offline pretabulation tractable.

D.1 Flat torus $\mathbb{T}^d$

As each coordinate of $\mathbb{T}^d$ satisfies $[\log_{x_t}(x_1)]_i \in [-\pi, \pi)$, we have $[u_t^{\text{cond}}]_i \in \left[-\frac{\pi}{1-t}, \frac{\pi}{1-t}\right)$. Since the metric is flat, the map from endpoints to velocities has constant Jacobian. Hence, the likelihood is a product of $\mathcal{N}(\eta_i, 1)$ distributions truncated to $\left[-\frac{\pi}{1-t}, \frac{\pi}{1-t}\right)$. Now, let $z_{\pm,i} = \eta_i \pm \frac{\pi}{1-t}$. With ϕ and Φ as the standard normal density and CDF respectively, the log-partition function separates across coordinates into

$$A_{x,t}(\eta) = \sum_{i=1}^{d} \log \int_{-\pi/(1-t)}^{\pi/(1-t)} e^{\eta_i u - u^2/2} \, \mathrm{d}u + \text{terms independent of } \eta \tag{53}$$

$$= \sum_{i=1}^{d} \left[\frac{\eta_i^2}{2} + \log\left(\Phi(z_{+,i}) - \Phi(z_{-,i})\right) \right] + \text{terms independent of } \eta. \tag{54}$$

The mean velocity is

$$v_i = \partial_{\eta_i} A_{x,t}(\eta) \tag{55}$$

$$= \eta_i + \frac{\phi(z_{+,i}) - \phi(z_{-,i})}{\Phi(z_{+,i}) - \Phi(z_{-,i})}. \tag{56}$$

The covariance is diagonal, with

$$[V_{x,t}(v)]_{ii} = \partial^2_{\eta_i} A_{x,t}(\eta) \tag{57}$$

$$= 1 + \frac{z_{-,i}\phi(z_{-,i}) - z_{+,i}\phi(z_{+,i})}{\Phi(z_{+,i}) - \Phi(z_{-,i})} - (v_i - \eta_i)^2. \tag{58}$$

Hence, we need just one scalar coefficient per coordinate to form the surrogate loss,

$$\ell = \frac{1}{2} \sum_{i=1}^{d} \frac{\Delta_i^2}{\operatorname{sg}([V_{x,t}(v)]_{ii})}. \tag{59}$$

As $t \to 1$, the interval $\left[-\frac{\pi}{1-t}, \frac{\pi}{1-t}\right)$ expands to all of $\mathbb{R}$, meaning $[V_{x,t}(v)]_{ii}$ tends to one. As a result, the loss approaches $\frac{1}{2}\|\Delta\|_g^2$, which is that of RFM up to a factor of $\frac{1}{2}$.

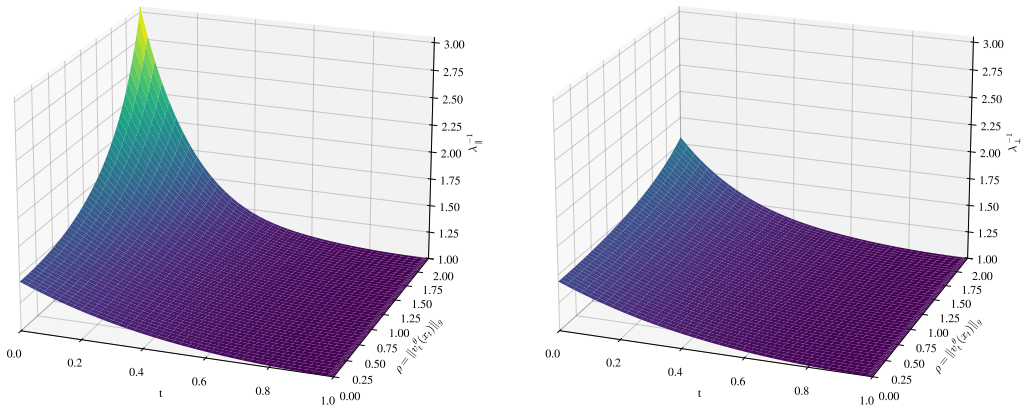
D.2 Complete simply connected manifolds of constant curvature

Here, let $\mathcal{M}$ be a complete simply connected d-dimensional manifold of constant sectional curvature κ where $d \geq 2$. Rotating around x_t preserves both distance and volume. Hence, the log-partition function depends on η only via $s = \|\eta\|_g$. To emphasize this radial dependence, we write this function as $A_t = A_t(s)$. The prediction v of the velocity is parallel to η and $\|v\|_g = A_t'(s)$. As $\nabla_\eta A_{x,t}(\eta) = A_t'(s)\frac{\eta}{s}$, the covariance is

$$V_{x,t}(v) = \nabla_\eta \left(A_t'(s) \frac{\eta}{s} \right) \tag{60}$$

$$= A_t''(s) \frac{\eta \otimes \eta}{s^2} + \frac{A_t'(s)}{s} \left(I - \frac{\eta \otimes \eta}{s^2} \right) \tag{61}$$

$$= \frac{A_t'(s)}{s} I + \left(A_t''(s) - \frac{A_t'(s)}{s} \right) \frac{v \otimes v}{\|v\|_g^2}. \tag{62}$$



(a) Weight parallel to v.

(b) Weight perpendicular to v.

Figure 7: Weighting of the surrogate loss on $\mathbb{S}^2$. The vertical axes show $\lambda_{\parallel}^{-1} = \frac{1}{A_t''(s)}$ and $\lambda_{\perp}^{-1} = \frac{s}{A_t'(s)}$, respectively. Errors parallel to the predicted velocity are given more weight than those perpendicular, but both tend to one as $t \to 1$.

At $v = 0$, the continuous extension is $V_{x,t}(0) = A_t''(0)I$; a directional decomposition is unneeded. Now, let $\Delta_{\parallel}$ and $\Delta_{\perp}$ be the components of Δ which are, respectively, parallel and perpendicular to v. Then, the surrogate loss is

$$\ell = \frac{1}{2}\left[\frac{\|\Delta_{\parallel}\|_g^2}{\mathrm{sg}(A_t''(s))} + \frac{\|\Delta_{\perp}\|_g^2}{\mathrm{sg}\left(\frac{A_t'(s)}{s}\right)}\right]. \tag{63}$$

We find that the loss weights errors parallel and perpendicular to the predicted velocity differently. As shown in fig. 7, the parallel component has greater weight than the perpendicular one on $\mathbb{S}^2$. This reflects the intuition that, when curvature is positive, geodesic convergence means that errors in the magnitude of the predicted velocity matter more than those in the direction.

Radial integration. We evaluate $A_t(s)$ in polar coordinates around x_t. Let $r = \frac{\mathrm{dist}_g(x_t, x_1)}{1-t}$ and ϑ be the angle between η and u_t^{cond}. The radial part of the Riemannian volume measure is (up to a factor independent of s)

$$J_t(r) = \begin{cases} \left(\frac{\sin((1-t)\sqrt{\kappa}\, r)}{(1-t)\sqrt{\kappa}}\right)^{d-1}, & \kappa > 0, \\ r^{d-1}, & \kappa = 0, \\ \left(\frac{\sinh((1-t)\sqrt{-\kappa}\, r)}{(1-t)\sqrt{-\kappa}}\right)^{d-1}, & \kappa < 0. \end{cases} \tag{64}$$

For a conditional velocity vector of magnitude r making an angle ϑ with η, the exponent in the likelihood is $-\frac{1}{2}r^2 + sr\cos\vartheta$. Over the sphere of possible directions, the area between angles ϑ and $\vartheta + \mathrm{d}\vartheta$ is proportional to $\sin^{d-2}\vartheta\,\mathrm{d}\vartheta$. Together with the radial factor in the volume measure $J_t(r)$, this means the log-partition function reduces to

$$A_t(s) = \log \int_0^{\frac{\mathrm{diameter}(\mathcal{M})}{1-t}} e^{-\frac{1}{2}r^2} J_t(r) \int_0^{\pi} e^{sr\cos\vartheta} \sin^{d-2}\vartheta\,\mathrm{d}\vartheta\,\mathrm{d}r + \text{terms independent of } s. \tag{65}$$

Computation of the derivatives. Employing the modified Bessel function I_ν of the first kind, we write the integral over ϑ as

$$\int_0^{\pi} e^{sr\cos\vartheta} \sin^{d-2}\vartheta\,\mathrm{d}\vartheta = \sqrt{\pi}\,\Gamma\left(\frac{d-1}{2}\right)\left(\frac{2}{sr}\right)^{\frac{d}{2}-1} I_{\frac{d}{2}-1}(sr). \tag{66}$$

For $\nu = \frac{d}{2} - 1$, the derivatives of the Bessel term satisfy the identities

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(z^{-\nu} I_\nu(z)\right) = z^{-\nu} I_{\nu+1}(z), \tag{67}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}z^2}\left(z^{-\nu} I_\nu(z)\right) = z^{-\nu}\left(I_{\nu+2}(z) + \frac{I_{\nu+1}(z)}{z}\right). \tag{68}$$

Differentiating eq. (65) under the integral sign enables the calculation of both $A_t'(s)$ and $A_t''(s)$ via quadrature in just one dimension. We tabulate these and the inverse of the map $s \mapsto A_t'(s)$ prior to the training loop. For numerical stability, we use exponentially scaled Bessel functions for large arguments and a series expansion near zero. We accumulate the quadrature sums via log-sum-exp.

Euclidean limit. As $t \to 1$, the radial factor in the volume measure converges as $J_t(r) \to r^{d-1}$. Hence, the integral tends to its Euclidean form for which

$$A_t(s) \longrightarrow \frac{s^2}{2} + \text{terms independent of } s. \tag{69}$$

As a result, $A_t'(s) \to s$ and $A_t''(s) \to 1$. Both weights in the loss tend to one, meaning $\ell \to \frac{1}{2}\|\Delta\|_g^2$. Up to a factor of $\frac{1}{2}$, this again recovers the loss of RFM.

D.3 Rotation groups and products $\mathrm{SO}(3)^n$

On $\mathrm{SO}(3)$, we use the canonical bi-invariant metric where the distance from the identity to a rotation is given by its angle. Let a tangent vector at R be denoted $R\widehat{u}$, where $\widehat{u}\, w = u \times w$. Identifying $R\widehat{u} \leftrightarrow u \in \mathbb{R}^3$ is isometric as

$$g_R(R\widehat{u}, R\widehat{w}) = u^\top w. \tag{70}$$

Under this metric, $\mathrm{SO}(3)$ has sectional curvature $\kappa = \frac{1}{4}$. The manifold is of constant curvature, albeit not simply connected. The largest possible geodesic distance between two rotations is π, meaning the conditional velocity satisfies $\|u_t^{\text{cond}}\|_g \leq \frac{\pi}{1-t}$.

Radial integration. As before, rotational symmetry means the log-partition function depends on η only via $s = \|\eta\|_g$. Let $A_t = A_t(s)$ denote this function (for a single rotation factor) and let r be the magnitude of the conditional velocity, corresponding to a rotation angle of $(1-t)r$. The radial part of the volume measure is proportional to $\sin^2(\frac{1-t}{2}r)$. In order to integrate over directions, let ϑ be the angle between η and the conditional velocity. In three dimensions, the average over directions is

$$\frac{1}{2}\int_0^\pi e^{sr\cos\vartheta}\sin\vartheta\,\mathrm{d}\vartheta = \frac{1}{2}\int_{-1}^1 e^{srz}\,\mathrm{d}z \tag{71}$$

$$= \frac{\sinh(sr)}{sr}. \tag{72}$$

These together reduce the log-partition function to

$$A_t(s) = \log\int_0^{\frac{\pi}{1-t}} e^{-r^2/2}\frac{\sinh(sr)}{sr}\sin^2\left(\frac{1-t}{2}r\right)\mathrm{d}r + \text{terms independent of } s. \tag{73}$$

To evaluate the covariance, we use some helpful derivatives,

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{\sinh z}{z}\right) = \frac{z\cosh z - \sinh z}{z^2}, \tag{74}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}z^2}\left(\frac{\sinh z}{z}\right) = \frac{(z^2+2)\sinh z - 2z\cosh z}{z^3}. \tag{75}$$

Differentiating (under the integral sign) lets us evaluate $A_t'(s)$ and $A_t''(s)$ via quadrature in one dimension. For numerical stability, we use series expansions near zero, scaled hyperbolic functions for large arguments, and log-sum-exp for the quadrature sums.

Product structure. Here, n is the number of rotation factors, so the intrinsic dimension is $d = 3n$. On $\mathrm{SO}(3)^n$, the metric is the sum of the metrics of the individual rotation groups. As both the likelihood and the volume measure factorize across rotation factors,

$$A_{R,t}(\eta) = \sum_{j=1}^{n} A_t(s_j), \tag{76}$$

where $s_j = \|\eta^{(j)}\|_g$ is the norm of the natural parameter for the jth rotation factor. The covariance has a 3×3 block per rotation factor. Each of these takes the form in eq. (62), where $A'_t(s_j) = \|v^{(j)}\|_g$ determines the s_j. Let $\Delta_{\parallel}^{(j)}$ and $\Delta_{\perp}^{(j)}$ be the components of the error in the jth factor parallel and perpendicular to $v^{(j)}$, respectively. Then, by eq. (63), the surrogate loss is

$$\ell = \frac{1}{2}\sum_{j=1}^{n}\left[\frac{\|\Delta_{\parallel}^{(j)}\|_g^2}{\mathrm{sg}(A''_t(s_j))} + \frac{\|\Delta_{\perp}^{(j)}\|_g^2}{\mathrm{sg}\left(\frac{A'_t(s_j)}{s_j}\right)}\right]. \tag{77}$$

The network uses all n rotation factors to predict the velocity in each factor. As $t \to 1$, both weights tend to one. Up to a factor of $\frac{1}{2}$, this again recovers the RFM loss. Due to the product structure and identical factors, we pretabulate just the two scalar functions, reusing them across all rotation factors.

E Experimental methodology and evaluation

E.1 Evaluation framework

Accurate measurement of generative performance on manifolds is important, but many tools from the Euclidean setting are not immediately to hand. Here we discuss validity of common evaluation metrics and lay out our choices.

Wasserstein distances. Wasserstein distances are a natural way to measure generative performance [Villani, 2009] but are costly to compute. Sliced Wasserstein mitigates this in Euclidean space by reducing optimal transport via projection, but generalizing requires injectivity of the Radon transform [Kolouri et al., 2019]. This is automatic in Euclidean space but for manifolds it requires choosing a family of projections. Hence, attempts to apply sliced Wasserstein to manifolds introduce additional geometric structure. For example, Bonet et al. [2025] propose global projections on Cartan–Hadamard manifolds, where geodesic structure is well behaved, while Rustamov and Majumdar [2023] suggest spectral slicing based on Laplace–Beltrami eigenfunctions. Their metric guarantees rely on the chosen family of projections being separating. In practice, the implementations use approximations via truncating to a finite number of terms.

Kernel methods. Methods based on kernels can work well, but their validity depends on RKHS conditions. The energy distance [ED; Székely and Rizzo, 2013] is a nonparametric discrepancy requiring the space to have strong negative type to be a valid metric. On $\mathbb{T}^1$, geodesic distance has negative type but not strong negative type [Lyons, 2020]. We can convert to strong negative type by square root [Lyons, 2013, Corollary 3.18], setting $d_{\mathrm{ED}}(x, y) = \sqrt{\mathrm{dist}_g(x, y)}$ in the ED formula. Kokkendorff [2002] proves that compact Riemannian manifolds of negative type and dimension ≥ 2 are simply connected, so geodesic distance on $\mathbb{T}^d$ is not of negative type and the same trick does not work for $d \geq 2$. Maximum mean discrepancy [MMD; Gretton et al., 2012] is a related discrepancy defined often with a Gaussian kernel parametrized by its bandwidth. Feragen et al. [2015] show the Riemannian generalization of this kernel is not positive definite on manifolds, suggesting the Laplace kernel specifically for $\mathbb{S}^d$ (which includes $\mathbb{T}^1$). In our experiments, we use the median heuristic to set the Laplace kernel bandwidth.

Our choices. Based on the above, we report Wasserstein distances for our analytic test case, where the data distribution is an equally weighted mixture of two Diracs, simplifying optimal transport (alongside the pedagogical reasons, we chose the singular form of the distributions precisely so that Wasserstein is feasible), alongside the unbiased forms of energy distance with the square-root trick and MMD with the Riemannian Laplace kernel. For the real-world geospatial benchmarks, we

found the small size of the test set commonly resulted in negative values for the unbiased kernel methods despite their population validity. As these are well-established benchmarks in the literature with many prior reported results, we think it reasonable to only report NLL. However, this means we cannot compare to RG-VFM as its NLLs are uninformative; see appendix B.2 for further discussion. As Zaghen et al. [2026] do not report NLLs in their experiments, we were unable to validate our replication directly according to this metric.

E.2 Evaluating on tori in high dimensions via an MMD with Fourier kernel

For high-dimensional product tori, we report NLL so as to compare to prior work, but felt it was necessary to supplement this with a more robust metric, one reason being for a fairer and clearer comparison to RG-VFM. However, the Riemannian energy distance and Laplace-kernel MMD are neither metrics nor pseudometrics for $d > 1$. Here, we construct an MMD pseudometric with a Fourier kernel designed specifically for the wrapped Gaussian target of our experiment. This uses the known closed form of the characteristic function to calculate the RKHS representation of the data distribution without estimation, thus eliminating this source of experimental uncertainty.

We first note that a translation-invariant kernel on $\mathbb{T}^d$ is characteristic if and only if its Fourier coefficients are positive on the whole Fourier spectrum $\mathbb{Z}^d$ [Fukumizu et al., 2008, Sriperumbudur et al., 2010]. We use the Fourier kernel

$$k(x, y) = \sum_{m \in \mathbb{Z}^d} a_m e^{\mathrm{i} m^\top (x-y)}, \tag{78}$$

with $a_m \geq 0$, $\sum_m a_m < \infty$, and wavevector m for which we choose $a_m = \exp(-\lambda_{\mathrm{MMD}} \|m\|_2^2)$ with parameter λ_{MMD}. The associated MMD is the RKHS distance between mean embeddings [Gretton et al., 2012],

$$\mathrm{MMD}^2 = \sum_{m \in \mathbb{Z}^d} a_m \left| \hat{P}(m) - \hat{Q}(m) \right|^2, \tag{79}$$

where $\hat{Q}(m) = \frac{1}{N} \sum_{i=1}^N e^{\mathrm{i} m^\top x_i}$ is the empirical Fourier embedding of generated samples $x_1, x_2, \ldots, x_N \sim p_1^\theta$. In the case of wrapped Gaussian data $x_1 \sim \mathcal{WN}(\mu, \Sigma)$, we use the closed-form expression for the Fourier embedding of the data distribution, $\hat{P}(m) = \mathbb{E}_{p_1}[e^{\mathrm{i} m^\top X}] = \exp\left(\mathrm{i} m^\top \mu - \frac{1}{2} m^\top \Sigma m\right)$. In practice, we truncate to a finite set of frequencies $\{m : m \in \Omega\}$ for some $\Omega \subseteq \mathbb{Z}^d$. Then the empirical MMD^2 V-statistic at parameter λ_{MMD} is

$$\widehat{\mathrm{MMD}}^2_{\lambda_{\mathrm{MMD}}} = \sum_{m \in \Omega} e^{-\lambda_{\mathrm{MMD}} \|m\|_2^2} \left| e^{\mathrm{i} m^\top \mu - \frac{1}{2} m^\top \Sigma m} - \frac{1}{N} \sum_{i=1}^N e^{\mathrm{i} m^\top x_i} \right|^2. \tag{80}$$

We use the set of wavevectors $\Omega = \{w e_i : i \in \{1, \ldots, d\}; w \in \{-10, -9, \ldots, 9, 10\}\} \setminus \{0\}$ with e_i the ith standard basis vector, and average over a grid centred at $\lambda_{\mathrm{MMD}} = \frac{1}{2}\sigma^2 = 0.02$, namely $\lambda_{\mathrm{MMD}} \in \Lambda_{\mathrm{MMD}} = \{0.005, 0.01, 0.02, 0.04, 0.08\}$ to get a more robust estimate of performance giving

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{|\Lambda_{\mathrm{MMD}}|} \sum_{\lambda_{\mathrm{MMD}} \in \Lambda_{\mathrm{MMD}}} \frac{1}{\alpha(\lambda_{\mathrm{MMD}})} \widehat{\mathrm{MMD}}^2_{\lambda_{\mathrm{MMD}}}, \tag{81}$$

where $\alpha(\lambda_{\mathrm{MMD}}) = \sum_{m \in \Omega} e^{-\lambda_{\mathrm{MMD}} \|m\|_2^2}$ normalizes for total Fourier weight. The result is a scalable squared pseudometric of finite resolution. As it detects discrepancies only in the marginal distributions of individual coordinates, it may miss pairwise interactions at high dimensions. Consideration of all pairs of basis vectors is infeasible at 400 dimensions and the data distribution is in any case isotropic. At low dimensions we found that pairs and higher-order interactions had negligible contribution to the MMD, especially compared with the effect of varying the bandwidth λ_{MMD}.

E.3 Protocols for training and evaluation

The analytic test case, torus sweep, and geospatial experiments use a single NVIDIA A100-SXM GPU with 80 GB of memory. The efficiency studies use the hardware specified in appendix F. We follow the RFM setup published by Chen and Lipman [2024] and extend its codebase to support

our method and additional evaluation metrics. This splits the data into $80\,\%$ train, $10\,\%$ validation, and $10\,\%$ test, giving us the same data splits as in the RFM paper. We use seed values 0–29, 0–4, and 0–4 for our training runs on the analytic test case, high-dimensional torus, and geospatial datasets respectively. Our backbone is a standard multilayer perceptron with time concatenated to the input. We use Swish [Ramachandran et al., 2018] with a learnable parameter and Adam with an exponential moving average [Polyak and Juditsky, 1992] (decay 0.999). For synthetic datasets, we train for $50\,000$ iterations with batch size 512 and 3 hidden layers of width 512. For the geospatial datasets, we train for $1\,000\,000$ iterations with batch size and hidden width following Chen and Lipman [2024]. The analytic test case uses dopri5 as the ODE solver with tolerance 1×10^{-5} while the high-dimensional torus uses $1\,000$ Euler steps with projection at each step for scalability following Chen and Lipman [2024]. In the latter case it was not feasible to use dopri5 (even if desired) as RG-VFM did not complete in a reasonable time beyond low dimensions. At these low dimensions we observe that the Euler solver appears most favourable for RG-VFM; this is also the solver used by Zaghen et al. [2026]. Evaluation on the geospatial datasets uses dopri5 with tolerance 1×10^{-7} also following Chen and Lipman [2024]. We modify the RFM codebase to `torch.compile` without graph breaks, training in default mode for these experiments. We disable compilation for the runtime benchmark in appendix F.2. In appendix F.3, we compare execution in both modes, eager and compiled. We are aware that RG-VFM is sensitive to solver choice and tolerance, and we investigate a variety of solvers and tolerances in appendix B.2.

E.4 Training and evaluation on human pose data

We use human pose data expressed in $\mathrm{SO}(3)^{19}$, following Dünkel et al. [2024], who train their normalizing flow model, HuProSO3, on a partition of the AMASS dataset collection [Mahmood et al., 2019], namely: CMU [Carnegie Mellon University], PosePrior [Akhter and Black, 2015], TotalCapture [Trumble et al., 2017], EyesJapanDataset [Eyes JAPAN Co. Ltd.], KIT and EKUT [Mandery et al., 2015, 2016, Krebs et al., 2021], BMLrub [Troje, 2002], TCDHands [Hoyet et al., 2012], and ACCAD [Advanced Computing Center for the Arts and Design]. Whereas Dünkel et al. [2024] are concerned with cross-dataset performance—the unconditional HuProSO3 is trained on the whole of the above partition and tested on an aggregate of remaining datasets from AMASS—we are concerned with in-distribution performance. We train RFM, RG-VFM, and our method with an 80/10/10 training/validation/test split of this partition, resulting in $11\,823\,971$ training samples. Using an MLP of 6 layers and width $1\,024$, we train for $150\,000$ steps.

We report NLL with respect to the Haar measure. For both energy distance and Gaussian MMD, we evaluate via the chordal distance (i.e., implicitly embedding into $\mathbb{R}^{171}$). The embedding into Euclidean space is injective, so these metrics are valid for distributions of poses. The Gaussian MMD metric uses the median heuristic for the bandwidth, calibrated uniformly using the test set.

F Computational performance

F.1 Timing dopri5 sampling on $\mathrm{SO}(3)^{19}$

Sampling for our human pose experimental results in table 5 used Euler integration with $1\,000$ steps (following the original RG-VFM implementation and for the reasons we outline in appendix B.2). Here, we also report the NFE and wall time required for sampling with dopri5 (table 8). Our method only differs from RFM in the backward pass, while the forward pass is through an MLP of the same architecture. Therefore, under Euler sampling they have the same sampling cost, while under adaptive sampling the sampling time depends on the neural approximation learned.

In many cases, flow matching under independent coupling has optimal solutions exhibiting significant spatiotemporal variation (motivating ideas around rectification and choosing better couplings). For methods possessing compatibility, we posit that their relative ordering measured by adaptive solver NFE depends on the true marginal field induced by the data. Notably, at a tolerance of 10^{-5}, the wall times for our method and RFM are comparable to that of a forward pass through HuProSO3's autoregressive normalizing flow. Although HuProSO3 does not require numerical integration, it has a relatively heavier neural architecture and its authors note they have not optimized all potential parallelism [Dünkel et al., 2024, §B.1].

Table 8: Sampling 20 000 poses. Mean±std, 5 runs after warm-up, using distinct trained checkpoints except for HuProSO3 (`id_0900`). Flow matching models use adaptive dopri5; HuProSO3 uses its autoregressive forward pass.

Method	Tolerance	**Time ↓** (s)	**NFE ↓**
HuProSO3	—	5.465±0.025	—
RFM	10^{-5}	4.153±0.132	550.4±17.3
RG-VFM	10^{-5}	57.456±1.581	1 212.0±30.2
Our method	10^{-5}	4.214±0.047	558.8±6.6
RFM	10^{-7}	9.479±0.279	1 254.8±37.1
RG-VFM	10^{-7}	118.052±17.562	2 584.0±385.6
Our method	10^{-7}	10.127±0.355	1 341.2±46.6

Table 9: Training iterations per second (it/s)↑. Mean±std, 5 runs. `torch.compile` disabled. We use the same full training loop for all methods and only change the loss. Thus this data does not reflect the high inference time of RG-VFM.

Method	$\mathbb{T}^2$	$\mathbb{T}^{500}$	$\mathbb{S}^2$	$\mathbb{S}^{500}$
RFM	**48.33**±1.11	**15.32**±0.08	**46.35**±0.65	**14.62**±0.04
RG-VFM	46.44±0.19	14.82±0.02	44.50±0.29	14.25±0.03
Our method	40.51±0.81	14.58±0.11	34.30±0.50	12.98±0.05

F.2 Throughput of training without compilation

We additionally benchmark the runtime of the full training loop in the same pipeline and report iterations per second (it/s). The results are in table 9. We measure wall time from the start to the end of each run. Across methods, we keep the model architecture and execution setup fixed and change only the training loss. The hardware is an NVIDIA GeForce RTX 3090 (24 GB) with an AMD Ryzen 9 5950X 16-Core CPU. All methods use the same architecture and run configuration: a three-layer MLP with width 512, batch size 512, with 1 000 training steps per run and 5 repetitions with `torch.compile` turned off. We disable the evaluation of validation NLL during this benchmark; differences in runtime reflect the cost of pretabulation, table lookups and training iterations. The reduction in throughput with our method is largest at low dimensions and shrinks at high dimensions. This is consistent with the $V_{x,t}$ lookup being a small per-sample cost relative to the initial pretabulation, reducing at high dimensions as other costs increase.

F.3 Wall time and memory use

In table 10, we report wall time and memory use for $\mathbb{T}^5$ and $\mathbb{S}^2$ in eager and compiled execution.

Timing methodology:

- We use eager execution and compilation via `torch.compile` for $\mathbb{T}^5$ and $\mathbb{S}^2$ (`wrapped_torus` and `flood` configs respectively).
- Each run has 40 steps for warm-up followed by 1 000 timed steps.
- After synchronizing CUDA, we use `time.perf_counter()` to time the construction of `ManifoldFMLitModule(cfg)`.
- We call `torch.cuda.synchronize()` at the end of each timed batch. This ensures that timing measures completed GPU work (not async CPU dispatch).

Memory methodology:

- We initialize the CUDA context with `torch.zeros(1, device="cuda:0")` first and then reset the peak memory statistics tracked by the allocator.
- CPK = peak memory recorded by the allocator during construction, i.e. `torch.cuda.max_memory_allocated()` at `on_train_start`.

- CLIVE = live memory recorded by the allocator after construction, i.e. `torch.cuda.memory_allocated()`, at `on_train_start`.
- CDRV = memory usage reported by the driver after construction, i.e. `(total - free)` from `torch.cuda.mem_get_info()` at `on_train_start`.
- MPK = peak memory recorded by the allocator during training, i.e. `torch.cuda.max_memory_allocated()` over the timed window after the reset following warm-up.
- MDRV = maximum memory usage reported by the driver during training, i.e. `total - min(free)` from `torch.cuda.mem_get_info()` over samples taken at the end of each batch timed, after synchronization.

We give both allocator and driver metrics as the former are only exact at byte level for live allocations tracked by PyTorch. (They omit memory pools for CUDA graphs and direct cuSOLVER-/cuBLAS or other CUDA allocations. Hence, they report apparently lower memory usage under `torch.compile`.) The latter also capture memory usage across the device. Measurements were taken on an AMD Ryzen Threadripper and NVIDIA RTX A6000. The allocator memory overhead is just a few MiB. We find it of particular note that compilation via `torch.compile` makes the runtime overhead per iteration *near negligible*.

Table 10: Wall time for construction, wall time per training iteration, and memory use on $\mathbb{T}^5$ and $\mathbb{S}^2$. Mean±std, 5 runs. We omit standard deviation where all runs have identical values. Memory identifiers as defined above.

			Time		Memory				
	Mode	Method	**Constructor ↓** (s)	**Step ↓** (ms)	**CPK ↓** (MiB)	**CLIVE ↓** (MiB)	**CDRV ↓** (MiB)	**MPK ↓** (MiB)	**MDRV ↓** (MiB)
$\mathbb{T}^5$	Eager	RFM	0.0146±0.0015	3.78±0.14	3.80	3.80	418	54.206	492
	Eager	RG-VFM	0.0155±0.0014	3.91±0.10	3.80	3.80	418	54.206	492
	Eager	Ours	0.0324±0.0002	6.50±0.07	4.84	4.84	420	55.268	494
	Compiled	RFM	0.7831±0.0070	2.30±0.06	21.08	11.93	508	7.717±0.005	492
	Compiled	RG-VFM	0.7701±0.0077	2.27±0.05	21.08	11.93	508	7.721±0.010	492
	Compiled	Ours	0.7915±0.0138	2.32±0.06	22.11	12.97	510	8.782	496
$\mathbb{S}^2$	Eager	RFM	0.2095±0.0062	91.82±0.48	166.39	166.39	582	4 493.520	5 112
	Eager	RG-VFM	0.2062±0.0068	92.13±0.13	166.39	166.39	582	4 493.520	5 110
	Eager	Ours	1.9158±0.0540	93.72±0.08	172.03	172.03	602	4 499.190	5 112
	Compiled	RFM	0.8567±0.0120	65.93±0.32	2 782.98	174.52	3 268	332.999	3 414
	Compiled	RG-VFM	0.8604±0.0048	66.52±0.38	2 782.98	174.52	3 266	332.999	3 412
	Compiled	Ours	2.5414±0.0509	65.63±0.36	2 788.62	180.16	3 268	338.669	3 436

F.4 Compute resources

For the $\mathbb{T}^1$ experiment, the 30 seeds complete in under $2\,\mathrm{h}$ across all three trained methods, as each run finishes in under $1\,\mathrm{min}$. On product tori in high dimensions, training takes approximately $1\,\mathrm{min}$ per run. We find this to be essentially independent of dimension. Evaluation of test/NLL becomes more expensive as d increases. This brings the total time per run from a few minutes at low d to around $1\,\mathrm{h}$ at $d = 400$. The full sweep over 5 seeds, 3 methods, and all dimensions that we report consumes roughly $30\,\mathrm{h}$ of A100 time. On the geospatial datasets, a single run of training and evaluation completes in under $5\,\mathrm{h}$. For one method on one dataset, the sweep over 5 seeds takes under $24\,\mathrm{h}$. Across the four datasets and the two methods we train ourselves (our method and our reproduction of RFM), this totals roughly $200\,\mathrm{h}$ of A100 time. The analytic test case, experiments on product tori in high dimensions, and geospatial experiments together account for approximately $250\,\mathrm{h}$ of A100 time. The training benchmark in appendix F.2 contributes a negligible amount of extra compute. Some additional compute was used for preliminary experiments, including the solver investigations in appendix B.2, as well as early prototypes.

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: See sections 3 to 5 for the supporting results.

   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: See section 6.

   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: Assumptions are stated with the results. Proofs and derivations are given in appendices A and D.

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: See appendices C and E for implementation details and experimental protocols.

   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: The supplementary material includes code and reproduction instructions. Dataset sources are given in appendix E.

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: See appendix E.

   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 report variability across seeds. The table and figure captions define the summary statistics.

   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: See appendices E and F for hardware, runtime measurements and compute requirements.

   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: We have reviewed the NeurIPS Code of Ethics and conducted the research in accordance with it.

   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: This methodological work is not tied to a particular deployment.

    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: The paper does not release pretrained generative models, scraped datasets, or any assets with high misuse potential.

    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: The RFM code license and the access terms for the geospatial datasets are recorded in the bibliography and accompanying `DATA_LICENSES.md` in the supplementary material.

    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: [Yes]

    Justification: The supplementary material includes a README with installation and reproduction instructions, and `DATA_LICENSES.md` with data provenance and access terms.

    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: The paper does not involve crowdsourcing or research with human subjects.

    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: The paper does not involve research with human subjects.

    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: LLMs were used only for writing assistance (rephrasing, formatting, and copy-editing) and did not contribute to the core methodology, scientific rigor, or originality of the research.

    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.