RETHINKING RIEMANNIAN MEANFLOW: GEOMETRIC GAPS AND GEOMETRY-ALIGNED FLOW MAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

MeanFlow enables efficient one-step or few-step generation by learning average velocities over finite time intervals. Recent works have extended this principle to Riemannian manifolds for data with intrinsic non-Euclidean structure, such as rotations and directions. However, under curvature, the Euclidean relationships among average velocity, endpoint displacement, and generation error no longer directly hold. We revisit two formulations of Riemannian MeanFlow and identify two geometric gaps. First, parallel-transported path averages generally differ from endpoint log-displacements, potentially inducing one-step endpoint errors and flow-map composition defects. Second, even with endpoint-based representations, metric distortion through the exponential map causes tangent-space residuals and geodesic endpoint errors to weight directions differently. Motivated by these findings, we propose **Barycentric Semigroup Riemannian MeanFlow (RMF-BSC)**, which combines endpoint differential consistency with multi-split geodesic semigroup consistency to align training with endpoint geometry. We establish barycentric stability bounds and a variance-reduction result for multi-split averaging. Experiments validate the identified gaps and show that RMF-BSC achieves strong one-step performance across four diverse manifold benchmarks.

1 INTRODUCTION

Diffusion and flow-matching models have become prominent paradigms for generative modeling (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020; Albergo & Vanden-Eijnden, 2022; Lipman et al., 2022; Chen et al., 2026). However, high-quality generation typically requires many iterative updates, resulting in substantial inference cost. Recent work therefore pursues one-step or few-step generation through distillation, consistency learning, shortcut models, and flow-map learning (Salimans & Ho, 2022; Song et al., 2023; Frans et al., 2025; Boffi et al., 2026). MeanFlow further introduces finite-interval average velocity and an identity linking it to the instantaneous field, enabling one-step generation without a pretrained teacher (Geng et al., 2026a). Yet many structured data types, including rotations, directional variables, and molecular conformations, naturally reside on non-Euclidean spaces, motivating efficient generative modeling on Riemannian manifolds (De Bortoli et al., 2022; Chen & Lipman, 2024).

Recent work has extended diffusion, flow matching, and few-step generation to Riemannian manifolds (De Bortoli et al., 2022; Chen & Lipman, 2024; Cheng et al., 2026; Davis et al., 2026). In particular, recent Riemannian MeanFlow (RMF) methods develop two distinct finite-interval formulations. RMF-PT (Zhong et al., 2026) defines average velocity by parallel transporting (PT) instantaneous velocities along trajectories to a common tangent space, whereas Endpoint RMF (Woo et al., 2026) characterizes finite-time flow maps through endpoint log-displacements together with Eulerian, Lagrangian, and semigroup identities. These formulations raise two fundamental questions when moving beyond Euclidean space: *whether a transported path average faithfully represents the finite-time displacement required by endpoint sampling*, and *whether tangent-space regression faithfully reflects the geometry of the resulting endpoint error*.

Based on these questions, we revisit Riemannian MeanFlow and show that the two Euclidean identifications generally fail on Riemannian manifolds. First, the transported path average need not coincide with the endpoint log-displacement, giving rise to a *path–endpoint representation gap* that can induce one-step endpoint errors and flow-map composition defects. Endpoint RMF resolves this

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

representation mismatch by learning the sampler-exact endpoint log-displacement within a normal domain. However, our analysis characterizes a remaining *metric gap*: tangent-space regression and geodesic endpoint error induce different error weightings through the exponential map. Motivated by this remaining mismatch, we propose **Barycentric Semigroup Riemannian MeanFlow (RMF-BSC)**, which extends Endpoint RMF by combining differential learning with multi-split semigroup consistency measured directly in endpoint geometry. We provide a Fréchet-barycentric interpretation, establish stability bounds under suitable geometric assumptions, and show that multi-split averaging reduces split-induced gradient variance. Controlled geometric experiments validate the identified gaps and their sampling consequences, while generation experiments across four diverse manifold benchmarks demonstrate strong one-step performance of RMF-BSC.

Specifically, our main contributions are as follows:

(1) We characterize two geometric mismatches in existing RMF formulations: a path–endpoint representation gap and a tangent–endpoint metric gap. We relate the former to one-step and composition errors, and characterize the latter through the pullback metric of the exponential map.

(2) We propose RMF-BSC, a geometry-aligned extension of Endpoint RMF that combines endpoint-aligned differential learning with multi-split geodesic semigroup consistency, enforcing finite-time consistency directly in endpoint geometry with a Fréchet-barycentric interpretation.

(3) We validate both gaps through controlled geometric diagnostics and demonstrate strong one-step generation performance across four diverse manifold benchmarks, supporting the effectiveness of geometry-aligned endpoint learning.

2 Related Work

Flow matching and MeanFlow for fast generation. Flow matching learns generative dynamics by regressing instantaneous velocity fields along probability paths (Lipman et al., 2022), with related formulations including Rectified Flow (Liu, 2022) and stochastic interpolants (Albergo et al., 2025). To reduce multi-step inference cost, prior work has explored distillation (Salimans & Ho, 2022; Geng et al., 2023; Luo et al., 2023; Sauer et al., 2024; Liu et al., 2023) and consistency-based approaches (Song et al., 2023; Song & Dhariwal, 2024; Kim et al., 2024; Wang et al., 2024; Yang et al., 2024; Lu & Song, 2025; Geng et al., 2025; Peng et al., 2026), as well as shortcut models (Frans et al., 2025) and flow-map formulations (Boffi et al., 2026). More recently, MeanFlow learns finite-interval average velocities through an identity with the instantaneous field, enabling one-step and few-step generation without pretrained teachers or multi-stage distillation (Geng et al., 2026a; Zhang et al., 2026; Geng et al., 2026b). However, many structured data domains are naturally non-Euclidean, motivating extensions of these fast generative frameworks to Riemannian manifolds.

Riemannian flow matching and Riemannian MeanFlow. Generative modeling on manifolds has been studied through Riemannian score-based models (De Bortoli et al., 2022), Riemannian Flow Matching (Chen & Lipman, 2024), Riemannian consistency learning (Cheng et al., 2026), and Generalised Flow Maps (Davis et al., 2026). The latter extend Euclidean flow-map self-distillation to Riemannian manifolds and include a single-split geodesic semigroup objective, whereas RMF-BSC averages losses over multiple splits. Riemannian MeanFlow has developed along two distinct finite-interval formulations: Zhong et al. (2026) define average velocity via parallel transport along trajectories, whereas Woo et al. (2026) use endpoint log-displacements. These complementary formulations raise a fundamental question of whether path-based average velocities and endpoint-based finite-time displacements remain equivalent on curved manifolds, and whether their training objectives are aligned with the geometry of finite-time endpoint sampling and composition. This motivates our rethinking of Riemannian MeanFlow from both representation and metric perspectives.

3 Preliminaries

In this section, we briefly review two geometric formulations of finite-interval average velocity in Riemannian MeanFlow: the transported path-average formulation and the endpoint log-displacement formulation. These two viewpoints provide the basis for our subsequent analysis. Let $(\mathcal{M}, g)$ be a smooth Riemannian manifold.

Path / Transported Riemannian MeanFlow. Consider a smooth trajectory $\gamma : [r, t] \to \mathcal{M}$, with $x_\tau = \gamma(\tau)$ and velocity $\dot{\gamma}(\tau) \in T_{x_\tau}\mathcal{M}$. Since these velocities lie in different tangent spaces, they must first be transported to a common tangent space before averaging (Zhong et al., 2026). Let $\mathcal{P}^\gamma_{\tau\to t} : T_{x_\tau}\mathcal{M} \to T_{x_t}\mathcal{M}$ denote parallel transport along γ. The transported path-average velocity is

$$u^\gamma_{\mathrm{PT}}(x_t, r, t) = \frac{1}{t-r}\int_r^t \mathcal{P}^\gamma_{\tau\to t}\dot{\gamma}(\tau)\, d\tau \in T_{x_t}\mathcal{M}. \tag{1}$$

This quantity is path-dependent, as it depends on the trajectory $\gamma|_{[r,t]}$ rather than only its endpoints.

Endpoint Riemannian MeanFlow. A different formulation starts directly from the endpoint flow map (Woo et al., 2026). Let $\exp_x$ and $\log_x$ denote the Riemannian exponential and logarithm maps. We write $\mathcal{D}_x \subset T_x\mathcal{M}$ for a normal domain on which $\exp_x$ is a diffeomorphism onto $\mathcal{U}_x = \exp_x(\mathcal{D}_x)$, and denote by $\|\cdot\|_{g_x}$ the norm induced by g_x. If $x_r \in \mathcal{U}_{x_t}$, the endpoint log-displacement average velocity is

$$u_{\log}(x_t, x_r, r, t) = -\frac{1}{t-r}\log_{x_t}(x_r) \in T_{x_t}\mathcal{M}. \tag{2}$$

Unlike u^γ_{PT}, this quantity depends only on the endpoints (x_t, x_r) and is directly aligned with the one-step update

$$x_r = \exp_{x_t}\left(-(t-r)u_{\log}(x_t, x_r, r, t)\right).$$

In Euclidean space, the two notions agree: $u^\gamma_{\mathrm{PT}}(x_t, r, t) = u_{\log}(x_t, x_r, r, t) = \frac{x_t - x_r}{t-r}$. On a curved manifold, however, this equality is not a general identity. The transported average is tied to the path followed by the dynamics, whereas the endpoint log-displacement is tied to the geodesic chord used by the one-step sampler. The gap between these two objects is the starting point of our analysis.

4 Rethinking Riemannian MeanFlow

Euclidean MeanFlow relies on two identifications that generally fail on curved manifolds: the path-averaged velocity need not coincide with the finite-time endpoint displacement, and tangent-space error need not reflect endpoint sampling error. These mismatches give rise to a *representation gap* and a *metric gap*. We first characterize the representation gap and its consequences for one-step sampling and finite-time composition, and then analyze the remaining metric mismatch in endpoint-aligned RMF. Let $v_t(x) \in T_x\mathcal{M}$ denote the marginal velocity field, and let γ be one of its integral curves, with $x_\tau = \gamma(\tau)$ and $h = t - r > 0$. We denote by $\Phi_{r,t}$ the associated backward flow map, so that $\Phi_{r,t}(x_t) = x_r$. Geometric assumptions are given in Appendix A.

4.1 The Path–Endpoint Representation Gap

In this section, we first examine whether the transported average velocity used in RMF actually represents the finite-time displacement required for one-step sampling. Although u^γ_{PT} satisfies the MeanFlow differential identity, this alone does not guarantee that its exponential update reaches the true endpoint. The transported average u^γ_{PT} in Eq. (1) satisfies

$$u^\gamma_{\mathrm{PT}} = v_t(x_t) - h\frac{D}{dt}u^\gamma_{\mathrm{PT}}, \tag{3}$$

where $\frac{D}{dt}$ denotes the total covariant derivative along the trajectory γ. Although RMF-PT (Zhong et al., 2026) is trained with conditional velocities, its population target reduces to Eq. (3) for the marginal field and uniquely recovers u^γ_{PT} under the vanishing boundary condition; see Lemma A.1. However, this differential identity does not imply that $\exp_{x_t}(-hu^\gamma_{\mathrm{PT}}) = x_r$. The sampler-exact object is characterized below.

Proposition 4.1 (Local sampler exactness). *Suppose* $\exp_{x_t} : \mathcal{D}_{x_t} \to \mathcal{U}_{x_t}$ *is a diffeomorphism and* $x_r \in \mathcal{U}_{x_t}$. *For any candidate* $u \in T_{x_t}\mathcal{M}$ *with* $-hu \in \mathcal{D}_{x_t}$,

$$\exp_{x_t}(-hu) = x_r \quad\Longleftrightarrow\quad u = u_{\log}, \qquad \textit{where} \quad u_{\log} \quad \textit{is defined in Eq. (2).} \tag{4}$$

Proof. Due to the page limit, we defer the proof to Appendix B. □

Thus, Endpoint RMF (Woo et al., 2026) learns the unique sampler-exact displacement in the chosen normal domain. The discrepancy from the transported view is the path–endpoint representation gap

$$\Delta^{\gamma}_{r,t} = u^{\gamma}_{\mathrm{PT}} - u_{\log} = \frac{\log_{x_t}(x_r) - \mathsf{A}_{\alpha}(h)}{h}, \tag{5}$$

where $\alpha(a) = \gamma(t-a)$, $\mathcal{P}^{\alpha}_{a\to 0}$ denotes parallel transport along α from $\alpha(a)$ to $\alpha(0)$, and $\mathsf{A}_{\alpha}(h) = \int_0^h \mathcal{P}^{\alpha}_{a\to 0}\dot{\alpha}(a)\,da$ is the anti-development endpoint. The two objects agree in Euclidean space and along minimizing geodesic segments, but can differ on general trajectories. We illustrate this discrepancy with a simple example on the sphere.

A spherical example. On a semicircle of latitude at height $z = 1/2$ on $\mathbb{S}^2$, the anti-development displacement has norm $\sqrt{6}$, while the endpoint log-displacement has norm $2\pi/3$. Hence $u^{\gamma}_{\mathrm{PT}} \neq u_{\log}$, giving a concrete nonzero path–endpoint representation gap. Since both lie within the injectivity radius, their exponential updates differ, so the transported average does not recover the true endpoint in one step. The full calculation is in Appendix B. Importantly, marginal ordinary differential equation (ODE) trajectories need not remain geodesic even when conditional paths are. Further discussion is provided in Appendix A.1.

Therefore, the MeanFlow differential identity alone does not ensure one-step endpoint exactness on curved manifolds. Within a normal domain, the endpoint log-displacement is the unique sampler-exact representation, whereas the transported average may deviate from it along non-geodesic trajectories. This path–endpoint representation gap motivates the analysis of its effects on one-step sampling and finite-time composition.

4.2 Consequences for One-Step Sampling and Composition

We next examine how the representation gap affects one-step sampling and finite-time composition. Assuming the transported target is known exactly, define

$$\Psi^{\mathrm{PT}}_{r,t}(x_t) := \exp_{x_t}(-h u^{\gamma}_{\mathrm{PT}}) = \exp_{x_t}\big(\log_{x_t}(x_r) - h\Delta^{\gamma}_{r,t}\big)\,. \tag{6}$$

Let d_g denote the Riemannian geodesic distance induced by g, and define

$$E_{r,t}(x_t) := d_g\big(\Psi^{\mathrm{PT}}_{r,t}(x_t), \Phi_{r,t}(x_t)\big)\,,$$
$$\mathfrak{S}^{\mathrm{PT}}_{r,s,t}(x_t) := d_g\big(\Psi^{\mathrm{PT}}_{r,t}(x_t), \Psi^{\mathrm{PT}}_{r,s}\big(\Psi^{\mathrm{PT}}_{s,t}(x_t)\big)\big)\,.$$

Proposition 4.2 (Sampling error and composition defect)**.** *Suppose* $\exp_{x_t}$ *is* $L_{\exp}$*-Lipschitz on a region containing the true and PT displacements. For* $r < s < t$*, further suppose* $\Psi^{\mathrm{PT}}_{r,s}$ *is* $L^{\Psi}_{r,s}$*-Lipschitz on a region containing the true and predicted intermediate endpoints. With* $x_s = \Phi_{s,t}(x_t)$*,*

$$E_{r,t}(x_t) \le L_{\exp} h \|\Delta^{\gamma}_{r,t}\|_{g_{x_t}}\,, \tag{7}$$

$$\mathfrak{S}^{\mathrm{PT}}_{r,s,t}(x_t) \le E_{r,t}(x_t) + E_{r,s}(x_s) + L^{\Psi}_{r,s} E_{s,t}(x_t). \tag{8}$$

Proof. Due to the page limit, we defer the proof to Appendix C. □

Hence, the representation gap directly controls the one-step endpoint error and can further propagate into composition error. Although the composition bound is one-sided and errors may cancel across subintervals, differential consistency of the transported average alone does not guarantee finite-time endpoint consistency.

4.3 The Remaining Metric Gap in Endpoint RMF

We next examine whether learning the sampler-exact endpoint representation is sufficient to align training with one-step sampling on Riemannian manifolds. Although Endpoint RMF (Woo et al., 2026) removes the representation gap by learning $u_{\log}$ directly, its tangent-space regression need not reflect the geometry of the resulting endpoint error. For a base point $x \in \mathcal{M}$, a target displacement $\xi \in T_x\mathcal{M}$, and a perturbation $\eta \in T_x\mathcal{M}$, compare

$$\ell_{\tan} = \|\eta\|^2_{g_x}, \qquad \ell_{\mathrm{geo}} = d^2_g(\exp_x(\xi+\eta), \exp_x(\xi))\,.$$

The following proposition characterizes their local discrepancy through the pullback metric of the exponential map, making its dependence on the displacement and the perturbation direction explicit.

Proposition 4.3 (Endpoint metric distortion). *Let $\xi \in \mathcal{D}_x$. As $\|\eta\|_{g_x} \to 0$,*

$$\ell_{\mathrm{geo}} = G_{x,\xi}(\eta,\eta) + O\big(\|\eta\|_{g_x}^3\big), \qquad G_{x,\xi}(\eta,\eta) = \big\|d\exp_x|_\xi[\eta]\big\|^2_{g_{\exp_x(\xi)}}, \tag{9}$$

where $G_{x,\xi}$ is positive definite on $T_x\mathcal{M}$. In particular, for constant sectional curvature κ, writing $\rho = \|\xi\|_{g_x} > 0$ and decomposing $\eta = \eta_\parallel + \eta_\perp$ into components parallel and orthogonal to ξ,

$$G_{x,\xi}(\eta,\eta) = \|\eta_\parallel\|^2_{g_x} + \Big(\frac{S_\kappa(\rho)}{\rho}\Big)^2 \|\eta_\perp\|^2_{g_x}, \qquad S_\kappa(\rho) = \begin{cases} \sin(\sqrt{\kappa}\rho)/\sqrt{\kappa}, & \kappa > 0, \\ \rho, & \kappa = 0, \\ \sinh(\sqrt{-\kappa}\rho)/\sqrt{-\kappa}, & \kappa < 0. \end{cases} \tag{10}$$

Proof. Due to the page limit, we defer the proof to Appendix D. □

Thus, even with a sampler-exact endpoint representation, tangent-space regression need not be equivalent to geodesic endpoint error away from the solution. When ξ and $\xi + \eta$ lie in the same normal domain $\mathcal{D}_x$, both vanish only at $\eta = 0$, but weight nonzero errors differently. Radial errors keep unit weight, whereas transverse errors are increasingly amplified ($\kappa < 0$) or attenuated ($\kappa > 0$) as displacement grows. Since the sampler's local endpoint error is governed by $G_{x,\xi}$ rather than g_x, we measure consistency directly through geodesic endpoint error.

5 Barycentric Semigroup Riemannian MeanFlow

Building on the preceding analysis, we propose RMF-BSC, which extends endpoint-aligned RMF with geometry-aware finite-time consistency. Specifically, RMF-BSC preserves sampler-exact endpoint learning through the log-displacement identity (LDI) objective and introduces multi-split geodesic semigroup consistency to reduce split dependence and enforce intrinsic consensus across flow decompositions.

5.1 Endpoint-Aligned Differential Learning

We first adopt the endpoint-aligned formulation of RMF. Let $u_\theta(x_t, r, t) \in T_{x_t}\mathcal{M}$ denote the model-predicted finite-interval velocity, parameterized by θ. For $h = t - r$, the sampler-exact endpoint log-displacement $u_{\log}$ defined in Eq. (2) satisfies the log-displacement identity

$$u_{\log} = -\frac{D}{dt}\log_{x_t}(x_r) - h\frac{D}{dt}u_{\log},$$

where x_r is held fixed when differentiating. For training, we use the detached predicted endpoint $\widehat{x}_r^\theta = \mathrm{sg}\big(\exp_{x_t}(-hu_\theta(x_t,r,t))\big)$, where sg means stop gradient. Let C index a sampled conditional path, $\widetilde{v}_t \in T_{x_t}\mathcal{M}$ denote its velocity, and define $D_t^C = \partial_t + \nabla_{\widetilde{v}_t}$, where $\nabla_{\widetilde{v}_t}$ denotes covariant differentiation along $\widetilde{v}_t$. The self-consistent LDI target is

$$\widehat{u}_{\mathrm{LDI}} = -\nabla^{(1)}_{\widetilde{v}_t}\log_{x_t}(\widehat{x}_r^\theta) - hD_t^C u_\theta(x_t,r,t), \tag{11}$$

yielding

$$\mathcal{L}_{\mathrm{LDI}} = \mathbb{E}\left[\|u_\theta(x_t,r,t) - \mathrm{sg}\,(\widehat{u}_{\mathrm{LDI}})\|^2_{g_{x_t}}\right]. \tag{12}$$

Here $\nabla^{(1)}_{\widetilde{v}_t}$ denotes the covariant derivative with respect to the logarithm's first argument along $\widetilde{v}_t$, while keeping $\widehat{x}_r^\theta$ fixed. This provides endpoint-aligned differential supervision, while regression error remains measured in the tangent space. At the population level, the conditional LDI target reduces to its marginal counterpart, and any self-consistent fixed point within a common normal domain induces the exact finite-time flow endpoint $\exp_{x_t}(-hu) = \Phi_{r,t}(x_t)$; see Proposition A.2.

5.2 Barycentric Geodesic Semigroup Consistency

To address the remaining metric mismatch, we impose consistency directly on the manifold endpoints induced by $\Phi^\theta_{r,t}(x) = \exp_x(-(t-r)u_\theta(x,r,t))$. Since the exact flow satisfies $\Phi_{r,t} = \Phi_{r,s}\circ\Phi_{s,t}$, for $Z = (x_t, r, t)$ and $s \sim \mathrm{Unif}(r,t)$, we define

$$y_s = \Phi^{\bar\theta}_{r,s}\Big(\Phi^{\bar\theta}_{s,t}(x_t)\Big), \qquad \ell_\theta(s;Z) = \frac{d_g^2\big(\Phi^\theta_{r,t}(x_t), \mathrm{sg}(y_s)\big)}{h_\varepsilon^2}, \tag{13}$$

where $h_\varepsilon = \max(t - r, \varepsilon)$ with $\varepsilon > 0$ a small denominator floor, and $\bar{\theta}$ is a detached copy of the current parameters. A single intermediate time produces a split-dependent target before exact consistency is reached. We therefore sample K independent splits $s_k \sim \mathrm{Unif}(r, t)$ and define

$$\mathcal{L}_{\mathrm{BSC}} = \mathbb{E}_Z \mathbb{E}_{s_{1:K}} \left[\frac{1}{K} \sum_{k=1}^{K} \ell_\theta(s_k; Z) \right]. \tag{14}$$

The multi-split objective seeks an intrinsic consensus across different flow decompositions. For fixed targets $y_k = y_{s_k}$, define their Fréchet barycenter by $\bar{y} \in \arg\min_{y \in \mathcal{M}} \frac{1}{K} \sum_{k=1}^{K} d_g^2(y, y_k)$. For any reference endpoint $y^\star \in \mathcal{M}$, on a Hadamard manifold,

$$d_g^2(\bar{y}, y^\star) \le \frac{1}{K} \sum_{k=1}^{K} d_g^2(y_k, y^\star).$$

Thus, the barycentric consensus does not amplify the average split-target error relative to the same reference. The final RMF-BSC objective combines endpoint-aligned differential learning, as introduced in Eq. (12), with geometry-aligned multi-split consistency, as established in Eq. (14):

$$\mathcal{L} = \lambda \mathcal{L}_{\mathrm{BSC}} + (1 - \lambda) \mathcal{L}_{\mathrm{LDI}}, \tag{15}$$

where $\lambda \in [0, 1]$ controls the weight of the BSC objective. Barycentric stability and multi-split variance-reduction guarantees, including a local non-Hadamard version, are given in Appendix E.

6 Experiments

In this section, we conduct experiments to validate the identified geometric gaps and their sampling consequences, and to evaluate the one-step generation performance of RMF-BSC across diverse manifold benchmarks. Additional experimental details and results are provided in Appendix F.

6.1 Experimental Details

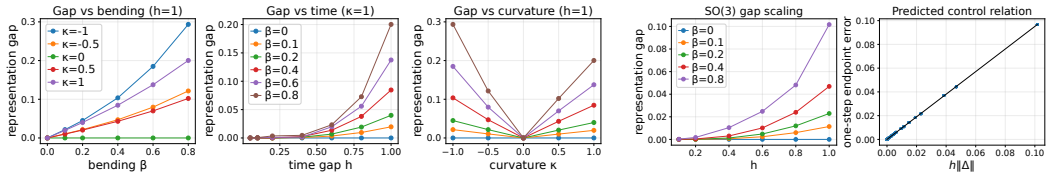
Baselines. We compare RMF-BSC with representative geometric generative models, including Riemannian Flow Matching (**RFM**) (Chen & Lipman, 2024), the directly trained, distillation-free variant of Riemannian Consistency Models (**RCT**) (Cheng et al., 2026), and **G-PSD**, the single-split geodesic semigroup variant of Generalised Flow Maps (Davis et al., 2026). We further include **RMF-Eulerian** and **RMF-Semi** (Woo et al., 2026), based on the endpoint and semigroup identities, respectively, and **RMF-PT** (Zhong et al., 2026), based on the parallel-transported average-velocity formulation, to compare the two RMF formulations examined in our analysis.

Datasets and metrics. We evaluate on four benchmarks: synthetic rotation distributions on $\mathrm{SO}(3)$ ($\mathbf{SO(3)}$; Cheng et al., 2026), Earth-event locations on $\mathbb{S}^2$ (**Earth**; Mathieu & Nickel, 2020), RNA torsion angles on $\mathbb{T}^7$ (**RNA**; Murray et al., 2003), and conditional DNA promoter sequence generation using FANTOM5 data (**DNA**; Hon et al., 2017). We report maximum mean discrepancy (**MMD**; lower is better) for $\mathrm{SO}(3)$, Earth, and RNA. For DNA, we report Fréchet Biological Distance (**FBD**; lower is better) between generated and test sequence embeddings, and k-mer correlation (***k*-mer corr.**; $k = 6$; higher is better). Dataset and metric details are provided in Appendix F.2.

6.2 Empirical Validation of Geometric Gaps

In this section, we validate the two geometric gaps in Section 4 through controlled experiments on $\mathrm{SO}(3)$ and constant-curvature spaces. We examine the representation gap's geometric origins and consequences for sampling and composition, then test the dependence of endpoint metric distortion on displacement length and perturbation direction. Few-step sampling results are in Appendix F.5.

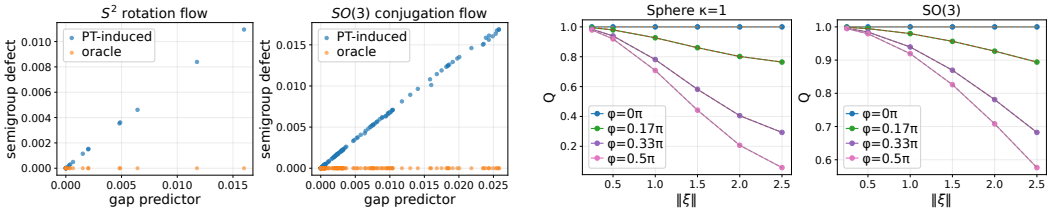
Existence and geometric origins of representation gap. We investigate the path–endpoint representation gap using controlled paths in constant-curvature spaces, without training a neural network. We sweep the sectional curvature $\kappa \in \{-1, -0.5, 0, 0.5, 1\}$, vary the bending parameter β, where $\beta = 0$ corresponds to the geodesic control and increasing β produces stronger trajectory bending



(a) Effects of curvature and path bending.

(b) Representation gap and sampling error.

Figure 1: **Geometric origins and sampling consequences of the representation gap.** (a): the gap $\|\Delta\| = \|u_{\mathrm{PT}}^{\gamma} - u_{\log}\|$ as curvature, path bending, and interval length vary. (b): the gap on $\mathrm{SO}(3)$ and the relationship between $h\|\Delta\|$ and one-step geodesic endpoint error, where $h = t - r$.



(a) Representation gap and composition defects.

(b) Directional distortion of tangent-space errors.

Figure 2: **Composition consistency and endpoint metric distortion.** (a) PT-induced composition defects versus the gap-based predictor on $\mathbb{S}^2$ and $\mathrm{SO}(3)$, with exact-flow defects near zero. (b) Endpoint-to-tangent error ratio across displacement lengths and perturbation directions, showing increasing transverse distortion.

while preserving the endpoints, and examine different interval lengths $h = t - r$. We measure $\|\Delta\|_{g_{x_t}} = \|u_{\mathrm{PT}}^{\gamma} - u_{\log}\|_{g_{x_t}}$. As shown in Fig. 1a, the gap is numerically negligible in the flat-space and geodesic-path controls, with maximum values of 9.59×10^{-11} and 8.19×10^{-15}, respectively. For the tested non-geodesic paths in curved spaces, the gap becomes nonzero and increases with stronger bending and longer intervals. These results verify the existence of the representation gap in controlled settings and demonstrate how curvature and path geometry contribute to the mismatch.

From representation gap to sampling error. To test whether the representation gap translates into endpoint error, we directly compute the transported average along each controlled $\mathrm{SO}(3)$ path and apply the one-step update $\widehat{x}_r^{\mathrm{PT}} = \exp_{x_t}(-h u_{\mathrm{PT}}^{\gamma})$. As shown in Fig. 1b, across the tested bending parameters and interval lengths, the geodesic error $d_g(\widehat{x}_r^{\mathrm{PT}}, x_r)$ is nearly proportional to $h\|\Delta\|_{g_{x_t}}$, with Pearson correlation 0.999983 and a through-origin slope of 0.9512. In contrast, replacing the transported average with $u_{\log}$ recovers the target endpoint up to numerical precision, with a maximum residual of 4.45×10^{-16}. These results demonstrate that the representation gap can induce one-step sampling error even when the transported average is computed directly, independently of neural-network approximation or optimization.

Composition consequences of representation gap. We test whether PT-induced updates remain consistent when the same interval is traversed directly or split at an intermediate time. Using analytic flows on $\mathbb{S}^2$ and $\mathrm{SO}(3)$, whose exact updates satisfy the semigroup property, we observe maximum one-step versus two-step endpoint discrepancies of 0.0110 and 0.0169, respectively (Fig. 2a). These discrepancies correlate strongly with the gap-based predictor (details in Appendix F.4), with Pearson correlation exceeding 0.9994, indicating that PT-induced updates can violate composition consistency even without learning error. This shows that composition consistency is not implied by the differential identity and must be imposed separately. RMF-BSC therefore penalizes the geodesic discrepancy between direct and composed predictions across multiple intermediate times.

Tangent–endpoint metric gap. We next examine whether equal tangent-space errors correspond to equal endpoint errors. On $\mathrm{SO}(3)$, with a spherical reference experiment, we vary the displacement length $\rho = \|\xi\|_{g_x}$ and the angle φ between ξ and a small perturbation η, and evaluate

Table 1: **One-step MMD (↓) on the $\mathbf{SO(3)}$ and Earth benchmarks.** Results are averaged over 5 random seeds; more details are reported in Appendix F.5. Best is **bold**; second best is underlined.

(a) $\mathrm{SO}(3)$

Method	Cone	Fisher	Line	Swiss Roll
RFM	0.1496	0.0780	0.0428	0.1258
RCT	0.0647	0.0777	0.0423	0.1235
G-PSD	0.0443	**0.0543**	0.0427	0.0845
RMF-PT	0.0276	0.0802	0.0353	0.1223
RMF-Eulerian	0.0275	0.0802	0.0353	0.0828
RMF-Semi	0.0305	0.0781	0.0350	0.0833
RMF-BSC	**0.0216**	0.0754	**0.0345**	**0.0582**

(b) Earth

Method	Volcano	Earthquake	Flood	Fire
RFM	0.1708	0.1666	0.1289	0.1463
RCT	0.1452	0.0501	0.0501	0.0487
G-PSD	0.1310	0.0408	0.0530	0.0393
RMF-PT	0.1248	0.0472	0.0588	0.0371
RMF-Eulerian	0.1397	0.0402	0.0649	0.0424
RMF-Semi	0.1146	**0.0329**	0.0603	0.0292
RMF-BSC	**0.1117**	0.0355	**0.0481**	**0.0227**

Table 2: **Results on the RNA and DNA benchmarks.** Results are averaged over 5 random seeds; more details are reported in Appendix F.5. Best is **bold**; second best is underlined.

Dataset	Metric	RFM	RCT	G-PSD	RMF-PT	RMF-Eulerian	RMF-Semi	RMF-BSC
RNA	1-step MMD ↓	0.1104	0.2286	0.1294	0.1193	0.1193	0.0846	**0.0825**
DNA	FBD ↓	9.4969	21.1888	9.6420	61.5611	3.7669	5.4410	**3.2888**
	k-mer corr. ↑	0.5935	0.2446	0.8807	0.2340	0.9109	0.7442	**0.9457**

$Q = d_g^2(\exp_x(\xi+\eta), \exp_x(\xi))/\|\eta\|_{g_x}^2$. Radial perturbations yield $Q \approx 1$, whereas transverse perturbations yield progressively smaller values as ρ increases. As shown in Fig. 2b, the numerical results agree with the local pullback-metric prediction; across the reported tests with $\|\eta\|_{g_x} \le 10^{-3}$, the maximum absolute discrepancy is 6.18×10^{-4}. By Proposition 4.3 with $\kappa = 1/4$ under the adopted SO(3) normalization, the transverse coefficient is $(\sin(\rho/2)/(\rho/2))^2$. Thus, tangent-space residuals of equal magnitude can induce different endpoint errors depending on their direction and displacement length, motivating the direct measurement of consistency in endpoint geometry.

6.3 Comparisons to Baselines

In this section, we evaluate RMF-BSC on four diverse manifold benchmarks to assess its effectiveness across different geometric structures. On $\mathbf{SO(3)}$ (Table 1a; full results in Table 5 of Appendix F.5), RMF-BSC achieves the lowest mean MMD on Cone, Line, and Swiss Roll, and ranks second to G-PSD on Fisher. On Swiss Roll, it reduces MMD to 0.0582, versus 0.1223 for RMF-PT and 0.0828 for RMF-Eulerian. One-step samples are shown in Fig. 3 and Fig. 8 in Appendix F.5. On **Earth** (Table 1b; full results in Table 6 of Appendix F.5), RMF-BSC achieves the lowest mean MMD on Volcano, Flood, and Fire, and ranks second to RMF-Semi on Earthquake. It outperforms RMF-PT and RMF-Eulerian on all four datasets. The one-step samples capture the main geographic patterns of held-out events (Fig. 4). On **RNA** (Table 2 and Fig. 7), RMF-BSC achieves the lowest one-step MMD (0.0825), ahead of RMF-Semi (0.0846) and RMF-PT/RMF-Eulerian (0.1193). This result on the flat torus $\mathbb{T}^7$ suggests that semigroup consistency and multi-split training also benefit non-curved manifolds. On **DNA** (Table 2), RMF-BSC achieves the best FBD (3.2888) and k-mer correlation (0.9457), improving on the strongest baseline, RMF-Eulerian (3.7669 and 0.9109, respectively). Overall, RMF-BSC delivers strong one-step performance across the rotation, spherical, toroidal, and conditional sequence benchmarks. Additional results are reported in Appendix F.5.

6.4 Ablation Study

In this ablation study, we investigate the sensitivity of RMF-BSC to the mixing weight λ and the number of splits K, together with the computational overhead introduced by multi-split training.

The mixing weight λ. We vary λ in Eq. (15) to evaluate sensitivity to the mixing weight. Performance remains relatively stable on Cone, Fisher, and Line, while larger λ generally reduces MMD on Swiss Roll (Fig. 5(a)). These results suggest limited sensitivity to λ on the first three datasets and additional benefits from stronger BSC weighting on Swiss Roll. More details are in Appendix F.5.3.

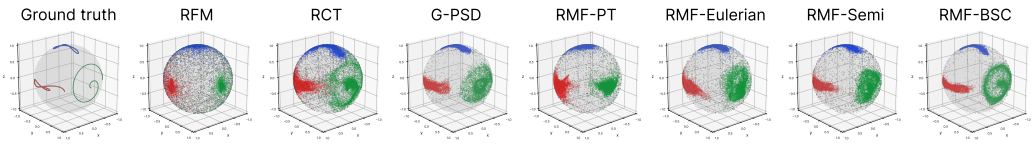
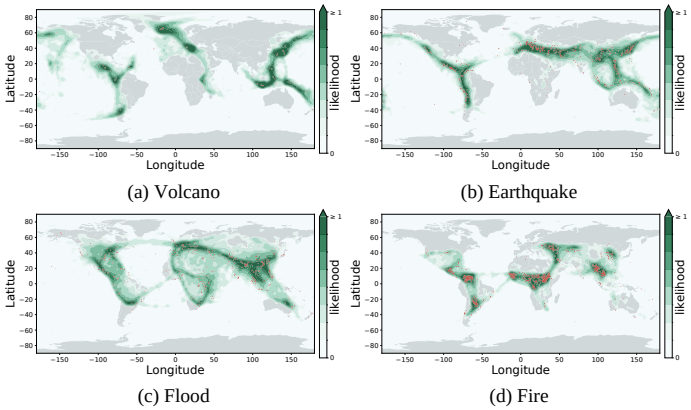
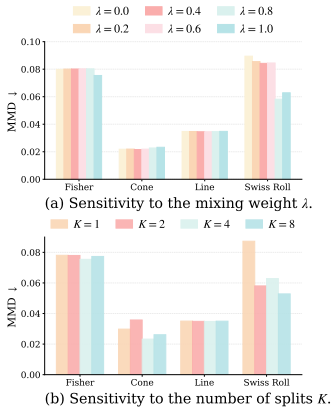


Figure 3: Ground truth and one-step samples from all baselines and RMF-BSC on Swiss Roll.



(a) Volcano (b) Earthquake (c) Flood (d) Fire

Figure 4: **Earth datasets.** One-step generated sample distributions from the selected BSC checkpoints, overlaid with ground-truth test data (red).



(a) Sensitivity to the mixing weight λ.

(b) Sensitivity to the number of splits K.

Figure 5: **Parameter sensitivity on SO(3).** MMD across different λ and K.

Tangent versus geodesic residual. To isolate the role of endpoint geometry, we compare RMF-BSC with RMF-Tan, which shares its K, LDI term, normalization, and λ but replaces the geodesic residual in Eq. (13) with the tangent residual of Appendix D.3. As shown in Table 3, the geodesic residual further reduces MMD on Cone ($0.0247 \rightarrow 0.0216$) and Swiss Roll ($0.0708 \rightarrow 0.0582$), a gain comparable to that from RMF-Semi to RMF-Tan, while the differences on Fisher and Line are marginal. This is consistent with Proposition 4.3: measuring consistency in endpoint geometry rather than in the tangent space provides a benefit beyond multi-split averaging and the LDI term.

Table 3: Residual ablation on SO(3). One-step MMD (↓), mean over 5 seeds.

Method	Cone	Fisher	Line	Swiss
RMF-Semi	0.0305	0.0781	0.0350	0.0833
RMF-Tan	0.0247	0.0767	0.0349	0.0708
RMF-BSC	**0.0216**	**0.0754**	**0.0345**	**0.0582**

The number of splits K and computational overhead. We vary $K \in \{1, 2, 4, 8\}$ with all other settings fixed to evaluate its impact on generation quality and training cost. Performance is generally stable, with larger K improving Swiss Roll, while increasing K from 1 to 8 raises per-step training time by $1.65\times$–$2.04\times$ (Fig. 5(b), Table 9 in Appendix F.5.3). These results suggest that K offers a practical trade-off between the benefits of multi-split averaging and computational cost.

7 Conclusions

In this paper, we revisit Riemannian MeanFlow and identify two geometric gaps underlying its finite-time formulations. The path–endpoint representation gap separates parallel-transported velocity averages from endpoint log-displacements, with consequences for one-step sampling and flow-map composition. The tangent–endpoint metric gap shows that even a sampler-exact representation does not make tangent-space regression equivalent to geodesic endpoint-error minimization. Together, these findings distinguish differential consistency, endpoint exactness, and metric alignment as separate requirements on curved manifolds. Guided by this analysis, we develop RMF-BSC, combining endpoint-aligned differential learning with multi-split geodesic semigroup consistency. Controlled experiments support the geometric analysis, while generation benchmarks demonstrate the practical benefits of endpoint-aligned training.

AI USE STATEMENT

AI-based tools were used during the preparation of this manuscript for limited purposes, including language editing, proofreading, and assisting with code implementation. All mathematical derivations, theoretical results, algorithmic designs, and scientific conclusions were developed by the authors. All AI-assisted code was manually reviewed, verified, and validated by the authors, and all experimental implementations and empirical analyses were independently checked by the authors.

REFERENCES

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

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

Nicholas Boffi, Michael Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. *Advances in Neural Information Processing Systems*, 38: 33346–33382, 2026.

G. R. Brakenridge. Global active archive of large flood events. `http://floodobservatory.colorado.edu/Archives/index.html`, 2017. Dartmouth Flood Observatory, University of Colorado.

Ricky TQ Chen and Yaron Lipman. Flow matching on general geometries. In *International Conference on Learning Representations*, volume 2024, pp. 47922–47945, 2024.

Yang Chen, Xiaowei Xu, Shuai Wang, Xinwen Zhang, Qiushi Guo, Tiezheng Ge, and Limin Wang. Mimflow: Integrating masked image modeling with normalizing flows for end-to-end image generation. *arXiv preprint arXiv:2606.26016*, 2026.

Chaoran Cheng, Yusong Wang, Yuxin Chen, Xiangxin Zhou, Nanning Zheng, and Ge Liu. Riemannian consistency model. *Advances in Neural Information Processing Systems*, 38:80745–80775, 2026.

Oscar Davis, Michael Albergo, Nicholas Boffi, Michael Bronstein, and Joey Bose. Generalised flow maps for few-step generative modelling on riemannian manifolds. In *International Conference on Learning Representations*, volume 2026, pp. 31119–31140, 2026.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. *Advances in neural information processing systems*, 35:2406–2422, 2022.

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

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, volume 2025, pp. 34668–34684, 2025.

Zhengyang Geng, Ashwini Pokle, and J Zico Kolter. One-step diffusion distillation via deep equilibrium models. *Advances in Neural Information Processing Systems*, 36:41914–41931, 2023.

Zhengyang Geng, Ashwini Pokle, Weijian Luo, Justin Lin, and Zico Kolter. Consistency models made easy. In *International Conference on Learning Representations*, volume 2025, pp. 75158–75185, 2025.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *Advances in Neural Information Processing Systems*, 38:75460–75482, 2026a.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30467–30476, 2026b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Chung-Chau Hon, Jordan A Ramilowski, Jayson Harshbarger, Nicolas Bertin, Owen JL Rackham, Julian Gough, Elena Denisenko, Sebastian Schmeier, Thomas M Poulsen, Jessica Severin, et al. An atlas of human long non-coding rnas with accurate 5′ ends. *Nature*, 543(7644):199–204, 2017.

Dongjun Kim, Chieh-Hsin Lai, WeiHsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *International Conference on Learning Representations*, volume 2024, pp. 44493–44525, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The twelfth international conference on learning representations*, 2023.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, volume 2025, pp. 50611–50649, 2025.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36:76525–76546, 2023.

Emile Mathieu and Maximilian Nickel. Riemannian continuous normalizing flows. *Advances in neural information processing systems*, 33:2503–2515, 2020.

Laura JW Murray, W Bryan Arendall III, David C Richardson, and Jane S Richardson. Rna backbone is rotameric. *Proceedings of the National Academy of Sciences*, 100(24):13904–13909, 2003.

NOAA. Global significant earthquake database. `https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ngdc.mgg.hazards:G012153`, 2020a. National Geophysical Data Center / World Data Service (NGDC/WDS): NCEI/WDS Global Significant Earthquake Database. NOAA National Centers for Environmental Information.

NOAA. Global significant volcanic eruptions database. `https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ngdc.mgg.hazards:G10147`, 2020b. National Geophysical Data Center / World Data Service (NGDC/WDS): NCEI/WDS Global Significant Volcanic Eruptions Database. NOAA National Centers for Environmental Information.

Yansong Peng, Kai Zhu, Yu Liu, Pingyu Wu, Hebei Li, Xiaoyan Sun, and Feng Wu. Facm: Flow-anchored consistency models. In *International Conference on Learning Representations*, volume 2026, pp. 7520–7543, 2026.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, volume 2024, pp. 15078–15097, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Fu-Yun Wang, Zhaoyang Huang, Alexander W Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency models. *Advances in neural information processing systems*, 37:83951–84009, 2024.

Dongyeop Woo, Marta Skreta, Seonghyun Park, Kirill Neklyudov, and Sungsoo Ahn. Riemannian meanflow. *arXiv preprint arXiv:2602.07744*, 2026.

Ling Yang, Zixiang Zhang, Zhilong Zhang, Xingchao Liu, Minkai Xu, Wentao Zhang, Chenlin Meng, Stefano Ermon, and Bin Cui. Consistency flow matching: Defining straight flows with velocity consistency. *arXiv preprint arXiv:2407.02398*, 2024.

Huijie Zhang, Aliaksandr Siarohin, Willi Menapace, Michael Vasilkovsky, Sergey Tulyakov, Qing Qu, and Ivan Skorokhodov. Alphaflow: Understanding and improving meanflow models. In *International Conference on Learning Representations*, volume 2026, pp. 144056–144094, 2026.

Zichen Zhong, Haoliang Sun, Yukun Zhao, Yongshun Gong, and Yilong Yin. Riemannian meanflow for one-step generation on manifolds. *arXiv preprint arXiv:2603.10718*, 2026.

A GEOMETRIC DOMAINS AND TRAJECTORY ASSUMPTIONS

Let $(\mathcal{M}, g)$ be a smooth connected Riemannian manifold with its Levi–Civita connection. Assume the smooth time-dependent field v_t generates a unique flow on the time intervals and spatial regions under consideration. We use the backward convention

$$\Phi_{r,t}(x) = \gamma_x(r), \qquad \dot{\gamma}_x(\tau) = v_\tau(\gamma_x(\tau)), \qquad \gamma_x(t) = x,$$

so that $\Phi_{t,t} = \mathrm{Id}$ and $\Phi_{r,t} = \Phi_{r,s} \circ \Phi_{s,t}$ for $r < s < t$. All exponentials used below are assumed to be defined. Completeness is not otherwise needed for the local arguments.

For each reference point x, choose an open star-shaped set $\mathcal{D}_x \subset T_x\mathcal{M}$ such that $\exp_x : \mathcal{D}_x \to \mathcal{U}_x$ is a diffeomorphism. Whenever a claim uses injectivity, both the target logarithm and the predicted displacement must belong to this same set. Endpoint membership alone does not restrict other preimages of $\exp_x$. A Hadamard manifold is complete, simply connected, and has nonpositive sectional curvature. In that setting $\exp_x$ is a global diffeomorphism, so $\mathcal{D}_x = T_x\mathcal{M}$ may be used. Global invertibility does not, however, provide a uniform global upper Lipschitz constant for $\exp_x$.

For a point $y \in \mathcal{M}$ and a direction $w \in T_x\mathcal{M}$, we write $\nabla_w^{(1)} \log_x(y)$ for the covariant derivative at x, in the direction w, of the vector field $x' \mapsto \log_{x'}(y)$, assuming the chosen logarithm branch is smooth in x' near x (e.g., the minimizing branch when $y \notin \mathrm{Cut}(x)$). That is, $\nabla^{(1)}$ differentiates only the base point, with the second argument frozen. When both arguments move, the dependence through the second argument enters separately via $d_2 \log_x|_y$, the differential of $y \mapsto \log_x(y)$; see Eq. (22). The second argument is $\hat{x}_r^\theta$ in the LDI target Eq. (11) and $Y_u(x, r, t)$ in Proposition A.2.

For a time-dependent field $u(x, r, t)$ and fixed r, our notation is

$$\frac{D}{dt} u(x_t, r, t) = \partial_t u(x_t, r, t) + \nabla_{v_t(x_t)} u(\,\cdot\,, r, t)|_{x_t}.$$

This is the total covariant derivative along the space–time trajectory, not only a spatial directional derivative.

Two Euclidean equivalences behind MeanFlow. The distinction studied in the main text comes from two identities that are automatic in Euclidean space. First, integrating velocity along a trajectory gives its endpoint displacement. Second, tangent displacement error is identical to the endpoint error after addition. On a curved manifold, the corresponding objects are respectively a transported path average versus a logarithmic endpoint displacement, and a tangent residual versus the geodesic error after applying the exponential map. The two gaps analyzed in Section 4 are precisely the failures of these two Euclidean equivalences.

A.1 CONDITIONAL PATHS, MARGINAL TRAJECTORIES, AND ENDPOINT SUPERVISION

Let C index a sampled conditional path, let $X_t = I_t(C)$, and let $\tilde{v}_t = \dot{I}_t(C) \in T_{X_t}\mathcal{M}$ denote its velocity. The marginal velocity field is

$$v_t(x) = \mathbb{E}[\dot{I}_t(C) \mid X_t = x]. \tag{16}$$

After conditioning, all vectors in this expectation belong to $T_x\mathcal{M}$. Individual interpolants can be geodesics, while integral curves of the resulting marginal field need not be. Thus, zero path–endpoint gap along each conditional interpolant does not establish zero gap along the marginal flow. Conversely, the geometric counterexample in Appendix B.3 does not by itself quantify the error of a trained model on a particular dataset.

Lemma A.1 (Conditional-to-marginal PT identity)**.** *Let $X_t = I_t(C)$ be a stochastic interpolant with conditional velocity $\widetilde{v}_t := \dot{I}_t(C) \in T_{X_t}\mathcal{M}$, and let*

$$v_t(x) = \mathbb{E}[\widetilde{v}_t \mid X_t = x]$$

be its marginal velocity field. For fixed $r < t$, define the conditional material derivative

$$D_t^C u = \partial_t u + \nabla_{\widetilde{v}_t} u$$

and the marginal material derivative

$$D_t u = \partial_t u + \nabla_{v_t} u.$$

Then, for any deterministic smooth vector field $u(x, r, t)$,

$$\mathbb{E}\big[\widetilde{v}_t - hD_t^C u(X_t, r, t) \,\big|\, X_t = x\big] = v_t(x) - hD_t u(x, r, t), \qquad h = t - r. \tag{17}$$

Consequently, any population self-consistent fixed point satisfying

$$u(x_t, r, t) = v_t(x_t) - hD_t u(x_t, r, t) \tag{18}$$

along a marginal trajectory, together with

$$\lim_{t\downarrow r}(t - r)u(x_t, r, t) = 0,$$

coincides with the transported path average $u^{\gamma}_{\mathrm{PT}}(x_t, r, t)$.

Proof. Since the Levi–Civita connection is linear in its direction argument, conditioning on $X_t = x$ gives

$$\begin{aligned}\mathbb{E}\big[D_t^C u(X_t, r, t) \,\big|\, X_t = x\big] &= \partial_t u(x, r, t) + \mathbb{E}[\nabla_{\widetilde{v}_t} u \,|\, X_t = x] \\ &= \partial_t u(x, r, t) + \nabla_{\mathbb{E}[\widetilde{v}_t | X_t = x]} u \\ &= \partial_t u(x, r, t) + \nabla_{v_t(x)} u = D_t u(x, r, t).\end{aligned} \tag{19}$$

Together with $\mathbb{E}[\widetilde{v}_t \mid X_t = x] = v_t(x)$, this proves Eq. (17).

Now let $x_t = \gamma(t)$ be an integral curve of the marginal field. If u satisfies Eq. (18), then

$$\frac{D}{dt}\big((t - r)u(x_t, r, t)\big) = u + (t - r)D_t u = v_t(x_t) = \dot{\gamma}(t).$$

Integrating this identity in a parallel frame along γ, and using the vanishing boundary condition at $t = r$, yields

$$(t - r)u(x_t, r, t) = \int_r^t \mathcal{P}^{\gamma}_{\tau\to t}\dot{\gamma}(\tau)\, d\tau.$$

Hence

$$u(x_t, r, t) = \frac{1}{t - r}\int_r^t \mathcal{P}^{\gamma}_{\tau\to t}\dot{\gamma}(\tau)\, d\tau = u^{\gamma}_{\mathrm{PT}}(x_t, r, t),$$

as claimed. □

The samplewise identity $\exp_{X_t}(\log_{X_t}(X_r)) = X_r$ also does not imply that a deterministic network recovers arbitrary random paired endpoints. If its only inputs are (X_t, r, t), squared tangent regression onto $-h^{-1}\log_{X_t}(X_r)$ has the population minimizer

$$u^{\star}(x, r, t) = -h^{-1}\mathbb{E}[\log_x(X_r) \mid X_t = x, r, t].$$

Geodesic endpoint regression instead selects a conditional Fréchet mean, when it exists. Neither object is automatically the endpoint of the marginal ODE in Eq. (16). The exactness results in the main text apply to specified endpoint pairs, and flow-map learning uses the deterministic endpoint $\Phi_{r,t}(x)$.

Proposition A.2 (Endpoint exactness of the population LDI fixed point)**.** *Let $u(x, r, t)$ be a smooth deterministic vector field and define*

$$Y_u(x, r, t) := \exp_x(-(t - r)u(x, r, t))\,.$$

Assume that, along the marginal trajectory $x_\tau = \gamma(\tau)$ for $\tau \in [r, t]$, the displacement $-(\tau - r)u(x_\tau, r, \tau)$ remains in a normal domain on which $\exp_{x_\tau}$ and $\log_{x_\tau}$ are mutually inverse.

Let

$$\widetilde{v}_t = \dot{I}_t(C), \qquad D_t^C = \partial_t + \nabla_{\widetilde{v}_t}, \qquad D_t = \partial_t + \nabla_{v_t}.$$

Then

$$\begin{aligned}&\mathbb{E}\Big[-\nabla^{(1)}_{\widetilde{v}_t}\log_{X_t}\big(Y_u(X_t, r, t)\big) - hD_t^C u(X_t, r, t)\,\Big|\, X_t = x\Big] \\ &\quad = -\nabla^{(1)}_{v_t(x)}\log_x\big(Y_u(x, r, t)\big) - hD_t u(x, r, t).\end{aligned} \tag{20}$$

Moreover, the marginal self-consistent LDI equation

$$u(x_t, r, t) = -\nabla^{(1)}_{v_t(x_t)} \log_{x_t}\big(Y_u(x_t, r, t)\big) - hD_t u(x_t, r, t) \tag{21}$$

holds along the marginal flow if and only if

$$Y_u(x_t, r, t) = \Phi_{r,t}(x_t).$$

Equivalently, within the chosen normal domain,

$$u(x_t, r, t) = -\frac{1}{t-r} \log_{x_t}\big(\Phi_{r,t}(x_t)\big).$$

Proof. We first establish the conditional-to-marginal relation. For fixed $Y_u(x, r, t)$, $\nabla^{(1)}_w \log_x(Y_u)$ is linear in the direction w. Hence

$$\begin{aligned}\mathbb{E}\Big[\nabla^{(1)}_{\widetilde{v}_t} \log_{X_t}\big(Y_u(X_t, r, t)\big) \,\Big|\, X_t = x\Big] &= \nabla^{(1)}_{\mathbb{E}[\widetilde{v}_t | X_t = x]} \log_x\big(Y_u(x, r, t)\big) \\ &= \nabla^{(1)}_{v_t(x)} \log_x\big(Y_u(x, r, t)\big).\end{aligned}$$

Likewise, by linearity of the Levi–Civita connection,

$$\mathbb{E}[D^C_t u(X_t, r, t) \mid X_t = x] = D_t u(x, r, t),$$

which proves Eq. (20).

Now fix r and let $x_\tau = \gamma(\tau)$ be a marginal trajectory. Write

$$Y(\tau) = Y_u(x_\tau, r, \tau) = \exp_{x_\tau}\left(-(\tau - r)u(x_\tau, r, \tau)\right).$$

By the normal-domain assumption,

$$\log_{x_\tau} Y(\tau) = -(\tau - r)u(x_\tau, r, \tau).$$

Taking the total covariant derivative along τ gives

$$\frac{D}{d\tau} \log_{x_\tau} Y(\tau) = \nabla^{(1)}_{v_\tau(x_\tau)} \log_{x_\tau} Y(\tau) + d_2 \log_{x_\tau}\big|_{Y(\tau)} \left[\dot{Y}(\tau)\right]. \tag{22}$$

On the other hand,

$$\frac{D}{d\tau}\left(-(\tau - r)u(x_\tau, r, \tau)\right) = -u(x_\tau, r, \tau) - (\tau - r)D_\tau u(x_\tau, r, \tau). \tag{23}$$

Combining Eqs. (22) and (23),

$$\begin{aligned}d_2 \log_{x_\tau}\big|_{Y(\tau)} \left[\dot{Y}(\tau)\right] = -\Big[&u(x_\tau, r, \tau) + \nabla^{(1)}_{v_\tau(x_\tau)} \log_{x_\tau} Y(\tau) \\ &+ (\tau - r)D_\tau u(x_\tau, r, \tau)\Big].\end{aligned} \tag{24}$$

If u satisfies Eq. (21), the right-hand side vanishes. Since $d_2 \log_{x_\tau}|_{Y(\tau)}$ is invertible within the normal domain,

$$\dot{Y}(\tau) = 0.$$

Thus $Y(\tau)$ is constant along the marginal trajectory. At $\tau = r$,

$$Y(r) = \exp_{x_r}(0) = x_r,$$

and therefore

$$Y(t) = x_r = \Phi_{r,t}(x_t).$$

Conversely, if $Y_u(x_t, r, t) = \Phi_{r,t}(x_t)$, then along the marginal trajectory $Y(\tau) = x_r$ is constant, so $\dot{Y}(\tau) = 0$. Substitution into Eq. (24) gives Eq. (21). Finally,

$$-(t - r)u(x_t, r, t) = \log_{x_t}\big(\Phi_{r,t}(x_t)\big),$$

which yields the endpoint log-displacement form. □

Remark A.3 (Role of stop-gradient). The stop-gradient operation in Eq. (12) changes differentiation through the bootstrap target but not its numerical value. Thus the proposition characterizes population self-consistent fixed points of the semi-gradient training rule, rather than minimizers of a parameter-independent static loss.

If a numerical teacher provides deterministic flow endpoints, direct supervision is possible. With $\xi^\star = \log_x(\Phi_{r,t}(x))$ and $\xi_\theta = -hu_\theta(x, r, t)$, the corresponding losses are

$$\mathcal{L}_{\mathrm{EA}} = \mathbb{E}\frac{d_g^2(\exp_x(\xi_\theta), \Phi_{r,t}(x))}{h_\varepsilon^2}, \qquad \mathcal{L}_{\tan} = \mathbb{E}\frac{\|\xi_\theta - \xi^\star\|_{g_x}^2}{h_\varepsilon^2}.$$

This is a teacher-distillation setting, separate from the semigroup self-training objective used in the main text. In Euclidean space the displayed losses agree. With clipped normalization, both equal $(h/h_\varepsilon)^2$ times the corresponding velocity MSE, rather than unweighted velocity MSE when $h < \varepsilon$.

B The Path–Endpoint Gap: Proofs and Counterexample

B.1 Transported identity and local sampler exactness

Proof of Eq. (3). Fix r and choose a parallel orthonormal frame $E_i(\tau)$ along γ. Write $\dot\gamma(\tau) = \sum_i v_i(\tau)E_i(\tau)$. Since parallel transport preserves frame components,

$$hu_{\mathrm{PT}}^\gamma(x_t, r, t) = \sum_i \left(\int_r^t v_i(\tau)\, d\tau\right) E_i(t).$$

The frame is parallel, so differentiating gives $D_t(hu_{\mathrm{PT}}^\gamma) = \sum_i v_i(t)E_i(t) = v_t(x_t)$. The product rule yields Eq. (3). Conversely, a smooth field along the trajectory satisfying this identity and $\lim_{t\downarrow r}(t-r)u(x_t, r, t) = 0$ has the same integral representation by integrating its frame components. The vanishing boundary term is necessary to exclude an integration constant. □

Proof of Proposition 4.1. For $-hu \in \mathcal{D}_{x_t}$, apply the inverse $\log_{x_t}$ to $\exp_{x_t}(-hu) = x_r$ to obtain $-hu = \log_{x_t}(x_r)$. The converse follows by applying $\exp_{x_t}$. Without the restriction on $-hu$, uniqueness can fail through geodesic wrapping, even when the endpoint has a unique logarithm in the chosen normal domain. □

B.2 Anti-development, vanishing cases, and short intervals

For $\alpha(a) = \gamma(t - a)$, the paths used by $\mathcal{P}^\alpha_{a\to 0}$ and $\mathcal{P}^\gamma_{t-a\to t}$ coincide with the same orientation. Substituting $\tau = t - a$ gives

$$\mathsf{A}_\alpha(h) = \int_0^h \mathcal{P}^\alpha_{a\to 0}\dot\alpha(a)\, da = -\int_r^t \mathcal{P}^\gamma_{\tau\to t}\dot\gamma(\tau)\, d\tau = -hu_{\mathrm{PT}}^\gamma.$$

Combining this with $-hu_{\log} = \log_{x_t}(x_r)$ proves Eq. (5).

Geometrically, $\mathsf{A}_\alpha(h)$ is the displacement obtained by developing the realized path into the fixed tangent space $T_{x_t}\mathcal{M}$, whereas $\log_{x_t}(x_r)$ is the geodesic chord selected only by the two endpoints. Thus the gap is not a notational discrepancy: it compares a path-dependent representation with the endpoint object required by the exponential sampler. The transported differential identity can hold exactly even when these two vectors differ.

In Euclidean space, $\mathsf{A}_\alpha(h) = x_r - x_t = \log_{x_t}(x_r)$ for every smooth path. If $\alpha(a) = \exp_{x_t}(a\xi/h)$ is the minimizing geodesic with $\xi = \log_{x_t}(x_r)$, its velocity transported back to x_t equals ξ/h for every a. Hence $\mathsf{A}_\alpha(h) = \xi$ and $\Delta^\gamma_{r,t} = 0$. The same conclusion holds for a smooth reparameterization of that segment whose parameter starts at zero and ends at one. These are sufficient conditions only: a non-geodesic path may also have matching developed and logarithmic endpoints.

For completeness, fix the terminal time t and let $r = t - h$ approach t. Smoothness gives, in a parallel frame, $\mathcal{P}^\gamma_{\tau\to t}\dot\gamma(\tau) = v_t(x_t) + O(h)$ uniformly on the interval. Thus $u_{\mathrm{PT}}^\gamma = v_t(x_t) + O(h)$. In normal coordinates at x_t,

$$\log_{x_t}(x_{t-h}) = -hv_t(x_t) + O(h^2), \qquad u_{\log} = v_t(x_t) + O(h).$$

It follows that $\Delta^{\gamma}_{t-h,t} = O(h)$ and tends to zero. This is a conservative continuity estimate, not a claim of a sharp curvature-dependent order. Uniform versions require uniform local derivative bounds.

B.3 An explicit spherical counterexample

Consider the unit sphere and the reversed latitude arc

$$\alpha(a) = \left(\frac{\sqrt{3}}{2} \cos a,\ \frac{\sqrt{3}}{2} \sin a,\ \frac{1}{2} \right), \qquad 0 \le a \le h = \pi.$$

Set $p = \alpha(0) = x_t$ and $q = \alpha(\pi) = x_r$. The forward curve $\gamma(\tau) = \alpha(\pi - \tau)$ is an integral curve of the smooth rotation field $v(x_1, x_2, x_3) = (x_2, -x_1, 0)$. Thus the example is an ODE trajectory, not merely an arbitrary endpoint coupling.

An orthonormal tangent frame along α is

$$E_1(a) = \left(\frac{1}{2} \cos a, \frac{1}{2} \sin a, -\frac{\sqrt{3}}{2} \right), \qquad E_2(a) = (-\sin a, \cos a, 0).$$

Tangential projection of their ambient derivatives gives $\nabla_{\dot\alpha} E_1 = \frac{1}{2} E_2$ and $\nabla_{\dot\alpha} E_2 = -\frac{1}{2} E_1$. Since $\dot\alpha(a) = \frac{\sqrt{3}}{2} E_2(a)$, solving the parallel-transport equation yields

$$\mathcal{P}^{\alpha}_{a \to 0} \dot\alpha(a) = \frac{\sqrt{3}}{2} \left[-\sin(a/2) E_1(0) + \cos(a/2) E_2(0) \right].$$

Integration and the sphere logarithm then give

$$\mathsf{A}_{\alpha}(\pi) = \sqrt{3}[-E_1(0) + E_2(0)], \qquad \log_p(q) = -\frac{2\pi}{3} E_1(0).$$

Indeed $\langle p, q \rangle = -1/2$, so $d_g(p, q) = 2\pi/3$. Therefore

$$\Delta^{\gamma}_{0,\pi} = \frac{1}{\pi} \left[\left(\sqrt{3} - \frac{2\pi}{3} \right) E_1(0) - \sqrt{3} E_2(0) \right] \neq 0.$$

Both $\|\mathsf{A}_{\alpha}(\pi)\| = \sqrt{6}$ and $\|\log_p(q)\| = 2\pi/3$ are smaller than the sphere's injectivity radius π. Their exponentials are consequently distinct. This example satisfies the transported identity exactly but has a nonzero one-step endpoint residual without crossing the cut locus. Rescaling time places the same geometric example on $[0, 1]$.

C Endpoint Error and Semigroup Defect

C.1 Local endpoint bounds

Proof of the endpoint bound in Proposition 4.2. Let $x = x_t$, $\xi = \log_x(\Phi_{r,t}(x))$, and $\zeta = -h u_{\mathrm{PT}}(x, r, t) = \xi - h\Delta_{r,t}(x)$. Here the path in $\Delta_{r,t}(x)$ is the unique integral curve terminating at x at time t. Suppose $\exp_x$ is $L_{\exp}$-Lipschitz on a set containing ξ and ζ. Then

$$E_{r,t}(x) = d_g(\exp_x(\zeta), \exp_x(\xi)) \le L_{\exp} \|\zeta - \xi\|_{g_x} = L_{\exp} h \|\Delta_{r,t}(x)\|_{g_x},$$

which proves Eq. (7). One sufficient hypothesis is that ξ, ζ lie in a convex tangent set on which $\|d \exp_x\|$ is bounded by $L_{\exp}$; integrating along their straight segment gives the required Lipschitz inequality. □

A lower bound is available with an additional inverse-map estimate. Assume $\xi, \zeta \in \mathcal{D}_x$ use the same inverse branch and their endpoints lie in a geodesically convex set $\mathcal{V} \subset \mathcal{U}_x$ on which $\|d \log_x\| \le L_{\log}$. Integration along the connecting geodesic gives

$$h \|\Delta_{r,t}(x)\|_{g_x} = \|\log_x(\exp_x(\zeta)) - \log_x(\exp_x(\xi))\|_{g_x} \le L_{\log} E_{r,t}(x).$$

Combining these estimates yields

$$\frac{h}{L_{\log}}\|\Delta_{r,t}(x)\|_{g_x} \le E_{r,t}(x) \le L_{\exp} h \|\Delta_{r,t}(x)\|_{g_x}.$$

The lower estimate is local and uses injectivity; the upper estimate alone does not establish nonzero endpoint error. Neither estimate asserts a universal constant over arbitrary intervals or unbounded regions.

For a learned field $u_\theta = u_{\mathrm{PT}} + e_\theta$, the analogous upper estimate is

$$d_g(\Phi^\theta_{r,t}(x), \Phi_{r,t}(x)) \le L_{\exp} h \|\Delta_{r,t}(x) + e_\theta(x,r,t)\|_{g_x},$$

under the same type of domain condition. Representation and approximation errors are distinct: a network need not incur the PT-induced error if its approximation error compensates for the representation discrepancy.

C.2 Composition bound with explicit evaluation points

Proof of the composition bound in Proposition 4.2. Fix $r < s < t$ and $x = x_t$. Write

$$x_s = \Phi_{s,t}(x), \qquad z_s = \Psi^{\mathrm{PT}}_{s,t}(x), \qquad x_r = \Phi_{r,s}(x_s).$$

Assume all displayed flow evaluations exist and that $\Psi^{\mathrm{PT}}_{r,s}$ is $L^\Psi_{r,s}$-Lipschitz on a region containing both x_s and z_s. The triangle inequality and the true semigroup relation give

$$\begin{aligned}
\mathfrak{S}^{\mathrm{PT}}_{r,s,t}(x) &= d_g\big(\Psi^{\mathrm{PT}}_{r,t}(x), \Psi^{\mathrm{PT}}_{r,s}(z_s)\big) \\
&\le d_g(\Psi^{\mathrm{PT}}_{r,t}(x), x_r) + d_g(x_r, \Psi^{\mathrm{PT}}_{r,s}(x_s)) + d_g(\Psi^{\mathrm{PT}}_{r,s}(x_s), \Psi^{\mathrm{PT}}_{r,s}(z_s)) \\
&\le E_{r,t}(x) + E_{r,s}(x_s) + L^\Psi_{r,s} E_{s,t}(x).
\end{aligned}$$

This proves Eq. (8). In particular, the subinterval error $E_{r,s}$ is evaluated at the true intermediate point x_s, while the map stability assumption also covers the perturbed point z_s. □

The endpoint quantity $E_{r,t}$ asks whether a single finite update recovers the reference flow endpoint, whereas $\mathfrak{S}^{\mathrm{PT}}_{r,s,t}$ compares a direct update with two composed updates. These are related but not equivalent diagnostics. A nonzero representation gap produces a local endpoint discrepancy whenever the two displacements remain in the same injective branch, while composition errors can cancel across subintervals. Therefore differential consistency of the transported average alone does not imply finite-time endpoint or semigroup exactness.

Each endpoint error can be bounded as above. With interval-dependent exponential Lipschitz constants,

$$\begin{aligned}
\mathfrak{S}^{\mathrm{PT}}_{r,s,t}(x) \le{}& L^{\exp}_{r,t}(t-r)\|\Delta_{r,t}(x)\|_{g_x} + L^{\exp}_{r,s}(s-r)\|\Delta_{r,s}(x_s)\|_{g_{x_s}} \\
&+ L^\Psi_{r,s} L^{\exp}_{s,t}(t-s)\|\Delta_{s,t}(x)\|_{g_x}.
\end{aligned}$$

The full-interval and first-applied subinterval gaps are evaluated at x, and the second-applied subinterval gap at x_s. The bound is one-sided: even inaccurate maps can be composition-consistent. For example, the identity family has zero semigroup defect regardless of its error relative to a nontrivial reference flow. These estimates concern recovery of a specified flow trajectory; they do not alone establish worse generated distributions.

D The Endpoint Pullback Metric

This appendix proves Proposition 4.3. Appendix D.1 establishes the local expansion Eq. (9) and the constant-curvature form Eq. (10). Appendix D.2 expands the latter for short displacements, and Appendix D.3 relates the result to existing semigroup objectives.

D.1 Proof of Proposition 4.3

Proof. Step 1: local expansion. Fix $x \in \mathcal{M}$ and $\xi \in \mathcal{D}_x$, and set $F = \exp_x$ and $y = F(\xi)$. Since F is smooth, $F(\xi + \eta) \to y$ as $\eta \to 0$. Hence, for $\|\eta\|_{g_x}$ sufficiently small, $F(\xi + \eta)$ lies in a geodesic ball $B(y, r)$ with $r < \mathrm{inj}(y)$, on which $\log_y$ is smooth and $d_g(y, z) = \|\log_y z\|_{g_y}$. Define $H(\eta) = \log_y \big(F(\xi + \eta)\big)$. Then $H(0) = 0$, and since $d \log_y |_y = \mathrm{Id}$, the chain rule gives

$$H(\eta) = dF|_\xi[\eta] + O\big(\|\eta\|_{g_x}^2\big).$$

Therefore

$$\ell_{\mathrm{geo}} = \|H(\eta)\|_{g_y}^2 = \|dF|_\xi[\eta]\|_{g_y}^2 + O\big(\|\eta\|_{g_x}^3\big),$$

where the cross term is $O(\|\eta\|^3)$ and the squared remainder is $O(\|\eta\|^4)$. The quadratic term is the pullback form $G_{x,\xi}(a, b) = g_y\big(dF|_\xi[a], dF|_\xi[b]\big)$, which proves Eq. (9).

The expansion only requires F to be smooth at ξ. The normal-domain assumption enters through positive definiteness: since $\exp_x |_{\mathcal{D}_x}$ is a diffeomorphism, $dF|_\xi$ is invertible, so $G_{x,\xi}$ is an inner product on $T_x\mathcal{M}$. At a conjugate point it would degenerate. Uniform remainder constants require bounded local derivatives on a common region of smoothness; the pointwise expansion alone does not supply them.

Step 2: constant curvature. We use the curvature convention $R(X, Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$, under which a Jacobi field J along a geodesic c satisfies $D_s^2 J + R(J, \dot{c})\dot{c} = 0$, and constant sectional curvature κ means

$$R(X, Y)Z = \kappa\big(\langle Y, Z\rangle X - \langle X, Z\rangle Y\big). \tag{25}$$

(a) Jacobi field representing $d \exp_x |_\xi$. Set $\rho = \|\xi\|_{g_x} > 0$, $e = \xi/\rho$, and $v_\epsilon = e + \epsilon\eta/\rho$. Consider the geodesic variation $c_\epsilon(s) = \exp_x(s v_\epsilon)$, $0 \le s \le \rho$, whose central curve $c(s) = c_0(s) = \exp_x(se)$ is the unit-speed geodesic from x to $\exp_x(\xi)$. Since each c_ϵ is a geodesic, the variation field $J(s) = \partial_\epsilon c_\epsilon(s)|_{\epsilon=0}$ is a Jacobi field along c. Its initial conditions and terminal value are as follows. First, $c_\epsilon(0) = x$ for all ϵ, so $J(0) = 0$. Second, by the symmetry lemma $D_s \partial_\epsilon c_\epsilon = D_\epsilon \partial_s c_\epsilon$, and since $\partial_s c_\epsilon(0) = v_\epsilon$ lies in the fixed space $T_x\mathcal{M}$,

$$D_s J(0) = \frac{d}{d\epsilon} v_\epsilon \Big|_{\epsilon=0} = \frac{\eta}{\rho}.$$

Third, $c_\epsilon(\rho) = \exp_x(\rho e + \epsilon\eta) = \exp_x(\xi + \epsilon\eta)$, so

$$J(\rho) = \frac{d}{d\epsilon} \exp_x(\xi + \epsilon\eta)\Big|_{\epsilon=0} = d \exp_x |_\xi[\eta].$$

Hence, by the definition in Eq. (9),

$$G_{x,\xi}(\eta, \eta) = \|J(\rho)\|_{g_{\exp_x(\xi)}}^2. \tag{26}$$

(b) Reduction to scalar equations. Let $n = \dim \mathcal{M}$ and let $\{e_1(s), \dots, e_n(s)\}$ be a parallel orthonormal frame along c with $e_1(s) = \dot{c}(s)$; this is possible because $\dot{c}$ is parallel along the geodesic c and has unit length. In particular $e_1(0) = e$, and $e_2(0), \dots, e_n(0)$ span the orthogonal complement of ξ. Write $J(s) = \sum_{i=1}^n f_i(s) e_i(s)$. Because the frame is parallel, $D_s^2 J = \sum_i f_i''(s) e_i(s)$. By Eq. (25) with $\|\dot{c}\| = 1$,

$$R(J, \dot{c})\dot{c} = \kappa\big(J - \langle J, \dot{c}\rangle \dot{c}\big) = \kappa \sum_{i=2}^n f_i(s) e_i(s).$$

The Jacobi equation therefore decouples into

$$f_1''(s) = 0, \qquad f_i''(s) + \kappa f_i(s) = 0 \quad (i = 2, \dots, n),$$

with initial conditions, from (a),

$$f_i(0) = 0, \qquad f_i'(0) = \frac{\langle \eta, e_i(0)\rangle_{g_x}}{\rho} \quad (i = 1, \dots, n).$$

(c) Solving the equations. The first equation gives $f_1(s) = s\,\langle \eta, e\rangle_{g_x}/\rho$. For $i \geq 2$, let

$$S_\kappa(s) = \begin{cases} \sin(\sqrt{\kappa}s)/\sqrt{\kappa}, & \kappa > 0, \\ s, & \kappa = 0, \\ \sinh(\sqrt{-\kappa}s)/\sqrt{-\kappa}, & \kappa < 0. \end{cases}$$

In each case $S_\kappa'' + \kappa S_\kappa = 0$ and $S_\kappa(0) = 0$, and $S_\kappa'(0) = 1$ since $S_\kappa'(s)$ equals $\cos(\sqrt{\kappa}s)$, 1, and $\cosh(\sqrt{-\kappa}s)$, respectively. Hence S_κ is the unique solution of $S'' + \kappa S = 0$ with $S(0) = 0$ and $S'(0) = 1$, and by linearity

$$f_i(s) = \frac{S_\kappa(s)}{\rho}\,\langle \eta, e_i(0)\rangle_{g_x}, \qquad i \geq 2.$$

(d) Evaluating the norm. At $s = \rho$, $f_1(\rho) = \langle \eta, e\rangle_{g_x}$ and $f_i(\rho) = \frac{S_\kappa(\rho)}{\rho}\langle \eta, e_i(0)\rangle_{g_x}$ for $i \geq 2$. Since $\{e_i(\rho)\}$ is orthonormal in $T_{\exp_x(\xi)}\mathcal{M}$,

$$\|J(\rho)\|^2_{g_{\exp_x(\xi)}} = \sum_{i=1}^n f_i(\rho)^2 = \langle \eta, e\rangle^2_{g_x} + \Big(\frac{S_\kappa(\rho)}{\rho}\Big)^2 \sum_{i=2}^n \langle \eta, e_i(0)\rangle^2_{g_x}.$$

Because $\{e_i(0)\}$ is an orthonormal basis of $T_x\mathcal{M}$ with $e_1(0) = e$, we have

$$\langle \eta, e\rangle^2_{g_x} = \|\eta_\parallel\|^2_{g_x}, \qquad \sum_{i=2}^n \langle \eta, e_i(0)\rangle^2_{g_x} = \|\eta_\perp\|^2_{g_x}.$$

Substituting into Eq. (26) gives

$$G_{x,\xi}(\eta,\eta) = \|\eta_\parallel\|^2_{g_x} + \Big(\frac{S_\kappa(\rho)}{\rho}\Big)^2 \|\eta_\perp\|^2_{g_x}, \qquad S_\kappa(\rho) = \begin{cases} \sin(\sqrt{\kappa}\rho)/\sqrt{\kappa}, & \kappa > 0, \\ \rho, & \kappa = 0, \\ \sinh(\sqrt{-\kappa}\rho)/\sqrt{-\kappa}, & \kappa < 0, \end{cases}$$

which is Eq. (10). Equivalently, in vector form,

$$d\exp_x|_\xi[\eta] = \sum_{i=1}^n f_i(\rho) e_i(\rho) = P^c_{0\to\rho}\Big[\eta_\parallel + \frac{S_\kappa(\rho)}{\rho}\eta_\perp\Big],$$

where we used $e_i(\rho) = P^c_{0\to\rho} e_i(0)$. Finally, at $\xi = 0$ we have $d\exp_x|_0 = \mathrm{Id}$, so $G_{x,0} = g_x$, consistent with $S_\kappa(\rho)/\rho \to 1$ as $\rho \to 0$. □

Remarks. The Jacobi-field formula holds for every $\rho > 0$. The normal-domain assumption guarantees $S_\kappa(\rho) > 0$: this is automatic for $\kappa \leq 0$, and for $\kappa > 0$ the first conjugate point along c occurs at $\rho = \pi/\sqrt{\kappa}$, which a normal domain excludes. For $\kappa < 0$ the transverse coefficient exceeds one and grows with $\sqrt{-\kappa}\rho$; for $\kappa > 0$ it lies in $(0, 1)$ on the normal domain. Radial perturbations have unit weight in either case. Thus displacement length and error direction matter in addition to curvature. A longer time interval alone does not guarantee a larger metric gap, and a flat manifold's global logarithm-branch issues are separate from this local distortion.

The two losses compare errors under two intrinsic metrics on $T_x\mathcal{M}$, namely g_x and $G_{x,\xi}$, rather than an intrinsic versus an extrinsic norm. When $\xi, \xi + \eta \in \mathcal{D}_x$, both vanish exactly at $\eta = 0$. With noisy targets or a restricted hypothesis class, however, identical zero sets do not imply identical population minimizers or optimization trajectories. For semigroup training, the target displacement is the logarithm of the composed endpoint: tangent regression measures the residual with g_x, whereas the sampler experiences it through $G_{x,\xi}$. The metric gap is therefore a finite-error weighting effect, not a claim that Endpoint RMF has the wrong realizable fixed point.

D.2 Small-Displacement Expansion

For fixed κ and $\rho \to 0$, the series of S_κ gives

$$\frac{S_\kappa(\rho)}{\rho} = 1 - \frac{\kappa\rho^2}{6} + O(\rho^4), \qquad \Big(\frac{S_\kappa(\rho)}{\rho}\Big)^2 = 1 - \frac{\kappa\rho^2}{3} + O(\rho^4).$$

Substitution into Eq. (10) yields

$$G_{x,\xi}(\eta,\eta) = \|\eta\|_{g_x}^2 - \frac{\kappa\rho^2}{3}\|\eta_\perp\|_{g_x}^2 + O\big(\rho^4\|\eta_\perp\|_{g_x}^2\big).$$

The leading curvature correction is quadratic in displacement length and acts only on transverse perturbations. It is positive in negative curvature and negative in positive curvature. This $\rho \to 0$ expansion concerns the target displacement and is distinct from the $\eta \to 0$ endpoint-error expansion in Eq. (9). For a fixed displacement, the latter is only a local approximation in prediction error; it does not identify the quadratic form with the full geodesic loss for arbitrary finite η.

D.3 Relation to Existing Semigroup Objectives

For a fixed composed endpoint y_s and $x = x_t$, let $\xi_s = \log_x(y_s)$ and $\xi_\theta = -hu_\theta(x,r,t)$. The tangent semigroup residual of Endpoint RMF (Woo et al., 2026), expressed in our backward convention, is

$$\big\|u_\theta(x,r,t) + h^{-1}\log_x(y_s)\big\|_{g_x}^2 = h^{-2}\|\xi_\theta - \xi_s\|_{g_x}^2.$$

The geodesic counterpart replaces the tangent residual by $h^{-2}d_g^2\big(\exp_x(\xi_\theta), y_s\big)$, and Eq. (13) replaces h^{-2} with h_ε^{-2} for numerical stability. With $\eta = \xi_\theta - \xi_s$ and $\xi_s \in \mathcal{D}_x$, Proposition 4.3 shows that, to leading order in η, the two residuals differ exactly by replacing g_x with G_{x,ξ_s}. This relation does not assert identical complete training procedures, which can differ in time sampling, anchors, and optimization. Our analysis concerns the geometry of the residual and the effect of averaging several splits.

The semigroup identity itself is not new: an exact flow already satisfies $\Phi_{r,t} = \Phi_{r,s} \circ \Phi_{s,t}$. Our modification concerns how deviations from this fixed point are measured. The stop-gradient target branch turns the composition into a bootstrap target for each update, while h_ε^{-2} places endpoint discrepancies on a velocity-like scale; clipping avoids arbitrarily large weights on very short intervals.

E Barycentric Stability and Multi-Split Averaging

E.1 Hadamard variance inequality

For fixed targets $y_1, \dots, y_K$, define $F(y) = K^{-1}\sum_{k=1}^K d_g^2(y, y_k)$. The positive normalization h_ε^{-2} does not affect its minimizer. On a finite-dimensional Hadamard manifold F is coercive and strongly geodesically convex, and hence has a unique minimizer $\bar{y}$. This is a statement about optimization over points of the manifold; a shared neural parameterization need not realize every pointwise minimizer.

Before exact semigroup consistency is reached, different intermediate times generally produce different composed endpoints for the same (X_t, R, T). A single-split objective therefore exposes training to an arbitrary decomposition-dependent bootstrap target. Averaging several splits replaces that choice by an intrinsic consensus objective. No explicit Karcher-mean solver is required in training: minimizing the averaged squared geodesic losses has the Fréchet barycenter as its pointwise minimizer.

Proposition E.1 (Barycentric variance inequality)**.** *On a Hadamard manifold, for any reference endpoint $y^\star$,*

$$d_g^2(\bar{y}, y^\star) + \frac{1}{K}\sum_{k=1}^K d_g^2(\bar{y}, y_k) \le \frac{1}{K}\sum_{k=1}^K d_g^2(y_k, y^\star). \tag{27}$$

Proof. The standard squared-distance convexity inequality on a Hadamard manifold is

$$d_g^2(\gamma(a), q) \le (1-a)d_g^2(z_0, q) + a d_g^2(z_1, q) - a(1-a)d_g^2(z_0, z_1),$$

where γ is the geodesic from z_0 to z_1. Averaging over $q = y_k$, setting $z_0 = \bar{y}$ and $z_1 = y^\star$, and using minimality gives

$$F(\bar{y}) \le F(\gamma(a)) \le (1-a)F(\bar{y}) + aF(y^\star) - a(1-a)d_g^2(\bar{y}, y^\star).$$

Divide by $a > 0$ and send $a \downarrow 0$ to obtain $F(y^\star) - F(\bar{y}) \ge d_g^2(\bar{y}, y^\star)$, which is Eq. (27). □

The inequality bounds the consensus error by the average target error, with target dispersion providing the subtractive term. It does not show improvement over the best individual target, remove a bias shared by all targets, or guarantee that neural optimization reaches $\bar{y}$. No independence assumption on the targets is needed for this deterministic inequality.

E.2 Local version on general manifolds

Normal-domain invertibility alone is insufficient for a unit-constant barycentric inequality. Suppose the targets and reference point lie in a common geodesically convex region Ω, the Fréchet functional attains its minimum over Ω at $\bar{y}_\Omega$, and for every geodesic γ joining $z_0, z_1 \in \Omega$,

$$F(\gamma(a)) \le (1-a)F(z_0) + aF(z_1) - \mu a(1-a)d_g^2(z_0, z_1)$$

for some $\mu > 0$. These conditions hold in sufficiently small regions with appropriate squared-distance Hessian bounds. The same proof yields uniqueness in Ω and

$$\mu\, d_g^2(\bar{y}_\Omega, y^\star) + \frac{1}{K}\sum_{k=1}^{K} d_g^2(\bar{y}_\Omega, y_k) \le \frac{1}{K}\sum_{k=1}^{K} d_g^2(y_k, y^\star). \tag{28}$$

In positive curvature μ can be less than one. Thus the local result carries a curvature- and region-dependent constant, and does not inherit the Hadamard non-amplification statement with coefficient one. Close to a cut locus, uniqueness and smoothness can fail.

E.3 Conditional gradient covariance

Let (R, T) be a random off-diagonal training time pair, set $X_T = I_T(C)$ and $Z = (X_T, R, T)$, and draw $S \mid Z \sim \mathrm{Unif}(R, T)$. Fix the parameter sets θ and $\bar{\theta}$. Let $Y_\theta(Z) = \Phi^\theta_{R,T}(X_T)$ and $y_S = \Phi^{\bar{\theta}}_{R,S}(\Phi^{\bar{\theta}}_{S,T}(X_T))$. Assume y_S avoids the cut locus of $Y_\theta(Z)$ almost surely and that the stopped parameter gradient has finite second moment. From Eq. (13),

$$g(S; Z) := \nabla_\theta \ell_\theta(S; Z) = -\frac{2}{H_\varepsilon^2}(D_\theta Y_\theta(Z))^* \log_{Y_\theta(Z)}(y_S),$$

where $H_\varepsilon = \max(T - R, \varepsilon)$ and the adjoint uses the Euclidean parameter metric and the output Riemannian metric. The target branch is held constant when differentiating.

Draw $S_1, \ldots, S_K$ conditionally independently from the same split distribution and set

$$G_K(Z) = \frac{1}{K}\sum_{k=1}^{K} g(S_k; Z), \qquad m(Z) = \mathbb{E}_S[g(S; Z) \mid Z], \qquad \Sigma(Z) = \mathrm{Cov}_S(g(S; Z) \mid Z).$$

Conditioning on both parameter sets is implicit throughout. Linearity of expectation and zero cross-covariances give

$$\mathbb{E}[G_K \mid Z] = m(Z), \qquad \mathrm{Cov}(G_K \mid Z) = \frac{1}{K}\Sigma(Z). \tag{29}$$

Indeed, the covariance of the sum is the sum of the K conditional covariances, divided by K^2. Applying the law of total covariance gives

$$\mathrm{Cov}(G_K) = \mathrm{Cov}_Z(m(Z)) + \frac{1}{K}\mathbb{E}_Z[\Sigma(Z)].$$

Only the split-induced component is reduced. Randomness from states and time pairs remains in the first term. With dependent splits, the additional term $K^{-2}\sum_{i\neq j}\mathrm{Cov}(g(S_i; Z), g(S_j; Z) \mid Z)$ must be retained. Any shared stochastic network randomness must also be conditioned on or included in this decomposition; shared deterministic network parameters alone do not invalidate conditional independence.

Finally, averaging K identically distributed per-split losses preserves their expectation, and Eq. (29) preserves the expected stopped update. Thus multi-split training is an estimator change for a fixed split distribution, not a new population optimum. Target-level barycentric stability and conditional gradient averaging are complementary statements; neither establishes unconditional training convergence or improved generation at a fixed compute budget.

F Experimental Details

F.1 Training Objective and Anchoring

We use the backward convention of the main text, with $r < t$ and

$$\Phi^{\theta}_{r,t}(x) = \exp_x(-(t-r)u_\theta(x,r,t))\,.$$

Implementations using the opposite time orientation use the algebraically equivalent sign convention.

Off-diagonal training. For each off-diagonal sample $Z = (X_t, r, t)$, we draw K independent splits $s_k \sim \mathrm{Unif}(r,t)$. For each split, the composed endpoint is evaluated using the current model under stop-gradient,

$$y_{s_k} = \mathrm{sg}\big[\Phi^{\theta}_{r,s_k}\big(\Phi^{\theta}_{s_k,t}(X_t)\big)\big]\,.$$

The direct endpoint retains gradients, and the resulting geodesic losses are averaged as in Eq. (14). The off-diagonal objective is

$$\mathcal{L} = \lambda\mathcal{L}_{\mathrm{BSC}} + (1-\lambda)\mathcal{L}_{\mathrm{LDI}}.$$

Diagonal anchor. For diagonal samples $r = t$, we replace the off-diagonal objective by instantaneous-velocity regression,

$$\mathcal{L}_{\mathrm{diag}} = \mathbb{E}\left[\left\|u_\theta(X_t,t,t) - \dot{I}_t(C)\right\|^2_{g_{X_t}}\right].$$

Since $v_t(x) = \mathbb{E}[\dot{I}_t(C) \mid X_t = x]$, this regression has the marginal velocity as its population minimizer. The diagonal anchor is counted once and is not added again to the off-diagonal LDI–BSC objective.

Numerical safeguards. We use $h_\varepsilon = \max(t-r, \varepsilon)$ in the geodesic semigroup loss to avoid unstable normalization over very short intervals, and exclude logarithms sufficiently close to the cut locus when required by the manifold implementation.

F.2 Datasets and Metrics

F.2.1 Datasets.

We evaluate on four benchmarks with distinct geometric and statistical structures. **(1) $\mathbf{SO(3)}$:** We use the synthetic Cone, Fisher, Line, and Swiss Roll datasets from Cheng et al. (2026), which exhibit different concentration patterns and geometric structures on the rotation group, allowing controlled evaluation of rotation distributions. **(2) Earth:** We consider the Earth-event benchmark introduced by Mathieu & Nickel (2020) and curated from multiple public sources (NOAA, 2020a;b; EOSDIS, 2020; Brakenridge, 2017), comprising Volcano, Earthquake, Flood, and Fire locations on $\mathbb{S}^2$. These real-world datasets exhibit nonuniform geographic distributions and test the recovery of spatial structure on a curved surface. **(3) RNA:** We use the seven-dimensional RNA backbone torsion-angle dataset of Murray et al. (2003). Each sample consists of seven periodic angles, forming a point on the flat torus $\mathbb{T}^7$, so generated samples must respect the periodicity of each angle. **(4) DNA:** We use human promoter sequences from FANTOM5 (Hon et al., 2017), with sequence length 1024 and four nucleotide categories per position. Conditioned on target transcription signal profiles, this benchmark tests high-dimensional sequence generation with both functional constraints and local sequence structure.

F.2.2 Metrics.

We use the same evaluation implementation and settings for all compared methods within each benchmark.

MMD. For $\mathrm{SO}(3)$, Earth, and RNA, we follow the benchmark MMD evaluation protocol of Zhong et al. (2026). Given generated samples $X = \{x_i\}_{i=1}^n$ and held-out test samples $Y = \{y_j\}_{j=1}^m$, we report the square root of the V-statistic

$$\widehat{\mathrm{MMD}}^2(X,Y) = \frac{1}{n^2}\sum_{i,i'=1}^{n} k(x_i, x_{i'}) + \frac{1}{m^2}\sum_{j,j'=1}^{m} k(y_j, y_{j'}) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i, y_j),$$

including within-sample diagonal terms. For $\mathrm{SO}(3)$, the implementation clips the estimated discrepancy at zero before taking the square root. The dataset-specific kernels are

$$k(x,y) = \begin{cases} \exp\left[-d_g(x,y)^2\right], & \mathrm{SO}(3) \text{ and Earth}, \\ \dfrac{1}{7}\displaystyle\sum_{\ell=1}^{7} \exp\left[-d_{\mathbb{S}^1}(x_\ell, y_\ell)^2\right], & \text{RNA}, \end{cases}$$

where $d_{\mathbb{S}^1}(a,b) = \min\{|a-b|,\ 2\pi - |a-b|\}$ for $a, b \in [0, 2\pi)$. The RNA score measures agreement across individual torsion-angle marginals. We retain these empirical evaluation conventions for comparability with the adopted benchmark protocol.

FBD. For DNA, generated sequences are decoded by positionwise argmax before evaluation. We compute Fréchet Biological Distance using the $d = 15360$ dimensional features obtained by flattening the spline transformation output immediately before the classifier of a fixed pretrained Sei model. Each sequence is padded on both sides with 1536 uniform-probability positions to obtain a length-4096 input. Following the evaluation implementation, we normalize the distance by the feature dimension:

$$\mathrm{FBD} = \frac{1}{d}\left[\|\mu_g - \mu_r\|_2^2 + \mathrm{Tr}\left(\widetilde{\Sigma}_g + \widetilde{\Sigma}_r - 2\left(\widetilde{\Sigma}_g^{1/2}\widetilde{\Sigma}_r\widetilde{\Sigma}_g^{1/2}\right)^{1/2}\right)\right],$$

where (μ_g, Σ_g) and (μ_r, Σ_r) are the empirical means and covariances of generated and test features, respectively. For numerical stability, the implementation uses $\widetilde{\Sigma}_g = \Sigma_g + 10^{-6}I$ and $\widetilde{\Sigma}_r = \Sigma_r + 10^{-6}I$.

k-mer correlation. We also evaluate DNA sequence statistics using the Pearson correlation between pooled 6-mer frequency vectors:

$$\rho_6 = \frac{\sum_{w\in\mathcal{V}}\left(f_g(w) - \bar{f}_g\right)\left(f_r(w) - \bar{f}_r\right)}{\sqrt{\sum_{w\in\mathcal{V}}\left(f_g(w) - \bar{f}_g\right)^2}\sqrt{\sum_{w\in\mathcal{V}}\left(f_r(w) - \bar{f}_r\right)^2}},$$

where $\mathcal{V}$ contains the 6-mers observed in either collection, f_g and f_r denote their pooled frequencies, and the bars denote averages over $\mathcal{V}$. Lower MMD and FBD and higher 6-mer correlation indicate better performance under the corresponding metrics.

F.3 Implementation and Hyperparameters

Data representation and splits. For $\mathrm{SO}(3)$, we use 20,000 rotations for Cone and 40,000 each for Fisher, Line, and Swiss Roll. Earth-event coordinates are mapped from latitude and longitude to the unit sphere. RNA is represented by seven torsion angles $(\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \chi)$, converted to radians in $[0, 2\pi)$ and modeled on $\mathbb{T}^7$. These datasets use an 80/10/10 training/validation/test split with a fixed split seed of 42.

DNA consists of 1,024-base promoter windows around transcription start sites from the first 100,000 entries of the ranked FANTOM-CAT annotation, with CAGE signals as conditioning information. Chromosome 10 is reserved for validation and chromosomes 8–9 for testing; the remaining chromosomes are used for training. Training windows receive random shifts of up to 10 bases. Nucleotide probabilities are represented on a product of positive spherical orthants, $(\mathbb{S}^3_+)^{1024}$.

Table 4: **Core training configurations.** MLP entries give the configured depth and width; CNN width denotes its hidden channels. Epochs indicate training budgets, with early stopping enabled for Earth and RNA.

Dataset	Network	Batch	Learning rate	Weight decay	Epochs
Cone/Fisher/Line	MLP 4×512	1,024	5×10^{-4}	10^{-2}	200
Swiss Roll	MLP 8×1024	1,024	5×10^{-4}	10^{-2}	200
Volcano	MLP 12×2048	8,192	10^{-4}	10^{-2}	2,000
Earthquake	MLP 10×2048	8,192	10^{-4}	10^{-3}	700
Flood	MLP 7×2048	8,192	10^{-4}	10^{-2}	900
Fire	MLP 10×2048	8,192	10^{-4}	10^{-2}	600
RNA	MLP 6×512	4,096	10^{-4}	10^{-2}	1,450
DNA	CNN, width 256	128	10^{-3}	0	200

Networks and optimization. Table 4 summarizes the core training configurations. The $\mathrm{SO}(3)$ network takes a flattened rotation matrix and predicts a three-dimensional body velocity. Earth and RNA use time-conditioned MLPs, while DNA uses a conditional one-dimensional CNN with Fourier time embeddings. All models are optimized with AdamW. Learning rates follow cosine schedules for $\mathrm{SO}(3)$ and Earth, with a minimum of 10^{-6}; RNA uses a constant learning rate. DNA uses a cosine decay horizon of 1,000 epochs and global gradient-norm clipping at 1.

Earth and RNA select checkpoints by one-step validation MMD. Their epoch budgets include repeated training examples, with repetition factors of 1,550/210/260/100 for Volcano/Earthquake/Flood/Fire and 14 for RNA. Repetition is applied after splitting and introduces no additional unique observations.

Early stopping. Training is stopped when the validation (MMD) does not improve for a predefined number of consecutive validation evaluations, preventing unnecessary over-training.

Time sampling and BSC settings. For $\mathrm{SO}(3)$, Earth, and RNA, we sample $t \sim \mathcal{U}(0,1)$ and $r \mid t \sim \mathcal{U}(0,t)$, replacing r by t with probability 0.25 for instantaneous-velocity supervision. DNA uses its forward-time convention: two logit-normal samples with normal parameters $(-0.4, 1)$ are ordered, with coincident times sampled with probability 0.75.

Unless otherwise stated, BSC uses $K = 4$ independent uniform intermediate times. The denominator floor ε is 0.05 for $\mathrm{SO}(3)$ and DNA, and 0.01 for Earth and RNA; the cut-locus margin is 10^{-3}. The mixture objective is

$$\mathcal{L} = \lambda \mathcal{L}_{\mathrm{BSC}} + (1-\lambda)\mathcal{L}_{\mathrm{LDI}},$$

with λ selected on validation data.

RNA additionally applies stopped adaptive weighting with exponent $1/2$ and caps the component losses at 5 after 10,000 updates. DNA clips the LDI material derivative at product-tangent norm 100 and multiplies per-example losses by $\min\{1, (1-t)/0.1\}^2$. Its BSC mixture uses an arithmetic mean; standalone LDI and Eulerian runs additionally use adaptive weighting with exponent $1/2$.

F.4 Gap-based predictor

For a temporal split $r < s < t$, let $h = t - r$, $h_1 = s - r$, and $h_2 = t - s$. For each interval $[a, b]$, we define the representation gap as

$$\Delta_{a,b} = u_{\mathrm{PT}}^{a,b} - u_{\log}^{a,b},$$

where $u_{\mathrm{PT}}^{a,b}$ is the parallel-transported average velocity and

$$u_{\log}^{a,b} = -\frac{\log_{x_b}(x_a)}{b-a}$$

is the endpoint-displacement velocity. We use

$$\mathcal{P}_{\mathrm{gap}}(r,s,t) = (t-r)\,\|\Delta_{r,t}\| + (s-r)\,\|\Delta_{r,s}\| + (t-s)\,\|\Delta_{s,t}\|$$

Table 5: **MMD (↓) on the $\mathbf{SO(3)}$ benchmarks**, reported as mean ± standard deviation over 5 random seeds. Best results are shown in **bold** and second-best results are underlined.

Method	Cone	Fisher	Line	Swiss Roll
RFM	0.1496 ± 0.0055	0.0780 ± 0.0011	0.0428 ± 0.0013	0.1258 ± 0.0035
RCT	0.0647 ± 0.0095	0.0777 ± 0.0013	0.0423 ± 0.0017	0.1235 ± 0.0136
G-PSD	0.0443 ± 0.0127	$\mathbf{0.0543 \pm 0.0010}$	0.0427 ± 0.0013	0.0845 ± 0.0054
RMF-PT	0.0276 ± 0.0056	0.0802 ± 0.0003	0.0353 ± 0.0007	0.1223 ± 0.0293
RMF-Eulerian	$\underline{0.0275 \pm 0.0051}$	0.0802 ± 0.0003	0.0353 ± 0.0007	$\underline{0.0828 \pm 0.0157}$
RMF-Semi	0.0305 ± 0.0040	0.0781 ± 0.0035	$\underline{0.0350 \pm 0.0002}$	0.0833 ± 0.0179
RMF-BSC	$\mathbf{0.0216 \pm 0.0003}$	$\underline{0.0754 \pm 0.0058}$	$\mathbf{0.0345 \pm 0.0001}$	$\mathbf{0.0582 \pm 0.0095}$

as a first-order proxy for the discrepancy between the direct PT-induced update and its two-step composition. The three terms correspond respectively to the direct interval and the two subintervals. This quantity is a diagnostic predictor rather than the BSC training objective itself.

F.5 Additional Experimental Results

F.5.1 Additional Results on the Empirical Validation of Geometric Gaps

Few-step sampling. We further evaluate multi-step generation on Swiss Roll by varying the number of function evaluations (NFE) in $\{1, 2, 4, 8, 16\}$ (Fig. 6). RMF-BSC achieves the lowest MMD at every NFE. With only two steps it reaches 0.0101, already below every baseline at 16 NFE (RMF-Semi 0.0155, RMF-Eulerian 0.0181, RMF-PT 0.0346). Beyond two steps, RMF-BSC remains essentially flat (0.0095–0.0109) and RMF-Semi varies only moderately (0.0138–0.0189), whereas RMF-Eulerian keeps improving up to 8 NFE ($0.0339 \rightarrow 0.0185$) and RMF-PT improves steadily throughout ($0.0645 \rightarrow 0.0346$). This is consistent with our analysis: once the learned flow map is approximately semigroup-consistent, further subdivision changes the composed endpoint only slightly, and the remaining error is dominated by approximation rather than composition. RMF-PT is the worst variant at every NFE. Its gap to RMF-Eulerian shrinks from 0.0548 at one step to 0.0165 at 16 steps, consistent with the representation gap: the per-step endpoint error is controlled by $h\|\Delta_{r,t}\|$ (Proposition 4.2), and $\Delta_{r,t}$ vanishes as the step size shrinks (Appendix B.2). The gap does not close within 16 steps, however, and RMF-PT remains roughly twice as high. For all methods, the largest improvement occurs from one to two steps, indicating that one-step quality is limited mainly by the accuracy of the full-interval map.

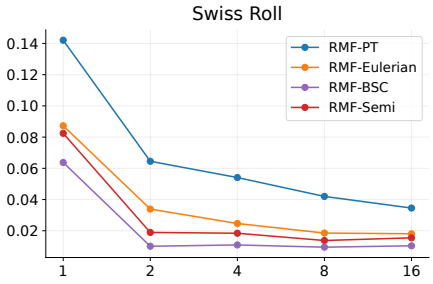


Figure 6: **Few-step sampling on Swiss Roll.** MMD (↓) versus NFE for the RMF variants. Single run per method, with RMF-BSC at $\lambda = 1$; values at NFE $= 1$ therefore differ from Table 1a.

F.5.2 Additional Baseline Comparison Results

In this section, we evaluate RMF-BSC on four diverse manifold benchmarks to assess its effectiveness across different geometric structures. On $\mathrm{SO}(3)$ (Table 1a and Table 5 in Appendix F.5), RMF-BSC achieves the lowest mean MMD on Cone, Line, and Swiss Roll, and ranks second to G-PSD on Fisher. The improvement is particularly pronounced on Swiss Roll, where MMD decreases from 0.1223 for RMF-PT and 0.0828 for RMF-Eulerian to 0.0582. The corresponding one-step samples are shown in Fig. 3 and Fig. 8 in Appendix F.5. On the Earth benchmarks (Table 1b and Table 6 in Appendix F.5), RMF-BSC achieves the best mean MMD on Volcano, Flood, and Fire, and ranks second to RMF-Semi on Earthquake. It consistently improves over RMF-PT and RMF-Eulerian across all four datasets, supporting the effectiveness of geodesic semigroup consistency on real-world spherical distributions. The visualizations in Fig. 4 further show that one-step samples capture the main geographic concentrations of held-out events, including the elongated patterns in Volcano and Earthquake and the broader regional clusters in Flood and Fire. On RNA (Table 7),

Table 6: **One-step MMD (↓) on the Earth benchmarks**, reported as mean ± standard deviation over 5 random seeds. Best results are shown in **bold** and second-best results are underlined.

Method	Volcano	Earthquake	Flood	Fire
RFM	0.1708 ± 0.0411	0.1666 ± 0.0082	0.1289 ± 0.0186	0.1463 ± 0.0034
RCT	0.1452 ± 0.0685	0.0501 ± 0.0108	$\underline{0.0501 \pm 0.0164}$	0.0487 ± 0.0057
G-PSD	0.1310 ± 0.0594	0.0408 ± 0.0083	0.0530 ± 0.0147	0.0393 ± 0.0094
RMF-PT	0.1248 ± 0.0446	0.0472 ± 0.0141	0.0588 ± 0.0184	0.0371 ± 0.0109
RMF-Eulerian	0.1397 ± 0.0398	0.0402 ± 0.0122	0.0649 ± 0.0145	0.0424 ± 0.0083
RMF-Semi	$\underline{0.1146 \pm 0.0378}$	$\mathbf{0.0329 \pm 0.0129}$	0.0603 ± 0.0254	$\underline{0.0292 \pm 0.0107}$
RMF-BSC	$\mathbf{0.1117 \pm 0.0395}$	$\underline{0.0355 \pm 0.0069}$	$\mathbf{0.0481 \pm 0.0140}$	$\mathbf{0.0227 \pm 0.0058}$

Table 7: **Results on the RNA and DNA benchmarks**, reported as mean with standard deviation in parentheses over 5 random seeds. Best means are shown in **bold** and second-best means are underlined.

Dataset	Metric	RFM	RCT	G-PSD	RMF-PT	RMF-Eulerian	RMF-Semi	RMF-BSC
RNA	1-step MMD ↓	0.1104 (0.0017)	0.2286 (0.0029)	0.1294 (0.0056)	0.1193 (0.0182)	0.1193 (0.0182)	<u>0.0846</u> (0.0028)	**0.0825** (**0.0081**)
DNA	FBD ↓	9.4969 (0.0475)	21.1888 (0.0278)	9.6420 (0.0369)	61.5611 (0.2806)	<u>3.7669</u> (0.0309)	5.4410 (0.0511)	**3.2888** (**0.0308**)
	k-mer corr. ↑	0.5935 (0.0009)	0.2446 (0.0050)	0.8807 (0.0011)	0.2340 (0.0007)	<u>0.9109</u> (0.0018)	0.7442 (0.0015)	**0.9457** (**0.0013**)

RMF-BSC achieves the best one-step MMD of 0.0825, compared with 0.0846 for RMF-Semi and 0.1193 for both RMF-PT and RMF-Eulerian. The improvement on the flat torus $\mathbb{T}^7$ suggests that the benefits of semigroup consistency and multi-split training extend beyond curved spaces. Since the reported RNA MMD aggregates per-angle marginal discrepancies, Fig. 7 complements this evaluation by visualizing the pairwise joint densities of (α, β) and (ϵ, ζ), estimated on identical bins from 947 held-out test samples and 947 one-step samples of each of the four RMF variants. Qualitatively, RMF-BSC is more concentrated around the dominant reference mode, most visibly for (ϵ, ζ), whereas the other variants spread mass over neighbouring bins. Nevertheless, the generated distributions remain more diffuse than the reference. On DNA (Table 7), RMF-BSC performs best on both metrics, reducing FBD from 3.7669 for the strongest baseline, RMF-Eulerian, to 3.2888, while increasing k-mer correlation from 0.9109 to 0.9457. These results extend its empirical benefits to high-dimensional conditional sequence generation.

F.5.3 Additional Ablation Study Results

The mixing weight λ. We vary λ in Eq. (15). As shown in Fig. 5(a), performance remains relatively stable across λ on Cone, Fisher, and Line, indicating limited sensitivity to the mixing weight on these distributions. In contrast, Swiss Roll shows a clearer MMD decrease as λ increases, likely because its stronger geometric nonlinearity amplifies composition errors, making BSC more effective. Overall, the results suggest that RMF-BSC is robust to the choice of λ, while larger BSC weights can provide additional gains on geometrically more challenging distributions.

Tangent versus geodesic residual. To isolate the effect of measuring semigroup consistency in endpoint geometry, we introduce RMF-Tan, which is identical to RMF-BSC ($K = 4$, the LDI term, h_ε^{-2} normalization, and the same λ) except that the geodesic residual in Eq. (13) is replaced by the tangent residual $h_\varepsilon^{-2}\|\xi_\theta - \xi_s\|_{g_x}^2$ with $\xi_s = \log_x(y_s)$ (Appendix D.3). RMF-Semi additionally differs from RMF-Tan in the number of splits, the LDI term, and the normalization. As shown in Table 8, moving from RMF-Semi to RMF-Tan improves Cone ($0.0305 \to 0.0247$) and Swiss Roll ($0.0833 \to 0.0708$), and replacing the tangent residual with the geodesic one yields a further gain of comparable size ($0.0247 \to 0.0216$ and $0.0708 \to 0.0582$). On Swiss Roll, the two changes contribute almost equally (-0.0125 and -0.0126). The geodesic residual also reduces seed variance on Cone (std $0.0036 \to 0.0003$). On Fisher and Line, the differences among the three variants are small in absolute terms (≤ 0.0027). These results are consistent with Proposition 4.3: the tangent

Table 8: **Ablation on the semigroup residual** on SO(3). One-step MMD (↓), mean ± std over 5 seeds. RMF-Tan matches RMF-BSC in K, the LDI term, normalization, and λ, and differs only in the residual geometry. RMF-Semi follows Woo et al. (2026).

Method	Residual	K	LDI	Cone	Fisher	Line	Swiss Roll
RMF-Semi	tangent	1	✗	0.0305±0.0040	0.0781±0.0035	0.0350±0.0002	0.0833±0.0179
RMF-Tan	tangent	4	✓	0.0247±0.0036	0.0767±0.0063	0.0349±0.0002	0.0708±0.0095
RMF-BSC	geodesic	4	✓	**0.0216**±0.0003	**0.0754**±0.0058	**0.0345**±0.0001	**0.0582**±0.0095

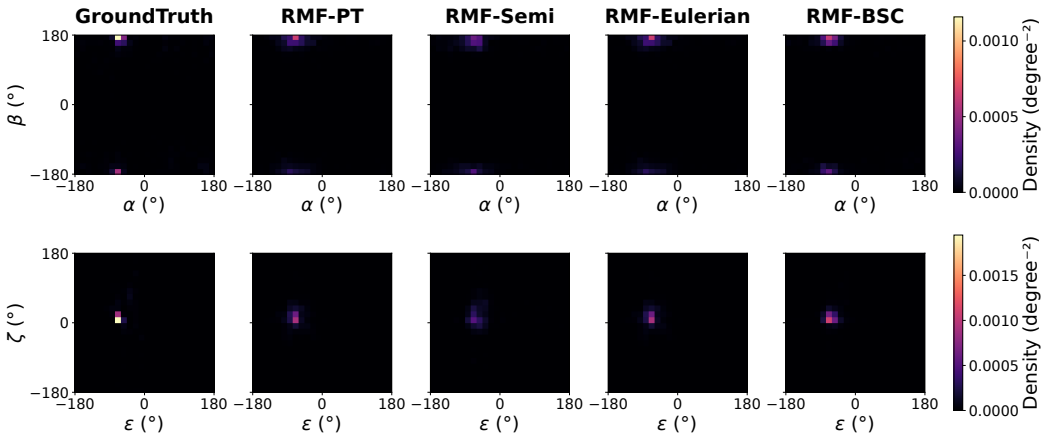


Figure 7: **One-step generation of RNA torsion-angle distributions.** The top and bottom rows show the joint densities of (α, β) and (ϵ, ζ), resspectively, comparing held-out data with samples from RMF-PT, RMF-Semi, RMF-Eulerian, and RMF-BSC. Each panel contains 947 samples, with identical binning and a shared color scale within each row. Brighter colors indicate higher probability density.

residual weights errors by g_x, whereas the sampler experiences them through $G_{x,\xi}$, and the two agree only to leading order as the displacement vanishes (Appendix D.2). The resulting mismatch is therefore expected to matter most on distributions that require long, strongly curved transport, such as Swiss Roll.

The number of splits K and computational overhead. We vary $K \in \{1, 2, 4, 8\}$ while keeping the remaining settings fixed. As shown in Fig. 5(b), performance remains largely stable across different choices of K, indicating limited sensitivity to the exact number of splits. Increasing K yields a clearer improvement on Swiss Roll, while the other datasets exhibit only minor fluctuations. Meanwhile, a larger K introduces additional computational overhead due to the increased number of forward passes. As reported in Table 9, increasing K from 1 to 8 raises the per-step training time by only $1.65\times$–$2.04\times$ across the four datasets, rather than scaling linearly with K. Together with the performance trends above, these results suggest that K provides a flexible trade-off between computational cost and the benefit of finer temporal splitting.

Table 9: **Runtime under different numbers of splits K.** Training time is measured in ms/step on a single NVIDIA RTX 4080S.

K	Cone	Fisher	Line	Swiss Roll
1	35.249	33.716	33.387	51.506
2	39.419	35.455	38.166	57.883
4	46.215	48.604	47.908	68.540
8	63.512	68.723	65.136	84.803
K=8/K=1	1.80×	2.04×	1.95×	1.65×

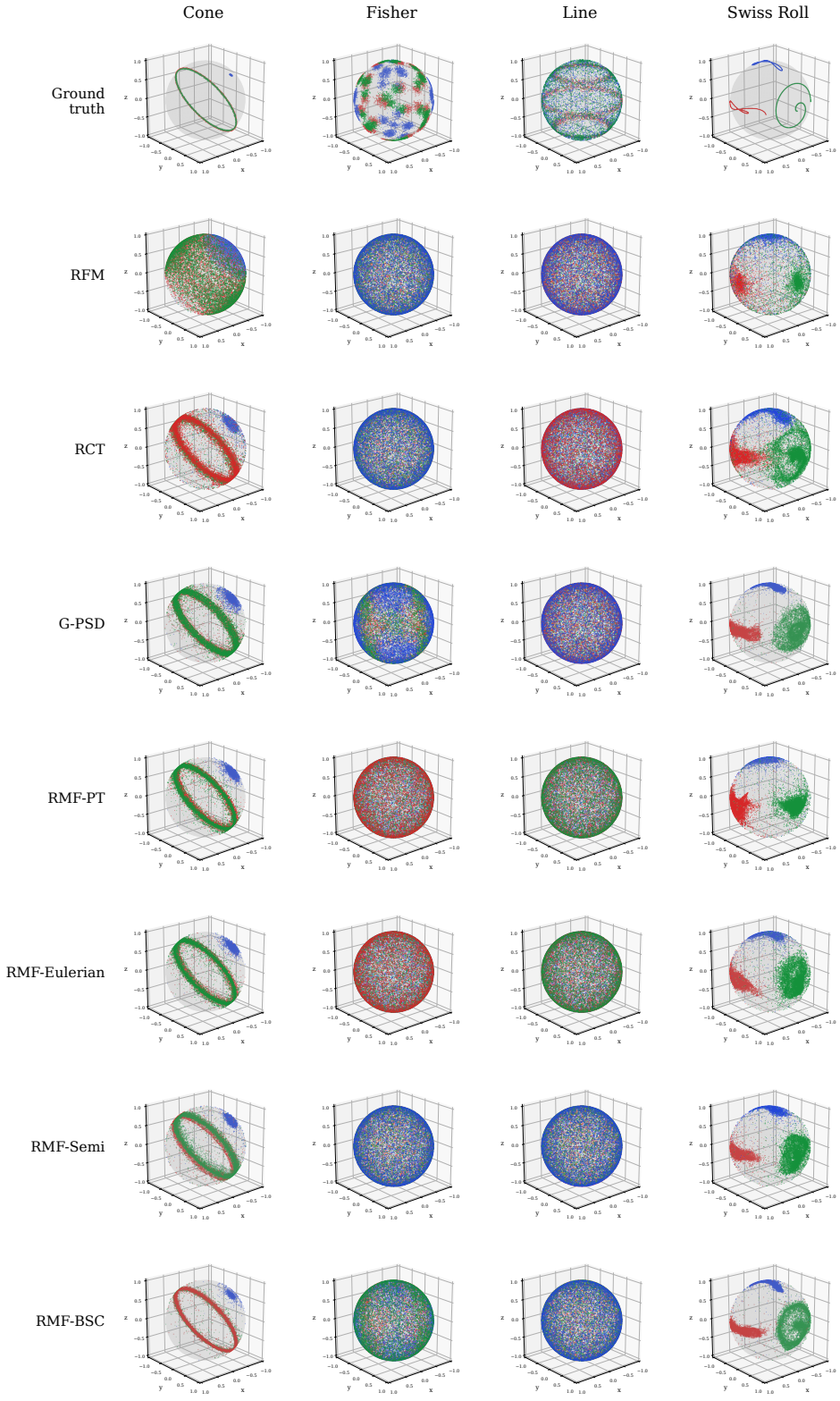


Figure 8: SO(3) one-step sample figure.