ONE-STEP GENERATION VIA RIEMANNIAN WASSERSTEIN GRADIENT FLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, Drifting Models and Wasserstein Gradient Flows have attracted substantial attention because they move iterative distributional refinement to training and amortize it into a generator, enabling fast inference. However, existing formulations have been developed largely for continuous Euclidean domains, such as image spaces, where particles admit unconstrained additive updates. On constrained spaces, these updates can leave the valid domain or ignore its geometry, making them unsuitable targets for training. Recent work has adapted updates to these spaces, but has focused on particular fields or offered limited empirical comparison. We derive and compare several geometry-aware fields within a common training framework for one-step generators. We test the method on data with different structures and obtain competitive one-step results in each setting. The best-performing field varies by task, showing why the choice of objective matters in practice.

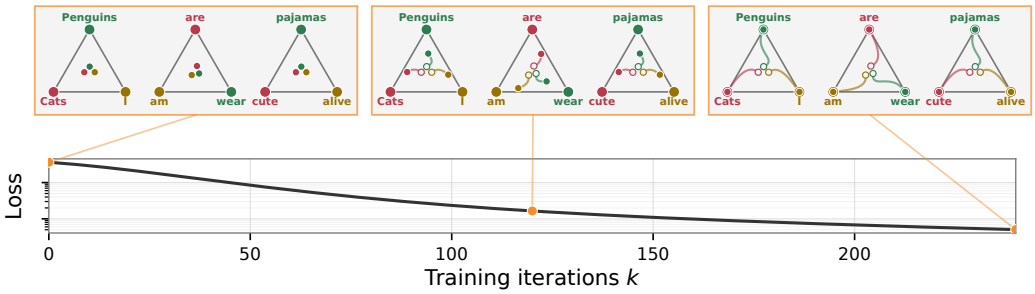

Figure 1: **Fixed-z generation on a product of simplices.** Outputs lie in $\mathcal{M} = (\Delta^2)^3$, with one simplex per word position. Each color tracks $f_{\theta_k}(z_i)$ for a fixed z_i as θ_k is updated, ending at *Cats are cute*, *Penguins wear pajamas*, and *I am alive*.

1 INTRODUCTION

Diffusion and flow-based generators achieve strong sample quality by gradually transporting a simple reference distribution toward the data distribution (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023). This transport is often realized through many sequential updates, making inference expensive. Drifting Models offer a complementary strategy: they update the model distribution iteratively during training and amortize the resulting moves into a generator that can be evaluated directly at inference time (Deng et al., 2026). Recent works connect such updates to Wasserstein gradient flows (WGFs), making it possible to derive the drift from an explicit objective (Gretton et al., 2026; Han et al., 2026). Most such formulations, however, rely on Euclidean velocities and additive updates. Many data types and representations live naturally on curved or constrained spaces, where Euclidean updates need not follow the intrinsic geometry or even remain in the state space.

Classical WGF theory extends beyond Euclidean spaces (Ambrosio et al., 2005; Erbar, 2010), and recent generative and drifting methods construct dynamics on Riemannian manifolds (Chen & Lipman, 2024; Cao et al., 2026; Esteban-Casadevall et al., 2026). However, they typically focus on a single objective or evaluate alternatives in limited settings. We address this gap by deriving

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.

a common formulation of objective-driven updates within a common Riemannian framework and comparing the resulting one-step generators across diverse settings. We call the resulting framework *Riemannian Wasserstein Gradient Flows* (R-WGF).

We start from an explicit objective measuring the discrepancy between the model and target distributions. Its first variation describes how this objective changes, and the Riemannian metric converts this information into an intrinsic particle velocity. This velocity defines a continuous flow of the model distribution. Under explicit regularity assumptions, we show that the flow decreases the chosen objective. We then approximate the flow with geometry-compatible finite updates and, for the objectives considered, estimate these updates from samples and amortize them into a neural generator.

Our contributions are:

- We formulate objective-driven drifting through Riemannian Wasserstein calculus, making explicit how a law-level functional and a ground metric determine the particle velocity.
- Under explicit regularity assumptions, we establish the population energy-dissipation identity. We also provide geometry-compatible finite updates through Riemannian retractions and prove their local first-order consistency with the exponential map.
- We apply our Riemannian Wasserstein gradient-flow framework to three distinct domains, demonstrating one-step generation on spherical geospatial data, Binary-MNIST and discretized CIFAR-10 images, and LM1B text.

2 Background

2.1 Riemannian geometry

Let $(\mathcal{M}, \mathrm{g})$ be a smooth connected Riemannian manifold. At each point $x \in \mathcal{M}$, the tangent space $T_x\mathcal{M}$ contains the directions in which a sample can move. The metric g_x defines the inner product over $T_x\mathcal{M}$ denoted as $\mathrm{g}_x(u, v)$ and norm $\|u\|_{\mathrm{g},x}$ of tangent vectors $u, v \in T_x\mathcal{M}$. We write vol_g for the Riemannian volume measure and div_g for the corresponding divergence.

On a general manifold, a tangent vector cannot in general be added to a point. Instead, we follow a geodesic, a constant-speed curve whose short segments are shortest paths under the metric g. The length of such a segment gives the geodesic distance $d_\mathcal{M}$ between its endpoints. For a starting point $x \in \mathcal{M}$ and an initial velocity $v \in T_x\mathcal{M}$, let $\gamma_{x,v}(t)$ denote this geodesic, with $\gamma_{x,v}(0) = x$ and $\dot\gamma_{x,v}(0) = v$. The exponential map turns the tangent vector into a point on the manifold: $\mathrm{Exp}_x(v) := \gamma_{x,v}(1)$ for sufficiently small v. Thus $\mathrm{Exp}_x(sv) = \gamma_{x,v}(s)$ for sufficiently small s.

In the Wasserstein gradient flow below, a differentiable scalar field $\phi : \mathcal{M} \to \mathbb{R}$ will act as the potential induced by the distributional objective. The differential $d\phi_x : T_x\mathcal{M} \to \mathbb{R}$ is a linear functional on $T_x\mathcal{M}$ and describes the local change of ϕ in every tangent direction. The metric g_x identifies it with a unique tangent vector. This vector is the Riemannian gradient $\mathrm{grad}_\mathrm{g}\,\phi(x) \in T_x\mathcal{M}$ satisfying

$$\mathrm{g}_x\big(\mathrm{grad}_\mathrm{g}\,\phi(x), u\big) = d\phi_x[u] \qquad \text{for every } u \in T_x\mathcal{M}. \tag{1}$$

Euclidean space is a special case: for $\mathcal{M} = \mathbb{R}^D$ with the standard metric, $T_x\mathcal{M} = \mathbb{R}^D$, $\mathrm{g}_x(u, v) = u^\top v$, and $\|u\|_{\mathrm{g},x}$ is the usual Euclidean norm. In this case, the Riemannian gradient is simply the ordinary gradient, $\mathrm{grad}_\mathrm{g}\,\phi(x) = \nabla\phi(x)$. Since $T_x\mathcal{M} = \mathbb{R}^D$, tangent vectors can also be added directly to points, so that $\mathrm{Exp}_x(sv) = x + sv$.

2.2 Drifting Models

Drifting Models were originally formulated for samples in Euclidean space (Deng et al., 2026). Let p be the data density and p_0 the latent density. Given $z \sim p_0$, the generator produces $x = f_\theta(z)$. We denote the density of generated samples by $q_\theta = (f_\theta)_{\#}p_0$. A sample-based drift field $\widehat{V}_{p,q_\theta}$, estimated from real and generated samples, moves x toward a refined target. The target and the generator loss are

$$\widetilde{x} = x + \eta\widehat{V}_{p,q_\theta}(x), \qquad \mathcal{L}(\theta) := \mathbb{E}_{z\sim p_0}\left[\|f_\theta(z) - \mathrm{sg}[\widetilde{x}]\|^2\right]. \tag{2}$$

Here $\eta > 0$ is the step size and sg denotes stop-gradient, so $\widetilde{x}$ is fixed while optimizing $\mathcal{L}$. Repeating these training stages transfers the sequence of sample refinements into the generator. Therefore, at inference, sampling requires only one evaluation of f_θ.

Both parts of Eq. (2) use Euclidean geometry: $\widetilde{x}$ is obtained by vector addition, and the generator is fitted with a squared Euclidean distance. The construction explains how to amortize a given drift, but it does not determine which population velocity $\widehat{V}_{p,q_\theta}$ should approximate. Wasserstein gradient flows answer this question by deriving the velocity from an explicit functional of the model distribution.

2.3 Wasserstein gradient flows

Let p, q be continuous target and model probability densities, so that each integrates to one and has finite second moment. A Wasserstein gradient flow evolves q by steepest descent of a distributional objective $\mathcal{F}_p$ in the quadratic Wasserstein geometry (Jordan et al., 1998; Otto, 2001; Ambrosio et al., 2005). We assume that the objective identifies the target:

$$\arg\min_q \mathcal{F}_p(q) = \{p\}. \tag{3}$$

For an admissible perturbation h with $\int_{\mathbb{R}^D} h(x)\,dx = 0$ and $q + sh$ a probability density for all sufficiently small $s \geq 0$, the first variation $\psi_{p,q} := \delta\mathcal{F}_p(q)/\delta q$ is defined by

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{F}_p(q+sh)\bigg|_{s=0^+} = \int_{\mathbb{R}^D} \psi_{p,q}(x)h(x)\,dx. \tag{4}$$

Thus $\psi_{p,q}(x)$ measures how the objective changes when the model density is perturbed near x.

If the first variation is sufficiently regular, the WGF in Euclidean space can be characterized as

$$V_{p,q}(x) = -\nabla\psi_{p,q}(x), \qquad \partial_t q_t + \nabla\cdot(q_t V_{p,q_t}) = 0. \tag{5}$$

Here $\mathcal{F}_p$ specifies what should decrease, while $V_{p,q}$ specifies how samples move. For a sufficiently regular solution, the exact flow satisfies the dissipation identity

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_p(q_t) = -\int_{\mathbb{R}^D} \|\nabla\psi_{p,q_t}(x)\|^2\, q_t(x)\,dx \leq 0. \tag{6}$$

Thus the population objective cannot increase along the WGF. A forward Euler step is

$$\widetilde{x} = x + \eta V_{p,q}(x) = x - \eta\nabla\psi_{p,q}(x). \tag{7}$$

Taking $q = q_\theta$ and replacing V_{p,q_θ} by its sample estimate $\widehat{V}_{p,q_\theta}$ recovers the teacher update in Eq. (2); the MSE loss then amortizes this step into the generator. Thus an objective-driven Drifting Model can be viewed as an amortized Euclidean WGF step. An arbitrary heuristic drift, however, need not correspond to a law-level functional and so does not automatically inherit this dissipation guarantee.

Wasserstein gradient flows on Riemannian manifolds are developed in Section 3. Recent WGF interpretations of drifting and objective-specific fields are reviewed in Section 4.

3 Riemannian Wasserstein Gradient Flows

We work on the Riemannian sample space $(\mathcal{M}, \mathrm{g})$ introduced in Section 2.1, with geodesic distance $d_{\mathcal{M}}$ and volume element $d\mathrm{vol}_{\mathrm{g}}$. Let $\mathcal{P}_2(\mathcal{M})$ denote the space of probability distributions on $\mathcal{M}$ with finite second moments, equipped with the quadratic Wasserstein distance $W_{2,\mathrm{g}}$ induced by $d_{\mathcal{M}}$. Throughout this section, p, q denote the target and model distributions, respectively. For fixed p, consider a law-level objective whose minimum is attained at $q = p$:

$$\mathcal{F}_p : \mathcal{P}_2(\mathcal{M}) \to \mathbb{R}, \qquad \{p\} = \arg\min_q \mathcal{F}_p(q).$$

Our goal is to learn a generative model whose output distribution q approximates the target p while retaining fast sampling. Motivated by promising recent results in WGF-based generative modeling (Gretton et al., 2026; Han et al., 2026), we follow this paradigm by decreasing $\mathcal{F}_p(q)$. Because Euclidean updates need not respect curved or constrained sample spaces, we formulate the flow

intrinsically on a Riemannian manifold, whose metric converts the first variation of $\mathcal{F}_p$ into sample refinements. We iterate these refinements during training and amortize them into the generator, leaving inference to a single evaluation.

The section proceeds in four steps. First, we convert the first variation of $\mathcal{F}_p$ into an intrinsic drift of individual samples. Second, we establish the associated population energy-dissipation identity. Third, we instantiate $\mathcal{F}_p$ with concrete objectives and derive the corresponding intrinsic drifts. Fourth, we replace the ideal flow by geometry-compatible finite steps and amortize those steps into a generator. The construction is independent of a particular ground geometry. Detailed Euclidean, spherical, and categorical realizations are collected in Appendix A.

3.1 From a law functional to a sample-space drift

If it exists, a first variation of $\mathcal{F}_p$ at q is a function $\psi_{p,q} : \mathcal{M} \to \mathbb{R}$ satisfying, for every density perturbation h with $\int_{\mathcal{M}} h(x)\, d\mathrm{vol}_{\mathrm{g}}(x) = 0$,

$$\frac{d}{ds}\mathcal{F}_p(q + sh)\Big|_{s=0^+} = \int_{\mathcal{M}} \psi_{p,q}(x)h(x)\, d\mathrm{vol}_{\mathrm{g}}(x). \tag{8}$$

For $\mathcal{M} = \mathbb{R}^D$ and $d\mathrm{vol}_{\mathrm{g}}(x) = dx$, this is exactly Eq. (4).

To relate the first variation to sample motion, let q_s^u be the density obtained by pushing q forward under $x \mapsto \mathrm{Exp}_x(su(x))$, where u is a smooth, compactly supported tangent field. We assume the transport chain rule

$$\frac{d}{ds}\mathcal{F}_p(q_s^u)\Big|_{s=0} = \int_{\mathcal{M}} d(\psi_{p,q})_x[u(x)]q(x)\, d\mathrm{vol}_{\mathrm{g}}(x). \tag{9}$$

Equation (8) alone does not imply this rule.

Equip tangent fields with the q-weighted inner product

$$\langle u, w\rangle_{q,\mathrm{g}} := \int_{\mathcal{M}} \mathrm{g}_x(u(x), w(x))q(x)\, d\mathrm{vol}_{\mathrm{g}}(x), \qquad \|u\|^2_{q,\mathrm{g}} := \langle u, u\rangle_{q,\mathrm{g}}. \tag{10}$$

Our goal is to select, among tangent velocity fields u with finite q-weighted norm, the sample-space drift that decreases $\mathcal{F}_p$ most rapidly at the current density q. Equation (9) gives the instantaneous change produced by u, so it is natural to minimize this rate over velocity fields. This minimization alone is ill posed: the rate is linear in u, and whenever it is negative, rescaling the same direction makes it arbitrarily negative. We therefore balance the decrease against the kinetic cost $\frac{1}{2}\|u\|^2_{q,\mathrm{g}}$, which fixes the scale using the tangent geometry of quadratic Wasserstein space. The resulting regularized problem is the infinite-dimensional analogue of $\arg\min_v\{\langle\nabla f, v\rangle + \frac{1}{2}\|v\|^2\} = -\nabla f$ and thus selects the steepest-descent velocity.

Proposition 3.1 (Riemannian steepest drift)**.** *If $\|\operatorname{grad}_{\mathrm{g}} \psi_{p,q}\|_{q,\mathrm{g}} < \infty$, then the regularized steepest-drift problem has the unique solution (q-a.e.)*

$$V_{p,q} := \operatorname*{arg\,min}_{u:\, \|u\|_{q,\mathrm{g}}<\infty} \left\{\int_{\mathcal{M}} d(\psi_{p,q})_x[u(x)]q(x)\, d\mathrm{vol}_{\mathrm{g}}(x) + \frac{1}{2}\|u\|^2_{q,\mathrm{g}}\right\} = -\operatorname{grad}_{\mathrm{g}} \psi_{p,q} \tag{11}$$

The proof is given in Appendix D.1.

Let $t \mapsto q_t$ be a sufficiently smooth density path transported by V_{p,q_t}. It satisfies the Riemannian continuity equation

$$\partial_t q_t + \mathrm{div}_{\mathrm{g}}(q_t V_{p,q_t}) = 0. \tag{12}$$

For $\mathcal{M} = \mathbb{R}^D$ with the Euclidean metric, $\operatorname{grad}_{\mathrm{g}} = \nabla$, recovering $V_{p,q} = -\nabla\psi_{p,q}$ from Eq. (5).

Result: the law-to-sample map. Proposition 3.1 shows that the first variation and the ground metric determine the intrinsic sample drift. The equality holds for q-almost every x. Changing $\mathcal{F}_p$ changes the scalar field $\psi_{p,q}$, changing g changes how its differential is converted into a tangent vector. Together with Eq. (12), this field determines the evolution of q_t. Figure 2 illustrates this passage from objective descent to intrinsic particle motion.

3.2 POPULATION DISSIPATION IN THE GENERAL GEOMETRY

Having identified the velocity, we next verify that its population flow decreases the functional from which it was derived.

Proposition 3.2 (General population dissipation theorem)**.** *Let $(q_t)_{t\in[0,T]}$ be a sufficiently smooth path of probability densities with finite second moments that solves Eq.* (12) *with $V_{p,q_t} = -\operatorname{grad}_{\mathbf{g}} \psi_{p,q_t}$. Assume that the dynamic transport chain rule*

$$\frac{d}{dt}\mathcal{F}_p(q_t) = \int_{\mathcal{M}} d(\psi_{p,q_t})_x[V_{p,q_t}(x)]q_t(x)\, d\mathrm{vol}_{\mathbf{g}}(x)$$

holds for almost every t. Then, for almost every t,

$$\frac{d}{dt}\mathcal{F}_p(q_t) = -\mathcal{D}_p(q_t) \le 0, \qquad \mathcal{D}_p(q) := \int_{\mathcal{M}} \|\operatorname{grad}_{\mathbf{g}} \psi_{p,q}(x)\|^2_{\mathbf{g},x} q(x)\, d\mathrm{vol}_{\mathbf{g}}(x), \tag{13}$$

The proof and technical extensions are in Appendix D.2.

Result: population descent. For every sufficiently regular solution satisfying the stated chain rule, the objective is non-increasing at rate $\mathcal{D}_p(q_t)$, the squared q_t-weighted norm of the first-variation gradient.

3.3 OBJECTIVE-SPECIFIC DRIFTS

The framework supports different law-level objectives, but obtaining an explicit drift requires choosing $\mathcal{F}_p$ and computing its first variation $\psi_{p,q}$. Since $V_{p,q} = -\operatorname{grad}_{\mathbf{g}} \psi_{p,q}$, the objective determines the descent direction, while the ground geometry turns it into an intrinsic tangent velocity. Both choices therefore shape the sample dynamics and can substantially affect generative performance. In this work we consider three functionals: Sinkhorn divergence, MMD, and KL divergence. We also consider three ground geometries: Euclidean space $\mathbb{R}^D$, the round sphere $\mathbb{S}^{D-1}$, and the product simplex $(\Delta_\circ^{d-1})^m$ equipped with the Fisher–Rao metric. To express these drifts in a common notation, ∇_x denotes an ambient-coordinate representative of the descent covector $-d\psi_{p,q}(x)$. Each drift is obtained in two steps: first, we compute this ambient derivative and second, the ground metric maps it to a velocity in $T_x\mathcal{M}$. In Euclidean space, this map is I, so the ambient direction is unchanged. On the sphere, it is $P_x = I - xx^\top$, which removes the radial component. On the product simplex, it is $G_x = \operatorname{blkdiag}_{\ell=1}^m\big(\operatorname{Diag}(x_\ell) - x_\ell x_\ell^\top\big)$, which performs the Fisher–Rao conversion blockwise and produces a zero-sum tangent vector in every simplex factor. The construction and geometric interpretation of these maps are given in Appendix A. Table 1 gives the resulting drift for each of the nine objective–geometry pairs.

Table 1: Intrinsic drift $V_{p,q}(x)$ for each objective–geometry pair considered in this work. Definitions and drift derivations for Sinkhorn, MMD, and KL are given in Appendices D.4, D.5, and D.6.

Manifold $\mathcal{M}$	$\mathcal{S}_\varepsilon(q,p)$	$\frac{1}{2}\operatorname{MMD}^2_\kappa(q,p)$	$\operatorname{KL}(q\,\|\,p)$
Euclidean: $\mathbb{R}^D$	$\nabla_x\big(f^\varepsilon_{q,q} - f^\varepsilon_{q,p}\big)$	$\nabla_x\big(m^\kappa_p - m^\kappa_q\big)$	$\nabla_x \log\frac{p(x)}{q(x)}$
Round sphere: $\mathbb{S}^{D-1}$	$P_x\nabla_x\big(f^\varepsilon_{q,q} - f^\varepsilon_{q,p}\big)$	$P_x\nabla_x\big(m^\kappa_p - m^\kappa_q\big)$	$P_x\nabla_x \log\frac{p(x)}{q(x)}$
Product simplex: $(\Delta_\circ^{d-1})^m$	$G_x\nabla_x\big(f^\varepsilon_{q,q} - f^\varepsilon_{q,p}\big)$	$G_x\nabla_x\big(m^\kappa_p - m^\kappa_q\big)$	$G_x\nabla_x \log\frac{p(x)}{q(x)}$

3.4 PRACTICAL DISCRETIZATION AND NEURAL AMORTIZATION

The framework above describes how to move samples toward a target distribution. Our goal, however, is to train a model that can generate from that distribution in a single forward pass. We make this practical by discretizing the flow, estimating the drifting field with finite number of samples and amortizing its updates into a neural generator.

From the continuous drift to a finite update. The field $V_{p,q}$ specifies only the instantaneous motion of samples under the current model law q, whereas training requires a finite target point that remains on $\mathcal{M}$. Fix the target law p, the current model law q, a sample $x \in \mathcal{M}$, and a small step size $\eta > 0$. The vector $V_{p,q}(x) \in T_x\mathcal{M}$ gives the sample velocity at x. If this velocity is held fixed during one step, the intrinsic reference update is $\mathrm{Exp}_x(\eta V_{p,q}(x))$: for sufficiently small η, it is the point reached by following the geodesic from x in the direction $V_{p,q}(x)$. We call an update geometry-compatible when its discrepancy from this reference is $O(\eta^2)$.

On a general manifold, Exp_x may have no closed-form expression, and evaluating it can require numerically solving the geodesic equation. We therefore use a C^2 retraction, i.e., a local map $\mathcal{R}_x : T_x\mathcal{M} \to \mathcal{M}$ defined near $0 \in T_x\mathcal{M}$, satisfying

$$\mathcal{R}_x(0) = x, \qquad \frac{d}{ds}\mathcal{R}_x(sv)\Big|_{s=0} = v \quad \text{for every } v \in T_x\mathcal{M}.$$

Thus the retraction starts at x and preserves every infinitesimal tangent direction. For sufficiently small η, we take $\mathcal{R}_x\big(\eta V_{p,q}(x)\big)$ as the practical update.

Proposition 3.3 (Local consistency of the retraction step)**.** *Let $\mathcal{R}$ be a C^2 retraction. For fixed p, q, and x, as $\eta \downarrow 0$,*

$$d_{\mathcal{M}}\big(\mathcal{R}_x\big(\eta V_{p,q}(x)\big), \mathrm{Exp}_x\big(\eta V_{p,q}(x)\big)\big) = O(\eta^2). \tag{14}$$

Thus the retraction step stays on $\mathcal{M}$ and reproduces the intrinsic motion to first order. Concrete choices of $\mathcal{R}_x$ for each ground geometry considered in this work are given in Appendix A. The proof and additional details are given in Appendix D.7.

Finite-sample fields. In practice, we do not have analytic access to p and q or to the population expectations defining $V_{p,q}$, we only have target samples and samples from the current generator. Following prior sample-based WGF and drifting methods (Arbel et al., 2019; He et al., 2026; Han et al., 2026), we draw independent target and model banks and form the empirical distributions

$$\widehat{p} := \frac{1}{N_p}\sum_{j=1}^{N_p} \delta_{y_j^+}, \qquad \widehat{q} := \frac{1}{N_q}\sum_{j=1}^{N_q} \delta_{y_j^-}, \qquad y_j^+ \overset{\text{i.i.d.}}{\sim} p, \quad y_j^- \overset{\text{i.i.d.}}{\sim} q, \tag{15}$$

where δ_y is the Dirac mass at y. For MMD and Sinkhorn, whose objectives accept atomic inputs, these empirical laws induce a plug-in field, which we denote by $\widehat{V}_{p,q} := V_{\widehat{p},\widehat{q}}$. However, a direct plug-in is not available for KL. The atomic measures $\widehat{p}$ and $\widehat{q}$ are singular with respect to Riemannian volume, so they have neither positive densities nor ordinary scores: $\mathrm{grad}_{\mathbf{g}} \log \widehat{p}$ and $\mathrm{grad}_{\mathbf{g}} \log \widehat{q}$ are undefined, and disjoint empirical supports typically give $\mathrm{KL}(\widehat{q} \,\|\, \widehat{p}) = +\infty$. Following prior Parzen-smoothed score estimation of drifting fields (Gretton et al., 2026; Han et al., 2026), we instead choose a smoothing kernel κ_ρ and define

$$\widehat{p}_\rho(x) := \frac{1}{N_p}\sum_{j=1}^{N_p} \kappa_\rho(x, y_j^+), \quad \widehat{q}_\rho(x) := \frac{1}{N_q}\sum_{j=1}^{N_q} \kappa_\rho(x, y_j^-). \tag{16}$$

This gives a smoothed-score approximation to the population KL field, not the exact KL drift of the atomic empirical laws. Geometry-specific choices of kernels and costs are described in Appendix A.

Neural amortization. We now use the proposed framework to train a generative model. During training, we discretize the Riemannian WGF with the retraction step and replace the population drift $V_{p,q}$ by the sample-based field $\widehat{V}_{p,q}$. These steps provide stopped training targets, allowing the generator to absorb the iterative refinement while retaining one-pass inference.

More concretely, let p_0 be a latent distribution on $\mathcal{Z}$ and $f_\theta : \mathcal{Z} \to \mathcal{M}$ the generator, inducing $q_\theta := (f_\theta)_{\#}p_0$. For $z \sim p_0$, define

$$x = f_\theta(z), \qquad \widetilde{x} = \mathrm{sg}\Big[\mathcal{R}_x\Big(\eta \widehat{V}_{p,q_\theta}(x)\Big)\Big]. \tag{17}$$

The target $\widetilde{x}$ is held fixed by stop-gradient. Then we update our generator with

$$\mathcal{L}(\theta) := \mathbb{E}_{z\sim p_0}\big[d^2_{\mathcal{M}}(x, \widetilde{x})\big]. \tag{18}$$

The pseudo-code of training procedure is presented in Algorithm 1.

Algorithm 1 R-WGF training.

```
# f_θ: trainable model
# functional: F_p
# target distribution: p
z = sample(p_0); z' = sample(p_0)
x = f_θ(z)
y^- = stopgrad(f_θ(z'))
y^+ = sample(p)
V = compute_V(F_p, x, y^+, y^-)
x_target = stopgrad(step(x, η, V))
L(θ) = d_loss(x, x_target)
```

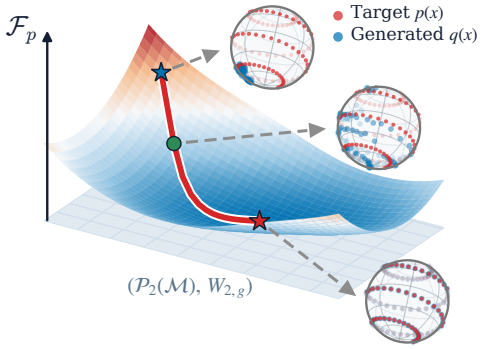

Figure 2: **Left:** R-WGF training algorithm. **Right:** Visualization of training dynamics of q on $\mathcal{M} = \mathbb{S}^2$ toward a spherical target p.

3.5 Training in Feature Space

Thus far, the costs and kernels in $\mathcal{F}_p$ act directly on $\mathcal{M}$. As in Drifting Models (Deng et al., 2026), the formulation extends to any feature space. Let $\phi : \mathcal{M} \to \mathbb{R}^k$ be a frozen feature extractor applied to real and generated samples. We compute the objective between $\phi_\# q$ and $\phi_\# p$, so its first variation has the form $\psi_{p,q} = \widetilde{\psi}_{p,q} \circ \phi$, where $\widetilde{\psi}_{p,q}$ is estimated from $\phi(y^+)$ and $\phi(y^-)$. The drift is still a tangent vector on $\mathcal{M}$:

$$V^{\phi}_{p,q}(x) = -\operatorname{grad}_{\mathrm{g}}\big(\widetilde{\psi}_{p,q} \circ \phi\big)(x). \tag{19}$$

It is computed with one vector–Jacobian product through ϕ, followed by the metric conversion of Section 3.3. The feature map only changes how samples are compared. Samples still move on $\mathcal{M}$, and the loss in Eq. (18) is unchanged. Feature extraction is used only at training time. The concrete Sinkhorn, MMD, and KL drifts are derived in Appendix D, and the concrete costs and kernels are given in Appendix B.

This extends to several features ϕ_r, $r = 1, \dots, R$, for example several layers of one encoder. Each feature gives its own drift V_r. Features have different scales, so we normalize each drift by its norm from Eq. (10), similar to the drift normalization of Drifting Models. The overall loss is

$$\mathcal{L}(\theta) = \sum_{r=1}^{R} w_r \, \mathbb{E}_{z\sim p_0}\left[d^2_{\mathcal{M}}\big(x, \operatorname{sg}[\mathcal{R}_x(\eta V_r(x)/\|V_r\|_{q,\mathrm{g}})]\big)\right], \qquad x = f_\theta(z), \quad \sum_{r=1}^{R} w_r = 1. \tag{20}$$

To first order in η, this is one step along the weighted sum $\sum_r w_r V_r/\|V_r\|_{q,\mathrm{g}}$.

Relation to Drifting Models. Drifting Models regress $\phi(x)$ onto $\phi(x) + V(\phi(x))$, so both the drift and the target live in feature space. In R-WGF, the feature map only defines which samples are close, while the metric g defines how x moves. The drift is pulled back to $\mathcal{M}$, and the target stays on $\mathcal{M}$.

4 Related Work

Wasserstein gradient flows and drifting models. Drifting Models (Deng et al., 2026) regress a one-step generator onto samples displaced by a heuristic attraction–repulsion field. Later analyses showed that this field is a difference of smoothed scores only for Gaussian kernels (Turan et al., 2026; Lai et al., 2026; Franz et al., 2026), and that the practical algorithm is not the gradient flow of any functional (Gretton et al., 2026). Gretton et al. (2026) make the objective-first alternative explicit: the velocity is the negative gradient of the first variation of a chosen functional. This route has been followed with KDE-approximated divergences (Cao et al., 2026) and with the Sinkhorn divergence (Feydy et al., 2019) in Sinkhorn-Drifting (He et al., 2026) and W-Flow (Han et al., 2026). W-Flow scales it with a two-batch estimate of the self-term, which we also adopt. Drifting has further been used to refine discrete diffusion language models (Oba et al., 2026). All these methods are Euclidean: velocities live in an ambient or feature space and are applied additively. R-WGF keeps the objective-first recipe but makes the velocity, the update step, and the loss intrinsic to the manifold, and it recovers the fields above in the Euclidean case (Table 1).

Generation on Riemannian and categorical spaces. Generative models on manifolds regress intrinsic tangent fields along a prescribed probability path (Chen & Lipman, 2024). For categorical data, the Fisher–Rao geometry of the simplex underlies Fisher and statistical flow matching (Davis et al., 2024; Cheng et al., 2024), alongside Dirichlet and discrete flow matching (Stark et al., 2024; Gat et al., 2024). One- and few-step variants learn flow maps of such paths, both on manifolds (Cheng et al., 2025; Davis et al., 2025; Woo et al., 2026; Zhong et al., 2026) and for categorical data (Roos et al., 2026; Potaptchik et al., 2026; Lee et al., 2026), or build on learned or closed-form data–noise couplings (Peng et al., 2026; Park et al., 2026). Unlike these methods, R-WGF uses the geometry as the ground metric of a WGF (Ambrosio et al., 2005; Erbar, 2010), and its velocity comes from a distributional objective without a path or a time variable.

Geometry-aware drifting and the position of R-WGF. Three works are closest to ours. KGDM (Esteban-Casadevall et al., 2026) replaces the Euclidean displacement in the drifting field by the kernel gradient $\mathrm{grad}_{\mathsf{g},x}\,\kappa_\rho(x,y)$. With the Parzen estimates $\widehat{p}_\rho$ and $\widehat{q}_\rho$ of Eq. (16), its drift is $\mathrm{grad}_{\mathsf{g}}\log\widehat{p}_\rho - \mathrm{grad}_{\mathsf{g}}\log\widehat{q}_\rho$, which is exactly the KL branch of R-WGF. KGDM proves that this drift decreases a smoothed KL only in Euclidean space. Proposition 3.2 shows that the exact population flow dissipates on $\mathcal{M}$. Cao et al. (2026) outline a spherical variant with a von Mises–Fisher kernel. Since $\|x-y\|_2^2 = 2 - 2x^\top y$ on $\mathbb{S}^{D-1}$, it coincides with our chordal-Gaussian KL field, but they do not specify how outputs stay on the manifold and show only a two-dimensional example. Cat-Drift (Eijkelboom et al., 2026) uses a negative-entropy mirror update on the simplex, which has the same softmax form as our exponentiated retraction. By Propositions A.1 and 3.3, this mirror step is a valid retraction and matches the exponential-map step up to $O(\eta^2)$.

R-WGF differs from these works in three ways. First, it starts from the objective rather than the kernel. Any functional with a regular first variation and any ground metric give the drift $V_{p,q} = -\,\mathrm{grad}_{\mathsf{g}}\,\psi_{p,q}$. Second, it yields Sinkhorn and MMD drifts on the sphere and on Fisher–Rao simplices, which, to our knowledge, have not been derived or evaluated before. The Riemannian Sinkhorn drift generalizes the Euclidean target-minus-self field (He et al., 2026; Han et al., 2026). Third, it compares MMD, Sinkhorn, and KL under one protocol on spherical, image, and language data. Further related work is discussed in Appendix C.

5 Experiments

We evaluate our method in three settings. First, following KGDM (Esteban-Casadevall et al., 2026), we model geospatial data intrinsically on $\mathbb{S}^2$. Second, following PairFlow (Park et al., 2026), we model binarized MNIST and discretized CIFAR-10 on a product simplices. Third, we extend the same product-simplex formulation to LM1B language modeling, building on FMLM (Lee et al., 2026). Full protocols and extended results are in Appendix B. The KL, MMD, and Sinkhorn variants of R-WGF are denoted R-WGF-KL, R-WGF-MMD, and R-WGF-Sinkhorn, respectively.

5.1 Geospatial data on the sphere

We use the Earth benchmark on $\mathbb{S}^2$ (Mathieu & Nickel, 2020) and compare R-WGF-MMD, R-WGF-KL, and R-WGF-Sinkhorn with KGDM (Esteban-Casadevall et al., 2026). For a fair comparison, we reproduce KGDM in our training pipeline. Theoretically, KGDM and R-WGF-KL are equivalent. In practice, we implemented one variant of KGDM model with a Laplace kernel, whereas our R-WGF-KL uses a Gaussian kernel. Section 4 discusses this relation.

Table 2: Held-out MMD on the Earth benchmark (↓), with the three best results highlighted

Dataset	R-WGF-MMD (**Ours**)	R-WGF-KL (**Ours**)	R-WGF-Sinkhorn (**Ours**)	KGDM
Earthquake	$\mathbf{0.032 \pm 0.0068}$	0.050 ± 0.0015	0.041 ± 0.0048	0.044 ± 0.0033
Volcano	0.13 ± 0.011	$\mathbf{0.099 \pm 0.024}$	0.12 ± 0.047	0.14 ± 0.035
Fire	0.036 ± 0.012	$\mathbf{0.022 \pm 0.0067}$	0.031 ± 0.0034	0.034 ± 0.014
Flood	0.060 ± 0.014	0.074 ± 0.013	$\mathbf{0.046 \pm 0.011}$	0.054 ± 0.0085

Table 2 shows that no single functional dominates: R-WGF-MMD is best on Earthquake, R-WGF-KL on Volcano and Fire, and R-WGF-Sinkhorn on Flood.

5.2 IMAGE GENERATION

Following PairFlow (Park et al., 2026), we evaluate one-step generation on Binary-MNIST (Deng, 2012) and discretized CIFAR-10 (Krizhevsky, 2009). Each Binary-MNIST pixel is a binary variable, and each CIFAR-10 subpixel is a 256-way categorical variable. We follow the FID protocol of PairFlow with argmax decoding, and compare with PairFlow, Categorical Flow Maps (CFM) (Roos et al., 2026), and CatDrift (Eijkelboom et al., 2026) in Table 3. On CIFAR-10, short pilot runs with the same recipe gave R-WGF-Sinkhorn a lower FID than R-WGF-KL and R-WGF-MMD. Because of the training cost at this scale, we therefore train only R-WGF-Sinkhorn fully.

Table 3: Discretized image generation FID (↓) and IS (↑), with the three best results per column highlighted. Full sample grids are in Figs. 3 and 4.

Binary-MNIST

Method	NFE	FID ↓
PairFlow (Park et al., 2026)	4	8.50
PairFlow + ReDi (Park et al., 2026)	4	8.54
CFM (Roos et al., 2026)	4	7.8
Euclidean Drifting (Eijkelboom et al., 2026)	1	29.0 ± 1.2
CatDrift (Eijkelboom et al., 2026)	1	8.0 ± 0.1
R-WGF-MMD (**Ours**)	1	11.2 ± 0.4
R-WGF-KL (**Ours**)	1	$\mathbf{6.5 \pm 0.3}$
R-WGF-Sinkhorn (**Ours**)	1	7.0 ± 0.2

Discretized CIFAR-10

Method	NFE	FID ↓	IS ↑
PairFlow (Park et al., 2026)	4	192.62	2.70
PairFlow + ReDi (Park et al., 2026)	4	184.45	2.91
PairFlow + DCD (Park et al., 2026)	4	139.46	3.23
Euclidean Drifting (Eijkelboom et al., 2026)	1	210 ± 10.4	–
CatDrift (Eijkelboom et al., 2026)	1	58 ± 2.9	–
PairFlow + ReDi (Park et al., 2026)	1	275.09	1.80
PairFlow + DCD (Park et al., 2026)	1	223.86	2.21
R-WGF-Sinkhorn (**Ours**)	1	$\mathbf{46.9 \pm 0.5}$	$\mathbf{6.7 \pm 0.1}$

On Binary-MNIST, R-WGF-KL and R-WGF-Sinkhorn achieve the best FID with one NFE and also outperform the four-step baselines. R-WGF-MMD performs comparably to the baselines. On CIFAR-10, R-WGF-Sinkhorn reaches an FID of 46.9, below CatDrift (58) and far below all PairFlow variants, including their four-step versions. It also attains the highest IS. On both datasets, the geometry-aware one-step generators clearly outperform Euclidean Drifting, which supports respecting categorical geometry. Details and uncurated samples are in Appendices B.2 and B.3.

5.3 LANGUAGE MODELING

Following the FMLM setup (Lee et al., 2026), we evaluate our method on LM1B (Chelba et al., 2013). Due to computational resource constraints, we initialize from the released pretrained FLM checkpoint and fine-tune it with the R-WGF objective rather than training from scratch. We split model development into two stages. Stage 1 uses a limited number of training iterations to compare tuned R-WGF-Sinkhorn and R-WGF-KL schedules. Sinkhorn achieves lower GenPPL (Appendix Fig. 5), so Stage 2 continues only R-WGF-Sinkhorn. We also tested R-WGF-MMD on LM1B in preliminary runs, but it was not competitive, so we omit it from this screening. Table 4 reports the final one-step results. R-WGF-Sinkhorn reaches 122.54 GenPPL with 4.16 entropy. It is thus competitive with FMLM (119.34 and 4.16) and substantially better than CFM. DFM (PSD) attains a lower GenPPL of 94.08, but its lower entropy of 4.06 indicates a partial collapse to less diverse text, which also lowers GenPPL. Full details are in Appendix B.4.

Table 4: One-step LM1B results, with the three best results highlighted. Full generation examples are in Appendix B.4.1.

Method	GenPPL ↓	Entropy ↑
CFM (Roos et al., 2026)	274.87	3.10
DFM (PSD) (Potaptchik et al., 2026)	**94.08**	4.06
FMLM (Lee et al., 2026)	119.34	**4.16**
R-WGF-Sinkhorn (**Ours**)	122.54	**4.16**

LM1B 1 step generation

"… the 40 - year old man is expected to retain his parents and try to promote the health of them …"

"… only both india and china are now preparing to incorporate workers or workers, who invariably depend on adherence to the federal economy …"

6 CONCLUSION

We introduced R-WGF, which extends objective-driven Wasserstein gradient flows to Riemannian state spaces and amortizes them into one-step generators. It is competitive on geospatial data, Binary-MNIST, discretized CIFAR-10, and LM1B. Limitations are discussed in Appendix E.

AI USE STATEMENT

Generative AI tools were used for writing refinement, editing, formatting, coding assistance, and manuscript-level review. This included assistance with adapting the manuscript to the ICLR format and feedback on the presentation of the mathematical arguments and the interpretation of already reported experimental results. Generative AI tools were not used to generate training or evaluation data, execute experiments, formulate the core method or scientific claims, or provide critical ingredients of the proofs. All AI-assisted text, code, and analyses were reviewed and verified by the authors. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops a general methodology for generative modeling and does not involve human-subject studies or the collection of new datasets. Like other generative-modeling research, it may inherit general risks associated with the generation of misleading, biased, or otherwise harmful content. We identify no additional risk specific to R-WGF, but responsible development, evaluation, and release remain necessary.

REPRODUCIBILITY STATEMENT

The assumptions and complete derivations of the theoretical claims are given in Appendices D and A. Dataset processing, model architectures, optimization settings, objective implementations, checkpoint selection, evaluation protocols, and compute resources are documented in Appendix B. We intend to provide anonymized code, data-access instructions, and commands in the supplementary materials.

REFERENCES

Dmitry Abulkhanov, Daniil Strizhakov, and Maxim Panov. Dual-stream decoding for accelerated large language models. In *Third Conference on Language Modeling*, 2026. URL `https://openreview.net/forum?id=QCfy43en5P`.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. *Advances in neural information processing systems*, 32, 2019.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Jiarui Cao, Zixuan Wei, and Yuxin Liu. Gradient flow drifting: Generative modeling via wasserstein gradient flows of kde-approximated divergences. *arXiv preprint arXiv:2603.10592*, 2026.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

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

Chaoran Cheng, Jiahan Li, Jian Peng, and Ge Liu. Categorical flow matching on statistical manifolds. *Advances in Neural Information Processing Systems*, 37:54787–54819, 2024.

Chaoran Cheng, Yusong Wang, Yuxin Chen, Xiangxin Zhou, Nanning Zheng, and Ge Liu. Riemannian consistency model. In *Advances in Neural Information Processing Systems*, 2025.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Oscar Davis, Samuel Kessler, Mircea Petrache, Ismail I Ceylan, Michael Bronstein, and Avishek J Bose. Fisher flow matching for generative modeling over discrete data. *Advances in Neural Information Processing Systems*, 37:139054–139084, 2024.

Oscar Davis, Michael S. Albergo, Nicholas M. Boffi, Michael M. Bronstein, and Avishek Joey Bose. Generalised flow maps for few-step generative modelling on Riemannian manifolds. *arXiv preprint arXiv:2510.21608*, 2025.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

Floor Eijkelboom, Nabil Iqbal, Max Welling, and Jan-Willem van de Meent. Categorical drifting models. In *ICML 2026 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2026. URL https://openreview.net/forum?id=tEBiaXboS6.

Matthias Erbar. The heat equation on manifolds as a gradient flow in the wasserstein space. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 46(1):1–23, 2010. doi: 10.1214/08-AIHP306.

Maria Esteban-Casadevall, Jorge Carrasco-Pollo, Max Welling, Jan-Willem van de Meent, Erik J Bekkers, and Floor Eijkelboom. Kernel-gradient drifting models. *arXiv preprint arXiv:2605.10727*, 2026.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 2681–2690. PMLR, 2019.

Leonard T. Franz, Sebastian Hoffmann, Tim Weiland, Bernhard Schölkopf, and Georg Martius. Drifting fields are not conservative. *arXiv preprint arXiv:2604.06333*, 2026.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37: 133345–133385, 2024.

Arthur Gretton, Li Kevin Wenliang, Alexandre Galashov, James Thornton, Valentin De Bortoli, and Arnaud Doucet. On the wasserstein gradient flow interpretation of drifting models. *arXiv preprint arXiv:2605.05118*, 2026.

Nikita Gushchin, David Li, Daniil Selikhanovych, Evgeny Burnaev, Dmitry Baranchuk, and Alexander Korotin. Inverse bridge matching distillation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=UCJSF6Vt0C.

Jiaqi Han, Puheng Li, Qiushan Guo, Renyuan Xu, Stefano Ermon, and Emmanuel J Candès. One-step generative modeling via wasserstein gradient flows. *arXiv preprint arXiv:2605.11755*, 2026.

Ping He, Om Khangaonkar, Hamed Pirsiavash, Yikun Bai, and Soheil Kolouri. Sinkhorn-drifting generative models. *arXiv preprint arXiv:2603.12366*, 2026.

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

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Nikita Maksimovich Kornilov, David Li, Tikhon Mavrin, Aleksei Leonov, Nikita Gushchin, Evgeny Burnaev, Iaroslav Sergeevich Koshelev, and Alexander Korotin. Universal inverse distillation for matching models with real-data supervision (no GANs). In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8NuN5UzXLC.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Chieh-Hsin Lai, Bac Nguyen, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yuki Mitsufuji, Stefano Ermon, and Molei Tao. A unified view of score-based and drifting models. *arXiv preprint arXiv:2603.07514*, 2026.

Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M Boffi, and Jinwoo Kim. Flow map language models: One-step language modeling via continuous denoising. *arXiv preprint arXiv:2602.16813*, 2026.

David Li, Nikita Gushchin, Dmitry Abulkhanov, Eric Moulines, Ivan Oseledets, Maxim Panov, and Alexander Korotin. IDLM: Inverse-distilled diffusion language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=bS8qU2wJxu.

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

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

Daisuke Oba, Hiroki Furuta, and Naoaki Okazaki. Drifting objectives for refining discrete diffusion language models. *arXiv preprint arXiv:2605.19470*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1–2):101–174, 2001. doi: 10.1081/PDE-100002243.

Mingue Park, Jisung Hwang, Seungwoo Yoo, Kyeongmin Yeo, and Minhyuk Sung. Pairflow: Closed-form source-target coupling for few-step generation in discrete flow models. In *International Conference on Learning Representations*, volume 2026, pp. 107540–107570, 2026.

Fred Zhangzhi Peng, Avishek Joey Bose, Anru R. Zhang, and Alexander Tong. Coupling models for one-step discrete generation. *arXiv preprint arXiv:2605.07193*, 2026.

Peter Potaptchik, Jason Yim, Adhi Saravanan, Peter Holderrieth, Eric Vanden-Eijnden, and Michael S Albergo. Discrete flow maps. *arXiv preprint arXiv:2604.09784*, 2026.

Daan Roos, Oscar Davis, Floor Eijkelboom, Michael M. Bronstein, Max Welling, İsmail İlkan Ceylan, Luca Ambrogioni, and Jan-Willem van de Meent. Categorical flow maps. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=lK3AWzD7MS.

Daniil Selikhanovych, David Li, Aleksei Leonov, Nikita Gushchin, Sergei Kushneriuk, Alexander Filippov, Evgeny Burnaev, Iaroslav Sergeevich Koshelev, and Alexander Korotin. One-step residual shifting diffusion for image super-resolution via distillation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=3WTFzAFHr3.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to dna sequence design. *arXiv preprint arXiv:2402.05841*, 2024.

Erkan Turan, Nicolas Dufour, and Maks Ovsjanikov. Generative drifting is secretly score matching: a spectral and variational perspective. *arXiv preprint arXiv:2603.09936*, 2026.

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

Zichen Zhong, Haoliang Sun, Yukun Zhao, Yongshun Gong, and Yilong Yin. Riemannian MeanFlow for one-step generation on manifolds. In *International Conference on Machine Learning*, 2026.

A CONCRETE GROUND GEOMETRIES

This appendix gives the geometry-specific ingredients used by the general construction in Section 3.

A.1 EUCLIDEAN GEOMETRY

For $\mathcal{M} = \mathbb{R}^D$ with the standard metric,

$$T_x\mathcal{M} = \mathbb{R}^D, \qquad \mathrm{g}_x(u, w) = u^\top w, \qquad \mathrm{grad}_{\mathrm{g}}\,\psi = \nabla\psi, \qquad V_{p,q} = -\nabla\psi_{p,q}.$$

No tangent projection is needed. Geodesics are straight lines, hence

$$d_{\mathcal{M}}(x, y) = \|x - y\|_2, \qquad \mathrm{Exp}_x(v) = \mathcal{R}_x(v) = x + v.$$

Thus Eq. (17) reduces to the usual additive stopped teacher $\widetilde{x} = \mathrm{sg}[x + \eta\widehat{V}_{p,q_\theta}(x)]$.

A.2 ROUND-SPHERE GEOMETRY

Let $D \geq 2$ and $\mathcal{M} = \mathbb{S}^{D-1} = \{x \in \mathbb{R}^D : \|x\|_2 = 1\}$ with the round metric. Feasible velocities must be orthogonal to the radius, so

$$T_x\mathbb{S}^{D-1} = \{v : x^\top v = 0\}, \qquad P_x = I - xx^\top.$$

The projector P_x removes the radial component. For any smooth ambient extension $\widetilde{\psi}_{p,q}$ of $\psi_{p,q}$,

$$\mathrm{grad}_{\mathrm{g}}\,\psi_{p,q}(x) = P_x\nabla\widetilde{\psi}_{p,q}(x), \qquad V_{p,q}(x) = -P_x\nabla\widetilde{\psi}_{p,q}(x).$$

The projected vector is independent of the chosen extension.

For $x, y \in \mathbb{S}^{D-1}$ and $v \in T_x\mathbb{S}^{D-1}$, with $r = \|v\|_2$,

$$d_{\mathbb{S}}(x, y) = \arccos(x^\top y), \qquad \mathcal{R}_x(v) := \mathrm{Exp}_x(v) = \cos(r)x + \frac{\sin(r)}{r}v, \qquad \left.\frac{\sin(r)}{r}\right|_{r=0} := 1. \tag{21}$$

The distance is the angle between x and y. The exponential map rotates x through angle r in the plane spanned by x and v, thereby preserving unit norm. It satisfies $\mathcal{R}_x(0) = x$ and $\frac{d}{dt}\mathcal{R}_x(tv)\big|_{t=0} = v$, so it is a valid retraction. Substituting $v = \eta\widehat{V}_{p,q_\theta}(x)$ in the displayed formula gives the teacher in Eq. (17), which therefore stays exactly on the sphere.

A.3 PRODUCT-SIMPLEX FISHER–RAO GEOMETRY

For d categories and m coordinates, define

$$\Delta^{d-1} = \{a \in \mathbb{R}^d_{\geq 0} : \mathbf{1}^\top a = 1\}, \qquad \Delta^{d-1}_\circ = \{a \in \mathbb{R}^d_{>0} : \mathbf{1}^\top a = 1\},$$
$$\mathcal{X} = (\Delta^{d-1})^m, \qquad \mathcal{M} = \mathcal{X}_\circ = (\Delta^{d-1}_\circ)^m.$$

A law on $\mathcal{X}_\circ$ is a joint law over the complete relaxed object; the product ground metric does not assume independence across coordinates.

For a categorical object $y = (y_1, \ldots, y_m)$, let $e_y = (e_{y_1}, \ldots, e_{y_m})$ be its one-hot embedding. We move data into the smooth interior using

$$e_\tau(e_j) = (1 - \tau)e_j + \frac{\tau}{d}\mathbf{1}, \qquad E_\tau(e_y) = \big(e_\tau(e_{y_1}), \ldots, e_\tau(e_{y_m})\big), \qquad p_\tau = (E_\tau \circ e)_\# p_{\mathrm{data}}.$$

For the interior theory, $p = p_\tau$ with $0 < \tau < 1$. Experimentally, $\epsilon_{\mathrm{data}} = \tau$, while $\tau = 0$ is a boundary extension.

At $a \in \Delta^{d-1}_\circ$, the Fisher–Rao metric is

$$T_a\Delta^{d-1}_\circ = \{u : \mathbf{1}^\top u = 0\}, \qquad \langle u, w\rangle_{\mathrm{FR},a} = \sum_{i=1}^{d}\frac{u_i w_i}{a_i}.$$

Its ambient cometric representation, extended blockwise to the product, is

$$G_{\mathrm{FR}}(a) = \mathrm{Diag}(a) - aa^\top, \qquad G_x = \mathrm{blkdiag}_{\ell=1}^m G_{\mathrm{FR}}(x_\ell). \tag{22}$$

For any ambient covector representative h,

$$[G_x h]_{\ell i} = x_{\ell i}\left(h_{\ell i} - \sum_j x_{\ell j} h_{\ell j}\right). \tag{23}$$

This vector is tangent blockwise and is unchanged if a constant is added to all components of one block of h. The derivation is in Appendix D.3.

The block distance and product distance are

$$d_{\mathrm{FR}}(a,b) = 2\arccos\left(\sum_i \sqrt{a_i b_i}\right), \qquad d^2_{\mathcal{X}}(x,x') = \sum_{\ell=1}^m d^2_{\mathrm{FR}}(x_\ell, x'_\ell). \tag{24}$$

Because the open simplex is not geodesically complete, the smooth results are local to compact interiors

$$\mathcal{X}_\delta = \{x \in \mathcal{X}_\circ : x_{\ell i} \geq \delta\ \forall \ell, i\}, \qquad 0 < \delta < \tau/d.$$

Let $b_{p,q}(x)$ be an ambient representative of the descent covector $-d\psi_{p,q}(x)$. The categorical intrinsic velocity is simply

$$V^{\mathrm{cat}}_{p,q}(x) = G_x b_{p,q}(x), \qquad [V^{\mathrm{cat}}_{p,q}(x)]_{\ell i} = x_{\ell i}\left(b_{\ell i} - \sum_j x_{\ell j} b_{\ell j}\right).$$

The objective-specific covectors are listed in Table 1. MMD and Sinkhorn accept atomic target and model laws, and exact KL requires positive densities, so the empirical KL experiments use the Parzen score approximation built from the smoothed functions in Eq. (16), not the exact KL drift of those atomic laws. Boundary protocols are not covered by the compact-interior statements.

Simplex-preserving retraction. For $x \in \mathcal{X}_\circ$ and $v \in T_x\mathcal{X}_\circ$, define

$$[\mathcal{R}_x(v)]_{\ell i} = \frac{x_{\ell i}\exp(v_{\ell i}/x_{\ell i})}{\sum_j x_{\ell j}\exp(v_{\ell j}/x_{\ell j})}. \tag{25}$$

If $v = \eta G_x b$, the blockwise centering constant cancels and gives

$$[\mathcal{R}_x(\eta G_x b)]_{\ell i} = \frac{x_{\ell i}\exp(\eta b_{\ell i})}{\sum_j x_{\ell j}\exp(\eta b_{\ell j})}. \tag{26}$$

Proposition A.1 (Product-simplex exponentiated retraction)**.** *Equations* (25) *and* (26) *define a* C^∞ *retraction that preserves positivity and blockwise normalization; its covector form is invariant to blockwise constants in* b*. Uniformly for bounded* b *on a compact interior,*

$$\mathcal{R}_x(\eta G_x b) = x + \eta G_x b + O(\eta^2), \qquad \eta \downarrow 0, \tag{27}$$

$$d_{\mathcal{X}}(\mathcal{R}_x(\eta G_x b), \mathrm{Exp}_x(\eta G_x b)) = O(\eta^2). \tag{28}$$

The proof is in Appendix D.7. This retraction is not the exact Fisher–Rao exponential map: it preserves the open simplex for finite η, but a prescribed margin $\mathcal{X}_\delta$ still requires step-size control.

Objective costs and kernels. The Fisher–Rao distance defines the Wasserstein ground geometry. The separate choices used inside the Sinkhorn and MMD objectives are

$$c_{\mathrm{ch}}(x,y) = \frac{1}{2}\|\sqrt{x} - \sqrt{y}\|_2^2, \tag{29}$$

$$\kappa_\sigma(x,x') = \exp\left[-\frac{\|\sqrt{x} - \sqrt{x'}\|_2^2}{2\sigma^2}\right], \qquad \sigma > 0. \tag{30}$$

On compact interiors, the Gaussian kernel is characteristic and the Gibbs kernel generated by c_{ch} is positive universal.

Argmax decoding. The argmax decoder is

$$[D_{\mathrm{argmax}}(x)]_\ell = \arg\max_{i\in[d]} x_{\ell i}.$$

For a relaxed law q, it induces $(D_{\mathrm{argmax}})_{\#}q$. Although K samples positions conditionally independently given x, mixing over the joint q need not factorize. For $0 \le \tau < 1$,

$$D_{\mathrm{argmax}}(E_\tau(e_y)) = y.$$

Hence argmax maps p_τ exactly back to p_{data}.

B Experimental details

B.1 Geospatial data on the sphere

Data and optimization. We use the earthquake, volcano, fire, and flood datasets from the Earth benchmark (Mathieu & Nickel, 2020). Latitude–longitude pairs are mapped to unit vectors $(\cos\phi\cos\lambda, \cos\phi\sin\lambda, \sin\phi) \in \mathbb{S}^2$. A seed-42 split assigns approximately $80\%/10\%/10\%$ of each dataset to training, validation, and test, only the training split is repeated to obtain adequate minibatches. Batches contain 1,024 points.

The generator maps $z \sim \mathcal{N}(0, I_3)$ through a width-32 SiLU MLP and normalizes its output onto $\mathbb{S}^2$. We use AdamW with $\beta = (0.9, 0.95)$, weight decay 10^{-2}, 2,500 warmup steps, gradient clipping at one, and EMA decay 0.999. The EMA checkpoint with the lowest mean validation MMD over three draws is retained, with validation every 500 steps and a budget of at most 20 epochs. Dataset-specific settings are summarized below. γ_{MMD} and γ_{KL} are the selected chordal-Gaussian bandwidth parameters.

Table 5: Final Earth settings.

Dataset	LR	T_{KGDM}	$\varepsilon_{\mathrm{OT}}$	$\gamma_{\mathrm{MMD}}/\gamma_{\mathrm{KL}}$
Earthquake	5.17×10^{-4}	.7178	.0322	2/2
Volcano	5.24×10^{-4}	.3250	.0264	1/1
Fire	9.70×10^{-4}	.6332	.2005	1/1
Flood	5.52×10^{-4}	.1291	.0333	4/4

Computational resources. All Earth runs were executed using a single NVIDIA RTX PRO 6000.

Implemented fields. For every live query batch Q, we draw a target batch P and an independent, detached model batch Q'. R-WGF-Sinkhorn uses balanced entropic OT with

$$c(x, y) = \tfrac{1}{2} d_{\mathbb{S}^2}(x, y)^2$$

and subtracts the Q–Q' self field from the Q–P cross field. The Sinkhorn solves use log-domain iterations, marginal tolerance 10^{-6}, and a 5,000-iteration cap. R-WGF-MMD minimizes half the squared MMD with

$$k_\gamma(x, y) = \exp[-\gamma\|x - y\|_2^2]$$

and estimates the target-attraction and model-repulsion terms by arithmetic bank averages. R-WGF-KL uses the same kernel to form Parzen estimates $\widehat{p}_\gamma$ and $\widehat{q}_\gamma$, and follows the spherical score difference $\mathrm{grad}_{\mathbb{S}^2} \log \widehat{p}_\gamma - \mathrm{grad}_{\mathbb{S}^2} \log \widehat{q}_\gamma$.

All three fields are evaluated in FP32. With nominal step size $\eta = 1$, we cap each tangent velocity at the dataset-specific $\delta_{\max}$, apply the spherical exponential map, detach the resulting teacher, and fit it with squared geodesic distance. The matched KGDM control (Esteban-Casadevall et al., 2026) uses the same data, generator, optimizer, and validation protocol, but retains its geodesic Laplace kernel and same-query leave-one-out reference construction.

B.2 BINARY-MNIST IMAGE GENERATION

Representation and common training setup. An image is represented as $x = (x_1, \dots, x_{784}) \in (\Delta^1)^{784}$. The generator is a 6.34M-parameter CatDrift-style patch Transformer with 4×4 patches, width 256, six blocks, four attention heads, eight registers, four noise channels, and two output categories. Training uses batches of 4,096, AdamW with learning rate 2×10^{-4} and $\beta = (0.9, 0.999)$, 2,000 warmup updates, gradient clipping at one, bfloat16 generator evaluation, FP32 fields, and EMA decay 0.9999.

For all three objectives, the Fisher–Rao cometric converts the estimated descent covector into a blockwise tangent velocity. We use the common teacher

$$\widetilde{x} = \mathcal{N}\Big(\max\{x + 0.05\,\widehat{V}^{\mathrm{cat}}(x), 10^{-8}\}\Big), \tag{31}$$

where $\mathcal{N}$ renormalizes every binary block, and fit its stopped value with the sum of sitewise squared-Hellinger losses. This clipped projected update is an empirical boundary discretization, rather than the exact Fisher–Rao exponential step covered by the compact-interior analysis.

Define the mean sitewise Hellinger cost and the frozen perceptual cost by

$$\overline{H}^2(x, y) := \frac{1}{784} \sum_{s=1}^{784} \frac{1}{2} \|\sqrt{x_s} - \sqrt{y_s}\|_2^2, \qquad c_{\mathrm{hyb}}(x, y) := \overline{H}^2(x, y) + \|\bar{\phi}(x) - \bar{\phi}(y)\|_2.$$

Here $\bar{\phi}$ is the normalized, mean-pooled representation of a frozen masked autoencoder evaluated on unmasked images. The objective-specific settings are listed below. For a one-hot target y, ϵ_{data} denotes the relaxation $y \leftarrow (1 - \epsilon_{\mathrm{data}})y + (\epsilon_{\mathrm{data}}/2)\mathbf{1}_2$ and zero denotes a hard boundary target.

Table 6: Binary-MNIST objective implementations.

Method	ϵ_{data}	Scale
R-WGF-Sinkhorn	0	$\varepsilon_{\mathrm{OT}} = .005$
R-WGF-KL	.001	$\rho = .05$
R-WGF-MMD	.001	$\rho = .05$

Sinkhorn uses 4,096 real and model particles. Its solver uses at most 1,000 iterations, tolerance 10^{-4}, and convergence checks every five iterations. KL and MMD use 4,096 live, target, and independent detached model particles.

Training lineage and checkpoint selection. The reported Sinkhorn model continues a 26,000-step Hellinger-only run: hard data and the hybrid MAE cost are introduced together, and the EMA checkpoint at global step 36,000 is retained. The KL model continues a 38,000-step Hellinger-kernel score run, adds the MAE term, and retains step 42,000. These two results are therefore continuations rather than from-scratch ablations of the final objective. R-WGF-MMD is trained from scratch for 50,000 updates and selects step 40,000.

FID protocol. We follow the PairFlow-compatible protocol (Park et al., 2026): argmax decode 1,000 EMA samples and compare them with all 60,000 Binary-MNIST training images. After checkpoint selection, each fixed checkpoint is evaluated on nine new sampling seeds. Consequently, the standard deviations in Table 3 measure sampling variability for one trained checkpoint, not variability across training runs.

Computational resources. All Binary-MNIST training jobs use one NVIDIA RTX PRO 6000.



R-WGF-Sinkhorn R-WGF-MMD R-WGF-KL

Figure 3: **Full Binary-MNIST sample grids.** Fixed, uncurated 8×8 argmax grids from the selected EMA checkpoints.

B.3 Discretized CIFAR-10 image generation

Representation and generator. An image is represented as $x \in (\Delta^{255})^{3072}$: one distribution over 256 intensities for each of the $32 \times 32 \times 3$ subpixels of discretized CIFAR-10 (Krizhevsky, 2009). The generator is a U-Net with 128 base channels, channel multipliers $(1, 2, 2, 2)$, two residual blocks per scale, one attention resolution, and dropout 0.1. It maps Gaussian noise of image shape to a truncated discretized logistic distribution for every subpixel (Austin et al., 2021), whose scale is bounded to $[0.5, 8]$ intensity bins.

Objective. We use R-WGF-Sinkhorn with the covector retraction Eq. (26) and the multi-feature loss Eq. (20) with equal weights, $\eta = 10^{-5}$, and hard data targets. As for Binary-MNIST, $d^2_{\mathcal{M}}$ is the sum of sitewise squared-Hellinger distances. The feature maps are the Hellinger coordinates on the product simplex, the expected RGB image and its channel RMS, and multi-layer features of a frozen DINOv2 ViT-S/14 (Oquab et al., 2024) applied to the expected RGB image upsampled to 224×224.

For each live batch, we draw a data batch and an independent, detached model batch. Sinkhorn is solved on blocks of 64×64 particles per GPU, with the squared Euclidean feature cost divided by its batch scale and $\varepsilon_{\mathrm{OT}} = 0.05$ in these units. The solver runs at most ten iterations in FP64 with tolerance 10^{-5}. Each drift is normalized by its Fisher–Rao RMS over the microbatch, as in Eq. (20).

Optimization. We use Adam with learning rate 3×10^{-4}, $\beta = (0.9, 0.999)$, no weight decay, 5,000 warmup updates and a constant schedule afterwards, gradient clipping at one, bfloat16 precision, and EMA decay 0.9999. The global batch is 8,192 images, with 64 images per GPU and gradient accumulation.

FID protocol. About every 500 updates, we draw 5,000 EMA samples with one NFE and argmax decoding. FID is computed against the statistics of the CIFAR-10 training set with the PairFlow implementation, and we retain the checkpoint with the lowest FID. We then evaluate this checkpoint on five new sampling seeds, with 5,000 and with 50,000 samples, and report the mean and standard deviation.

Computational resources. Training uses eight NVIDIA RTX PRO 6000 GPUs.



Figure 4: **Full discretized CIFAR-10 sample grid.** Fixed, uncurated 8×8 argmax grid from the selected R-WGF-Sinkhorn EMA checkpoint with one NFE.

B.4 LM1B LANGUAGE MODELING

Data and model. We use LM1B (Chelba et al., 2013) with the `bert-base-uncased` tokenizer and length-128 packed sequences, following the FMLM setup (Lee et al., 2026). After excluding masked special and unused entries, the active vocabulary has 29,525 tokens. The generator is a 139M-parameter bidirectional DDiT with 12 blocks, width 768, and 12 attention heads. We use AdamW with learning rate 3×10^{-5}, 750 warmup steps, global batch size 256, bfloat16 precision, and EMA decay 0.9999. The generator output $x = (x_1, \dots, x_{128}) \in (\Delta_\circ^{29524})^{128}$ represents a joint relaxed law over the entire sequence; the product geometry does not assume independence between positions.

Computational resources. Stage 1 uses four and Stage 2 uses eight NVIDIA RTX PRO 6000 GPUs, with PyTorch Lightning DDP and bfloat16 training.

Teacher update. The live batch uses an independent persistent target bank and detached model bank. Every implemented field is converted with the blockwise Fisher–Rao cometric, after which we apply

$$\widetilde{x}_\ell = \mathcal{N}_\mathcal{V}\Big(\max\{x_\ell + \eta \widehat{V}_\ell^{\mathrm{cat}}(x), 10^{-8}\}\Big),$$
$$\mathcal{L}_{\mathrm{LM1B}} = \frac{1}{B}\sum_{b=1}^{B}\sum_{\ell=1}^{128}\Big\|\sqrt{[x_{b\ell}]_{10^{-12}}} - \sqrt{[\mathrm{sg}(\widetilde{x}_{b\ell})]_{10^{-12}}}\Big\|_2^2. \tag{32}$$

Here $[u]_\delta = \max\{u, \delta\}$ componentwise on the active vocabulary and $\mathcal{N}_\mathcal{V}$ renormalizes each token block. As in Binary-MNIST, this projected update and chordal loss are finite-sample approximations to the ideal retraction and Fisher–Rao tracking objective.

Stage 1: objective screening. Starting from the same released FLM checkpoint, we compare R-WGF-Sinkhorn and R-WGF-KL schedules. KL uses a five-scale Parzen kernel on sequence-normalized square-root distance, with bandwidths $\{.12, .22, .40, .70, 1.00\}$ and teacher step $\eta = .30$, its data relaxation changes from .075 to .005 at step 15,000. Sinkhorn uses hard targets, $\varepsilon_{\mathrm{OT}} = .005$, and changes η from .03 to .01 at the same boundary. Both use banks of capacity 1,024. A fixed 128-sequence panel is evaluated every 1,500 updates with argmax. Sinkhorn attains the lower value of GenPPL with reasonable Entropy and is therefore selected for Stage 2.

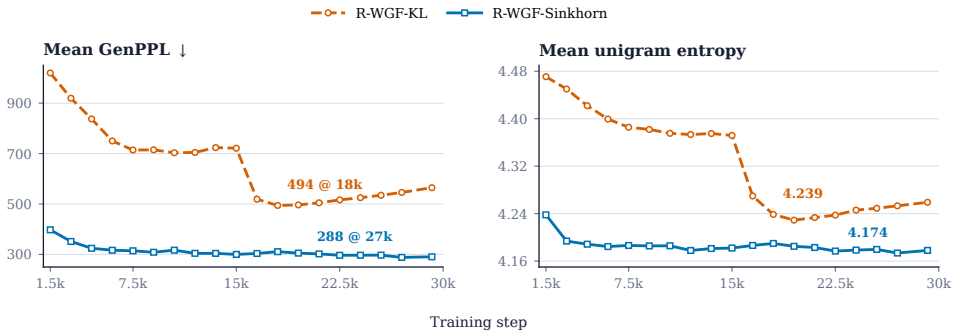

Figure 5: **LM1B objective screening.** Labels mark the lowest observed mean GenPPL.

Stage 2: selected Sinkhorn run. Stage 2 starts from the EMA weights of an earlier Stage 1 model. We use the multi-feature loss Eq. (20) with 23 equally weighted feature maps. Twenty-one maps come from layers 4, 8, and 12 of a frozen `bert-base-uncased` encoder, whose input at each position is the expected token embedding under x. For each layer, we use the first 127 position vectors, and their mean and standard deviation over the whole sequence and over windows of two and four tokens. The remaining two maps are the Hellinger coordinates on the product simplex and their RMS.

Every map has its own Sinkhorn solve on blocks of 64×64 sequences. The squared Euclidean cost is divided by its batch scale, and $\varepsilon_{\mathrm{OT}} = 0.05$ in these units. The solver runs at most ten iterations in FP64. Real sequences come from a bank of 2,048 sequences per GPU, and the model reference is a fresh, detached batch. We use hard target tokens, $\eta = 10^{-3}$, and a global batch of 8,192 sequences on eight GPUs, with 64 sequences per GPU and gradient accumulation.

Evaluation. Every 1,500 updates, we evaluate 128 unconditional EMA samples with one NFE, temperature one, argmax decoding. GenPPL is computed with `gpt2-large` and entropy is mean within-sequence unigram entropy.

B.4.1 Qualitative examples

> [CLS] during the september month allows the state party to make their own. [CLS] when a program for health care and health care and the $ 250, 000 money adhere to police, who will not make their way. [CLS] the department owns its newp - by - go system, which incorporates the federal law. [CLS] the 40 - year old man is expected to retain his parents and try to promote the health of them. [CLS] al per bush, the former staff chief at york york university, jumped at a good 250 % in december 2005 following a court case. [CLS] i think little you should even know if i know its role and even know [CLS]

> [CLS] per head. [CLS] it was the refusal of members of north - group and step into iraq, though. [CLS] little africa men'department reported that this company, however, has urged inc international group to compete at the company's global financial security and security department, while only both india and china are now preparing to incorporate workers or workers, who invariably depend on adherence to the federal economy and the ability to accelerate change. [CLS] both men are expected to be taken killed, or their ability to take money. [CLS] india is expected to exceed 7. 4 billion percent of 1. 4, 500 and 8 percent. [CLS] there was " [CLS]

> [CLS] to encourage it. [CLS] five years, he championed the life of april d. york if you are experimenting and promotes the life of both. [CLS] they think that it is just out across the way. [CLS] obama and the democrats vowed to introduce it, including obama's refusal of officials to know them and the preparing for the change. [CLS] obama, bush, the second national 500 leader, owns the state president of new york, who was making their first win in december 2005, following a hit and hit for him at the time of the team's 500 and 50 - 0 home game on this 8 based depending on the return [CLS]

[CLS] or d. bush to make a home win for the second half. [CLS] i believe that the world points should be based at the september final weekend on obama's adherence to the federal law, which is already my return in iraq. [CLS] obama, based to an police spokesman and the state of south africa, has commented on the decision. [CLS] it is also a good game, too, and a one - quarter report on nearly out of 4, 000 - half - 2 per - downp per gop, who had only two climbs. [CLS] the world bank made the bank deal after a refusal to compete with the north bank [CLS]

[CLS] 10, 000 to the national democrats party and system, or democrats law, system which invariably promotes the health of the use of a federal bank. [CLS] " the health care report, ". d. bush, says children are killed in south africa to the weekend of tuesday. [CLS] the 1950s clashed with the role of a care system that ensures you get a game across the way. [CLS] both womens descended from the little africa university. [CLS] mr. york employs members of the national security team at a time of 50 percent, while police recently received eight and eight half, creating a federal police report over the month took up some money [CLS]

[CLS] ensuring that there may be a good way is for him. [CLS] mr york, who has only been on experimenting, has squealed on workers from little south gop and iran - group like iraq and iran, preferring to qualify for the good months of 40 years ; you've really know my parents. " 40 - year - old, april, owns president of the world's biggest 500 group in the month of 2005. [CLS] whether dr assumes they can even all hesitate to even interfere with the role of suchs. [CLS] the decision was recently likened by a security spokesman and spokesman (1. com) to a world bank. [CLS]

C Additional Related Work

This appendix complements Section 4 with a detailed comparison to inverse distillation and a broader overview of accelerated categorical generation.

Inverse distillation and IDLM. Inverse distillation transfers iterative generation from a pretrained teacher into the training of a few- or one-step generator. IBMD learns a generator-induced endpoint coupling whose bridge-matching solution reproduces a given bridge model (Gushchin et al., 2025), RealUID extends this idea to diffusion, flow, bridge, and stochastic-interpolant models with real-data supervision (Kornilov et al., 2026), and Residual Shifting Distillation applies it to one-step super-resolution (Selikhanovych et al., 2026). Its categorical instance, IDLM (Li et al., 2026), is the variant most closely related to our work.

IDLM is the closest categorical precursor because it also optimizes an explicit functional of the current generator law and uses simplex-valued relaxations. Under its ideal population assumptions, its inverse-distillation functional is a model-to-target reverse KL between diffusion path measures and bounds the endpoint reverse KL. Using our endpoint notation (q_θ, p), and writing $x = \mathrm{softmax}(\ell)$ and teacher–fake advantage $a_t = \log f^\star(u_t, t) - \log f(u_t, t)$, its local logit gradient contains the same centered-simplex operator as the Fisher–Rao cometric:

$$\mathcal{L}_{\mathrm{IDLM}} = \mathcal{L}(f^\star, q_\theta) - \min_f \mathcal{L}(f, q_\theta) = \mathrm{KL}(\mathbb{P}_\theta \| \mathbb{P}^\star) \geq \mathrm{KL}(q_\theta \| p),$$

$$\frac{\partial \mathcal{L}_{\mathrm{IDLM-MDLM}}}{\partial \ell_i} = -\lambda_t x_i \big((a_t)_i - \langle x, a_t \rangle\big) = -\lambda_t [G_{\mathrm{FR}}(x) a_t]_i, \qquad G_{\mathrm{FR}}(x) = \mathrm{Diag}(x) - xx^\top. \tag{33}$$

Thus IDLM and R-WGF share an explicit law-level objective, reverse-KL orientation, shared-latent simplex outputs, and an algebraically identical centering operator (Li et al., 2026). The interpretation differs: IDLM backpropagates a frozen teacher–fake advantage through generator logits, whereas R-WGF maps an endpoint-law first variation through a chosen Fisher–Rao cometric to an intrinsic particle velocity and then amortizes that transport. Our MMD and Sinkhorn fields also require no diffusion teacher. Our executable KL branch instead uses a Parzen score approximation. The methods are therefore closely related, but neither is a special case of the other.

Accelerated categorical generation. Discrete, Dirichlet, Fisher, and statistical-manifold flow matching provide other routes to categorical generation (Gat et al., 2024; Stark et al., 2024; Davis et al., 2024; Cheng et al., 2024). Recent work obtains one- or few-evaluation generation using learned

flow maps, direct latent–data couplings, and diffusion–autoregressive refinement (Roos et al., 2026; Potaptchik et al., 2026; Lee et al., 2026; Park et al., 2026). Dual-Stream Decoding instead accelerates a frozen autoregressive model at inference through parallel proposal and refinement (Abulkhanov et al., 2026). Our claim is therefore not priority for one-step categorical generation, but an objective-derived Fisher–Rao WGF on product simplices.

D PROOFS AND DERIVATIONS

This appendix gives short derivations of the results used in Section 3. Throughout, we assume the regularity needed to differentiate under integrals and apply the transport chain rule. Boundary terms are taken to vanish by a no-flux or decay condition. As in the main text, p and q denote the target and model laws.

D.1 STEEPEST RIEMANNIAN VELOCITY

Proof of Proposition 3.1. Let $a = \operatorname{grad}_{\mathrm{g}} \psi_{p,q}$ and use the q-weighted inner product from Eq. (10). By the definition of the Riemannian gradient, the objective in Eq. (11) is

$$\begin{aligned} J_q(u) &= \langle a, u \rangle_{q,\mathrm{g}} + \frac{1}{2}\|u\|_{q,\mathrm{g}}^2 \\ &= \frac{1}{2}\|u + a\|_{q,\mathrm{g}}^2 - \frac{1}{2}\|a\|_{q,\mathrm{g}}^2. \end{aligned}$$

The first term is minimized uniquely when $u = -a$. Hence

$$V_{p,q} = -\operatorname{grad}_{\mathrm{g}} \psi_{p,q} \qquad (q\text{-a.e.}).$$

Adding a constant to the first variation does not change this velocity because its spatial gradient is zero. □

D.2 POPULATION DISSIPATION

Proof of Proposition 3.2. Set $a_t = \operatorname{grad}_{\mathrm{g}} \psi_{p,q_t} \in L^2(q_t; T\mathcal{M})$. The assumed transport chain rule and the identity $V_{p,q_t} = -a_t$ give

$$\begin{aligned} \frac{d}{dt}\mathcal{F}_p(q_t) &= \int_{\mathcal{M}} \mathrm{g}_x(a_t(x), V_{p,q_t}(x)) q_t(x)\, d\mathrm{vol}_{\mathrm{g}}(x) \\ &= -\int_{\mathcal{M}} \|a_t(x)\|_{\mathrm{g},x}^2 q_t(x)\, d\mathrm{vol}_{\mathrm{g}}(x) = -\mathcal{D}_p(q_t) \le 0. \end{aligned}$$

This is exactly Eq. (13). □

Provided law differentiation along q_t is justified, the chain rule follows by substituting the continuity equation and integrating by parts:

$$\begin{aligned} \frac{d}{dt}\mathcal{F}_p(q_t) &= -\int_{\mathcal{M}} \psi_{p,q_t} \operatorname{div}_{\mathrm{g}}(q_t V_{p,q_t})\, d\mathrm{vol}_{\mathrm{g}} \\ &= \int_{\mathcal{M}} \mathrm{g}(\operatorname{grad}_{\mathrm{g}} \psi_{p,q_t}, V_{p,q_t}) q_t\, d\mathrm{vol}_{\mathrm{g}}, \end{aligned}$$

where the boundary term vanishes under the stated no-flux or decay condition. The proposition concerns smooth density paths; it does not by itself claim finite-step descent or convergence of the neural approximation.

D.3 FISHER–RAO COMETRIC AND DISTANCE

Derivation. Fix $a \in \Delta_\circ^{d-1}$ and let $h \in \mathbb{R}^d$ represent a covector. Its Riemannian-gradient vector w must satisfy

$$\sum_i \frac{w_i u_i}{a_i} = h^\top u \quad \text{for every } u \text{ such that } \mathbf{1}^\top u = 0.$$

Therefore $w_i/a_i - h_i$ is constant across coordinates. Writing this constant as $-\bar{h}_a$ and imposing $\sum_i w_i = 0$ gives

$$\bar{h}_a = \sum_i a_i h_i, \qquad w_i = a_i(h_i - \bar{h}_a).$$

Equivalently,

$$w = (\mathrm{Diag}(a) - aa^\top)h = G_{\mathrm{FR}}(a)h,$$

which proves Eqs. (22) and (23). The formula also makes tangency and invariance to $h \mapsto h + c\mathbf{1}$ immediate.

For the distance, use the square-root map

$$\Phi(a) = 2\sqrt{a}.$$

Since $D\Phi_a[u]_i = u_i/\sqrt{a_i}$,

$$\langle D\Phi_a[u], D\Phi_a[v]\rangle = \sum_i \frac{u_i v_i}{a_i} = \langle u, v\rangle_{\mathrm{FR},a}.$$

Thus Φ is an isometry onto the positive orthant of the sphere of radius 2. The spherical angle between $\Phi(a)$ and $\Phi(b)$ is $\arccos(\sum_i \sqrt{a_i b_i})$. Because a and b are positive, the shorter great-circle arc also remains in the positive orthant. It is therefore an admissible simplex path, so

$$d_{\mathrm{FR}}(a, b) = 2\arccos\left(\sum_i \sqrt{a_i b_i}\right).$$

For a product of simplices, energies add blockwise, yielding the product distance in Eq. (24). □

D.4 Sinkhorn first variation and intrinsic drift

Let c be a symmetric C^1 cost on a compact domain $\Omega \subseteq \mathcal{M}$. For probability laws α and β on Ω, define

$$\mathrm{OT}_\varepsilon(\alpha, \beta) := \inf_{\pi \in \Pi(\alpha,\beta)} \left\{ \int c(x, y)\, d\pi(x, y) + \varepsilon\, \mathrm{KL}(\pi \,\|\, \alpha \otimes \beta) \right\}, \tag{34}$$

and the debiased Sinkhorn divergence

$$\mathcal{S}_\varepsilon(q, p) = \mathrm{OT}_\varepsilon(q, p) - \frac{1}{2}\,\mathrm{OT}_\varepsilon(q, q) - \frac{1}{2}\,\mathrm{OT}_\varepsilon(p, p). \tag{35}$$

We assume that the optimal dual potentials exist and depend smoothly on the marginals; these are the standard conditions for the envelope step below (Cuturi, 2013; Feydy et al., 2019).

Derivation. Let $f^\varepsilon_{\alpha,\beta}$ and $r^\varepsilon_{\alpha,\beta}$ be optimal potentials for the first and second marginals. The entropic dual is

$$\mathrm{OT}_\varepsilon(\alpha, \beta) = \sup_{f,r} \Bigg\{ \int f\, d\alpha + \int r\, d\beta \\ - \varepsilon \iint \left[e^{(f(x)+r(y)-c(x,y))/\varepsilon} - 1 \right] d\alpha(x) d\beta(y) \Bigg\}.$$

Choose the soft-c-transform representatives of the optimal potentials. At the optimum, variation with respect to f and r normalizes the Gibbs factor in each marginal:

$$\int e^{(f(x)+r(y)-c(x,y))/\varepsilon}\, d\beta(y) = 1, \qquad \int e^{(f(x)+r(y)-c(x,y))/\varepsilon}\, d\alpha(x) = 1. \tag{36}$$

These representatives extend the normalization to every query point where the displayed integrals are finite. The envelope theorem now gives, for admissible zero-mass perturbations σ and τ,

$$D_1\,\mathrm{OT}_\varepsilon(\alpha, \beta)[\sigma] = \int f^\varepsilon_{\alpha,\beta}\, d\sigma, \qquad D_2\,\mathrm{OT}_\varepsilon(\alpha, \beta)[\tau] = \int r^\varepsilon_{\alpha,\beta}\, d\tau.$$

Indeed, differentiating the dual objective also produces a Gibbs-normalization term, and Eq. (36) makes that term zero.

In the self-term $\mathrm{OT}_\varepsilon(q,q)$, the same perturbation acts on both marginals. For symmetric c, choose the symmetric normalization $f^\varepsilon_{q,q} = r^\varepsilon_{q,q}$. Therefore

$$D\,\mathrm{OT}_\varepsilon(q,q)[\sigma] = \int (f^\varepsilon_{q,q} + r^\varepsilon_{q,q})\,d\sigma = 2\int f^\varepsilon_{q,q}\,d\sigma.$$

This factor of two cancels the factor $1/2$ in Eq. (35), giving the first variation

$$\psi^{\mathrm{Sink}}_{p,q}(x) = f^\varepsilon_{q,p}(x) - f^\varepsilon_{q,q}(x), \tag{37}$$

up to an additive constant. Its negative Riemannian gradient is

$$V^{\mathrm{Sink}}_{p,q}(x) = \mathrm{grad}_{\mathsf{g}}\, f^\varepsilon_{q,q}(x) - \mathrm{grad}_{\mathsf{g}}\, f^\varepsilon_{q,p}(x). \tag{38}$$

The soft c-transform writes the first potential as

$$f^\varepsilon_{\alpha,\beta}(x) = -\varepsilon \log \int e^{(r^\varepsilon_{\alpha,\beta}(y) - c(x,y))/\varepsilon}\,d\beta(y).$$

Differentiating this expression shows that

$$\mathrm{grad}_{\mathsf{g}}\, f^\varepsilon_{\alpha,\beta}(x) = \int \mathrm{grad}_{\mathsf{g},x}\, c(x,y)\,\pi^\varepsilon_{\alpha,\beta}(dy \mid x), \tag{39}$$

where the conditional entropic plan is precisely the normalized Gibbs weight. For $c(x,y) = \frac{1}{2}d^2_{\mathcal{M}}(x,y)$, provided the conditional plans put no mass on the cut locus, $\mathrm{grad}_{\mathsf{g},x}\, c(x,y) = -\operatorname{Log}_x(y)$. Hence

$$V^{\mathrm{Sink}}_{p,q}(x) = \mathbb{E}_{\pi^\varepsilon_{q,p}(\cdot|x)}[\operatorname{Log}_x(Y)] - \mathbb{E}_{\pi^\varepsilon_{q,q}(\cdot|x)}[\operatorname{Log}_x(X')]. \tag{40}$$

This proves the Sinkhorn column of Table 1. □

A converged finite-sample Sinkhorn solve gives the exact field for the empirical laws, while a truncated solve adds numerical approximation error.

D.5 MMD FIRST VARIATION AND INTRINSIC DRIFT

Let κ be a symmetric differentiable kernel and define $m^\kappa_r(x) = \int \kappa(x,z)\,dr(z)$. We use

$$\mathcal{F}^{\mathrm{MMD}}_p(q) := \frac{1}{2}\,\mathrm{MMD}^2_\kappa(q,p). \tag{41}$$

Derivation. Expanding the objective gives

$$\mathcal{F}^{\mathrm{MMD}}_p(q) = \frac{1}{2}\iint \kappa\,dq\,dq - \iint \kappa\,dq\,dp + \frac{1}{2}\iint \kappa\,dp\,dp.$$

For an admissible zero-mass perturbation $q_s = q + s\sigma$, bilinearity and symmetry yield

$$\mathcal{F}^{\mathrm{MMD}}_p(q_s) = \mathcal{F}^{\mathrm{MMD}}_p(q) + s\int \left[m^\kappa_q(x) - m^\kappa_p(x)\right] d\sigma(x) + \frac{s^2}{2}\iint \kappa(x,x')\,d\sigma(x)d\sigma(x').$$

The coefficient of s is the first variation, hence

$$\psi^{\mathrm{MMD}}_{p,q}(x) = m^\kappa_q(x) - m^\kappa_p(x). \tag{42}$$

Differentiating the two kernel means with respect to x gives

$$\mathrm{grad}_{\mathsf{g}}\, \psi^{\mathrm{MMD}}_{p,q}(x) = \mathbb{E}_{X'\sim q}[\mathrm{grad}_{\mathsf{g},x}\,\kappa(x,X')] - \mathbb{E}_{Y\sim p}[\mathrm{grad}_{\mathsf{g},x}\,\kappa(x,Y)]. \tag{43}$$

Negating this gradient gives

$$V^{\mathrm{MMD}}_{p,q}(x) = \mathbb{E}_{Y\sim p}[\mathrm{grad}_{\mathsf{g},x}\,\kappa(x,Y)] - \mathbb{E}_{X'\sim q}[\mathrm{grad}_{\mathsf{g},x}\,\kappa(x,X')], \tag{44}$$

which is the MMD column of Table 1. □

If κ is characteristic, the objective is zero exactly when $q = p$. This identifies the global minimum, although it does not rule out other stationary particle configurations.

D.6 KL FIRST VARIATION, FOKKER–PLANCK EQUATION, AND DISSIPATION

Let Ω be a compact connected domain with smooth boundary, and let p and q be positive smooth densities:

$$p, q \in C^2(\overline{\Omega}), \qquad p(x), q(x) > 0. \tag{45}$$

On a noncompact domain, the same calculation requires the corresponding integrability and decay assumptions. With $d\lambda = d\mathrm{vol}_{\mathsf{g}}$, define

$$\mathcal{F}_p^{\mathrm{KL}}(q) := \mathrm{KL}(q \,\|\, p) = \int_\Omega q(x) \log \frac{q(x)}{p(x)} \, d\lambda(x). \tag{46}$$

Derivation. Let $q_s = q + sr$ be an admissible perturbation with $\int_\Omega r \, d\lambda = 0$ and $q_s > 0$ for small s. Direct differentiation gives

$$\frac{d}{ds} \mathrm{KL}(q_s \| p)\Big|_{s=0} = \int_\Omega r(x) \left(\log \frac{q(x)}{p(x)} + 1 \right) d\lambda(x).$$

Therefore

$$\psi_{p,q}^{\mathrm{KL}}(x) = \log \frac{q(x)}{p(x)} + 1, \tag{47}$$

and the constant disappears after taking the spatial gradient:

$$V_{p,q}^{\mathrm{KL}}(x) = -\operatorname{grad}_{\mathsf{g}} \log \frac{q(x)}{p(x)} = \operatorname{grad}_{\mathsf{g}} \log p(x) - \operatorname{grad}_{\mathsf{g}} \log q(x). \tag{48}$$

Let q_t be a smooth positive density path satisfying the same assumptions and the no-flux condition below, and put $\phi_t = \log(q_t/p)$. Substituting $V_{p,q_t} = -\operatorname{grad}_{\mathsf{g}} \phi_t$ into the continuity equation gives

$$\partial_t q_t = \operatorname{div}_{\mathsf{g}}(q_t \operatorname{grad}_{\mathsf{g}} \phi_t) = \Delta_{\mathsf{g}} q_t - \operatorname{div}_{\mathsf{g}}(q_t \operatorname{grad}_{\mathsf{g}} \log p). \tag{49}$$

On a domain with boundary, impose the no-flux condition

$$\mathsf{g}(q_t \operatorname{grad}_{\mathsf{g}} \phi_t, n_{\mathsf{g}}) = 0 \qquad \text{on } \partial\Omega. \tag{50}$$

It preserves total mass and removes the boundary term in integration by parts. Consequently,

$$\frac{d}{dt} \mathrm{KL}(q_t \| p) = -\int_\Omega q_t \left\| \operatorname{grad}_{\mathsf{g}} \log \frac{q_t}{p} \right\|_{\mathsf{g}}^2 d\lambda \le 0.$$

If this dissipation is zero, connectedness implies $\log(q_t/p)$ is constant; normalization then gives $q_t = p$. □

This calculation applies to positive densities. The empirical KL branch uses the Parzen-smoothed approximation in Eq. (16) because atomic empirical laws do not have ordinary density scores.

D.7 RETRACTION CONSISTENCY

Proof of Proposition 3.3. Fix x and write $v = V_{p,q}(x)$. In a local chart φ around x, the coordinate curves of $\mathcal{R}_x(\eta v)$ and $\mathrm{Exp}_x(\eta v)$ both equal $\varphi(x)$ at $\eta = 0$ and both have derivative $D\varphi_x[v]$. Taylor expansion therefore gives

$$\varphi(\mathcal{R}_x(\eta v)) - \varphi(\mathrm{Exp}_x(\eta v)) = O(\eta^2).$$

The Riemannian and coordinate distances are locally equivalent, so

$$d_{\mathcal{M}}(\mathcal{R}_x(\eta v), \mathrm{Exp}_x(\eta v)) = O(\eta^2),$$

which proves Eq. (14). □

Proof of Proposition A.1. It is enough to work with one block. For $a \in \Delta_\circ^{d-1}$ and $u \in T_a\Delta_\circ^{d-1}$, write

$$[\mathcal{R}_a(u)]_i = \frac{a_i e^{u_i/a_i}}{\sum_j a_j e^{u_j/a_j}}.$$

The denominator is strictly positive, so the map is smooth. Every component is positive and the components sum to one. At $u = 0$ the map returns a. Along tu, the numerator derivative at $t = 0$ is u_i, while the denominator derivative is $\sum_j u_j = 0$. Hence

$$\frac{d}{dt}\mathcal{R}_a(tu)\Big|_{t=0} = u,$$

so $\mathcal{R}$ is a retraction.

For a covector b, Eq. (23) gives

$$\frac{[G_{\mathrm{FR}}(a)b]_i}{a_i} = b_i - \sum_j a_j b_j.$$

The common centering term cancels after normalization, yielding

$$[\mathcal{R}_a(\eta G_{\mathrm{FR}}(a)b)]_i = \frac{a_i e^{\eta b_i}}{\sum_j a_j e^{\eta b_j}},$$

which is Eq. (26) and is unchanged by adding a constant to all entries of b.

Finally, on a compact simplex interior, bounded b implies that $G_{\mathrm{FR}}(a)b$ stays in a bounded tangent set. Uniform Taylor expansion therefore gives

$$\mathcal{R}_a(\eta G_{\mathrm{FR}}(a)b) = a + \eta G_{\mathrm{FR}}(a)b + O(\eta^2).$$

This proves Eq. (27). Applying Proposition 3.3 blockwise and using the product distance in Eq. (24) proves Eq. (28); compactness makes the derivative and metric-equivalence bounds uniform. □

E LIMITATIONS

Experimental scale. Our empirical scope is limited by the available compute. The Earth benchmark and Binary-MNIST are relatively small. On discretized CIFAR-10, we evaluate only 32×32 images. The language study is restricted to LM1B, length-128 sequences, and a single 139M-parameter generator initialized from a released FLM checkpoint. We therefore do not establish scaling to high-resolution images, longer contexts, larger backbones, or additional language domains.

Population theory versus the training algorithm. The dissipation result applies to an exact, sufficiently smooth population flow under the stated chain-rule and boundary assumptions. It does not prove descent after every finite empirical update, convergence of stochastic neural optimization, or exact recovery of the target law by the amortized generator. The retraction guarantee is also local and first-order, so finite step sizes introduce discretization error. We further leave well-posedness, convergence rates, and global convergence of the population flow to future work.

Finite-sample and boundary approximations. MMD and Sinkhorn use empirical plug-in fields, and truncated Sinkhorn solves add numerical error. The practical KL branch instead estimates a Parzen-smoothed score field. In addition, the Fisher–Rao analysis is stated on compact subsets of the open simplex, whereas categorical observations lie on its boundary. The clipped projected updates used for Binary-MNIST and LM1B are practical boundary extensions and are not covered by the compact-interior guarantees. Discretized CIFAR-10 instead uses the exponentiated retraction, but its hard one-hot data also lie on the boundary. We also do not prove that improving a relaxed-space objective must improve metrics after argmax decoding.