RIEMANNIAN ENERGY MATCHING FOR FIXED-DENSITY TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

An energy-based model defines a target density, but efficient sampling also depends on how particles move through it. *Energy Matching* uses the same scalar function to define both the density and the transport field. Improving transport by changing the energy therefore also changes the distribution being sampled. We ask whether learning the geometry of transport can improve sampling while keeping the energy fixed. We introduce *Riemannian Energy Matching* (REM), which learns a symmetric positive-definite mobility from optimal-transport velocities. This mobility redirects the frozen energy gradient, while a unit-determinant constraint rules out uniform time rescaling. We characterize the velocities this model can represent and derive an objective for fitting them along transport paths. A divergence correction ensures that the resulting diffusion preserves the original Boltzmann density in continuous time. On a curved multimodal target, full mobility captures motion that diagonal models miss. Removing the correction from the same fitted mobility substantially increases sampling error. A high-dimensional study shows that geometry can improve field fitting without changing the energy-defined density. For image generation, we use learned geometry during deterministic transport and return to Euclidean dynamics during stochastic refinement. This phase-gated strategy improves generation quality under matched solver budgets. REM provides a framework for separating errors in density learning from limitations in transport geometry, and for testing whether a better fitted field leads to better samples.

1 INTRODUCTION

Energy-based models (EBMs) define densities through a scalar potential (Song & Kingma, 2021), while Flow Matching learns transport fields directly (Lipman et al., 2023). Energy Matching (EM) couples these roles in one potential, whose negative gradient approximates an optimal-transport (OT) velocity and whose Boltzmann law defines the sampling target (Balcerak et al., 2025). After density learning, the Euclidean gradient may align poorly with a curved transport field, yet updating the energy also changes the density. We ask whether geometry can improve transport while the energy stays fixed.

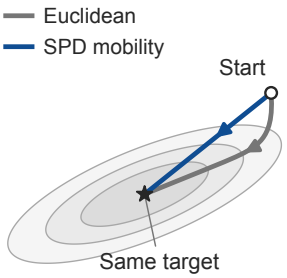


Figure 1: **Same density, different routes.** From a common start, Euclidean (gray) and determinant-one symmetric positive-definite (SPD, blue) descent approach the same Gaussian mode.

Projected Energy Matching (PEM) uses an auxiliary field to absorb rotational components during distillation (Barco et al., 2026). In our frozen-energy setting, an unrestricted additive field can improve velocity fit but does not by itself ensure density preservation. Classical state-dependent preconditioning redirects gradients and preserves the target density in continuous time through a divergence correction (Girolami & Calderhead, 2011; Ma et al., 2015). Diagonal mobility, however, is limited to coordinatewise scaling. Figure 1 illustrates the directional freedom of matrix-valued mobility at a fixed energy.

We introduce *Riemannian Energy Matching* (REM), which learns transport geometry while keeping the energy V fixed. A symmetric positive-definite (SPD) mobility matrix G_ϕ, parameterized by ϕ, maps the energy gradient to $-G_\phi \nabla V$ and is trained to fit target velocities from optimal transport. A

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.

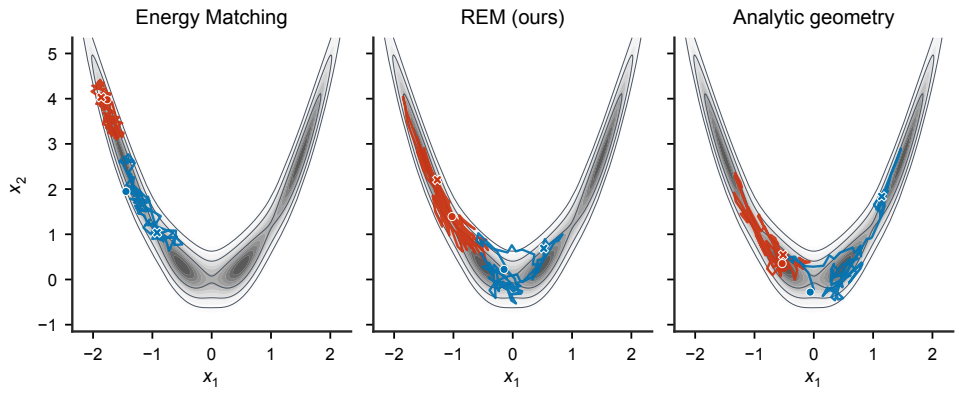


Figure 2: **Dynamics at the strongest coordinate warp.** REM follows the curved high-density channel and resembles the analytic reference dynamics, whereas Euclidean EM cuts across the coordinate-induced geometry.

full matrix couples coordinate directions through its off-diagonal entries, while the determinant-one constraint removes uniform time rescaling. With matching noise and a divergence correction, the resulting diffusion preserves the Boltzmann density of V in continuous time. Figure 2 illustrates the geometric effect in a controlled curved-target test. Here, *analytic geometry* uses the known mobility induced by the coordinate warp, and REM is fitted to its analytic transport field to test geometry recovery. The learned full mobility explores the curved channel similarly to this reference, whereas the Euclidean baseline shows more localized exploration in the displayed trajectories.

Our contributions.

- **A fixed-energy transport formulation.** We learn determinant-one SPD mobility by projecting external OT velocities onto a frozen energy gradient, with full, diagonal, and structured low-rank parameterizations.
- **Representation and sampling guarantees.** In dimension $d \geq 2$ with nonzero energy gradient, we constructively characterize the representable velocities as the strict descent half-space, establish regularity of the construction, and state the invariant-law conditions for the corrected diffusion, including an exact criterion for omitting the drift.
- **Controlled empirical evidence.** Paired curved-target experiments connect field fitting to mode passage and sampling accuracy, with a matched ablation isolating the divergence correction. High-dimensional and image studies separately assess representation capacity and finite-time generation.

2 Related Work

Energy-based models and transport learning. EBMs learn unnormalized densities through contrastive divergence or score matching (Hinton, 2002; LeCun et al., 2007; Song & Kingma, 2021). Flow Matching learns a time-dependent transport field (Lipman et al., 2023; Tong et al., 2024). Iterated Denoising Energy Matching (iDEM) learns a diffusion sampler from a prescribed Boltzmann energy (Akhound-Sadegh et al., 2024). Energy Matching fits the gradient of a scalar energy to transport supervision and refines the associated density (Balcerak et al., 2025). REM freezes that energy and learns mobility, changing the transport field while retaining the Boltzmann density.

Residual and flux formulations. PEM trains a flow teacher, absorbs rotational components in an auxiliary field during Helmholtz distillation, and updates the scalar energy (Barco et al., 2026). Our PEM-inspired control holds the energy fixed to isolate the residual representation. It does not reproduce that full training pipeline. Flux Matching learns generative fields by matching probability-flux divergence and can select nonconservative dynamics with desired properties (Pao-Huang et al., 2026). REM selects SPD mobility from external transport velocities. Its corrected diffusion is reversible, with freedom in the mobility and noise covariance.

Riemannian generative modeling. Geometry can enter these models at several levels. Riemannian score and flow models generate on a given intrinsic manifold (De Bortoli et al., 2022; Chen & Lipman, 2024). Learned metrics can also define geodesic paths or support trajectory inference (Kapuśniak et al., 2024; Bethune et al., 2025; Zweig et al., 2025; Sun et al., 2025; Scarvelis & Solomon, 2023). In REM, the geometry acts on a frozen ambient energy gradient and preserves a density relative to ordinary Lebesgue measure.

Preconditioned and mirror sampling. Position-dependent preconditioning and its invariant-law correction are classical (Girolami & Calderhead, 2011; Ma et al., 2015; Patterson & Teh, 2013). Learned mirror samplers such as the Convex Potential Mirror Langevin Algorithm (CPMLA) (Yang et al., 2025), time-dependent preconditioners, irreversible perturbations, score-prior inverse methods, and Hessian-based diffusion schedules likewise exploit geometry (Bhattacharya & Jiang, 2024; Falk et al., 2026; Zhang et al., 2022; Baldassari et al., 2025; Cao et al., 2026). Our question concerns the representation of a supplied velocity as $u = -G\nabla V$. A mirror map imposes inverse-Hessian integrability, whereas unrestricted SPD mobility need not.

3 Riemannian Energy Matching

3.1 Preliminaries on Energy Matching

Energy Matching (EM) learns a time-independent scalar potential $V_\theta : \mathbb{R}^d \to \mathbb{R}$ that serves both as a density model and as a generator (Balcerak et al., 2025). At a fixed temperature $\varepsilon > 0$, these roles are

$$p_\theta(x) = \frac{1}{Z_\theta} \exp\left(-\frac{V_\theta(x)}{\varepsilon}\right), \qquad v_\theta(x) = -\nabla V_\theta(x), \tag{1}$$

where θ contains the energy parameters and Z_θ is the finite normalizing constant.

Let a minibatch OT coupling γ_{OT} pair noise $x_0 \sim p_0$ and data $x_1 \sim p_{\mathrm{data}}$. The paired interpolation $x_t = (1-t)x_0 + tx_1$ has velocity $u_t = x_1 - x_0$. The transport term fits the negative energy gradient to these velocities by minimizing

$$\mathcal{L}_{\mathrm{EM}}^{\mathrm{OT}}(\theta) = \int_0^1 \mathbb{E}_{(x_0,x_1)\sim\gamma_{\mathrm{OT}}}\Big(\big\|\nabla V_\theta(x_t) + u_t\big\|^2\Big)\, dt. \tag{2}$$

After transport warm-up, EM jointly optimizes this term and a contrastive-divergence objective, using Langevin negatives to refine the density. Its sampling schedule combines deterministic transport with positive-temperature stochastic refinement.

Updating θ to change v_θ generally also changes p_θ. We instead freeze an analytic or pretrained energy and learn mobility acting on its gradient. Identity mobility recovers Eq. (1). The next subsection characterizes the additional directions available to REM.

3.2 Representing transport over a fixed energy

With V_θ fixed, we learn an SPD mobility $G_\phi : \mathbb{R}^d \to \mathbb{S}^d_{++}$, with transport and metric

$$v_{\theta,\phi}(x) = -G_\phi(x)\nabla V_\theta(x), \qquad g_\phi(x) = G_\phi(x)^{-1}, \qquad \det G_\phi(x) = 1. \tag{3}$$

Here $\mathbb{S}^d_{++}$ denotes the $d \times d$ SPD matrices. The determinant fixes mobility scale. Write $g = \nabla V_\theta(x)$ for the energy gradient and u for a target velocity. The classical gradient-like-field principle (Borodzik et al., 2016) characterizes the available pointwise freedom.

Proposition 3.1 (Fixed-energy representability)**.** *For $d \geq 2$ and $g \neq 0$, determinant-one SPD mobilities represent exactly the descent half-space,*

$$\big\{-Gg : G \in \mathbb{S}^d_{++},\ \det G = 1\big\} = \big\{u \in \mathbb{R}^d : u^\top g < 0\big\}. \tag{4}$$

If g, u are C^k fields on a compact set and $-u^\top g \geq \delta > 0$, a solution can be chosen C^k and uniformly elliptic there.

The result is constructive. Set $v = -u$, $r = \|g\|$, $e = g/r$, $a = e^\top v > 0$, $b = v - ae$, and $P = I - ee^\top$. One admissible mobility is

$$G_*(g,u) = \frac{a}{r}ee^\top + \frac{eb^\top + be^\top}{r} + \frac{bb^\top}{ar} + \left(\frac{r}{a}\right)^{1/(d-1)} P. \tag{5}$$

It satisfies $G_* g = -u$ and $\det G_* = 1$. The strict descent margin controls every denominator. Appendix A verifies the mapping, Schur complement, determinant, smoothness, and uniform ellipticity one step at a time.

At $g = 0$, an invertible mobility represents only $u = 0$, and smooth extension across joint zeros requires additional compatibility. Proposition 3.1 certifies existence within the full SPD class. A bounded network provides a restricted approximation. Along $\dot{X} = -G\nabla V$ with $G \succeq mI$, the chain rule gives $\frac{d}{ds}V(X_s) = -\nabla V^\top G \nabla V \le -m\|\nabla V\|^2$, so mobility cannot fit supervision that points uphill. We measure this obstruction through the held-out descent-violation rate (Appendix Tables 8 and 9).

3.3 Learning within the mobility class

We use full SPD mobility in two dimensions and diagonal mobility for images. In larger dimensions, we combine a determinant-one diagonal D_ϕ with a rank-four factor update,

$$Q_\phi = I + BC_\phi(x)^\top, \qquad G_\phi = \frac{D_\phi^{1/2} Q_\phi Q_\phi^\top D_\phi^{1/2}}{(\det Q_\phi)^{2/d}}. \tag{6}$$

Here r is the update rank, $B \in \mathbb{R}^{d\times r}$ has orthonormal columns, and the bounded output $C_\phi \in \mathbb{R}^{d\times r}$ keeps Q_ϕ invertible with positive determinant. Structured operations cost $O(dr^2 + r^3)$ without a dense image-sized matrix. All classes initialize at identity and have bounded eigenvalues. Appendix B.1 specifies their bounds and noise factors.

Given the OT coupling γ_{OT}, define $r_t = u_t + G_\phi(x_t)\nabla V_\theta(x_t)$. We learn geometry by minimizing the path-integrated projection loss

$$\mathcal{L}_{\mathrm{REM}}(\phi; W) = \int_0^1 \mathbb{E}_{\gamma_{\mathrm{OT}}}\Big(Z_u^{-1} \big\| u_t + G_\phi(x_t)\nabla V_\theta(x_t) \big\|^2_{W_\phi(x_t)} + \lambda_{\mathrm{geo}} \mathcal{R}_\phi(x_t) \Big)\, dt. \tag{7}$$

Figure 3 shows the loss at one interpolation site. With $\|r\|_W^2 = r^\top W r$, the Euclidean choice $W_\phi = I$ measures velocity error directly, while $W_\phi = G_\phi^{-1}$ measures it in the learned metric. Both vanish at an exact representer but can select different approximations in a restricted class. We use Euclidean fitting for synthetic tests, intrinsic fitting for images, and compare both in the ablations. Z_u normalizes velocity scale, while λ_{geo} weights the mobility penalty $\mathcal{R}_\phi$.

In practice we draw paired minibatches and a uniform time $t \sim U[0,1]$, evaluate the frozen energy gradient, and update only ϕ. This Monte Carlo estimate of Eq. (7) requires no trajectory integration during training. Algorithm 1 records the complete update, normalization, and optimizer. The fitted field is then coupled to a stochastic sampler.

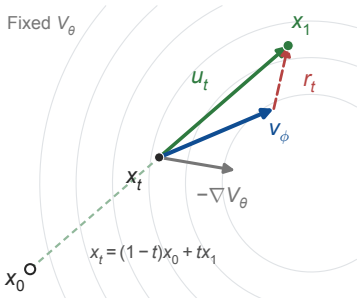


Figure 3: **One training site (schematic).** u_t (green), $-\nabla V_\theta$ (gray), $v_\phi = -G_\phi \nabla V_\theta$ (blue), and r_t (red).

3.4 Sampling the same density with learned geometry

For fixed temperature $\varepsilon > 0$, use the classical corrected diffusion (Girolami & Calderhead, 2011; Ma et al., 2015; Patterson & Teh, 2013), written as

$$dX_s = \Big(\underbrace{-G(X_s)\nabla V(X_s)}_{\text{transport}} + \underbrace{\varepsilon \operatorname{div} G(X_s)}_{\text{correction}} \Big) ds + \underbrace{\sqrt{2\varepsilon G(X_s)}}_{\text{noise factor}}\, dW_s. \tag{8}$$

Here W_s is standard Brownian motion and $(\operatorname{div} G)_i = \sum_j \partial_j G_{ij}$. The invariant density below does not depend on the OT fit or the determinant-one gauge.

Proposition 3.2 (Invariant density and reversibility)**.** *Let V and G be smooth, G uniformly elliptic, and $Z = \int_{\mathbb{R}^d} e^{-V/\varepsilon} dx < \infty$. Assume nonexplosive dynamics and sufficient tail decay for the boundary terms to vanish. Then Eq. (8) is reversible with invariant Lebesgue density $\pi = Z^{-1} e^{-V/\varepsilon}$. Its generator satisfies, for smooth compactly supported f, h,*

$$\int f \mathcal{L}_G h \, d\pi = -\varepsilon \int \nabla f^\top G \nabla h \, d\pi = \int h \mathcal{L}_G f \, d\pi. \tag{9}$$

For a time-dependent density ρ, the Fokker–Planck equation has probability current

$$\partial_s \rho = -\nabla \cdot J_G[\rho], \qquad J_G[\rho] = -G\Big(\rho \nabla V + \varepsilon \nabla \rho\Big). \tag{10}$$

The identity $\varepsilon \nabla \pi = -\pi \nabla V$ gives $J_G[\pi] = 0$. Writing $\mathcal{L}_G h = \varepsilon \pi^{-1} \nabla \cdot (\pi G \nabla h)$ and integrating by parts gives Eq. (9). Appendix A expands every derivative and boundary step and then uses symmetry of G to conclude reversibility.

Corollary 3.3 (When the correction can be omitted)**.** *Under the same regularity and boundary assumptions, the diffusion with drift $-G\nabla V$ and noise covariance $2\varepsilon G$ retains π exactly when*

$$\nabla \cdot ((\operatorname{div} G)\pi) = 0, \qquad J_{\text{no div}}[\pi] = -\varepsilon (\operatorname{div} G)\pi. \tag{11}$$

The criterion follows by removing $\varepsilon \operatorname{div} G$ from the current. Determinant one alone does not imply it. Reducing an integration step cannot repair a violated stationary equation. This gives a direct diagnostic for the ablation of the divergence drift.

A coordinate construction makes both claims testable. For a volume-preserving C^3 diffeomorphism T with Jacobian J_T, Euclidean Langevin dynamics for U push forward to Eq. (8) with $V = U \circ T^{-1}$ and $G = (J_T J_T^\top) \circ T^{-1}$. For the shear used in our experiments,

$$T_c(z) = (z_1, z_2 + c z_1^2), \qquad G_c(x) = \begin{bmatrix} 1 & 2cx_1 \\ 2cx_1 & 1 + 4c^2 x_1^2 \end{bmatrix}. \tag{12}$$

Here $\det G_c = 1$ and $\operatorname{div} G_c = (0, 2c)^\top$, yet the omitted current has divergence $-2\varepsilon c \, \partial_{x_2} \pi$, generally nonzero for $c \neq 0$. The controlled target therefore isolates a missing-drift effect from mobility-learning error. Appendix A derives the Itô pushforward and the additional Jacobian terms for a general coordinate map.

For image generation we use the EM schedule with deterministic transport before τ^*, followed by stochastic refinement. We switch from $G_s = G_\phi$ to $G_s = I$ at the geometry cutoff $\tau_g = \tau^* = 1$, confining learned mobility to its supervised phase. Every positive-temperature endpoint then has $G = I$ and zero divergence, so Euler–Heun needs no high-dimensional Jacobian probe. The main image benchmarks test this finite-time schedule. Synthetic targets and a small ungated CIFAR-10 screen test the active stochastic correction. Algorithm 2 gives the complete update.

4 Experiments

4.1 Experimental setup

We evaluate whether geometry learned over a frozen energy improves the fitted field, the sampled law, and finite-budget generation. Synthetic experiments use exact energies. The main curved target is a four-component Gaussian mixture transformed by a volume-preserving shear. It supports both ordinary minibatch-OT supervision and a known analytic field. We train identity, diagonal, full, constant-SPD, mirror, and residual representations for five seeds. An ablation with the same fitted mobility removes only $\varepsilon \operatorname{div} G$ from full REM. A correlated 64-dimensional Gaussian then compares a parameter-matched diagonal model, rank-four structured mobility, and continued energy fitting.

Within each synthetic comparison, samplers share initial states, budgets, temperatures, and reference draws. We measure held-out relative velocity mean squared error (MSE), first passage between modal cores (conditioned on arrival, with completion rates in the appendix), effective sample size (ESS), covariance error, and maximum mean discrepancy (MMD). We use the biased squared MMD estimator with a Gaussian kernel, in latent coordinates for the 2D shear and observed coordinates otherwise. Seeds are the independent analysis units. Figures show all paired seeds.

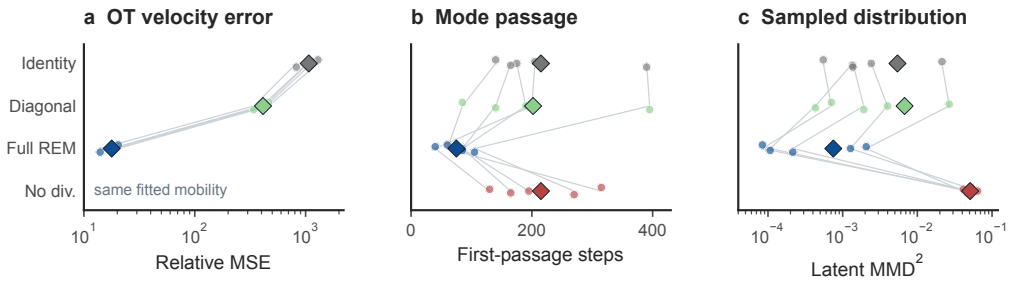


Figure 4: **OT fitting, mode passage, and sampled distribution.** Five paired seeds. Dots are runs, diamonds are means, gray lines connect seeds. Full REM and REM-NoDiv share one fitted model. Passage is the median among chains reaching another modal core. Completion rates and costs appear in Appendix Table 6.

For images, we freeze the published EM energies, which combine a U-Net with a vision transformer, and learn a 57,635-parameter diagonal mobility from OT chords. CIFAR-10 uses energy weights from step 147k and 325 Euler–Heun intervals. ImageNet32 uses weights from step 641k and 250 intervals. Three independently trained mobilities are each evaluated on three matched 50k-sample streams using Fréchet Inception distance (FID) and kernel Inception distance (KID). Controls compare identity, constant mobility, an additive field, objective weighting, rank-four structure, and solver budget. The principal sampler activates mobility only during deterministic transport. A separate 10k screen tests the active-temperature correction. Appendix C records objectives, schedules, sample counts, and hardware.

4.2 Learning curved transport from OT chords

We shear a four-component Gaussian mixture (means $(-1.5, 0), \ldots, (1.5, 0)$, variance .1) with $T_c(z) = (z_1, z_2 + cz_1^2)$. At $c = 1.2$, we fit ordinary minibatch-OT chords with identity, diagonal, and full SPD mobility. The learned paths in Appendix Figure 10 follow the same density contours differently. Figure 4 asks whether the field change also helps a finite-budget sampler. Each condition uses five paired seeds and 64 chains, with 2,000 burn-in and 6,000 measurement steps at $\Delta t = .002$.

Full REM reaches another mode in 75 steps, compared with 215 for identity and 202 for diagonal mobility. Its latent MMD is 7.5×10^{-4}, versus 5.41×10^{-3} for identity. These measurements connect the improved velocity fit to both passage and the sampled distribution under the same budget. To distinguish the learned geometry from its stochastic implementation, we then remove only the divergence drift from the same fitted full mobilities, giving the REM-NoDiv ablation. MMD rises to 5.088×10^{-2}, showing that the mobility alone is insufficient, even though its deterministic transport field is unchanged.

The fitted field nevertheless remains a projection of the OT evidence. On held-out chords, 21.7% violate the pointwise descent condition in Proposition 3.1. Full REM retains relative residual 17.8. The residual is not a geometry-learning failure alone. Even at zero shear, full and diagonal OT residuals are 1.57 and 22.67, so off-diagonal mobility can help OT projection without coordinate curvature. A separate test with a known geometric target therefore follows.

We isolate geometric recovery by supervising the known analytic field $u_{\mathrm{ref}} = -G_c \nabla V_c$, varying the shear c. Every representation uses the same sites and target velocities. We measure mobility recovery against G_c. At zero curvature all methods require 97 first-passage steps (Figure 5). At $c = 1.2$, full REM needs 111 steps, compared with 196 for identity and 256 for diagonal mobility. The corresponding relative metric errors, 0.00167 and 0.748 for full and diagonal mobility, explain why the diagonal class does not recover the same curved motion. Full path, occupancy, and return statistics appear in Appendix Table 14 and Figure 9.

Discretization error is a remaining alternative explanation for the uncorrected sampler. Figure 5(c) removes mobility-fitting error by holding the analytic geometry fixed while refining the step size. At $\Delta t = .001$, corrected MMD is 1.00×10^{-3} versus 1.746×10^{-2} uncorrected, and across $\Delta t \in \{.001, .002, .004, .008\}$ the uncorrected error stays near $.018$. Refinement therefore does not remove

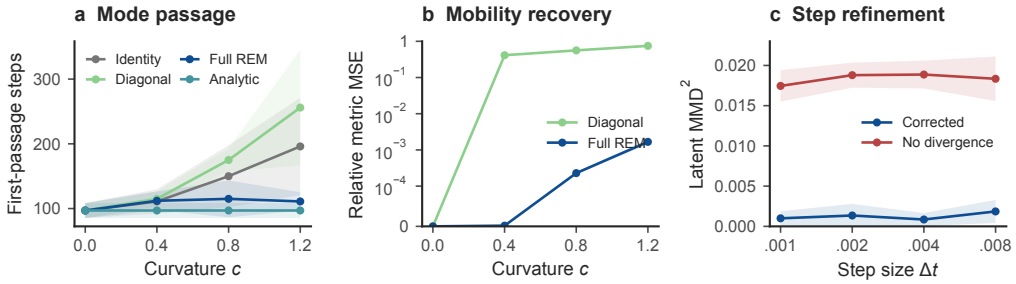


Figure 5: **Isolating geometry and correction.** Analytic-field supervision, five seeds. (a–b) Curvature sweeps over five seeds. (c) Step refinement with analytic mobility isolates the divergence correction from fitting error. Trajectories appear in Figure 2.

the discrepancy. Together with the OT ablation above, this supports the distinction in Eq. (10), which separates changing the field from preserving its target law.

4.3 Separating representation and density preservation

Table 1 relaxes the representation while keeping the analytic target field and training sites fixed, comparing full REM with free and penalized residual fields, mirror mobility, and constant SPD mobility. Residual and mirror networks approximately match full REM in parameter count. The free residual fits the field almost as well, but its additive stochastic extension has substantially higher distribution error. Mirror and constant-SPD mobilities preserve the target law under the corrected diffusion while fitting the curved field less accurately. The PEM-inspired and mirror controls isolate these representation classes rather than reproduce the full PEM or CPMLA pipelines.

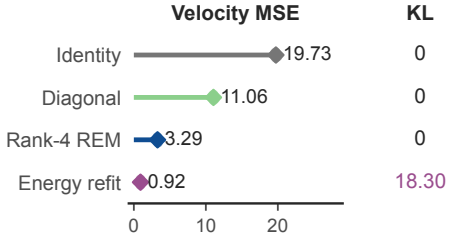


Figure 6: 64D field fit and density change, five seeds. Methods: identity, matched diagonal, rank-four REM, and continued energy fitting. Analytic Kullback–Leibler (KL) divergence is zero for frozen-energy mobilities.

This distinction also matters when increasing dimension. On a correlated 64D Gaussian with ordinary OT supervision, rank-four REM lowers held-out velocity MSE by 70.3% relative to the 66.1k-parameter diagonal model (Figure 6). Continuing to fit the exact quadratic energy achieves a still smaller residual, but incurs 18.301 analytic KL to the original density. The mobility models retain zero analytic KL because the energy is frozen. Finite-step sampling needs its own check, which currently gives a more limited picture. On a curved 16D mixture, structured mobility reduces field error by 20.7% without a consistent sampling advantage over identity, and the 64D Gaussian comparison likewise shows only a small correction gap (MMD 0.0135 versus 0.0141, Appendix Tables 12 and 13). We therefore

Table 1: **Matched representation controls.** Analytic-field supervision at $c = 1.2$, means over five seeds. All rows share energy, target velocities, and sampling budget. Velocity MSE is relative. Passage is measured in steps. The Law column denotes target-law preservation in continuous time, with correction where applicable.

Method	Representation			Fit and sampling		
	Class	Law	Params	Vel. MSE ↓	Latent MMD^2 ↓	Passage ↓
Identity (EM)	none	yes	0	1.968	1.61×10^{-3}	196
Free residual	free $r(x)$	no	33.7k	1.34×10^{-3}	1.47×10^{-2}	166
Penalized residual	penalized $r(x)$	no	33.7k	0.967	3.61×10^{-2}	1316
Mirror mobility	$[\nabla^2\Phi]^{-1}$	yes	32.8k	0.982	1.63×10^{-3}	418
Constant SPD	constant G	yes	3	1.244	2.64×10^{-3}	279
Full REM	SPD $G(x)$	yes	33.8k	$\mathbf{1.08\times10^{-3}}$	$\mathbf{1.08\times10^{-3}}$	**111**

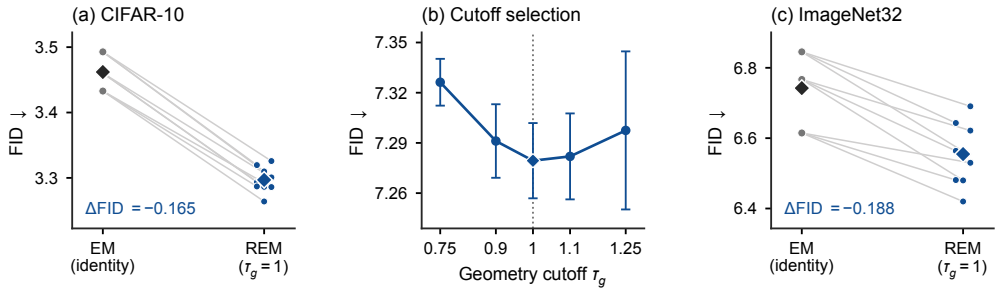


Figure 7: **Matched image results.** (a,c) Three fits, three shared streams, 50k samples/run. Lines pair runs, diamonds mark means, REM offsets separate fits. (b) Cutoff: mean ± SD, three 10k streams. Dotted line: selected $\tau_g = 1$. Negative ΔFID favors REM. Grids: Appendix G.

distinguish high-dimensional field-fitting capacity from the stronger mode-passage and correction evidence on the curved 2D target. The image experiments then examine how the learned field affects finite-time generation from an already trained energy.

4.4 Transport over frozen image energies

We fit a 57,635-parameter diagonal mobility for 150k steps on each frozen energy, using intrinsic residuals. The energy, initialization streams, and solver budget are shared within each comparison (Section 4). Each of three mobility fits is evaluated on three matched streams of 50k samples. Mobility acts during deterministic transport, with identity mobility during positive-temperature refinement, so these benchmarks evaluate finite-time generation.

Table 2: **Frozen image energies.** 50k per stream. Means over three fits for REM/additive and three streams otherwise.

Method	**FID** ↓	**KID** $\times 10^4$ ↓
CIFAR-10 (325 intervals)		
Matched EM ($G = I$)	3.462	9.19
Constant SPD	3.420	9.04
Additive residual	3.686	9.35
REM ($\tau_g = 1$)	**3.297**	**7.41**
ImageNet32 (250 intervals)		
Matched EM ($G = I$)	6.742	–
REM ($\tau_g = 1$)	**6.555**	–

At the main budgets, REM improves FID in all 9/9 paired streams on each dataset (Table 2). Figure 7 retains the per-fit and per-stream variation behind the aggregates. The matched constant-SPD and additive-residual controls share the same CIFAR energy and schedule, so the benefit is not explained by added parameters or an arbitrary velocity residual.

Independent of the intrinsic training objective, we evaluate the learned field in the Euclidean norm on 24,576 held-out OT pairs per dataset. Error falls 8.3% on CIFAR-10 and 11.2% on ImageNet32, improving 77.9% and 82.7% of pairs (Table 3a), with fit-to-fit log-mobility map correlations of .895 and .904 and fewer than .21% of coordinates near the imposed bound (Appendix Table 9). These diagnostics show that the network learns spatially varying geometry beyond uniform gradient rescaling.

The fitting norm still matters, as intrinsic diagonal training reaches 3.297 FID versus 3.544 for Euclidean-residual training, and rank-four intrinsic mobility gives 3.299 (Table 3b), so extra off-diagonal capacity does not improve this image setting. The matched additive field reaches 3.686 under its Euclidean objective. The synthetic loss comparison (Appendix Table 15) is complementary. Euclidean weighting recovers the prescribed metric more accurately, while intrinsic weighting gives comparable corrected-chain MMD and passage. Thus, field recovery and generation quality need not select the same restricted projection.

The solver budget places a further limit on the image gain. At 400 energy-gradient evaluations, REM is worse than identity by 0.064 FID. Budgets of 100 and 650 favor REM, while 200 is less consistent. The cutoff $\tau_g = 1$ was fixed by an independent 10k screen and not retuned per budget. Appendix

Table 3: **Field fit and objective controls.** (a) Euclidean MSE on 24,576 held-out OT pairs. ρ is cross-fit log-mobility correlation. (b) CIFAR-10 FID at 325 Euler–Heun intervals. Learned rows are means over three fits. Identity uses three streams.

(a) Held-out field fit

Dataset	Velocity MSE ↓		ρ
	Identity	REM	
CIFAR-10	0.12952	0.11882	0.895
ImageNet32	0.17432	0.15485	0.904

(b) Objective and mobility

Mobility		FID ↓
Class	Norm	
Identity	–	3.462
Diagonal	intrinsic	**3.297**
Diagonal	Euclidean	3.544
Rank-4	intrinsic	3.299

Tables 11 and 21 give the sweep and the post-hoc diagnostic. The main result is therefore conditional on the finite-time schedule, despite its consistency across paired runs at that schedule.

Finally, Appendix Table 22 keeps mobility active during stochastic refinement, holding the CIFAR energy, fitted mobility, sampling seed, and temperature schedule fixed. Including the explicit Heun correction lowers FID by 2.076 without numerical failure, though phase gating remains best. This one-fit, one-stream 10k screen supports retaining the correction when mobility is active but does not establish the stationary image distribution or variation across fits. Appendix D.3 records the settings and estimator.

5 Discussion

Exact geometry can cost more per step. In the ordinary-OT mixture, ESS rises from 660 to 1351, but 8,000 steps take 3.28 s (identity), 22.43 (diagonal), and 51.45 (corrected full) on the same processor. At image scale, phase gating avoids divergence computation by using learned geometry only during deterministic transport. These results establish finite-budget gains under matched steps, not wall-clock speedups. Invariance holds in continuous time, while experiments use finite unadjusted integrators. Field supervision identifies only $G\nabla V$, so we report field error, passage, sample discrepancy, and compute separately. Resolution and sampling dimension limit the current evidence. Image experiments are restricted to 32×32 transient generation. The 64-dimensional result establishes structured field-fitting capacity while its Gaussian sampling gap is small. Learned mobility inherits the target of the frozen energy rather than repairing a misspecified density.

These distinctions make REM a controlled diagnostic for energy-driven transport. Holding the energy fixed isolates the effect of geometry. Comparing one fitted model with and without the divergence drift isolates the correction. Step refinement then probes numerical error. These interventions separate explanations for a poor sampler without assuming that the learned density is correct. Positive diagonal mobility preserves the sign of each negative-gradient component, whereas off-diagonal terms permit coordinate coupling while maintaining net energy descent. The shear tests expose a restriction that additional diagonal capacity cannot remove under a fixed energy. Such representational freedom need not improve every target, as the image comparison shows.

6 Conclusion

Riemannian Energy Matching (REM) fits determinant-one SPD mobility to transport velocities while keeping the energy fixed. Our representability result characterizes the available directions, and the divergence-corrected diffusion preserves the energy's Boltzmann density in continuous time. On curved multimodal targets, full mobility recovers coordinate coupling that diagonal models miss, improving field fitting and accelerating mode passage. Removing the divergence drift from the same fitted mobility increases sampling error. In 64 dimensions, rank-four mobility reduces held-out velocity error by 70.3% relative to a parameter-matched diagonal model while retaining the frozen energy. On CIFAR-10 and ImageNet32, phase-gated mobility improves FID in all nine paired evaluations per dataset under the main matched solver budgets. Together, these experiments show how learned mobility improves transport and generation under a fixed density model.

AI USE STATEMENT

Generative AI tools assisted with mathematical formulations and proof drafts, experimental design, implementation and debugging of analysis code, interpretation of numerical results, and drafting and editing the manuscript. We checked the mathematical arguments and verified AI-assisted numerical summaries against recorded outputs. The authors take responsibility for the final methods, claims, citations, and text.

REPRODUCIBILITY STATEMENT

Code is available at https://anonymous.4open.science/r/REM-0056/. The repository includes mobility models, training and sampling routines, evaluation, tests, experiment configurations, and reproduction instructions. The appendix provides the derivations, experimental protocols, and per-run diagnostics. Synthetic targets are generated from specified seeds. Image reproduction requires the public datasets and pretrained Energy Matching weights listed in the code documentation. Datasets, model weights, and experimental outputs are not bundled with the code.

REFERENCES

Tara Akhound-Sadegh, Jarrid Rector-Brooks, Joey Bose, Sarthak Mittal, Pablo Lemos, Cheng-Hao Liu, Marcin Sendera, Siamak Ravanbakhsh, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, and Alexander Tong. Iterated denoising energy matching for sampling from boltzmann densities. In *International Conference on Machine Learning*, 2024.

Michal Balcerak, Tamaz Amiranashvili, Antonio Terpin, Suprosanna Shit, Lea Bogensperger, Sebastian Kaltenbach, Petros Koumoutsakos, and Bjoern Menze. Energy matching: Unifying flow matching and energy-based models for generative modeling. In *Advances in Neural Information Processing Systems*, 2025.

Lorenzo Baldassari, Josselin Garnier, Knut Solna, and Maarten V. de Hoop. Preconditioned langevin dynamics with score-based generative models for infinite-dimensional linear bayesian inverse problems. In *Advances in Neural Information Processing Systems*, 2025.

Daniel Barco, Michal Balcerak, Suprosanna Shit, Chinmay Prabhakar, Philipp Denzel, Bjoern Menze, and Frank-Peter Schilling. Projected energy matching for generative 3d priors. *arXiv preprint arXiv:2607.07749*, 2026.

Louis Bethune, David Vigouroux, Yilun Du, Rufin VanRullen, Thomas Serre, and Victor Boutin. Follow the energy, find the path: Riemannian metrics from energy-based models. In *Advances in Neural Information Processing Systems*, 2025.

Riddhiman Bhattacharya and Tiefeng Jiang. Fast sampling and inference via preconditioned langevin dynamics. *arXiv preprint arXiv:2310.07542v2*, 2024.

Maciej Borodzik, András Némethi, and Andrew Ranicki. Morse theory for manifolds with boundary. *Algebraic & Geometric Topology*, 16(2):971–1023, 2016.

Haozhuo Cao, Xiaorui Wang, Liyang Yu, and Wangcai Ding. GeomFlow: Geometry-aware adaptive diffusion model via hessian information. *Neurocomputing*, 676:132934, 2026. doi: 10.1016/j.neucom.2026.132934.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *International Conference on Learning Representations*, 2024.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of ImageNet as an alternative to the CIFAR datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *Advances in Neural Information Processing Systems*, 2022.

Alexander Falk, Laurenz Nagler, Andreas Habring, and Thomas Pock. Time-inhomogeneous preconditioned langevin dynamics. *arXiv preprint arXiv:2605.06091*, 2026.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B*, 73(2):123–214, 2011.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Kacper Kapuśniak, Peter Potaptchik, Teodora Reu, Leo Zhang, Alexander Tong, Michael Bronstein, Avishek Joey Bose, and Francesco Di Giovanni. Metric flow matching for smooth interpolations on the data manifold. In *Advances in Neural Information Processing Systems*, 2024.

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

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. In *Predicting Structured Data*. MIT Press, 2007.

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

Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, 2015.

Peter Pao-Huang, Xiaojie Qiu, and Stefano Ermon. Generative modeling with flux matching. *arXiv preprint arXiv:2605.07319*, 2026. URL https://arxiv.org/abs/2605.07319.

Sam Patterson and Yee Whye Teh. Stochastic gradient riemannian langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, 2013.

Christopher Scarvelis and Justin Solomon. Riemannian metric learning via optimal transport. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=v3y68gz-WEz.

Yang Song and Diederik P. Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

Xingzhi Sun, Danqi Liao, Kincaid MacDonald, Yanlei Zhang, Guillaume Huguet, Guy Wolf, Ian Adelstein, Tim G. J. Rudner, and Smita Krishnaswamy. Geometry-aware generative autoencoders for warped riemannian metric learning and generative modeling on data manifolds. In *International Conference on Artificial Intelligence and Statistics*, 2025.

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

Zitao Yang, Amin Ullah, Shuai Li, Fuxin Li, and Jun Li. Convex potential mirror langevin algorithm for efficient sampling of energy-based models. In *Advances in Neural Information Processing Systems*, 2025.

Benjamin J. Zhang, Youssef M. Marzouk, and Konstantinos Spiliopoulos. Geometry-informed irreversible perturbations for accelerated convergence of langevin dynamics. *Statistics and Computing*, 32(5):78, 2022. doi: 10.1007/s11222-022-10147-6.

Aaron Zweig, Mingxuan Zhang, Elham Azizi, and David Knowles. Energy guided geometric flow matching. *arXiv preprint arXiv:2509.25230*, 2025.

A Supporting Calculations

A.1 SPD feasibility under a fixed descent direction

Proof of Proposition 3.1. We prove necessity, construct a solution, and then establish its field regularity.

Necessity. Suppose $g \neq 0$ and $G \succ 0$. From $u = -Gg$,

$$u^\top g = (-Gg)^\top g = -g^\top Gg < 0. \tag{13}$$

If $g = 0$, invertibility gives $-Gg = 0$, so only $u = 0$ is representable. Conversely, when $g = u = 0$, any SPD matrix is admissible.

Construction. Assume $g \neq 0$ and $u^\top g < 0$. Define

$$v = -u, \quad r = \|g\|, \quad e = \frac{g}{r}, \quad a = e^\top v = -\frac{u^\top g}{r} > 0, \quad b = v - ae, \quad P = I - ee^\top. \tag{14}$$

Then $e^\top b = e^\top v - a = 0$. For an arbitrary $\lambda > 0$, set

$$G_\lambda = \frac{a}{r} ee^\top + \frac{eb^\top + be^\top}{r} + \frac{bb^\top}{ar} + \lambda P. \tag{15}$$

Using $e^\top e = 1$, $b^\top e = 0$, and $Pe = 0$ gives

$$G_\lambda g = rG_\lambda e = r\left(\frac{a}{r}e + \frac{b}{r}\right) = ae + b = v = -u. \tag{16}$$

Positive definiteness. Extend e to an orthonormal basis of $\mathbb{R}^d$. In that basis,

$$G_\lambda = \begin{bmatrix} a/r & b^\top/r \\ b/r & bb^\top/(ar) + \lambda I_{d-1} \end{bmatrix}. \tag{17}$$

The leading block satisfies $a/r > 0$. Its Schur complement is

$$S = \frac{bb^\top}{ar} + \lambda I_{d-1} - \frac{b}{r}\left(\frac{a}{r}\right)^{-1}\frac{b^\top}{r} \tag{18}$$

$$= \frac{bb^\top}{ar} + \lambda I_{d-1} - \frac{bb^\top}{ar} = \lambda I_{d-1} \succ 0. \tag{19}$$

The Schur-complement criterion therefore gives $G_\lambda \succ 0$.

Determinant gauge. Block determinant factorization and Eq. (19) give

$$\det G_\lambda = \det(a/r)\det(S) = \frac{a}{r}\lambda^{d-1}. \tag{20}$$

For $d \geq 2$, the choice $\lambda = (r/a)^{1/(d-1)}$ is positive and makes $\det G_\lambda = 1$. Substitution yields Eq. (5).

Smooth field and uniform ellipticity. Let K be compact and assume $g, u \in C^k(K; \mathbb{R}^d)$ with $-u^\top g \geq \delta > 0$. Set $M_g = \max_K \|g\|$ and $M_u = \max_K \|u\|$. Cauchy–Schwarz and the strict margin imply

$$\|g\| \geq \frac{\delta}{M_u}, \qquad \|u\| \geq \frac{\delta}{M_g}, \qquad a = -\frac{u^\top g}{\|g\|} \geq \frac{\delta}{M_g}. \tag{21}$$

Thus r, a, $1/r$, $1/a$, and $\lambda = (r/a)^{1/(d-1)}$ are bounded on K, and the coordinate-free expression in Eq. (5) is C^k. Continuity on the compact set gives $\lambda_{\max}(G_*) \leq M$ for some finite M. Since $1 = \det G_* = \prod_{i=1}^d \lambda_i(G_*)$, every eigenvalue satisfies $\lambda_i(G_*) \geq M^{-(d-1)}$. Hence $M^{-(d-1)}I \preceq G_* \preceq MI$, completing the proof. □

At joint zero sets, the pointwise solution $G = I$ need not join smoothly to the construction above. Additional compatibility conditions are required. The smooth field statement above therefore uses a uniform margin. Our implementation logs the descent-violation rate before fitting a mobility, and the nonrepresentable control in Appendix E verifies that REM does not evade this boundary.

Energy dissipation. Suppose $G(x) \succeq mI$ and $\dot{X} = -G(X)\nabla V(X)$. Then along every classical trajectory, $\frac{d}{ds}V(X_s) \leq -m\|\nabla V(X_s)\|^2$. Indeed,

$$\frac{d}{ds}V(X_s) = \nabla V(X_s)^\top \dot{X}_s = -\nabla V(X_s)^\top G(X_s)\nabla V(X_s) \leq -m\|\nabla V(X_s)\|^2, \tag{22}$$

Thus $dV(X_s)/ds = 0$ holds exactly at critical points of V.

A.2 Stationarity and reversibility of corrected dynamics

Proof of Proposition 3.2. Write $b_i = -\sum_j G_{ij}\partial_j V + \varepsilon \sum_j \partial_j G_{ij}$ and $a_{ij} = 2\varepsilon G_{ij}$. The Fokker–Planck equation of Eq. (8) is

$$\partial_s \rho = -\sum_i \partial_i(b_i\rho) + \frac{1}{2}\sum_{ij}\partial_i\partial_j(a_{ij}\rho) \tag{23}$$

$$= \sum_{ij}\partial_i(G_{ij}\rho\,\partial_j V) - \varepsilon\sum_{ij}\partial_i((\partial_j G_{ij})\rho) + \varepsilon\sum_{ij}\partial_i\partial_j(G_{ij}\rho) \tag{24}$$

$$= \sum_{ij}\partial_i(G_{ij}\rho\,\partial_j V) + \varepsilon\sum_{ij}\partial_i(G_{ij}\partial_j\rho) \tag{25}$$

$$= \nabla\cdot\left(G\left(\rho\nabla V + \varepsilon\nabla\rho\right)\right). \tag{26}$$

The two derivative-of-G terms cancel in the third line. For $\pi = Z^{-1}e^{-V/\varepsilon}$,

$$\nabla\pi = -\varepsilon^{-1}\pi\nabla V \quad\Longrightarrow\quad \pi\nabla V + \varepsilon\nabla\pi = 0, \tag{27}$$

so Eq. (26) gives $\partial_s\pi = 0$ pointwise.

It remains to prove reversibility. Expanding the divergence-form generator gives

$$\varepsilon\pi^{-1}\nabla\cdot(\pi G\nabla h) = \varepsilon\pi^{-1}(\nabla\pi)^\top G\nabla h + \varepsilon(\operatorname{div} G)^\top\nabla h + \varepsilon G : \nabla^2 h \tag{28}$$

$$= -\nabla V^\top G\nabla h + \varepsilon(\operatorname{div} G)^\top\nabla h + \varepsilon G : \nabla^2 h = \mathcal{L}_G h. \tag{29}$$

For compactly supported smooth f, h, integration by parts and the boundary assumption yield

$$\int f\mathcal{L}_G h\,d\pi = \varepsilon\int f\nabla\cdot(\pi G\nabla h)\,dx \tag{30}$$

$$= -\varepsilon\int\nabla f^\top G\nabla h\,d\pi = -\varepsilon\int\nabla h^\top G\nabla f\,d\pi = \int h\mathcal{L}_G f\,d\pi. \tag{31}$$

The middle equality uses $G = G^\top$, proving reversibility. Here π is a density relative to dx. A density relative to Riemannian volume would describe a different model. □

Proof of Corollary 3.3. Without the correction, $b_0 = -G\nabla V$. The probability current is

$$J_{0,i}[\rho] = b_{0,i}\rho - \varepsilon\sum_j\partial_j(G_{ij}\rho) \tag{32}$$

$$= -(G\nabla V)_i\rho - \varepsilon(\operatorname{div} G)_i\rho - \varepsilon(G\nabla\rho)_i. \tag{33}$$

At $\rho = \pi$, the first and third terms cancel because $\varepsilon\nabla\pi = -\pi\nabla V$, leaving

$$J_0[\pi] = -\varepsilon(\operatorname{div} G)\pi, \qquad \partial_s\pi = -\nabla\cdot J_0[\pi] = \varepsilon\nabla\cdot\left((\operatorname{div} G)\pi\right). \tag{34}$$

Therefore π is stationary exactly when $\nabla\cdot((\operatorname{div} G)\pi) = 0$. Detailed balance further requires zero current under the usual boundary conditions. Decreasing the numerical step size removes integration bias but cannot repair a violated stationary equation. □

A.3 Volume-preserving coordinate calculation

Write $J_{ia}(z) = \partial_{z_a}T_i(z)$ and let $K = J^{-1}$. Applying Itô's formula to $X = T(Z)$ gives

$$dX_i = \left(-\sum_a J_{ia}\partial_{z_a}U + \varepsilon\sum_a\partial_{z_a}J_{ia}\right)ds + \sqrt{2\varepsilon}\sum_a J_{ia}\,dW_a. \tag{35}$$

Since $V = U \circ T^{-1}$, $\nabla_z U = J^\top \nabla_x V$, so the first drift term is $-(JJ^\top \nabla_x V)_i$ and the diffusion covariance is $2\varepsilon JJ^\top = 2\varepsilon G$. Indeed, $Q(z) = G(T(z))^{-1/2} J(z)$ is orthogonal, so $\widetilde{W}_s = \int_0^s Q(Z_r)\, dW_r$ is a Brownian motion by Lévy's characterization, and $J\, dW = G^{1/2} d\widetilde{W}$. It remains to identify the Itô drift. Evaluating at $x = T(z)$ and using $\partial_{x_j} = \sum_b K_{bj} \partial_{z_b}$,

$$(\mathrm{div}_x\, G)_i = \sum_{j,b,a} K_{bj} \partial_{z_b} (J_{ia} J_{ja}) \tag{36}$$

$$= \sum_a \partial_{z_a} J_{ia} + \sum_a J_{ia} \sum_{j,b} K_{bj} \partial_{z_b} J_{ja}. \tag{37}$$

Mixed partials give $\partial_{z_b} J_{ja} = \partial_{z_a} J_{jb}$, and Jacobi's formula gives $\sum_{j,b} K_{bj} \partial_{z_a} J_{jb} = \partial_{z_a} \log\det J = 0$. The second term in Eq. (37) vanishes, while the first is $\Delta_z T_i$. Substitution into Eq. (35) gives the corrected drift in transformed coordinates. Finally, $\det J = 1$ implies $dx = dz$, so the pushforward of $e^{-U/\varepsilon} dz$ is $e^{-V/\varepsilon} dx$ after normalization. For a general diffeomorphism the pushforward Lebesgue density instead has energy $V_T = U \circ T^{-1} - \varepsilon \log|\det J_{T^{-1}}|$. The nonzero $\nabla \log|\det J|$ term also remains in Eq. (37), so a general coordinate change alters the fixed-energy representation.

A.4 What the controlled warp supervises

The controlled experiment makes geometric recovery identifiable. Let U be the latent mixture energy, $x = T_c(z)$, and $V_c(x) = U(T_c^{-1}(x))$. At every sampled training site x, the code evaluates

$$z = T_c^{-1}(x), \qquad G_c(x) = J_{T_c}(z) J_{T_c}(z)^\top, \qquad u_{\mathrm{ref}}(x) = -G_c(x) \nabla_x V_c(x). \tag{38}$$

For the shear, $z_1 = x_1$ and hence $G_c(x) = \begin{bmatrix} 1 & 2cx_1 \\ 2cx_1 & 1+4c^2x_1^2 \end{bmatrix}$. Half of each batch is drawn from the warped mixture. The other half is obtained by warping a prespecified latent channel with z_1 uniform over the modal range (plus a margin) and z_2 Gaussian around zero. The learned network is therefore tested on whether it can approximate Eq. (38) over this site distribution. This provides independent ground truth for field and mobility error. The image experiments complement it with the ordinary minibatch-OT supervision of Algorithm 1.

B Algorithms and Parameterizations

B.1 Mobility parameterizations

An unconstrained mobility can trade global scale against traversal time. For diagonal and full mobility we isolate anisotropy by imposing

$$\log\det G_\phi(x) = 0, \qquad e^{-2\alpha} I \preceq G_\phi(x) \preceq e^{2\alpha} I. \tag{39}$$

For images, a convolutional network outputs $h_\phi(x)$ and

$$\ell_i = \alpha \tanh h_i - \frac{1}{d} \sum_j \alpha \tanh h_j, \qquad G_\phi(x) = \mathrm{diag}(e^{\ell_1}, \ldots, e^{\ell_d}). \tag{40}$$

For full low-dimensional geometry, we exponentiate a trace-free, smoothly norm-bounded symmetric matrix. A zero-initialized output layer makes both parameterizations exactly identity at initialization.

The structured class uses a determinant-one diagonal matrix $D_\phi(x)$, a fixed orthonormal cosine basis $B \in \mathbb{R}^{d \times r}$, and network output $C_\phi(x) \in \mathbb{R}^{d \times r}$ to form

$$Q_\phi(x) = I + BC_\phi(x)^\top, \qquad G_\phi(x) = s_\phi(x) D_\phi(x)^{1/2} Q_\phi(x) Q_\phi(x)^\top D_\phi(x)^{1/2}, \tag{41}$$

where $s_\phi = |\det Q_\phi|^{-2/d}$ enforces $\det G_\phi = 1$. We bound each network output as $C_\phi = c \tanh(H_\phi)/\sqrt{dr}$ with $c = .75$, giving $\|C_\phi\|_2 < c < 1$ and thus an invertible Q_ϕ. This class has uniform spectral bounds. Application, inversion, noise transformation, and log determinant cost $O(dr^2 + r^3)$, including the small-matrix construction. Unlike an additive $D + UU^\top$ factorization, Eq. (41) has a nonzero off-diagonal derivative at its exact identity initialization.

B.2 FROZEN-ENERGY REM TRAINING

For Algorithm 1 and the image experiments, draw an OT-coupled pair (x_0, x_1) and $t \sim U[0,1]$, set $x_t = (1-t)x_0 + tx_1$ and $u_t = x_1 - x_0$, and update only ϕ using

$$\frac{1}{Z_u}\left(u_t + G_\phi(x_t)\nabla V_\theta(x_t)\right)^\top W_\phi(x_t)\left(u_t + G_\phi(x_t)\nabla V_\theta(x_t)\right) + \lambda_{\text{geo}}\mathcal{R}_\phi(x_t). \tag{42}$$

Here $W_\phi = I$ is the canonical Euclidean objective and $W_\phi = G_\phi^{-1}$ is the intrinsic image variant. The image and analytic-field fits use $Z_u = d$, matching the implementation's per-coordinate mean. Synthetic OT fits set Z_u to the current minibatch mean of $\|u_t\|_2^2$. For diagonal and full mobility, $\mathcal{R}_\phi = \|\log G_\phi\|_F^2/d$. For structured mobility, write $G = sD^{1/2}QQ^\top D^{1/2}$ as in Eq. (41). The implementation uses the inexpensive factor penalty $\mathcal{R}_{\text{str}} = d^{-1}\sum_i(\log D_{ii} + \log s)^2 + 2d^{-1}\|C\|_F^2$. This surrogate is not the exact squared matrix logarithm. The energy is in evaluation mode with `requires_grad=False`. Image training saves both raw and exponential moving average (EMA) mobility weights. Model selection in the synthetic experiments follows the protocols below. The output layer is zero-initialized, so training begins at the exact EM vector field. The intrinsic residual $r^\top G^{-1} r$ transforms as a scalar under coordinate changes, while the regularizer and neural class remain coordinate dependent.

Algorithm 1 Frozen-energy Riemannian Energy Matching

Require: pretrained V_θ; data $\mathcal{D}$; noise p_0; mobility G_ϕ; W_ϕ; λ_{geo}

1: Freeze θ; initialize the output of G_ϕ to zero so that $G_\phi \equiv I$
2: **for** $n = 1, \ldots, N_{\text{train}}$ **do**
3: Draw minibatches $\{x_{1,b}\} \sim \mathcal{D}$, $\{x_{0,b}\} \sim p_0$ and compute their OT coupling
4: Draw $t_b \sim U[0,1]$; set $x_{t_b} = (1-t_b)x_{0,b} + t_b x_{1,b}$ and $u_b = x_{1,b} - x_{0,b}$
5: Compute $\mathcal{L}_{\text{REM}}$ from Eq. (7); update ϕ only
6: **return** fitted mobility G_ϕ (EMA for images) and the unchanged energy V_θ

B.3 CORRECTED SAMPLING STEP

Algorithm 2 Phase-aligned REM generation (Euler–Heun)

Require: $x_0 \sim p_0$, frozen V_θ, mobility G_ϕ, step h, schedule $\varepsilon(s)$, gate τ_g

1: **for** $k = 0, \ldots, N-1$ **do**
2: $s_k \leftarrow kh$; $M_k \leftarrow G_\phi$ if $s_k < \tau_g$, otherwise I; $\epsilon_k \leftarrow \varepsilon(s_k)$
3: $a_k(x) \leftarrow -M_k(x)\nabla V_\theta(x) + \frac{1}{2}\epsilon_k \operatorname{div} M_k(x)$
4: Draw $\xi_k \sim \mathcal{N}(0, I)$; $q_k(x) \leftarrow \sqrt{2h\epsilon_k}\, M_k(x)^{1/2}\xi_k$
5: $\widetilde{x} \leftarrow x_k + ha_k(x_k) + q_k(x_k)$ ▷ predictor
6: Form $M_{k+1}, \epsilon_{k+1}, a_{k+1}, q_{k+1}$ at s_{k+1}, reusing ξ_k
7: $x_{k+1} \leftarrow x_k + \frac{h}{2}\left(a_k(x_k) + a_{k+1}(\widetilde{x})\right) + \frac{1}{2}\left(q_k(x_k) + q_{k+1}(\widetilde{x})\right)$
8: **return** x_N

For Euler–Maruyama with step h, temperature ε_k, and Gaussian ξ_k,

$$x_{k+1} = x_k + h\left(-G_k\nabla V(x_k) + \varepsilon_k \operatorname{div} G_k\right) + \sqrt{2h\varepsilon_k}\, G_k^{1/2}\xi_k. \tag{43}$$

Algorithm 2 assumes diagonal mobility whenever the temperature is positive. This covers the ungated diagonal screen and all phase-gated runs, where $G = I$ during stochastic refinement. For diagonal diffusion, the equivalent Stratonovich drift contains one half of the Itô divergence term. A general non-diagonal noise factor requires its own Stratonovich conversion. Predictor and corrector reuse the same Brownian increment. When $G = I$, both updates reduce exactly to the original EM sampler.

Disabling the explicit Heun correction sets the Stratonovich drift to $-G\nabla V$. For diagonal noise, its equivalent Itô drift remains $-G\nabla V + \frac{1}{2}\varepsilon \operatorname{div} G$. This solver-term ablation differs from removing the entire Itô correction, which the synthetic experiments implement with Euler–Maruyama.

B.4 Efficient divergence estimation

For Rademacher ξ, the vector Jacobian–vector identity

$$\mathbb{E}_\xi \left[J_x(G(x)\xi)\, \xi \right]_i = \sum_j \partial_j G_{ij}(x) = (\operatorname{div} G)_i \tag{44}$$

provides an unbiased estimator. Low-dimensional experiments use exact Jacobians. Phase-gated image runs require no probes because $G \neq I$ only when $\varepsilon = 0$.

C Experimental Details

Table 4: Geometry stress-test hyperparameters.

Quantity	Value	Quantity	Value
Curvatures	$0, .4, .8, 1.2$	seeds	5 independent
Mixture variance	.1	chains	64
Mobility train steps	3,000	batch size	512
Hidden width/depth	128 / 3	learning rate	10^{-3}
Log bound	5	geometry weight	10^{-3}
Main step size	.002	validity steps	$.001, .002, .004, .008$
Burn/draw time	4 / 12	save interval	.01

In the 2D shear tests, passage starts at the first retained post-burn-in state. Its nearest latent mean defines the initial mode. Subsequent labels change only inside radius-$.65$ modal cores. Reported passage averages per-seed medians conditioned on arrival, with completion rates reported separately. MMD uses 2,048 randomly subsampled retained states and equally many target draws in latent coordinates. Higher-dimensional checks use the first 1,024 flattened states and target draws in observed coordinates. The squared Gaussian-kernel bandwidth is the median positive squared pairwise distance in the pooled comparison sample. ESS uses mode indicators in 2D and individual coordinates in higher dimensions.

For the inductive-bias comparison at $c = 1.2$, all learned classes use the same five seeds, 3,000 optimization steps, batch size 512, learning rate 10^{-3}, training-point generator, and held-out velocity set. Full REM has 33,795 parameters. The unrestricted and PEM Helmholtz-residual multilayer perceptrons each have 33,666. In 2D we compute the normalized divergence penalty exactly, removing Hutchinson variance while retaining PEM's Phase-2 constraint. Its velocity MSE is divided by target-velocity energy. Writing the learned residual as r_ψ and dimension as $D = 2$, its fitted objective is

$$\frac{\mathbb{E}\|-\nabla V + r_\psi - u\|_2^2}{\mathbb{E}\|u\|_2^2} + \lambda_{\mathrm{div}} \mathbb{E}\left[\left(\frac{\operatorname{tr} J_{r_\psi}}{D} \right)^2 \right] + 10^{-3} \frac{\mathbb{E}\|r_\psi\|_2^2}{\mathbb{E}\|u\|_2^2}, \qquad \lambda_{\mathrm{div}} = 1. \tag{45}$$

The divergence penalty therefore applies to the residual field, not the total drift. Dividing the velocity terms by target-velocity energy and the divergence by D makes the unit weight a dimensionless controlled calibration. The mirror potential is a strongly convex sum of 8,192 softplus ridge functions. Its 32,769 parameters make it capacity matched to the two neural controls while retaining the inverse-Hessian constraint. The learned constant-SPD preconditioner has only its three intrinsic packed-matrix parameters. Each stochastic evaluation uses 64 chains with the same burn, draw, and step-size settings as the geometry stress test. The residual row isolates PEM's frozen-energy additive-field component. Full PEM also trains a teacher, performs Helmholtz distillation and negative caching, and refines the energy. The mirror row isolates inverse-Hessian mobility, whereas full CPMLA jointly trains the EBM and convex potential flow. The present controls keep the energy fixed to compare representation classes.

C.1 High-dimensional OT-chord setup

The OT-supervised structured experiment uses $d = 64$, rank $r = 4$, five independent seeds, 1,200 training minibatches, batch size 128, and 32 held-out minibatches per seed. Let B contain the first

four orthonormal discrete-cosine vectors and $P = BB^\top$. The target is the centered Gaussian with transport square root

$$A = .72I + (.18 - .72)P, \qquad \Sigma = AA^\top, \tag{46}$$

and frozen exact energy $V(x) = \frac{1}{2}x^\top\Sigma^{-1}x$. Each training batch independently draws $x_0 \sim \mathcal{N}(0, I)$ and $x_1 \sim \mathcal{N}(0, \Sigma)$, computes a minibatch optimal transport coupling, samples $t \sim U[0, 1]$, and returns the ordinary observed-space chord $(x_t, u_t) = ((1-t)x_0+tx_1, x_1-x_0)$. No analytic mobility or population transport map is provided to any learner.

All learned variants minimize the canonical Euclidean objective in Eq. (7) with learning rate 2×10^{-3} and geometry weight 10^{-4}. Both the structured rank-four model and the capacity-matched diagonal control have 66.1k trainable parameters. The ordinary diagonal model has 33.1k. The energy-continuation control is a full-SPD quadratic with 4,160 trainable parameters, initialized at the exact target precision and updated against the same OT velocities. Its learned equilibrium is available in closed form, so Table 5 reports $D_{\mathrm{KL}}(\mathcal{N}(0, \Sigma) \,\|\, p_{V_{\mathrm{continued}}})$ exactly rather than through finite-chain sampling. This Gaussian setting isolates representation error and density change.

Table 5: **Minibatch-OT supervision in 64 dimensions** (means over five seeds). The target energy is exact before fitting. Structured REM and the matched diagonal control have equal capacity and keep it fixed. The energy-fit row starts from the same exact potential and receives the same velocities. The Fixed column refers to the represented density. Velocity MSE is relative. KL and covariance error are analytic.

Method	Fixed?	Params	vel. MSE ↓	KL ↓	cov. err. ↓	off-diag.
Identity	yes	0	19.733	0	0	0
Diagonal REM	yes	33.1k	11.237	0	0	0
Diagonal REM (matched)	yes	66.1k	11.055	0	0	0
Structured REM (rank 4)	yes	66.1k	**3.286**	0	0	0.113
Continue energy fit	no	4.2k	0.924	18.301	3.667	–

C.2 OT SUPERVISION ON CURVED TARGETS AND CORRECTED SAMPLING

The curved 16D target is an equally weighted four-mode Gaussian mixture in the first two latent coordinates, with means $(\pm 2, 0), (0, \pm 2)$ and component variance .25. We apply the unit-Jacobian shear $z_2 \mapsto z_2 + .2z_1^2$, append 14 Gaussian coordinates of standard deviation .7, and rotate all 16 coordinates by a fixed random orthogonal matrix. This construction provides exact sampling and energy evaluation only. Training uses 2,000 independent minibatch-OT chord batches of size 256, and evaluation uses 32 fresh chord batches. The capacity-matched diagonal and rank-four structured models use the canonical Euclidean objective, learning rate 2×10^{-3}, geometry weight 10^{-4}, and five independent seeds. No analytic velocity, latent coordinate, or reference mobility is supplied.

For a same-target bridge to the controlled 2D dynamics, we also train on the exactly sampleable four-mode shear of Appendix Figure 9, now using independent standard-Gaussian base and target draws with exact squared-Euclidean minibatch-OT assignments instead of the analytic field. The frozen energy, component variance .1, and shear $c \in \{0, 1.2\}$ are unchanged. Full and diagonal determinant-one mobilities use width 128, depth three, 1,500 independent training batches of 256 pairs, learning rate .002, log-eigenvalue bound 3, and minimal-distortion weight 10^{-4}. They have 33,795 and 33,666 trainable parameters, respectively. We select each fitted model using eight fresh OT batches and evaluate velocity error on 16 disjoint batches. The same five training seeds and chord batches are used for full and diagonal models at $c = 1.2$. The $c = 0$ field control has three seeds. The exact-divergence Langevin comparison uses 64 matched chains per seed, $\Delta t = .002$, physical burn-in 4 and draw duration 12, with the same modal-core first-passage definition as the analytic-field experiment. All variants use 8,000 integration steps. This small extension ran on a four-thread Intel Xeon Gold 6530 central processing unit (CPU) with PyTorch 2.4.

Full REM improves first passage and latent MMD over both controls in all five paired seeds, while its mode-indicator ESS rises from 660 (identity) to 1351. The fitted field remains a projection. In the held-out set, 21.7% of chords violate pointwise descent, and full REM has relative residual 17.8. With no shear ($c = 0$), full versus diagonal OT residual is 1.57 versus 22.67. Ordinary OT projection can favor off-diagonal mobility even without coordinate curvature. The curvature-dependent claim in Figure 9 therefore pertains to its analytic-field diagnostic.

Table 6: OT-supervised 2D shear at $c = 1.2$ (means over five paired seeds). The median first-passage steps are computed per seed before aggregation. All sampling rows use the same 8,000 steps. CPU seconds show the cost of exact divergence. The no-divergence row reuses the fitted full mobilities and therefore has the same OT MSE.

Variant	Relative OT MSE	First passage	Latent MMD	CPU seconds
Identity	1068	215	5.41×10^{-3}	3.28
Diagonal REM	413	202	6.75×10^{-3}	22.43
Full REM	17.8	75	7.5×10^{-4}	51.45
Full REM, no div G	same	215	5.088×10^{-2}	12.53

For the positive-temperature extension, we fit diagonal and rank-four structured mobilities to the correlated-Gaussian chord setup in dimensions 16 and 64, with 1,200 training batches of size 128 and five seeds. Each fitted mobility is then used at temperature one in 64 parallel chains, with 1,000 burn-in steps followed by 400 draws thinned every five steps. We test $\Delta t \in \{.002, .005\}$. Dimension 16 computes $\operatorname{div} G$ exactly. Dimension 64 uses eight Hutchinson probes. The no-divergence row reuses the identical fitted structured mobility, so its stationary discrepancy isolates the correction rather than retraining. The 16D discrepancy is large. In 64D, the absolute MMD gap is small, but all five paired seeds favor correction in MMD and covariance error. The cost of exact or eight-probe divergence at image resolution remains unmeasured.

Image datasets are CIFAR-10 (Krizhevsky, 2009) and the downsampled ImageNet benchmark (Chrabaszcz et al., 2017). The image ablations keep the frozen CIFAR energy, 150k mobility-training budget, determinant gauge, spectral bound, phase cutoff, 325-interval sampler, and three-by-three nested evaluation fixed. One replaces intrinsic weighting by Euclidean residual training. Another uses rank-four congruence mobility under intrinsic weighting. The solver sweep evaluates three paired fit/stream combinations at 50, 100, and 200 intervals (100, 200, and 400 energy-gradient evaluations) and uses the corresponding one-stream-per-fit subset of the 325-interval grid for a matched 650-evaluation row.

Table 7: Image settings and model training steps. Mobility training seeds index independent fits. Sampling seeds determine the FID uncertainty.

	CIFAR-10	ImageNet32
Energy training step	147,000	641,000
Mobility training step	150,000	150,000
Mobility training seeds	3	3
Sampling seeds	3	3
Generated images / seed	50,000	50,000
Euler–Heun intervals	325	250
Terminal time / Δt	3.25 / .01	2.50 / .01
Temperature maximum	.01	.01
Geometry cutoff	1.0	1.0
Real statistic	CIFAR standard	1,281,167 train images
Batch size / device	64 / H100	64 / H100

All final image runs used Python 3.12.3, PyTorch `2.7.0a0+ecf3bae40a.nv25.02`, CUDA 12.8, and one NVIDIA H100 graphics processing unit (GPU) with 80 GB of memory. The code provides protocol settings and run commands. Appendix D reports per-fit summaries. The ImageNet statistic uses floor quantization and is cached before paired evaluation. For identity mobility, $G_s = I$ for every cutoff. Optimizing that cutoff therefore leaves the matched EM schedule unchanged.

C.3 Representability on the supervision distribution

For image pairs drawn from the same augmentation and minibatch-OT distribution as mobility training, we measure the normalized margin $m = -u_t^\top \nabla V(x_t)/(\|u_t\|_2 \|\nabla V(x_t)\|_2)$. Positive m is exactly the nonzero pointwise SPD-representability condition. For the synthetic row, we apply the same formula to held-out sites and u_{ref} from Eq. (38), not to OT pairs. The margin measures pointwise representability on the supervision distribution. Table 8 reports no violations among 69,632 held-out examples. The lower tail remains positive, although it approaches zero for a very small

fraction of image pairs, consistent with treating Eq. (4) as a local existence statement rather than a uniform-margin guarantee for the neural model.

Table 8: Normalized descent margin on held-out supervision. Image quantiles pool three independent 8,192-pair draws. Synthetic values at the strongest warp are means over quantiles across five independent 4,096-pair draws.

Setting	pairs	violation rate	1st pct.	5th pct.	median
Synthetic, $c = 1.2$	20,480	0	0.073	0.094	0.451
CIFAR-10	24,576	0	.608	.838	.956
ImageNet32	24,576	0	.523	.788	.938

C.4 Image mobility learned from OT supervision

What intrinsic image weighting selects. Writing $g = \nabla V$ gives $r^\top G^{-1} r = u^\top G^{-1} u + 2u^\top g + g^\top G g$. The cross term is independent of G. The other terms compare the target tangent vector and energy covector in dual norms. Under $r_x = J r_z$ and $G_x = J G_z J^\top$, the residual is the coordinate scalar $r_x^\top G_x^{-1} r_x = r_z^\top G_z^{-1} r_z$. Any exact representer zeros both metric-weighted and Euclidean residuals, but a bounded diagonal network generally cannot fit exactly, so the two losses define different projections. The descent condition in Eq. (4) does not select between them. Table 15 confirms this distinction empirically.

Table 9: Held-out image OT diagnostics. Relative MSE is Euclidean and therefore does not use the learned metric to score its own residual. REM variation is across three mobility fits. Eigenvalue percentiles sample 64 fixed coordinates per image. κ is the per-image condition number. ρ is mean pairwise per-image correlation of log-mobility maps across fits. The near-bound column gives the fraction of coordinates with $|\log G_i| \geq .95\alpha$.

Dataset	Identity rel. MSE	REM rel. MSE	reduction	improved pairs	eig. 5–95%	med. κ	map ρ	near bound
CIFAR-10	0.12952	0.11882	8.3%	77.9%	0.831–1.200	1.89	0.895	0.08%
ImageNet32	0.17432	0.15485	11.2%	82.7%	0.762–1.307	2.39	0.904	0.20%

Table 10: Association with OT phase. The diagonal network receives no explicit time input. Activation is mean absolute log eigenvalue. Residual reduction is relative to the frozen-energy Euclidean field within each held-out time bin. Values average three fits.

Statistic	$[0, .2)$	$[.2, .4)$	$[.4, .6)$	$[.6, .8)$	$[.8, 1]$
CIFAR-10, $\mathbb{E}\lvert \log G \rvert$	0.039	0.051	0.061	0.092	0.180
CIFAR-10, residual reduction	0.7%	2.2%	3.8%	8.0%	13.9%
ImageNet32, $\mathbb{E}\lvert \log G \rvert$	0.061	0.078	0.093	0.136	0.252
ImageNet32, residual reduction	2.1%	3.1%	4.5%	10.7%	18.4%

Table 9 evaluates the three final mobility fits on 24,576 held-out OT pairs per dataset. Unlike the training objective, the reported residual is ordinary Euclidean relative MSE, so G cannot reduce the score merely by changing the norm. The learned diagonal mobility reduces this external residual by 8.3% on CIFAR-10 and 11.2% on ImageNet32, with nearly identical values across fits and improvement on 77.9% and 82.7% of pairs. Log-mobility maps are also strongly correlated across independently trained fits ($\rho = .895$ and $.904$). Thus the OT-supervised networks learn reproducible, nonidentity structure rather than only fit-specific scale noise.

The determinant-one gauge keeps the median coordinate eigenvalue near one, while the 5th–95th percentile ranges show anisotropy. Fewer than $.21\%$ of coordinates approach the imposed log bound. Table 10 shows that activation increases toward the data end of the OT path, where the Euclidean residual and REM's reduction are largest. The network has no explicit time input. In the matched retraining of Table 3, intrinsic training retains the image FID gain, while Euclidean training does not.

D ADDITIONAL QUANTITATIVE RESULTS

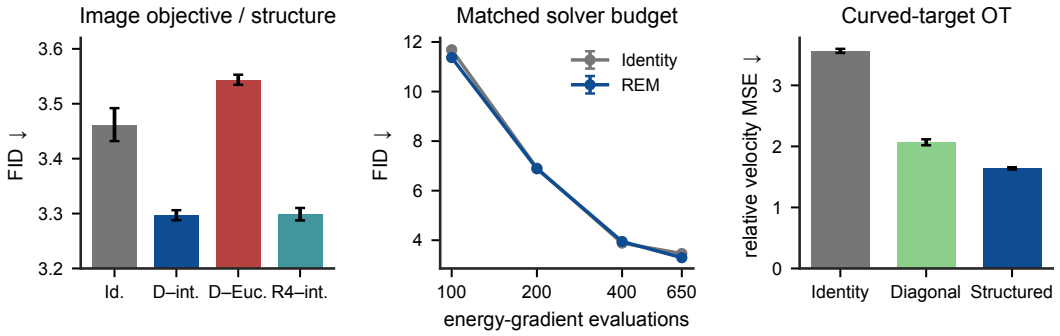


Figure 8: **Additional controls.** Intrinsic image training selects the FID gain (left), paired FID varies with the energy-gradient budget (center), and standard OT chords on a curved 16D target favor structured off-diagonal lifting over a capacity-matched diagonal class (right). Image-objective abbreviations: Id. identity, D–int. intrinsic diagonal, D–Euc. Euclidean diagonal, R4–int. rank-four intrinsic mobility.

Table 11: CIFAR-10 50k solver-budget sweep. Each row uses three paired runs. ΔFID is REM minus identity, so negative values favor REM. The 650-evaluation row selects one matched stream from each of the three mobility fits to keep the replication unit aligned.

Energy-gradient evals.	Identity FID	REM FID	paired ΔFID	wins
100	11.686	11.372	-0.314	3/3
200	6.909	6.883	-0.025	2/3
400	3.884	3.948	0.064	0/3
650	3.462	3.296	-0.166	3/3

Table 12: OT-supervised curved multimodal target in 16D. All learned rows receive only the frozen energy gradient and ordinary minibatch-OT straight chords. Held-out Euclidean metrics are means over five seeds.

Method	rel. velocity MSE ↓	cosine ↑	off-diag. fraction
Identity	3.565	0.355	0.000
Capacity-matched diagonal REM	2.067	0.333	0.000
Rank-4 structured REM	1.640	0.342	0.185

Table 13: Corrected positive-temperature sampling with OT-trained mobilities at $\Delta t = .005$. The 16D correction is exact. The 64D correction uses eight Hutchinson probes. Five seeds.

d	Mobility	mean ESS ↑	MMD^2 ↓	rel. cov. error ↓
16	Identity	855.5	8.7×10^{-3}	0.362
16	Diagonal REM	854.4	8.9×10^{-3}	0.367
16	Structured REM	939.8	7.81×10^{-3}	0.346
16	Structured, no div.	676.3	2.64×10^{-2}	0.991
64	Identity	839.2	1.42×10^{-2}	0.771
64	Diagonal REM	846.5	1.36×10^{-2}	0.769
64	Structured REM	868.2	1.35×10^{-2}	0.755
64	Structured, no div.	856.8	1.41×10^{-2}	0.779

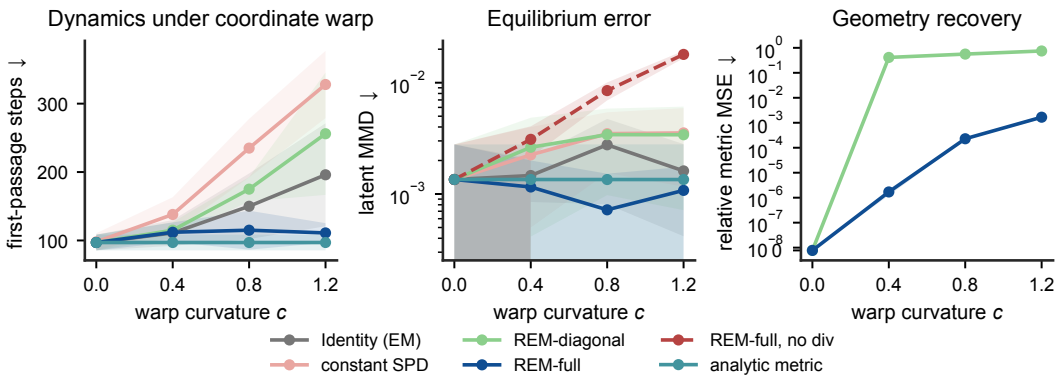


Figure 9: **Analytic-field geometry diagnostic (means over five seeds).** An area-preserving warp bends a fixed four-mode equilibrium. Under supervision by its analytic pushforward field, full REM tracks the analytic reference as curvature rises. Identity, constant, and diagonal mobility slow down. The no-divergence sampler has higher latent MMD.

Table 14: Strongest warp ($c = 1.2$). Occupancy is maximum absolute error from target mass 0.25. Lower is better except transitions.

Method	latent MMD $\times 10^3$	occupancy	trans./1k	first passage
Identity (EM)	1.61	3.2×10^{-2}	10.09	196
Constant SPD	3.54	4.8×10^{-2}	6.10	328
REM-diagonal	3.41	4.5×10^{-2}	8.18	256
REM-full	**1.08**	$\mathbf{2.2\times10^{-2}}$	11.25	**111**
REM-NoDiv	17.96	2.6×10^{-2}	10.38	117
Analytic metric	1.35	3.1×10^{-2}	**11.49**	97

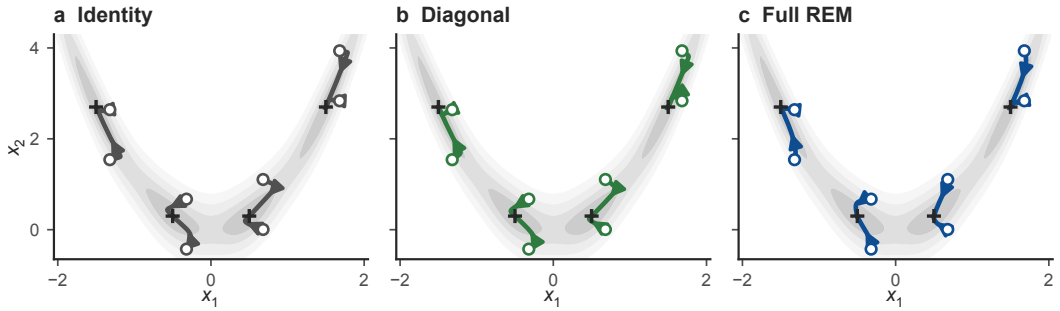


Figure 10: **One energy, different transport paths.** Deterministic $-G\nabla V$ paths from eight shared starts (open circles), integrated with the identity baseline and trained diagonal/full mobility models under OT supervision. Gray shading is the fixed target density. Crosses mark component centers. Figure 4 evaluates the stochastic samplers.

D.1 Synthetic loss-weighting ablation

Table 15: Loss-weighting ablation at the strongest warp (means over five independent fits). All optimization, architecture, held-out, and chain settings are identical.

Training residual	rel. velocity MSE	rel. metric MSE	latent MMD^2	first passage
Euclidean (canonical)	1.08×10^{-3}	1.67×10^{-3}	1.08×10^{-3}	111
Intrinsic ($W_\phi = G_\phi^{-1}$)	0.393	0.339	7.5×10^{-4}	117

The canonical Euclidean objective is markedly sharper for recovery of the known field and metric, while intrinsic weighting reaches comparable corrected-chain MMD and first passage. The two norms therefore provide complementary projections within Eq. (7). The controlled experiment emphasizes exact recovery, and the image experiment uses intrinsic weighting while reporting external Euclidean residuals in Table 9.

D.2 IMAGE PER-RUN VALUES AND CUTOFF SCREEN

Table 16: Nested CIFAR-10 replication. Each mobility-training seed is evaluated on the same three sampling seeds. The final row averages the three within-fit means.

Mobility fit	Seeds / fit	REM FID	paired ΔFID	paired ΔKID $\times 10^4$
Fit A	3	3.3000	−0.1620	−1.74
Fit B	3	3.2867	−0.1752	−1.84
Fit C	3	3.3042	−0.1578	−1.76
Across fits	nested	3.2970	−0.1650	−1.78

Table 17: Nested ImageNet32 50k replication. Each mobility fit is evaluated on the same three sampling streams. The final row averages the three within-fit means.

Mobility fit	Streams / fit	REM FID	paired ΔFID
Fit A	3	6.5626	−0.1797
Fit B	3	6.4875	−0.2548
Fit C	3	6.6140	−0.1282
Across fits	nested	6.5547	−0.1876

Table 18: CIFAR-10 learned constant-SPD control under the shared sampling streams.

Run	Identity FID	Constant-SPD FID	paired ΔFID
Run A	3.4602	3.4133	-0.0470
Run B	3.4928	3.4525	-0.0403
Run C	3.4328	3.3941	-0.0387

Table 19: Matched CIFAR-10 additive-transport control. A 57,635-parameter residual network receives the same frozen energy, OT chords, training budget, phase cutoff, solver, and three 50k sampling streams per fit as REM. Values in the first three rows are within-fit stream means. The last row averages independent fit means. Positive deltas favor the corresponding reference method.

Residual fit	Streams / fit	FID	Δ vs. identity	Δ vs. REM
Fit A	3	3.677	+0.215	+0.377
Fit B	3	3.691	+0.229	+0.404
Fit C	3	3.690	+0.228	+0.386
Across fits	nested	3.686	+0.224	+0.389

Table 20: Three-seed 10k cutoff screen. This screen selected the final cutoff before 50k tests.

τ_g	.75	.90	1.00	1.10	1.25
FID	7.326	7.291	**7.279**	7.282	7.297
KID $\times 10^4$	7.486	7.210	7.158	7.129	7.130

Table 21: Post-hoc CIFAR-10 cutoff diagnostic at 400 energy-gradient evaluations. Each row uses the same three paired mobility-fit/sampling-stream seeds and 50k samples per seed. Negative ΔFID favors REM. The final cutoff of 1.0 came from the independent 10k screen above.

Mobility cutoff τ_g	FID	Paired ΔFID vs. identity	Wins
.50	3.876	-8×10^{-3}	3/3
.75	3.960	$+7.5\times10^{-2}$	0/3
.90	3.993	$+1.09\times10^{-1}$	0/3
1.00	3.948	$+6.4\times10^{-2}$	0/3

D.3 Ungated image correction screen

We complete a single-seed CIFAR-10 comparison by disabling the explicit Heun correction while keeping the learned diagonal mobility active throughout sampling. All three conditions use the same frozen energy, fitted mobility, sampling seed, batch size 64, and 325 Euler–Heun intervals with $\Delta t = .01$ and terminal time 3.25. The temperature schedule is unchanged, with a maximum of .01. The corrected ungated sampler uses one Hutchinson probe per drift evaluation and the diagonal Stratonovich correction described in Appendix B. The off row retains the noise-induced half correction in Itô form, as derived in Appendix B. The gated and corrected rows reuse earlier runs. The off row completes this comparison. All three runs use the same H100 GPU model and PyTorch/CUDA versions.

Table 22: CIFAR-10 correction screen with 10k generated images per condition, one mobility fit, and one sampling seed. All rows share the frozen energy and fitted mobility. The gated row uses identity during stochastic refinement. Off removes the explicit Stratonovich correction. The equivalent Itô drift retains $\frac{1}{2}\varepsilon \operatorname{div} G$. These are individual runs, not averages across seeds.

Sampling	Explicit Heun correction	FID ↓
Phase-gated mobility	Not needed	7.259
Ungated mobility	On	9.018
Ungated mobility	Off	11.094

Both ungated samplers finish without numerical failure. Correction lowers FID by 2.076, but the gated sampler remains better. The comparison concerns finite-time generation with one mobility fit and one sampling seed. It does not measure the stationary distribution or establish that the difference persists across seeds. Its 10k-sample FID values are separate from the 50k-sample results in Table 2.

D.4 Validity step-size grid

Table 23: Five-seed equilibrium validity at curvature 1.2.

Metric	Variant	.001	.002	.004	.008
latent MMD	corrected	1.00×10^{-3}	1.35×10^{-3}	8.5×10^{-4}	1.85×10^{-3}
	no divergence	1.746×10^{-2}	1.879×10^{-2}	1.887×10^{-2}	1.834×10^{-2}
moment error	corrected	.044	.021	.037	.075
	no divergence	.252	.370	.379	.362

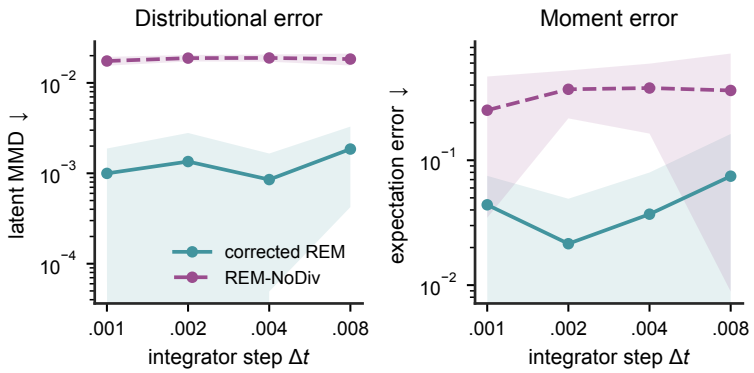


Figure 11: **Equilibrium validity at $c = 1.2$.** Corrected dynamics with analytic mobility stay near the target across step sizes. The no-divergence bias persists as $\Delta t \to 0$, so it is not integrator error.

E Representability Negative Control

We also fit full REM to a constructed field that violates $u^\top \nabla V < 0$. The SPD constraint cannot repair this sign contradiction. Relative velocity error rises from 2.549 for Euclidean Energy Matching to 3.994 for full REM, consistent with the descent boundary in Eq. (4).

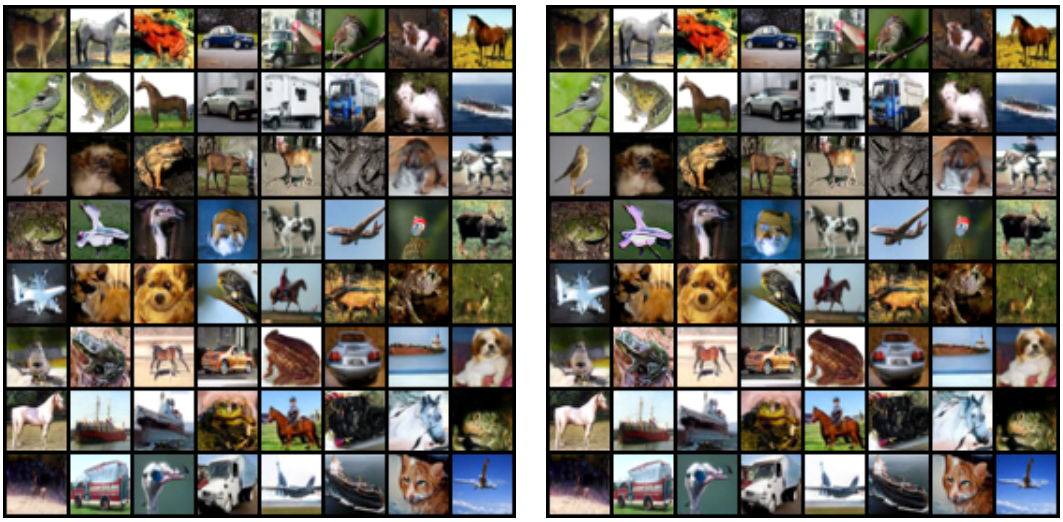
F Compute, Scope, and Broader Impact

The geometry stress suite and final image evaluations used single H100 80GB jobs. CIFAR 50k sampling takes 4.46 hours for identity and 4.55 hours for phase-aligned REM on the measured system. ImageNet32 takes 3.42 and 3.52 hours, respectively. These wall times include sample generation and metric evaluation as recorded by the evaluation code. At 650 energy-gradient evaluations, the 0.058M-parameter image mobility adds 2.0% time and 2.9% memory on H100.

Both image benchmarks have 32×32 resolution, so they test transfer across datasets but not to higher-resolution image geometry. The 2D OT-trained corrected sampler improves mixing at matched steps while requiring substantially more CPU time for exact state-dependent divergence. Practical corrected stochastic speedup at image scale remains untested. External learned samplers and parallel tempering have not been compared at matched frozen energy and number of energy-gradient evaluations, so their relative performance remains open.

REM inherits the limitations of its fixed energy, including biases or unsafe correlations in the training data. Geometry does not correct such misspecification.

G Additional Qualitative and Diagnostic Figures



(a) Energy Matching (b) REM

Figure 12: **Full CIFAR-10 sample grids.** These are the complete fixed 64-image prefixes under matched initial noise. No sample was selected or removed after generation.



(a) Energy Matching (b) REM

Figure 13: **Full ImageNet32 sample grids.** Complete fixed 64-image prefixes under matched initial noise. The grids permit a visual comparison of sample variety, while Table 2 quantifies the distributional difference.