Geometry-Aware Energy-Based Image Modelling

Zoe Wu

[1]Harvard University

Motivation

Despite their ability to produce high-fidelity images, modern text-to-image diffusion models often exhibit limited sample diversity under fixed prompts.



Figure 1. Generating "a cow on a blue pasture with red skies" using Stable Diffusion XL

A creative exploring design choices may want the trained image generative model to produce more diverse samples, even if they may be somewhat less likely in the data.

In this work, we explore how we can create a geometry-aware image sampling process to improve the tradeoff between diversity and fidelity.

Existing Works

Energy-Based Models estimate the probability distribution using a learned energy function $E_\theta(x)$. [2] It then uses Langevin dynamics during sampling to update

$$\tilde{x}^k = \tilde{x}^{k-1} - \frac{\lambda}{2}\nabla_x E_\theta\left(\tilde{x}^{k-1}\right) + \omega^k, \qquad \omega^k \sim \mathcal{N}(0, \lambda)$$

Can we make our scalar step size λ directionally dependent to better exploit/explore?

Stein Score Metric Tensor provides a smooth, data-dependent Riemannian structure over the energy landscape [1]

$$\mathbf{G}_\tau(\mathbf{x}) = \tau\,\mathbf{s}(\mathbf{x})\mathbf{s}(\mathbf{x})^\top + \mathbf{I}$$

where $\tau > 0$ is chosen to normalize the outer product and tune the weight of Riemannian geometry. We update:

$$\tilde{x}^k = \tilde{x}^{k-1} - \mathbf{G}_\tau^{-1}\frac{\lambda}{2}\nabla_x E_\theta\left(\tilde{x}^{k-1}\right) + \omega^k, \qquad \omega^k \sim \mathcal{N}(0, \mathbf{G}_\tau^{-1}\lambda)$$

Intuition for Why We Use Geometry

Say you are given the below probability distribution without any knowledge of local geometry. You want to evaluate a $\nabla_x E(x)$ that shifts your distribution

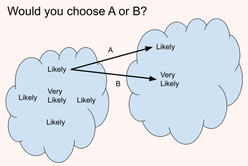


Figure 2. Without learned geometry

Say you are given both the probability distribution and learn that we are just looking at a stress ball being flattened. With more knowledge about geometry,

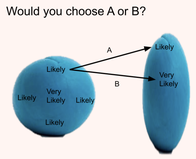


Figure 3. With learned geometry

Estimating the anisotropy of the parameter space with a data-based metric tensor encourages better traversal of the space and hence better fidelity-diversity tradeoffs [3].

2D MVN samples reflect local anisotropic geometry

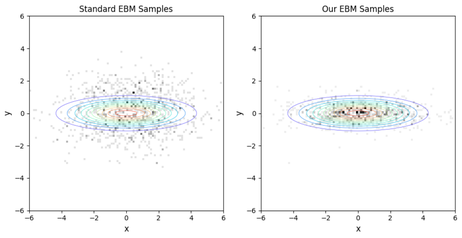


Figure 4. We ran standard EBM and Riemannian EBM with the same setup. We normalized our Riemannian metric tensor to control for how magnitude may affect exploration/exploitation.



Figure 5. Left: standard Euclidean EBM. Right: Riemannian EBM with $\tau = 0.5$ allows both structure and color, even for large step sizes. From top to bottom: smaller step sizes correspond to more muted, less structured images.

Effects of tuning τ on image fidelity and diversity



Figure 6. Top row shows standard EqM ($\tau = 0$) samples for a class. Increasing τ (middle and last row) improves diversity with minimal fidelity tradeoffs.

Interactive Image Generation Tool Example



Varied Samples after 100 Langevin Steps
Using high τ encourages efficient exploration before mode convergence in subsequent steps.



Single sample altered after 100 Langevin Steps
A single sample is selected here. Due to the composable nature of equilibrium dynamics models, mid-generation samples can be altered by users. Alterations for this example are highlighted in blue.



Sample after 200 Langevin Steps
Lowering τ in subsequent steps enhances "smoothing" behavior. Previous user changes are iterated upon.



Sample after 300 Langevin Steps
Samples reflect user preferences while maintaining both fidelity and diversity

Ethics We view image generative models as tools to augment rather than replace creative work, supporting exploration and iteration without displacing human authorship. Beyond fidelity to typical outcomes, we emphasize the importance of representing *atypical but plausible* samples. In a broader societal context, this corresponds to breaking visual stereotypes, avoiding perspective pigeonholing, and enabling more diverse and creative modes of expression in AI systems.

References

[1] Simone Azeglio and Arianna Di Bernardo. What's inside your diffusion model? a score-based riemannian metric to explore the data manifold. 2025. URL `https://arxiv.org/abs/2505.11128`.
[2] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019.
[3] Mark Girolami, Ben Calderhead, and Siu A Chin. Riemannian manifold hamiltonian monte carlo. *arXiv preprint arXiv:0907.1100*, 2009.
[4] Runqian Wang and Yilun Du. Equilibrium matching: Generative modeling with implicit energy-based models. *arXiv preprint arXiv:2510.02300*, 2025.