Particle Dynamics for Latent-Variable Energy-Based Models

Shiqin Tang[*1] **Shuxin Zhuang**[*1,2] **Runsheng Yu**[3] **Rong Feng**[1,2] **Shujian Yu**[4,5]
Hongzong Li[3] **Mingyang Zhao**[6] **Gaofeng Meng**[†1]

[1]CAIR-CAS, Hong Kong [2] City University of Hong Kong, Hong Kong
[3]The Hong Kong University of Science and Technology, Hong Kong
[4]Vrije Universiteit Amsterdam, Netherlands [5] UiT The Arctic University of Norway, Norway
[6]Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Abstract

Latent-variable energy-based models (LV-EBMs) assign a single normalized energy to joint pairs of observed data and latent variables, offering expressive generative modeling while capturing hidden structure. We recast maximum-likelihood training as a saddle problem over distributions on the latent and joint manifolds and view the inner updates as coupled Wasserstein gradient flows. The resulting algorithm alternates overdamped Langevin updates for a joint negative pool and for conditional latent particles with stochastic parameter ascent, requiring no discriminator or auxiliary networks. We prove existence and convergence under standard smoothness and dissipativity assumptions, with decay rates in KL divergence and Wasserstein-2 distance. The saddle-point view further yields an ELBO strictly tighter than bounds obtained with restricted amortized posteriors. Our method is evaluated on numerical approximations of physical systems and performs competitively against comparable approaches.

1 INTRODUCTION

Energy-based models (EBMs) define a globally normalized probability density via an energy function, avoiding rigid decoder parameterizations while enabling flexible learning of complex structure [LeCun et al., 2006, Du and Mordatch, 2019, Grathwohl et al., 2020]. Throughout, we write x for observed data and z for latent states. In many domains, however, the observed signal x is shaped by unobserved causes: style/content factors, nuisance variables, attributes, or even missing modalities. Latent-variable EBMs (LV-EBMs) address this mismatch by scoring pairs (x, z) under a joint Gibbs density $p_\theta(x, z) \propto \exp\{-E_\theta(x, z)\}$, then marginalizing z to obtain $p_\theta(x)$; the latent path grants both representational headroom and conditional control that EBMs on x alone struggle to match. This paper develops a simple and principled learning framework for LV-EBMs grounded in optimal-transport dynamics, yielding an implementable particle algorithm with no discriminator, auxiliary posterior, or explicit decoder.

Why LV-EBMs over EBMs on x alone? Firstly, explicit latents let the model factorize causes: by integrating over z, the marginal $p_\theta(x) = \int p_\theta(x, z)\, dz$ can represent richer multi-modality than a direct energy on x [Yin and Zhou, 2018, Du and Mordatch, 2019]. Secondly, LV-EBMs provide native conditionals $p_\theta(z \mid x)$ and $p_\theta(x \mid z)$ via clamping or conditional energies, enabling attribute manipulation, imputation, and semi-supervised learning without retrofitting [Grathwohl et al., 2020]. In short, the latent states improve controllability while being integrated within a simple unnormalized energy function.

Why LV-EBMs over common generative families? Likelihood-based decoders in VAEs are often paired with simple priors and amortized posteriors, which can limit expressivity or induce posterior collapse [Kingma and Welling, 2014, Cremer et al., 2018]. Normalizing flows require invertibility and Jacobian constraints that trade off architectural freedom against tractable likelihoods [Rezende and Mohamed, 2015, Dinh et al., 2017]. Diffusion and score-based models excel at synthesis but rely on carefully scheduled noise processes and learned score fields over x alone [Ho et al., 2020, Song and Ermon, 2019]. By contrast, LV-EBMs score joint configurations (x, z) with a nonlinear energy, decoupling representation from decoder parameterization while retaining conditional pathways through z.

Our approach in a sentence. We cast maximum-likelihood learning for $p_\theta(x, z)$ as a saddle problem over distributions on the latent manifold and the joint (x, z) manifold, and interpret the inner dynamics as coupled Wasserstein gradient flows [Jordan et al., 1998a, Ambrosio et al., 2008b, Chewi

[*]Equal contribution.
[†]Corresponding author.

et al., 2024]. The flows induce explicit Fokker–Planck evolutions with practical overdamped Langevin updates [Risken, 1996a, Welling and Teh, 2011]. Instantiated with particles, the algorithm alternates (i) per-example conditional latent moves approximating $p_\theta(z \mid x)$ and (ii) joint negative-sample refreshes for $p_\theta(x, z)$, followed by a stochastic ascent step on θ. Under standard smoothness, dissipativity, and log-Sobolev-type conditions, these dynamics are well-posed and enjoy stability controls that tie KL/Wasserstein distances to the training objective [Gross, 1975, Bakry and Émery, 1985, Durmus and Moulines, 2017].

Why particle dynamics over VI-based training of LV-EBMs? Classical latent-variable training builds and optimizes a separate approximate posterior (e.g., amortized Gaussian families) and often couples learning to a decoder network [Kingma and Welling, 2014]. Our formulation eliminates the auxiliary inference network and decoder: particles and Langevin steps realize the inner Wasserstein flows directly, using the energy to drive both joint and conditional updates. This avoids approximation bias from restrictive variational families while keeping the implementation compact. Conceptually, it parallels particle-based inference [Liu and Wang, 2016] but operates in a coupled latent-and-joint space aligned with MLE for EBMs.

Paper roadmap. Section 2 reviews related work such as (LV)-EBMs and particle-based training. Section 3 sets up notation and assumptions. Section 4 presents our training as a saddle problem realized by coupled Wasserstein gradient flows. Section 5 provides theoretical guarantees. Section 6 details practical conditional-generation procedures. Section 7 reports empirical results on numerical physical simulations *and* real-world UCI density-estimation benchmarks, with diagnostics for mixing and conditioning.

Contributions. ① We cast LV-EBM maximum-likelihood learning as a *coupled saddle problem* over conditional latent measures and joint (x, z) measures, and interpret the resulting inner updates as coupled Wasserstein gradient flows, yielding explicit Fokker–Planck/Langevin dynamics implemented by particles. ② We characterize the inner optima as the joint and conditional Gibbs targets and show that (unrestricted) particle training is *exact* in the sense that it recovers the dataset log-likelihood, while any VI restriction yields a weakly smaller ELBO with a strict gap under misspecification. ③ We provide well-posedness and convergence guarantees for the inner flows: exponential KL/W_2 contraction under log–Sobolev inequalities (LSIs), and a weaker (rate-free) convergence guarantee under mild coercivity/dissipativity.

2 BACKGROUND AND RELATED WORK

Optimization on probability manifolds. In variational inference, one typically minimizes a Kullback–Leibler divergence over a parametric family,

$$\min_{\phi \in \Phi} \; D_{\mathrm{KL}}\big(q_\phi(z) \,\|\, p^*(z)\big), \tag{1}$$

with $\phi \in \Phi \subset \mathbb{R}^m$ (see Blei et al. [2017] for a review). This is a finite-dimensional Euclidean optimization problem. A complementary viewpoint — enabled by optimal transport — optimizes directly over probability measures endowed with the L^2–Wasserstein (Otto) Riemannian structure [Otto, 2001, Ambrosio et al., 2005]:

$$\min_{q \in \mathcal{P}_2(\mathbb{R}^\ell)} \; D_{\mathrm{KL}}\big(q(z) \,\|\, p^*(z)\big), \tag{2}$$

where $\mathcal{P}_2(\mathbb{R}^\ell)$ is the set of distributions with finite second moments. This measure-space formulation removes the restriction to a variational family and links the minimization of $D_{\mathrm{KL}}(\cdot\|p^*)$ to Wasserstein gradient flows: the steepest–descent dynamics for (2) coincide with the Fokker–Planck/Langevin flow targeting p^* and admit the JKO implicit–Euler scheme for time discretization [Jordan et al., 1998c, Ambrosio et al., 2005, Peyré and Cuturi, 2019]. In parallel, natural-gradient methods view Euclidean VI as Riemannian optimization on statistical manifolds equipped with the Fisher metric [Amari, 1998], offering a complementary geometric perspective.

This manifold view naturally motivates particle procedures. Kuntz et al. [2023] propose a particle algorithm to solve the classical Expectation Maximization (EM) problem with formulation $\mathcal{P}_2(\mathbb{R}^\ell)$,

$$\max_{\theta \in \Theta, q \in \mathcal{P}_2(\mathbb{R}^\ell)} \; \mathbb{E}_{q(z)}[\log p_\theta(x|z)] - D_{\mathrm{KL}}\big(q(z) \,\|\, p_\theta(z|x)\big). \tag{3}$$

and realize it via finite-particle gradient flows, alternating with an M-step that updates θ through stochastic gradient descent. Related extensions apply particle flows to enrich the variational family in semi-implicit VI (SIVI), improving posterior expressivity while retaining tractable updates (cf. the SIVI framework of Yin and Zhou [2018] and particle-based refinements such as Lim and Johansen [2024]).

Particle methods for latent-variable models. Kuntz et al. [2023] study particle algorithms for maximum-likelihood training of latent-variable models from an EM perspective. Pang et al. [2020] learn an energy-based prior in the latent space of a generator model, and Marks et al. [2025] develop an interacting-particle Langevin method for latent energy-based models. These works are closely related in their use of particle or Langevin dynamics for latent-variable learning. However, Pang et al. and Marks et al. retain an explicit decoder or likelihood $p_\beta(x \mid z)$, whereas our

model directly defines a globally normalized joint energy $p_\theta(x, z) \propto \exp(-E_\theta(x, z))$ without an explicit decoder, generator, or amortized posterior. Consequently, our method couples per-example conditional latent particles with a persistent joint negative pool over (x, z), targeting a different saddle structure induced by the joint LV-EBM objective.

Adversarial training and particle dynamics for EBMs. VERA [Grathwohl et al., 2021] frames EBM learning as a bi-level min–max problem: an entropy-regularized generator adversarially approximates the model's negative phase, yielding a variational upper bound on NLL and stable training without Markov chains. In contrast to amortized variational approximations $q_\phi(x)$, Neklyudov et al. [2021] replace the parametric family with distributions $q \in \mathcal{P}_2(\mathbb{R}^d)$ and construct deterministic particle flows whose finite-time transport tracks the evolving model density, effectively substituting short-run MCMC with Wasserstein gradient dynamics.

3 PROBLEM SETTINGS

This section fixes notation and definitions and states the standing assumptions used throughout.

Ambient spaces and measures. Suppose latent variable z and observed variable x take values in $\mathbb{R}^\ell$ and $\mathbb{R}^d$, respectively. Let ν and μ denote the Lebesgue measures on $\mathbb{R}^d$ and $\mathbb{R}^\ell$, respectively; on the product we use $\nu \otimes \mu$. Whenever gradients or Wasserstein geometry are invoked, we identify $\mathbb{R}^d \times \mathbb{R}^\ell \cong \mathbb{R}^{d+\ell}$ with its standard inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$.

Entropy functionals and Wasserstein metrics. For a probability measure r absolutely continuous w.r.t. a reference measure λ, we write

$$H_\lambda(r) \;:=\; -\int \log\Big(\frac{dr}{d\lambda}\Big)\, dr. \tag{4}$$

We denote by $W_{2,z}$ the 2-Wasserstein metric on $\mathcal{P}_2(\mathbb{R}^\ell)$ and by $W_{2,j}$ the 2-Wasserstein metric on $\mathcal{P}_2(\mathbb{R}^{d+\ell})$. When the ambient space can be implied from context, we abbreviate $H_\mu(\cdot)$, $H_{\nu\otimes\mu}(\cdot)$ and write simply W_2.

Probability classes. We use $\mathcal{P}_2(\mathcal{X})$ to denote the set of Borel probability measures on $\mathcal{X}$ with finite second moment and let

$$\mathcal{Q} := \big\{\, q \in \mathcal{P}_2(\mathbb{R}^\ell) :\; q \ll \mu,\; H_\mu(q) > -\infty \,\big\}$$
$$\tilde{\mathcal{Q}} := \big\{\, \tilde{q} \in \mathcal{P}_2(\mathbb{R}^{d+\ell}) :\; \tilde{q} \ll (\nu \otimes \mu),\; H_{\nu\otimes\mu}(\tilde{q}) > -\infty \,\big\}$$

For each data point x_i, we associate a measure $q_i \in \mathcal{Q}$ over z, while $\tilde{q} \in \tilde{\mathcal{Q}}$ is a joint pool over (x, z). Following contrastive EBM terminology (not supervised labels), we will refer to samples from q_i as *positive* and samples from $\tilde{q}$ as *negative*; operational use of this terminology appears in Section 4.

Latent-variable EBM. For fixed θ, let $E : \mathbb{R}^d \times \mathbb{R}^\ell \to \mathbb{R} \cup \{+\infty\}$ be measurable and define the Gibbs density

$$p_\theta(x, z) \;=\; \frac{1}{Z(\theta)}\, \exp\big(-E(x, z; \theta)\big), \tag{5}$$

where the partition function $Z : \Theta \to \mathbb{R}_+$ is defined as

$$Z(\theta) \;=\; \int_{\mathbb{R}^d \times \mathbb{R}^\ell} \exp\big(-E(x, z; \theta)\big)\, (\nu \otimes \mu)(dx\, dz). \tag{6}$$

Allowing $E = \infty$ encodes hard constraints. We denote the marginal distribution as $p_\theta(x) = \int p_\theta(x, z)\, \mu(dz)$.

Assumption 1 (Regularity and tails of the joint energy)**.** $E(\cdot, \cdot; \theta)$ is C^1 with L-Lipschitz joint gradient and is (m, b)-dissipative:

$$\|\nabla E(u; \theta) - \nabla E(v; \theta)\| \le L\|u - v\|, \tag{7}$$
$$\langle u, \nabla E(u; \theta)\rangle \;\ge\; m\|u\|^2 - b, \quad \forall\, u, v \in \mathbb{R}^{d+\ell}. \tag{8}$$

In particular, this assumption ensures that $Z(\theta) < \infty$ in (5). Let $\mathcal{D} = \{x^i : i \in [N]\}$ be the training set with empirical distribution $p_{\mathcal{D}}(x) = \frac{1}{N}\sum_{i=1}^N \delta_{x^i}(x)$.

Assumption 2 (Functional inequalities)**.** For fixed θ, the target p_θ satisfies a log-Sobolev inequality (LSI) with constant $\rho > 0$ on $\mathbb{R}^{d+\ell}$. Consequently, Talagrand's $T_2(\rho)$ holds: $W_{2,j}^2(\tilde{q}, p_\theta) \le \frac{2}{\rho} D_{\mathrm{KL}}(\tilde{q}\|p_\theta)$ for all $\tilde{q} \in \tilde{\mathcal{Q}}$. When x is clamped, the conditional targets $p_\theta(\cdot \mid x)$ satisfy an LSI with constant $\rho_z > 0$ on $\mathbb{R}^\ell$ (for p_D-a.e. x), yielding $W_{2,z}$–Talagrand bounds on $\mathcal{Q}$.

Remark 1. In practice, we can parameterize E as

$$E_\theta(x, z) = U_\theta(x, z) + \frac{\lambda_x}{2}\|x\|^2 + \frac{\lambda_z}{2}\|z\|^2, \tag{9}$$

with $\lambda_x, \lambda_z > 0$ and keep ∇E_θ L-Lipschitz via spectral norm or gradient clipping. This ensures normalizability and tail dissipativity; the quadratic envelope makes standard LSI and T_2 conditions plausible.

Notational economy. We maintain two distinct Wasserstein spaces throughout: the latent manifold $(\mathcal{Q}, W_{2,z})$ on $\mathbb{R}^\ell$ and the joint manifold $(\tilde{\mathcal{Q}}, W_{2,j})$ on $\mathbb{R}^{d+\ell}$. To avoid redundancy, we (i) suppress subscripts on H and W_2 when unambiguous from the argument, and (ii) use the same font ∇ for gradients in z or (x, z), specifying the variable by a subscript only when needed (e.g., $\nabla_z E$, $\nabla_{x,z} E$).

4 PARTICLE DYNAMICS

In this section, we cast maximum-likelihood learning of the joint Gibbs model $p_\theta(x, z)$ as a saddle problem over

distributions on the latent manifold and the joint (x, z) manifold. This viewpoint replaces amortized inference with coupled Wasserstein gradient flows: per-example latent particles approximate $p_\theta(z|x)$ while a joint negative pool tracks $p_\theta(x, z)$. Discretizing the flows with overdamped Langevin dynamics yields a simple, fully energy-driven particle algorithm that alternates inner particle moves with a stochastic ascent step on θ. We begin by deriving the saddle objective and its inner optima, then instantiate the corresponding flows and their particle updates.

Saddle-point MLE for LV-EBMs ("adversarial" view). We train the LV-EBMs by maximizing the log likelihood $p_\theta(x)$ represented in the following form,

$$\begin{aligned}\log p_\theta(x) &= \log \int \exp(-E(x,z;\theta))\, dz \\ &\quad - \log \iint \exp(-E(x,z;\theta))\, dx\, dz \\ &= \log \mathbb{E}_{q(z)}\Big[\frac{\exp(-E(x,z;\theta))}{q(z)}\Big] \\ &\quad - \log \mathbb{E}_{\tilde{q}(x,z)}\Big[\frac{\exp(-E(x,z;\theta))}{\tilde{q}(x,z)}\Big].\end{aligned}$$

We use the Donsker–Varadhan variational identity $\log \int \exp(f)\, d\lambda = \sup_{q \ll \lambda}\{\mathbb{E}_q[f] + H_\lambda(q)\}$. Applying it to the latent term and to the joint normalizer gives

$$\begin{aligned}\log p_\theta(x) &= \sup_{q \in Q} \mathbb{E}_{q(z)}\Big[\log \frac{\exp(-E(x,z;\theta))}{q(z)}\Big] \\ &\quad - \sup_{\tilde{q} \in \tilde{Q}} \mathbb{E}_{\tilde{q}(x,z)}\Big[\log \frac{\exp(-E(x,z;\theta))}{\tilde{q}(x,z)}\Big] \\ &= \sup_{q \in Q} \inf_{\tilde{q} \in \tilde{Q}} \Big\{ - \mathbb{E}_{q(z)}[E(x,z;\theta)] \\ &\quad + \mathbb{E}_{\tilde{q}(x,z)}[E(x,z;\theta)] + H(q) - H(\tilde{q}) \Big\}.\end{aligned}$$

where $H(\cdot)$ represents entropy defined in (4). We use the notation $q^{\neg i} = \{q_j : (j \in [N]) \wedge (j \neq i)\}$ and $\bar{q} = \{q_j : j \in [N]\}$. Consequently, we establish the following saddle-point formulation:

$$\begin{aligned}\max_{\theta, q^i \in Q} \min_{\tilde{q} \in \tilde{Q}} \quad &\Big\{ F(\bar{q}, \tilde{q}, \theta) := \mathbb{E}_{\tilde{q}(x,z)}[E(x,z;\theta)] - H(\tilde{q}) \\ &- \frac{1}{N} \sum_{i=1}^{N} \big(\mathbb{E}_{q^i(z)}[E(x^i,z;\theta)] - H(q^i) \big) \Big\}\end{aligned} \tag{10}$$

The first variations of F are given by

$$\begin{aligned}\frac{\delta F}{\delta q^i}(z) &= -E_\theta(x^i, z) \;-\; \log q^i(z) \;+\; c_i, \\ \frac{\delta F}{\delta \tilde{q}}(x, z) &= \;\; E_\theta(x, z) \;+\; \log \tilde{q}(x, z) \;+\; c,\end{aligned} \tag{11}$$

with c_i and c being constants. Following Otto calculus [Ambrosio et al., 2008a, Santambrogio, 2015, Jordan et al., 1998b], we denote the Wasserstein gradient by $\mathbb{W}_q F := \nabla \frac{\delta F}{\delta q}$. Since each q^i is maximized and $\tilde{q}$ is minimized in (10), we take ascent for q^i and descent for $\tilde{q}$:

$$\mathbb{W}_{q^i} F(\bar{q}, \tilde{q}, \theta) = -\nabla_z E_\theta(x^i, z) \;-\; \nabla_z \log q^i(z), \tag{12}$$

$$\mathbb{W}_{\tilde{q}} F(\bar{q}, \tilde{q}, \theta) = \;\; \nabla_{x,z} E_\theta(x, z) \;+\; \nabla_{x,z} \log \tilde{q}(x, z). \tag{13}$$

Coupled with the continuity equation $\partial_t q_t = -\nabla \cdot (q_t v_t)$ and the choices $v_t^i = \mathbb{W}_{q_t^i} F$ and $\tilde{v}_t = -\mathbb{W}_{\tilde{q}_t} F$, we obtain the Fokker–Planck equations (FPEs)

$$\partial_t q_t^i(z) = \nabla_z \cdot \big(q_t^i(z)\, \nabla_z E_\theta(x^i, z) \big) + \Delta_z q_t^i(z), \tag{14}$$

$$\partial_t \tilde{q}_t(x, z) = \nabla_{x,z} \cdot \big(\tilde{q}_t(x, z)\, \nabla_{x,z} E_\theta(x, z) \big) + \Delta_{x,z} \tilde{q}_t(x, z), \tag{15}$$

for $i \in [N]$. Under standard smoothness and confining growth of E_θ, (14)–(15) are the W_2–gradient flows of the corresponding free energies [Ambrosio et al., 2008a, Jordan et al., 1998b]. The FPEs (14)–(15) are realized by the Itô SDEs

$$\mathrm{d}z_t^i = -\nabla_z E_\theta(x^i, z_t^i)\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}B_t^{(i)}, \tag{16}$$

$$\mathrm{d}\tilde{z}_t = -\nabla_z E_\theta(\tilde{x}_t, \tilde{z}_t)\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}B_t^{(z)}, \tag{17}$$

$$\mathrm{d}\tilde{x}_t = -\nabla_x E_\theta(\tilde{x}_t, \tilde{z}_t)\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}B_t^{(x)}, \tag{18}$$

where B_t represents standard Brownian motions. Their marginal laws solve (14)–(15) in the weak sense [Risken, 1996b, Pavliotis, 2014, Ch. 5].

Particle algorithm. We approximate the Wasserstein flows using persistent particle systems: for each datum x^i we maintain latent particles $\{z_t^{ik}\}_{k=1}^M$ evolving under conditional Langevin dynamics, and we maintain a joint negative pool $\{(\tilde{x}_t^j, \tilde{z}_t^j)\}_{j=1}^D$ evolving under joint Langevin dynamics. Particles are initialized i.i.d. but thereafter are correlated across iterations due to persistence. The corresponding empirical measures are

$$\hat{q}_t^i \;=\; \frac{1}{M} \sum_{k=1}^{M} \delta_{z_t^{ik}} \quad \text{and} \quad \hat{\tilde{q}}_t \;=\; \frac{1}{D} \sum_{j=1}^{D} \delta_{(\tilde{x}_t^j, \tilde{z}_t^j)}\,,$$

which approximate q^i and $\tilde{q}$, respectively. At each iteration we take overdamped Langevin steps for all particles and perform a stochastic ascent step on θ that increases the energy contrast (negative minus positive):

$$\theta_{t+1} = \theta_t - \alpha \nabla_\theta \Big[\frac{1}{N} \sum_i \mathbb{E}_{\hat{q}_{t+1}^i} E(x^i, z; \theta_t) - \mathbb{E}_{\hat{\tilde{q}}_{t+1}} E(x, z; \theta_t) \Big] \tag{19}$$

The complete particle dynamics and parameter updates are summarized in Algorithm 1.

5 THEORETICAL GUARANTEES

We first identify the inner optima of the saddle objective as joint and conditional Gibbs measures. We then show that

Algorithm 1: Langevin Dynamics for LV-EBMs

Input: data $\{x^i\}_{i=1}^N$; particle counts M (latent), D (joint); step sizes η_z, η_{xz}; learning rate α; iterations T

Init: $z_0^{ik} \sim \mathcal{N}(0, I_l)$ for all i, k;
$\tilde{z}_0^j \sim \mathcal{N}(0, I_l), \quad \tilde{x}_0^j \sim \mathcal{N}(0, I_d)$ for all j;

for $t = 0$ **to** $T - 1$ **do**

`// Latent (conditional) Langevin:`

for $i = 1$ **to** N **do**

for $k = 1$ **to** M **do**

sample $\xi_t^{ik} \sim \mathcal{N}(0, I_l)$

$z_{t+1}^{ik} \leftarrow z_t^{ik} - \eta_z \nabla_z E(x^i, z_t^{ik}; \theta_t) + \sqrt{2\eta_z}\, \xi_t^{ik}$

`// Joint (negative) Langevin:`

for $j = 1$ **to** D **do**

sample $\xi_t^{j,(x)} \sim \mathcal{N}(0, I_d)$, $\xi_t^{j,(z)} \sim \mathcal{N}(0, I_l)$;

$\tilde{x}_{t+1}^j \leftarrow \tilde{x}_t^j - \eta_{xz} \nabla_x E(\tilde{x}_t^j, \tilde{z}_t^j; \theta_t) + \sqrt{2\eta_{xz}}\, \xi_t^{j,(x)}$

$\tilde{z}_{t+1}^j \leftarrow \tilde{z}_t^j - \eta_{xz} \nabla_z E(\tilde{x}_t^j, \tilde{z}_t^j; \theta_t) + \sqrt{2\eta_{xz}}\, \xi_t^{j,(z)}$

`// Parameter ascent on the energy contrast`

$$\theta_{t+1} \leftarrow \theta_t + \alpha \Big[\frac{1}{D} \sum_{j=1}^{D} \nabla_\theta E(\tilde{x}_{t+1}^j, \tilde{z}_{t+1}^j; \theta_t) - \frac{1}{NM} \sum_{i=1}^{N} \sum_{k=1}^{M} \nabla_\theta E(x^i, z_{t+1}^{ik}; \theta_t) \Big]$$

the associated Fokker–Planck (FP) flows contract exponentially in D_{KL} and W_2 under log–Sobolev inequalities (LSIs), formalizing that our Langevin inner loops are Wasserstein gradient flows toward the model targets [Jordan et al., 1998c, Ambrosio et al., 2008b]. To clarify the role of LSIs, we also state a weaker convergence guarantee (without rates) under standard coercivity/dissipativity conditions. Finally, we describe a variational-inference (VI) training variant for LV-EBMs—obtained by restricting the inner distributions to amortized parametric families—and prove a tighter-than-VI result.

Proposition 1 (Inner optima for (10): joint and conditional Gibbs principles)**.** Under Assumption 1, for any $\tilde{q} \in \tilde{\mathcal{Q}}$ and any $x \in \mathbb{R}^d$, $q \in \mathcal{Q}$,

$$\mathbb{E}_{\tilde{q}}[E(x, z; \theta)] - H_{\nu\otimes\mu}(\tilde{q}) = D_{\mathrm{KL}}(\tilde{q} \| p_\theta) - \log Z(\theta), \tag{20}$$

$$\mathbb{E}_q[E(x, z; \theta)] - H_\mu(q) = D_{\mathrm{KL}}(q \| p_\theta(\cdot \mid x)) - \log p_\theta(x). \tag{21}$$

Hence the inner optima in (10) are

$$\tilde{q}^\star(x, z) = p_\theta(x, z) \quad \text{and} \quad q^\star(z \mid x) = p_\theta(z \mid x), \tag{22}$$

and in each case the minimizer is unique up to $\nu \otimes \mu$ (resp. μ) null sets.

Theorem 1 (Convergence of inner Fokker–Planck flows)**.** For a fixed θ, assume:

(J) (Joint LSI) The target p_θ on $\mathbb{R}^{d+\ell}$ satisfies a log–Sobolev inequality with constant $\rho > 0$.

(C) (Conditional LSI) For every $x \in \mathbb{R}^d$ for which $p_\theta(\cdot \mid x)$ is defined, the conditional target on $\mathbb{R}^\ell$ satisfies a log–Sobolev inequality with constant $\rho_z > 0$.

Let $\tilde{q}_t \in \tilde{\mathcal{Q}}$ solve the joint Fokker–Planck equation

$$\partial_t \tilde{q}_t = \nabla_{x,z} \cdot \big(\tilde{q}_t \nabla_{x,z} E(\cdot; \theta) \big) + \Delta_{x,z} \tilde{q}_t, \quad \tilde{q}_{t=0} = \tilde{q}_0 \in \tilde{\mathcal{Q}},$$

and, for any fixed $x \in \mathbb{R}^d$, let $q_t^x \in \mathcal{Q}$ solve the conditional Fokker–Planck equation

$$\partial_t q_t^x = \nabla_z \cdot \big(q_t^x \nabla_z E(x, \cdot; \theta) \big) + \Delta_z q_t^x, \quad q_{t=0}^x = q_0^x \in \mathcal{Q}.$$

Then, for all $t \geq 0$,

$$D_{\mathrm{KL}}(\tilde{q}_t \,\|\, p_\theta) \leq e^{-2\rho t}\, D_{\mathrm{KL}}(\tilde{q}_0 \,\|\, p_\theta),$$
$$W_{2,j}(\tilde{q}_t, p_\theta) \leq \sqrt{\tfrac{2}{\rho}}\, e^{-\rho t}\, D_{\mathrm{KL}}(\tilde{q}_0 \,\|\, p_\theta)^{1/2},$$

and, for every such x,

$$D_{\mathrm{KL}}\big(q_t^x \,\|\, p_\theta(\cdot \mid x)\big) \leq e^{-2\rho_z t}\, D_{\mathrm{KL}}\big(q_0^x \,\|\, p_\theta(\cdot \mid x)\big), \tag{23}$$
$$W_{2,z}\big(q_t^x, p_\theta(\cdot \mid x)\big) \leq \sqrt{\tfrac{2}{\rho_z}}\, e^{-\rho_z t}\, D_{\mathrm{KL}}\big(q_0^x \,\|\, p_\theta(\cdot \mid x)\big)^{1/2}. \tag{24}$$

Proof sketch. For the joint flow, the entropy dissipation identity gives $\frac{d}{dt} D_{\mathrm{KL}}(\tilde{q}_t \| p_\theta) = -\mathcal{I}(\tilde{q}_t \| p_\theta)$, where $\mathcal{I}$ is the relative Fisher information. By (J), $\mathcal{I} \geq 2\rho\, D_{\mathrm{KL}}$, hence $\frac{d}{dt} D_{\mathrm{KL}} \leq -2\rho\, D_{\mathrm{KL}}$; Grönwall yields the KL decay, and Talagrand's $T_2(\rho)$ gives the $W_{2,j}$ bound [Santambrogio, 2015, Ch. 20]. The conditional case is identical with x fixed and (C) replacing (J). □

Theorem 2 (Weak convergence without LSI)**.** Fix x and consider the conditional dynamics in $z \in \mathbb{R}^\ell$ with energy $E(x, z; \theta)$. Assume, for a.e. x:

1. **Coercive growth:** $E(x, z; \theta) \to \infty$ as $|z| \to \infty$.
2. **Smoothness:** $E(x, \cdot; \theta) \in C^2(\mathbb{R}^\ell)$ and $\nabla_z E(x, \cdot; \theta)$ is locally Lipschitz.
3. **Dissipativity:** $\langle z, \nabla_z E(x, z; \theta) \rangle \geq m|z|^2 - b$ for some $m > 0$, $b \geq 0$.
4. **Well-posedness (FP/SDE):** The conditional FP equation in Theorem 1 (equivalently, the Langevin SDE) is non-explosive and admits a unique weak solution for any $q_0^x \in \mathcal{P}_2(\mathbb{R}^\ell)$.

Let q_t^x denote the solution of the conditional Fokker–Planck equation. Let $p_\theta(z \mid x) \propto e^{-E(x,z;\theta)}$ be the conditional Gibbs measure (well-defined by coercivity). Then $q_t^x \rightharpoonup p_\theta(\cdot \mid x)$ weakly as $t \to \infty$.

Remark (rates vs. qualitative convergence). The LSI assumptions in Theorem 1 are used to obtain *explicit exponential* contraction rates in KL and Wasserstein distance. Theorem 2 shows that, even without LSIs, the conditional Langevin inner loop still converges to the correct Gibbs target under mild tail/dissipativity conditions (typically satisfied by standard regularized neural energies).

Variational training for LV-EBMs. Variational inference (VI) for LV-EBMs restricts the variational families to parametric classes: $q^i(z) \in \mathcal{S}_{\mathrm{pos}}(x^i) = \{q_\phi(z \mid x^i) : \phi \in \Phi\}$ and $\tilde{q}(x,z) \in \mathcal{S}_{\mathrm{neg}} = \{q_\psi(x,z) : \psi \in \Psi\}$, e.g., Gaussian/amortized posteriors and a tractable joint family for the global normalizer. The resulting (dataset) ELBO is then

$$\begin{aligned}\mathrm{ELBO}_{\mathrm{VI}}(\theta,\phi,\psi) &:= \big[\mathbb{E}_{q_\psi} E(x,z;\theta) - H(q_\psi)\big] \\ &+ \frac{1}{N}\sum_{i=1}^{N}\big[-\mathbb{E}_{q_\phi(z|x^i)}E(x^i,z;\theta) + H(q_\phi(\cdot \mid x^i))\big],\end{aligned} \tag{25}$$

which is maximized jointly over (θ,ϕ,ψ). In contrast, particle dynamics allows q^i and $\tilde{q}$ to range over nonparametric particle measures that, in the limit of enough particles/steps, approximate any element of $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$.

Theorem 3. Define, for each x^i,

$$\begin{aligned}\mathcal{A}_i(q) &:= -\mathbb{E}_{q(z)}[E(x^i,z;\theta)] + H(q), \\ \mathcal{B}(\tilde{q}) &:= -\mathbb{E}_{\tilde{q}(x,z)}[E(x,z;\theta)] + H(\tilde{q}).\end{aligned}$$

For any families $\mathcal{S}_{\mathrm{pos}}(x) \subseteq \mathcal{Q}$ and $\mathcal{S}_{\mathrm{neg}} \subseteq \tilde{\mathcal{Q}}$, let

$$\mathrm{ELBO}(\theta;\mathcal{S}_{\mathrm{pos}},\mathcal{S}_{\mathrm{neg}}) := \frac{1}{N}\sum_{i=1}^{N}\sup_{q\in\mathcal{S}_{\mathrm{pos}}(x^i)}\mathcal{A}_i(q) - \sup_{\tilde{q}\in\mathcal{S}_{\mathrm{neg}}}\mathcal{B}(\tilde{q}).$$

Consider the following cases:

(1) **Exactness with unrestricted sets.** If $\mathcal{S}_{\mathrm{pos}}(x) = \mathcal{Q}$ and $\mathcal{S}_{\mathrm{neg}} = \tilde{\mathcal{Q}}$, then

$$\begin{aligned}\mathrm{ELBO}(\theta;\mathcal{Q},\tilde{\mathcal{Q}}) &= \frac{1}{N}\sum_{i=1}^{N}\log\int e^{-E(x^i,z;\theta)}dz \\ &- \log\iint e^{-E(x,z;\theta)}dx\,dz = \frac{1}{N}\sum_{i=1}^{N}\log p_\theta(x^i).\end{aligned}$$

(2) **Monotonicity & domination.** For any $\mathcal{S}_{\mathrm{pos}}(x) \subseteq \mathcal{Q}$ and $\mathcal{S}_{\mathrm{neg}} \subseteq \tilde{\mathcal{Q}}$, we have

$$\begin{aligned}&\mathrm{ELBO}(\theta;\mathcal{S}_{\mathrm{pos}},\mathcal{S}_{\mathrm{neg}}) \\ &\le \mathrm{ELBO}(\theta;\mathcal{Q},\tilde{\mathcal{Q}}) = \frac{1}{N}\sum_{i=1}^{N}\log p_\theta(x^i).\end{aligned}$$

Moreover, the inequality is strict if either $p_\theta(z \mid x^i) \notin \mathcal{S}_{\mathrm{pos}}(x^i)$ for a set of i with positive frequency, or $p_\theta(x,z) \notin \mathcal{S}_{\mathrm{neg}}$.

Remark. The full proofs for proposition 1 and theorem 2 and 3 are shown in Appendix A, B, and C, respectively.

6 CONDITIONAL GENERATION

This section clarifies how reconstruction and conditional generation are performed in a joint LV-EBM without introducing an explicit decoder. Since the model directly defines a joint energy over (x,z), reconstructing x from z amounts to drawing from or optimizing under the conditional distribution

$$x \sim p_\theta(x|z) \propto \exp(-E(x,z;\theta)). \tag{26}$$

Consider the following ways:

(i) Deterministic generation (MAP reconstruction): We can find the most likely x according to $x^* = \arg\max_x \log p_\theta(x|z)$. To get x^*, we can perform gradient descent given an randomly initiated x_0:

$$x_{t+1} = x_t - \eta\nabla_x E(x_t,z;\theta). \tag{27}$$

(ii) Stochastic generation (Langevin dynamics): Based on (26), we can find the set x to be the one that minimizes the energy, $x^* = \arg\min_x E(x,z;\theta)$. In practice, we can run overdamped Langevin dynamics in the x-coordinates while holding z fixed:

$$x_{t+1} = x_t - \eta\nabla_x E(x_t,z;\theta) + \sqrt{2\eta}\epsilon_t. \tag{28}$$

Alternatively, for better mixing in high-dimensional x, we may introduce momentum:

$$\begin{aligned}v_{k+1} &= (1-\gamma)v_k - \eta\nabla_x E(x_k,z;\theta) + \sqrt{2\gamma\eta}\epsilon_k, \\ x_{k+1} &= x_k + v_{k+1}.\end{aligned} \tag{29}$$

The former approach gives a sharp deterministic MAP estimation of x, while the latter preserves multimodality/uncertainty.

7 EXPERIMENTS

We aim to evaluate the effectiveness of our method for both *posterior inference* and *generative reconstruction*. To this end, we consider three controllable geometric families: **2D concentric rings**, **3D concentric spheres**, and **2D harmonic-modulated rings** (Fourier-based radius modulation). These datasets allow us to assess whether the models can accurately approximate the conditional posterior. We compare against **VAE**, **Non-Amortized VI (NAVI)** and **Hard EM** on multiple metrics. All our experiments were conducted on the server with 48-core 3.00GHz Intel(R) Xeon(R) Gold 6248R CPU and 8 NVIDIA A30 GPUs. The code is made available at `https://github.com/sxzhuang/LV_EBM`.

		Ours	VAE	NAVI	Hard EM
LCR-2D	ELBO	**2.5048**	2.3038	2.3039	2.3025
	RMSE	**0.1612**	0.7585	0.7651	0.7655
	MMD	**0.1549**	0.5712	0.5259	0.5680
	W_2^2	**0.2161**	1.0821	1.0787	1.0834
LCS-3D	ELBO	**3.4984**	2.8499	2.8505	2.8274
	RMSE	**0.1303**	0.6247	0.6189	0.6225
	MMD	**0.0129**	0.5376	0.5471	0.5474
	W_2^2	**0.2567**	1.0837	1.0748	1.0869
HMR-2D	ELBO	**5.2234**	3.1866	3.1813	3.1850
	RMSE	**0.0375**	0.8565	1.1890	1.1881
	MMD	**0.0012**	0.3485	0.4843	0.5620
	W_2^2	**0.1663**	1.1998	1.6820	1.6739

Table 1: Results across three scenarios, four methods, and four evaluation metrics.

DATASETS

We consider three datasets and use fixed train/test splits with shared random seeds across methods. Each synthetic dataset contains $50{,}000$ samples, with $40{,}000$ used for training and $10{,}000$ held out for testing. In all synthetic experiments, x denotes the observed data point and z denotes the latent control variable. The learned parameter is the neural energy parameter θ in $E_\theta(x, z)$. For evaluation, the ground truth is the held-out observation x generated by the simulator; the latent variable z is not used as a supervised training target.

Here $\mathcal{C}$ denotes a finite set of base radii, and R_c is the base radius sampled from $\mathcal{C}$. The noise variables ε_r and ε_x denote radial noise and observation noise, respectively. In HMR-2D, the coefficients a_k and b_k are harmonic coefficients that control the angle-dependent radius modulation.

1. **2D concentric rings (LCR-2D).** Here the latent control is $z = (r, \phi)$, and the observation is a noisy Cartesian point generated from polar coordinates. Sample R_c from $\mathcal{C}$, $\phi \sim \mathrm{Unif}[0, 2\pi)$, $\varepsilon_r \sim \mathcal{N}(0, \sigma_r^2)$, $\varepsilon_x \sim \mathcal{N}(0, \sigma_x^2 I_2)$; set $r = R_c + \varepsilon_r$, $x = r\,[\cos\phi, \sin\phi]^\top + \varepsilon_x$.
2. **3D concentric spheres (LCS-3D).** Here the latent control is $z = (r, \theta, \phi)$, and the observation is a noisy Cartesian point generated from spherical coordinates. Sample R_c from $\mathcal{C}$; draw a uniform direction on $\mathbb{S}^2$ by $u \sim \mathrm{Unif}[-1, 1]$, $\phi \sim \mathrm{Unif}[0, 2\pi)$, $\theta = \arccos u$. With $\varepsilon_r \sim \mathcal{N}(0, \sigma_r^2)$, $\varepsilon_x \sim \mathcal{N}(0, \sigma_x^2 I_3)$: $r = R_c + \varepsilon_r$, $x = r\,[\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta]^\top + \varepsilon_x$.
3. **2D harmonic-modulated rings (HMR-2D).** Here the latent control is $z = (R_c, \phi, a_1, \ldots, a_K, b_1, \ldots, b_K)$. The radius is not a single latent coordinate but is determined by the base radius, angle, and harmonic coefficients. On top of R_c, define $r(\phi) = R_c + \sum_{k=1}^{K} \big[a_k \cos(k\phi) + b_k \sin(k\phi)\big] + \varepsilon_r$ with $\phi \sim \mathrm{Unif}[0, 2\pi)$, $\varepsilon_r \sim \mathcal{N}(0, \sigma_r^2)$. Let $\varepsilon_x \sim \mathcal{N}(0, \sigma_x^2 I_2)$, $x = r(\phi)\,[\cos\phi, \sin\phi]^\top + \varepsilon_x$.

7.1 BASELINES

VAE. Encoder $q_\phi(z \mid x)$ and decoder $p_\theta(x \mid z)$ are diagonal Gaussians; we optimize the ELBO with the reparameterization trick.

Non-Amortized VI (NAVI). Per-sample variational parameters $(\mu_i, \log\sigma_i^2)$ are optimized directly via K inner gradient-ascent steps; then θ is updated to maximize the ELBO. No encoder is used, removing the amortization gap.

Hard EM. E-step: MAP optimization of z to maximize $\log p(x \mid z; \theta) + \log p(z)$. M-step: update θ by maximizing the conditional likelihood.

7.2 EVALUATION METRICS

We report four metrics (arrows indicate the preferred direction):

1. **ELBO / NLL (↑).** We report absolute ELBO for our method, VAE, and NAVI, and absolute NLL for Hard EM.
2. **RMSE (↓).**Root mean squared error between the reconstruction $\hat{x}$ and the ground truth x.
3. **MMD (RBF) (↓).** Gaussian-kernel MMD; bandwidth via median heuristic or multi-scale.
4. W_2^2 **(Sinkhorn) (↓).** Entropy-regularized OT approximation to squared 2-Wasserstein; common ε and iteration budget across settings.

7.3 MAIN RESULTS

Table 1 summarizes all results across the three settings and four methods. We observe: 1) **ELBO / NLL:** Our method is better than or competitive with the strongest baseline on all datasets, consistent with direct, energy-driven particle updates in both the joint and conditional spaces that avoid approximation bias from restricted variational families. 2) **RMSE and MMD (RBF):** Our method shows a clear advantage in generative reconstruction, achieving *lower* RMSE and MMD, indicating more reliable preservation of and alignment with the ground-truth data structure. 3) W_2^2 **(Sinkhorn):** Our method yields *lower* distances across all three datasets, suggesting better global alignment of the distributions in both summary statistics and distributional shape.

To illustrate the training dynamics and convergence of our method, we plot the training loss for each setting together with the mean $\pm$ standard deviation over three independent runs. As shown in Figure 1, the objective decreases before plateauing, and variance decreases steadily, indicating stable

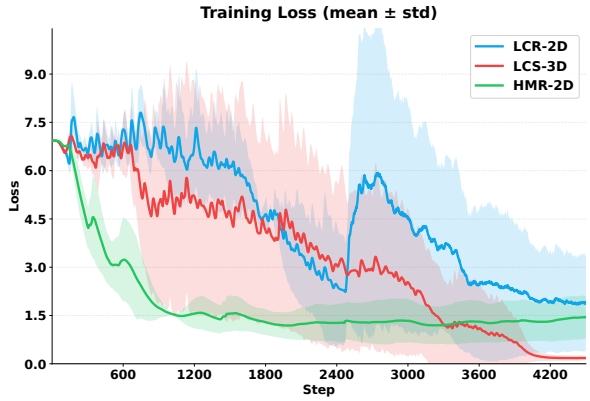

Figure 1: Training loss vs. step (up to 4500) for three scenarios. Solid lines depict the mean over three runs; shaded regions show ±1 standard deviation.

convergence. This behavior is consistent with theoretical results on entropy decay and Wasserstein contraction for particle Langevin updates under standard assumptions.

Reconstruction Method. For each observation x, we run a short-run Langevin posterior probe under the energy $E(x, z)$ to produce M candidates and pick the MAP code $\hat{z} = \arg\min_z E(x, z)$. With the energy network frozen, we train a decoder g_θ by minimizing $\|g_\theta(\hat{z}) - x\|_2^2$ over the training set. At inference, we repeat the probe to obtain $\hat{z}$ and output the reconstruction $x_{\text{rec}} = g_\theta(\hat{z})$. We visualize the training snapshots for LCR-2D as shown in Figure 2. Points are colored by three radius-based classes. Test particles evolve from a scattered layout to tight alignment with the target rings; inter-class boundaries become clear and within-class density becomes more uniform, showing good mixing and mode preservation.

Figure 3 demonstrates the alignment between reconstructed and ground-truth distributions under three settings. The reconstructions closely track the ground truth in ring or shell thickness.

7.4 REAL-WORLD DENSITY ESTIMATION ON UCI BENCHMARKS

We additionally evaluate on two standard UCI density-estimation benchmarks using the widely adopted MAF-preprocessed setup: **POWER** (dimension D=6) and **MINIBOONE** (dimension D=43). Following common practice and to match compute budgets across methods, we subsample 20,000 training points and evaluate on 2,000 held-out test points (fixed seed; full provenance and preprocessing details are in the supplementary material). Table 2 reports the same metrics as above. Our method achieves the best performance across metrics and datasets, consistent with the advantage of unrestricted (particle) inner distributions over parametric VI restrictions.

		Ours	VAE	NAVI	Hard EM
UCI-Power	ELBO	**8.2616**	7.5651	7.5971	3.8666
	RMSE	**0.3368**	0.3705	0.3630	0.6671
	MMD	**0.0957**	0.1159	0.1133	0.2231
	W_2^2	**0.2331**	0.2855	0.2854	0.2570
UCI-Miniboone	ELBO	**35.7191**	32.3384	31.5001	31.9774
	RMSE	**0.2953**	0.3410	0.3273	0.3224
	MMD	**0.0542**	0.0751	0.0766	0.1032
	W_2^2	**0.0012**	0.0014	0.0024	0.0021

Table 2: Results on UCI POWER and MINIBOONE using four evaluation metrics.

7.5 DISCUSSION AND SUMMARY

Across the three tasks and four metrics, our method achieves stable gains in both *probabilistic consistency* (ELBO/NLL) and *geometric/perceptual consistency* (RMSE, MMD (RBF), W_2^2 (Sinkhorn)). We attribute this to two key factors: (i) coupled particle–Langevin updates for the joint (x, z) and the conditional $z \mid x$, which fit the target Gibbs distribution more directly and reduce bias from restricted variational families or decoder assumptions; and (ii) MAP/Langevin reconstruction along the energy landscape, which preserves multi-modal geometric structure and frequency-related patterns. These observations are consistent with the theoretical motivation and guarantees discussed earlier (particle implementation, Fokker–Planck evolution, and KL/W_2 contraction), and they suggest that the approach scales to more complex data.

8 CONCLUSION

We presented a principled framework for latent-variable energy-based models (LV-EBMs) via coupled Wasserstein gradient flows, casting MLE as a saddle problem over latent and joint distributions. This yields a compact, particle-based learning algorithm that alternates conditional latent updates with joint negative sampling, requiring no auxiliary posterior, discriminator, or decoder, and whose induced Langevin dynamics enjoy existence and stability guarantees, with exponential KL/W_2 contraction under LSIs and a weaker (rate-free) convergence guarantee under standard coercivity/dissipativity. Empirically, across controlled geometric systems and standard UCI density-estimation benchmarks, the method is consistently competitive or superior to VAE, non-amortized VI, and hard-EM baselines, reflecting stronger preservation of multimodal structure and tighter likelihood objectives. Looking ahead, we see opportunities to scale to high-dimensional real data, enrich energy parameterizations, and interface with diffusion/flow-based models, further unifying theory and practice for LV-EBMs.

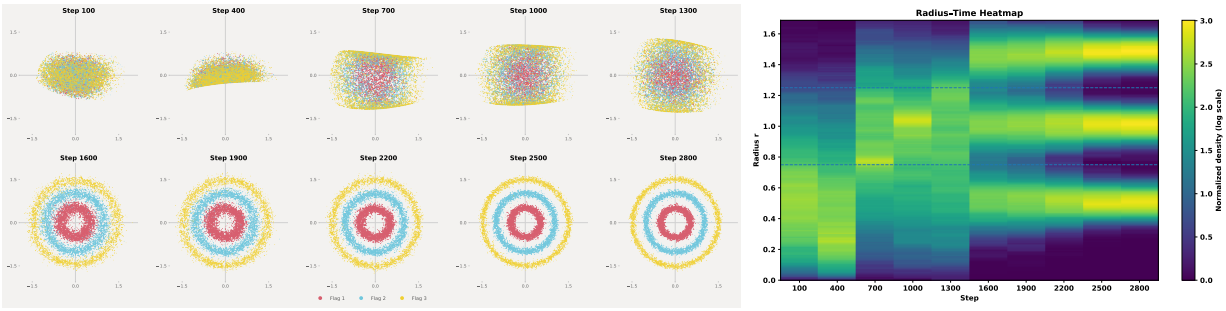

Figure 2: **Left**: Ten snapshots of LCR-2D reconstructions across training steps; Colors denote ground-truth radial classes. **Right**: Radius–step heatmap computed from the same reconstructions. The color intensity encodes the per-step normalized radial density.

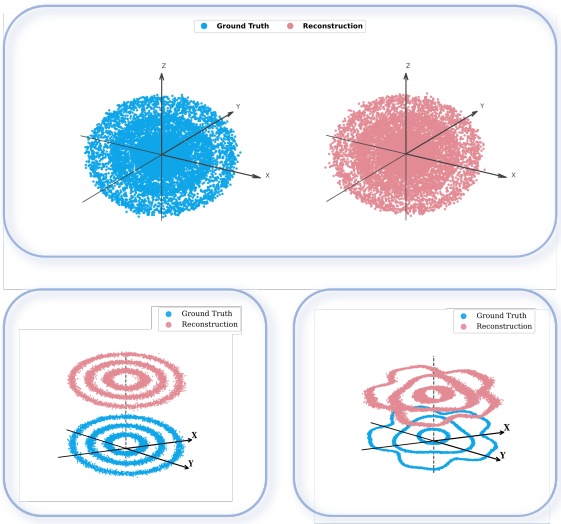

Figure 3: Reconstruction results: LCS-3D (top), LCR-2D (bottom-left), and HMR-2D (bottom-right).

References

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2005.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Birkhäuser, Basel, 2nd edition, 2008a.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2 edition, 2008b.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de probabilités de Strasbourg*, volume 19, pages 177–206. 1985.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 3, 2024.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. *Proceedings of the 35th International Conference on Machine Learning (ICML) Workshop*, 2018. arXiv:1801.03558.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations (ICLR)*, 2017.

Monroe Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time. i. *Comm. Pure Appl. Math.*, 1975.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy-based models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Alain Durmus and Eric Moulines. Nonasymptotic convergence of the unadjusted langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1912.03263.

Will Grathwohl, Jacob Kelly, Milad Hashemi, Mohammad Norouzi, Kevin Swersky, and David Duvenaud. No mcmc for me: Amortized sampling for fast and stable training of energy-based models. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://arxiv.org/abs/2010.04230`.

Leonard Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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, 1998a.

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

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998c. doi: 10.1137/S0036141096303359.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Juan Kuntz, Jen Ning Lim, and Adam M. Johansen. Particle algorithms for maximum likelihood training of latent variable models. *arXiv preprint arXiv:2204.12965*, 2023. doi: 10.48550/arXiv.2204.12965. URL `https://arxiv.org/abs/2204.12965`.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu-Jie Huang. A tutorial on energy-based learning. In *Predicting Structured Data*. 2006. Tutorial overview of EBMs.

Jen Ning Lim and Adam Johansen. Particle semi-implicit variational inference. *Advances in Neural Information Processing Systems*, 37:123954–123990, 2024.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Joanna Marks, Tim YJ Wang, and O Deniz Akyildiz. Learning latent energy-based models via interacting particle langevin dynamics. *arXiv preprint arXiv:2510.12311*, 2025.

Kirill Neklyudov, Priyank Jaini, and Max Welling. Particle dynamics for learning EBMs. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. In *Lectures on the Mathematics of Diffusion*, pages 1–62. Springer, 2001.

Bo Pang, Tian Han, Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. Learning latent space energy-based prior model. *Advances in Neural Information Processing Systems*, 33:21994–22008, 2020.

Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*. Springer, New York, 2014.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*, volume 11. Foundations and Trends in Machine Learning, 2019. doi: 10.1561/2200000073.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

H. Risken. *The Fokker–Planck Equation: Methods of Solution and Applications*. Springer, 2 edition, 1996a.

Hannes Risken. *The Fokker–Planck Equation: Methods of Solution and Applications*. Springer, Berlin, 2nd edition, 1996b.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhäuser, Cham, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 2008.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011.

Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. In *International conference on machine learning*, pages 5660–5669. PMLR, 2018.

Particle Dynamics for Latent-Variable Energy-Based Models (Supplementary Material)

Shiqin Tang[*1] **Shuxin Zhuang**[*1,2] **Runsheng Yu**[3] **Rong Feng**[1,2] **Shujian Yu**[4,5]
Hongzong Li[3] **Mingyang Zhao**[6] **Gaofeng Meng**[†1]

[1]CAIR-CAS, Hong Kong [2] City University of Hong Kong, Hong Kong
[3]The Hong Kong University of Science and Technology, Hong Kong
[4]Vrije Universiteit Amsterdam, Netherlands [5] UiT The Arctic University of Norway, Norway
[6]Academy of Mathematics and Systems Science, Chinese Academy of Sciences

A PROOF FOR PROPOSITION 1

Proof. Define the Radon–Nikodym derivative $\tilde{r} = \frac{d\tilde{q}}{d(\nu\otimes\mu)}$. We have

$$\begin{aligned} D_{\mathrm{KL}}(\tilde{q}\|p_\theta) &= \int \log\Big(\tfrac{\tilde{r}}{dp_\theta/d(\nu\otimes\mu)}\Big)\, d\tilde{q} \\ &= \int \log\tilde{r}\, d\tilde{q} + \int E\, d\tilde{q} + \log Z(\theta) \\ &= -H_{\nu\otimes\mu}(\tilde{q}) + \mathbb{E}_{\tilde{q}}[E] + \log Z(\theta). \end{aligned}$$

The conditional identity is identical with the product base measure replaced by μ and $p_\theta(\cdot \mid x) \propto e^{-E(x,\cdot)}$; $\log p_\theta(x)$ is the (negative) log-partition for the conditional. Uniqueness follows from $D_{\mathrm{KL}}(\cdot\|\cdot) \geq 0$ with equality iff the arguments agree almost surely. □

B PROOF FOR THEOREM 2

Proof. **1. Entropy functional.** Define the relative entropy $\mathcal{H}(t) := D_{\mathrm{KL}}(q_t^x\|p^*(\cdot \mid x))$.

2. Entropy dissipation. Using $\partial_t q_t^x = \nabla_z\cdot(q_t^x\nabla_z E) + \Delta_z q_t^x$ and integrating by parts,

$$\begin{aligned} \frac{d\mathcal{H}}{dt} &= \int \big(\partial_t q_t^x\big)\Big(\log\frac{q_t^x}{p^*} + 1\Big)\, dz \\ &= \int \big[\nabla_z\cdot(q_t^x\nabla_z E) + \Delta_z q_t^x\big]\, \log\frac{q_t^x}{p^*}\, dz \\ &= -\int q_t^x\, \nabla_z E\cdot\nabla_z\log\frac{q_t^x}{p^*}\, dz - \int \nabla_z q_t^x\cdot\nabla_z\log\frac{q_t^x}{p^*}\, dz \\ &= -\int \big(q_t^x\nabla_z E + \nabla_z q_t^x\big)\cdot\nabla_z\log\frac{q_t^x}{p^*}\, dz \\ &= -\int q_t^x\big(\nabla_z E + \nabla_z\log q_t^x\big)\cdot\nabla_z\log\frac{q_t^x}{p^*}\, dz \\ &= -\int q_t^x\big(-\nabla_z\log p^* + \nabla_z\log q_t^x\big)\cdot\nabla_z\log\frac{q_t^x}{p^*}\, dz \\ &= -\int q_t^x\left|\nabla_z\log\frac{q_t^x}{p^*}\right|^2 dz = -I(q_t^x \mid p^*), \end{aligned} \tag{30}$$

where $I(q \mid p) := \int q\,|\nabla\log(q/p)|^2 dz$ is the relative Fisher information. Hence $\mathcal{H}$ is nonincreasing.

3. Integrability of Fisher information. From (30),

$$\mathcal{H}(t) = \mathcal{H}(0) - \int_0^t I(q_s^x \mid p^*)\, ds, \quad \Rightarrow \quad \int_0^\infty I(q_s^x \mid p^*)\, ds \le \mathcal{H}(0) < \infty.$$

4. Second moment and tightness. Let $M_2(t) := \int |z|^2 q_t^x(z)\, dz$. Then

$$\begin{aligned}\frac{dM_2}{dt} &= \int |z|^2\, \partial_t q_t^x\, dz = \int |z|^2 \big[\nabla_z \cdot (q_t^x \nabla_z E) + \Delta_z q_t^x\big]\, dz \\ &= -\int \nabla_z(|z|^2) \cdot (q_t^x \nabla_z E)\, dz \;+\; \int q_t^x\, \Delta_z(|z|^2)\, dz \\ &= -2\int q_t^x \langle z, \nabla_z E\rangle dz \;+\; 2\ell,\end{aligned}$$

since $\nabla_z(|z|^2) = 2z$ and $\Delta_z(|z|^2) = 2\ell$. By dissipativity,

$$\frac{dM_2}{dt} \;\le\; -2m\, M_2(t) + 2b + 2\ell,$$

so by Grönwall,

$$M_2(t) \;\le\; e^{-2mt} M_2(0) + \frac{b+\ell}{m}\big(1 - e^{-2mt}\big) \;\le\; M_2(0) + \frac{b+\ell}{m},$$

which yields $\sup_{t\ge 0} M_2(t) < \infty$ and tightness of $\{q_t^x\}_{t\ge 0}$.

5. Subsequence limits and identification. Tightness gives $t_n \to \infty$ with $q_{t_n}^x \rightharpoonup q_\infty$. Since $\int_0^\infty I(q_s^x \mid p^*)\, ds < \infty$, there exists a (relabeled) subsequence with $I(q_{t_n}^x \mid p^*) \to 0$. Lower semicontinuity of entropy and Fisher information under weak convergence implies

$$D_{\mathrm{KL}}(q_\infty \| p^*) \le \liminf_n D_{\mathrm{KL}}(q_{t_n}^x \| p^*), \qquad I(q_\infty \mid p^*) \le \liminf_n I(q_{t_n}^x \mid p^*) = 0,$$

hence $I(q_\infty \mid p^*) = 0$, which forces $q_\infty = p^*$.

6. Conclusion. Every convergent subsequence has the same limit p^*, so $q_t^x \rightharpoonup p^*$. □

C PROOF FOR THEOREM 3

Proof. The proof applies the Donsker–Varadhan variational identity [Donsker and Varadhan, 1975, Wainwright and Jordan, 2008, Cover and Thomas, 2006] $\log \int e^f\, d\lambda = \sup_{q \ll \lambda}\{\mathbb{E}_q[f] + H(q)\}$ twice:

(i) With $f_x(z) = -E(x, z; \theta)$ and base measure dz to express $\log \int e^{-E(x^i, z;\theta)} dz$ as $\sup_{q\in\mathcal{Q}} \mathcal{A}_i(q)$, attained at $q^\star(\cdot) = p_\theta(\cdot \mid x^i)$.

(ii) With $f(x, z) = -E(x, z; \theta)$ and base measure $dx\, dz$ to express $\log \iint e^{-E(x,z;\theta)} dx\, dz$ as $\sup_{\tilde{q}\in\tilde{\mathcal{Q}}} \mathcal{B}(\tilde{q})$, attained at $\tilde{q}^\star = p_\theta(x, z)$.

Subtracting the second identity from the first and averaging over i yields Part (1). Part (2) follows because taking a supremum over a smaller feasible set cannot increase its value. Strictness holds since the optimizers in Part (1) are (a.e.) unique, which follows from the strict convexity of relative entropy [Cover and Thomas, 2006, Rockafellar, 1970]. □

D EXPERIMENTS ON REAL-WORLD DATA

UCI POWER

Origin & semantics. The POWER dataset records household electrical power consumption over 47 months at one-minute resolution. Although it is a time series, prior density-estimation work treats each record as an i.i.d. sample from the marginal distribution. The commonly used preprocessed version converts time-of-day to minutes, adds small uniform jitter to avoid duplicate values, discards the date field, and removes "global reactive power" due to many exact zeros. The resulting dimensionality is $D = 6$.

Table 3: Reference statistics for the MAF-preprocessed POWER and MINIBOONE datasets (as distributed), plus our experimental subsampling protocol.

Dataset	Dim.	Train	Valid	Test	Provenance
POWER	6	1,659,917	184,435	204,928	MAF bundle[4]
MINIBOONE	43	29,556	3,284	3,648	MAF bundle[4]
Our protocol (for experiments): train 20k, test 2k (fixed seed).					

UCI MINIBOONE

Origin & semantics. From the MiniBooNE neutrino oscillation experiment, originally a classification task. The preprocessed density-estimation version uses only the positive class (electron neutrinos), removes 11 obvious outliers with -1000 in every column, and drops seven features with extreme value counts. The resulting dimensionality is $D = 43$.

Provenance of the preprocessed data. We use the exact public bundle released for the Masked Autoregressive Flow (MAF) experiments (Zenodo record #1161203), which standardizes the community setup across POWER, GAS, and MINIBOONE.[1]

HOW WE USE THE DATASETS (TRAINING/EVALUATION PROTOCOL)

For comparability across methods and compute budgets, we randomly sample (with a fixed seed) $20{,}000$ examples from the train pool for training and $2{,}000$ examples from the test split for evaluation.

E EXPERIMENT DETAILS

In this section, we provide more details about the experimental configurations and the training procedure used for our Latent-Variable Energy-Based Model (LV-EBM). We evaluate our method on both synthetic datasets (LCR-2D, LCS-3D, HMR-2D) and real-world UCI benchmarks (POWER, MINIBOONE).

ADDITIONAL PARTICLE-BASED COMPARISONS

Table 4 reports additional comparisons with particle-based or decoder-based references discussed in the rebuttal. The Marks-style model is a decoder-based latent EBM reference rather than a same-class baseline, since it includes an explicit decoder. The Kuntz-style posterior-only and Joint-CD variants are ablations within the joint-energy setting.

Table 5 summarizes the hyperparameters used in each dataset. We maintain a consistent configuration for most parameters, while tuning the number of training steps and Langevin steps to accommodate the varying complexity of the data distributions.

[1] Zenodo record #1161203: `https://zenodo.org/record/1161203`

Table 4: Additional particle-based comparisons and ablations on synthetic datasets.

Dataset	Method	RMSE	MMD	W_2^2
LCR-2D	Ours	0.1612	0.1549	0.2161
LCR-2D	Marks-style	0.0603	0.0144	0.0389
LCR-2D	Kuntz-style posterior-only	2.5331	0.6116	8.2686
LCR-2D	Joint-CD with one posterior particle	0.9333	0.3785	0.6085
LCS-3D	Ours	0.1303	0.0129	0.2567
LCS-3D	Marks-style	0.1264	0.0435	0.0941
LCS-3D	Kuntz-style posterior-only	2.5704	0.6176	14.1360
LCS-3D	Joint-CD with one posterior particle	0.6479	0.4282	0.6331
HMR-2D	Ours	0.0375	0.0012	0.1663
HMR-2D	Marks-style	0.0622	0.0048	0.0374
HMR-2D	Kuntz-style posterior-only	2.5576	0.5763	8.0292
HMR-2D	Joint-CD with one posterior particle	0.8174	0.4012	0.5215

Hyperparameter	**Notation**	**LCR-2D**	**LCS-3D**	**HMR-2D**	**UCI**
Joint buffer size	N	4096	4096	4096	4096
Latent per data point	M	16	16	16	16
Batch size	B	512	512	512	512
Training steps	T	4000	5000	7000	6000
Langevin steps	K	30	25	20	25
Step size (joint)	η_x	10^{-3}	8×10^{-4}	10^{-3}	10^{-3}
Step size (latent)	η_z	10^{-3}	8×10^{-4}	10^{-3}	10^{-3}
Refresh prob. (joint)	p_{joint}	0.02	0.02	0.02	0.02
Refresh prob. (bank)	p_{bank}	0.10	0.10	0.10	0.10
Noise scale (joint)	ξ_x	1.0	1.0	1.0	1.0
Noise scale (latent)	ξ_z	1.0	1.0	1.0	1.0

Table 5: Hyperparameter settings for LV-EBM experiments. The table lists the configurations used for LCR-2D, LCS-3D, HMR-2D, and UCI datasets.