Learning Energy-Based Generative Models via Potential Flow: A Variational Principle Approach to Probability Density Homotopy Matching

Anonymous authors
Paper under double-blind review

Abstract

Energy-based models (EBMs) are a powerful class of probabilistic generative models due to their flexibility and interpretability. However, relationships between potential flows and explicit EBMs remain underexplored, while contrastive divergence training via implicit Markov chain Monte Carlo (MCMC) sampling is often unstable and expensive in high-dimensional settings. In this paper, we propose Variational Potential Flow Bayes (VPFB), a new energy-based generative framework that eliminates the need for implicit MCMC sampling and does not rely on auxiliary networks or cooperative training. VPFB learns an energy-parameterized potential flow by constructing a flow-driven density homotopy that is matched to the data distribution through a variational loss minimizing the Kullback-Leibler divergence between the flow-driven and marginal homotopies. This principled formulation enables robust and efficient generative modeling while preserving the interpretability of EBMs. Experimental results on image generation, interpolation, out-of-distribution detection, and compositional generation confirm the effectiveness of VPFB, showing that our method performs competitively with existing approaches in terms of sample quality and versatility across diverse generative modeling tasks.

1 Introduction

Energy-based models (EBMs) have emerged as a flexible and expressive class of probabilistic generative models (Nijkamp et al., 2019; Du & Mordatch, 2019; Grathwohl et al., 2020b; Gao et al., 2020; Du et al., 2021; Gao et al., 2021; Grathwohl et al., 2020a; Yang et al., 2023; Zhu et al., 2024). By assigning a potential energy that correlates with the unnormalized data likelihood (Song & Kingma, 2021), EBMs offer a structured energy landscape for probability density estimation, providing several notable advantages. First, EBMs are interpretable, as the underlying energy function can be visualized in terms of energy surfaces. Second, they are highly expressive and do not impose strong architectural constraints (Bond-Taylor et al., 2022), enabling them to capture complex data distributions. Third, EBMs exhibit inherent robustness to Out-of-Distribution (OOD) inputs, given that regions with low likelihood are naturally penalized (Du & Mordatch, 2019; Grathwohl et al., 2020a). Building on their origins in Boltzmann machines (Hinton, 2002), EBMs also share conceptual ties with statistical physics, allowing practitioners to adapt physical insights and tools for model design and analysis (Feinauer & Lucibello, 2021). They have demonstrated promising performance in various applications beyond image modeling, including text generation (Deng et al., 2020), robot learning (Du et al., 2020), point cloud synthesis (Xie et al., 2021a), trajectory prediction (Pang et al., 2021; Wang et al., 2023), molecular design (Liu et al., 2021), and anomaly detection (Yoon et al., 2023).

Despite these advantages, training deep EBMs often relies on implicit Markov Chain Monte Carlo (MCMC) sampling for contrastive divergence. In high-dimensional settings, MCMC suffers from poor mode mixing and slow mixing (Du & Mordatch, 2019; Nijkamp et al., 2019; Gao et al., 2020; Grathwohl et al., 2020a; Nijkamp et al., 2022; Bond-Taylor et al., 2022), yielding biased estimates that may optimize unintended objectives (Grathwohl et al., 2020b). Truncated chains, in particular, can lead models to learn an implicit sampler rather than a true density, which prevents valid steady-state convergence and inflates computational overhead. As a result, the generated samples can deviate significantly from the target distribution (Grathwohl et al., 2020b). To mitigate these issues, some works propose auxiliary or cooperative strategies that learn complementary models to either avoid MCMC via variational inference (Xiao et al., 2021) or combine short-run MCMC refinements with learned generator distributions (Xie et al., 2020; Grathwohl et al., 2021; Hill et al., 2022). However, these approaches can complicate model architectures and training procedures. In parallel, flow-based models have advanced generative modeling by leveraging continuous normalizing flows and optimal transport techniques to surpass diffusion models in sample quality and efficiency Kim et al. (2021); Song et al. (2021). Notable examples include Flow Matching Lipman et al. (2023), which models diffeomorphic mappings between noise and data; Rectified Flow Liu et al. (2023b), which optimizes sampling paths; Stochastic Interpolants Albergo & Vanden-Eijnden (2023); Rezende & Mohamed (2015); Chen et al. (2018), which incorporate stochastic processes into flows for complex data geometries; Schrödinger Bridge Matching Shi et al. (2023), which integrates entropy-regularized optimal transport with diffusion; and Poisson Flow Generative Model (PFGM) Xu et al. (2022), which introduces an augmented space governed by the Poisson equation. However, these methods do not directly parameterize probability density and lack the theoretical advantages of EBMs, such as generating conservative vector fields aligned with log-likelihood gradients Salimans & Ho (2021). Recent approaches, like Action Matching Neklyudov et al. (2023), explicitly model the energy (action) to generate data-recovery vector fields, providing a structured approach to learning conservative dynamics. Meanwhile, Diffusion Recovery Likelihood (DRL) Gao et al. (2021) and Denoising Diffusion Adversarial EBMs (DDAEBM) Geng et al. (2024) refine conditional EBMs by improving sampling efficiency and training stability through diffusion-based probability paths. However, a direct connection between energy-parameterized flow models and explicit (marginal) EBMs remains unexplored, limiting the application of flow-based techniques for learning EBMs. Furthermore, existing generative models have yet to adopt variational formulations, such as the Deep Ritz approach, to align the evolution of density paths.

To address the computational challenges of existing energy-based methods, we propose Variational Potential Flow Bayes (VPFB), a novel generative framework grounded in variational principles that eliminates the need for auxiliary models and implicit MCMC sampling. VPFB employs the Deep Ritz method to learn an energy-parameterized potential flow, ensuring alignment between the flow-driven density homotopy and the data-recovery likelihood homotopy. To address the intractability of homotopy matching, we formulate a variational loss function that minimizes the Kullback-Leibler (KL) divergence between these density homotopies. Additionally, we validate the learned potential energy as an effective parameterization of the stationary Boltzmann energy. Through empirical validations, we benchmark VPFB against state-of-the-art generative models, showcasing its competitive performance in Fréchet Inception Distance (FID) for image generation and excellent OOD detection with high Area Under the Receiver Operating Characteristic Curve (AUROC) scores across multiple datasets.

2 Background and Related Works

In this section, we provide an overview of EBMs, particle flow, and the Deep Ritz approach, collectively forming the cornerstone of the proposed VPFB framework.

2.1 Energy-based Models (EBMs)

Denote $\bar{x} \in \Omega \subseteq \mathbb{R}^n$ as the training data, EBMs approximate the data likelihood $p_{\text{data}}(\bar{x})$ via defining a Boltzmann distribution

$$p_B(x) = \frac{e^{\Phi_B(x)}}{Z} \tag{1}$$

where Φ_B is the Boltzmann energy parameterized via neural networks and $Z = \int_\Omega e^{\Phi_B(x)}\, dx$ is the normalizing constant. Given that this partition function is analytically intractable for high-dimensional data, EBMs perform the Maximum Likelihood Estimation (MLE) by minimizing the negative log-likelihood loss $\mathcal{L}_{\text{MLE}}(\theta) = -\mathbb{E}_{p_{\text{data}}(\bar{x})}[\log p_B(\bar{x})] = \mathbb{E}_{p_{\text{data}}(\bar{x})}\big[\Phi_B(\bar{x})\big] - \mathbb{E}_{p_{\text{data}}(\bar{x})}\big[\log Z\big]$. The gradient of this MLE loss with respect to model parameters θ is approximated via the contrastive divergence (Hinton, 2002) loss $\nabla_\theta \mathcal{L}_{\text{MLE}} = \mathbb{E}_{p_{\text{data}}(\bar{x})}\big[\nabla_\theta \Phi_B(\bar{x})\big] - \mathbb{E}_{p_B(x)}\big[\nabla_\theta \Phi_B(x)\big]$. Nonetheless, EBMs are computationally intensive due to the implicit MCMC generating procedure, required for generating negative samples $x \sim p_B(x)$ implicitly during training.

2.2 Particle Flow

Particle flow, introduced by Daum & Huang (2007), is a class of nonlinear Bayesian filtering (sequential inference) methods designed to approximate the posterior distribution $p(x_t|\bar{x}_{\leq t})$ of the sampling process given observations. While closely related to normalizing flows (Rezende & Mohamed, 2015) and neural ordinary differential equations (ODEs) (Chen et al., 2018), these frameworks do not explicitly accommodate a Bayes update. Instead, particle flow achieves Bayes update $p(x_t|\bar{x}_{\leq t}) \propto p(x_t|\bar{x}_{<t})\, p(\bar{x}_t|x_t, \bar{x}_{<t})$ by transporting the prior samples $x_t \sim p(x_t|\bar{x}_{<t})$ through an ODE $\frac{dx}{dt} = v(x,t)$ parameterized by a velocity field $v(x,t)$, over pseudo-time $t \in [0,1]$. The velocity field is designed such that the sample density follows a log-homotopy that induces the Bayes update. Despite its effectiveness in time-series inference (Pal et al., 2021b; Chen et al., 2019b; Yang et al., 2014) and its robustness against the curse of dimensionality (Surace et al., 2019), particle flow, particularly potential flow where the velocity field $v(x,t) = \Phi(x,t)$ is the gradient of potential energy, remains largely unexplored in energy-based generative modeling.

2.3 Deep Ritz Approach

The Deep Ritz approach is a deep learning-based variational numerical approach, originally proposed by E & Yu (2018), for solving scalar elliptic partial differential equations (PDEs) in high dimensions. Consider the following Poisson's equation, fundamental to many physical models:

$$\Delta u(x) = \Gamma(x),\, x \in \Omega, \quad u(x) = 0,\ x \in \partial\Omega \tag{2}$$

where Δ is the Laplace operator, and $\partial\Omega$ denotes the boundary of Ω. For a Sobolev function $u \in \mathcal{H}_0^1(\Omega)$ (definition in Proposition 2) and square-integrable $\Gamma \in L^2(\Omega)$, the variational principle ensures that a weak solution of the Euler-Lagrange boundary value equation (2) is equivalent to the variational problem of minimizing the Dirichlet energy (Müller & Zeinhofer, 2019), as follows:

$$u = \arg\min_{v} \int_{\Omega} \left(\frac{1}{2} \|\nabla_x v(x)\|^2 - \Gamma(x)\, v(x) \right) dx \tag{3}$$

where ∇_x denotes the Del operator (gradient). In particular, the Deep Ritz approach parameterizes the trial function v using neural networks and performs the optimization (3) via stochastic gradient descent. Due to its versatility and effectiveness in handling high-dimensional PDE systems, the Deep Ritz approach is predominantly applied for finite element analysis (Liu et al., 2023a). In (Olmez et al., 2020), the Deep Ritz approach is employed to solve the density-weighted Poisson equation arising from the feedback particle filter (Yang et al., 2013). However, its application in generative modeling remains unexplored.

3 Variational Potential Flow Bayes (VPFB)

In this section, we introduce VPFB, a novel generative modeling framework inspired by particle flow and the Deep Ritz approach. VPFB encompasses four key elements: constructing a Bayesian marginal homotopy between the Gaussian prior and data likelihood (Section 3.1), designing a potential flow that aligns the flow-driven homotopy with the marginal homotopy (Section 3.2), formulating a variational loss function using the Deep Ritz approach (Section 3.4), and establishing connections between homotopy matching, diffusion, and EBMs (Section 3.3).

3.1 Interpolating Between Prior and Data Likelihood: Log-Homotopy Bayesian Transport

Let $\bar{x} \in \Omega$ denote the training data, with likelihood $p_{\text{data}}(\bar{x})$, and let $x \in \Omega$ represent the generative samples. First, we define a Gaussian prior $q(x) = \mathcal{N}(0, \omega^2 I)$ and a Gaussian conditional data likelihood $p(\bar{x} \mid x) = \mathcal{N}(\bar{x}; x, \nu^2 I)$, both with isotropic covariances. This data likelihood satisfies the state space model $x = \bar{x} + \nu\, \epsilon$, where ϵ is the standard Gaussian noise. The standard deviation ν is usually set to be small so that x closely resembles $\bar{x}$. The aim of flow-based generative modeling is to learn a density homotopy (path) interpolating between the prior and the data likelihood for generative modeling. On that account, consider the following conditional (data-conditioned) probability density log-homotopy $\rho : \Omega^2 \times [0,1] \to \mathbb{R}$:

$$\rho(x \mid \bar{x}, t) = \frac{e^{h(x|\bar{x},t)}}{\int_\Omega e^{h(x|\bar{x},t)}\, dx} \tag{4}$$

where $f : \Omega^2 \times [0,1] \to \mathbb{R}$ is a log-linear function:

$$f(x \mid \bar{x}, t) = \alpha(t)\, \log q(x) + \beta(t)\, \log p(\bar{x} \mid x) \tag{5}$$

where $\alpha : [0,1] \to [0,1]$ and $\beta : [0,1] \to [0,1]$ are both monotonically increasing functions parameterized by time t. The following proposition shows that this log-homotopy transformation results in a Gaussian perturbation kernel.

Proposition 1. *Consider a Gaussian prior $q(x) = \mathcal{N}(x; 0, \omega^2 I)$ and a conditional data likelihood $p(\bar{x} \mid x) = \mathcal{N}(\bar{x}; x, \nu^2 I)$. The log-homotopy transport (4) corresponds to a Gaussian perturbation kernel $\rho(x \mid \bar{x}, t) = \mathcal{N}\big(x; \mu(t)\, \bar{x}, \sigma(t)^2 I\big)$, characterized by the time-varying mean and standard deviation:*

$$\mu(t) = \text{sigmoid}\left(\log\left(\frac{\beta(t)}{\alpha(t)}\, \frac{\omega^2}{\nu^2}\right)\right), \quad \sigma(t) = \sqrt{\frac{\nu^2}{\beta(t)}\, \mu(t)} \tag{6}$$

where $\text{sigmoid}(z) = \frac{1}{1+e^{-z}}$ *denotes the logistic (sigmoid) function.*

Proof. Refer to Appendix A.1. □

Hence, the density homotopy equation 4 represents a tempered Bayesian transport mapping from the Gaussian prior $q(x)$ to the posterior kernel

$$\rho(x \mid \bar{x}, 1) = \frac{e^{h(x|\bar{x},1)}}{\int_\Omega e^{h(x|\bar{x},1)}\, dx} = \frac{p(\bar{x} \mid x)\, q(x)}{\int_\Omega p(\bar{x} \mid x)\, q(x)\, dx} = p(x \mid \bar{x}) \tag{7}$$

which is the maximum a posteriori estimation centered on discrete data samples. To approximate the intractable data likelihood, we can then consider the following marginal probability density homotopy:

$$\bar{\rho}(x, t) = \int_\Omega p_{\text{data}}(\bar{x})\, \rho(x \mid \bar{x}, t)\, d\bar{x}, \tag{8}$$

where it remains that $p(x, 0) = q(x)$, and we have $\bar{\rho}(x, 1) = \int_\Omega p_{\text{data}}(\bar{x})\, p(x \mid \bar{x})\, d\bar{x} = \bar{\rho}(x)$. Therefore, this marginal homotopy defines a data-recovery path interpolation between the Gaussian prior $q(x)$ and the approximate data likelihood $\bar{\rho}(x)$. In particular, $\bar{\rho}(x)$ represents a Bayesian approximation of the true data likelihood, by constructing a kernel density estimation of the discrete data likelihood $p_{\text{data}}(\bar{x})$ with perturbation kernel $\rho(x \mid \bar{x}, t)$. Nevertheless, the marginalization in (8) is intractable, thereby precluding a closed-form solution for the marginal homotopy. To overcome this challenge, we propose a potential flow-driven density homotopy, whose time evolution is aligned with this data-recovery marginal homotopy.

3.2 Modeling Potential Flow in a Data-Recovery Homotopy Landscape

Our goal is to model a potential flow whose density evolution aligns with the marginal homotopy, thereby directing samples toward the data likelihood. We begin by deriving the time evolution of the marginal homotopy in the following proposition.

Proposition 2. *Consider the conditional homotopy $\rho(x \mid \bar{x}, t)$ in (4) with Gaussian conditional data likelihood $p(\bar{x} \mid x) = \mathcal{N}(\bar{x}; x, \nu^2 I)$. Then, the time evolution (derivative) of the marginal homotopy $\bar{\rho}(x, t)$ is given by the following partial differential equation (PDE):*

$$\frac{\partial \bar{\rho}(x,t)}{\partial t} = -\frac{1}{2}\, \mathbb{E}_{p_{data}(\bar{x})} \left[\rho(x \mid \bar{x}, t) \left(\gamma(x, \bar{x}, t) - \bar{\gamma}(x, \bar{x}, t) \right) \right] \tag{9}$$

where γ denotes the innovation function

$$\gamma(x, \bar{x}, t) = \frac{\dot{\alpha}(t)}{\omega^2} \|x\|^2 + \frac{\dot{\beta}(t)}{\nu^2} \|x - \bar{x}\|^2 \tag{10}$$

Here, $\dot{\alpha}(t)$ and $\dot{\beta}(t)$ denote the time-derivatives, and $\bar{\gamma}(x, \bar{x}, t) = \mathbb{E}_{\rho(x|\bar{x},t)}[\gamma(x, \bar{x}, t)]$ denotes the expectation.

Proof. Refer to Appendix A.2. □

A potential flow involves subjecting the prior samples to an energy-generated velocity field, where their trajectories $(x(t))$ satisfy the following ODE:

$$\frac{dx(t)}{dt} = \nabla_x \Phi(x, t) \tag{11}$$

where $\Phi : \Omega \times [0, 1] \to \mathbb{R}$ is a scalar potential energy, and ∇_x denotes the Del operator (gradient) with respect to the data samples $x(t)$. The vector field $\nabla_x \Phi \in \Omega$ represents the divergence (irrotational) component in the Helmholtz decomposition. By incorporating this potential flow, the flow-driven density homotopy $\rho_\Phi(x, t)$ evolves via the continuity equation (Gardiner, 2009):

$$\frac{\partial \rho_\Phi(x,t)}{\partial t} = -\nabla_x \cdot \left(\rho_\Phi(x, t)\, \nabla_x \Phi(x, t) \right) \tag{12}$$

which corresponds to the transport equation for modeling fluid advection. Our aim is to model the potential energy such that the evolution of the prior density under the potential flow emulates the evolution of the marginal homotopy. In other words, we seek to achieve homotopy matching, $\rho_\Phi \equiv \bar{\rho}$, by aligning their respective time evolutions as described in (9) and (12). This leads to the following PDE, which takes the form of a density-weighted Poisson equation:

$$\nabla_x \cdot \left(\rho_\Phi(x, t)\, \nabla_x \Phi(x, t) \right) = \frac{1}{2}\, \mathbb{E}_{p_{\text{data}}(\bar{x})} \left[\rho(x \mid \bar{x}, t) \left(\gamma(x, \bar{x}, t) - \bar{\gamma}(x, \bar{x}, t) \right) \right] \tag{13}$$

However, this Poisson equation remains intractable due to the lack of a closed-form expression for ρ_Φ. To overcome this limitation, we substitute the intractable ρ_Φ with the target marginal homotopy $\bar{\rho}$, enabling direct sampling and a variational principle approach. In the following proposition, we demonstrate that the revised Poisson's equation minimizes the KL divergence between the flow-driven and conditional homotopies, yielding statistically optimal homotopy matching.

Proposition 3. *Consider a potential flow of the form (11) and given that $\Phi \in \mathcal{H}_0^1(\Omega, p)$, where $\mathcal{H}_0^n$ denotes the (Sobolev) space of n-times differentiable functions that are compactly supported, and square-integrable with respect to marginal homotopy $\bar{\rho}(x, t)$. Solving for the potential energy $\Phi(x)$ that satisfies the following density-weighted Poisson's equation:*

$$\nabla_x \cdot \left(\bar{\rho}(x, t)\, \nabla_x \Phi(x, t) \right) = \frac{1}{2}\, \mathbb{E}_{p_{data}(\bar{x})} \left[\rho(x \mid \bar{x}, t) \left(\gamma(x, \bar{x}, t) - \bar{\gamma}(x, \bar{x}, t) \right) \right] \tag{14}$$

is then equivalent to minimizing the KL divergence $\mathcal{D}_{\text{KL}} \left[\rho_\Phi(x, t) \| \bar{\rho}(x, t) \right]$ between the flow-driven homotopy and the conditional homotopy.

Proof. Refer to Appendix A.3. □

Therefore, solving this density-weighted Poisson's equation corresponds to performing a homotopy matching $\rho_\Phi \equiv \bar{\rho}$. In the following section, we demonstrate that this homotopy matching gives rise to a Boltzmann energy expressed in terms of the potential energy Φ when the marginal homotopy $\bar{\rho}$ reaches its stationary equilibrium, thereby establishing a connection between our proposed potential flow framework and EBMs.

3.3 Connections to Diffusion Process and Energy-Based Modeling

In this section, we clarify the relationship between diffusion models and flow matching within the homotopy matching framework. Based on this insight, we establish a link between our proposed potential flow framework and energy-based modeling.

First, we present results from diffusion models. It has been outlined in (Song et al., 2021) that the conditional density homotopy, represented by the Gaussian perturbation kernel $\rho(x \mid \bar{x}, t) = \mathcal{N}\big(x; \mu(t)\, \bar{x}, \sigma(t)^2 I\big)$, characterizes a diffusion process governed by the following stochastic differential equation (SDE):

$$dx(t) = -f(t)\, x(t)\, dt + g(t)\, dW(t) \tag{15}$$

where $W(t) \in \mathbb{R}^n$ denote the standard Wiener process. Note that the time parameterization with respect to t here is the reverse of the conventional parameterization used in diffusion models, where the diffusion process transitions from $x(1) \sim p_{\text{data}}(\bar{x})$ to $x(0) \sim q(x) = \mathcal{N}(0, \omega^2 I)$ as defined in Section 3.1. In addition, the time-varying drift $f : [0,1] \to \mathbb{R}$ and diffusion $g : [0,1] \to \mathbb{R}$ coefficients are shown (Karras et al., 2022) to be given by

$$f(t) = -\frac{\dot{\mu}(t)}{\mu(t)}, \quad g(t) = -\sqrt{2\, \sigma(t) \left(\dot{\sigma}(t) + f(t)\, \sigma(t)\right)} \tag{16}$$

where $\dot{\mu}(t)$ and $\dot{\sigma}(t)$ denote the time-derivatives. It has also been shown in (Song et al., 2021) that the following deterministic probability flow ODE:

$$\frac{dx(t)}{dt} = -f(t)\, x(t) + \frac{1}{2}\, g(t)^2\, \nabla_x \log \bar{\rho}(x, t) \tag{17}$$

results in the same marginal probability homotopy $\bar{\rho}(x, t)$ as the forward-time diffusion SDE (16). Subsequently, we highlight the link between the diffusion process and the vector field modeled in flow matching.

Proposition 4. *The conditional vector field in flow matching (Lipman et al., 2023), given by*

$$\frac{dx(t)}{dt} = v(x \mid \bar{x}, t) = \dot{\mu}(t)\, \bar{x} + \dot{\sigma}(t)\, \epsilon \tag{18}$$

with standard Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, satisfies the conditional probability flow ODE governing the diffusion process conditioned on boundary condition $x(1) \sim p_{data}(\bar{x})$. It follows that the marginal vector field, given by the law of iterated expectation (tower property) $\mathbb{E}[U|X = x] = \mathbb{E}[\mathbb{E}\big[U|X = x, Y\big]\,|X = x]$:

$$\frac{dx(t)}{dt} = v(x, t) = \mathbb{E}_{p_{data}(\bar{x}|x)}\big[v(x \mid \bar{x}, t) \mid x\big] = \int_{\Omega} v(x \mid \bar{x}, t)\, \frac{\rho(x \mid \bar{x}, t)\, p_{data}(\bar{x})}{\bar{\rho}(x, t)}\, d\bar{x} \tag{19}$$

also satisfies the marginal probability flow ODE (17).

Proof. Refer to Appendix A.4. □

Building on this result, we establish a connection between the proposed potential flow framework and EBMs. The following proposition demonstrates that homotopy matching, e.g., $\rho_\Phi \equiv \bar{\rho}$ leads to an energy-parameterized Boltzmann equilibrium.

Proposition 5. *Given that the flow-driven homotopy ρ_Φ matches the data-recovery marginal homotopy $\bar{\rho}$, their dynamics satisfy a Fokker-Planck equation. As the time-varying marginal density $\bar{\rho}(x,t)$ converges to its stationary equilibrium $\bar{\rho}_\infty(x)$ at $t = t_\infty$, this Fokker-Planck dynamics attain the Boltzmann distribution (1), where the Boltzmann energy $\Phi_B(x)$ is defined in terms of the steady-state potential energy $\Phi(x)$ as follows:*

$$\Phi_B(x) = \frac{4\, \Phi_\infty(x) + f_\infty\, \|x\|^2}{g_\infty^2} \tag{20}$$

where $\Phi_\infty(x)$, f_∞, and g_∞ are the corresponding steady states of the potential energy and the coefficients at stationary equilibrium.

Proof. Refer to Appendix A.5. □

On that note, we uncover the connection between the proposed VPFB framework and EBMs, demonstrating the validity of the potential energy as a parameterization of a Boltzmann energy. This holds provided that $\bar{\rho}$ converges to its stationary equilibrium and ρ_Φ learns to match these convergent dynamics. In the following section, we introduce a variational principle approach to solving the density-weighted Poisson equation (14), thereby addressing the intractable homotopy matching problem.

3.4 Variational Potential Energy Loss Formulation: Deep Ritz Approach

Solving the density-weighted Poisson's equation (14) is particularly challenging in high-dimensional settings. Numerical approximation struggles to scale with higher dimensionality, as selecting suitable basis functions, such as in the Galerkin approximation, becomes increasingly complex (Yang et al., 2016). Similarly, a diffusion map-based algorithm demands an exponentially growing number of particles to ensure error convergence (Taghvaei et al., 2020). To address these challenges, we propose a variational loss function using the Deep Ritz approach. This approach casts Poisson's equation as a variational problem compatible with stochastic gradient descent. Consequently, the proposed approach solves equation (14), effectively aligning the flow-driven homotopy with the marginal homotopy. Directly solving Poisson's equation (14) is challenging. Therefore, we first consider the following weak formulation:

$$\int_\Omega \frac{1}{2}\, \mathbb{E}_{p_{\text{data}}(\bar{x})} \Big[\rho(x \mid \bar{x}, t) \left(\gamma(x, \bar{x}, t) - \bar{\gamma}(x, \bar{x}, t)\right)\Big] \Psi \, dx = \int_\Omega \nabla_x \cdot \Big(\bar{\rho}(x, t)\, \nabla_x \Phi(x, t)\Big) \Psi \, dx \tag{21}$$

This PDE must hold for all differentiable trial functions Ψ. In the following proposition, we introduce a variational loss function that is equivalent to solving this weak formulation of the density-weighted Poisson's equation.

Proposition 6. *The variational problem of minimizing the following loss function:*

$$\mathcal{L}(\Phi, t) = \operatorname{Cov}_{\rho(x|\bar{x},t)\, p_{data}(\bar{x})} \Big[\Phi(x, t), \gamma(x, \bar{x}, t)\Big] + \mathbb{E}_{\bar{\rho}(x,t)} \Big[\big\|\nabla_x \Phi(x, t)\big\|^2\Big] \tag{22}$$

with respect to the potential energy Φ, is equivalent to solving the weak formulation (21) of the density-weighted Poisson's equation (14). Here, $\|\cdot\|$ denotes the Euclidean norm, and Cov *denotes the covariance. Furthermore, the variational problem (22) admits a unique solution $\Phi \in \mathcal{H}_0^1(\Omega; p)$ if the marginal homotopy p satisfy the Poincaré inequality:*

$$\mathbb{E}_{\bar{\rho}(x,t)} \Big[\big\|\nabla_x \Phi(x, t)\big\|^2\Big] \geq \eta\, \mathbb{E}_{\bar{\rho}(x,t)} \Big[\big\|\Phi(x, t)\big\|^2\Big] \tag{23}$$

for some positive scalar constant $\eta > 0$ (spectral gap).

Proof. Refer to Appendix A.6. □

Therefore, Propositions 2 and 3 reformulate the intractable task of minimizing the KL divergence between the flow-driven homotopy and the marginal homotopy as an equivalent variational problem of solving the loss function (22). By optimizing the potential energy with respect to this loss and transporting the prior samples through the ODE (11), the prior particles evolve along a trajectory that aligns with the marginal homotopy. In particular, the covariance loss here plays an important role by ensuring that the normalized innovation (residual sum of squares) is inversely proportional to the potential energy. As a result, the energy-generated velocity field $\nabla_x \Phi$ consistently points in the direction of greatest potential ascent, thereby driving the flow of prior particles towards high likelihood regions. Given that homotopy matching is performed over the entire time horizon, we apply stochastic integration to the loss function over time, where $t \sim \mathcal{U}(0, t_{\text{end}})$ is drawn from the uniform distribution.

Table 1: Comparison of FID scores on unconditional CIFAR-10 image generation.

Energy-based Models	**FID ↓**	**Other Likelihood-based Models**	**FID ↓**
EBM-IG (Du & Mordatch, 2019)	38.2	ResidualFlow (Chen et al., 2019a)	47.4
EBM-FCE (Gao et al., 2020)	37.3	Glow (Kingma & Dhariwal, 2018)	46.0
CoopVAEBM (Xie et al., 2021b)	36.2	DC-VAE (Parmar et al., 2021)	17.9
CoopNets (Xie et al., 2020)	33.6	**GAN-based Models**	
Divergence Triangle (Han et al., 2019)	30.1	SN-GAN (Miyato et al., 2018)	21.7
VERA (Grathwohl et al., 2021)	27.5	SNGAN-DDLS Che et al. (2020)	15.4
EBM-CD (Du et al., 2021)	25.1	BigGAN (Brock et al., 2019)	14.8
GEBM (Arbel et al., 2021)	19.3	**Score-based and Diffusion Models**	
HAT-EBM (Hill et al., 2022)	19.3	NCSN-v2 (Song & Ermon, 2020)	10.9
CF-EBM (Zhao et al., 2020)	16.7	DDPM Distil (Luhman et al., 2021)	9.36
CoopFlow (Xie et al., 2022)	15.8	DDPM (Ho et al., 2020)	3.17
VAEBM (Xiao et al., 2021)	12.2	NCSN++ (Song et al., 2021)	2.20
DRL (Gao et al., 2021)	9.58	**Flow-based Models**	
CLEL (Lee et al., 2023)	8.61	Action Matching (Neklyudov et al., 2023)	10.0
DDAEBM (Geng et al., 2024)	4.82	Flow Matching (Lipman et al., 2023)	6.35
CDRL (Zhu et al., 2024)	3.68	Rectified Flow (Liu et al., 2023b)	4.85
VPFB (Autonomous)	**15.4**	DSBM (Shi et al., 2023)	4.51
VPFB (Time-varying)	**7.08**	PFGM (Xu et al., 2022)	2.35

Table 2: Comparison of AUROC scores ↑ for OOD detection on several datasets.

Models	**CIFAR-10 Interpolation**	**CIFAR-100**	**CelebA**	**SVHN**
PixelCNN (Salimans et al., 2017)	0.71	0.63	-	0.32
GLOW (Kingma & Dhariwal, 2018)	0.51	0.55	0.57	0.24
NVAE (Vahdat & Kautz, 2020)	0.64	0.56	0.68	0.42
EBM-IG (Du & Mordatch, 2019)	0.70	0.50	0.70	0.63
VAEBM (Xiao et al., 2021)	0.70	0.62	0.77	0.83
CLEL (Lee et al., 2023)	0.72	0.72	0.77	0.98
DRL (Gao et al., 2021)	-	0.44	0.64	0.88
CDRL (Zhu et al., 2024)	0.75	0.78	0.84	0.82
VPFB (Ours)	0.78	0.67	0.84	0.61

3.5 Training Implementation

During training, we implement the Optimal Transport Flow Matching (OT-FM) (Lipman et al., 2023), which corresponds to the SDE parameterization given by (Kingma & Gao, 2023): $f(t) = -\frac{1}{t}, \quad g(t) = \sqrt{\frac{2\,(1-t)}{t}}$ or equivalently, $\alpha(t) = \frac{\omega^2}{1-t}, \quad \beta(t) = \frac{\nu^2\, t}{(1-t)^2}$ in the log-homotopy transformation. Given that OT-FM does not stabilize beyond the predetermined ODE boundary (Sprague et al., 2024), we extend the time horizon to $[0, t_{\text{end}}]$ for training and instead sample from a stationary-enforced marginal $\bar{\rho}_\infty$ beyond a cutoff time $t_{\max} < t_{\text{end}}$. Here, we define $\bar{\rho}_\infty = \bar{\rho}(x, t_{\max})$, corresponding to $f_\infty = f(t_{\max}), \quad g_\infty = g(t_{\max})$. Additionally, we set $t_{\max} < 1$, ensuring that $g(t_{\max})$ remains non-zero, preserving the diffusion component in the Fokker–Planck equation, which is essential for establishing the Boltzmann distribution. This adjustment promotes the convergence of the Fokker–Planck dynamics of ρ_Φ to a stationary Boltzmann equilibrium, aligning with the convergent dynamics of $\bar{\rho}$ through homotopy matching. Rather than freezing the time input of the potential energy, we enforce quasi-static dynamics by regularizing the Euclidean norm of the time derivative $\left\|\frac{\partial \Phi}{\partial t}\right\|^2$ during training, thereby encouraging convergence to the steady-state potential $\Phi_\infty(x)$.

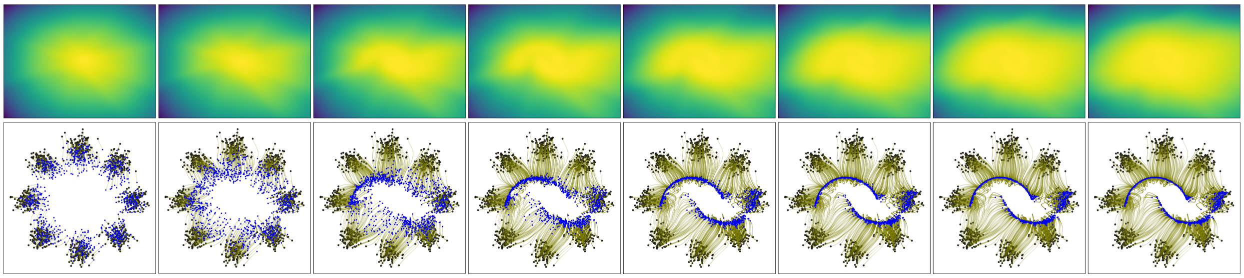

Figure 1: Visualization of the learned energy on 2D density datasets. Top: Density estimation over time using Boltzmann energy parameterization Φ_B. Bottom: Sample trajectories from the 8-Gaussian distribution (black) to the 2-Moons distribution (blue), driven by the potential energy Φ.

In particular, our VPFB loss function is implemented as follows:

$$\mathcal{L}^{\mathrm{VPFB}}(\Phi) = \int_0^{t_{\mathrm{end}}} \mathcal{L}(\Phi, t)\, dt = \mathbb{E}_{\mathcal{U}(0,t_{\mathrm{end}})}\left[\mathcal{L}(\Phi, t)\right] \tag{24}$$

where

$$\begin{aligned}\mathcal{L}(\Phi, t) = w(t)\ \mathrm{Cov}_{\rho(x|\bar{x},t)\, p_{\mathrm{data}}(\bar{x})}\left[\Phi(x, t), \gamma(x, \bar{x}, t)\right]\ - \frac{\nabla_x \Phi(x, t) \cdot v(x \mid \bar{x}, t)}{\left\|\nabla_x \Phi(x, t)\right\| \left\|v(x \mid \bar{x}, t)\right\|} \\ + \mathbb{E}_{\rho(x|\bar{x},t)\, p_{\mathrm{data}}(\bar{x})}\left[\left\|\nabla_{(x,t)} \Phi(x, t)\right\|^2 + \eta \left\|\Phi(x, t)\right\|^2\right]\end{aligned} \tag{25}$$

Here, we incorporate an additional cosine distance between the potential gradient $\nabla_x \Phi$ and the conditional vector field in (18) to the loss function. While this cosine distance does not influence the learning of the potential energy's magnitude (magnitude learning is entirely supervised by the covariance loss), it enforces directional alignment between the gradient and the vector field. Our empirical observations indicate that such gradient alignment improves performance. Considering that the marginal homotopy may not satisfy the Poincaré inequality (23), we include the right-hand side of this inequality in the loss function to enforce the uniqueness of the minimizer. Finally, a weighting $w(t) = (1 - t)^2$ is applied to balance the covariance loss across time to stabilize training. The cutoff time $t_{\max}$, terminal time t_{end}, and the spectral gap constant η are hyperparameters to be determined during implementation. Training Algorithm 1 of VPFB is provided in Appendix B.

4 Experiments

In this section, we validate the energy-based generative modeling capabilities of VPFB across several key tasks. Section 4.1 explores 2D density estimation. Section 4.2 presents the unconditional generation and spherical interpolation results on CIFAR-10 and CelebA. Section 4.3 evaluates mode coverage and model generalization through energy histograms of train and test data and the nearest neighbors of generated samples. Section 4.4 examines unsupervised OOD detection performance on various datasets. Section 4.5 verifies the convergence of long-run ODE samples to a Boltzmann equilibrium. Detailed implementation information, including architecture, training, numerical solvers, datasets, and FID evaluation, is provided in Appendix B.

4.1 Density Estimation on 2D Data

To verify the convergence properties of the potential energy and to assess the validity of the Boltzmann energy (20), we conduct density estimation on two-dimensional (2D) synthetic datasets. Specifically, we learn a potential flow that transforms an 8-Gaussian prior density into a 2-Moons target distribution. Figure 1 illustrates the log-likelihood (density) estimated by the Boltzmann distribution (1) and the sample

Figure 2: Uncurated and unconditional samples generated for CIFAR-10 (left) and CelebA (right).

trajectories generated by the potential flow 11 throughout ODE sampling. Our results demonstrate that the density estimation accurately approximates the target 2-Moons distribution at the ODE boundary, stabilizing after the cutoff time $t = 0.9$. Furthermore, the samples generated by the potential flow remain stationary beyond this point, indicating that the potential energy converges to its steady state and facilitates the formation of the Boltzmann energy. The dispersion of the energy-parameterized density beyond the cutoff time correctly captures the diffusion component. These results highlight the effectiveness of our variational principle approach in learning the Boltzmann stationary distribution through homotopy matching with the stationary-constrained marginal homotopy.

4.2 Image Generation

For image generation, we consider three VPFB model variants: an autonomous (independent of time) energy model $\Phi(x)$ parameterized by Zagoruyko & Komodakis (2016), and a time-varying energy model $\Phi(x,t)$ parameterized by U-Net (Ronneberger et al., 2015). Figure 2 shows the uncurated and unconditional image samples generated using the time-varying energy model on CIFAR-10 32×32 and CelebA 64×64. The

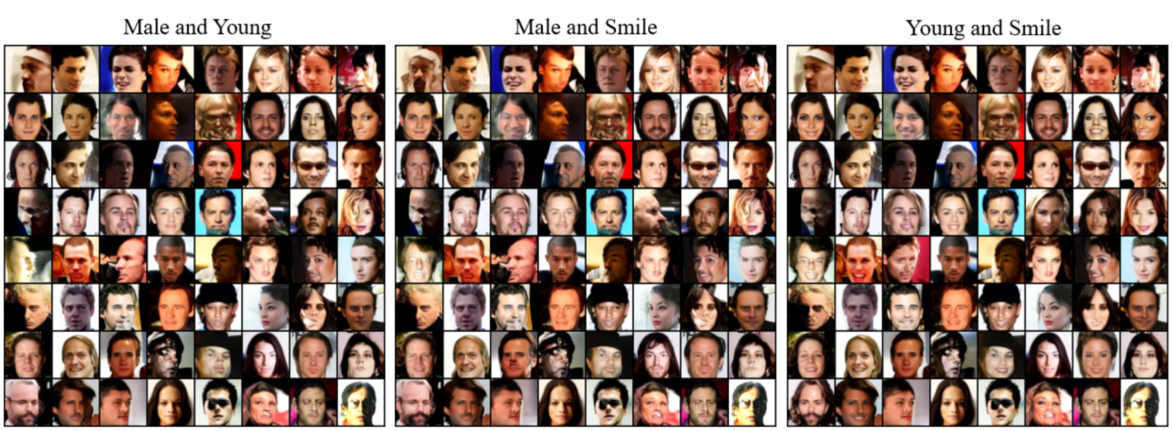

Figure 3: Compositional and conditional CelebA samples generated based on three attribute pairs.

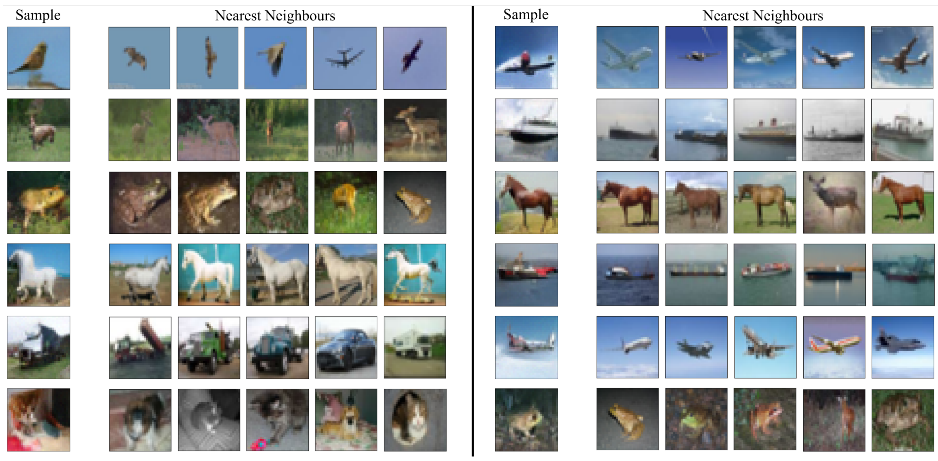

Figure 4: Generated CIFAR-10 samples and their five nearest neighbors in train set based on pixel distance.

generated samples are of decent quality and resemble the original datasets, despite not having the highest fidelity as achieved by state-of-the-art models. Table 1 summarizes the quantitative evaluations of our framework in terms of FID (Heusel et al., 2017) scores on the CIFAR-10. In particular, the VPFB models achieved FID scores competitive to existing generative models. Additional unconditional generation and image interpolation results on CIFAR-10 and CelebA are provided in Appendix C. To perform smooth and semantically coherent image interpolation, we construct the spherical interpolation between two Gaussian noises and then subject these interpolated noises to the ODE sampling. Figure 3 presents the results of compositional generation conditioned on composite CelebA attributes, specifically (*Male, Young*), (*Male, Smile*), and (*Young, Smile*). The compositional samples are obtained by first training a class-conditioned energy model $\Phi(x, c)$, and then sample using an average of the conditional energies across the selected classes. However, certain samples show limited variation across attribute pairs, suggesting that composition weight should be incorporated to enhance conditioning on specific class attributes.

4.3 Model Generalization and Mode Evaluation

To evaluate the model generalization capability of VPFB, Figure 4 presents the nearest neighbors of the generated samples in the train set of CIFAR-10. The results show that nearest neighbors are significantly different from the generated samples, thus suggesting that our models do not over-fit the training data and generalize well across the underlying data distribution. To validate the mode coverage and over-fitting ability, Figure 5 plots the histogram of the energy outputs on the CIFAR-10 train and test dataset. The energy histogram shows that the learned energy model assigns similar energy values to both train and test set images. This indicates that the VPFB model generalizes well to unseen test data and extensively covers all the modes in the training data.

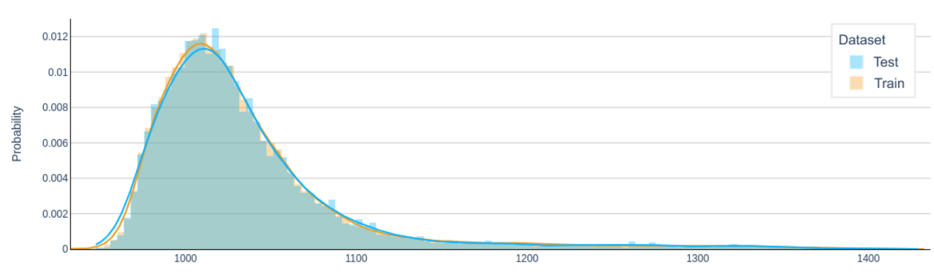

Figure 5: Histogram of energy-parameterized density estimation on CIFAR-10 train and test set.

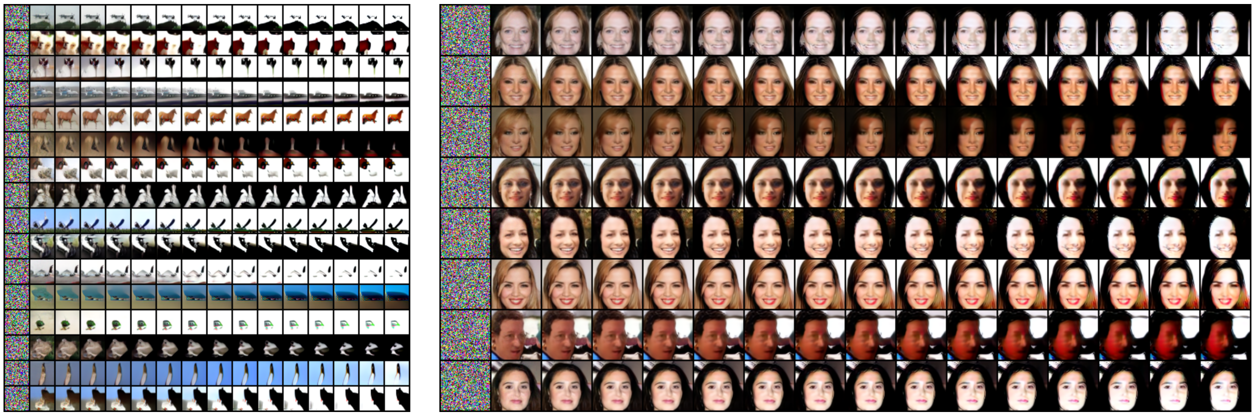

Figure 6: Long-run ODE sampling over an extended time horizon $t \in [0, 15]$ for CelebA, with a time interval of one second between consecutive samples.

4.4 Out-of-Distribution Detection

Given that the potential flow corresponds to a stationary Boltzmann distribution, the Boltzmann energy Φ_B from (20) can be used to differentiate between in-distribution and OOD samples based on their assigned energy values. Specifically, the potential energy model trained on the CIFAR-10 training set assigns energy values to both in-distribution samples (CIFAR-10 test set) and OOD samples from various other image datasets. The OOD performance is evaluated using the AUROC metric, where a higher score reflects the model's efficacy in accurately assigning lower energy values to OOD samples. Table 2 compares the AUROC scores of VPFB against various likelihood-based and energy-based models. The results show that our model performs exceptionally well on interpolated CIFAR-10 and CelebA 32×32 while achieving average performance on CIFAR-100 and SVHN.

4.5 Long-Run Steady-State Equilibrium

Figure 6 illustrates the long-run ODE sampling over an extended time horizon. The results indicate that the ODE samples remain stationary for the initial few seconds beyond the predetermined ODE boundary $[0, t_{\max}]$, suggesting that the potential flow is converging toward the Boltzmann equilibrium. However, as ODE sampling extends beyond $t = 5$, image quality deteriorates due to excessive saturation and loss of background detail. This suggests that extending the training time horizon to $[0, t_{\text{end}}]$ may be necessary to better enforce the stationarity of the marginal homotopy. Nevertheless, our observations align with the results of (Agoritsas et al., 2023) and (Nijkamp et al., 2020), which demonstrates that non-convergent MCMC sampling achieves optimal performance within a finite number of ODE iterations.

5 Conclusion

We propose VPFB, a novel energy-based potential flow framework that reduces the computational cost and instability in EBM training. Empirical results show that VPFB outperforms many existing EBMs in unconditional image generation and achieves competitive performance in OOD detection, demonstrating its versatility across generative modeling tasks. Despite these successes, future work will focus on refining the training strategy to enhance scalability for higher-resolution images and other data modalities. Also, we plan to investigate sampling directly from the stationary Boltzmann distribution using MCMC techniques, such as Hamiltonian Monte Carlo and Metropolis-Adjusted Langevin Algorithm (Pal et al., 2021a).

Broader Impact Statement

Generative models represent a rapidly growing field of study with overarching implications in science and society. Our work proposes a new generative model designed for efficient data generation and OOD detection, with potential applications in fields such as medical imaging, entertainment, and content creation. However, like any powerful technology, generative models come with substantial risks, including the potential misuse in creating deepfakes or misleading content that could undermine social security and trust. Given this dual-use nature, it is essential to implement safeguards, such as classifier-based guidance, to prevent the generation of biased or harmful content. Moreover, generative models are vulnerable to backdoor adversarial attacks and can inadvertently amplify biases present in the training data, reinforcing social inequalities. Although our work uses standard datasets, it is important to address how such biases are handled. We are actively exploring methods to identify and mitigate biases during both the training and generation phases. This includes employing fairness-aware training algorithms and evaluating the model's output for biased patterns. One potential solution is incorporating privacy-preserving encryption techniques to safeguard sensitive data and ensure that generative models do not expose private information. Furthermore, while this work demonstrates the potential benefits of generative models, the ethical concerns surrounding their deployment must be considered. Addressing these issues will require ongoing collaboration to develop frameworks for responsible use, including transparency, model interpretability, and robust safeguards against malicious applications. By proactively engaging with these ethical concerns, the broader community can contribute to the responsible advancement of generative modeling technologies.

References

Elisabeth Agoritsas, Giovanni Catania, Aurélien Decelle, and Beatriz Seoane. Explaining the effects of non-convergent MCMC in the training of energy-based models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 322–336. PMLR, 2023.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023.

Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. ISSN 0304-4149.

Michael Arbel, Liang Zhou, and Arthur Gretton. Generalized energy based models. In *International Conference on Learning Representations*, 2021.

Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7327–7347, 2022.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Tong Che, Ruixiang ZHANG, Jascha Sohl-Dickstein, Hugo Larochelle, Liam Paull, Yuan Cao, and Yoshua Bengio. Your gan is secretly an energy-based model and you should use discriminator driven latent sampling. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12275–12287. Curran Associates, Inc., 2020.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019a.

Xinshi Chen, Hanjun Dai, and Le Song. Particle flow Bayes' rule. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1022–1031, 2019b.

Fred Daum and Jim Huang. Nonlinear filters with log-homotopy. In *Signal and Data Processing of Small Targets*, volume 6699, pp. 669918, 2007. doi: 10.1117/12.725684.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. Residual energy-based models for text generation. In *International Conference on Learning Representations*, 2020.

J.R. Dormand and P.J. Prince. A family of embedded runge-kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, 1980. ISSN 0377-0427. doi: https://doi.org/10.1016/0771-050X(80)90013-3.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yilun Du, Toru Lin, and Igor Mordatch. Model-based planning with energy-based models. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura (eds.), *Proceedings of the Conference on Robot Learning*, volume 100, pp. 374–383, 2020.

Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy-based models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 2837–2848, 2021.

Weinan E and Bing Yu. The deep ritz method: A deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6(1):1–12, 2018.

Christoph Feinauer and Carlo Lucibello. Reconstruction of pairwise interactions using energy-based models*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124007, 2021.

Ruiqi Gao, Erik Nijkamp, Diederik P. Kingma, Zhen Xu, Andrew M. Dai, and Ying Nian Wu. Flow contrastive estimation of energy-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P Kingma. Learning energy-based models by diffusion recovery likelihood. In *International Conference on Learning Representations*, 2021.

C. Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer Series in Synergetics. Springer Berlin Heidelberg, 2009. ISBN 9783540707127.

Cong Geng, Tian Han, Peng-Tao Jiang, Hao Zhang, Jinwei Chen, Søren Hauberg, and Bo Li. Improving adversarial energy-based model via diffusion process. In *Forty-first International Conference on Machine Learning*, 2024.

Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020a.

Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, and Richard Zemel. Learning the stein discrepancy for training and evaluating energy-based models without sampling. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 3732–3747, 2020b.

Will Sussman Grathwohl, Jacob Jin 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*, 2021.

Tian Han, Erik Nijkamp, Xiaolin Fang, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Divergence triangle for joint training of generator model, energy-based model, and inferential model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8670–8679, 2019.

Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units, 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Mitch Hill, Erik Nijkamp, Jonathan Craig Mitchell, Bo Pang, and Song-Chun Zhu. Learning probabilistic models from generator latent spaces with hat EBM. In *Advances in Neural Information Processing Systems*, 2022.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022.

Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. *arXiv preprint arXiv:2106.05527*, 2021.

Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 65484–65516. Curran Associates, Inc., 2023.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

Richard S. Laugesen, Prashant G. Mehta, Sean P. Meyn, and Maxim Raginsky. Poisson's equation in nonlinear filtering. *SIAM Journal on Control and Optimization*, 53(1):501–525, 2015.

Hankook Lee, Jongheon Jeong, Sejun Park, and Jinwoo Shin. Guiding energy-based models via contrastive latent variables. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Meng Liu, Keqiang Yan, Bora Oztekin, and Shuiwang Ji. GraphEBM: Molecular graph generation with energy-based models. In *Energy Based Models Workshop - ICLR 2021*, 2021.

Min Liu, Zhiqiang Cai, and Karthik Ramani. Deep ritz method with adaptive quadrature for linear elasticity. *Computer Methods in Applied Mechanics and Engineering*, 415:116229, 2023a.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023b.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Eric Luhman, Troy Luhman, and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Johannes Müller and Marius Zeinhofer. Deep ritz revisited. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2019.

Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: Learning stochastic dynamics from samples. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25858–25889. PMLR, 2023.

Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5272–5280, 2020.

Erik Nijkamp, Ruiqi Gao, Pavel Sountsov, Srinivas Vasudevan, Bo Pang, Song-Chun Zhu, and Ying Nian Wu. MCMC should mix: Learning energy-based model with neural transport latent space MCMC. In *International Conference on Learning Representations*, 2022.

S. Yagiz Olmez, Amirhossein Taghvaei, and Prashant G. Mehta. Deep fpf: Gain function approximation in high-dimensional setting. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 4790–4795, 2020.

Arghya Pal, Raphael C.-W. Phan, and KokSheik Wong. Synthesize-it-classifier: Learning a generative classifier through recurrent self-analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5161–5170, June 2021a.

Soumyasundar Pal, Liheng Ma, Yingxue Zhang, and Mark Coates. Rnn with particle flow for probabilistic spatio-temporal forecasting. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8336–8348, 2021b.

Bo Pang, Tianyang Zhao, Xu Xie, and Ying Nian Wu. Trajectory prediction with latent belief energy-based model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11814–11824, June 2021.

Gaurav Parmar, Dacheng Li, Kwonjoon Lee, and Zhuowen Tu. Dual contradistinctive generative autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 823–832, 2021.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 1530–1538, 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing.

Tim Salimans and Jonathan Ho. Should EBMs model the energy or the score? In *Energy Based Models Workshop - ICLR 2021*, 2021.

Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Yang Song and Diederik P. Kingma. How to train your energy-based models, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Christopher Iliffe Sprague, Arne Elofsson, and Hossein Azizpour. Stable autonomous flow matching. *arXiv preprint arXiv:2402.05774*, 2024.

Simone Carlo Surace, Anna Kutschireiter, and Jean-Pascal Pfister. How to avoid the curse of dimensionality: Scalability of particle filters with and without importance weights. *SIAM Review*, 61(1):79–91, 2019.

Amirhossein Taghvaei, Prashant G. Mehta, and Sean P. Meyn. Diffusion map-based algorithm for gain function approximation in the feedback particle filter. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):1090–1117, 2020.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19667–19679, 2020.

Dafeng Wang, Hongbo Liu, Naiyao Wang, Yiyang Wang, Hua Wang, and Seán McLoone. Seem: A sequence entropy energy-based model for pedestrian trajectory all-then-one prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1070–1086, 2023.

Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. {VAEBM}: A symbiosis between variational autoencoders and energy-based models. In *International Conference on Learning Representations*, 2021.

Jianwen Xie, Yang Lu, Ruiqi Gao, Song-Chun Zhu, and Ying Nian Wu. Cooperative training of descriptor and generator networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):27–45, 2020.

Jianwen Xie, Yifei Xu, Zilong Zheng, Song-Chun Zhu, and Ying Nian Wu. Generative pointnet: Deep energy-based learning on unordered point sets for 3d generation, reconstruction and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14976–14985, 2021a.

Jianwen Xie, Zilong Zheng, and Ping Li. Learning energy-based model with variational auto-encoder as amortized sampler. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10441–10451, 2021b.

Jianwen Xie, Yaxuan Zhu, Jun Li, and Ping Li. A tale of two flows: Cooperative learning of langevin flow and normalizing flow toward energy-based model. *arXiv preprint arXiv:2205.06924*, 2022.

Yilun Xu, Ziming Liu, Max Tegmark, and Tommi Jaakkola. Poisson flow generative models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 16782–16795, 2022.

Tao Yang, Prashant G. Mehta, and Sean P. Meyn. Feedback particle filter. *IEEE Transactions on Automatic Control*, 58(10):2465–2480, 2013.

Tao Yang, Henk A. P. Blom, and Prashant G. Mehta. The continuous-discrete time feedback particle filter. In *American Control Conference (ACC)*, pp. 648–653, 2014.

Tao Yang, Richard S. Laugesen, Prashant G. Mehta, and Sean P. Meyn. Multivariable feedback particle filter. *Automatica*, 71:10–23, 2016. ISSN 0005-1098.

Xiulong Yang, Qing Su, and Shihao Ji. Towards bridging the performance gaps of joint energy-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15732–15741, 2023.

Sangwoong Yoon, Young-Uk Jin, Yung-Kyun Noh, and Frank C. Park. Energy-based models for anomaly detection: A manifold diffusion recovery approach. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. In *British Machine Vision Conference 2016*, York, France, January 2016. British Machine Vision Association. doi: 10.48550/arXiv.1605.07146.

Yang Zhao, Jianwen Xie, and Ping Li. Learning energy-based generative models via coarse-to-fine expanding and sampling. In *International Conference on Learning Representations*, 2020.

Yaxuan Zhu, Jianwen Xie, Ying Nian Wu, and Ruiqi Gao. Learning energy-based models by cooperative diffusion recovery likelihood. In *The Twelfth International Conference on Learning Representations*, 2024.

A Proofs and Derivations

A.1 Proof of Proposition 1

Proof. Based on the definitions of $q(x)$ and $p(\bar{x} \mid x)$, we can expand their logarithms (ignoring additive constants) as follows:

$$\log q(x) = -\frac{1}{2\,\omega^2}\,\frac{1}{2\,\omega^2}\,\|x\|^2 + (\text{terms independent of } x) \tag{26}$$

$$\log p(\bar{x} \mid x) = -\frac{1}{2\,\nu^2}\,\|\bar{x} - x\|^2 + (\text{terms independent of } x) \tag{27}$$

Substituting these into (5), we obtain:

$$f(x \mid \bar{x}, t) = -\frac{\alpha(t)}{2\,\omega^2}\,\|x\|^2 - \frac{\beta(t)}{2\,\nu^2}\,\|\bar{x} - x\|^2 \tag{28}$$

Expanding the squared term:

$$\|\bar{x} - x\|^2 = \|x\|^2 - 2\,x^T\bar{x} + \|\bar{x}\|^2 \tag{29}$$

and substituting it back into $f(x \mid \bar{x}, t)$:

$$f(x \mid \bar{x}, t) = -\left(\frac{\alpha(t)}{\omega^2} + \frac{\beta(t)}{\nu^2}\right)\|x\|^2 + \frac{\beta(t)}{\nu^2}\,x^T\bar{x} \tag{30}$$

Recognizing the quadratic form in terms of x, we identify that $\rho(x \mid \bar{x}, t)$ is a Gaussian density:

$$\rho(x \mid \bar{x}, t) = \mathcal{N}\big(x; \mu(t)\bar{x}, \sigma(t)^2 I\big) \tag{31}$$

whose mean $\mu(t)$ and variance $\sigma(t)^2$ can be obtained by completing the square.

Define

$$A := \frac{\alpha(t)}{\omega^2} + \frac{\beta(t)}{\nu^2}, \qquad B := \frac{\beta(t)}{\nu^2} \tag{32}$$

Then the exponent becomes

$$f(x \mid \bar{x}, t) = -\frac{1}{2}\Big[A\,\|x\|^2 - 2B\,x^T\bar{x}\Big] \tag{33}$$

and we wish to express this quadratic form as follows

$$A\,\|x - \mu(t)\,\bar{x}\|^2 + (\text{terms independent of } x) \tag{34}$$

Expanding $A\,\|x - \mu(t)\,\bar{x}\|^2$, we obtain

$$A\,\|x - \mu(t)\,\bar{x}\|^2 = A\,\|x\|^2 - 2A\,\mu(t)\,x^T\bar{x} + A\,\mu(t)^2\,\|\bar{x}\|^2 \tag{35}$$

To match the linear term, the mean of the Gaussian is thus given by

$$\mu(t) = \frac{B}{A} = \frac{\beta(t)/\nu^2}{\alpha(t)/\omega^2 + \beta(t)/\nu^2} = \text{sigmoid}\left(\log\left(\frac{\beta(t)}{\alpha(t)}\,\frac{\omega^2}{\nu^2}\right)\right) \tag{36}$$

where $\text{sigmoid}(z) = \frac{1}{1+e^{-z}}$ denotes the standard logistic (sigmoid) function.

By comparing with the standard Gaussian exponent

$$-\frac{1}{2\sigma^2}\,\|x - \mu(t)\,\bar{x}\|^2, \tag{37}$$

we deduce that the variance is given by

$$\sigma(t)^2 = \frac{1}{A} = \frac{1}{\alpha(t)/\omega^2 + \beta(t)/\nu^2}. \tag{38}$$

Using the expression obtained for $\mu(t)$, the standard deviation can also be written as

$$\sigma(t) = \sqrt{\frac{\nu^2}{\beta(t)}\,\mu(t)}. \tag{39}$$

□

A.2 Proof of Proposition 2

Proof. Differentiating the conditional homotopy $\rho(x \mid \bar{x}, t)$ in (4) with respect to t, we have:

$$\begin{aligned}
\frac{\partial \rho(x \mid \bar{x}, t)}{\partial t} &= \frac{1}{\int_\Omega e^{h(x|\bar{x},t)}\,dx}\,\frac{\partial[e^{h(x|\bar{x},t)}]}{\partial t} - \frac{e^{h(x|\bar{x},t)}}{[\int_\Omega e^{h(x|\bar{x},t)}\,dx]^2}\,\frac{\partial[\int_\Omega e^{h(x|\bar{x},t)}\,dx]}{\partial t} \\
&= \frac{1}{\int_\Omega e^{h(x|\bar{x},t)}\,dx}\,\frac{\partial[e^{h(x|\bar{x},t)}]}{\partial f}\,\frac{\partial h(x \mid \bar{x}, t)}{\partial t} - \frac{e^{h(x|\bar{x},t)}}{[\int_\Omega e^{h(x|\bar{x},t)}\,dx]^2}\int_\Omega \frac{\partial[e^{h(x|\bar{x},t)}]}{\partial f}\,\frac{\partial h(x \mid \bar{x}, t)}{\partial t}\,dx \\
&= \frac{e^{h(x|\bar{x},t)}}{\int_\Omega e^{h(x|\bar{x},t)}\,dx}\,\frac{\partial h(x \mid \bar{x}, t)}{\partial t} - \frac{e^{h(x|\bar{x},t)}}{\int_\Omega e^{h(x|\bar{x},t)}\,dx}\int_\Omega \frac{e^{h(x|\bar{x},t)}}{\int_\Omega e^{h(x|\bar{x},t)}\,dx}\,\frac{\partial h(x \mid \bar{x}, t)}{\partial t}\,dx \\
&= \rho(x \mid \bar{x}, t)\left(\frac{\partial h(x \mid \bar{x}, t)}{\partial t} - \int_\Omega \rho(x \mid \bar{x}, t)\,\frac{\partial h(x \mid \bar{x}, t)}{\partial t}\,dx\right) \\
&= -\frac{1}{2}\,\rho(x \mid \bar{x}, t)\left(\frac{d\alpha(t)}{dt}\,\frac{x^T x}{\omega^2} + \frac{d\beta(t)}{dt}\,\frac{(x-\bar{x})^T(x-\bar{x})}{\nu^2}\right. \\
&\qquad\left. - \int_\Omega \rho(x \mid \bar{x}, t)\,\frac{d\alpha(t)}{dt}\,\frac{x^T x}{\omega^2} + \frac{d\beta(t)}{dt}\,\frac{(x-\bar{x})^T(x-\bar{x})}{\nu^2}\,dx\right)
\end{aligned} \tag{40}$$

where we have applied the quotient rule in the first equation and the chain rule in the second equation.

Subsequently, define

$$\gamma(x, \bar{x}, t) = \frac{d\alpha(t)}{dt}\,\frac{\|x\|^2}{\omega^2} + \frac{d\beta(t)}{dt}\,\frac{\|x-\bar{x}\|^2}{\nu^2} \tag{41}$$

and using the fact that:

$$\frac{\partial \bar{\rho}(x,t)}{\partial t} = \frac{\partial \int_\Omega \rho(x \mid \bar{x}, t)\,p_{\text{data}}(\bar{x})\,d\bar{x}}{\partial t} = \int_\Omega \frac{\partial \rho(x \mid \bar{x}, t)}{\partial t}\,p_{\text{data}}(\bar{x})\,d\bar{x} \tag{42}$$

we can substitute (40) into (42) to obtain:

$$\frac{\partial \bar{\rho}(x,t)}{\partial t} = -\int_\Omega p_{\text{data}}(\bar{x})\,\rho(x \mid \bar{x}, t)\left(\gamma(x, \bar{x}, t) - \int_\Omega \rho(x \mid \bar{x}, t)\,\gamma(x, \bar{x}, t)\,dx\right)d\bar{x} \tag{43}$$

Given that both $\rho(x \mid \bar{x}, t)$ and $p_{\text{data}}(\bar{x})$ are normalized (proper) density functions, writing (43) in terms of expectations yields the PDE in (9).

□

A.3 Proof of Proposition 3

Here, we used the Einstein tensor notation interchangeably with the conventional notation for vector dot product and matrix-vector multiplication in PDE. Also, we write for example, ρ_t^Φ instead of $\rho_\Phi(x, t)$, for brevity.

Proof. Applying forward Euler to the particle flow ODE (11) using step size Δ_t, we obtain:

$$x_{t+\Delta_t} = \alpha(x_t) = x_t + \Delta_t \, u(x_t) \tag{44}$$

where

$$u(x_t) = \nabla_x \Phi(x_t) \tag{45}$$

where x_t denotes the discretization of $x(t)$.

Assuming that the $\alpha : \Omega \to \Omega$ is a diffeomorphism (bijective function with differentiable inverse), the push-forward operator $\alpha_\# : \mathbb{R} \to \mathbb{R}$ on the density function $\rho_t^\Phi \mapsto \rho_{t+\Delta_t}^\Phi := \alpha_\# \rho_t^\Phi$ is defined as follows:

$$\int_\Omega \rho_{t+\Delta_t}^\Phi(x)\, g(x)\, dx = \int_\Omega \alpha_\# \rho_t^\Phi(x)\, g(x)\, dx = \int_\Omega \rho_t^\Phi(x)\, g\big(\alpha(x)\big)\, dx \tag{46}$$

for any measurable function g.

Associated with this change of variable formula is the following density transformation:

$$\rho_{t+\Delta_t}^\Phi\big(\alpha(x)\big) = \frac{1}{|D\alpha|}\, \rho_t^\Phi(x) \tag{47}$$

where $|D\alpha|$ denotes the Jacobian determinant of α.

From (9) and (43), we obtain:

$$\frac{\partial \log \bar{\rho}_t(x)}{\partial t} = \frac{1}{\bar{\rho}_t(x)} \frac{\partial \bar{\rho}_t(x)}{\partial t} = -\frac{1}{\bar{\rho}_t(x)} \frac{1}{2}\, \mathbb{E}_{p_{\text{data}}(\bar{x})} \Big[\rho_t(x \mid \bar{x}) \big(\gamma(x, \bar{x}, t) - \bar{\gamma}(x, \bar{x}, t)\big)\Big] \tag{48}$$

Applying the forward Euler method to (48), we obtain:

$$\log \bar{\rho}_{t+\Delta_t}(x) \geq \log \bar{\rho}_t(x) - \frac{\Delta_t}{2} \frac{1}{\bar{\rho}_t(x)}\, \mathbb{E}_{p_{\text{data}}(\bar{x})} \Big[\rho_t(x \mid \bar{x}) \big(\gamma(x, \bar{x}, t) - \bar{\gamma}(x, \bar{x}, t)\big)\Big] \tag{49}$$

Applying the change-of-variables formula (46) and density transformation (47), then substituting (49) into the KL divergence $\mathcal{D}_{\text{KL}}\left[\rho_{t+\Delta_t} \| \bar{\rho}_{t+\Delta_t}\right]$ at time $t + \Delta_t$, we obtain:

$$\begin{aligned}
\mathcal{D}_{\text{KL}}\left[\rho_{t+\Delta_t}^\Phi(x) \| \bar{\rho}_{t+\Delta_t}(x)\right] &= \int_\Omega \rho_t^\Phi(x)\, \log\left(\frac{\rho_{t+\Delta_t}^\Phi\big(\alpha(x)\big)}{\bar{\rho}_{t+\Delta_t}\big(\alpha(x)\big)}\right) dx \\
= \int_\Omega \rho_t^\Phi(x) \Bigg(&\log \rho_t^\Phi(x) - \log |D\alpha| - \log \bar{\rho}_t\big(\alpha(x)\big) \\
&+ \frac{\Delta_t}{2} \frac{1}{\bar{\rho}_t\big(\alpha(x)\big)}\, \mathbb{E}_{p_{\text{data}}(\bar{x})} \Big[\rho_t(\alpha(x), \bar{x}) \Big(\gamma\big(\alpha(x), \bar{x}\big) - \bar{\gamma}\big(\alpha(x), \bar{x}\big)\Big)\Big] + C \Bigg) dx
\end{aligned} \tag{50}$$

Consider minimizing the KL divergence (50) with respect to α as follows:

$$\begin{aligned}
&\min_\alpha \;\; \mathcal{D}_{\text{KL}}(\alpha) \\
&= \min_\alpha \;\; \underbrace{\frac{\Delta_t}{2} \int_\Omega \rho_t^\Phi(x) \frac{1}{\bar{\rho}_t\big(\alpha(x)\big)}\, \mathbb{E}_{p_{\text{data}}(\bar{x})} \Big[\rho_t\big(\alpha(x), \bar{x}\big) \Big(\gamma\big(\alpha(x), \bar{x}\big) - \bar{\gamma}\big(\alpha(x), \bar{x}\big)\Big)\Big] dx}_{\mathcal{D}_1^{\text{KL}}(\alpha)} \\
&\qquad - \underbrace{\int_\Omega \rho_t^\Phi(x)\, \log \bar{\rho}_t(\alpha(x))\, dx}_{\mathcal{D}_2^{\text{KL}}(\alpha)} \;\; - \underbrace{\int_\Omega \rho_t^\Phi(x)\, \log |D\alpha|\, dx}_{\mathcal{D}_3^{\text{KL}}(\alpha)}
\end{aligned} \tag{51}$$

where we have neglected the constant terms that do not depend on α.

To solve the optimization (51), we consider the following optimality condition in the first variation of $\mathcal{D}_{\mathrm{KL}}$:

$$\mathcal{I}(\alpha,\nu) = \frac{d}{d\epsilon}\, \mathcal{D}_{\mathrm{KL}}\left(\alpha(x) + \epsilon\,\nu(x)\right)\Big|_{\epsilon=0} = 0 \tag{52}$$

This condition must hold for all trial functions $\nu(x)$.

Taking the variational derivative of the first functional $\mathcal{D}_1^{\mathrm{KL}}$ in (51), we obtain:

$$\begin{aligned}
\mathcal{I}^1(\alpha,\nu) &= \frac{d}{d\epsilon}\,\mathcal{D}_1^{\mathrm{KL}}(\alpha+\epsilon\nu)\Big|_{\epsilon=0} \\
&= \frac{\Delta}{2}\int_\Omega \rho^\Phi(x)\,\frac{d}{d\epsilon}\left\{\frac{1}{\bar\rho(\alpha+\epsilon\nu)}\,\mathbb{E}_{p_{\mathrm{data}}(\bar x)}\left[\rho(\alpha+\epsilon\nu,\bar x)\left(\gamma(\alpha+\epsilon\nu,\bar x)-\bar\gamma(\alpha+\epsilon\nu,\bar x)\right)\right]\right\}\Big|_{\epsilon=0} dx \\
&= \frac{\Delta}{2}\int_\Omega \rho^\Phi(x)\,D\left\{\frac{1}{\bar\rho(x)}\,\mathbb{E}_{p_{\mathrm{data}}(\bar x)}\left[\rho(x\mid\bar x)\left(\gamma(x,\bar x,t)-\bar\gamma(x,\bar x,t)\right)\right]\right\}\nu\,dx
\end{aligned} \tag{53}$$

where $Dg := \nabla_x^T g$ denotes the Jacobian of the function $g(x)$ with respect to x.

A Taylor series expansion of the derivative $\frac{\partial g}{\partial x_i}(\alpha)$ with respect to x_i yields:

$$\frac{\partial g(\alpha)}{\partial x_i} = \frac{\partial g(x+\Delta u)}{\partial x_i} = \frac{\partial g(x)}{\partial x_i} + \Delta\sum_j \frac{\partial^2 g(x)}{\partial x_i\,\partial x_j}\,u_j + O(\Delta^2) \tag{54}$$

Using the Taylor series expansion (54), (53) can be written in tensor notation as follows:

$$\mathcal{I}^1(\alpha,\nu) = \frac{\Delta}{2}\int_\Omega \rho^\Phi(x)\sum_i \frac{\partial}{\partial x_i}\left\{\frac{1}{\bar\rho(x)}\,\mathbb{E}_{p_{\mathrm{data}}(\bar x)}\left[\rho(x\mid\bar x)\left(\gamma(x,\bar x,t)-\bar\gamma(x,\bar x,t)\right)\right]\right\}\nu_i\,dx \;+\; O(\Delta^2) \tag{55}$$

Taking the variational derivative of the second functional $\mathcal{D}_2^{\mathrm{KL}}$ in (51) yields:

$$\begin{aligned}
\mathcal{I}^2(\alpha,\nu) &= \frac{d}{d\epsilon}\,\mathcal{D}_2^{\mathrm{KL}}(\alpha+\epsilon\nu)\Big|_{\epsilon=0} \\
&= \int_\Omega \rho^\Phi(x)\,\frac{d}{d\epsilon}\log\bar\rho(\alpha+\epsilon\nu)\Big|_{\epsilon=0} dx \\
&= \int_\Omega \rho^\Phi(x)\,\frac{1}{\bar\rho(\alpha)}\,\nabla_x\bar\rho(\alpha)\cdot\nu\,dx \\
&= \int_\Omega \rho^\Phi(x)\,\nabla_x\log\bar\rho(\alpha)\cdot\nu\,dx
\end{aligned} \tag{56}$$

where we have used the derivative identity $d\log g = \frac{1}{g}\,dg$ to obtain the second equation.

Using the Taylor series expansion (54), (56) can be written in tensor notation as follows:

$$\begin{aligned}
\mathcal{I}^2(\alpha,\nu) &= -\int_\Omega \rho^\Phi(x)\sum_i\left(\frac{\partial\log\bar\rho(x)}{\partial x_i} - \Delta\sum_j\frac{\partial^2\log\bar\rho(x)}{\partial x_i\,\partial x_j}\,u_j\right)\nu_i\,dx \;+\; O(\Delta^2) \\
&= -\int_\Omega \rho^\Phi(x)\sum_i\left(\frac{\partial\log\bar\rho(x)}{\partial x_i} - \Delta\sum_j\frac{\partial^2\log\bar\rho(x)}{\partial x_i\,\partial x_j}\,u_j\right)\nu_i\,dx \;+\; O(\Delta^2)
\end{aligned} \tag{57}$$

Similarly, taking the variational derivative of the $\mathcal{D}_3^{\mathrm{KL}}$ term in (51), we obtain:

$$\begin{aligned}\mathcal{I}^3(\alpha,\nu) &= \frac{d}{d\epsilon}\,\mathcal{D}_3^{\mathrm{KL}}(\alpha+\epsilon\nu)\Big|_{\epsilon=0} \\ &= \int_\Omega \rho^\Phi(x)\,\frac{d}{d\epsilon}\log\big|D(\alpha+\epsilon\nu)\big|\Big|_{\epsilon=0}\,dx \\ &= \int_\Omega \rho^\Phi(x)\,\frac{1}{\big|D\alpha\big|}\,\frac{d}{d\epsilon}\big|D(\alpha+\epsilon\nu)\big|\Big|_{\epsilon=0}\,dx \\ &= \int_\Omega \rho^\Phi(x)\;\mathrm{tr}\left(D\alpha^{-1}D\nu\right)dx\end{aligned} \tag{58}$$

where we have used the following Jacobi's formula:

$$\frac{d}{d\epsilon}\big|D(\alpha+\epsilon\nu)\big|\Big|_{\epsilon=0} = |D\alpha|\,\mathrm{tr}\left(D\alpha^{-1}D\nu\right) \tag{59}$$

to obtain the last equation in (58).

The inverse of Jacobian $D\alpha^{-1}$ can be expanded via the Neuman series to obtain:

$$D\alpha^{-1} = \left(\mathrm{I}+\Delta\,Du\right)^{-1} = \mathrm{I}-\Delta\,Du \;+\; O(\Delta^2) \tag{60}$$

Substituting in (60) and using the Taylor series expansion (54), (56) can be written in tensor notation as follows:

$$\begin{aligned}\mathcal{I}^3(\alpha,\nu) &= \int_\Omega \sum_i \left(\rho^\Phi(x)\,\frac{\partial\nu_i}{\partial x_i} - \Delta\sum_j \rho^\Phi(x)\,\frac{\partial u_j}{\partial x_i}\,\frac{\partial \nu_i}{\partial x_j}\right)dx \;+\; O(\Delta^2) \\ &= \int_\Omega \sum_i \left(\frac{\partial\rho^\Phi(x)}{\partial x_i}\,\nu_i - \Delta\sum_j \frac{\partial}{\partial x_j}\left\{\rho^\Phi(x)\,\frac{\partial u_j}{\partial x_i}\right\}\nu_i\right)dx \;+\; O(\Delta^2) \\ &= \int_\Omega \sum_i \left(\frac{\partial\rho^\Phi(x)}{\partial x_i} - \Delta\sum_j \frac{\partial}{\partial x_j}\left\{\rho^\Phi(x)\,\frac{\partial u_j}{\partial x_i}\right\}\right)\nu_i\,dx \;+\; O(\Delta^2)\end{aligned} \tag{61}$$

where we have used integration by parts to obtain the second equation.

Taking the limit $\lim\Delta\to 0$, the terms $O(\Delta^2)$ that approach zero exponentially vanish. Subtracting (55) by (57) and (61), then equating to zero, we obtain the first-order optimality condition (52) as follows:

$$\begin{aligned}\int_\Omega \bar\rho(x)\sum_i\Bigg(\sum_j -\frac{\partial}{\partial x_i}\left\{\frac{1}{\bar\rho(x)}\,\frac{\partial}{\partial x_j}\left\{\bar\rho(x)\,u_j\right\}\right\} \\ +\frac{1}{2}\,\frac{\partial}{\partial x_i}\left\{\frac{1}{\bar\rho(x)}\;\mathbb{E}_{p_{\mathrm{data}}(\bar x)}\Big[\rho(x\mid\bar x)\left(\gamma(x,\bar x,t)-\bar\gamma(x,\bar x,t)\right)\Big]\right\}\Bigg)\,\nu_i\,dx = 0\end{aligned} \tag{62}$$

where we have assumed that $\rho^\Phi(x)\equiv\bar\rho(x)$ holds and have used the following identities:

$$\begin{aligned}\frac{\partial\log\bar\rho(x)}{\partial x_i} &= \frac{1}{\bar\rho(x)}\,\frac{\partial\bar\rho(x)}{\partial x_i} \\ \frac{\partial^2\log\bar\rho(x)}{\partial x_i\,\partial x_j} &= \frac{\partial}{\partial x_i}\left(\frac{1}{\bar\rho(x)}\,\frac{\partial\bar\rho(x)}{\partial x_j}\right)\end{aligned} \tag{63}$$

Given that ν_i can take any value, equation (62) holds (in the weak sense) only if the terms within the round bracket vanish. Integrating this term with respect to the x_i, we obtain:

$$\sum_j \frac{\partial}{\partial x_j}\left\{\bar\rho(x)\,u_j\right\} = \frac{1}{2}\;\mathbb{E}_{p_{\mathrm{data}}(\bar x)}\Big[\rho(x\mid\bar x)\left(\gamma(x,\bar x,t)-\bar\gamma(x,\bar x,t)\right)\Big] \;+\; \bar\rho(x)\,C \tag{64}$$

which can also be written in vector notation as follows:

$$\nabla_x \cdot \big(\bar{\rho}(x)\, u\big) = \frac{1}{2}\, \mathbb{E}_{p_{\text{data}}(\bar{x})} \Big[\rho(x \mid \bar{x}) \big(\gamma(x, \bar{x}, t) - \bar{\gamma}(x, \bar{x}, t)\big)\Big] \,+\, \bar{\rho}(x)\, C \tag{65}$$

To find the scalar constant C, we integrate both sides of (65) to obtain:

$$\begin{aligned} \int_\Omega \nabla_x \cdot \big(\bar{\rho}(x)\, u\big)\, dx &= \frac{1}{2} \int_\Omega \mathbb{E}_{p_{\text{data}}(\bar{x})} \Big[\rho(x \mid \bar{x}) \big(\gamma(x, \bar{x}, t) - \bar{\gamma}(x, \bar{x}, t)\big)\Big]\, dx \,+\, \int_\Omega \bar{\rho}(x)\, C\, dx \\ &= \frac{1}{2} \int_\Omega \mathbb{E}_{p_{\text{data}}(\bar{x})} \Big[\rho(x \mid \bar{x}) \big(\gamma(x, \bar{x}, t) - \bar{\gamma}(x, \bar{x}, t)\big)\Big]\, dx \,+\, C \end{aligned} \tag{66}$$

Applying the divergence theorem to the left-hand side of (66), we obtain:

$$\int_\Omega \nabla_x \cdot \big(\bar{\rho}(x)\, u\big)\, dx = \int_{\partial\Omega} \bar{\rho}(x)\, u \cdot \hat{n}\, dx \tag{67}$$

where $\hat{n}$ is the outward unit normal vector to the boundary $\partial\Omega$ of Ω.

Given that $\bar{\rho}(x)$ is a normalized (proper) density with compact support (vanishes on the boundary), the term (67) becomes zero, leading to $C = 0$. Substituting this result along with $u(x) = \nabla_x \Phi(x)$ into (65), we arrive at the following PDE:

$$\nabla_x \cdot \big(\bar{\rho}_t(x)\, \nabla_x \Phi(x)\big) = \frac{1}{2}\, \mathbb{E}_{p_{\text{data}}(\bar{x})} \Big[\rho_t(x \mid \bar{x}) \big(\gamma(x, \bar{x}, t) - \bar{\gamma}(x, \bar{x}, t)\big)\Big] \tag{68}$$

Therefore, assuming that the base case $\rho_0(x) \equiv \bar{\rho}_0(x)$ holds and that a solution to (68) exists at every t, the proposition follows by the principle of induction.

□

A.4 Proof of Proposition 4

To show that the conditional and marginal homotopies satisfy the reverse diffusion process, we first express the forward-time SDE and ODE of Song et al. (2021):

$$\begin{aligned} dx(\tau) &= f(\tau)\, x(\tau)\, d\tau + g(\tau)\, dW(\tau) \\ \frac{dx(\tau)}{d\tau} &= f(\tau)\, x(\tau) - \frac{1}{2}\, g(\tau)^2\, \nabla_x \log p(x, \tau) \end{aligned} \tag{69}$$

in terms of reverse time $t = 1 - \tau$, via applying the change of variable $dt = -d\tau$ as follows:

$$\begin{aligned} dx(t) &= -f(t)\, x(t)\, dt + g(t)\, dW(t) \\ \frac{dx(t)}{dt} &= -f(t)\, x(t) + \frac{1}{2}\, g(t)^2\, \nabla_x \log \bar{\rho}(x, t) \end{aligned} \tag{70}$$

which gives (15) and (17).

Substituting the marginal score $\nabla_x \log \bar{\rho}(x, t)$ with the conditional score:

$$\nabla_x \log \rho(x \mid \bar{x}, t) = \frac{1}{\rho(x \mid \bar{x}, t)}\, \nabla_x \rho(x \mid \bar{x}, t) = -\frac{\epsilon}{\sigma(t)} \tag{71}$$

and applying reparameterization $x(t) = \mu(t)\,\bar{x} + \sigma(t)\,\epsilon$ and (16), we can write the conditional ODE as follows:

$$\begin{aligned}
\frac{dx(t)}{dt} &= v(x \mid \bar{x}, t) \\
&= -f(t)\,x(t) + \frac{1}{2}\,g(t)^2\,\nabla_x \log \rho(x \mid \bar{x}, t) \\
&= -f(t)\,x(t) + \frac{1}{2}\,g(t)^2\,\frac{\epsilon}{\sigma(t)} \\
&= -f(t)\,x(t) + \sigma(t)\,\Big(\dot{\sigma}(t) + f(t)\,\sigma(t)\Big)\,\frac{\epsilon}{\sigma(t)} \\
&= \frac{\dot{\mu}(t)}{\mu(t)}\,\Big(x(t) - \sigma(t)\,\epsilon\Big) + \dot{\sigma}(t)\,\epsilon \\
&= \dot{\mu}(t)\,\bar{x} + \dot{\sigma}(t)\,\epsilon
\end{aligned} \tag{72}$$

and thus corresponds to the conditional vector field defined in flow matching (Lipman et al., 2023).

Marginalizing (72) with respect to

$$p_{\text{data}}(\bar{x} \mid x) = \frac{\rho(x \mid \bar{x}, t)\,p_{\text{data}}(\bar{x})}{\bar{\rho}(x, t)} \tag{73}$$

and substituting (19) and applying (71), we obtain

$$\begin{aligned}
v(x, t) &= \int_\Omega \Big(-f(t)\,x(t) + \frac{1}{2}\,g(t)^2\,\nabla_x \log \rho(x \mid \bar{x}, t)\Big)\,\frac{\rho(x \mid \bar{x}, t)\,p_{\text{data}}(\bar{x})}{\bar{\rho}(x, t)}\,d\bar{x} \\
&= -f(t)\,x(t) + \frac{1}{2}\,g(t)^2 \int_\Omega \nabla_x \log \rho(x \mid \bar{x}, t)\,\frac{\rho(x \mid \bar{x}, t)\,p_{\text{data}}(\bar{x})}{\bar{\rho}(x, t)}\,d\bar{x} \\
&= -f(t)\,x(t) + \frac{1}{2}\,g(t)^2 \int_\Omega \frac{1}{\rho(x \mid \bar{x}, t)}\,\frac{\rho(x \mid \bar{x}, t)\,p_{\text{data}}(\bar{x})}{\bar{\rho}(x, t)}\,\nabla_x \rho(x \mid \bar{x}, t)\,d\bar{x} \\
&= -f(t)\,x(t) + \frac{1}{2}\,g(t)^2\,\frac{1}{\bar{\rho}(x, t)}\,\nabla_x \bar{\rho}(x, t) \\
&= -f(t)\,x(t) + \frac{1}{2}\,g(t)^2\,\frac{1}{\bar{\rho}(x, t)}\,\nabla_x \log \bar{\rho}(x, t)
\end{aligned} \tag{74}$$

and thus corresponds to the marginal probability flow ODE 17.

A.5 Proof of Proposition 5

Proof. Based on the result of Proposition 4 and using (12), we can express the homotopy matching problem

$$\frac{\partial \rho_\Phi(x, t)}{dt} = \frac{\partial \bar{\rho}(x, t)}{dt} \tag{75}$$

equivalently as

$$\nabla_x \cdot \Big(\rho_\Phi\,\nabla_x \Phi(x, t)\Big) = \nabla_x \cdot \Big(\rho_\Phi\,\Big(-f(t)\,x(t) + \frac{1}{2}\,g(t)^2\,\nabla_x \log \bar{\rho}(x, t)\Big)\Big) \tag{76}$$

Given that this matching holds identically, we have

$$\nabla_x \Phi(x, t) = -f(t)\,x(t) + \frac{1}{2}\,g(t)^2\,\nabla_x \log \bar{\rho}(x, t) \tag{77}$$

Furthermore, given that both the forward-time ODE and SDE of Song et al. (2021) exhibit the same marginal probability density $\bar{\rho}(x, t)$, it is shown that they satisfy the following reverse-time SDE:

$$dx(\tau) = \Big(f(\tau)\,x(\tau) - g(\tau)^2\,\nabla_x \log \bar{\rho}(x, t)\Big)\,d\tau + g(\tau)\,dW(\tau) \tag{78}$$

which reverses the diffusion process as outlined by Anderson (1982) and Song et al. (2021). Applying the change of variable $dt = -d\tau$, this reverse-time SDE can similarly be written in terms of $t = 1 - \tau$ as

$$dx(t) = -\Big(f(t)\,x(t) - g(t)^2\,\nabla_x \log \bar{\rho}(x,t)\Big)\,dt + g(t)\,dW(t) \tag{79}$$

where $dW(t)$ does not change sign, since the Wiener process is invariant under time reversal.

Subsequently, the Fokker-Plank dynamic that governs the time evolution of the marginal density homotopy $\bar{\rho}(x,t)$ is given by

$$\frac{\partial \bar{\rho}(x,t)}{\partial t} = -\,\nabla_x \cdot \Bigg(\bar{\rho}(x,t)\,\Big(-\,f(t)\,x(t) + g(t)^2\,\nabla_x \log \bar{\rho}(x,t)\Big)\Bigg) + \frac{1}{2}\,g(t)^2\,\Delta_x\,\bar{\rho}(x,t) \tag{80}$$

where $\Delta_x = \nabla_x \cdot \nabla_x$ denotes the Laplacian. By substituting (77) into this Fokker-Plank equation, we then have

$$\frac{\partial \bar{\rho}(x,t)}{\partial t} = -\,\nabla_x \cdot \Bigg(\bar{\rho}(x,t)\,\Big(2\,\nabla_x \Phi(x,t) + f(t)\,x(t)\Big)\Bigg) + \frac{1}{2}\,g(t)^2\,\Delta_x\,\bar{\rho}(x,t) \tag{81}$$

At equilibrium $\frac{\partial \bar{\rho}(x,t)}{\partial t} = 0$, the Fokker-Planck equation admits a unique normalized steady-state solution, given by the Boltzmann distribution:

$$p_B(x) \propto \exp\left(\frac{2}{g_\infty^2}\left(2\,\Phi_\infty(x) + \frac{f_\infty}{2}\,x(t)^T x(t)\right)\right) \tag{82}$$

when the potential energy function, the drift coefficient, and the diffusion coefficient reach their time-independent steady states $\Phi_\infty(x)$, f_∞ and g_∞ at equilibrium. The Boltzmann distribution can then be written in terms of a coherent Boltzmann energy Φ_B considered in EBMs, as follows:

$$p_B(x) = \frac{e^{\Phi_B}}{Z} \tag{83}$$

where

$$\Phi_B(x) = \frac{4\,\Phi_\infty(x) + f_\infty\,\|x\|^2}{g_\infty^2} \tag{84}$$

□

and $Z = \int_\Omega e^{\Phi_B(x)}\,dx$ is the normalizing constant.

A.6 Proof of Proposition 6

Proof. The variational loss function in (22) can be written as follows:

$$\mathcal{L}(\Phi,t) = \frac{1}{2}\,\mathbb{E}_{\rho(x|\bar{x},t)\,p_{\text{data}}(\bar{x})}\Big[\Phi(x)\,\big(\gamma(x,\bar{x},t) - \bar{\gamma}(x,\bar{x},t)\big)\Big] + \frac{1}{2}\,\mathbb{E}_{\bar{\rho}(x,t)}\,\Big[\big\|\nabla_x \Phi(x)\big\|^2\Big] \tag{85}$$

where we have assumed, without loss of generality, that a normalized energy $\bar{E}_\theta(x,t) = 0$. For an unnormalized solution $\Phi(x)$, we can always obtain a normalization by subtracting its mean.

The optimal solution Φ of the functional (85) is given by the first-order optimality condition:

$$\mathcal{I}(\Phi,\Psi) = \frac{d}{d\epsilon}\,\mathcal{L}(\Phi(x) + \epsilon\Psi(x),t)\bigg|_{\epsilon=0} = 0 \tag{86}$$

which must hold for all trial function Ψ.

Taking the variational derivative of the particle flow objective (86) with respect to ϵ, we have:

$$
\begin{aligned}
\mathcal{I}(\Phi, \Psi) &= \frac{d}{d\epsilon}\, \mathcal{L}(\Phi + \epsilon\Psi)\bigg|_{\epsilon=0} \\
&= \frac{1}{2}\int_{\Omega\times\Omega} p_{\text{data}}(\bar{x})\, \rho(x \mid \bar{x})\, \big(\gamma(x, \bar{x}, t) - \bar{\gamma}(x, \bar{x}, t)\big)\, \frac{d}{d\epsilon}(\Phi + \epsilon\Psi)\; d\bar{x}\, dx \\
&\quad + \frac{1}{2}\int_{\Omega} \bar{\rho}(x)\, \frac{d}{d\epsilon}\big\|\nabla_x(\Phi + \epsilon\Psi)\big\|^2\, dx \\
&= \frac{1}{2}\int_{\Omega\times\Omega} p_{\text{data}}(\bar{x})\, \rho(x \mid \bar{x})\, \big(\gamma(x, \bar{x}, t) - \bar{\gamma}(x, \bar{x}, t)\big)\, \Psi\; d\bar{x}\, dx \;+\; \int_{\Omega} \bar{\rho}(x)\, \nabla_x\Phi \cdot \nabla_x\Psi\, dx
\end{aligned}
\tag{87}
$$

Given that $\Phi \in \mathcal{H}^1_0(\Omega; p)$, its value vanishes on the boundary $\partial\Omega$. Therefore, the second summand of the last expression in (87) can be written, via multivariate integration by parts, as

$$
\int_{\Omega} \bar{\rho}(x)\, \nabla_x\Phi \cdot \nabla_x\Psi = -\int_{\Omega} \nabla_x \cdot \big(\bar{\rho}(x)\, \nabla_x\Phi\big)\, \Psi\, dx \tag{88}
$$

By substituting (88) into (87), we get

$$
\mathcal{I}(\Phi, \Psi) = \int_{\Omega} \left(\frac{1}{2}\int_{\Omega} p_{\text{data}}(\bar{x})\, \rho(x \mid \bar{x})\, \big(\gamma(x, \bar{x}, t) - \bar{\gamma}(x, \bar{x}, t)\big)\, d\bar{x} \;-\; \int_{\Omega} \nabla_x \cdot \big(\bar{\rho}(x)\, \nabla_x\Phi\big) \right) \Psi\, dx \tag{89}
$$

and equating it to zero, we obtain the weak formulation (21) of the density-weighted Poisson's equation.

Given that the Poincaré inequality (23) holds, (Laugesen et al., 2015, Theorem 2.2) presents a rigorous proof of existence and uniqueness for the solution of the weak formulation (21), based on the Hilbert-space form of the Riesz representation theorem. □

B Experimental Details

B.1 Model architecture

Our network architectures for the autonomous and time-varying VPFB models are based on the WideResNet (Zagoruyko & Komodakis, 2016) and the U-Net (Ronneberger et al., 2015), respectively. For WideResNet, we include a spectral regularization loss for model training to penalize the spectral norm of the convolutional layer. Also, we apply weight normalization with data-dependent initialization (Salimans & Kingma, 2016) on the convolutional layers to regularize model output. Our WideResNet architecture adopts the model hyperparameters reported by Xiao et al. (2021). For U-Net, we remove the final scale-by-sigma operation (Kim et al., 2021; Song et al., 2021) and replace it with the Euclidean norm $\frac{1}{2}\|x - f_\theta(x)\|^2$ between the input $x(t)$ and the output of the U-Net $f_\theta(x)$. Our U-Net architecture adopts the same model hyperparameters used by Lipman et al. (2023). Here, we replace the LeakyReLU activations with Gaussian Error Linear Unit (GELU) activations (Hendrycks & Gimpel, 2017) in both the WideResNet and U-Net models, which we found improves training stability and convergence.

B.2 Training

We use the Lamb optimizer (You et al., 2020) and a learning rate of 0.001 for all the experiments. We find that Lamb performs better than Adam over large learning rates. We use a batch size of 128 and 64 for training CIFAR-10 and CelebA, respectively. For all experiments, we set a spectral gap constant of $\eta = 0.0001$, a cutoff time of $t_{\max} = 0.9$, and a terminal time of $t_{\text{end}} = 2$ during training. Here, the mean and standard deviation scheduling functions $\mu(t) = t$ and $\sigma(t) = 1 - t$ follow those defined by the OT-FM path. All models are trained on a single NVIDIA A100 (80GB) GPU until the FID scores, computed on 2,500 samples, no longer show improvement. We observe that the models converge within 800k training iterations.

Algorithm 1 VPFB Training

Input: Initial energy model Φ_θ, spectral gap constant η, mean and standard deviation scheduling functions $\mu(t)$ and $\sigma(t)$, cutoff time $t_{\max}$, terminal time t_{end}, and batch size B.
repeat
  Sample observed data $\bar{x}_i \sim p_{\text{data}}(\bar{x})$, $t_i \sim \mathcal{U}(0, t_{\text{end}})$, and $\epsilon_i \sim \mathcal{N}(0, I)$
  if $t_i < t_{\max}$ **then**
    Sample $x_i \sim \rho(x \mid \bar{x}, t_i)$ via reparameterization $x_i = \mu(t_i)\,\bar{x}_i + \sigma(t_i)\,\epsilon_i$
  else
    Sample $x_i \sim \rho(x \mid \bar{x}, t_{\max})$ via reparameterization $x_i = \mu(t_{\max})\,\bar{x}_i + \sigma(t_{\max})\,\epsilon_i$
  end if
  Compute gradient $\nabla_x \Phi_\theta(x_i, t_i)$ w.r.t. x_i via backpropagation
  Calculate VPFB loss $\frac{1}{B}\sum_{i=1}^{B} \mathcal{L}(\Phi_\theta, t_i)$
  Backpropagate and update model parameters θ
until FID converges

B.3 Numerical Solver

In our experiments, the default solver of ODEs used is the black box solver in the Scipy library with the RK45 method (Dormand & Prince, 1980), following Xu et al. (2022). Here, we allow additional ODE iterations to further refine the samples within regions of high likelihood, which we observe that it improves the quality of generated images. We achieve this by extending the ODE time horizon from $[0, t_{\max}]$ to the $[0, t_{\text{end}}]$. Our observations indicate that setting $t_{\max} = 0.99$ and $t_{\text{end}} = 1.625$ yields the best generation results.

B.4 Datasets

We use the CIFAR-10 (Krizhevsky, 2009) and CelebA (Liu et al., 2015) datasets for our experiments CIFAR-10 is of resolution 32×32 and contains $50,000$ training images and $10,000$ test images. CelebA contains $202,599$ face images, of which $162,770$ are training images and $19,962$ are test images. For processing, we first clip each image to 178×178 and then resize it to 64×64. For processing, we first crop each image to a square image whose side is of length which is the minimum of the height and weight, and then we resize it to 64×64 or 128×128. For resizing, we set the anti-alias to True. We applied random horizontal flipping as a data augmentation technique for our datasets.

B.5 Quantitative Evaluation

We employ the FID and inception scores as quantitative evaluation metrics for assessing the quality of generated samples. For CIFAR-10, we compute the FID between $50,000$ samples and the pre-computed statistics on the training set Heusel et al. (2017). For CelebA 64×64, we follow the setting of Song & Ermon (2020) where the distance is computed between $5,000$ samples and the pre-computed statistics on the test set. For model selection, we follow Song et al. (2021) and pick the checkpoint with the smallest FID scores, computed on $2,500$ samples every $10,000$ iteration.

C Additional Results

Figures 7 and 8 show additional examples of image interpolation on CIFAR-10 and CelebA 64×64, respectively. Figures 9 and 10 show additional uncurated examples of unconditional image generation on CIFAR-10 and CelebA 64×64, respectively.

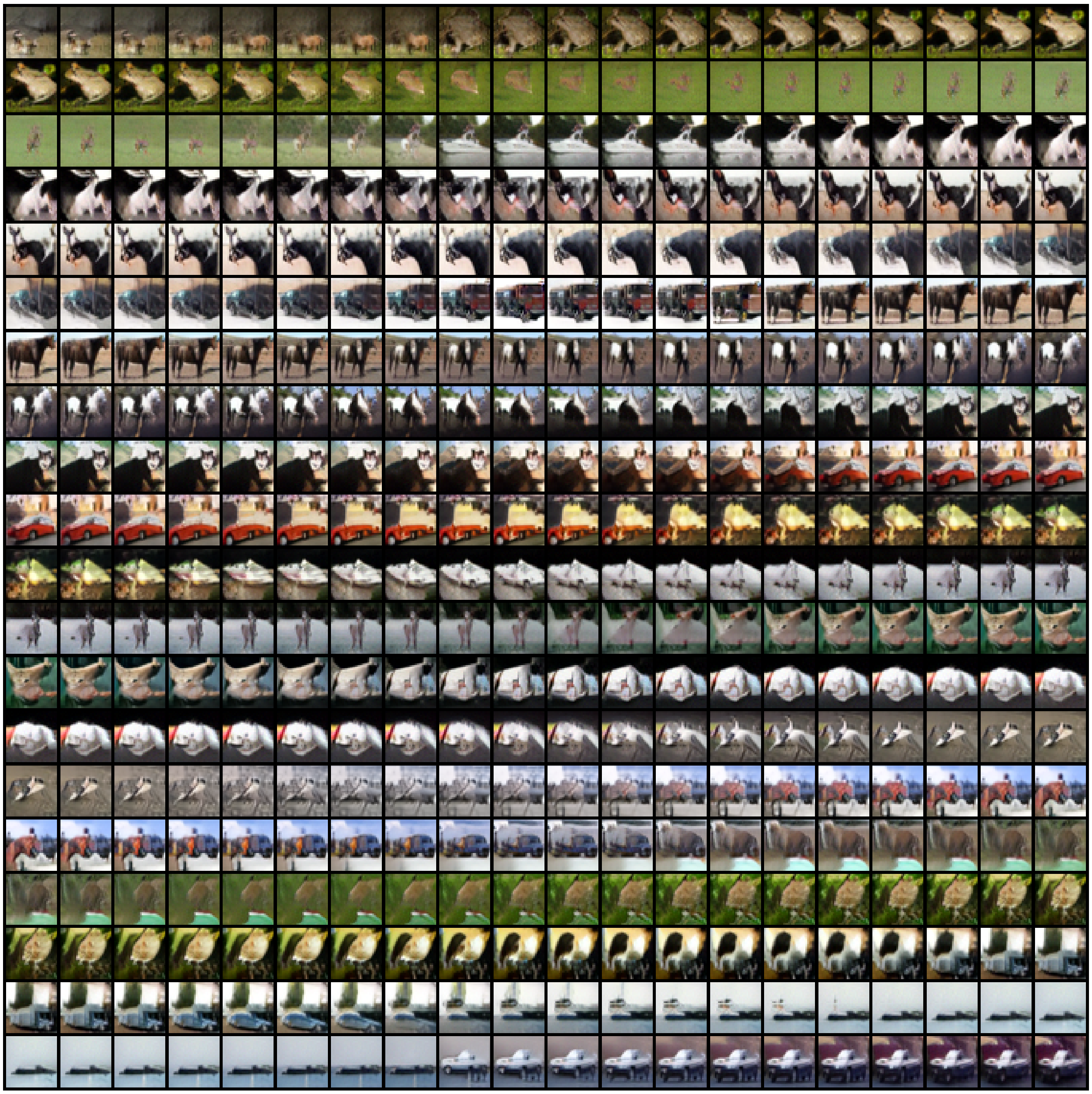

Figure 7: Additional interpolation results on unconditional CIFAR-10 32×32.

Figure 8: Additional interpolation results on unconditional CelebA 64×64.

Figure 9: Additional uncurated samples on unconditional CIFAR-10 32×32.

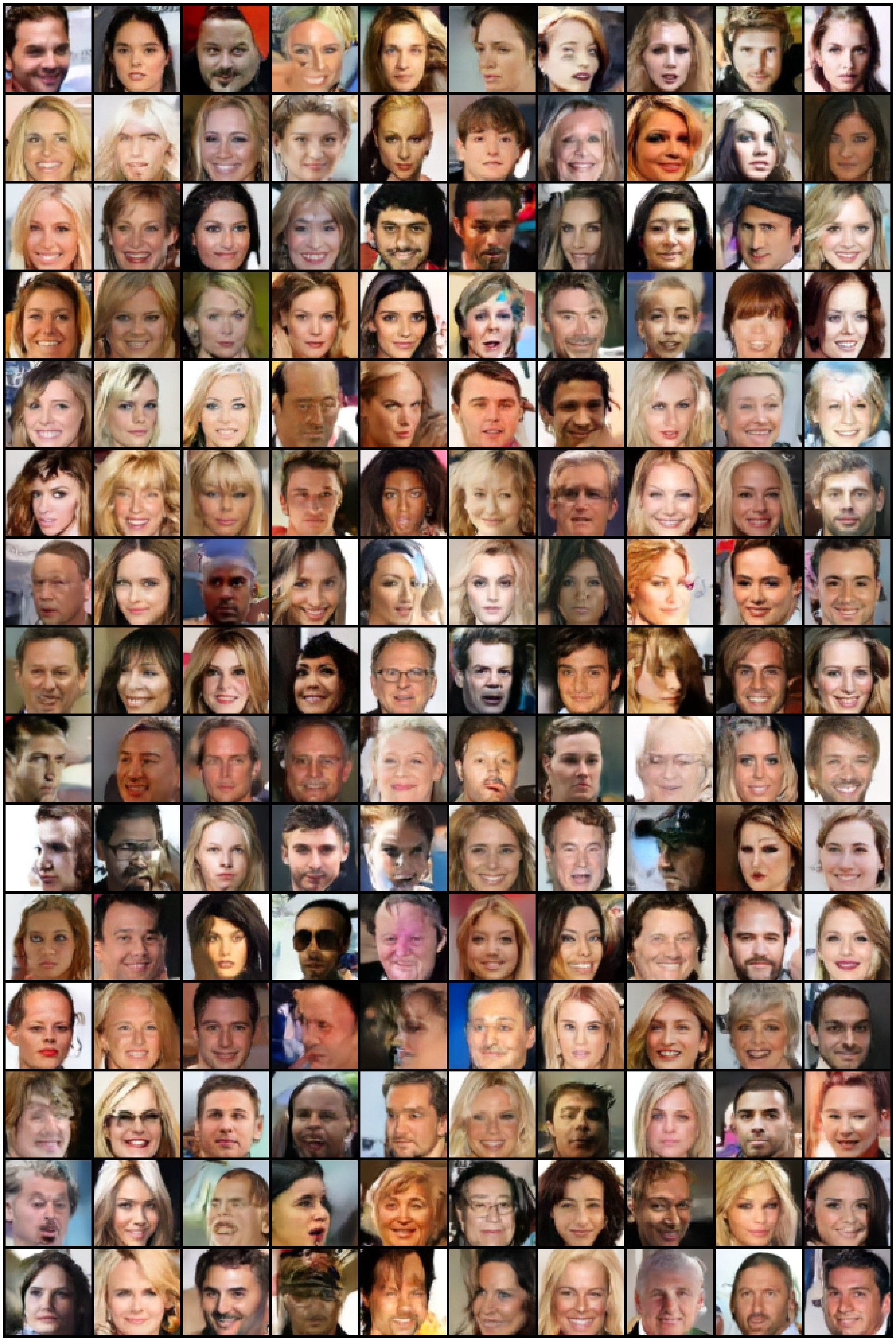

Figure 10: Additional uncurated samples on unconditional CelebA 64 × 64.