Learning Latent Space Hierarchical EBM Diffusion Models

Jiali Cui [1] **Tian Han** [1]

Abstract

This work studies the learning problem of the energy-based prior model and the multi-layer generator model. The multi-layer generator model, which contains multiple layers of latent variables organized in a top-down hierarchical structure, typically assumes the Gaussian prior model. Such a prior model can be limited in modelling expressivity, which results in a gap between the generator posterior and the prior model, known as the prior hole problem. Recent works have explored learning the energy-based (EBM) prior model as a second-stage, complementary model to bridge the gap. However, the EBM defined on a multi-layer latent space can be highly multi-modal, which makes sampling from such marginal EBM prior challenging in practice, resulting in ineffectively learned EBM. To tackle the challenge, we propose to leverage the diffusion probabilistic scheme to mitigate the burden of EBM sampling and thus facilitate EBM learning. Our extensive experiments demonstrate a superior performance of our diffusion-learned EBM prior on various challenging tasks.

1. Introduction

The hierarchical generative model with multiple layers of latent variables (a.k.a., *multi-layer generator model*) has made significant progress in learning complex data distribution (Sø nderby et al., 2016; Vahdat & Kautz, 2020) and has garnered particular interest for its top-down hierarchical structure, where multiple layers of latent variables that are organized from the top to the bottom layers tend to capture levels of (hierarchical) data representations, with high-level semantic representations captured by the latent variables at the top layers and low-level detail representations by those at the bottom layers (Maaløe et al., 2019; Child, 2020). Learning such hierarchical representation can be essential and crucial to various downstream applications (Havtorn et al., 2021; Nijkamp et al., 2020b). However, such multi-layer generator models typically assume the Gaussian prior model, which can be limited in statistical expressivity by primarily focusing on the inter-layer relation between layers of latent variables while largely ignoring the intra-layer relation between latent units within each layer (Cui et al., 2023a;b). This may result in the *prior hole problem* (Rosca et al., 2018; Hoffman & Johnson, 2016; Takahashi et al., 2019; Bauer & Mnih, 2019) where the non-expressive Gaussian prior fails to match the aggregated generator posterior.

Recent studies (Aneja et al., 2021; Cui et al., 2023a) have investigated the utilization of the energy-based (EBM) prior model as a complementary model to address this limitation. The EBM prior is typically trained with a fixed generator model (referred to as the *Two-Stage* learning scheme) to tilt the non-expressive Gaussian prior to match the generator posterior. However, learning a *single* (marginal) EBM is challenging because the generator posterior is often multi-modal, and more importantly, the Markov Chain Monte Carlo (MCMC) sampling required to maximize the marginal EBM likelihood can be difficult, as multiple layers of latent variables are interwoven and require exploration at different latent scales. In addition, MCMC sampling, such as Langevin dynamics, usually starts from a noise-initialized point, which is hard to explore the energy landscape and mix between different local modes. Therefore, for multi-layer latent variables, EBM prior sampling may serve as the bottleneck for effective EBM learning, which still poses a challenge.

Inspired by recent diffusion probabilistic frameworks (Ho et al., 2020; Gao et al., 2020; Zhu et al., 2023; Du et al., 2023), we propose learning the EBM prior of multi-layer latent variables in a diffusion learning scheme. We construct a series of conditional EBMs prior to gradually matching the highly multi-modal generator posterior, with each conditional EBM prior only matching perturbed samples at each step. Compared to marginal EBM prior, such a conditional EBM prior can be less multi-modal, leading to more tractable conditional likelihood learning. For EBM sampling, the proposed conditional EBM prior can render a smoother energy landscape, which mitigates the burden of MCMC sampling and thus further facilitates effective EBM learning. However, for multi-layer latent variables,

[1]Department of Computer Science, Stevens Institute of Technology. Correspondence to: Tian Han <than6@stevens.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

MCMC sampling needs to account for their different latent scales at different layers (i.e., the scales of latent variables at the top and bottom layers can be very different); moreover, directly perturbating the latent samples may destroy their inter-layer relation (i.e., conditional dependency formulated in the hierarchical structure). Therefore, we further employ a uni-scale $\tilde{\mathbf{u}}$-space (see definition in Sec. 3.1) converted from the multi-scale latent space, which allows us to preserve the hierarchical dependency along the forward process while at the same time, further reducing the burden of MCMC sampling by traversing a uni-scale latent space. Our experiments demonstrate the effectiveness of the proposed method in various challenging tasks and show that our model is capable of generating high-quality samples and capturing hierarchical representations at different layers.

Contribution: 1) We develop a learning framework that incorporates the diffusion probabilistic scheme for learning the joint EBM prior for the multi-layer generator model; 2) To preserve hierarchical structures and enable more effective EBM sampling, we adopt a uni-scale space to further mitigate the burden of MCMC sampling; 3) We conduct various experiments to examine our model in generating high-quality samples and learning effective hierarchical representations.

2. Preliminary

2.1. Multi-layer Latent Variable Model

Let $\mathbf{x} \in R^D$ be the high-dimensional observed example and $\mathbf{z} \in R^d$ be the low-dimensional latent variable. The multi-layer generator model contains multiple latent variables (i.e., $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_L$) organized in a top-down hierarchical structure and can be specified as a joint distribution. We denote $\tilde{\mathbf{z}} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_L)$, then

$$\begin{aligned} p_\beta(\mathbf{x}, \tilde{\mathbf{z}}) =& p_{\beta_0}(\mathbf{x}|\tilde{\mathbf{z}}) p_{\beta_{>0}} p(\tilde{\mathbf{z}}) \quad \text{where} \\ p_{\beta_{>0}}(\tilde{\mathbf{z}}) =& \prod_{i=1}^{L-1} p_{\beta_i}(\mathbf{z}_i|\mathbf{z}_{i+1}) p(\mathbf{z}_L) \end{aligned} \tag{1}$$

in which $p_{\beta_0}(\mathbf{x}|\tilde{\mathbf{z}})$ is the generation model that maps from the latent space to the data space, and $p_{\beta_{>0}}(\tilde{\mathbf{z}})$ is the prior model that factories consecutive layers of latent variables with conditional Gaussian distribution (i.e., $p_{\beta_i}(\mathbf{z}_i|\mathbf{z}_{i+1}) \sim \mathcal{N}(\mu_{\beta_i}(\mathbf{z}_{i+1}), \sigma^2_{\beta_i}(\mathbf{z}_{i+1}))$) parameterized by learnable parameter β_i. The $p(\mathbf{z}_L) \sim \mathcal{N}(0, \mathbf{I}_d)$ is assumed to be unit Gaussian at the top layer.

Learning such a hierarchical generative model can be achieved using maximum likelihood estimation (MLE) with the gradient estimated as $\nabla_\beta \mathbb{E}_{p_\beta(\tilde{\mathbf{z}}|\mathbf{x})}[\log p_\beta(\mathbf{x}, \tilde{\mathbf{z}})]$. For the generator posterior $p_\beta(\tilde{\mathbf{z}}|\mathbf{x})$, prior works utilize MCMC sampling to obtain approximated posterior samples (Nijkamp et al., 2020b), and (Sø nderby et al., 2016; Child, 2020; Maaløe et al., 2019) propose the variational learning that introduces a parameterized inference network (e.g., $q_\phi(\tilde{\mathbf{z}}|\mathbf{x})$) learned to approximate the generator posterior distribution.

However, such multi-layer generator models often fall short in generating high-quality image synthesis, as the Gaussian prior typically only focuses on the *inter-layer* relation modelling while largely ignoring the *intra-layer* relation modelling (Cui et al., 2023a), resulting in the *prior hole problem* with mismatch regions between the prior and aggregate posterior distribution (Dai & Wipf, 2019; Ghosh et al., 2019).

2.2. Energy-based Prior Model.

Another generative model, the energy-based model (EBM), is shown to be expressive in capturing the intra-layer relation and representing data uncertainty. In general, on data space $\mathbf{x}$, the EBM can be defined as

$$p_\omega(\mathbf{x}) = \frac{1}{\mathrm{Z}_\omega} \exp\left[f_\omega(\mathbf{x})\right] \tag{2}$$

where Z_ω is the normalizing constant or partition function, $f_\omega(\mathbf{x})$ is the energy function parameterized with ω.

Learning the EBM via MLE estimates the gradient as $\mathbb{E}_{p_{\text{data}}(\mathbf{x})}[\nabla_\omega f_\omega(\mathbf{x}) - \mathbb{E}_{p_\omega(\mathbf{x})}[f_\omega(\mathbf{x})]]$. For EBM samples from $p_\omega(\mathbf{x})$, (Du & Mordatch, 2019; Du et al., 2020; Nijkamp et al., 2019) adopt MCMC sampling such as Langevin dynamics (LD). In particular, it is applied as

$$\mathbf{x}_{\tau+1} = \mathbf{x}_\tau + s \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau) + \sqrt{2s} U_\tau \tag{3}$$

where τ indexes the time step, s is the step size and $U_\tau \sim \mathcal{N}(0, \mathbf{I}_D)$, and $\mathbf{x}_{\tau=0}$ is usually initialized from the Gaussian noise. However, in practice, it may take a long time to explore the energy landscape and mix between different local modes. To mitigate the burden of EBM sampling, recent advances have explored EBMs on low-dimensional latent space $p_\omega(\mathbf{z})$ (Pang et al., 2020a; Xiao & Han, 2022; Yu et al., 2022), but a single-layer $p_\omega(\mathbf{z})$ prior model can still be limited in modelling compacity of the whole model.

Two-stage complementary EBM prior. Learning the EBM prior for multi-layer of latent variables $\tilde{\mathbf{z}} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_L)$ can be more expressive than single-layer latent variables, but jointly learning both the multi-layer generator model and EBM prior can be extremely inefficient, especially with a deep hierarchical structure involved (Vahdat & Kautz, 2020; Child, 2020). This motivates a *Two-Stage* learning scheme (Xiao et al., 2020; Aneja et al., 2021; Cui et al., 2023a) that learns the Gaussian prior generator model at the *first stage* (see Sec. 2.1) and then learns the EBM, as a complementary model at the *second stage* with the fixed generator backbone. In our work, we adopt such a learning scheme for its efficiency.

With multi-layer of latent variables, the NCP-VAE (Aneja et al., 2021) factories a conditional EBM prior

$$p_{\omega_i,\beta_i}(\mathbf{z}_i|\mathbf{z}_{i+1}) = \frac{1}{\mathrm{Z}_{\omega_i,\beta_i}(\mathbf{z}_{i+1})} \exp\left[f_{\omega_i}(\mathbf{z}_i)\right] p_{\beta_i}(\mathbf{z}_i|\mathbf{z}_{i+1})$$

which aims to tilt the Gaussian prior toward the generator posterior distribution. The noise-contrastive estimation (NCE) is used for learning, which treats EBM as a *classifier* and thus does not need MCMC approximation. However, the NCE scheme can render *suboptimal learning* with a large gap between the two distributions (Xiao & Han, 2022), which exists between the generator posterior and the Gaussian prior (Arjovsky et al., 2017; Dai & Wipf, 2019).

The recent work (Cui et al., 2023a) considers learning EBM prior via MLE by jointly modelling all layers of latent variables

$$p_{\omega,\beta_{>0}}(\tilde{\mathbf{z}}) = \frac{1}{\mathrm{Z}_{\omega,\beta_{>0}}} \exp\left[F_\omega(\tilde{\mathbf{z}})\right] p_{\beta_{>0}}(\tilde{\mathbf{z}}) \tag{4}$$

where the energy function $F_\omega(\tilde{\mathbf{z}}) = \sum_{i=1}^{L} f_{\omega_i}(\mathbf{z}_i)$. However, MCMC sampling for Eqn. 4 can be practically challenging as layers of $\mathbf{z}_i$ can have different scales (i.e., $\mathbf{z}_L \sim \mathcal{N}(0, \mathbf{I}_d)$ and $\mathbf{z}_1 \sim p_{\beta_{>0}}(\mathbf{z}_1) = \int p_{\beta_{>0}}(\mathbf{z}_1, \ldots, \mathbf{z}_L) d\mathbf{z}_2, \ldots, d\mathbf{z}_L)$, which requires special designs for MCMC sampling to account for such variation.

Learning both the multi-layer EBM prior can be viewed to minimize the Kullback-Leibler (KL) divergence between the generator posterior distribution and the EBM prior, i.e., $\mathrm{KL}(p_\theta(\tilde{\mathbf{z}}|\mathbf{x})||p_{\omega,\beta_{>0}}(\tilde{\mathbf{z}}))$, which is difficult due to the highly multi-modal generator posterior and the multi-scale latent space, resulting in ineffective MCMC sampling for EBM learning.

3. Methodology

Inspired by recent diffusion probabilistic methods that focus on learning a sequence of parameterized models to gradually match target data distribution, we study a probabilistic framework that can leverage such diffusion scheme with a sequence of *conditional EBMs prior* for the *multi-layer generator models*.

3.1. Diffusion with Multi-layer Latent Variables

Attempt on $\tilde{\mathbf{z}}$-space. The diffusion probabilistic scheme assumes a sequence of perturbed samples $\mathbf{z}_{0:T} = (\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_T)$ for each diffusion step $t = 0, 1, \ldots, T$. In particular, the noisy sample $\tilde{\mathbf{z}}_t$ is generated by a pre-defined Gaussian perturbation kernel as

$$q(\tilde{\mathbf{z}}_{t+1}|\tilde{\mathbf{z}}_t) \sim \mathcal{N}(\alpha_{t+1}\tilde{\mathbf{z}}_t, \sigma_{t+1}^2 \mathbf{I}_{|d|}) \tag{5}$$

where α_{t+1} is typically set to be $\sqrt{1-\sigma_{t+1}^2}$ to ensure a spherical interpolation between samples and noise.

However, directly employing Eqn. 5 does not suit for multi-layer latent variables $\tilde{\mathbf{z}}$, since it does not take into account the hierarchical structure between layers of latent variables. Their inter-layer relation is consequently *destroyed* during the progress, i.e., each $\mathbf{z}_i$ becomes independently distributed as standard Gaussian noise at the final diffusion step. Our goal is to reach the Gaussian prior model $p_{\beta_{>0}}(\tilde{\mathbf{z}})$ (Eqn. 1) at the final step, such that the reverse process can start from the Gaussian prior model to approximate the generator posterior distribution gradually.

Toward $\tilde{\mathbf{u}}$-space. Instead of latent space, we formulate our diffusion model on $\tilde{\mathbf{u}}$-*space*. In particular, for multi-layer generator models, the Gaussian prior model $p_{\beta_{>0}}(\tilde{\mathbf{z}})$ is factorized to be the multiplication of consecutive layers of conditional Gaussian distribution $p_{\beta_i}(\mathbf{z}_i|\mathbf{z}_{i+1}) \sim \mathcal{N}(\mu_{\beta_i}(\mathbf{z}_{i+1}), \sigma_{\beta_i}^2(\mathbf{z}_{i+1}))$, which features the re-parametrization sampling, i.e., $\mathbf{z}_i = \mu_{\beta_i}(\mathbf{z}_{i+1}) + \sigma_{\beta_i}(\mathbf{z}_{i+1}) \cdot \mathbf{u}_{i+1}$ where $\mathbf{u}_{i+1} \sim \mathcal{N}(0, \mathbf{I}_d)$. We denote such an invertible deterministic transformation function to be $T_{\beta_{>0}}$, i.e., $\tilde{\mathbf{z}} = T_{\beta_{>0}}(\tilde{\mathbf{u}})$ and $\tilde{\mathbf{u}} = T_{\beta_{>0}}^{-1}(\tilde{\mathbf{z}})$ (see App. A.1), which allows us to adapt our diffusion model on $\tilde{\mathbf{u}}$-space. The $\tilde{\mathbf{u}}$-space enables the hierarchical structure to be maintained during the forward progress. When $q(\tilde{\mathbf{u}}_T) \sim \mathcal{N}(0, \mathbf{I}_{|d|})$ becomes standard Gaussian noise after the forward diffusion process, the corresponding latent variables reach the desired Gaussian prior model at the final step.

Forward on $\tilde{\mathbf{u}}$-space. The perturbation kernel is defined as

$$q(\tilde{\mathbf{u}}_{t+1}|\tilde{\mathbf{u}}_t) \sim \mathcal{N}(\alpha_{t+1}\tilde{\mathbf{u}}_t, \sigma_{t+1}^2 \mathbf{I}_d) \tag{6}$$

With a designed diffusion σ-schedule (e.g., signal-to-noise ratio, SNR), $\tilde{\mathbf{u}}_T$ at final step $t = T$ becomes Gaussian noise and $q(\tilde{\mathbf{u}}_T) = \mathcal{N}(0, \mathbf{I}_{|d|})$ is then the stationary distribution.

For the case of $t = 0$, we obtain $\tilde{\mathbf{u}}_0$ via transformation function $\tilde{\mathbf{u}}_0 = T_{\beta_{>0}}^{-1}(\tilde{\mathbf{z}}_0)$ with $\tilde{\mathbf{z}}_0$ sampled from the generator posterior. For variational-based generator models (see Sec. 2.1), $\tilde{\mathbf{z}}_0$ can be inferred from the inference model, while we can also perform MCMC posterior sampling for $\tilde{\mathbf{z}}_0$. For deep hierarchical structures, the inference model is preferred for efficiency. The forward trajectory then becomes

$$q_\beta(\tilde{\mathbf{u}}_{0:T}|\mathbf{x}) = \prod_{t=0}^{T-1} q(\tilde{\mathbf{u}}_{t+1}|\tilde{\mathbf{u}}_t) q_\beta(\tilde{\mathbf{u}}_0|\mathbf{x}) \tag{7}$$

where $q_\beta(\tilde{\mathbf{u}}_0|\mathbf{x})$ is the unknown target distribution. The goal is now to reverse from $\tilde{\mathbf{u}}_T$ to $\tilde{\mathbf{u}}_0$, which, in view of distributions, reverses from the Gaussian prior model $p_{\beta_{>0}}(\tilde{\mathbf{z}})$ to the generator posterior $p_\beta(\tilde{\mathbf{z}}|\mathbf{x})$. This strategy circumvents the problem of destroying hierarchical patterns during the forward process and thus satisfies our goal to bridge the gap between Gaussian prior model and the generator posterior.

3.2. Reverse with Multi-layer Latent Variables

$\tilde{\mathbf{u}}$-space for marginal EBM prior. The uni-scale $\tilde{\mathbf{u}}$-space is adopted in works (Xiao et al., 2020; Cui et al., 2023a) for EBM sampling. Specifically, sampling from the marginal EBM (Eqn. 4) is equivalent to sampling from

$$p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}) = \frac{1}{\mathrm{Z}_{\omega,\beta_{>0}}} \exp\left[F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}))\right] p_0(\tilde{\mathbf{u}}) \quad (8)$$

where $p_0(\tilde{\mathbf{u}})$ is standard Gaussian. The uni-scale $\tilde{\mathbf{u}}$-space can make EBM sampling easier than the multi-scale latent space $\tilde{\mathbf{z}}$. However, this is still challenging as it aims to match the highly multi-modal generator posterior and Gaussian prior with a single (marginal) EBM.

To tackle this challenge, prior works (Gao et al., 2020; Yu et al., 2022) leverage diffusion scheme and learn sequential conditional EBMs, which has seen some success in modelling the high-dimensional $\mathbf{x}$-space and single-layer $\mathbf{z}$-space. Inspired by their work, we propose to learn a sequence of conditional EBMs prior but focus on hierarchical generative models with $\tilde{\mathbf{u}}$-space to further alleviate the burden of EBM sampling and learning.

$\tilde{\mathbf{u}}$-space for conditional EBM prior. For our diffusion model, we formulate the marginal EBM prior (Eqn. 8) to a sequence of conditional EBMs prior, where each marginal EBM prior at each diffusion step (i.e., $p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t)$) is constrained by the forward generated $\tilde{\mathbf{u}}_{t+1}$, i.e.,

$$p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t+1}) \propto p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t) p(\tilde{\mathbf{u}}_{t+1}|\tilde{\mathbf{u}}_t) = \quad (9)$$
$$\frac{1}{\mathrm{Z}_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_{t+1})} \exp\left[F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t)\right] p_0(\tilde{\mathbf{u}}_t) \cdot p(\tilde{\mathbf{u}}_{t+1}|\tilde{\mathbf{u}}_t)$$

where we slightly abuse the notation and use $p(\tilde{\mathbf{u}}_{t+1}|\tilde{\mathbf{u}}_t)$ for the perturbation kernel as in Eqn. 6. The energy function essentially couples all layers of latent variables (i.e., $F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t) = \sum_{i=1}^{L} f_{\omega_i}(\mathbf{z}_i, t)$), and thus the inter-layer and intra-layer relation of each $\mathbf{z}_i$ can be effectively modelled. Each $f_{\omega_i}(\mathbf{z}_i, t)$ corresponding to $\mathbf{z}_i$ at each layer can also capture the representation of different layers.

Compared to Eqn. 8 that directly models the complex $\tilde{\mathbf{u}}_0$ (generator posterior) with a single marginal EBM, the conditional EBM only models $\tilde{\mathbf{u}}_t$, reversing step by step until $\tilde{\mathbf{u}}_0$. $p(\tilde{\mathbf{u}}_{t+1}|\tilde{\mathbf{u}}_t)$ from Gaussian noise perturbation kernel can serve to localize $\tilde{\mathbf{u}}_t$ to $\tilde{\mathbf{u}}_{t+1}$, making our conditional EBM less multi-modal and easier to be sampled than the marginal EBM. The reverse trajectory constitutes

$$p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_{0:T}) = \prod_{t=0}^{T-1} p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t+1}) p(\tilde{\mathbf{u}}_T) \quad (10)$$

where $p(\tilde{\mathbf{u}}_T) \sim N(0, \mathbf{I}_{|d|})$ is standard Gaussian as $q(\tilde{\mathbf{u}}_T)$ in forward trajectory.

EBM learning. The proposed method now contains a sequence of parameterized EBMs prior. With forward trajectory Eqn. 7, we minimize $\mathrm{KL}(q_\beta(\tilde{\mathbf{u}}_{0:T}|\mathbf{x})||p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_{0:T}))$ for EBM learning. The gradient is estimated as

$$\nabla_\omega \mathbb{E}_{q_{\beta_{>0}}(\tilde{\mathbf{u}}_{0:T}|\mathbf{x})}[\sum_{t=0}^{T-1} F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t) - \mathbb{E}_{p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_{0:T})}[\sum_{t=0}^{T-1} F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t)]] \quad (11)$$

which involves sampling from the whole forward and reverse trajectories. To provide more efficient sampling and learning, we follow the strategy used in (Ho et al., 2020) and utilize random diffusion steps t at each iteration of optimization. Thus, the learning gradient is

$$\mathbb{E}_{q_{\beta_{>0}}(\tilde{\mathbf{u}}_t, \tilde{\mathbf{u}}_{t+1}|\mathbf{x})}[\nabla_\omega F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t) - \mathbb{E}_{p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t+1})}[\nabla_\omega F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t)]] \quad (12)$$

where we only need perturbed sample $\tilde{\mathbf{u}}_t, \tilde{\mathbf{u}}_{t+1}$ from forward trajectory (see Eqn. 7) and prior samples $\tilde{\mathbf{u}}_t$ from the conditional EBM $p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t+1})$. To obtain prior samples, we perform Langevin dynamics (Eqn. 3) with the gradient

$$\nabla_{\tilde{\mathbf{u}}_t} \log p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t+1}) = \quad (13)$$
$$\nabla_{\tilde{\mathbf{u}}_t}[F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t) - \frac{||\tilde{\mathbf{u}}_t||^2}{2} - \frac{||\alpha_{t+1}\tilde{\mathbf{u}}_t - \tilde{\mathbf{u}}_{t+1}||^2}{2\sigma_{t+1}^2}]$$

Compared to MCMC sampling of marginal EBM (see Eqn. 8) that can be hard to mix between different local modes with noise initial points, MCMC sampling of conditional EBM can start from the given $\tilde{\mathbf{u}}_{t+1}$ and only needs to search for the local modes around $\tilde{\mathbf{u}}_{t+1}$. Specifically, the quadratic term is from the noise-aware term in Eqn. 9 and constrains the exploration of energy landscape to be localized around $\tilde{\mathbf{u}}_{t+1}$, which is much easier to obtain EBM samples than marginal EBMs. This conditional sampling can be more effective and efficient, which in turn benefits the learning of the proposed EBM prior.

The overall training algorithm and sampling process are shown in Alg. 1 and Alg. 2.

3.3. Coupling with symbol vector

In this section, we present the applicability of the proposed EBM prior. For multi-layer generator models, they are typically learned in an unsupervised scheme and thus are only feasible to generate random samples. However, the controllable and compositional ability to generate desired synthesis nowadays becomes a key requirement for many downstream tasks, yet it would be computationally expensive to *re-train* these models to fit the job.

The proposed method is flexible in coupling labels or attributes, which in turn empowers controllability and composition ability for the learned multi-layer generator models.

Specifically, we can adapt our model with given labels $\mathbf{y}$ as

$$p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t, \mathbf{y}|\tilde{\mathbf{u}}_{t+1}) = \frac{1}{\mathrm{Z}_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_{t+1})} \cdot \exp\left[\langle F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t), \mathbf{y}\rangle\right] p_0(\tilde{\mathbf{u}}_t) p(\tilde{\mathbf{u}}_{t+1}|\tilde{\mathbf{u}}_t) \tag{14}$$

where the energy function couples both $\mathbf{y}$ and $\tilde{\mathbf{u}}_t$, forming an associative memory that allows sampling $\tilde{\mathbf{u}}_t$ with a given $\mathbf{y}$. Generating images by such sampled $\tilde{\mathbf{u}}_t$ is known as the controllable generation. (Pang et al., 2020b; Yu et al., 2022) adopt similar formulations but only focus on single-layer latent space, while we focus on a multi-layer latent space with a hierarchical structure, allowing coupling $\mathbf{y}$ at different layers for specific tasks. We formulate such EBM prior at diffusion step $t = 0$ as the signal of $\mathbf{y}$ correlates strongly with clean $\tilde{\mathbf{u}}_0$ samples, while noisy $\tilde{\mathbf{u}}_{t>0}$ can be less correlated with $\mathbf{y}$.

The proposed EBM prior is capable of capturing inter-layer and intra-layer relations for different layers of latent variables, thus rendering better controllability and compositionality with effectively learned latent representations. We refer to learning derivation and details in App. A.2.

Algorithm 1 Learning EBM parameter ω

Require:
Training images $\mathbf{x}$; Number of learning iterations M; Hierarchical generator model β; Diffusion steps T; Langevin steps k;.

Let $m \leftarrow 0$, initialize EBM parameters ω.
repeat
  Sample $\tilde{\mathbf{u}}_0$: obtain $\tilde{\mathbf{z}} \sim p_\beta(\tilde{\mathbf{z}}|\mathbf{x})$ and $\tilde{\mathbf{u}}_0 = T_{\beta_{>0}}(\tilde{\mathbf{z}})$
  Diffusion step t: $t \sim U(0, T-1)$
  Noise sample: sample $\tilde{\mathbf{u}}_t^*$, $\tilde{\mathbf{u}}_{t+1}^*$ from Eqn. 7.
  Prior sample: sample $\tilde{\mathbf{u}}_t^k$ by Eqn. 3 and Eqn. 13 with $\tilde{\mathbf{u}}_{t+1}^*$ being the initialization
  Learn ω: Update ω with $\tilde{\mathbf{u}}_t^*$ and $\tilde{\mathbf{u}}_t^k$ using Eqn. 12
  Let $m \leftarrow m + 1$.
until $m = M$

Algorithm 2 Sampling and Image Synthesis

Require:
Diffusion steps T; Hierarchical generator model β; EBM prior ω.

Let $t \leftarrow T - 1$ and $\tilde{\mathbf{u}}_T \sim N(0, \mathbf{I}_{|d|})$.
repeat
  Sample $\tilde{\mathbf{u}}_t$: sample $\tilde{\mathbf{u}}_t \sim p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t+1})$ using Eqn. 3 and Eqn. 13
  Let $t \leftarrow t - 1$.
until $t = -1$
Sample $\tilde{\mathbf{z}}$: obtain $\tilde{\mathbf{z}} = T^{-1}_{\beta_{>0}}(\tilde{\mathbf{u}}_0)$.
Generate $\mathbf{x}$: generate $\mathbf{x} \sim p_{\beta_0}(\mathbf{x}|\tilde{\mathbf{z}})$ with obtained $\tilde{\mathbf{z}}$.

4. Related Work

Energy-based model. The EBM is expressive in representing the data uncertainty. Most existing works learn EBM on data space by maximizing EBM likelihood, which involves challenging MCMC sampling for the EBM. To tackle the challenge, (Du & Mordatch, 2019; Du et al., 2020) propose the use of a replay buffer, while (Cui & Han, 2023; Han et al., 2020b; Xie et al., 2021; 2018) consider learning a complementary generator model to jump-start MCMC sampling. (Xiao & Han, 2022; Pang et al., 2020a;b) focus on learning EBM prior on low-dimensional latent space to mitigate the burden of EBM sampling, but these works only deal with a single-layer latent space. We consider learning EBM prior for multi-layer latent variables, which enables hierarchical representation learning and better sample generation.

Hierarchical generative model. For hierarchical generative models (Vahdat & Kautz, 2020; Maaløe et al., 2019; Sø nderby et al., 2016), they typically assume a Gaussian prior model, which can be less informative, resulting in the prior hole problem and poor generation quality. Our work aims to learn effective EBM prior for hierarchical generative models, which is related to joint EBM prior (see Sec. 2.2). These works intend to learn a single (marginal) EBM on $\tilde{\mathbf{z}}$-space, while we study learning a sequence of conditional EBMs on $\tilde{\mathbf{u}}$-space. NCP-VAE (Aneja et al., 2021) discuss a *autoregressive*-style model, which can be *intractable* for MLE learning, as the normalization constant includes the top layer $\mathbf{z}_{i+1}$ and needs an additional inner loop for sampling. Our model leverages diffusion probabilistic models such that we can directly draw $\tilde{\mathbf{u}}_{t+1}$ through the forward trajectory. Compared to (Cui et al., 2023a) (Eqn. 4), our EBM prior only models $\tilde{\mathbf{u}}_t$ conditioned on the perturbed sample $\tilde{\mathbf{u}}_{t+1}$ generated by $\tilde{\mathbf{u}}_{t+1} = \alpha_{t+1}\tilde{\mathbf{u}}_t + \epsilon_{t+1}\sigma_{t+1}$, which can be much easier than directly recovering from $\tilde{\mathbf{u}}_T$ to $\tilde{\mathbf{u}}_0$ (if consider equivalent form Eqn. 8).

EBM diffusion model. Other diffusion EBMs (Gao et al., 2020; Yu et al., 2022; 2023) motivate EBM learning with a diffusion scheme. They build energy-based *recovery* model on data space and single-layer latent space, respectively. In this work, we focus on multi-layer generator model with a top-down hierarchical structure, which is shown to be capable of learning meaningful hierarchical representations (Child, 2020; Zhao et al., 2017; Havtorn et al., 2021). It is typically challenging for the diffusion scheme to maintain such hierarchical structures as the goal of the forward process is to destroy the data pattern, which includes the inter-layer relation (conditional dependency) among multi-layer latent variables. To preserve the hierarchical structure of latent variables, we conduct the forward and reverse processes on $\tilde{\mathbf{u}}$-space, which also leads to more effective learning and sampling than multi-scale $\mathbf{x}$-space (Gao et al., 2020) and $\mathbf{z}$-space (Yu et al., 2022).

Figure 1: Image synthesis on CelebA-HQ-256 (left), LSUN-Church-64 (center) and CIFAR-10 (right).

Table 1: IS(↑) and FID(↓) on CIFAR-10. Model* indicates our backbone model.

Method	IS	FID
Ours ($T = 3$)	9.03	8.93
Joint-EBM (Cui et al., 2023a)	8.99	11.34
DRL-EBM ($T = 6$) (Gao et al., 2020)	8.40	9.58
NCP-VAE (Aneja et al., 2021)	-	24.08
Hierarchical Generative Models w Gaussian Prior		
NVAE* (Vahdat & Kautz, 2020)	5.30	37.73
NVAE*-Recon	-	0.68
HVAE (Sø nderby et al., 2016)	-	81.44
BIVA (Maaløe et al., 2019)	-	66.37
Energy-based Models		
Architectural-EBM (Cui et al., 2023b)	-	63.42
Dual-MCMC (Cui & Han, 2023)	8.55	9.26
Adaptive-CE (Xiao & Han, 2022)	-	65.01
VAEBM (Xiao et al., 2020)	8.43	12.19
Hat EBM (Hill et al., 2022a)	-	19.15
ImprovedCD (Du et al., 2020)	7.85	25.1
Divergence Triangle (Han et al., 2020a)	-	30.10
Adv-EBM (Yin et al., 2020)	9.10	13.21
GANs+Score+Diffusion Models		
StyleGANv2 w/o ADA (Karras et al., 2020)	8.99	9.9
Diffusion-Amortized (Yu et al., 2023)	-	57.72
NCSN (Song & Ermon, 2019)	8.87	25.32
LSGM (Vahdat et al., 2021)	-	2.10
DDPM ($T = 1000$) (Ho et al., 2020)	9.46	3.17

Table 2: FID on CelebA-HQ-256 and LSUN-Church-64.

Method	CelebA-HQ-256	LSUN-Church-64
Ours ($T = 3$)	8.78	7.34
Joint-EBM (Cui et al., 2023a)	9.89	8.38
DRL-EBM ($T = 6$) (Gao et al., 2020)	-	7.04
NCP-VAE (Aneja et al., 2021)	24.79	-
NVAE* (Vahdat & Kautz, 2020)	30.25	38.13
NVAE*-Recon	1.64	2.45
Adv-EBM (Yin et al., 2020)	17.31	10.84
GLOW (Kingma & Dhariwal, 2018)	68.93	59.35
PGGAN (Karras et al., 2017)	8.03	6.42

5. Experiment

5.1. Image Synthesis

First, we examine the sample quality of our model. The proposed model is learned as a reverse approximation model in a diffusion probabilistic scheme, which allows sampling reverse steps $\tilde{\mathbf{u}}_t$ to reach $\tilde{\mathbf{u}}_0$. With sampled $\tilde{\mathbf{u}}_0$, we obtain $\tilde{\mathbf{z}}_0$ through the transformation function and then generate images $\mathbf{x}$ through the generator model (see Alg. 2).

Our project page is available at https://jcui1224.github.io/diffusion-hierarchical-ebm-proj/.

We assess our model on the standard benchmark CIFAR-10 and the challenging high-resolution CelebA-HQ-256 and large-scale LSUN-Church-64. We compare with our direct baseline model Joint-EBM (Cui et al., 2023a), NCP-VAE (Aneja et al., 2021) that learn signal (marginal) EBM prior for hierarchical generative models, and Diffusion Recovery (DRL) EBM (Gao et al., 2020) which learn EBM with diffusion probabilistic scheme on data space, as well as hierarchical generative model with the Gaussian prior and other powerful advanced generative models.

We recruit Fréchet Inception Distance (FID) and Inception Score (IS) metrics to evaluate the quality of image synthesis. We report our results in Tab. 1 and Tab. 2 as well as the FID score of the reconstructed images. It can be observed that our hierarchical EBM prior shows superior performance compared to our baseline models and can even be competitive with those powerful GANs and diffusion-based methods. For a fair comparison, we adopt the NVAE model (Vahdat & Kautz, 2020) as the backbone generator model as our direct baseline Joint-EBM and NCP-VAE. More quantitative and qualitative results can be found in the ablation studies and App. B.

5.2. Hierarchical Representation

Different from other generator models, the hierarchical generator model has an appealing structure, in which the latent variables at the top layers tend to learn high-level semantic representations, while low-level details representation can be learned by latent variables at lower layers. We examine our model in learning such hierarchical representations.

Hierarchical sampling. First, we demonstrate our model by performing *hierarchical sampling*, which generates variations of image synthesis that can visualize the different levels of data representation learned by different layers of latent variables. In particular, for multi-layer latent variables, we only generate random samples for some layers of latent variables while fixing other layers; hence, the corresponding variation of features is captured by the latent variables randomly sampled. We show visualization results on CelebA-HQ-256 in Fig. 2. It can be observed that by randomly sampling for the top layers of latent variables, the general structure (e.g., genders and face identities) would be changed, while for lower layers of latent variables, low-level

features (e.g., hair color, skin color) can vary correspondingly. We note that this is a challenging task where some minority of features are entangled across layers, but the majority of features can be successfully captured. This showcases the capability of our model in learning hierarchical representations.

Figure 2: Hierarchical sampling. Visualization of representations learned by latent variables from the top to bottom layers, arranged as top-left, top-right, bottom-left and bottom-right.

Out-of-distribution detection. Then, we evaluate our model in out-of-distribution (OOD) detection task to further demonstrate learned hierarchical representations. Typically, EBMs can be applied to the OOD task by computing the energy score as the decision function (Hill et al., 2022b; Cui & Han, 2023). For latent space EBMs, the inferred latent sample (i.e., $\tilde{\mathbf{u}}_0 \sim q_\beta(\tilde{\mathbf{u}}_0|\mathbf{x})$ for our case) from in-distribution (ID) data is usually assigned with a lower energy than from the OOD data. In this work, we compute the energy score of inferred latent samples at the top layers as the decision function for using high-level semantic representations learned at the top layers of latent variables to distinguish the OOD and the ID data (Havtorn et al., 2021).

To better leverage the diffusion probabilistic scheme, we propose conducting EBM sampling based on perturbed inference samples at the top layers, together with prior samples at the bottom layers. Specifically, with testing images $\mathbf{x}$ and layer index k, we first obtain $\tilde{\mathbf{u}}_1^{>k} \sim q_{\beta_{>0}}(\tilde{\mathbf{u}}_{0:T}|\mathbf{x})$ (see Eqn. 7) and $\tilde{\mathbf{u}}_1^{\leq k} \sim p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_{0:T})$ (see Eqn. 10) where we choose diffusion step $t = 1$ to ensure that only minor noise is added. Then, we perform reverse sampling conditioned on $\tilde{\mathbf{u}}_1^{>k}$ and $\tilde{\mathbf{u}}_1^{\leq k}$ for jointly sampling final $\tilde{\mathbf{u}}_0$, i.e., $\tilde{\mathbf{u}}_0 \sim p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_0|[\tilde{\mathbf{u}}_1^{>k}, \tilde{\mathbf{u}}_1^{\leq k}])$ where $[,]$ is the operation of concatenation. If $\tilde{\mathbf{u}}_1^{>k}$ is from ID data, then $\tilde{\mathbf{u}}_0^{>k}$ should render lower energy scores as both the $\tilde{\mathbf{u}}_1^{>k}$ and $\tilde{\mathbf{u}}_1^{\leq k}$ are from similar local modes of the learned energy landscape. If $\tilde{\mathbf{u}}_1^{>k}$ is from OOD data, then $\tilde{\mathbf{u}}_1^{>k}$ and $\tilde{\mathbf{u}}_1^{\leq k}$ can be in different modes, making the final reverse sampling difficult to traverse the energy landscape and thus rendering higher energy score of sampled $\tilde{\mathbf{u}}_0^{>k}$. We follow the standard protocol and evaluate by AUROC score for our EBM prior trained on CIFAR-10 with SVHN dataset being the OOD dataset. The result is reported in Fig. 3 where the performance of our model indeed improves as the layer of k increases, which agrees with the observation in (Havtorn et al., 2021) that our model can capture hierarchical representations at different layers. Compared to using inferred latent samples (inference scheme in Fig. 3), the proposed diffusion-based method (diffusion scheme in Fig. 3) can render better performance by conducting additional MCMC sampling of the learned EBM.

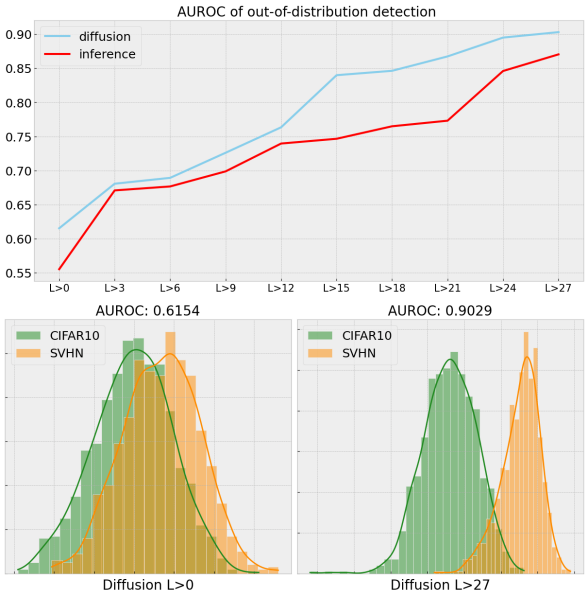

Figure 3: AUROC results for energy scores of different layers (denoted as $L > k$ for using top layers above k-th layer). **Top** figure visualizes the comparison between the diffusion scheme ($\tilde{\mathbf{u}}_0$ sampled from EBM) and the inference scheme ($\tilde{\mathbf{u}}_0$ inferred from inference model) in different layers. **Bottom** figure is the histogram of energy scores using all layers $L > 0$ and top layers $L > 27$. Total number of layers is 30.

5.3. Controllable Synthesis

For hierarchical generator models that are typically learned without labels (unsupervised learning), they can only generate random synthesis. Our diffusion EBM prior, as a flexible complementary model, can be coupled with labels to make hierarchical generative models more applicable in downstream tasks, such as controllable and compositional generation. In practice, we fix the hierarchical generative model and train our adapted model (Eqn. 14) with given $\mathbf{y}$ (supervised learning), and then we assess if our model can render controllability and compositionality with the captured hierarchical representations. We refer to App. A.2 for details of training and sampling.

Categorical labels. First, we evaluate our model with categorical label information, e.g., $\mathbf{y}$ represents labels of objects in CIFAR-10. To better examine our model, we only consider coupling $\mathbf{y}$ at the top layers by which the semantic representations are captured. With the learned model, we conduct controllable generation by specifying target labels

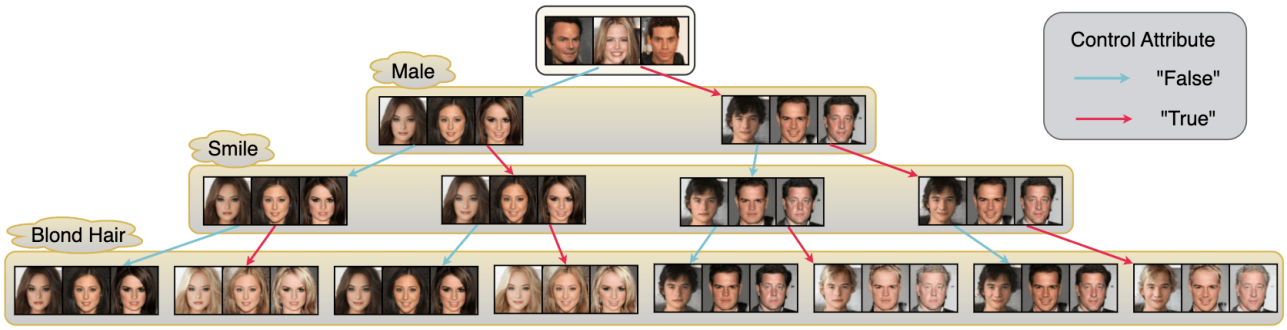

Figure 4: Fine-tuned image synthesis with multiple attributes on CelebA-64.

$\mathbf{y}$ and sampling corresponding $\tilde{\mathbf{u}}$ for generating the images. We show the results in Fig. 5, and it can be seen that our model can capture data representations and thus can generate synthesis with specific categories.

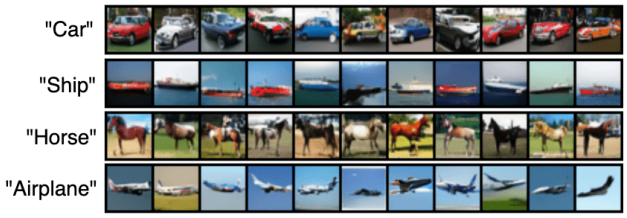

Figure 5: Controllable synthesis on CIFAR-10.

Multiple attributes. Then, we showcase our model by utilizing multiple data attributes for the challenging *fine-tuned* image synthesis. In particular, for example of CelebA-64 dataset, we can have multiple attributes $\mathbf{y}_{1:N}$ for different levels of data features. To better suit the hierarchical structure, we choose to couple high-level attributes, such as gender information, at the top layers and gradually couple lower-level attributes, such as face and hair features, at the lower layers. In Fig. 4, we first specify gender attributes (e.g., "Male") and sample $\tilde{\mathbf{u}}$ at the corresponding top layers for generating gender-specified images. Then, we fix these $\tilde{\mathbf{u}}$ of top layers and specify lower-level attributes (e.g., "Smile" and "Blond Hair") for sampling $\tilde{\mathbf{u}}$ at lower layers. We observe our EBM prior successfully renders the desired compositional synthesis by gradually adding specific features to the sampled images without changing the majority of previously fixed features. This suggests learned hierarchical representation of our EBM prior.

5.4. Langevin Transition for Energy Landscape

We examine the energy landscape of our learned EBM prior by visualizing the Langevin transition. Our conditional EBM prior should render a smooth energy landscape such that Langevin dynamics can effectively explore with a smooth Langevin transition. We visualize the corresponding image synthesis of the Langevin trajectory of each diffusion step, i.e., $t = T-1, \ldots, 1, 0$. We show in Fig. 7 for each step of the short-run Langevin dynamics (e.g., 30 Langevin steps) and in Fig. 6 for a challenging long-run setting (e.g., 300 Langevin steps for each diffusion step). We observe in Fig. 7 that the quality of image synthesis becomes better as the Langevin progresses, with large improvement at diffusion step $t = T-1$ and minor improvement at diffusion step $t = 0$; while, in Fig. 6, we do not see an *oversaturated problem* of EBM learning as observed in (Nijkamp et al., 2020a). We conduct such experiments to demonstrate a smooth energy landscape learned for our EBMs prior.

Figure 7: Short-run Langevin transition on CelebA-HQ-256. The top rows of figures show the transition for the diffusion step at $t = T-1$, and the bottom rows show the transition for the diffusion step at $t = 0$.

5.5. Ablation Studies

Diffusion step T. First, we train our diffusion-based EBM prior with more diffusion steps, e.g., $T = 6$. By doing so, our model should render better performance with easier EBM sampling by matching less perturbed samples at each

Figure 6: Long-run Langevin transition on CelebA-HQ-256. Visualization for each 30 steps.

step. We report the FID score and sampling time in Tab.3 where the synthesis quality indeed improves with more diffusion steps but also requires more sampling time. We thus report $T = 3$ as our result.

Table 3: Langevin steps K and diffusion steps T.

	$K = 30$	$K = 100$	$K = 50, T = 3$	$T = 6$
FID	9.98	8.13	**8.93**	8.13
Time (seconds)	75.17	166.34	94.23	193.45

Langevin step K. By using more Langevin steps, we should explore the energy landscape better and obtain more effective EBM samples for learning. The learned EBMs can thus generate high-quality samples for image synthesis. We show our results in Tab. 3 where using 50 steps (denoted as $K = 50$) delivers better synthesis than using 30 steps, while using 100 steps only shows a minor improvement but costs much more training and sampling time.

Figure 8: Images synthesis on CIFAR-10 with different backbones. Top-row images are generated with the NVAE (30 layers) backbone, and bottom-row images are generated with the HVAE (3 layers) backbone.

Other hierarchical generator models. In addition to the NVAE backbone model, we also train a simple backbone hierarchical VAE with 3-layer latent variables. We visualize images generated by the Gaussian prior and our model in Fig. 8, where our model still improves the quality of the generation in a large way (from FID 81.44 to 35.13), suggesting the effectiveness of the proposed method.

6. Limitation

In this work, the proposed method still renders inferior performance compared to state-of-the-art models (e.g., modern diffusion probabilistic models (Vahdat et al., 2021)). In addition, sampling from our EBM prior requires an iterative Langevin dynamics sampler, which can be further improved or even bypassed; we will consider it in our future studies.

7. Conclusion

We propose learning EBM prior for hierarchical generative model with a diffusion probabilistic scheme, which features more tractable conditional likelihood learning and more effective EBM sampling. We employ a uni-scale $\tilde{\mathbf{u}}$-space to maintain the hierarchical structure and further mitigate the burden of MCMC sampling. Such learned EBM prior can generate high-quality samples for image synthesis and can capture hierarchical representations for downstream tasks.

Impact Statement

This paper presents a generative probabilistic framework whose goal is to advance the field of Machine Learning and may share the limitations and negative impact as other advanced generative models. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Aneja, J., Schwing, A., Kautz, J., and Vahdat, A. A contrastive learning approach for training variational autoencoder priors. *Advances in neural information processing systems*, 34:480–493, 2021.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Bauer, M. and Mnih, A. Resampled priors for variational autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 66–75. PMLR, 2019.

Child, R. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

Cui, J. and Han, T. Learning energy-based model via dual-mcmc teaching. *Advances in Neural Information Processing Systems*, 36:28861–28872, 2023.

Cui, J., Wu, Y. N., and Han, T. Learning joint latent space ebm prior model for multi-layer generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3603–3612, June 2023a.

Cui, J., Wu, Y. N., and Han, T. Learning hierarchical features with joint latent space energy-based prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2218–2227, October 2023b.

Dai, B. and Wipf, D. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019.

Du, Y. and Mordatch, I. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019.

Du, Y., Li, S., Tenenbaum, J., and Mordatch, I. Improved contrastive divergence training of energy based models. *arXiv preprint arXiv:2012.01316*, 2020.

Du, Y., Durkan, C., Strudel, R., Tenenbaum, J. B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., and Grathwohl, W. S. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pp. 8489–8510. PMLR, 2023.

Gao, R., Song, Y., Poole, B., Wu, Y. N., and Kingma, D. P. Learning energy-based models by diffusion recovery likelihood. *arXiv preprint arXiv:2012.08125*, 2020.

Ghosh, P., Sajjadi, M. S., Vergari, A., Black, M., and Schölkopf, B. From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436*, 2019.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Norouzi, M., and Swersky, K. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

Han, T., Nijkamp, E., Zhou, L., Pang, B., Zhu, S.-C., and Wu, Y. N. Joint training of variational auto-encoder and latent energy-based model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020a.

Han, T., Nijkamp, E., Zhou, L., Pang, B., Zhu, S.-C., and Wu, Y. N. Joint training of variational auto-encoder and latent energy-based model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7978–7987, 2020b.

Havtorn, J. D. D., Frellsen, J., Hauberg, S., and Maaløe, L. Hierarchical vaes know what they don't know. In *International Conference on Machine Learning*, pp. 4117–4128. PMLR, 2021.

Hill, M., Nijkamp, E., Mitchell, J., Pang, B., and Zhu, S.-C. Learning probabilistic models from generator latent spaces with hat ebm. *Advances in Neural Information Processing Systems*, 35:928–940, 2022a.

Hill, M., Nijkamp, E., Mitchell, J., Pang, B., and Zhu, S.-C. Learning probabilistic models from generator latent spaces with hat ebm. *Advances in Neural Information Processing Systems*, 35:928–940, 2022b.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Hoffman, M. D. and Johnson, M. J. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, 2016.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.

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

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Maaløe, L., Fraccaro, M., Liévin, V., and Winther, O. Biva: A very deep hierarchy of latent variables for generative modeling. *Advances in neural information processing systems*, 32, 2019.

Nie, W., Vahdat, A., and Anandkumar, A. Controllable and compositional generation with latent-space energy-based models. *Advances in Neural Information Processing Systems*, 34:13497–13510, 2021.

Nijkamp, E., Hill, M., Zhu, S.-C., and Wu, Y. N. Learning non-convergent non-persistent short-run mcmc toward energy-based model. *Advances in Neural Information Processing Systems*, 32, 2019.

Nijkamp, E., Hill, M., Han, T., Zhu, S.-C., and Wu, Y. N. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5272–5280, 2020a.

Nijkamp, E., Pang, B., Han, T., Zhou, L., Zhu, S.-C., and Wu, Y. N. Learning multi-layer latent variable model via variational optimization of short run mcmc for approximate inference. In *European Conference on Computer Vision*, pp. 361–378. Springer, 2020b.

Pang, B., Han, T., Nijkamp, E., Zhu, S.-C., and Wu, Y. N. Learning latent space energy-based prior model. *Advances in Neural Information Processing Systems*, 33: 21994–22008, 2020a.

Pang, B., Nijkamp, E., Cui, J., Han, T., and Wu, Y. N. Semi-supervised learning by latent space energy-based model of symbol-vector coupling. *arXiv preprint arXiv:2010.09359*, 2020b.

Rosca, M., Lakshminarayanan, B., and Mohamed, S. Distribution matching in variational inference. *arXiv preprint arXiv:1802.06847*, 2018.

Sø nderby, C. K., Raiko, T., Maalø e, L., Sø nderby, S. r. K., and Winther, O. Ladder variational autoencoders. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/6ae07dcb33ec3b7c814df797cbda0f87-Paper.pdf.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Takahashi, H., Iwata, T., Yamanaka, Y., Yamada, M., and Yagi, S. Variational autoencoder with implicit optimal priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5066–5073, 2019.

Vahdat, A. and Kautz, J. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020.

Vahdat, A., Kreis, K., and Kautz, J. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.

Xiao, Z. and Han, T. Adaptive multi-stage density ratio estimation for learning latent space energy-based model. *arXiv preprint arXiv:2209.08739*, 2022.

Xiao, Z., Kreis, K., Kautz, J., and Vahdat, A. Vaebm: A symbiosis between variational autoencoders and energy-based models. *arXiv preprint arXiv:2010.00654*, 2020.

Xie, J., Lu, Y., Gao, R., Zhu, S.-C., and Wu, Y. N. Cooperative training of descriptor and generator networks. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):27–45, 2018.

Xie, J., Zheng, Z., and Li, P. Learning energy-based model with variational auto-encoder as amortized sampler. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10441–10451, 2021.

Yin, X., Li, S., and Rohde, G. K. Analyzing and improving generative adversarial training for generative modeling and out-of-distribution detection. *arXiv preprint arXiv:2012.06568*, 2020.

Yu, P., Xie, S., Ma, X., Jia, B., Pang, B., Gao, R., Zhu, Y., Zhu, S.-C., and Wu, Y. Latent diffusion energy-based model for interpretable text modeling. In *International Conference on Machine Learning (ICML 2022).*, 2022.

Yu, P., Zhu, Y., Xie, S., Ma, X., Gao, R., Zhu, S.-C., and Wu, Y. N. Learning energy-based prior model with diffusion-amortized mcmc. *arXiv preprint arXiv:2310.03218*, 2023.

Zhao, S., Song, J., and Ermon, S. Learning hierarchical features from generative models. *arXiv preprint arXiv:1702.08396*, 2017.

Zhu, Y., Xie, J., Wu, Y., and Gao, R. Learning energy-based models by cooperative diffusion recovery likelihood. *arXiv preprint arXiv:2309.05153*, 2023.

A. Theorectical Derivation

A.1. Between $\tilde{\mathbf{u}}$-space and $\tilde{\mathbf{z}}$-space

For conditional Gaussian distribution $p_{\beta_i}(\mathbf{z}_i|\mathbf{z}_{i+1})$ (see Eqn. 1), we can have an invertible transformation function $T_{\beta_{>0}}$. For example of 2-layer latent variables, it is defined specifically as

$$\mathbf{z}_2 = T^{\mathbf{z}_2}_{\beta_{>0}}(\mathbf{u}_2) = \mathbf{u}_2 \;\text{ and }\; \mathbf{z}_1 = T^{\mathbf{z}_1}_{\beta_{>0}}(\mathbf{u}_1, \mathbf{u}_2) = \mu_{\beta_1}(\mathbf{z}_2) + \sigma_{\beta_1}(\mathbf{z}_2) \cdot \mathbf{u}_1 \tag{15}$$

where $\mathbf{u}_1$ and $\mathbf{u}_2$ are distributed as independent Gaussian noise, i.e., $(\mathbf{u}_1, \mathbf{u}_2) \sim p_0(\mathbf{u}_1, \mathbf{u}_2)$ and $p_0(\mathbf{u}_1, \mathbf{u}_2) = p_0(\mathbf{u}_1)p_0(\mathbf{u}_2)$ with each $p_0(\mathbf{u}_i) \sim \mathcal{N}(0, I_d)$. By change-of-variable rule, we can have

$$p_{\beta_{>0}}(\mathbf{z}_1, \mathbf{z}_2) = p_0(\mathbf{u}_1, \mathbf{u}_2)|\det(J_{T^{-1}_{\beta_{>0}}})| \;\text{ and }\; p_0(\mathbf{u}_1, \mathbf{u}_2) = p_{\beta_{>0}}(\mathbf{z}_1, \mathbf{z}_2)|\det(J_{T_{\beta_{>0}}})| \tag{16}$$

where $J_{T_{\beta_{>0}}}$ is the Jacobian of $T_{\beta_{>0}}$.

For joint EBM prior on $\tilde{\mathbf{z}}$-space $p_{\omega,\beta_{>0}}(\tilde{\mathbf{z}})$ (see Eqn. 4), we can apply the change-of-variable rule and Eqn. 16 as

$$\begin{aligned} p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}) &= p_{\omega,\beta_{>0}}(\tilde{\mathbf{z}})|\det(J_{T_{\beta_{>0}}})| \\ &= \frac{1}{\mathrm{Z}_{\omega,\beta_{>0}}} \exp[F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}))] p_{\beta_{>0}}(\tilde{\mathbf{z}})|\det(J_{T_{\beta_{>0}}})| \\ &= \frac{1}{\mathrm{Z}_{\omega,\beta_{>0}}} \exp[F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}))] p_0(\tilde{\mathbf{u}}) \end{aligned} \tag{17}$$

which is the Eqn. 8. With such marginal EBM prior on $\tilde{\mathbf{u}}$-space, we construct our conditional EBM prior as shown in Eqn. 9.

A.2. Coupling with symbol vector

Energy-based model can be flexible to couple with symbol vector (Pang et al., 2020b; Yu et al., 2022; Nie et al., 2021; Grathwohl et al., 2019). For Eqn. 14, we adapt our model to couple with symbol vector $\mathbf{y}$. The marginal version of Eqn. 14 is given as

$$\hat{p}_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t+1}) = \frac{1}{\mathrm{Z}_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_{t+1})} \exp\left[\hat{F}_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t)\right] p_0(\tilde{\mathbf{u}}_t) p(\tilde{\mathbf{u}}_{t+1}|\tilde{\mathbf{u}}_t) \tag{18}$$

where $\hat{F}_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t) = \log \sum_{\mathbf{y}} \exp(\langle F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t), \mathbf{y}\rangle)$. This forms a softmax classifier, i.e.,

$$p_{\omega,\beta_{>0}}(\mathbf{y}|\tilde{\mathbf{u}}_t, \tilde{\mathbf{u}}_{t+1}) = \frac{p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t, \mathbf{y}|\tilde{\mathbf{u}}_{t+1})}{\hat{p}_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t+1})} = \frac{\exp(\langle F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t), \mathbf{y}\rangle)}{\sum_{\mathbf{y}} \exp(\langle F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t), \mathbf{y}\rangle)} \tag{19}$$

where energy function $F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t)$ outputs the logit score of categories. Recall that $F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t) = \sum_{i=1}^{L} f_{\omega_i}(\mathbf{z}_i, t)$, we therefore can couple $\mathbf{y}$ at different layers and let the energy score $f_{\omega_i}(\mathbf{z}_i, t)$ at i-th layer serve as the softmax classifier for $\mathbf{y}$.

Recall that we only couple $\mathbf{y}$ at $t = 0$, learning such model can be achieved by maximizing the likelihood of $\log p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_{0:T}, \mathbf{y})$. Specifically, we have

$$\log p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_{0:T}, \mathbf{y}) = \underbrace{\log \hat{p}_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_0|\tilde{\mathbf{u}}_1) + \log p_{\omega,\beta_{>0}}(\mathbf{y}|\tilde{\mathbf{u}}_0, \tilde{\mathbf{u}}_1)}_{\text{with symbol vector}} \tag{20}$$

$$+ \underbrace{\sum_{t=1}^{T-1} \log p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t+1})}_{\text{without symbol vector}} \tag{21}$$

where Eqn. 21 computes the gradient similar as Eqn. 12, while for Eqn. 20, it is learned with an extra term computing the gradient for the softmax classifier $p_{\omega,\beta_{>0}}(\mathbf{y}|\tilde{\mathbf{u}}_0, \tilde{\mathbf{u}}_1)$, i.e., optimizing using standard cross-entropy.

For sampling $\tilde{\mathbf{u}}_0$ with specified $\mathbf{y}$, we first obtain $\tilde{\mathbf{u}}_1$ by reversing step by step from $\tilde{\mathbf{u}}_T$ via Langevin dynamics (see Alg. 2). Then, we perform Langevin dynamics (see Eqn. 3) to sample from $p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_0, \mathbf{y}|\tilde{\mathbf{u}}_1)$. The gradient is then computed as $\nabla_{\tilde{\mathbf{u}}_0} \log p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_0, \mathbf{y}|\tilde{\mathbf{u}}_1) = \nabla_{\tilde{\mathbf{u}}_0} \left[\log \hat{p}_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_0|\tilde{\mathbf{u}}_1) + \log \hat{p}_{\omega,\beta_{>0}}(\mathbf{y}|\tilde{\mathbf{u}}_0, \tilde{\mathbf{u}}_1)\right]$

B. Additional Result

B.1. Additional Result

Figure 9: Images synthesis on CIFAR-10 and LSUN-Chruch-64.

Figure 10: Images synthesis on CelebA-HQ-256. Left figure shows temperature=1.0. Right figure shows temperature=0.7.