Learning Latent Space Hierarchical EBM Diffusion Models

Jiali Cui   Tian Han
Stevens Institute of Technology

Preliminary

Multi-layer Generator 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 can be specified as a joint distribution,

$$p_\beta(\mathbf{x}, \tilde{\mathbf{z}}) = p_{\beta_0}(\mathbf{x}|\tilde{\mathbf{z}}) p_{\beta_{>0}}(\tilde{\mathbf{z}}) \quad \text{where}$$

$$p_{\beta_{>0}}(\tilde{\mathbf{z}}) = \prod_{i=1}^{L-1} p_{\beta_i}(\mathbf{z}_i|\mathbf{z}_{i+1}) p(\mathbf{z}_L)$$

where $\tilde{\mathbf{z}}$ collects $(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_L)$, and $p_{\beta_{>0}}(\tilde{\mathbf{z}})$ is the prior model that factories consecutive layers of latent variables with conditional Gaussian distribution, i.e., $p_{\beta_i}(\mathbf{z}_i|\mathbf{z}_{i+1}) \sim \mathcal{N}(\mu_{\beta_i}(\mathbf{z}_{i+1}), \sigma^2_{\beta_i}(\mathbf{z}_{i+1}))$.

Limitation. The Gaussian prior typically only focuses on the *inter-layer* relation modelling while largely ignoring the *intra-layer* relation modelling, resulting in the *prior hole problem* with mismatch regions between the prior and aggregate posterior distribution.

Joint Energy-based Prior Model. The joint energy-based (EBM) prior model is shown to be expressive in capturing the intra-layer relation. With multi-layer of latent variables $\tilde{\mathbf{z}}$,

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

where $Z_{\omega,\beta_{>0}}$ is the normalizing constant or partition function, $F_\omega(\tilde{\mathbf{z}}) = \sum_{i=1}^{L} f_{\omega_i}(\mathbf{z}_i)$ is the energy function parameterized with ω.

Limitation. Learning such 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, leading to ineffective MCMC sampling for EBM learning.

Diffusion 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 $t = 0, 1, \ldots, T$. The noisy sample $\tilde{\mathbf{z}}_t$ is generated by pre-defined Gaussian perturbation kernel

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

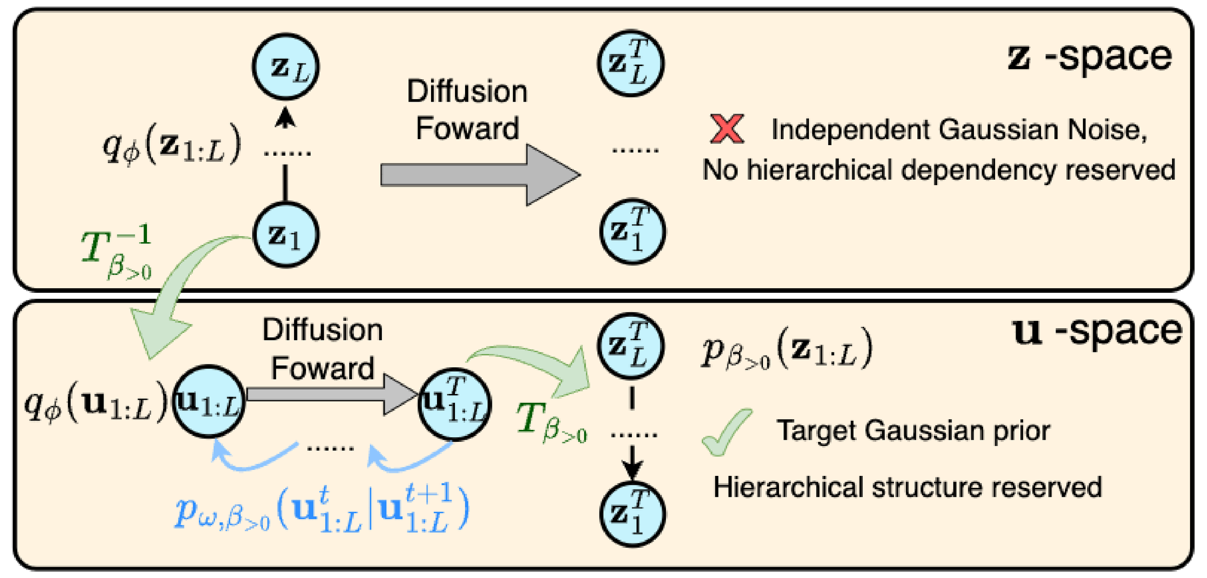
Limitation. It does not suit for multi-layer latent variables $\tilde{\mathbf{z}}$, since it does not take into account the hierarchical structure between layers of latent variables. Their inter-layer relation is consequently *destroyed* during the progress, i.e., each $\mathbf{z}_i$ becomes independently distributed as standard Gaussian noise at the final diffusion step.

Proposed Method

Diffusion on $\tilde{\mathbf{u}}$-space. The conditional Gaussian distribution $p_{\beta_i}(\mathbf{z}_i|\mathbf{z}_{i+1}) \sim \mathcal{N}(\mu_{\beta_i}(\mathbf{z}_{i+1}), \sigma^2_{\beta_i}(\mathbf{z}_{i+1}))$ features the re-parametrization sampling, for which we can define an invertible deterministic transformation function to be $T_{\beta_{>0}}$, i.e., $\tilde{\mathbf{z}} = T_{\beta_{>0}}(\tilde{\mathbf{u}})$ and $\tilde{\mathbf{u}} = T^{-1}_{\beta_{>0}}(\tilde{\mathbf{z}})$. We can adapt our diffusion model on $\tilde{\mathbf{u}}$-space.

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

Illustration on Diffusion Process.



Reverse on $\tilde{\mathbf{u}}$-space. For marginal EBM prior on $\tilde{\mathbf{u}}$-space, we have

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

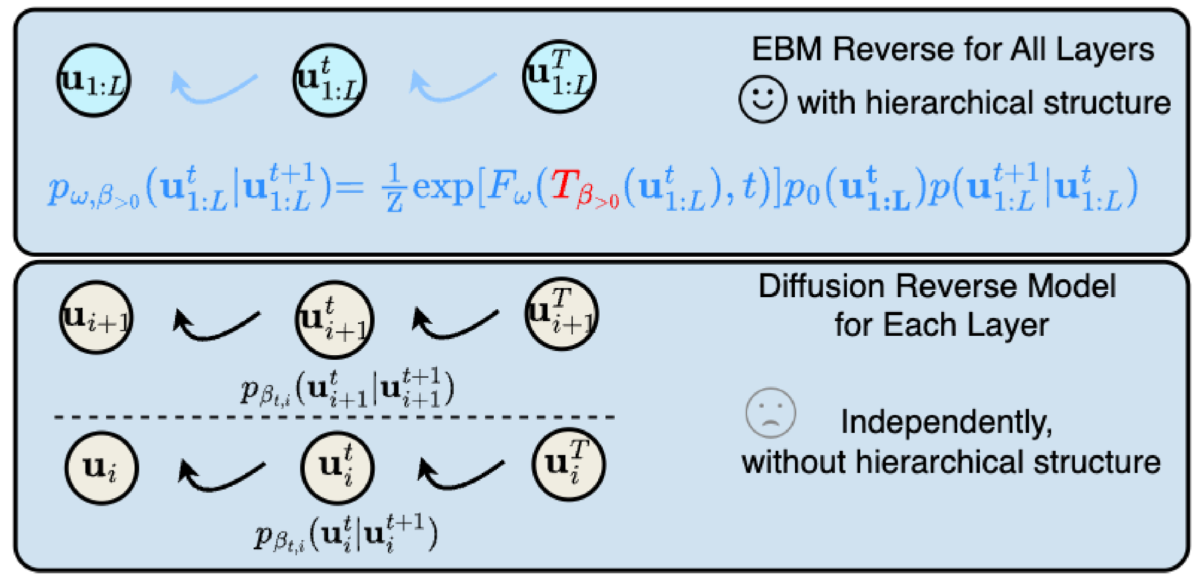
For our reverse model, we formulate the marginal EBM prior to a sequence of conditional EBM prior, i.e.,

$$p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t+1}) \propto p_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_t) p(\tilde{\mathbf{u}}_{t+1}|\tilde{\mathbf{u}}_t) =$$

$$\frac{1}{Z_{\omega,\beta_{>0}}(\tilde{\mathbf{u}}_{t+1})} \exp\left[F_\omega(T_{\beta_{>0}}(\tilde{\mathbf{u}}_t), t)\right] p_0(\tilde{\mathbf{u}}_t) \cdot p(\tilde{\mathbf{u}}_{t+1}|\tilde{\mathbf{u}}_t)$$

where we abuse the notation and use $p(\tilde{\mathbf{u}}_{t+1}|\tilde{\mathbf{u}}_t)$ for forward kernel.

Illustration on Reverse Process.



Experiment: Image Synthesis

Quantiative (FID score) Comparison with Direct Baselines.

FID(↓)	CIFAR-10	CelebA-HQ-256	LSUN-Church-64
NVAE*	37.73	30.25	38.13
NVAE*-Recon	0.68	1.64	2.45
Ours ($T = 3$)	**8.93**	**8.78**	**7.34**
Joint-EBM	11.34	9.89	8.38
DRL EBM ($T = 6$)	9.58	-	8.38
NCP-VAE	24.08	24.79	-

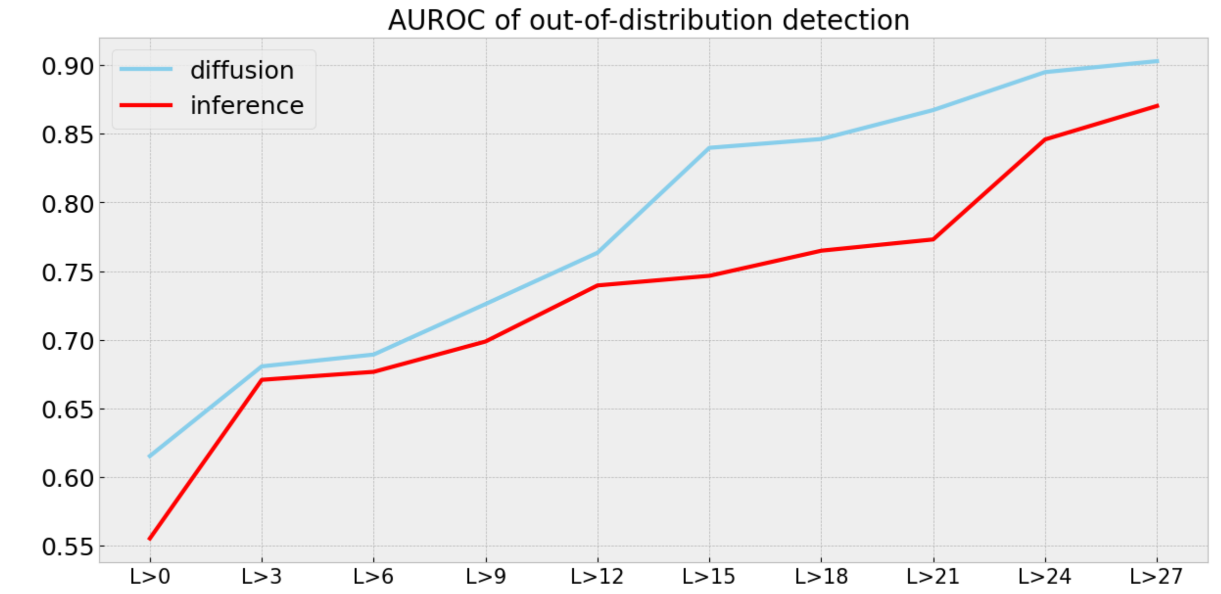
Qualititative Results.

Experiment: Hierarchical Representation

Hierarchical Sampling.

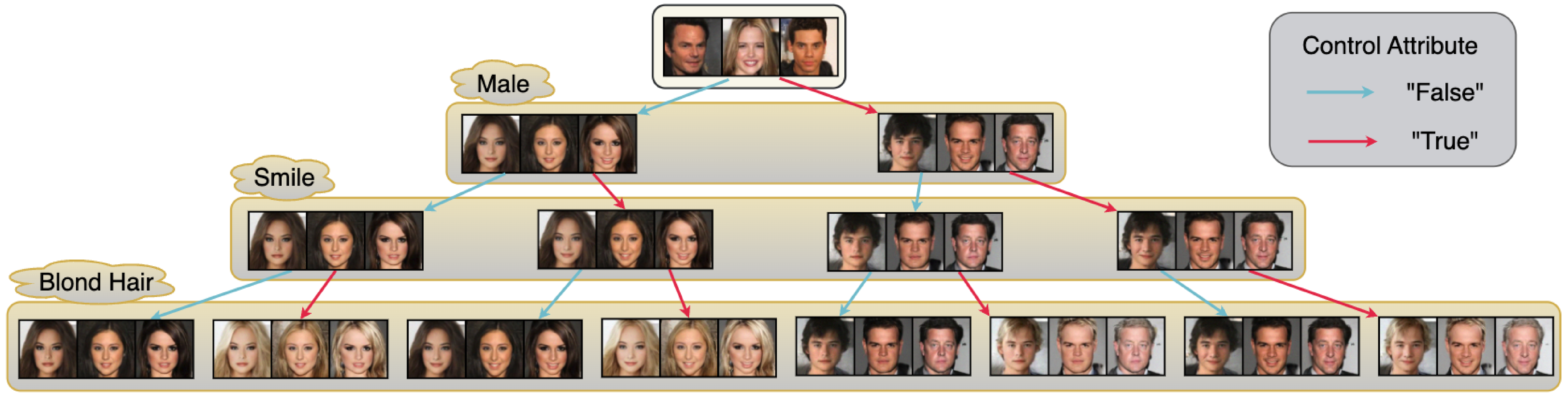
Visualization of representations learned by latent variables from the top to bottom layers, arranged as top-left, top-right, bottom-left and bottom-right.

Hierarchical Out-of-distribution Detection.



The AUROC results for using energy scores of different layers (denoted as $L > k$ for using top layers above k-th layer) as the decision function.

Experiment: Controllable Synthesis



Fine-tuned (hierarchical) controllable image synthesis with multiple attributes on CelebA-64.