Learning Multimodal Latent Generative Models with Energy-Based Prior

Shiyu Yuan[1], Jiali Cui[2], Hanao Li[2], Tian Han[2]

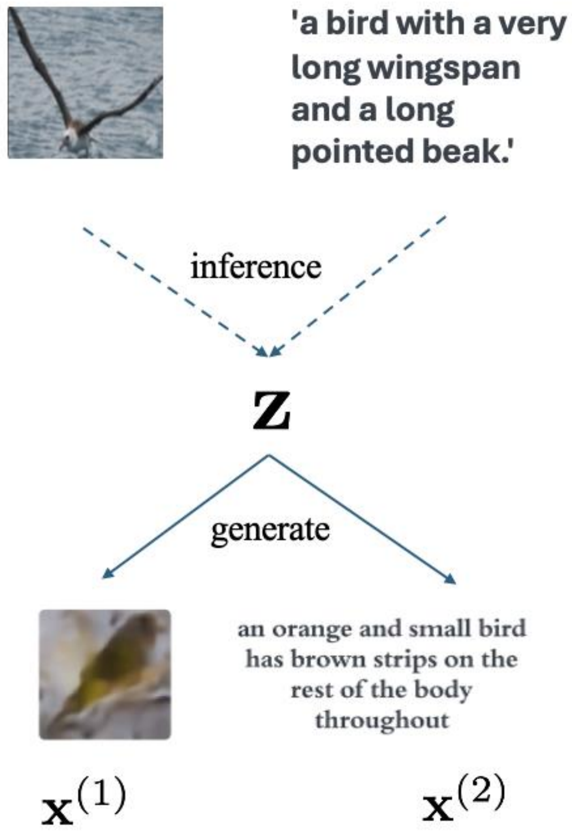
Department of Systems and Enterprises[1] Department of Computer Science[2]

Research Problem

$$\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$$

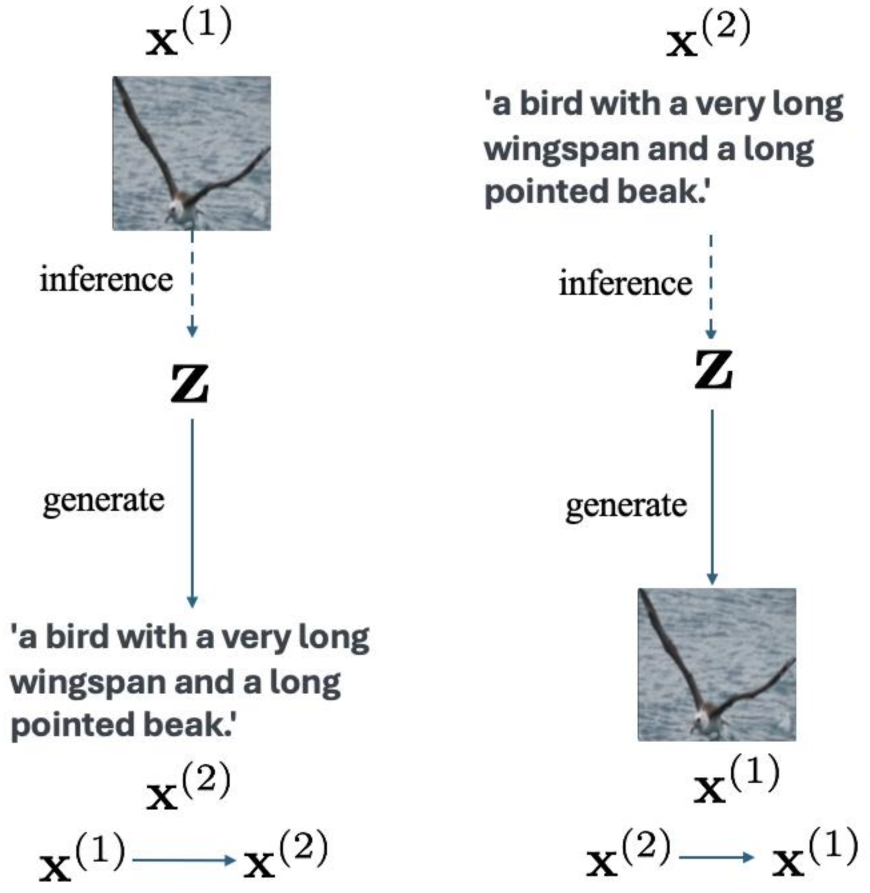
1) Joint Generation:

$$p(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots \mathbf{x}^{(m)})$$



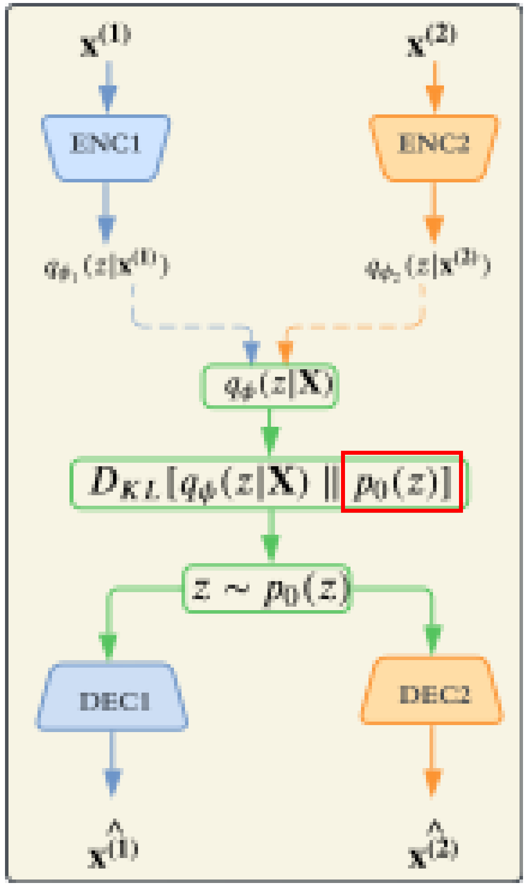
2) Cross Generation:

$$p(\mathbf{x}^{(i)} \mid \mathbf{x}^{(j)}) \quad \text{where} \quad (i \neq j)$$

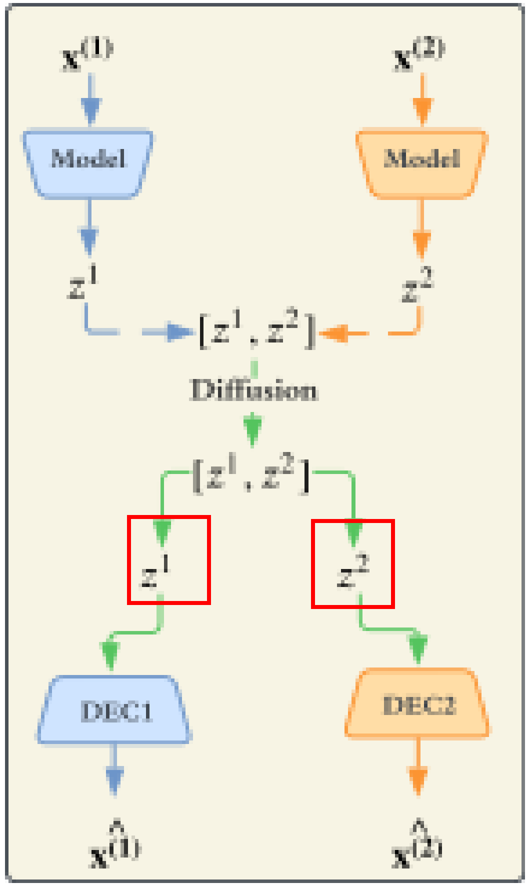


Related Work

Variational Autoencoder-based [1,2,3,4]



Diffusion-based [5,6,7]



[1] Shi, Y et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. **NeurIPS2019**

[2]Wu, M et al. Multimodal generative models for scalable weakly-supervised learning. **NeurIPS2018**

[3] Sutter, T. M., et al. Generalized multimodal ELBO. **ICLR2021**

[4] Palumbo, E., et al. MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises. **ICLR2023**

[5]Bao, F., et al. One transformer fits all distributions in multi-modal diffusion at scale. **ICML2023**

[6]Hu, M., Zheng, et al. UniD3: unified discrete diffusion for simultaneous vision-language generation. **ICLR2023**

[7]Ramesh, A., et al. DALL-E 2

Multimodal Latent Generative Model

$$p_\theta(\mathbf{X}, \mathbf{z}) = p_{\beta_{(1)}}(\mathbf{x}^{(1)}|\mathbf{z}) p_{\beta_{(2)}}(\mathbf{x}^{(2)}|\mathbf{z}) \cdots p_{\beta_{(m)}}(\mathbf{x}^{(m)}|\mathbf{z}) p(\mathbf{z})$$

$\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$	m : number of modalities
$\mathbf{z}$	shared representation among modalities
$p_{\beta_{(m)}}(\mathbf{x}^{(m)}\|\mathbf{z})$	generation model : modality-specific information
$p(\mathbf{z})$	prior: common information

Generation Model

generation model : modality-specific information

$$p_\theta(\mathbf{X}, \mathbf{z}) = \boxed{p_{\beta_{(1)}}(\mathbf{x}^{(1)}|\mathbf{z})p_{\beta_{(2)}}(\mathbf{x}^{(2)}|\mathbf{z}) \cdots p_{\beta_{(m)}}(\mathbf{x}^{(m)}|\mathbf{z})} p(\mathbf{z})$$

assume: $\mathbf{x}^{(1)} \perp\!\!\!\perp \mathbf{x}^{(2)} \perp\!\!\!\perp \cdots \perp\!\!\!\perp \mathbf{x}^{(m)} \mid \mathbf{z}$

$$p_{\beta_{(m)}}(\mathbf{x}^{(m)}|\mathbf{z}) \sim \mathcal{N}(G_{\beta_{(m)}}(\mathbf{z}), I_{D^{(m)}})$$

$$\mathbf{x}^{(m)} = G_{\beta_{(m)}}(\mathbf{z}) + \epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, I_{D^{(m)}})$$

modality-specific generator

Towards more Expressive Prior: Energy-based prior

$$p_\theta(\mathbf{X}, \mathbf{z}) = p_{\beta_{(1)}}(\mathbf{x}^{(1)}|\mathbf{z})p_{\beta_{(2)}}(\mathbf{x}^{(2)}|\mathbf{z})\cdots p_{\beta_{(m)}}(\mathbf{x}^{(m)}|\mathbf{z})p_\alpha(\mathbf{z})$$

- **prior in existing work**: less-informative unimodal distribution such as Gaussian, Laplacian
- **EBM prior:** expressive prior to capture complexity of multimodal shared information
- **Exponential Tilting:** exponentially tilt modification of base distribution via energy function

exponentially tilting

$$p_\alpha(\mathbf{z}) = \frac{1}{\mathbb{Z}(\alpha)} \boxed{\exp[-f_\alpha(\mathbf{z})]} p_0(\mathbf{z})$$

partition function

energy function

base distribution

$p_0(\mathbf{z}) \sim \mathcal{N}(0, I_{D^{(m)}})$

Learning: Maximum Likelihood Estimation

$$\max_{\theta=\{\beta,\alpha\}} L(\theta) = \log \int_{\mathbf{z}} p_{\beta_{(1)}}(\mathbf{x}^{(1)}|\mathbf{z}) \cdots p_{\beta_{(m)}}(\mathbf{x}^{(m)}|\mathbf{z}) p_\alpha(\mathbf{z}) d\mathbf{z}$$

with sufficient data

$$\min_\theta \mathrm{KL}(p_{\text{data}}(\mathbf{X}) \parallel p_\theta(\mathbf{X}))$$

β : generator parameter

α : EBM prior parameter

$$\frac{\partial}{\partial\theta} L(\theta) = \mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{X})}[\frac{\partial}{\partial\theta} \log p_\theta(\mathbf{X},\mathbf{z})]$$

$$q_\phi(\mathbf{z}|\mathbf{X}) = \frac{1}{m} \sum_{m=1}^{M} q_{\phi_{(m)}}(\mathbf{z}|\mathbf{x}^{(m)})$$ [1]

Mixture of Expert (MOE)

ϕ : inference model parameter

[1] Shi, Y et,.al Variational mixture-of-experts autoencoders for multi-modal deep generative models. **NeurIPS2019**

Final Objective for Joint Learning

$$\min_{\theta} \mathrm{KL}(p_{\mathrm{data}}(\mathbf{X}) \parallel p_{\theta}(\mathbf{X}))$$

$$q_{\phi}(\mathbf{z}|\mathbf{X}) \Big\downarrow$$

$$\min_{\beta,\phi,\alpha} \mathrm{KL}(p_{\mathtt{data}}(\mathbf{X}) q_{\phi}(\mathbf{z}|\mathbf{X}) || p_{\beta}(\mathbf{X}|\mathbf{z}) p_{\alpha}(\mathbf{z}))$$

$$\Big\downarrow$$

$$\max_{\beta,\phi,\alpha} \frac{1}{m} \sum_{m=1}^{M} \mathbb{E}_{q_{\phi_{(m)}}(\mathbf{z}|\mathbf{x}^{(m)})} [\log \frac{p_{\beta,\alpha}(\mathbf{X},\mathbf{z})}{q_{\phi}(\mathbf{z}|\mathbf{X})}]$$

β : generator parameter

ϕ : inference model parameter

α : EBM prior parameter

Learning Generator and Inference Model

$$\frac{\partial}{\partial \beta, \phi} L(\beta, \phi, \alpha) = \frac{\partial}{\partial \beta, \phi} \frac{1}{m} \sum_{m=1}^{M} [\mathbb{E}_{q_{\phi_{(m)}}(\mathbf{z}|\mathbf{x}^{(m)})}[\log p_{\beta_{(m)}}(\mathbf{x}^{(m)}|\mathbf{z})]$$

Modality-Specific Reconstruction

Cross-Modality Generation

$$+ \sum_{n=1, n \neq m}^{M} \mathbb{E}_{q_{\phi_{(m)}}(\mathbf{z}|\mathbf{x}^{(m)})}[\log p_{\beta_{(n)}}(\mathbf{x}^{(n)}|\mathbf{z})]$$

Regularization with EBM Prior

$$+ \mathbb{E}_{q_{\phi_{(m)}}(\mathbf{z}|\mathbf{x}^{(m)})}[\log \frac{p_\alpha(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{X})}]]$$

Learning EBM

$$\frac{\partial}{\partial\alpha}L(\beta,\phi,\alpha) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{X})}}_{\text{MOE: variation inference}}\left[\frac{\partial}{\partial\alpha}f_\alpha(\mathbf{z})\right] - \underbrace{\mathbb{E}_{p_\alpha(\mathbf{z})}}_{\text{Langevin dynamics}}\left[\frac{\partial}{\partial\alpha}f_\alpha(\mathbf{z})\right]$$

MOE: variation inference

Langevin dynamics

Sampling from EBM: Langevin dynamics

$$\mathbf{z}_{\tau+1} = \mathbf{z}_\tau + \frac{s^2}{2}\frac{\partial}{\partial\mathbf{z}}[\log p_\alpha(\mathbf{z}_\tau)] + s\cdot\epsilon_\tau \quad \text{where} \quad \epsilon_\tau \sim \mathcal{N}(0, I_d)$$

latent variable at τ step size sampled distribution Gaussian noise

τ : time step EBM prior

[1] Shi, Y et,.al Variational mixture-of-experts autoencoders for multi-modal deep generative models. **NeurIPS2019**

Model Generalization

EBM Prior: Base Version

$$p_\theta(\mathbf{X}, \mathbf{z}) = p_\beta(\mathbf{X}|\mathbf{z}) p_\alpha(\mathbf{z})$$

MOE with modality prior

$$p_\theta(\mathbf{X}, \mathbf{z}, \mathbf{W}) = p_\beta(\mathbf{X}|\mathbf{z}, \mathbf{W}) p_0(\mathbf{z}) p_0(\mathbf{W})$$

$$\mathbf{W} = \{\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(m)}\}$$

Laplacian prior modality-specific prior

EBM Prior: Generalized Version

$$p_\theta(\mathbf{X}, \mathbf{z}, \mathbf{W}) = p_\beta(\mathbf{X}|\mathbf{z}, \mathbf{W}) p_\alpha(\mathbf{z}) p_0(\mathbf{W})$$

EBM prior modality-specific prior

[1] Palumbo, E., et,.al MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises. **ICLR2023**

PolyMNIST[3] : Coherence

- **Joint Coherence**: generated samples modalities alignment and mutually consistent
- **Cross Coherence:** capacity of one modality infer other modalities

EBM Prior: Base Version

Model	Joint Coherence ↑	Cross Coherence↑
	PolyMNIST	
Ours	$\boxed{0.746}$	$\boxed{0.853}$
MMVAE[1]	0.232	0.844

EBM Prior: Generalized Version

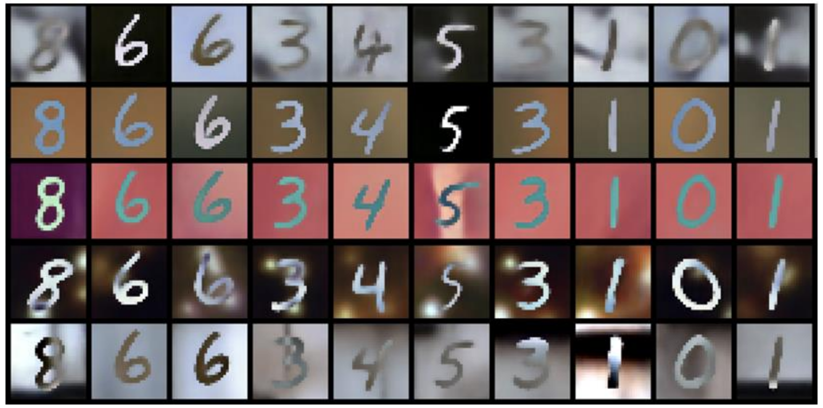
Model	Joint Coherence↑	Cross Coherence↑
	PolyMNIST	
Ours	$\boxed{0.878}$	$\boxed{0.897}$
MMVAE+[2]	0.344	0.869
MoPoE[3]	0.141	0.720
MVTCAE[4]	0.003	0.591
mmJSD[5]	0.060	0.778

[1] Shi, Y et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. **NeurIPS2019**
[2] Palumbo, E., et al. MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises.**ICLR2023**
[3] Sutter, T. M., et al. Generalized multimodal ELBO. **ICLR2021**
[4]Hwang, H., et al. Multi-view representation learning via total correlation objective. **NeurIPS2021**
[5]Sutter, T., et al. Multimodal generative learning utilizing jensen-shannon-divergence. **NeurIPS2020**

PolyMNIST: Joint Generation Visual Result

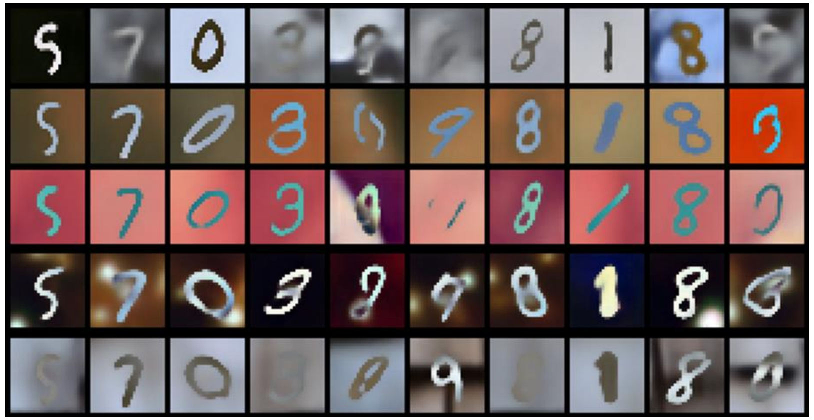
EBM Prior: Base Version

EBM Prior: Generalized Version



MMVAE

MMVAE+



CUB[1] : Markov Transition

$$\mathbf{z}_{\tau+1} = \mathbf{z}_\tau + \frac{s^2}{2}\frac{\partial}{\partial \mathbf{z}}[\log p_\alpha(\mathbf{z}_\tau)] + s \cdot \epsilon_\tau$$

$$z \sim p_0(\mathbf{z}) \dashrightarrow z \sim p_\alpha(\mathbf{z})$$

this bird is a black and are and and a very red ..

this bird has wings that are black and are very red beak

this small has has a are breast and white belly

this small has a orange breast and has a black belly

$z \sim p_0(\mathbf{z})$

$z \sim p_\alpha(\mathbf{z})$

[1]Wah, C., ET AL. The caltech-ucsd birds-200-2011 dataset.2011

Questions please reach out to:

syuan14@stevens.edu