

Learning Multimodal Energy-Based Model with Multimodal Variational Auto-Encoder via MCMC Revision

Jiali Cui

*Basking Ridge, New Jersey, USA
Futurewei Technologies Inc*

jcui2@futurewei.com

Zhiqiang Lao

*Basking Ridge, New Jersey, USA
Futurewei Technologies Inc*

zlao@futurewei.com

Heather Yu

*Basking Ridge, New Jersey, USA
Futurewei Technologies Inc*

hyu@futurewei.com

Reviewed on OpenReview: <https://openreview.net/forum?id=ZVD7bHNpY1>

Abstract

Energy-based models (EBMs) are a flexible class of deep generative models and are well-suited to capture complex dependencies in *multimodal* data. However, learning multimodal EBM by maximum likelihood requires Markov Chain Monte Carlo (MCMC) sampling in the joint data space, where noise-initialized Langevin dynamics often mixes poorly and fails to discover coherent *inter-modal* relationships. Multimodal VAEs have made progress in capturing such inter-modal dependencies by introducing a shared latent generator and a joint inference model. However, both the shared latent generator and joint inference model are parameterized as unimodal Gaussian (or Laplace), which severely limits their ability to approximate the complex structure induced by multimodal data. In this work, we study the learning problem of the multimodal EBM, shared latent generator, and joint inference model. We present a learning framework that effectively interweaves their MLE updates with corresponding MCMC refinements in both the data and latent spaces. Specifically, the generator is learned to produce coherent multimodal samples that serve as strong initial states for EBM sampling, while the inference model is learned to provide informative latent initializations for generator posterior sampling. Together, these two models serve as complementary models that enable effective EBM sampling and learning, yielding realistic and coherent multimodal EBM samples. Extensive experiments demonstrate superior performance for multimodal synthesis quality and coherence compared to various baselines. We conduct various analyses and ablation studies to validate the effectiveness and scalability of the proposed multimodal framework.

1 Introduction

Deep generative models (DGMs) have achieved remarkable success in modelling complex data distributions for single modalities (Ho et al., 2020; Vahdat & Kautz, 2020; Karras et al., 2020). In recent years, these advances have rapidly extended to single-flow multimodal frameworks (e.g., text-to-image) (Ramesh et al., 2022; Alayrac et al., 2022; Li et al., 2023) and to multi-flow multimodal models capable of supporting multiple generation flows within a single model (Hu et al., 2023; Bao et al., 2023; Le et al., 2025; Li et al., 2025). Among various DGMs, energy-based models (EBMs) form a particularly flexible class and are well suited to capture rich contextual structure in data space (Du et al., 2020; Gao et al., 2020; Cui et al., 2023a).

Yet, most existing EBMs focus on single-modality data and remain largely underexplored in the multimodal domain, falling behind other generative approaches.

Learning multimodal EBMs using maximum likelihood estimation (MLE) requires sampling from the EBM distribution, typically via Markov Chain Monte Carlo (MCMC) methods such as Langevin dynamics Neal et al. (2011). When the sampling chains are initialized from random noise, Langevin dynamics is often ineffective and may take a long time to mix between local modes (Nijkamp et al., 2020a). More critically, in multimodal modelling, effective EBM samples should additionally exhibit high coherence and consistency across modalities during the Langevin tranverse. Notably, to tackle EBM sampling challenge, various methods for *single-modal* EBMs have been proposed, which can be broadly grouped into two categories: (i) using auxiliary models to amortize EBM sampling and thus bypass explicit MCMC during training, and (ii) using complementary models (Xie et al., 2018; 2021; 2022; Cui & Han, 2023) to provide informative initial states, thereby enabling more effective EBM sampling. While both strategies can provide meaningful learning signals for single-modal EBMs, *multimodal* data additionally contains complex inter-modal dependencies (i.e., relationships shared across modalities) and modality-specific variations (i.e., inductive biases unique to each modality). As a result, the development of multimodal EBM is needed yet remains largely underexplored.

Multimodal VAE (Wu & Goodman, 2018; Shi et al., 2019) has emerged as a promising approach for multimodal modelling. These models factorize a shared latent space and modality-specific generation models that map the low-dimensional shared latent space to the high-dimensional multimodal data space. The shared latent variables are thus learned to capture common structure across modalities, while the generators can preserve modality-specific characteristics. Learning typically follows the standard variational framework (Kingma & Welling, 2013), using an inference model to approximate the generator posterior. Unlike the single-modality VAE, designing an effective joint inference model in the multimodal setting is nontrivial and remains an active research topic (Palumbo et al., 2023; 2024). A prominent family of works adopts a mixture-of-experts (MoE) formulation (Shi et al., 2019), in which each modality contributes an expert posterior, typically parameterized as a unimodal Gaussian (or Laplace) over the shared latent variable. However, the multimodal generator posterior has a product-of-experts (PoE) structure that is often sharp and complex (see analysis in Section. 2.2.1), making these MoE-based approximations inherently limited (Daunhawer et al., 2021). Alternatively to the variational learning scheme, one can sample from the generator posterior using MCMC methods such as Langevin dynamics (Han et al., 2017). Nevertheless, when chains are initialized from noise, Langevin dynamics again struggle to effectively explore the shared latent space (Nijkamp et al., 2020b), especially in complex multimodal regimes.

To address the learning and sampling challenges above, we propose a cooperative framework that incorporates the multimodal EBM, a shared latent generator, and a joint inference model into a unified probabilistic framework. We present a novel learning scheme that can seamlessly integrate these models by interweaving their MLE updates. Specifically, to enable effective multimodal EBM sampling, the shared latent generator is learned to approximate the EBM distribution so that it can provide informative and coherent initial states for MCMC-based EBM sampling. For the mismatch between the MoE-based joint inference model and the PoE-style generator posterior, we perform a finite number of MCMC posterior update steps, initialized from the joint inference model, which is itself learned to approximate the generator posterior. This cooperative interplay among the three models yields effective sampling, accurate posterior inference, and stable training dynamics. Various experiments demonstrate that our learned EBM and shared latent generator model produce realistic and highly coherent multimodal samples, and consistently outperform strong multimodal variational baselines in both synthesis quality and cross-modal coherence. We further provide ablations and diagnostic analyses to clarify how each component contributes to sampling efficiency, multimodal coherence, and overall performance.

Our contributions can be summarized as: **(i)** We present a novel learning methodology that facilitates effective EBM sampling and learning toward multimodality. **(ii)** We integrate the multimodal EBM, shared latent generator, and joint inference model into a unified probabilistic framework, interleaving their MLE updates so that each component benefits from the others. **(iii)** We conduct extensive experiments, demonstrating superior performance of our multimodal EBM and effectiveness of our learning method.

2 Preliminary

In this section, we briefly review energy-based models, shared latent generator models for multimodal data, and joint inference, which together form the foundation of our framework.

2.1 Preliminary: Multimodal Energy-based Model

Let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ denote an observed multimodal data example consisting of M modalities, and let $p_{\text{data}}(\mathbf{X})$ represent the unknown empirical data distribution. Energy-based models (EBMs) (Du et al., 2020; Gao et al., 2020; Cui et al., 2023b) represent a flexible class of generative models that define an undirected probability distribution

$$\pi_\alpha(\mathbf{X}) = \frac{1}{\mathbf{Z}(\alpha)} \exp[F_\alpha(\mathbf{X})] \quad (1)$$

where $\mathbf{Z}(\alpha) = \int \exp[F_\alpha(\mathbf{X})] d\mathbf{X}$ is the normalizing constant (or partition function), and $F_\alpha(\mathbf{X})$ is a scalar energy function parameterized by α . For multimodal \mathbf{X} , the energy function takes all inputs $\{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ and outputs a single energy value.

While EBMs offer considerable modelling flexibility, their application to multimodal data remains relatively underexplored. A central challenge is to design energy functions that effectively capture the joint structure and dependencies across heterogeneous modalities. In this work, we adopt a simple yet general parameterization: $F_\alpha(\mathbf{X}) = \bar{f}([f_1(\mathbf{x}_1), \dots, f_M(\mathbf{x}_M)])$, where each f_i maps modality \mathbf{x}_i to a fixed-dimensional feature vector, and \bar{f} aggregates the concatenated features to produce the final energy score. More sophisticated architectures could further improve performance, but our focus is on the learning methodology rather than architectural innovation. Implementation details are provided in Appendix. D.

EBM Learning and Sampling. Given N multimodal data $\{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ drawn from $p_{\text{data}}(\mathbf{X})$, the EBM can be learned via maximum likelihood estimation (MLE). The log-likelihood is computed as $\mathcal{L}_\pi(\alpha) = \frac{1}{N} \sum_{i=1}^N \log \pi_\alpha(\mathbf{X}_i)$. When N becomes sufficiently large, maximizing $\mathcal{L}_\pi(\alpha)$ is equivalent to minimizing the KL divergence between the true data density and EBM density, i.e.,

$$-\mathcal{L}_\pi(\alpha) = D_{\text{KL}}(p_{\text{data}}(\mathbf{X}) || \pi_\alpha(\mathbf{X})) \quad \text{where} \quad (2)$$

$$\frac{\partial}{\partial \alpha} \mathcal{L}_\pi(\alpha) = \mathbb{E}_{p_{\text{data}}(\mathbf{X})} \left[\frac{\partial}{\partial \alpha} F_\alpha(\mathbf{X}) \right] - \mathbb{E}_{\pi_\alpha(\mathbf{X})} \left[\frac{\partial}{\partial \alpha} F_\alpha(\mathbf{X}) \right]$$

Computing Eqn. 2 requires EBM samples, i.e., $\mathbf{X} \sim \pi_\alpha(\mathbf{X})$, which can be achieved via MCMC methods, such as Langevin dynamics (Neal et al., 2011) that iteratively updates

$$\mathbf{X}^{k+1} = \mathbf{X}^k + s \frac{\partial}{\partial \mathbf{X}^k} \log \pi_\alpha(\mathbf{X}^k) + \sqrt{2s} \cdot \epsilon^k \quad (3)$$

where k denotes the iteration index, s is the step size, and ϵ^k is Gaussian noise. In the limit as $s \rightarrow 0$ and $k \rightarrow \infty$, this process will converge to the stationary distribution $\pi_\alpha(\mathbf{X})$ (Neal et al., 2011).

Multimodal Challenge. Common practices typically adopt short-run Langevin dynamics (Nijkamp et al., 2019), which perform certain steps (e.g., $k = 30$) of Langevin dynamics to obtain approximate EBM samples. Although this yields useful learning signals, it is still difficult to obtain high-quality samples when the chain is initialized from non-informative states¹ (Grathwohl et al., 2021; Kumar et al., 2019).

For multimodal data, this difficulty is amplified. The input $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ exhibits complex inter-modal dependencies (i.e., relationships among modalities), and successful sampling should traverse coherent local modes that preserve consistent cross-modal structure. As a result, informative and *coherent* initializations become especially critical for effective multimodal EBM sampling and learning.

¹For example, $\mathbf{X}^{k=0}$ drawn from a unit Gaussian or a uniform distribution.

2.2 Preliminary: Multimodal Variational AutoEncoder

To capture inter-modal relationships, shared latent generator models have emerged as a promising approach for multimodal modelling (Wu & Goodman, 2018; Shi et al., 2019). Let \mathbf{z} denote a low-dimensional latent variable. The shared latent generator specifies a joint distribution over multimodal inputs as

$$\begin{aligned} p_\omega(\mathbf{X}, \mathbf{z}) &= p_\omega(\mathbf{X}|\mathbf{z})p_0(\mathbf{z}) \quad \text{where} \\ p_\omega(\mathbf{X}|\mathbf{z}) &= p_{\omega_1}(\mathbf{x}_1|\mathbf{z})p_{\omega_2}(\mathbf{x}_2|\mathbf{z}) \cdots p_{\omega_M}(\mathbf{x}_M|\mathbf{z}) \end{aligned} \quad (4)$$

Here, $p_0(\mathbf{z})$ is the prior distribution (e.g., Gaussian or Laplace distribution) over a shared latent variable \mathbf{z} , and $p_\omega(\mathbf{X}|\mathbf{z})$ is the conditional likelihood given such shared latent variable and factorizes a product of M modality-specific generation models. Each $p_{\omega_i}(\mathbf{x}_i|\mathbf{z}) \sim \mathcal{N}(\mu_{\omega_i}(\mathbf{z}), \sigma^2 I_d)$ represents a conditional Gaussian parameterized by ω_i , mapping the low-dimensional latent space to high-dimensional data space.

This shared latent generator model is designed to capture modality-invariant representations (i.e., high-level semantics) across different modalities through the shared latent space \mathbf{z} , while also being capable of modelling modality-specific biases through separate generation models for each modality.

Multimodal Joint Inference Model. For learning Eqn. 4, multimodal VAEs (Sutter et al., 2020; Hwang et al., 2021; Palumbo et al., 2023; 2024) employ variational learning schemes by introducing a joint inference model. For multimodal data, factorizing effective joint inference models remains challenging and is an active research area (see details in Section. 4). Among various methods, one major paradigm is the mixture-of-experts (MoE) (Shi et al., 2019) defined as

$$q_\phi(\mathbf{z}|\mathbf{X}) = \frac{1}{M} \sum_{i=1}^M q_{\phi_i}(\mathbf{z}|\mathbf{x}_i) \quad (5)$$

Each $q_{\phi_i}(\mathbf{z}|\mathbf{x}_i) \sim \mathcal{N}(\mu_{\phi_i}(\mathbf{x}_i), V_{\phi_i}(\mathbf{x}_i))$ is modeled as conditional Gaussian (or Laplace), where $\mu_{\phi_i}(\mathbf{x}_i)$ and $V_{\phi_i}(\mathbf{x}_i)$ denote the mean and diagonal covariance matrix parameterized by ϕ_i .

This mixture-based joint inference model offers a tractable approximation to the generator posterior, particularly useful in scenarios with missing modalities. However, both the mixture formulation and the assumed individual posteriors are limited in statistical expressivity. They often induce an overly smooth latent space, which may fail to capture the intricate structure of the true multimodal generator posterior, ultimately resulting in a suboptimal generator model.

2.2.1 Revisiting Learning from MLE Perspective

In this section, we revisit the learning of the shared latent generator from the maximum-likelihood perspective. Our goal is to clarify the use of MCMC-based revision and a joint inference model in our framework.

MLE Learning: Consider maximizing the log-likelihood of the shared latent generator model, i.e., $\mathcal{L}_p(\omega) = \frac{1}{N} \sum_{i=1}^N \log p_\omega(\mathbf{X}_i)$, where $p_\omega(\mathbf{X}_i) = \int_{\mathbf{z}} p_\omega(\mathbf{X}, \mathbf{z}) d\mathbf{z}$ is its marginal distribution. With a sufficiently large number of N , it is equivalent to minimizing the KL-divergence as

$$\mathcal{L}_p(\omega) = D_{\text{KL}}(p_{\text{data}}(\mathbf{X}) || p_\omega(\mathbf{X})) \quad (6)$$

$$\text{where } \frac{\partial}{\partial \omega} \mathcal{L}_p(\omega) = \mathbb{E}_{p_{\text{data}}(\mathbf{X})p_\omega(\mathbf{z}|\mathbf{X})} \left[\frac{\partial}{\partial \omega} \log p_\omega(\mathbf{X}, \mathbf{z}) \right]$$

Here, $p_\omega(\mathbf{z}|\mathbf{X})$ is the generator posterior. To shed further light, we can decompose it into

$$p_\omega(\mathbf{z}|\mathbf{X}) = \frac{p_\omega(\mathbf{X}|\mathbf{z})p_0(\mathbf{z})}{p_\omega(\mathbf{X})} = \frac{p_0(\mathbf{z})}{p_\omega(\mathbf{X})} \prod_{i=1}^M \frac{p_{\omega_i}(\mathbf{z}|\mathbf{x}_i)p_{\omega_i}(\mathbf{x}_i)}{p_0(\mathbf{z})} \propto \frac{\prod_{i=1}^M p_{\omega_i}(\mathbf{z}|\mathbf{x}_i)}{\prod_{i=1}^{M-1} p_0(\mathbf{z})} \quad (7)$$

which reveals that the generator posterior is effectively a product of individual posteriors, modulated by the prior, leading to sharp and complex structures in the latent space. Approximating this product-based

posterior using a mixture-based joint inference model (Eqn. 5) can be suboptimal due to the smoothing effect inherent in averaging (Daunhawer et al., 2021). Moreover, the unimodal $q_{\phi_i}(\mathbf{z}|\mathbf{x}_i)$ (e.g., Gaussian or Laplace) can be limited in statistical expressivity and may fail to capture the intricate structure of the complex individual posteriors (Pang et al., 2021; Xie et al., 2022).

MCMC Posterior Sampling. Alternatively, one can obtain posterior samples by MCMC methods, such as Langevin dynamics (Han et al., 2017; Kong et al., 2024a;b), i.e.,

$$\mathbf{z}^{k+1} = \mathbf{z}^k + s \frac{\partial}{\partial \mathbf{z}^k} \log p_{\omega}(\mathbf{z}^k|\mathbf{X}) + \sqrt{2s} \cdot \epsilon^k \quad (8)$$

The target distribution is the generator posterior $p_{\omega}(\mathbf{z}|\mathbf{X})$, and the gradient term can be computed as $\frac{\partial}{\partial \mathbf{z}} \log p_{\omega}(\mathbf{z}|\mathbf{X}) \propto \frac{\partial}{\partial \mathbf{z}} \log p_{\omega}(\mathbf{X}|\mathbf{z})p_0(\mathbf{z})$. The log-likelihood gradient decomposes as $\log p_{\omega}(\mathbf{X}|\mathbf{z}) = \sum_{i=1}^M \log p_{\omega_i}(\mathbf{x}_i|\mathbf{z})$, which updates shared latent variable \mathbf{z} to explain all modality observations $\{\mathbf{x}_1, \dots, \mathbf{x}_M\}$. As $s \rightarrow 0$ and $k \rightarrow \infty$, this process converges to the stationary distribution $p_{\omega}(\mathbf{z}|\mathbf{X})$ (Neal et al., 2011).

Multimodal Challenge. While MCMC-based posterior sampling can yield more accurate approximations than variational methods, it often suffers from poor mixing and slow convergence when using short-run Langevin dynamics (e.g., $k = 10$) initialized from non-informative points² (Nijkamp et al., 2020b). More critically, the product-based formulation of the generator posterior requires complete observations from all modalities, and the individual unimodal posteriors are often undertrained and poorly calibrated. With inconsistent initializations, this becomes ill-defined in cross-modal inference scenarios, where only a subset of modalities is available (Shi et al., 2019; Daunhawer et al., 2021).

3 Multimodal Learning via MCMC Revision

To address the learning and sampling challenges discussed above, we propose a multimodal learning framework that jointly learns the multimodal EBM, the shared latent generator, and the joint inference model through a cooperative mechanism. The key idea is to leverage the *complementary* strengths of each component: the joint inference model provides informative and coherent initial states for MCMC posterior sampling; the shared latent generator produces consistent multimodal samples that initialize EBM sampling; and the EBM, in turn, supplies revision signals that refine both the generator and the inference model. By interweaving their maximum-likelihood updates with MCMC-based revision, each model benefits from the others, which improves sampling effectiveness and leads to more effective multimodal modelling.

3.1 Revision Signal of MCMC Kernels

Let $\mathcal{M}_{\alpha}^{k_{\mathbf{X}}}(\cdot)$ denote the Markov transition kernel corresponding to $k_{\mathbf{X}}$ steps of Langevin dynamics on EBM density (Eqn. 3), and let $\mathcal{M}_{\omega}^{k_{\mathbf{z}}}(\cdot)$ denote the Markov transition kernel for $k_{\mathbf{z}}$ steps of Langevin dynamics on generator posterior (Eqn. 8). We define two joint densities that describe these MCMC revision processes:

$$\begin{aligned} \Omega_{\omega, \alpha}(\mathbf{X}, \mathbf{z}) &= \mathcal{M}_{\alpha}^{k_{\mathbf{X}}} \cdot p_{\omega}(\mathbf{X}|\mathbf{z})p_0(\mathbf{z}) \\ \Phi_{\omega, \phi}(\mathbf{X}, \mathbf{z}) &= p_{\text{data}}(\mathbf{X}) \cdot \mathcal{M}_{\omega}^{k_{\mathbf{z}}} \cdot q_{\phi}(\mathbf{z}|\mathbf{X}) \end{aligned} \quad (9)$$

where $\Omega_{\omega, \alpha}(\mathbf{X}, \mathbf{z})$ is obtained by first drawing $\mathbf{z} \sim p_0(\mathbf{z})$, generating an initial multimodal sample $\mathbf{X} \sim p_{\omega}(\mathbf{X}|\mathbf{z})$, and then applying $k_{\mathbf{X}}$ steps of Langevin dynamics under the EBM to obtain \mathbf{X} . The resulting marginal $\mathcal{M}_{\alpha}^{k_{\mathbf{X}}} p_{\omega}(\mathbf{X}) = \int_{\bar{\mathbf{X}}} \int_{\mathbf{z}} \mathcal{M}_{\alpha}^{k_{\mathbf{X}}}(\bar{\mathbf{X}}) p_{\omega}(\bar{\mathbf{X}}, \mathbf{z}) d\mathbf{z} d\bar{\mathbf{X}}$ is generally more expressive than the Gaussian (or Laplace) generator. Similarly, $\Phi_{\omega, \phi}(\mathbf{X}, \mathbf{z})$ is obtained by first taking a data example $\mathbf{X} \sim p_{\text{data}}(\mathbf{X})$, initializing the latent state via the joint inference model $\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{X})$, and then applying $k_{\mathbf{z}}$ steps of Langevin dynamics targeting the generator posterior. This yields a revised posterior marginal $\mathcal{M}_{\omega}^{k_{\mathbf{z}}} q_{\phi}(\mathbf{z}|\mathbf{X}) = \int_{\bar{\mathbf{z}}} \mathcal{M}_{\omega}^{k_{\mathbf{z}}}(\bar{\mathbf{z}}) q_{\phi}(\bar{\mathbf{z}}|\mathbf{X}) d\bar{\mathbf{z}}$, which can better approximate the sharp, PoE structure of the multimodal generator posterior than the mixture-based approximate posterior and its unimodal (Gaussian or Laplace) experts.

Related MCMC-revised densities have been used in earlier work (Xie et al., 2018; 2022; 2021; Cui & Han, 2023) on single-modality EBMs, where a generator provides initial states that are refined by an EBM,

²e.g., $\mathbf{z}^{k=0}$ drawn from a unit Gaussian or uniform distribution.

and the resulting samples are then used to update both models. In contrast, our formulation operates directly on multimodal data \mathbf{X} and the MCMC kernels act on both the EBM and the multimodal generator posterior. In particular, $\Omega_{\omega,\alpha}(\mathbf{X}, \mathbf{z})$ leverages the shared latent generator to produce coherent multimodal initializations (across all \mathbf{x}_i), which are then refined by the EBM. This improves mixing and learning for multimodal EBM sampling while respecting inter-modal dependencies. $\Phi_{\omega,\phi}(\mathbf{X}, \mathbf{z})$ uses the mixture-based joint inference model as an amortized initializer for latent MCMC. The subsequent Langevin refinement yields more accurate posterior samples than those obtained by purely variational multimodal VAE schemes (Shi et al., 2019; Palumbo et al., 2023), especially in regimes where the true posterior is sharp and highly structured. These two revised joint densities provide the core “revision signals” in our framework: the EBM refines generator samples in data space, and the generator posterior refines inference samples in latent space. In Section. 3.2, we show how to use $\Omega_{\omega,\alpha}(\mathbf{X}, \mathbf{z})$ and $\Phi_{\omega,\phi}(\mathbf{X}, \mathbf{z})$ to define coupled learning objectives that jointly update all three models.

3.2 Learning Objectives with MCMC-revised Densities

The two MCMC-revised joint densities serve as intermediate targets that couple the three models. At optimization step t , MCMC kernels are evaluated at the current parameters for revised joint densities denoted as $\Omega_{\omega_t,\alpha_t}(\mathbf{X}, \mathbf{z})$ and $\Phi_{\omega_t,\phi_t}(\mathbf{X}, \mathbf{z})$. During parameter updates, we adopt a stop-gradient operation for the revision process, so that the revised samples are treated as fixed guidance, and gradients are not propagated through the Langevin transition kernels. Each model is then updated by minimizing a KL divergence between these revised densities and its own current distribution. In this way, the EBM, the shared latent generator, and the joint inference model are all guided toward the more accurate “revised” distributions.

(i) for multimodal EBM α , we define the learning objective $\mathcal{L}_\pi(\alpha)$ as

$$-\mathcal{L}_\pi(\alpha) = D_{\text{KL}}(\Phi_{\omega_t,\phi_t}(\mathbf{X}, \mathbf{z})||\pi_\alpha(\mathbf{X})q_\phi(\mathbf{z}|\mathbf{X})) - D_{\text{KL}}(\Omega_{\omega_t,\alpha_t}(\mathbf{X}, \mathbf{z})||\pi_\alpha(\mathbf{X})q_\phi(\mathbf{z}|\mathbf{X})) \quad (10)$$

$$\text{with } \frac{\partial}{\partial \alpha} \mathcal{L}_\pi(\alpha) = \mathbb{E}_{\Phi_{\omega_t,\alpha_t}(\mathbf{X}, \mathbf{z})} \left[\frac{\partial}{\partial \alpha} F_\alpha(\mathbf{X}) \right] - \mathbb{E}_{\Omega_{\omega_t,\alpha_t}(\mathbf{X}, \mathbf{z})} \left[\frac{\partial}{\partial \alpha} F_\alpha(\mathbf{X}) \right]$$

Intuitively, our multimodal EBM is updated to increase the energy (i.e., lower probability) of samples from Ω to those from Φ . Since Φ is anchored on real data and refined in latent space, while Ω is obtained from generator-initialized chains refined in data space, this contrast encourages the EBM to better align with the data-driven revised distribution. We optimize $\mathcal{L}_\pi(\alpha)$ via stochastic gradient ascent (SGA).

(ii) For shared latent generator model ω , the learning objective $\mathcal{L}_p(\omega)$ is

$$-\mathcal{L}_p(\omega) = D_{\text{KL}}(\Phi_{\omega_t,\phi_t}(\mathbf{X}, \mathbf{z})||p_\omega(\mathbf{X}, \mathbf{z})) + D_{\text{KL}}(\Omega_{\omega_t,\alpha_t}(\mathbf{X}, \mathbf{z})||p_\omega(\mathbf{X}, \mathbf{z})) \quad (11)$$

$$\text{with } \frac{\partial}{\partial \omega} \mathcal{L}_p(\omega) = \mathbb{E}_{\Phi_{\omega_t,\alpha_t}(\mathbf{X}, \mathbf{z})} \left[\frac{\partial}{\partial \omega} \log p_\omega(\mathbf{X}, \mathbf{z}) \right] - \mathbb{E}_{\Omega_{\omega_t,\alpha_t}(\mathbf{X}, \mathbf{z})} \left[\frac{\partial}{\partial \omega} \log p_\omega(\mathbf{X}, \mathbf{z}) \right]$$

The shared latent generator is learned toward both revised distributions: Φ , which reflects data-driven posterior refinement, and Ω , which reflects EBM-refined samples. Matching these two targets guides the generator to better approximate the EBM density while maintaining consistency with data and posterior structure. We again use SGA to update ω .

(iii) For multimodal joint inference model ϕ , the learning objective $\mathcal{L}_q(\phi)$ is

$$-\mathcal{L}_q(\phi) = D_{\text{KL}}(\Phi_{\omega_t,\phi_t}(\mathbf{X}, \mathbf{z})||p_{\text{data}}(\mathbf{X})q_\phi(\mathbf{z}|\mathbf{X})) + D_{\text{KL}}(\Omega_{\omega_t,\alpha_t}(\mathbf{X}, \mathbf{z})||\pi_\alpha(\mathbf{X})q_\phi(\mathbf{z}|\mathbf{X})) \quad (12)$$

$$\text{with } \frac{\partial}{\partial \phi} \mathcal{L}_q(\phi) = \mathbb{E}_{\Phi_{\omega_t,\alpha_t}(\mathbf{X}, \mathbf{z})} \left[\frac{\partial}{\partial \phi} \log q_\phi(\mathbf{z}|\mathbf{X}) \right] + \mathbb{E}_{\Omega_{\omega_t,\alpha_t}(\mathbf{X}, \mathbf{z})} \left[\frac{\partial}{\partial \phi} \log q_\phi(\mathbf{z}|\mathbf{X}) \right]$$

$$= \mathbb{E}_{\Phi_{\omega_t,\alpha_t}(\mathbf{X}, \mathbf{z})} \left[\frac{\partial}{\partial \phi} \text{LSE}(\log q_{\phi_i}(\mathbf{z}|\mathbf{x}_i)) \right] + \mathbb{E}_{\Omega_{\omega_t,\alpha_t}(\mathbf{X}, \mathbf{z})} \left[\frac{\partial}{\partial \phi} \text{LSE}(\log q_{\phi_i}(\mathbf{z}|\mathbf{x}_i)) \right]$$

where $\text{LSE}(\cdot)$ denotes the log-sum-exp operation, i.e., $\log \sum_{i=1}^M \exp(\cdot)$, corresponding to the mixture-based joint inference over M modalities. In this case, the inference model is encouraged to align with both revised

joint densities: one anchored on the true data distribution p_{data} and one on the current EBM π_α . Because the mixture-of-experts structure is preserved through the log-sum-exp form, the gradients naturally propagate to each modality-specific encoder. We update ϕ using SGA based on the computed gradient.

3.2.1 Why MCMC-revised Kernels?

The MCMC-revised kernels act as bridges between the three components and determine how they interact during cooperative learning. Consider the long-run behavior of the transition kernels $\mathcal{M}_\alpha^{k_x}(\cdot)$ and $\mathcal{M}_\omega^{k_z}(\cdot)$, the marginal distribution induced by the generator and inference models converge as $\mathcal{M}_\alpha^{k_x} p_\omega(\mathbf{X}) \rightarrow \pi_{\alpha_t}(\mathbf{X})$ and $\mathcal{M}_\omega^{k_z} q_\phi(\mathbf{z}|\mathbf{X}) \rightarrow p_{\omega_t}(\mathbf{z}|\mathbf{X})$, respectively. In this limit, learning the multimodal EBM via Eqn. 10 seeks to approximate the true data distribution $p_{\text{data}}(\mathbf{X})$ while simultaneously contrasting the current model against its previous state $\pi_{\alpha_t}(\mathbf{X})$. In particular, Eqn. 10 amounts to

$$-\mathcal{L}_\pi(\alpha) \equiv \underbrace{D_{\text{KL}}(p_{\text{data}}(\mathbf{X})||\pi_\alpha(\mathbf{X}))}_{\text{match data density}} - \underbrace{D_{\text{KL}}(\pi_{\alpha_t}(\mathbf{X})||\pi_\alpha(\mathbf{X}))}_{\text{criticize itself}} \quad (13)$$

This objective reflects a form of *self-adversarial* learning, where the EBM is encouraged both to fit the data and to move away from its earlier approximation. At the same time, the difference of KL terms cancels the intractable partition function $\log \mathbf{Z}(\alpha)$, yielding a tractable, stable EBM learning objective.

An analogous interpretation holds for the generator objective in Eqn. 11, which can be expressed as

$$-\mathcal{L}_p(\omega) \equiv \underbrace{D_{\text{KL}}(p_{\text{data}}(\mathbf{X})||p_\omega(\mathbf{X}))}_{\text{match data density}} + \underbrace{D_{\text{KL}}(\pi_{\alpha_t}(\mathbf{X})||p_\omega(\mathbf{X}))}_{\text{match EBM density}} + \quad (14)$$

$$\underbrace{\mathbb{E}_{p_{\text{data}}(\mathbf{X})} [D_{\text{KL}}(p_{\omega_t}(\mathbf{z}|\mathbf{X})||p_\omega(\mathbf{z}|\mathbf{X}))] + \mathbb{E}_{\pi_{\alpha_t}(\mathbf{X})} [D_{\text{KL}}(p_{\omega_t}(\mathbf{z}|\mathbf{X})||p_\omega(\mathbf{z}|\mathbf{X}))]}_{\text{additional surrogate KL perturbation terms}}$$

Thus, the shared latent generator is learned to align its marginal distribution with both the empirical data and current EBM density. The additional KL terms over the latent posterior can be viewed as a majorization step Han et al. (2019). They treat \mathbf{z} as part of the complete data inferred at the current iteration and provide an upper bound on the marginal likelihood, which makes optimization more tractable. Unlike prior cooperative schemes defined for single-modality data, our formulation explicitly operates on multimodal inputs. The shared latent generator is therefore encouraged to synthesize *coherent* multimodal samples that are consistent across modalities and well matched to the EBM, which in turn improves EBM learning.

Learning the joint inference model by minimizing Eqn. 12 matches the corresponding latent samples from multimodal real data $\mathbf{X} \sim p_{\text{data}}(\mathbf{X})$ and multimodal synthesis $\mathbf{X} \sim \pi_{\alpha_t}(\mathbf{X})$. Specifically, given the optimal $\mathcal{M}_\alpha^{k_x}(\cdot)$ and $\mathcal{M}_\omega^{k_z}(\cdot)$ kernel Eqn. 12 becomes

$$-\mathcal{L}_q(\phi) \equiv \underbrace{\mathbb{E}_{p_{\text{data}}(\mathbf{X})} [D_{\text{KL}}(p_{\omega_t}(\mathbf{z}|\mathbf{X})||q_\phi(\mathbf{z}|\mathbf{X}))]}_{\text{real latent sample inference}} + \underbrace{\mathbb{E}_{\pi_{\alpha_t}(\mathbf{X})} [D_{\text{KL}}(p_{\omega_t}(\mathbf{z}|\mathbf{X})||q_\phi(\mathbf{z}|\mathbf{X}))]}_{\text{synthesis latent sample inference}} \quad (15)$$

Both terms push the joint inference model $q_\phi(\mathbf{z}|\mathbf{X})$ toward the generator posterior $p_{\omega_t}(\mathbf{z}|\mathbf{X})$, for both real and EBM-synthesized multimodal inputs. Since the posterior samples produced by MCMC (cf. Eqn. 7) are typically sharper and more accurate than the initial mixture-of-experts approximation, this objective gradually improves the quality of the latent initializations. In turn, better initializations make subsequent MCMC posterior refinement more effective, allowing the generator to explore the latent space more thoroughly and to model complex multimodal structure.

3.3 Model Generalization to Modal-specific Latent Variable

Notably, prior works (Sutter et al., 2020; Palumbo et al., 2023) extend the shared latent generative model by introducing additional modality-specific latent variables $\mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_M\}$, i.e.,

$$p_\omega(\mathbf{X}, \mathbf{z}, \mathbf{W}) = p_0(\mathbf{z}) \prod_{i=1}^M p_{\omega_i}(\mathbf{x}_i|\mathbf{z}, \mathbf{w}_i) p_0(\mathbf{w}_i) \quad q_\phi(\mathbf{z}, \mathbf{W}|\mathbf{X}) = q_{\phi_z}(\mathbf{z}|\mathbf{X}) \prod_{i=1}^M q_{\phi_{\mathbf{w}_i}}(\mathbf{w}_i|\mathbf{x}_i) \quad (16)$$

The modality-specific latent variable \mathbf{W} is introduced to capture inductive biases unique to each modality, thereby enhancing the representational capacity of the latent space and improving robustness in cross-modal inference scenarios. Our proposed learning framework can extend to this setting. The same MCMC-revised kernels can be applied jointly over $(\mathbf{X}, \mathbf{z}, \mathbf{W})$, so all model components are updated under the same cooperative scheme.

4 Related Work

Energy-Based Models. EBMs offer high modeling flexibility and have been widely studied in the context of maximum-likelihood training (Nijkamp et al., 2019; Du & Mordatch, 2019; Du et al., 2020; Xiao et al., 2020; Marks et al., 2025; Dutta et al., 2025). Beyond standard MLE, recent work has explored amortized sampling using generator networks (Han et al., 2019; Grathwohl et al., 2021; Kumar et al., 2019; Luo et al., 2024). For example, Luo et al. (2024) proposes learning conditional EBMs, while Han et al. (2019); Grathwohl et al. (2021); Kumar et al. (2019); Schröder et al. (2023b) learn marginal EBMs with auxiliary or complementary generators to approximate the EBM distribution. These amortized-MCMC methods differ fundamentally from our MCMC-based training in their learning objectives (see Appendix. B.1 for discussion). Another line of work, termed cooperative learning, uses a generator to initialize MCMC chains for more efficient EBM training (Xie et al., 2018; 2021). However, existing methods in this family have been developed primarily for single-modal data. Extending EBMs to multimodal settings introduces additional difficulty because the sampler should also account for complex inter-modal dependencies to produce coherent, cross-modally consistent samples.

Multimodal VAE. Multimodal VAEs have become a central paradigm for multimodal representation learning. MVAE (Wu & Goodman, 2018) introduces a product-of-experts (PoE) inference scheme that combines unimodal posteriors into a single joint posterior, enabling scalable training and principled handling of missing modalities. MMVAE (Shi et al., 2019) instead adopts a mixture-of-experts (MoE) inference to improve robustness under partial observations. MoPoE (Sutter et al., 2020; 2021) unifies these ideas via a mixture-of-products formulation that interpolates between PoE and MoE, balancing flexibility and expressivity. Various recent works have also demonstrated other informative strategies for factorizing the joint posterior, such as CoDEVAE Mancisidor et al. (2025), MWBVAE Qiu et al. (2025), HELVAE Vo & Valera (2026), and InvariantVAE Hirt et al. (2024). Beyond inference structure, several works focus on refining the latent space. MVTCVAE (Hwang et al., 2021) enforces cross-modal consistency through total-correlation regularization. MMVAE+ (Palumbo et al., 2023) augments the shared latent space with modality-specific priors, and MVEBM (Yuan et al., 2024) replaces the Gaussian prior with an energy-based prior to capture latent structure. CMVAE (Palumbo et al., 2024) explicitly promotes semantic clustering in the shared latent space.

Multimodal VAE with Diffusion. Recent multimodal methods further couple VAEs with diffusion or score-based models for refinement (Pandey et al., 2022a; Palumbo et al., 2024; Wesego & Rooshenas, 2024; 2023). Specifically, Diff-CMVAE (Palumbo et al., 2024) integrates DiffuseVAE (Pandey et al., 2022a), applying diffusion-based refinement to modality-specific outputs to significantly improve generation quality. ScoreMVAE (Wesego & Rooshenas, 2023) applies score-based refinement at the latent level. These approaches demonstrate that powerful diffusion backbones can substantially boost generation quality, but they typically operate as post-hoc refinement stages. Instead, our work directly learns a multimodal EBM over the joint data space and couples it with a shared latent generator and inference model through MCMC revision, so that the core generative model itself is improved without relying on a separate refinement step.

5 Experiment

Experiment Setting. Following prior multimodal VAE work Palumbo et al. (2023; 2024), we benchmark our method on PolyMNIST (Thomas M. Sutter, 2021) and Caltech-Birds (CUB) Image-Captions (Shi et al., 2019). PolyMNIST contains five visual modalities that depict the same digit class under different styles and backgrounds. CUB provides paired image and text modalities with rich semantics and substantial modality-specific variation, and is regarded as a challenging benchmark for multimodal generation and alignment. Additional experiments and supplementary results are provided in Appendix. A.

5.1 Multimodal Data Modelling

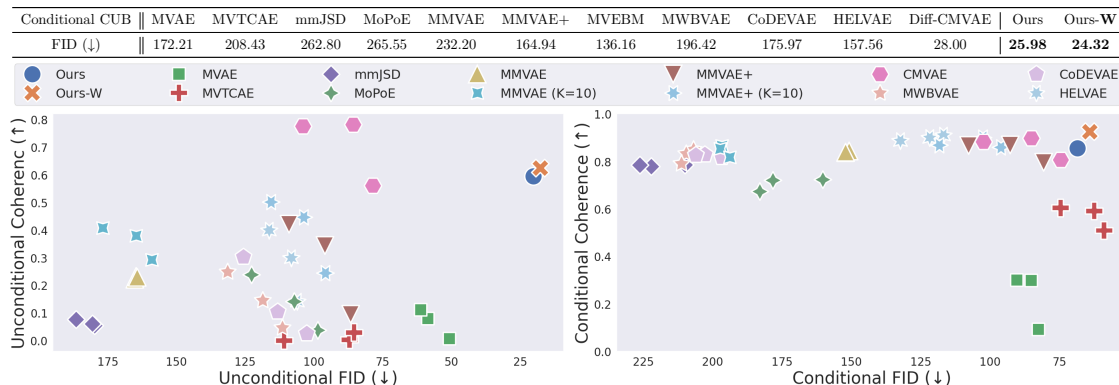


Figure 1: Comparison for unconditional and conditional multimodal synthesis on PolyMNIST (bottom), and comparison for conditional FID score on CUB (top). Additional quantitative and qualitative results are provided in Appendix.E.

We first assess whether our proposed learning scheme successfully yields high-quality, coherent multimodal synthesis. In our method, the generator is encouraged to approximate the EBM density and provide good initial states for EBM sampling, while the joint inference model offers informative latent initializations for posterior MCMC sampling. This cooperative interplay should contribute to improved synthesis quality and semantic coherence across modalities.

To quantitatively assess synthesis coherence, we follow standard practice and apply pre-trained classifiers³ to the generated samples. These classifiers evaluate whether generated samples correspond to the correct digit class; higher classification accuracy indicates stronger cross-modal alignment. We compare against strong multimodal VAE baselines on both unconditional and conditional generation. For CUB, the CMVAE results are taken from its diffusion-based variant (Diff-CMVAE), which integrates a diffusion model to substantially boost image quality. We denote our variant with modality-specific variables (Eqn. 16) as Ours-W.

Figure. 1 includes our variational baselines trained under different configurations (e.g., importance-weight sampling) for a broad comparison over the landscape. Across this spectrum, our method consistently achieves better coherence–quality performance, even when compared to diffusion-based methods (e.g., Diff-CMVAE Palumbo et al. (2024)). These results demonstrate that MCMC-revised cooperative learning effectively captures shared semantics, preserves modality-specific details, and yields superior performance.

5.2 Analysis of Learning Multimodal EBM with Complementary Models

In this section, we study how the complementary models affect the learning of multimodal EBM.

Learning Multimodal EBM *without* Complementary Models. We begin with the challenging setting in which the complementary multimodal shared generator is removed. In this case, the EBM learning objective in Eqn. 10 can be viewed to minimizing only the first KL term as $\min_{\alpha} D_{\text{KL}}(\Phi_{\omega_t, \phi_t}(\mathbf{X}, \mathbf{z}) || \pi_{\alpha}(\mathbf{X}) q_{\phi}(\mathbf{z} | \mathbf{X}))$, which reduces to standard MLE learning of EBM, i.e., $= \min_{\alpha} D_{\text{KL}}(p_{\text{data}}(\mathbf{X}) || \pi_{\alpha}(\mathbf{X}))$ since $\Phi_{\omega_t, \phi_t}(\mathbf{X}, \mathbf{z})$ is anchored on the data distribution (see details in Appendix. B). In this setting, we keep the same EBM architecture and perform EBM sampling using short-run Langevin dynamics initialized from random noise.

As shown in Figure. 2, the resulting EBM loss trajectory fluctuates strongly throughout training, even when we increase the number of Langevin steps. This behavior indicates that short-run MCMC from non-informative initial states does not adequately explore the high-dimensional multimodal space (see corresponding FID in Appendix. 9). In contrast, when we use our proposed learning framework with complementary models, the EBM loss curve becomes smoother, suggesting that the learned initializers are essential for multimodal EBM training.

³Pre-trained classifiers for each modality are provided by Palumbo et al. (2024).

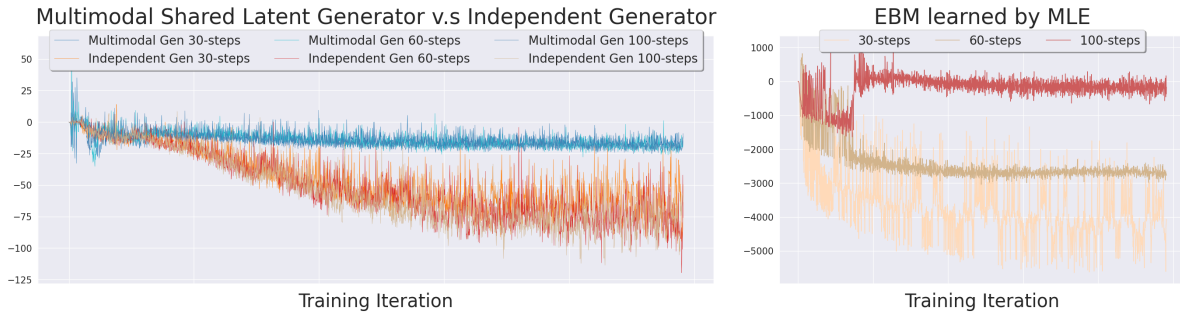


Figure 2: **Left:** EBM loss when using our shared latent generator versus independent per-modality generators; **Right:** EBM learned by MLE with noise-initialized Langevin dynamics with different sampling steps.

Learning Multimodal EBM with *Independent* Generators. We then examine the role of the multimodal shared latent generator. In our framework, the shared latent generator (Eqn. 4) factorizes a single latent variable \mathbf{z} to capture inter-modal dependencies, thereby producing coherent multimodal initializations for EBM sampling. To assess its importance, we replace it with M *independent* generators $p_{\omega_i}(\mathbf{x}_i, \mathbf{z}_i)$, each with its own latent variable \mathbf{z}_i and no explicit coupling across modalities.

Figure. 2 reports the corresponding EBM loss profiles (Eqn. 10). With independent generators, the EBM loss again shows strong fluctuations over training, even when we increase the number of EBM sampling steps (e.g., $k_{\mathbf{x}} = 60$ or 100). This behavior indicates that independently trained generators do not provide coherent multimodal initializations, making it difficult for the EBM to learn consistent cross-modal structure. In contrast, the shared latent generator yields smoother loss curves, highlighting that a shared latent representation is crucial for generating multimodal initial states and for effective multimodal EBM learning.

Learning Multimodal EBM with *Independent* Inference Models.

Finally, we investigate the effect of the multimodal joint inference model, which aggregates information from all available modalities to produce a shared latent initialization for posterior MCMC. To assess its impact, we replace the joint inference model with M *independent* inference models $p_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)$, each predicting a separate latent variable \mathbf{z}_i from a single modality and without explicit cross-modal coupling in latent space.

The corresponding generator loss profiles (Eqn. 11) are reported in Figure. 3. With independent inference models, the generator loss remains high and can even increase with more posterior sampling steps (e.g., $k_{\mathbf{z}} = 60$ or 100), indicating that the resulting latent initializations are not well aligned across modalities. In contrast, our multimodal joint inference model leads to a consistent decrease in generator loss. This suggests that shared latent initialization substantially improves generator learning, thereby supporting more effective multimodal EBM learning.

Taken together, these three analyses highlight that (i) learning multimodal EBMs directly from noise-initialized MCMC is difficult, and (ii) both the shared latent generator and the multimodal joint inference model are important for providing coherent initial states that facilitate effective multimodal EBM learning.

5.3 How Complementary Models Match with Their MCMC Revision?

It is crucial that the shared latent generator and joint inference model are well aligned with their corresponding MCMC-refined samples, so that they can provide good initial states for EBM sampling and posterior sampling, respectively. To assess this, we visualize the trajectories of EBM sampling and posterior sampling, each initialized from its complementary model and then refined via a finite number of Langevin steps.



Figure 3: Generator loss profiles for joint vs. independent inference models.

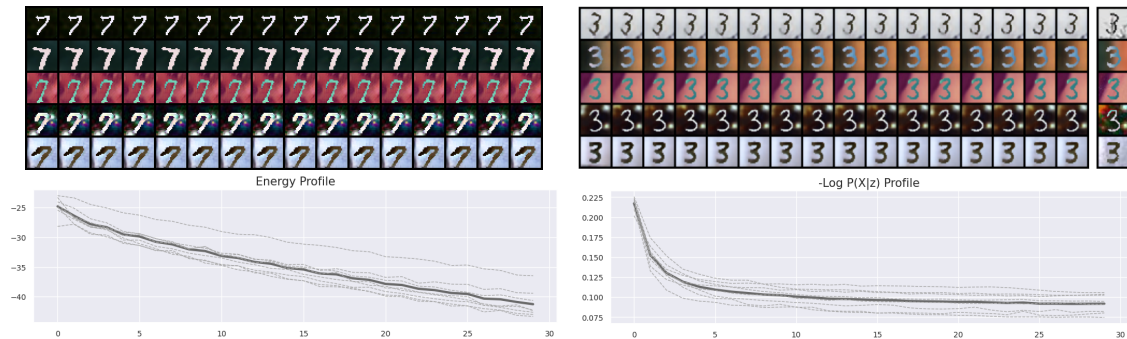


Figure 4: Trajectories of EBM sampling (**left**) and posterior sampling (**right**). Each row corresponds to one modality. The first column shows initial states from the initializer models; intermediate columns show intermediate samples (every 2 steps, up to 30 steps); the final column shows the MCMC-refined outputs. For posterior sampling, the rightmost column shows the observed input examples.

As shown in Figure. 4, the initializations from both the shared latent generator and the joint inference model are already semantically coherent across modalities (capturing the same digit class). Along the Langevin trajectories, only modest visual changes are observed, indicating that the initializer models closely match their MCMC-refined counterparts. At the same time, the energy profile $F_\alpha(\mathbf{X}^k)$ and the log-likelihood $\log p_\omega(\mathbf{X}|\mathbf{z}^k)$ continue to improve over iterations, demonstrating that the MCMC kernels still provide useful signals rather than merely reconstructing the initial states.

Overall, these observations show that the complementary models and their MCMC revisions are well matched, with the revision steps further guiding and sharpening both the EBM and the shared latent generator. In contrast to VAE-based approaches that rely on an external diffusion stage for high-quality refinement, our method learns the multimodal EBM, shared latent generator, and joint inference model in a unified framework, enabling both the EBM and the generator to produce high-quality multimodal samples directly (see additional results in Section. 5.4).

5.4 What Does MCMC Revision Cost and Provide?

Our framework introduces MCMC revision, which incurs computation cost compared to purely variational generators. To make this trade-off clear, we compare sampling time, NFE (Number of Function Evaluations), and parameter overhead, together with FID and CLIP scores on CUB.

Table 1: Comparison for sampling time (second / batch), NFE (\downarrow), FID (\downarrow), and CLIP scores (\uparrow).

	Ours (Gen)	Ours (EBM)	CMVAE	Diff-CMVAE	MMVAE
Sampling	0.001	0.08	0.001	38.12	0.001
NFE	1	31	1	251	1
FID (T2I)	26.15	25.98	155.11	28.00	232.20
FID (uncond)	21.45	20.72	141.00	N/A	213.89
CLIP (uncond)	0.280	0.284	0.263	N/A	0.231
CLIP (T2I)	0.278	0.282	0.260	0.272	0.242
CLIP (I2T)	0.286	0.290	0.272	N/A	0.235
Parameter Overhead	-	2M	-	14M	-

Table. 1 shows that MCMC revision yields consistent gains in both generation quality (lower FID) and cross-modal alignment (higher CLIP) at relatively modest cost. The EBM-augmented variant improves substantially over CMVAE and MMVAE, while adding only a small number of additional parameters and keeping sampling fast. At the same time, the shared latent generator alone already achieves strong performance with essentially one-step sampling, indicating that cooperative training makes the generator itself an effective multimodal model rather than just an auxiliary sampler. In comparison, diffusion-refined VAEs (e.g., Diff-CMVAE) rely on a separately trained, second-stage diffusion model, which entails higher sampling time and a larger parameter budget.

5.5 Scale-up to High-resolution and Large-scale Dataset



Figure 5: Unconditional synthesis on high-resolution CUB (left) and large-scale MSCOCO (right).

We test the scalability of our proposed method on the challenging high-resolution image (256x256) CUB data and the large-scale MSCOCO datasets. To better understand and assess the effectiveness endowed by our proposed learning method, we use the same network structures for all experiments. We visualize the unconditional and conditional multimodal synthesis in Figure. 5, suggesting our method effectively scales to higher resolutions and large-scale datasets while maintaining faithful multimodal synthesis quality. To further quantify this performance, we evaluate the generation quality and show results in Table. 2.

Table 2: FID (\downarrow) on challenging dataset.

	Ours (Gen)	Ours (EBM)	MMVAE+
CUB (256x256)	56.32	55.81	213.74
MSCOCO	68.94	68.10	187.22

5.6 Ablation Study

We investigate how the number of Langevin steps in the two MCMC kernels, $\mathcal{M}_\alpha^{k_x}$ for EBM sampling and $\mathcal{M}_\omega^{k_z}$ for posterior sampling, affects both performance and training cost (see Table. 3).

MCMC Steps of $\mathcal{M}_\alpha^{k_x}$. For EBM sampling, increasing k_x allows Langevin dynamics to explore the energy landscape more thoroughly, which in turn provides a stronger revision signal for the shared latent generator and yields a better learned EBM. On CUB, using a small number of steps ($k_x = 10$) leads to noticeably worse FID but lower training time, while a larger value ($k_x = 60$) improves FID but also doubles the cost. We thus report the setting ($k_x = 30$ and $k_z = 30$) of a favorable balance.

MCMC Steps of $\mathcal{M}_\omega^{k_z}$. For posterior sampling, increasing k_z improves the accuracy of samples from the generator posterior, which helps the joint inference model better match $p_\omega(\mathbf{z}|\mathbf{X})$ and thus supports more effective generator learning. When k_z is small (e.g., $k_z = 10$), FID degrades accordingly, whereas a larger value (e.g., $k_z = 60$) slightly improves FID but also incurs higher computational cost. The intermediate setting ($k_z = 30$ with $k_x = 30$) offers a good trade-off between performance and efficiency. These ablations suggest that MCMC revisions can yield most of the gains while keeping training practical.

Table 3: MCMC Steps for FID and training Time.

CUB	$k_x=60$	$k_x=10$	$k_x=30$ and $k_z=30$	$k_z=10$	$k_z=60$
FID	25.16	30.40	25.98	35.46	25.78
Time (seconds/iteration)	2.62	1.34	1.78	1.03	3.15

6 Conclusion

We propose a joint learning scheme that effectively learns the multimodal EBM by interweaving MLE updates of EBM, shared latent generator, and joint inference model via MCMC-based revision. The shared latent generator is learned to provide coherent initializations for EBM sampling, while the joint inference model is learned to offer starting points for posterior sampling. These MCMC samples serve as revision signals, guiding the complementary models, which in turn facilitate effective multimodal EBM sampling and learning.

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pp. 1692–1717. PMLR, 2023.
- Jiali Cui and Tian Han. Learning energy-based model via dual-mcmc teaching. *arXiv preprint arXiv:2312.02469*, 2023.
- Jiali Cui, Ying Nian Wu, and Tian Han. 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.
- Jiali Cui, Ying Nian Wu, and Tian Han. 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.
- Imant Daunhawer, Thomas M Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E Vogt. On the limitations of multimodal vaes. *arXiv preprint arXiv:2110.04121*, 2021.
- Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019.
- Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy based models. *arXiv preprint arXiv:2012.01316*, 2020.
- Debottam Dutta, Chaitanya Amballa, Zhongweiyang Xu, Yu-Lin Wei, and Romit Roy Choudhury. Learning energy-based variational latent prior for vaes. *arXiv preprint arXiv:2510.00260*, 2025.
- Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P Kingma. Learning energy-based models by diffusion recovery likelihood. *arXiv preprint arXiv:2012.08125*, 2020.
- 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. URL <https://openreview.net/forum?id=ixpSx09f1k3>.
- Tian Han, Yang Lu, Song-Chun Zhu, and Ying Nian Wu. Alternating back-propagation for generator network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
- 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 Conference on Computer Vision and Pattern Recognition*, pp. 8670–8679, 2019.
- Marcel Hirt, Domenico Campolo, Victoria Leong, and Juan-Pablo Ortega. Learning multi-modal generative models with permutation-invariant encoders and tighter variational objectives. *Trans. Mach. Learn. Res.*, 2024, 2024. URL <https://openreview.net/forum?id=1M4nHnxGfL>.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- Minghui Hu, Chuanxia Zheng, Zuopeng Yang, Tat-Jen Cham, Heliang Zheng, Chaoyue Wang, Dacheng Tao, and Ponnuthurai N. Suganthan. Unified discrete diffusion for simultaneous vision-language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL <https://openreview.net/forum?id=8JqINxA-2a>.

- HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Multi-view representation learning via total correlation objective. *Advances in Neural Information Processing Systems*, 34:12194–12207, 2021.
- Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Deqian Kong, Yuhao Huang, Jianwen Xie, Edouardo Honig, Ming Xu, Shuanghong Xue, Pei Lin, Sanping Zhou, Sheng Zhong, Nanning Zheng, and Ying Nian Wu. Molecule design by latent prompt transformer. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/a229cb89a98a84b2373496bb3cfc3570-Abstract-Conference.html.
- Deqian Kong, Dehong Xu, Minglu Zhao, Bo Pang, Jianwen Xie, Andrew Lizarraaga, Yuhao Huang, Sirui Xie, and Ying Nian Wu. Latent plan transformer for trajectory abstraction: Planning as latent space inference. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/df22a19686a558e74f038e6277a51f68-Abstract-Conference.html.
- Rithesh Kumar, Sherjil Ozair, Anirudh Goyal, Aaron Courville, and Yoshua Bengio. Maximum entropy generators for energy-based models. *arXiv preprint arXiv:1901.08508*, 2019.
- Duong H Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. One diffusion to generate them all. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2671–2682, 2025.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.
- Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Zichun Liao, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Omniflow: Any-to-any generation with multi-modal rectified flows. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13178–13188, 2025.
- Yihong Luo, Siya Qiu, Xingjian Tao, Yujun Cai, and Jing Tang. Energy-calibrated vae with test time free lunch. In *European Conference on Computer Vision*, pp. 326–344. Springer, 2024.
- Rogelio A Mancisidor, Robert Jenssen, Shujian Yu, and Michael Kampffmeyer. Aggregation of dependent expert distributions in multimodal variational autoencoders. In *Forty-second International Conference on Machine Learning*, pp. 2, 2025.
- Joanna Marks, Tim YJ Wang, and O Deniz Akyildiz. Learning latent energy-based models via interacting particle langevin dynamics. *arXiv preprint arXiv:2510.12311*, 2025.
- Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.
- Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. *Advances in Neural Information Processing Systems*, 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5272–5280, 2020a.
- Erik Nijkamp, Bo Pang, Tian Han, Linqi Zhou, Song-Chun Zhu, and Ying Nian Wu. Learning multi-layer latent variable model via variational optimization of short run MCMC for approximate inference. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, pp. 361–378. Springer, 2020b. doi: 10.1007/978-3-030-58539-6_22. URL https://doi.org/10.1007/978-3-030-58539-6_22.
- Emanuele Palumbo, Imant Daunhawer, and Julia E Vogt. Mmvae+: Enhancing the generative quality of multimodal vaes without compromises. In *The Eleventh International Conference on Learning Representations*. OpenReview, 2023.
- Emanuele Palumbo, Laura Manduchi, Sonia Laguna, Daphné Chopard, and Julia E Vogt. Deep generative clustering with multimodal diffusion variational autoencoders. In *International Conference on Learning Representations*, 2024.
- Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *Transactions on Machine Learning Research*, 2022a.
- Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *arXiv preprint arXiv:2201.00308*, 2022b.
- Bo Pang, Erik Nijkamp, Tian Han, and Ying Nian Wu. Generative text modeling through short run inference. *arXiv preprint arXiv:2106.02513*, 2021.
- Peijie Qiu, Wenhui Zhu, Sayantan Kumar, Xiwen Chen, Jin Yang, Xiaotong Sun, Abolfazl Razi, Yalin Wang, and Aristeidis Sotiras. Multimodal variational autoencoder: A barycentric view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20060–20068, 2025.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- Tobias Schröder, Zijing Ou, Jen Lim, Yingzhen Li, Sebastian Vollmer, and Andrew Duncan. Energy discrepancies: a score-independent loss for energy-based models. *Advances in Neural Information Processing Systems*, 36:45300–45338, 2023a.
- Tobias Schröder, Zijing Ou, Jen Lim, Yingzhen Li, Sebastian Vollmer, and Andrew Duncan. Energy discrepancies: a score-independent loss for energy-based models. *Advances in Neural Information Processing Systems*, 36:45300–45338, 2023b.
- Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in neural information processing systems*, 32, 2019.
- Thomas Sutter, Imant Daunhawer, and Julia Vogt. Multimodal generative learning utilizing jensen-shannon-divergence. *Advances in neural information processing systems*, 33:6100–6110, 2020.
- Thomas M Sutter, Imant Daunhawer, and Julia E Vogt. Generalized multimodal elbo. *arXiv preprint arXiv:2105.02470*, 2021.
- Julia E Vogt Thomas M. Sutter, Imant Daunhawer. Generalized multimodal elbo. In *9th International Conference on Learning Representations, ICLR, 2021*.
- Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

Huyen Khanh Vo and Isabel Valera. Hellinger multimodal variational autoencoders. *arXiv preprint arXiv:2601.06572*, 2026.

Daniel Wesego and Amirmohammad Rooshenas. Score-based multimodal autoencoders. *arXiv preprint arXiv:2305.15708*, 2023.

Daniel Wesego and Pedram Rooshenas. Multimodal elbo with diffusion decoders. *arXiv preprint arXiv:2408.16883*, 2024.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018.

Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. Vaebm: A symbiosis between variational autoencoders and energy-based models. *arXiv preprint arXiv:2010.00654*, 2020.

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, 2018.

Jianwen Xie, Zilong Zheng, and Ping Li. 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.

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. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=31d5RLCUuXC>.

Shiyu Yuan, Jiali Cui, Hanao Li, and Tian Han. Learning multimodal latent generative models with energy-based prior. In *European Conference on Computer Vision (ECCV)*, 2024.

A Additional Experiment

A.1 Latent Classification

In our method, the joint inference model is continuously updated to catch up with the MCMC-revised posterior, which makes its latent predictions more consistent across modalities and should therefore yield a more semantically structured shared latent space. We further examine whether the inferred latent variables capture shared high-level semantics across modalities.

Table 4: Accuracy for latent classifier.

Method	Ours	MVAE	MMVAE	MoPoE
Accuracy	0.962	0.926	0.835	0.944

Following prior work (Sutter et al., 2021), we train simple classifiers on the inferred latent representations and report their classification accuracy. If the shared latent space encodes meaningful semantic information, these classifiers should perform well and achieve high accuracy. Using our mixture-based joint inference model, we compute accuracy for each modality and report the average across all modalities. As shown in Table. 4, our method attains the highest latent classification accuracy among the compared multimodal VAEs, suggesting that the learned latent space is more semantically structured and better aligned with the underlying label information.

A.2 Latent Space Interpolation



Figure 6: Visualization of unconditional synthesis via Latent space interpolation.

We evaluate whether the shared latent generator model can produce smooth interpolations in the shared latent space, leading to gradual transitions in the multimodal data space. To this end, we perform linear interpolation in the latent space, $\tilde{\mathbf{z}} = (1 - \alpha) \cdot \mathbf{z}_1 + \alpha \cdot \mathbf{z}_2$. As shown in Figure. 6, the shared latent generator produces smooth and coherent transitions across modalities, indicating its ability to capture shared semantics and effectively explore the energy landscape.

A.3 MCMC Refinement on Cross-modal Inference

Following the evaluation protocols of our variational baselines, we assessed conditional coherence when only one modality is available in Figure. 1 in the main text. In our framework, MCMC posterior sampling offers an additional capability: it can refine latent variables inferred from multiple subsets of available modalities. This refinement step, which iteratively adjusts the latent variables toward better cross-modal consistency, is, however, not feasible for standard variational approaches.

In particular, given a set of observed modality data, our proposed method performs MCMC refinement in the latent space to further improve inference quality. Formally, let $\mathbf{X}_{\mathcal{O}}$ denote the set of observed modalities indexed by $\mathcal{O} \in \{1, \dots, M\}$. The generator posterior conditioned on this partial observation can be defined as $p_{\omega}(\mathbf{z}|\mathbf{X}_{\mathcal{O}}) \propto p_0(\mathbf{z}) \prod_{m \in \mathcal{O}} p_{\omega_m}(\mathbf{x}_m|\mathbf{z})$. The corresponding Langevin posterior refinement uses the gradient

$$\frac{\partial}{\partial \mathbf{z}} \log p_{\omega}(\mathbf{z}|\mathbf{X}_{\mathcal{O}}) = \frac{\partial}{\partial \mathbf{z}} \log p_0(\mathbf{z}) + \sum_{m \in \mathcal{O}} \frac{\partial}{\partial \mathbf{z}} \log p_{\omega_m}(\mathbf{x}_m|\mathbf{z}) \quad (17)$$

This subset-conditioned posterior follows from the factorization of the multimodal shared latent generative model and is defined under partial observations. We note that our multimodal generator is learned using full multimodal data, which learns a shared latent representation that captures inter-modal semantic structure.

Table 5: Coherence with MCMC refinement.

Number of Modality	1 ($n = 0$)	2 ($n = 1$)	3 ($n = 2$)	4 ($n = 3$)
Coherence	0.921	0.930	0.938	0.940

If our model is well-learned, the role of Langevin refinement is to improve the latent representation by incorporating evidence from the additional available modalities. Starting from an initialization $\mathbf{z}_0 \sim q_{\phi}(\mathbf{z}|\mathbf{X}_{\mathcal{O}})$, the refinement step adjusts the latent variable toward regions that better explain the observed subset under the generator likelihood. In this way, the refinement process seeks a latent representation that more closely corresponds to a shared multimodal mode consistent with the available observations. We report our results in Table. 5, where the conditional coherence becomes better with increasing number of available modalities under our MCMC refinement process.

B Theoretical Derivation

For MLE learning of the EBM objective (Eqn. 2), the gradient is derived as

$$\frac{\partial}{\partial \alpha} \mathcal{L}_{\pi}(\alpha) = \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \left[\frac{\partial}{\partial \alpha} \log \pi_{\alpha}(\mathbf{X}) \right] = \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \left[\frac{\partial}{\partial \alpha} F_{\alpha}(\mathbf{X}) \right] - \frac{\partial}{\partial \alpha} \log \mathbf{Z}(\alpha) \quad (18)$$

where $\frac{\partial}{\partial \alpha} \log \mathbf{Z}(\alpha)$ is derived as

$$\frac{\partial}{\partial \alpha} \log \mathbf{Z}(\alpha) = \frac{1}{\mathbf{Z}(\alpha)} \int \frac{\partial}{\partial \alpha} \exp [F_\alpha(\mathbf{X})] d\mathbf{X} = \int \pi_\alpha(\mathbf{X}) \frac{\partial}{\partial \alpha} [F_\alpha(\mathbf{X})] d\mathbf{X} = \mathbb{E}_{\pi_\alpha(\mathbf{X})} \left[\frac{\partial}{\partial \alpha} F_\alpha(\mathbf{X}) \right] \quad (19)$$

By applying Eqn. 19 to Eqn. 18, we have derived Eqn. 2.

For our EBM learning objective in Eqn. 10, the gradient is derived from the KL difference formulation

$$-\hat{\mathcal{L}}_\pi(\alpha) = D_{\text{KL}}(\Phi_{\omega_t, \phi_t}(\mathbf{X}, \mathbf{z}) || \pi_\alpha(\mathbf{X}) q_\phi(\mathbf{z} | \mathbf{X})) - D_{\text{KL}}(\Omega_{\omega_t, \alpha_t}(\mathbf{X}, \mathbf{z}) || \pi_\alpha(\mathbf{X}) q_\phi(\mathbf{z} | \mathbf{X})) \quad (20)$$

$$\begin{aligned} \text{with } \frac{\partial}{\partial \alpha} \hat{\mathcal{L}}_\pi(\alpha) &= \underbrace{\mathbb{E}_{\Phi_{\omega_t, \alpha_t}(\mathbf{X}, \mathbf{z})} \left[\frac{\partial}{\partial \alpha} F_\alpha(\mathbf{X}) \right]}_{\text{First KL Term}} - \frac{\partial}{\partial \alpha} \log \mathbf{Z}(\alpha) - \underbrace{\mathbb{E}_{\Omega_{\omega_t, \alpha_t}(\mathbf{X}, \mathbf{z})} \left[\frac{\partial}{\partial \alpha} F_\alpha(\mathbf{X}) \right]}_{\text{Second KL Term}} + \frac{\partial}{\partial \alpha} \log \mathbf{Z}(\alpha) \\ &= \mathbb{E}_{\Phi_{\omega_t, \alpha_t}(\mathbf{X}, \mathbf{z})} \left[\frac{\partial}{\partial \alpha} F_\alpha(\mathbf{X}) \right] - \mathbb{E}_{\Omega_{\omega_t, \alpha_t}(\mathbf{X}, \mathbf{z})} \left[\frac{\partial}{\partial \alpha} F_\alpha(\mathbf{X}) \right] \end{aligned}$$

in which, with respect to the gradient on α , the first KL divergence term corresponds to the MLE learning of EBM with samples anchored by the data distribution through $\Phi_{\omega_t, \alpha_t}(\mathbf{X}, \mathbf{z}) = p_{\text{data}}(\mathbf{X}) \cdot \mathcal{M}_{\omega_t}^{k_z} \cdot q_{\phi_t}(\mathbf{z} | \mathbf{X})$. For the second KL-divergence of its own critique, the partition function is canceled out, and samples are obtained from the MCMC-revised distribution $\Omega_{\omega_t, \alpha_t}(\mathbf{X}, \mathbf{z}) = \mathcal{M}_{\alpha_t}^{k_x} \cdot p_{\omega_t}(\mathbf{X} | \mathbf{z}) p_0(\mathbf{z})$ at the current optimization step under the stop-gradient update scheme.

For MLE learning of the shared latent generator objective (Eqn. 6), the gradient is derived as

$$\begin{aligned} \frac{\partial}{\partial \omega} \log p_\omega(\mathbf{X}) &= \mathbb{E}_{p_\omega(\mathbf{z} | \mathbf{X})} \left[\frac{\partial}{\partial \omega} \log p_\omega(\mathbf{X}) \right] \\ &= \mathbb{E}_{p_\omega(\mathbf{z} | \mathbf{X})} \left[\frac{\partial}{\partial \omega} \log p_\omega(\mathbf{X}) \right] + \mathbb{E}_{p_\omega(\mathbf{z} | \mathbf{X})} \left[\frac{\partial}{\partial \omega} \log p_\omega(\mathbf{z} | \mathbf{X}) \right] \\ &= \mathbb{E}_{p_\omega(\mathbf{z} | \mathbf{X})} \left[\frac{\partial}{\partial \omega} \log p_\omega(\mathbf{X}, \mathbf{z}) \right] \end{aligned} \quad (21)$$

where $\mathbb{E}_{p_\omega(\mathbf{z} | \mathbf{X})} \left[\frac{\partial}{\partial \omega} \log p_\omega(\mathbf{z} | \mathbf{X}) \right] = \int p_\omega(\mathbf{z} | \mathbf{X}) \left[\frac{\partial}{\partial \omega} \log p_\omega(\mathbf{z} | \mathbf{X}) \right] d\mathbf{z} = \frac{\partial}{\partial \omega} \int p_\omega(\mathbf{z} | \mathbf{X}) d\mathbf{z} = 0$.

B.1 Compared to Amortized-MCMC Method

Several recent advances have investigated EBM learning without explicit MCMC sampling Grathwohl et al. (2021); Han et al. (2019); Luo et al. (2024); Schröder et al. (2023a). These works study *single-modal* EBMs that employ amortized samplers to replace MCMC, thereby avoiding iterative sampling. In contrast, our focus is on the *multimodal setting*, which introduces two additional challenges: (i) effectively capturing the shared inter-modal relationships across heterogeneous modalities, and (ii) mitigating the mismatch induced by multimodal joint inference models. To address these challenges, we incorporate MCMC revision as a key component of our cooperative framework, which allows both the EBM and the generator posterior to be iteratively refined by each other, ensuring coherent multimodal alignment that cannot be achieved by amortized single-pass updates.

Moreover, the inclusion of MCMC revision makes our **learning objectives fundamentally different** from previous amortizing formulations. For clarity, and to directly illustrate the difference in learning dynamics independent of modality notation, we denote Ω, Φ as shorthand for MCMC-revised densities $\Omega_{\omega, \alpha}(\mathbf{X}, \mathbf{z}), \Phi_{\omega, \phi}(\mathbf{X}, \mathbf{z})$ and denote $Q = q_\phi(\mathbf{z} | \mathbf{X}) p_{\text{data}}(\mathbf{X}), \Pi = \pi_\alpha(\mathbf{X}) q_\phi(\mathbf{z} | \mathbf{X}), P = p_\omega(\mathbf{X}, \mathbf{z})$ for joint densities of amortized models. We denote AM for methods using amortized models without MCMC.

Learning the EBM (α): The corresponding KL terms in our method (Eqn. 10) and AM are: $\min_\alpha KL(\Phi || \Pi) - KL(\Omega || \Pi)$ v.s. $\min_\alpha KL(Q || \Pi) - KL(P || \Pi)$. Our formulation leverages MCMC-revised samples (i.e., joint densities of Ω and Φ), whereas AM relies solely on ancestral samples (Q and P). Because our samples are refined by the EBM itself, they provide a more accurate approximation of the target energy landscape $KL(M_{\alpha_t} p_{\omega_t}(\mathbf{X}) || \pi_{\alpha_t}(\mathbf{X})) \leq KL(p_{\omega_t}(\mathbf{X}) || \pi_{\alpha_t}(\mathbf{X}))$. This results in more effective and stable

EBM learning and leverages the contextual modelling capability of EBM to effectively guide the multimodal generator model.

Learning the (shared) generator model (ω): For learning the generator model (Eqn. 11), KL terms for ours and AM are $\min_{\theta} KL(\Phi\|P) + KL(\Omega\|P)$ v.s. $\min_{\theta} KL(Q\|P) + KL(P\|\Pi)$. The learning dynamics differ substantially. In our case, the generator is trained with MCMC-revised latent samples, yielding a closer match to the true generator posterior: $KL(M_{\omega_t, q_{\phi_t}}(\mathbf{z}|\mathbf{X})\|p_{\omega_t}(\mathbf{z}|\mathbf{X})) \leq KL(q_{\phi_t}(\mathbf{z}|\mathbf{X})\|p_{\theta_t}(\mathbf{z}|\mathbf{X}))$, which aims to address the mismatch between the generator posterior and joint inference model (analysis in Sec. 3.1). In addition, $KL(\Omega\|P)$ learns to match the revised MCMC samples from EBM-refined samples, while the Amortizer method intends to chase the major modes of $\pi_{\alpha}(\mathbf{X})$ through variational approximation (i.e., $KL(P\|\Pi)$). Hence, our generator directly learns from revised multimodal samples that can better capture inter-modal consistency.

Learning the (joint) inference model (ϕ): For the inference model (Eqn. 12), learning objectives for ours and AM are: $\min_{\phi} KL(\Phi\|Q) + KL(\Omega\|\Pi)$ v.s. $\min_{\phi} KL(Q\|P) + KL(P\|\Pi)$. The two approaches differ in both learning source and optimization target. Our inference network amortizes latent MCMC refinement on observed data (i.e., $KL(\Phi\|Q)_{CompilerError}$), while AM performs pure variational inference (i.e., $KL(Q\|P)$). On generated samples, our model matches EBM-revised generator samples (i.e., $KL(\Omega\|\Pi)$), whereas AM directly uses ancestral generator outputs ($KL(P\|\Pi)$), which can lead to sub-optimal inference quality.

C Inference Mechanism under Missing Modalities

Our framework follows the standard inference mechanism established in multimodal VAEs for handling missing modalities. Given any available observed modality \mathbf{x}_i , we first infer its shared latent variable, and then use this latent variable to generate the missing modalities through the shared latent generator for \mathbf{x}_j where $j \neq i$. This mechanism is identical to prior multimodal VAE baselines, ensuring fair comparison and consistent inference behavior. In our experiments, we evaluate using the same inference mechanism as in baseline models to ensure fairness, and the results consistently show superior reconstruction and coherence.

D Implementation

EBM Block (in_ch, out_ch, downsample, head)	EBM Network on PolyMNIST (nef)
Input: x ReLU if head Conv(in_ch, out_ch), ReLU, Conv(out_ch, out_ch) Downsample(factor=2) if downsample output: h	Input: \mathbf{X} h = concat(Conv(each \mathbf{x}) along channel dim EBM Block(nc, nef, downsample=True, head=True) EBM Block (nef, nef, downsample=True) EBM Block (nef, nef, downsample=False) EBM Block (nef, nef, downsample=False) ReLU, Downsample(factor=8), Linear(nef, 1) output: h
Input: x Downsample(factor=2) if head Conv(in_ch, out_ch) if downsample Downsample(factor=2) if downsample and not head output: y	Input: \mathbf{X} img = Conv(img), txt=Linear(ReLu(Linear(txt emb))) EBM Block(nc, nef, downsample=True, head=True) EBM Block (nef, nef, downsample=True) EBM Block (nef, nef, downsample=False) EBM Block (nef, nef, downsample=False) ReLU, Downsample(factor=8), Linear(Concat(h,txt), 1) output: h
output: h + y	

Table 6: We use generator and inference network structures from (Palumbo et al., 2023; 2024). For our EBM energy function structures, we denote the operation of convolution as **Conv** (input channel, output channel, k=3, s=1, p=1), where k is the kernel size, s is the stride number, and p is padding value. We conduct Upsample and Downsample via *interpolate* and *avg_pool2d* operations.

E Supplementary Result

Corresponding to our Figure. 1, we additionally report quantitative results in Table. 7 and Table. 8, where we report only the best performance of our baselines. Qualitative results can be seen in Figure. 7, Figure. 8, Figure. 9, and Figure. 10.

Table 7: Comparison of synthesis coherence.

Methods	PolyMNIST	
	Unconditional	Conditional
MVAE	0.112	0.301
MVTCAE	0.029	0.604
mmJSD	0.076	0.785
MoPoE	0.238	0.723
MMVAE	0.232	0.844
MMVAE+	0.421	0.869
MVEBM	0.735	0.857
CMVAE	0.781	0.897
MWBVAE	0.297	0.850
CoDEVAE	0.431	0.828
HELVAE	0.508	0.910
Ours	0.594	0.855
Ours-W	0.624	0.921

Table 8: Comparison of multimodal synthesis quality.

Methods	PolyMNIST	
	Unconditional	Conditional
MVAE	50.65	82.59
MVTCAE	85.43	58.95
mmJSD	179.76	178.27
MoPoE	98.56	160.29
MMVAE	164.29	150.83
MMVAE+	86.64	80.75
MVEBM	75.43	70.45
CMVAE	78.52	74.53
MWBVAE	111.45	206.88
CoDEVAE	102.61	196.64
HELVAE	106.05	116.83
Ours	20.12	68.52
Ours-W	17.65	64.12

Table 9: Unconditional FID on EBM learned by MLE.

Ours	MLE learned		
	30-steps	60-steps	100-steps
20.12	125.23	101.34	122.92



1. A black bird is up with a short, short bill.
2. The bird has a small surface and oak tree which are black yellowed branches.
3. This bird has yellow with brown on its chest and has a very short beak.
4. This bird has wings that are black and have a brown crown.
5. This is a blue bird bird with white chest.
6. The bird has a green chest and black eye rings.
7. This particular bird has a belly that has white and yellow color.
8. The bird has a small brown bill with brown shoulder that also appear to be juvenile.
9. A blue bird with a chevron and something.
10. This bird has a white neck and wings that are grey and has a short bill.
11. This bird is brown coloured with a redhead and has a long crest.
12. This bird is white and grey in color, with it having few black wings.

Figure 7: Unconditional generation on CUB.

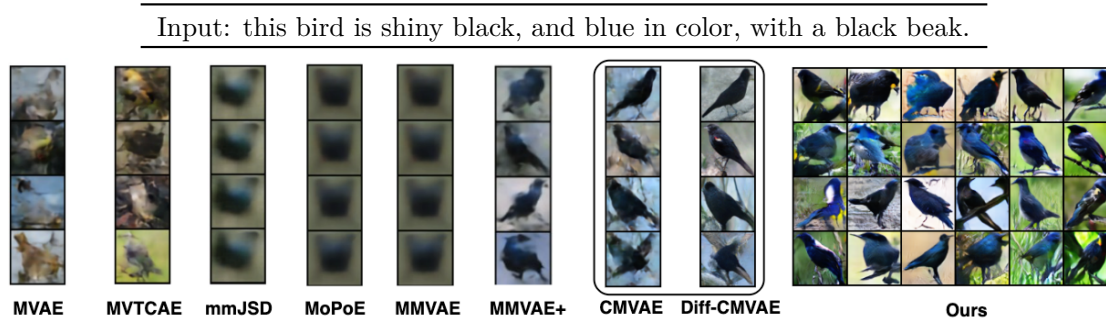


Figure 8: Conditional generation on CUB. Baseline results are taken from (Palumbo et al., 2023). CMVAE and Diff-CMVAE results are reproduced with codes provided by Palumbo et al. (2024); Pandey et al. (2022b).

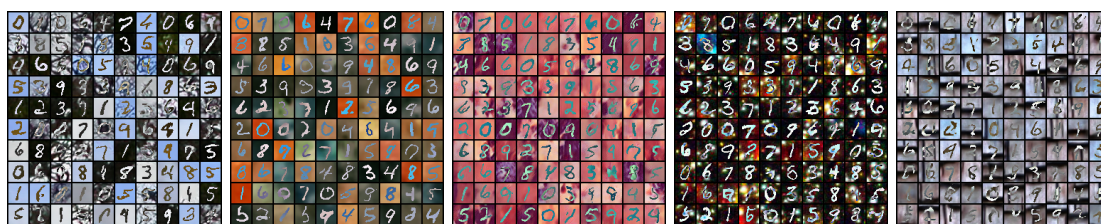


Figure 9: Unconditional generation on PolyMNIST.

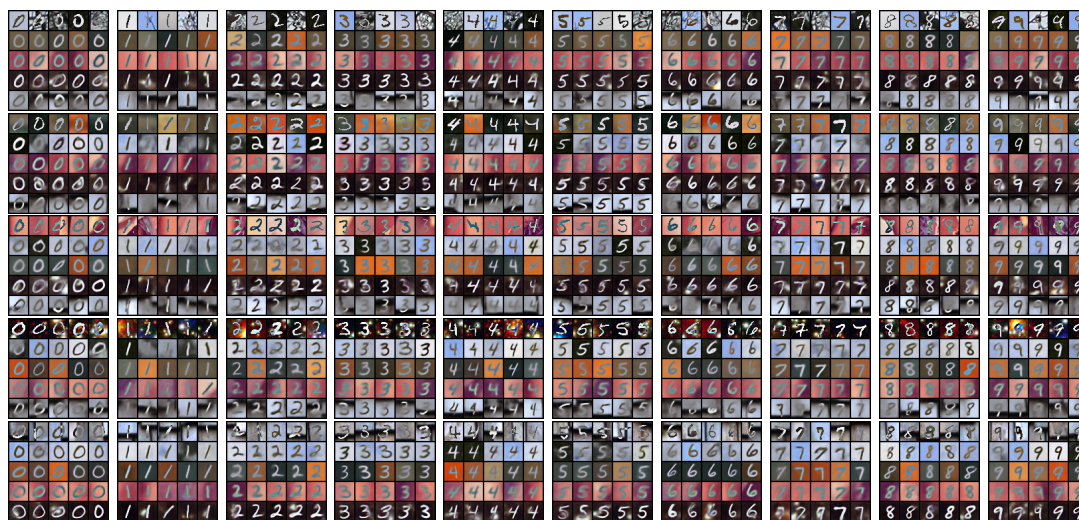


Figure 10: Conditional generation on PolyMNIST. From top to bottom, available modality from 1 to 5. In each block, the first row shows the given input modality, while the subsequent rows display the generated outputs for the remaining missing modalities.

F Broader Impact Statement

This work advances multimodal energy-based generative modeling, and therefore shares the usual risks associated with powerful generative models, such as potential misuse for creating misleading synthetic content or amplifying biases present in training data. At the same time, better joint modeling of multimodal structure may benefit applications such as modality prediction, provided that practitioners use appropriate safeguards and responsible data practices.

G Algorithm

Algorithm 1 Learning Scheme

Require: a multimodal EBM α , a multimodal shared latent generator ω , a joint inference model ϕ , Markov transition kernels with stop-gradient for both $\hat{\mathcal{M}}_\alpha^{k_x}(\cdot)$ and $\hat{\mathcal{M}}_\omega^{k_z}(\cdot)$ with Langevin steps k_x and k_z ;

- 1: **for** # training iteration **do**
 - 2: Draw training sample $\mathbf{X}_{\text{data}} \sim p_{\text{data}}(\mathbf{X})$.
 - 3: Draw latent prior sample $\mathbf{z}_{\text{prior}} \sim p_0(\mathbf{z})$.
 \triangleright *MCMC-revised sample.*
 - 4: Draw inference latent sample $\mathbf{z}_{\text{Inf}} \sim q_\phi(\mathbf{z}|\mathbf{X}_{\text{data}})$ with \mathbf{X}_{data} .
 - 5: Draw generator revised latent sample $\mathbf{z}_{\text{Gen-revised}} \sim \hat{\mathcal{M}}_\omega^{k_z}(\cdot)$ by Eqn. 8 based on \mathbf{z}_{Inf} .
 - 6: Draw generator sample $\mathbf{X}_{\text{Gen}} \sim p_\omega(\mathbf{X}|\mathbf{z}_{\text{prior}})$ with prior latent sample $\mathbf{z}_{\text{prior}}$.
 - 7: Draw EBM revised sample $\mathbf{X}_{\text{EBM-revised}} \sim \hat{\mathcal{M}}_\alpha^{k_x}(\cdot)$ by Eqn. 3 based on \mathbf{X}_{Gen} .
 - 8: Draw EBM revised latent sample $\mathbf{z}_{\text{EBM-revised}} \sim q_\phi(\mathbf{z}|\mathbf{X}_{\text{EBM-revised}})$ based on $\mathbf{X}_{\text{EBM-revised}}$.
 \triangleright *Model update.*
 - 9: Update EBM parameter α using Eqn. 10 with \mathbf{X}_{data} and $\mathbf{X}_{\text{EBM-revised}}$.
 - 10: Update Generator parameter ω using Eqn. 11 with \mathbf{X}_{data} , $\mathbf{z}_{\text{Gen-revised}}$, \mathbf{X}_{Gen} , and $\mathbf{X}_{\text{EBM-revised}}$.
 - 11: Update Inference parameter ϕ using Eqn. 12 with \mathbf{z}_{Inf} , $\mathbf{z}_{\text{Gen-revised}}$, $\mathbf{z}_{\text{prior}}$, and $\mathbf{z}_{\text{EBM-revised}}$.
 - 12: **end for**
-

H Pytorch PseudoCode

```

1 import torch as t
2 import torch.nn as nn
3
4 data_loader = get_dataloader(dataset, batch_size)
5 netG, netI, netE = get_networks(dataset)
6
7 optG = t.optim.Adam(netG.parameters(), lr=1e-3)
8 optE = t.optim.Adam(netE.parameters(), lr=4e-4)
9 optI = t.optim.Adam(netI.parameters(), lr=1e-3)
10
11 e_l_steps, e_l_step_size, e_n_step_size = 30, 0.1, 0.001
12 z_l_steps, z_l_step_size, z_n_step_size = 30, 0.1, 0.1
13
14 latent_dim = 32
15 pz = get_distribution(t.distributions.Normal, latent_dim)
16 qz = get_distribution(t.distributions.Normal, latent_dim)
17
18 dataset = "PolyMNIST"
19 batch_size = 256
20
21 mse = nn.MSELoss(reduction='none').cuda()
22
23 def log_mean_exp(value, dim=0, keepdim=False):
24     return t.logsumexp(value, dim, keepdim=keepdim) - math.log(value.size(dim))
25
26 def langevin_x(x_init):
27     x = [x.clone().detach().requires_grad_(True) for x in x_init]
28
29     for steps in range(e_l_steps):
30         energy = netE(x)
31         energy = energy.sum()
32         grad = t.autograd.grad(energy, x)
33         for d, x_i in enumerate(x):

```

```

34         x_i.data = x_i.data - 0.5 * e_l_step_size * e_l_step_size * grad[d] + e_n_step_size * t.
           randn_like(x_i).data
35
36     return [x_i.detach() for x_i in x]
37
38 def langevin_z(z_init, x_data):
39     z = [z.clone().detach().requires_grad(True) for z in z_init]
40     views = len(z_init)
41     for steps in range(z_l_steps):
42         recon_value = [[None for _ in range(views)] for _ in range(views)]
43
44         for e in range(views):
45             for d in range(views):
46                 rec = netG(z[e], v_idx=d)
47                 recon_value[e][d] = mse(rec, x_data[d])
48
49         nls = []
50         for r in range(views):
51             lpz = pz.log_prob(z[r])
52             nlpx = [px_u for px_u in recon_value[r]]
53             nlpxu = t.stack(nlpxu).sum(0)
54             nl = nlpxu - lpz
55             nls.append(nlw)
56         nls = t.stack(nls).mean(0)
57         nls = nls.sum(0)
58
59         grad = t.autograd.grad(nls, z)
60         for d, z_i in enumerate(z):
61             z_i.data = z_i.data - 0.5 * z_l_step_size * z_l_step_size * grad[d] + z_n_step_size * t.
               randn_like(z_i).data
62
63     return [z_i.detach() for z_i in z]
64
65 for i, x in enumerate(data_loader):
66     x = [x_i.cuda() for x_i in x]
67     views = len(x)
68
69     z_prior = pz.rsample()
70     samples_init = netG(z_prior)
71     samples_corr = langevin_x(samples_init)
72
73     z_q_mu_init, z_q_lv_init = netI(x)
74     z_q_init = qz(z_q_mu_init, z_q_lv_init)
75     z_q_corr = langevin_z(z_q_init, x)
76
77     optG.zero_grad()
78     recon_value = [[None for _ in range(views)] for _ in range(views)]
79
80     for e in range(views):
81         for d in range(views):
82             rec = netG(z[e], v_idx=d)
83             recon_value[e][d] = mse(rec, x[d])
84
85     nls = []
86     for r in range(views):
87         nlpx = [px_u for px_u in recon_value[r]]
88         nlpxu = t.stack(nlpxu).sum(0)
89         nls.append(nlpxu)
90     nls = t.stack(nls).mean(0)

```

```
91     nls = nls.mean(0)
92
93     errS = mse(samples_init, samples_corr)
94     errG = nls + errS
95     errG.backward()
96     optG.step()
97
98     optI.zero_grad()
99     z_p_mu, z_p_lv = netI(samples_corr)
100
101     nlqz_true = []
102     nlqz_fake = []
103     for r in range(views):
104         lqz_true = log_mean_exp(t.stack([sum_flat(qz_x.log_prob(z_q_corr[r])) for qz_x in qz(z_q_mu_init,
105             z_q_lv_init)]))
106         nlqz_true.append(- lqz_true)
107         lqz_gen = log_mean_exp(t.stack([sum_flat(qz_x.log_prob(z_prior)) for qz_x in qz(z_p_mu, z_p_lv)]))
108         nlqz_fake.append(- lqz_gen)
109
110     nlqz_true = t.stack(nlqz_true).mean(0)
111     nlqz_fake = t.stack(nlqz_fake).mean(0)
112     errI = nlqz_true + nlqz_fake
113     errI.backward()
114     optI.step()
115
116     optE.zero_grad()
117     E_t = netE(x)
118     E_f = netE(samples_corr)
119     errE = (E_t - E_f) / (e_n_step_size/e_l_step_size)**2
120     errE.backward()
121     optE.step()
```

Listing 1: PyTorch code used in our experiments.