ShaLa: Multimodal Shared Latent Generative Modelling

Jiali Cui[1], **Yan-Ying Chen**[2], **Yanxia Zhang**[2], **Matthew Klenk**[2]

[1]Stevens Institute of Technology, Hoboken, NJ
[2]Toyota Research Institute, Los Altos, CA
jcui7@stevens.edu, {yan-ying.chen, yanxia.zhang, matt.klenk}@tri.global

Abstract

This paper presents a novel generative framework for learning shared latent representations across multimodal data. Many advanced multimodal methods focus on capturing all combinations of modality-specific details across inputs, which can inadvertently obscure the high-level semantic concepts that are shared across modalities. Notably, Multimodal VAEs with low-dimensional latent variables are designed to capture shared representations, enabling various tasks such as joint multimodal synthesis and cross-modal inference. However, multimodal VAEs often struggle to design expressive joint variational posteriors and suffer from low-quality synthesis. In this work, ShaLa addresses these challenges by integrating a novel architectural inference model and a second-stage expressive diffusion prior, which not only facilitates effective inference of shared latent representation but also significantly improves the quality of downstream multimodal synthesis. We validate ShaLa extensively across multiple benchmarks, demonstrating superior coherence and synthesis quality compared to state-of-the-art multimodal VAEs. Furthermore, ShaLa scales to many more modalities while prior multimodal VAEs have fallen short in capturing the increasing complexity of the shared latent space.

Code — https://jcui1224.github.io/shared-latent-space-proj

1 Introduction

Deep generative models have demonstrated remarkable success in generating high-fidelity outputs across a variety of single modalities (Ho, Jain, and Abbeel 2020a; Karras et al. 2020). These advances have then rapidly extended to multimodal generation tasks, such as text-to-image, image-to-text, and other cross-modal flows (Ramesh et al. 2022; Alayrac et al. 2022; Li et al. 2023). However, many of these approaches are tailored to specific generation directions (i.e., *single-flow* model) and often require dedicated models for each modality pair (Chen et al. 2020; Bao et al. 2023). To address this limitation, recent work (Xu et al. 2023; Le et al. 2025) has explored unified frameworks capable of handling multiple generation flows within a single model (i.e., *multi-flow* model). Yet, such models tend to emphasize capturing modality-specific combinations and fine-grained details, while often ignoring the semantic structure that conceptually links different modalities (Wu and Goodman 2018; Thomas M. Sutter 2021; Han, Xing, and Wu 2018) (See Fig.1).

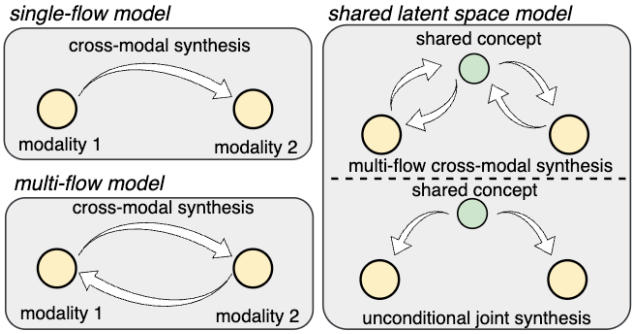

Figure 1: Comparison of multimodal modeling paradigms. **Left**: Single-flow and multi-flow models typically learn direct mappings between modalities such as images, text, or views. **Right**: Shared latent space models capture modality-invariant shared concepts, enabling both multi-flow cross-modal synthesis and unconditional joint generation.

To align with the way human cognition abstracts high-level concepts across different sensories, Multimodal Variational Autoencoders (VAEs) have emerged as a promising class of generative frameworks (Shi et al. 2019; Palumbo, Daunhawer, and Vogt 2023; Palumbo et al. 2024). By projecting multimodal inputs into a low-dimensional shared latent space, they aim to capture modality-invariant semantic representations, forming the basis of what we refer to as shared latent variable generative modeling.

However, a major bottleneck in multimodal VAEs lies in the design of the joint inference model. Most existing approaches are built upon two dominant paradigms: Product of Experts (PoE) (Wu and Goodman 2018) and Mixture of Experts (MoE) (Shi et al. 2019). While PoE models enforce agreement across modalities, they struggle under missing-modality scenarios. Alternatively, MoE models are more robust to missing inputs but lack the expressivity needed to accurately capture complex joint posteriors (Daunhawer et al. 2021). Besides this trade-off, the variational learning paradigm itself introduces an additional challenge: the prior-hole problem, a distributional mismatch between the aggre-

[1]Work Done during an internship at Toyota Research Institute.

gated joint posterior and the assumed prior (e.g., Gaussian or Laplacian). This gap often leads to poor sample quality and severely limits the effectiveness of multimodal VAEs in downstream generation tasks (Aneja et al. 2021; Cui and Han 2024).

The recent success of latent diffusion models has opened new horizons for generative modeling (Rombach et al. 2022; Vahdat, Kreis, and Kautz 2021). These models are typically trained in two stages: first, a VAE is used to encode data into a compact latent space; second, a diffusion model is trained to model the aggregated posterior distribution in that space, thereby bridging the prior-posterior gap and improving synthesis quality. However, most existing works focus on single-flow mappings (e.g., text-to-image) (Rombach et al. 2022) or use attention mechanisms to implicitly fuse inputs in multi-flow settings (Xu et al. 2023; Le et al. 2025), without learning a shared latent variable that captures common semantic structure across modalities.

In this work, we introduce **ShaLa** — short for **Sha**red **La**tent space modeling — a generative framework for learning multimodal shared latent representation. ShaLa brings together the strengths of Multimodal VAEs and diffusion models, explicitly learning shared semantic representations through latent variables while simultaneously addressing two primary challenges: inference under missing modalities and low-quality synthesis due to the prior-hole problem.

At its core, ShaLa adopts an architectural inference model that encodes each modality into deterministic features and fuses them into a joint representation. This fused representation acts as an information bottleneck, conditioning the shared latent variable and promoting semantic alignment across modalities. On top of this, ShaLa employs a second-stage diffusion prior over the shared latent space, conditioned on these deterministic representations. The diffusion model is trained to approximate the aggregated joint posterior with modality-aware guidance, thereby overcoming the prior-hole issue and enabling robust, coherent generation. Importantly, ShaLa supports flexible inference even in the presence of missing modalities and scales to complex multi-view settings with a large number of modality instances.

In summary, our contributions include: (i) We propose ShaLa, a novel generative framework that unifies the architectural inference and diffusion model for jointly modelling the multimodal shared latent space. (ii) We conduct extensive experiments on standard datasets, demonstrating that ShaLa achieves state-of-the-art coherence and synthesis quality compared to other shared latent variable generative models. (iii) We performed evaluations on the challenging multi-view dataset, which features a significantly larger number of modalities (16 views), suggesting the scalability of ShaLa for learning complex shared latent variables.

2 Related Work

2.1 Multimodal Deep Generative Model

To place multimodal VAEs in context, it is important to recognize the broader trajectory of research on multimodal generative modeling. In recent years, multimodal advances have enabled applications such as text-to-image synthesis (Ramesh et al. 2021; Saharia et al. 2022), audio-visual generation (Owens et al. 2016; Mo and Raj 2023), and multi-sensory robotics (Lee et al. 2019). However, many of these models are tailored for specific cross-modal tasks (i.e., *single-flow* models), typically using modality-specific architectures. To address this limitation, recent studies have proposed unified frameworks that support multiple generation flows within a single model, referred to as *multi-flow* models (Xu et al. 2023; Le et al. 2025). Our approach differs in that it centers on explicitly learning a shared latent space. This line of work complements existing trends in unified modeling by providing a probabilistic framework grounded in the shared representation of multimodal data. See Fig.1 for comparison.

2.2 Multimodal VAE

Multimodal VAEs have emerged as a principled and flexible class of models for learning joint representations across modalities within a shared latent space. Early efforts (Suzuki, Nakayama, and Matsuo 2016; Hsu and Glass 2018; Vedantam et al. 2018) proposed factorized latent structures and separate inference networks for each modality subset. However, such designs suffer from poor scalability, as the number of inference networks grows exponentially with the number of modalities.

To address this, MVAE (Wu and Goodman 2018) introduced the Product-of-Experts (PoE) formulation, which combines unimodal posteriors into a single joint distribution, enabling efficient training. This formulation was extended by MoPoE (Sutter, Daunhawer, and Vogt 2020), which generalized the encoder to a Mixture-of-Products, allowing richer combinations of modality subsets. Parallelly, MMVAE (Shi et al. 2019) proposed the Mixture-of-Experts (MoE) to improve robustness for missing modalities, and follow-up work (Sutter, Daunhawer, and Vogt 2021) proposed hybrid PoE/MoE formulations to balance expressivity and flexibility.

Other directions focused on improving the latent space itself. MVTCAE (Hwang et al. 2021) encouraged cross-modal consistency via total correlation regularization. MMVAE+ (Palumbo, Daunhawer, and Vogt 2023) introduced modality-specific priors, while MVEBM (Yuan et al. 2024) proposed energy-based priors to capture richer structures than the Gaussian prior. CMVAE (also known as D-CMVAE) (Palumbo et al. 2024) enforced clustering in the latent space for better semantic separation. Note that CMVAE integrates DiffuseVAE (Pandey et al. 2022) to the data generated by each modality, but the diffusion process is applied to multiple high-dimensional data outputs. Instead, ShaLa employs a diffusion process in a single low-dimensional shared latent space, allowing faster sampling and scaling to an increasing number of modalities.

Despite these advancements, a major challenge remains: most multimodal VAE variants either rely on rigid inference structures or suffer from limitations in modeling complex joint distributions, particularly in the presence of modality dropout or missing data. Our work addresses this by proposing a more expressive prior and a flexible inference structure enabled by diffusion-based modeling.

3 Methodology

In this section, we review two inference paradigms: PoE and MoE, widely adopted in Multimodal VAEs, and summarize their strengths and limitations (cf. Tab.1). We then discuss the prior-hole problem, followed by our approach ShaLa.

3.1 Preliminary: Shared Latent Variable Model

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$ represent an observation composed of M modalities, where each $\mathbf{x}_i$ corresponds to one modality. The true multimodal data distribution is denoted $p_{\text{data}}(\mathbf{X})$, and our goal is to approximate it using a parameterized model $p_\theta(\mathbf{X})$. Shared latent variable models introduce a global latent variable $\mathbf{z}$ to capture common semantic information across modalities, i.e., $p_\theta(\mathbf{X}) = \int p_\theta(\mathbf{X}, \mathbf{z}) d\mathbf{z}$.

$$\begin{aligned} p_\theta(\mathbf{X}, \mathbf{z}) &= p_\theta(\mathbf{X}|\mathbf{z}) p_0(\mathbf{z}) \quad \text{where} \\ p_\theta(\mathbf{X}|\mathbf{z}) &= p_{\theta_1}(\mathbf{x}_1|\mathbf{z}) p_{\theta_2}(\mathbf{x}_2|\mathbf{z}) \cdots p_{\theta_M}(\mathbf{x}_M|\mathbf{z}) \end{aligned} \tag{1}$$

Here, $p_0(\mathbf{z})$ is the prior distribution over the latent variable, typically assumed to be a standard Gaussian: $p_0(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, where d is the latent dimensionality. The conditional distribution $p_\theta(\mathbf{X}|\mathbf{z})$ defines a set of modality-specific decoders parameterized by $\theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$, each mapping the latent code $\mathbf{z}$ to an individual modality $\mathbf{x}_M$.

To train this model, one can use maximum likelihood estimation (MLE), which maximizes the log-likelihood, i.e., $\max_\theta \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{N} \log p_\theta(\mathbf{X}_i)$. With a sufficiently large N, this is equivalent to minimizing the KL divergence between the true data distribution and the model distribution, i.e., $\text{KL}(p_{\text{data}}(\mathbf{X}) || p_\theta(\mathbf{X}))$. The gradient can be written as $\frac{\partial}{\partial \theta} \mathcal{L}(\theta) = \mathbb{E}_{p_{\text{data}}(\mathbf{X}) p_\theta(\mathbf{z}|\mathbf{X})} \left[\frac{\partial}{\partial \theta} \log p_\theta(\mathbf{X}, \mathbf{z}) \right]$. However, evaluating this gradient requires access to the true joint generator posterior $p_\theta(\mathbf{z}|\mathbf{X})$, which is typically *intractable*.

3.2 Preliminary: Multimodal VAE

To address the intractability of the true posterior, Multimodal VAEs leverage Variational Autoencoders (VAEs) (Kingma and Welling 2013) as an inference network (or encoder) and optimize a tractable surrogate objective known as the Evidence Lower Bound (ELBO). Hence, the central question becomes how to design an effective joint inference distribution $q_\phi(\mathbf{z}|\mathbf{X})$. Two main paradigms have emerged:

Product-of-Experts (PoE) The PoE introduced by MVAE (Wu and Goodman 2018), defines the joint posterior approximation as the product of modality-specific encoders:

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

Each encoder provides complementary evidence about $\mathbf{z}$, and the product enforces consistency across modalities, leading to sharp and coherent posteriors when all modalities are present. However, the product becomes ill-defined when one or more modalities are missing. Therefore, PoE alone does not naturally support flexible inference under partial observations, e.g, cross-modal inference, and the training procedure remains heuristic.

Method	Modelling Expressivity	Cross-modal Inference
Product of Expert	✓	✗
Mixture of Expert	✗	✓
ShaLa	✓	✓

Table 1: Comparison with PoE and MoE in terms of expressivity and support for cross-modal inference. In contrast, ShaLa facilitates both capabilities. See Sec.3.3 and Sec.3.4.

Mixture-of-Experts (MoE) To better facilitate missing modalities, MMVAE (Shi et al. 2019) proposes the MoE.

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

This formulation inherently supports partial input by excluding missing modalities from the mixture. However, due to its averaging nature, MoE often produces over-smoothed posteriors and weak modality alignment. As such, it may struggle to model fine-grained semantic interactions between modalities, limiting its expressiveness (Daunhawer et al. 2021).

Prior-hole Problem Another major challenge in multimodal VAEs arises from the mismatch between the aggregated posterior and the fixed prior, known as the prior-hole problem (Aneja et al. 2021; Cui and Han 2024). The aggregated posterior is defined as $q_\phi(\mathbf{z}) = \int q_\phi(\mathbf{z}|\mathbf{X}) p_{\text{data}}(\mathbf{X}) d\mathbf{X}$. In practice, this distribution often occupies only a narrow subset of the latent space, leaving large regions of the prior unaligned with any training data. As a result, prior samples may fall into these low-density "holes" (see Fig.2), leading to low-quality generations.

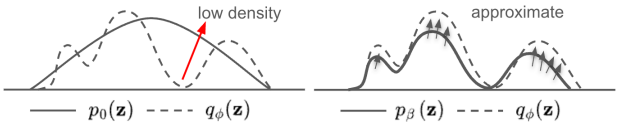

Figure 2: Prior-hole problem: the aggregated posterior (dashed line) occupies a narrow region of the latent space, while the prior (solid line in the left figure) covers a broader area, leading to mismatched samples. Our diffusion prior (solid line in the right figure) is learned to approximate this aggregated posterior, bridging the distribution gap.

3.3 ShaLa: Architectural Inference Model

We now introduce ShaLa to overcome two major challenges in multimodal VAEs: (1) the design of a flexible and expressive joint inference model, and (2) the prior–posterior mismatch known as the prior-hole problem. ShaLa addresses these issues by combining a novel architectural inference model with a second-stage diffusion-based prior.

Unlike PoE and MoE, ShaLa adopts a more direct approach by parameterizing the joint posterior $q_{\phi_i}(\mathbf{z}|\mathbf{X})$ as a single conditional Gaussian distribution, whose parameters are inferred from the entire set of input modalities. Rather than relying on explicit unimodal posteriors, we encode each modality $\mathbf{x}_i$ into a deterministic representation $\mathbf{h}_i$ using a

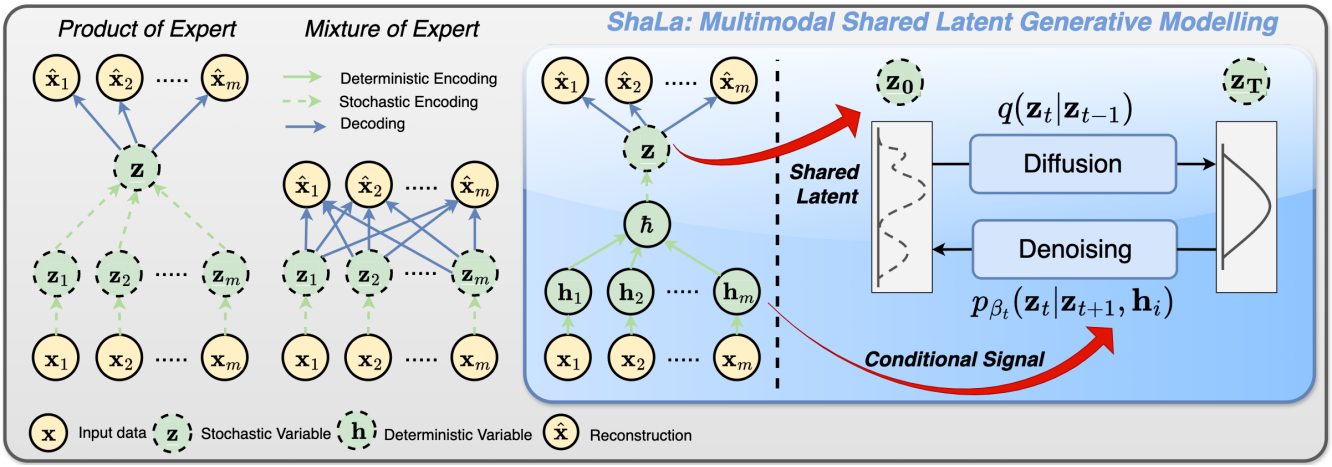

Figure 3: **Left**: PoE and MoE models construct the joint posterior by combining modality-specific *stochastic* encoders, each introducing limitations under missing modalities or expressivity constraints. **Right**: ShaLa replaces stochastic encoders with *deterministic* modality-specific embeddings fused into a shared representation, which then conditions a latent diffusion prior, enabling robust and flexible inference while bridging the prior-posterior gap for high-quality multimodal generation.

modality-specific encoder composed of multiple convolutional and down-sampling layers. These deterministic embeddings are then fused via a shared function $\odot(\cdot)$, implemented in practice as concatenation followed by several linear layers, into a global summary $\hbar$, which conditions the final posterior:

$$\begin{aligned} \mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M &= I_\phi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M) \\ \hbar &= \odot(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M) \\ \mathbf{z} &\sim \mathcal{N}(\mu_\phi(\hbar), V_\phi(\hbar)) \end{aligned} \tag{2}$$

Here, $\hbar$ serves as an information bottleneck, summarizing modality-specific semantics into a compact, high-level abstraction. By conditioning solely on this fused representation $\hbar$, our approach avoids the complexity of combining individual posteriors while retaining the ability to capture rich cross-modal interactions. Importantly, this formulation naturally supports guiding the generative process from the fused features and also forms the foundation for our diffusion-based prior in the next stage.

3.4 ShaLa: Expressive Prior via Latent Diffusion

While our architectural inference model offers a compact parameterization of the joint posterior $q_\phi(\mathbf{z}|\mathbf{X})$, it inherently assumes access to all modalities at inference time. In particular, since we do not model per-modality posteriors (as in PoE or MoE), our inference pathway cannot readily accommodate missing modalities at test time.

To overcome this limitation and simultaneously address the prior-hole problem, we introduce a second-stage latent diffusion model that serves as a flexible prior over the shared latent space. This diffusion-based prior enables cross-modal inference by conditioning its generative process on deterministic modality-specific features $\mathbf{h}_i$, allowing generation from any subset of modalities without retraining the encoder.

Forward and Reverse Process We adopt the denoising diffusion probabilistic model (DDPM) (Ho, Jain, and Abbeel 2020b), applied in shared latent space. Given a latent sample $\mathbf{z}_0 \sim q_\phi(\mathbf{z})$ from the aggregated posterior, we define a forward diffusion trajectory $\mathbf{z}_{0:T} = \{\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_T\}$ using the Gaussian transition:

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \alpha_t \mathbf{z}_{t-1}, \sigma_t^2 \mathbf{I}) \tag{3}$$

Here, $\alpha_t := \sqrt{1-\sigma_t^2}$ controls signal preservation, and σ_t is defined by a pre-specified schedule. As $t \to T$, $\mathbf{z}_T$ approaches an isotropic Gaussian $q(\mathbf{z}_T) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. The full forward trajectory is:

$$q(\mathbf{z}_{0:T}) = q(\mathbf{z}_0) \prod_{t=1}^{T} q(\mathbf{z}_t|\mathbf{z}_{t-1}) \tag{4}$$

where $q(\mathbf{z}_0) = q_\phi(\mathbf{z}) = \int q_\phi(\mathbf{z}|\mathbf{X}) p_{\text{data}}(\mathbf{X}) d\mathbf{X}$. This process gradually transforms latent representations into noise, which is our prior distribution defined in Eqn.1.

To model the generative prior, we learn a reverse process that bridges from our prior distribution back to the aggregated posterior distribution of interest as

$$p_{\beta_t}(\mathbf{z}_t|\mathbf{z}_{t+1}) = \mathcal{N}(\mathbf{z}_t; \mu_{\beta_t}(\mathbf{z}_{t+1}), V_{\beta_t}(\mathbf{z}_{t+1})) \tag{5}$$

The full generative trajectory is

$$p_\beta(\mathbf{z}_{0:T}) = p(\mathbf{z}_T) \prod_{t=0}^{T-1} p_{\beta_t}(\mathbf{z}_t|\mathbf{z}_{t+1}) \tag{6}$$

This learned prior thus bridges the mismatch between the aggregated posterior and the assumed Gaussian prior, resolving the prior-hole problem and improving generation quality.

Cross-modal Conditioning To support cross-modal inference, i.e., sampling from $p_\beta(\mathbf{z}|\text{subset of modalities})$, we condition the reverse process on deterministic modality-specific embeddings, $\mathbf{h}_1, \ldots, \mathbf{h}_M$ from Eqn.2. During training, we adopt a random conditioning strategy inspired by

prior work (Zhang et al. 2023; Bie et al. 2024), which improves flexibility in multimodal generation. Concretely, we randomly sample one available modality-specific embedding $\mathbf{h}_j \in \{\mathbf{h}_1, \ldots, \mathbf{h}_M\}$ and condition the reverse step as

$$p_{\beta_t}(\mathbf{z}_t|\mathbf{z}_{t+1}, \mathbf{h}_j) = \mathcal{N}(\mu_{\beta_t}(\mathbf{z}_{t+1}, \mathbf{h}_j), V_{\beta_t}(\mathbf{z}_{t+1}, \mathbf{h}_j)) \quad \text{where} \quad \mathbf{h}_j \sim \text{Uniform}(\mathbf{h}_{1:M}) \tag{7}$$

This approach ensures that any single modality can serve as conditioning input, enabling robust inference even when other modalities are missing.

We train the diffusion prior via MLE, optimizing the marginal likelihood of latent trajectories under the learned generative model. Following common DDPM practices (Ho, Jain, and Abbeel 2020b), we optimize a time-averaged surrogate by randomly sampling diffusion step t and computing $\frac{\partial}{\partial\beta}\mathbb{E}_{\text{Uni}(t),\text{Uni}(\mathbf{h}_{1:m}),q(\mathbf{z}_{0:T})}\left[\log p_{\beta_t}(\mathbf{z}_t|\mathbf{z}_{t+1}, \mathbf{h}_j)\right]$.

Joint and Cross-modal Sampling: Our framework supports both unconditional generation and conditional cross-modal synthesis. During training, we randomly drop conditioning signals to enable the model to learn both conditional and unconditional pathways within a unified diffusion framework. This allows the model to flexibly control the strength of conditioning using a guidance scale, enabling high-quality generation under both complete and incomplete observations.

4 Experiment

Our experimental objective is to evaluate the learned latent representation of ShaLa in terms of multimodal coherence and scalability compared to several established baselines of shared latent variable generative modeling. Toward that goal, we assess ShaLa in the following tasks: (1) generate semantically coherent outputs across modalities, (2) perform conditional cross-modal inference, (3) scale to many more modalities in multi-view generation tasks, and (4) maintain the quality and structure of the shared latent space.

Baselines To ensure fair and consistent comparisons, we benchmark ShaLa against a diverse set of representative shared latent variable generative models, including MVAE (Wu and Goodman 2018), MVTCAE (Hwang et al. 2021), mmJSD (Sutter, Daunhawer, and Vogt 2020), MoPoE (Sutter, Daunhawer, and Vogt 2021), MMVAE (Shi et al. 2019), MMVAE+ (Palumbo, Daunhawer, and Vogt 2023), CMVAE (Palumbo et al. 2024), and MVEBM (Yuan et al. 2024).

Datasets We utilize standard multimodal datasets used in our baselines, such as PolyMNIST (Sutter, Daunhawer, and Vogt 2021), MNIST-SVHN-Text (MST) (Sutter, Daunhawer, and Vogt 2020), and Caltech UCSD Birds (CUB) (Wah et al. 2011), which provide clear and interpretable benchmarks for assessing generation quality and semantic coherence across modalities.

4.1 Shared Latent for Synthesis Coherence

A central objective of shared latent variable generative models is to render semantically coherent outputs across multiple modalities. This includes both unconditional joint generation, i.e., sampling all modalities from the latent prior, and conditional cross-modal inference, i.e., generating target modalities from given modalities.

Method	PolyMNIST		MST	
	Unconditional	Conditional	Unconditional	Conditional
MVAE	0.008	0.298	0.12	0.27
MVTCAE	0.003	0.591	-	-
mmJSD	0.060	0.778	-	0.72
MoPoE	0.141	0.720	0.31	0.69
MMVAE	0.232	0.844	0.28	0.68
MMVAE+	0.344	0.869	-	-
MVEBM*	0.735	0.857	0.42	0.43
CMVAE	0.781	**0.897**	-	-
Ours	**0.815**	**0.897**	**0.44**	**0.75**

Table 2: Coherence (↑) for unconditional and conditional cross-modal generation. Note that several baselines do not report results on the MST benchmark. We nonetheless include these models to present a broad landscape of multimodal generative approaches.

Unconditional Joint Coherence We assess ShaLa's capacity for coherent multimodal synthesis on the two benchmarks, PolyMNIST and MST, widely adopted in our direct baselines which use quantitative metrics computed using pre-trained modality-specific classifiers (Sutter, Daunhawer, and Vogt 2021). These classifiers assess whether the different modalities generated from the same latent code correspond to the same semantic class. Higher coherence indicates a better alignment of modalities in the latent space and thus a more accurate modeling of the joint distribution.

Unconditional generation forms the bedrock of generative models, and reveals the quality of the shared latent. We evaluate ShaLa's ability to generate all modalities simultaneously from samples drawn from the learned latent prior $\mathbf{z}$. As shown in Tab.2, ShaLa demonstrates superior coherence in this setting, consistently outperforming our baselines.

Conditional Cross-modal Coherence We next assess ShaLa in conditional generation, where one modality is used to infer a latent representation that is then used to generate unseen modalities. Specifically, given an input $\mathbf{x}_i$, we encode it into a deterministic representation $\mathbf{h}_i$ and use it to condition the diffusion-based prior, which samples the latent variable $\mathbf{z}$. This latent is then decoded into the target modality $\mathbf{x}_j$. The semantic coherence of $\mathbf{x}_j$ is again measured using a pre-trained classifier to determine whether it aligns with the semantic label of the observed input $\mathbf{x}_i$.

ShaLa consistently achieves high coherence across all modality pairs, reflecting strong alignment and semantic consistency between observed and generated modalities. This effectiveness is largely due to ShaLa's conditioning mechanism, which enables flexible cross-modal generation by leveraging deterministic modality-specific embeddings within a learned diffusion prior.

4.2 Shared Latent for Synthesis Quality

Besides multimodal coherence, a major challenge in variational frameworks is the prior-hole problem, where latent samples drawn from the prior often fall into low-density regions of the aggregated posterior. This misalignment leads to

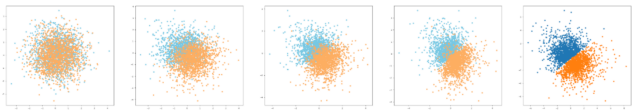

Figure 4: Visualization of shared latent space. The **leftmost panel** shows Gaussian prior samples. The **center panels** show diffusion approximation. The **rightmost panel** shows posterior (encoded) samples.

degraded sample quality and unrealistic generations. ShaLa addresses this issue through the integration of a second-stage diffusion-based latent prior, bridging the distributional gap between the learned posterior and the assumed prior.

To assess the learned latent prior, we conduct a 2D latent space visualization on PolyMNIST using only digits 0 and 1. As shown in Fig.4, the posterior samples form two clusters that correspond to digits 0 and 1, which indicates that ShaLa successfully extracts *semantic structure* shared across multimodal inputs. As sampling progresses, the learned prior distribution approximates the posterior, bridging the distributional gap and producing informative latents that are semantically coherent with the underlying digit identity.

We evaluate the quality of the data generated from this shared latent using full PolyMNIST and CUB, which is considered a more complex and realistic multimodal benchmark. We report Fréchet Inception Distance (FID) as our primary quantitative metric. As shown in Tab.3, ShaLa achieves lower FID scores compared to competitive baselines of multimodal shared representation learning, indicating superior sample quality. Qualitative results on CUB in Fig.**??** illustrate that ShaLa's shared latent can generate visually realistic and semantically meaningful multimodal samples in both unconditional and conditional settings. These results demonstrate that ShaLa achieves high perceptual fidelity, validating the effectiveness of its diffusion-based latent prior in overcoming key limitations of traditional variational approaches. We report the best results as originally published in our baselines to ensure consistency.

Method	PolyMNIST		CUB
	Unconditional	Conditional	Conditional
MVAE	50.65	82.59	172.21
MVTCAE	85.43	58.98	208.43
mmJSD	179.76	178.27	262.80
MoPoE	98.56	160.29	265.55
MMVAE	164.29	150.83	232.20
MMVAE+	86.64	80.75	164.94
MVEBM	-	-	136.16
CMVAE	78.52	74.53	28.00
Ours	**47.30**	**40.18**	**25.58**

Table 3: Generation quality by FID (↓).

4.3 Scalability for Multi-View Challenge

To further assess the scalability of ShaLa, we evaluate its performance on multi-view datasets, where different viewpoints of the same object are treated as distinct modalities. Specifically, we conduct experiments on ShapeNet (Chang et al. 2015), a widely used benchmark for view-consistent synthesis. In this setting, each view captures complementary visual information of the same 3D object, analogously to how multimodal data share the same semantic content.

Figure 5: Example 16-view samples from ShapeNet Car.

Experiment Setting We frame this task as a challenging instance of multimodal generative modeling, due to the significantly increased number of modalities (i.e., views). Following standard practice (Liu et al. 2023; Anciukevičius et al. 2023), we render 16 canonical views per object, each resized to a resolution of 128×128. We assess cross-view synthesis quality, i.e., generating unseen views, using two standard image similarity metrics: peak-signal-to-noise-ratio (PSNR) and structural-similarity-index-measure (SSIM).

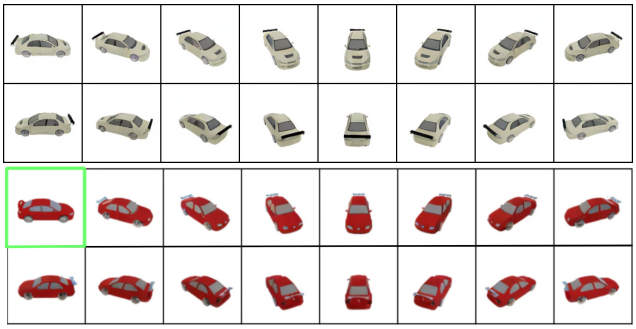

Figure 6: ShaLa unconditional and cross-modal multi-view generation. The input views are denoted by green boxes.

Benchmark Performance We adapt MMVAE+ (Palumbo, Daunhawer, and Vogt 2023) and CMVAE (Palumbo et al. 2024), both of which serve as prominent shared latent variable backbones in multimodal VAEs. As a reference, we also report results from *task-specific view synthesis baselines*, including SyncD (Liu et al. 2023), Px-NeRF (Yu et al. 2021), EG3D (Chan et al. 2022), and RenderD (Anciukevičius et al. 2023). These methods incorporate *explicit 3D priors or inductive biases* to maintain geometric consistency and are tailored for the view synthesis task. In contrast, ShaLa is trained *without any 3D supervision or assumptions*, relying solely on image views and latent alignment. As shown in Tab.4, ShaLa significantly outperforms MMVAE+ and CMVAE. Compared to task-specific methods, it achieves competitive performance, highlighting the flexibility and effectiveness of ShaLa as a unified generative modeling framework for both traditional multimodal and multi-view data.

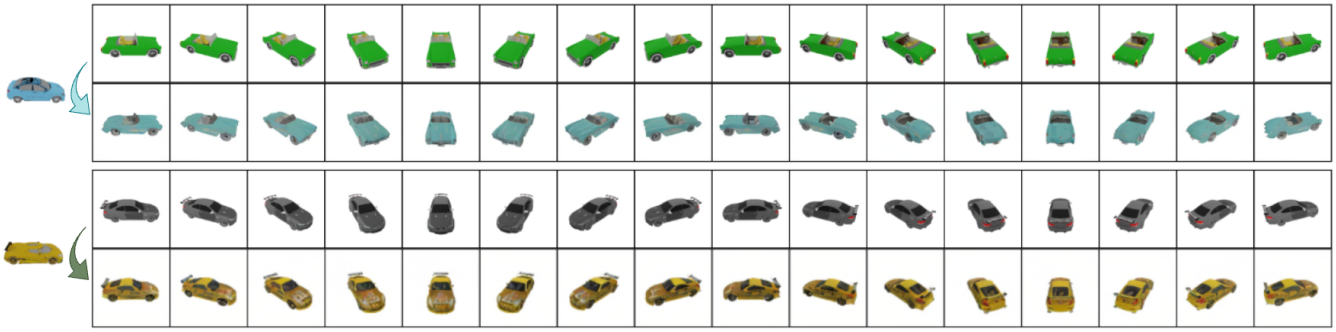

Figure 7: Visualization of latent transfer. The left column is $\mathbf{X}^{\text{ref}}$, the first row is $\mathbf{X}^{\text{source}}$, and the second row is the result.

	Ours	MMVAE+	CMVAE	Px-NeRF	EG3D	RenderD	Ours*	SyncD
PSNR	24.7	19.3	20.5	23.2	21.8	25.4	22.2	21.9
SSIM	0.89	0.59	0.64	0.90	0.71	0.81	0.88	0.88

Table 4: PSNR (↑) and SSIM (↑) for novel view synthesis. Ours* use image size 256×256 as SyncD for fairness.

4.4 Latent-Based Multi-View Style Transfer

By modeling a meaningful and flexible shared latent space, ShaLa can support a range of downstream tasks that go beyond traditional joint and conditional generation. In this section, we explore multi-view style transfer, a novel application setting that transfers a reference style to multiviews where the generated views not only carry the style (e.g., texture of car) but also maintain the structure presented in the source views (e.g, shape of car). This can be challenging for prior multimodal VAEs lack a conditioning mechanism, or task-specific models such as SyncD, which are typically constrained to only novel-view synthesis.

Given a source instance $\mathbf{X}^{\text{source}}$ and a reference instance $\mathbf{X}^{\text{ref}}$, each comprising multiple views, the goal is to transfer low-level visual features (e.g., color, texture) from the reference to the source while preserving the source's structural semantics (e.g., shape). We achieve this by first sampling a latent code $\mathbf{z}^{\text{source}} \sim q_\phi(\mathbf{z}|\mathbf{X}^{\text{source}})$ and applying K steps of forward diffusion to obtain a partially perturbed latent $\mathbf{z}_K^{\text{source}}$. This serves to remove view-specific low-level information while retaining higher-level semantic structure. Then, we condition the reverse diffusion process on a deterministic feature $\mathbf{h}_j^{\text{ref}}$ randomly selected from $\mathbf{X}^{\text{ref}}$ to resample the latent code for the decoder to synthesize outputs while adopting style features from the reference.

Qualitative results in Fig.7 demonstrate ShaLa's ability to produce faithful structural preservation with successful style transfer across views. These results highlight the expressive and compositional nature of the shared latent representation learned by ShaLa.

4.5 Ablation Study

Architectural Inference In Eqn.2, our architectural inference model encodes each modality $\mathbf{x}_i$ into a deterministic embedding $\mathbf{h}_i$, which is then fused to produce a shared latent posterior. This formulation treats $\mathbf{h}_{1:m}$ as an information bottleneck, providing a compact and semantically meaningful conditioning signal for the second-stage diffusion prior. To validate this, we compare performance when conditioning the diffusion model directly on $\mathbf{x}_{1:m}$ versus on $\mathbf{h}_{1:m}$. As shown in Tab.5, using $\mathbf{h}_{1:m}$ improves both unconditional and conditional synthesis, suggesting the benefit of our architectural inference model.

Condition	PolyMNIST		MST	
	Unconditional	Conditional	Unconditional	Conditional
$\mathbf{x}_{1:m}$	0.584	0.612	0.23	0.35
$\mathbf{h}_{1:m}$	**0.815**	**0.897**	**0.44**	**0.75**

Table 5: Coherence of using different conditional signals.

Ablation of Diffusion Prior We assess the effect of the diffusion prior by varying the capacity of the diffusion network. As shown in Tab.6, reducing the model size leads to noticeable drops in generation quality (as measured by FID), while increasing the model size yields consistent improvements. This suggests that ShaLa's shared latent space provides a stable foundation across different diffusion configurations, and that our second-stage diffusion prior effectively leverages this representation for high-quality synthesis.

FID / #P (parameter)	2× #P	1× #P	1/2× #P	1/4× #P
Unconditional	**45.24**	47.30	50.24	56.44
Conditional	**37.15**	40.18	45.86	49.75

Table 6: Varying diffusion configuration.

5 Conclusion

In this work, we present a novel framework for learning shared latent space, coined as ShaLa. We leverage an architectural inference model for learning the shared latent variables, where the conditional deterministic variables can facilitate the second-stage shared latent diffusion model. Extensive experiments demonstrate favorable results of ShaLa and the scalability in modelling the difficult multi-view data.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Anciukevičius, T.; Xu, Z.; Fisher, M.; Henderson, P.; Bilen, H.; Mitra, N. J.; and Guerrero, P. 2023. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12608–12618.

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

Bao, F.; Nie, S.; Xue, K.; Cao, Y.; Li, C.; Su, H.; and Zhu, J. 2023. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22669–22679.

Bie, F.; Yang, Y.; Zhou, Z.; Ghanem, A.; Zhang, M.; Yao, Z.; Wu, X.; Holmes, C.; Golnari, P.; Clifton, D. A.; et al. 2024. Renaissance: A survey into ai text-to-image generation in the era of large model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16123–16133.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; Xiao, J.; Yi, L.; and Yu, F. 2015. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Cui, J.; and Han, T. 2024. Learning Latent Space Hierarchical EBM Diffusion Models. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 9633–9645. PMLR.

Daunhawer, I.; Sutter, T. M.; Chin-Cheong, K.; Palumbo, E.; and Vogt, J. E. 2021. On the limitations of multimodal vaes. *arXiv preprint arXiv:2110.04121*.

Han, T.; Xing, X.; and Wu, Y. N. 2018. Learning Multi-view Generator Network for Shared Representation. In *2018 24th International Conference on Pattern Recognition (ICPR)*, 2062–2068.

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

Ho, J.; Jain, A.; and Abbeel, P. 2020b. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*.

Hsu, W.-N.; and Glass, J. 2018. Disentangling by partitioning: A representation learning framework for multimodal sensory data. *arXiv preprint arXiv:1805.11264*.

Hwang, H.; Kim, G.-H.; Hong, S.; and Kim, K.-E. 2021. Multi-view representation learning via total correlation objective. *Advances in Neural Information Processing Systems*, 34: 12194–12207.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Le, D. H.; Pham, T.; Lee, S.; Clark, C.; Kembhavi, A.; Mandt, S.; Krishna, R.; and Lu, J. 2025. One diffusion to generate them all. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2671–2682.

Lee, M. A.; Zhu, Y.; Srinivasan, K.; Shah, P.; Savarese, S.; Fei-Fei, L.; Garg, A.; and Bohg, J. 2019. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International conference on robotics and automation (ICRA)*, 8943–8950. IEEE.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2023. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*.

Mo, S.; and Raj, B. 2023. Weakly-supervised audio-visual segmentation. *Advances in Neural Information Processing Systems*, 36: 17208–17221.

Owens, A.; Isola, P.; McDermott, J.; Torralba, A.; Adelson, E. H.; and Freeman, W. T. 2016. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2405–2413.

Palumbo, E.; Daunhawer, I.; and Vogt, J. E. 2023. MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises. In *The Eleventh International Conference on Learning Representations*. OpenReview.

Palumbo, E.; Manduchi, L.; Laguna, S.; Chopard, D.; and Vogt, J. E. 2024. Deep Generative Clustering with Multimodal Diffusion Variational Autoencoders. In *International Conference on Learning Representations*.

Pandey, K.; Mukherjee, A.; Rai, P.; and Kumar, A. 2022. DiffuseVAE: Efficient, Controllable and High-Fidelity Generation from Low-Dimensional Latents. *Transactions on Machine Learning Research*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Shi, Y.; Paige, B.; Torr, P.; et al. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in neural information processing systems*, 32.

Sutter, T.; Daunhawer, I.; and Vogt, J. 2020. Multimodal generative learning utilizing jensen-shannon-divergence. *Advances in neural information processing systems*, 33: 6100–6110.

Sutter, T. M.; Daunhawer, I.; and Vogt, J. E. 2021. Generalized multimodal ELBO. *arXiv preprint arXiv:2105.02470*.

Suzuki, M.; Nakayama, K.; and Matsuo, Y. 2016. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*.

Thomas M. Sutter, J. E. V., Imant Daunhawer. 2021. Generalized Multimodal ELBO. In *9th International Conference on Learning Representations, ICLR*.

Vahdat, A.; Kreis, K.; and Kautz, J. 2021. Score-based Generative Modeling in Latent Space. In *Neural Information Processing Systems (NeurIPS)*.

Vedantam, R.; Fischer, I.; Huang, J.; and Murphy, K. 2018. Generative Models of Visually Grounded Imagination. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

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

Xu, X.; Wang, Z.; Zhang, E.; Wang, K.; and Shi, H. 2023. Versatile Diffusion: Text, Images and Variations All in One Diffusion Model . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4578–4587.

Yuan, S.; Cui, J.; Li, H.; and Han, T. 2024. Learning Multimodal Latent Generative Models with Energy-Based Prior. In *European Conference on Computer Vision (ECCV)*.

Zhang, C.; Zhang, C.; Zhang, M.; and Kweon, I. S. 2023. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*.