Multimodal ELBO with Diffusion Decoders

Daniel Wesego
University of Illinois Chicago
Chicago, IL 60607
dweseg2@uic.edu

Pedram Rooshenas
University of Illinois Chicago
Chicago, IL 60607
pedram@uic.edu

Abstract

Multimodal variational autoencoders have demonstrated their ability to learn the relationships between different modalities by mapping them into a latent representation. Their design and capacity to perform any-to-any conditional and unconditional generation make them appealing. However, different variants of multimodal VAEs often suffer from generating low-quality output, particularly when complex modalities such as images are involved. In addition to that, they frequently exhibit low coherence among the generated modalities when sampling from the joint distribution. To address these limitations, we propose a new variant of the multimodal VAE ELBO that incorporates a better decoder using a diffusion generative model. The diffusion decoder enables the model to learn complex modalities and generate high-quality outputs. The multimodal model can also seamlessly integrate with a standard feed-forward decoder for different types of modality, facilitating end-to-end training and inference. Furthermore, we introduce an auxiliary score-based model to enhance the unconditional generation capabilities of our proposed approach. This approach addresses the limitations imposed by conventional multimodal VAEs and opens up new possibilities to improve multimodal generation tasks. Our model provides state-of-the-art results compared to other multimodal VAEs in different datasets with higher coherence and superior quality in the generated modalities.

1. Introduction

Deep learning has revolutionized many fields through its ability to extract and learn patterns that are otherwise difficult through other means [16, 29]. One area that has seen immense progress is generative models, where neural networks learn to generate samples that are similar to the training distribution. Variational autoencoders (VAEs) have emerged as one powerful framework for this task [14, 26]. VAEs were initially built to process a single modality, such as an image. However, in the real world, data often comes from multiple modalities that are interconnected and expressed in different formats [2]. Multimodal learning aims to build models that can process and relate information from various modalities, such as vision, language, and other modalities together [21]. Incorporating this multimodal information can have a huge impact on the performance of the model in tasks that require understanding from multiple perspectives [33].

Multimodal variational autoencoders extend the VAE framework to learn joint representations across modalities [21, 36, 43]. By capturing the correlations between modalities, these models can learn robust and informative representations. These learned multimodal representations can be used in different ways, including transfer learning in downstream tasks [11, 12], and they also provide a framework to perform unconditional and conditional generation across the different modalities [30, 42].

Despite their promising potential, multimodal VAEs continue to face significant challenges in effectively integrating diverse modalities, generating high-quality samples, and scaling to high-dimensional data [5, 42]. Various prior works have attempted to address these issues by proposing alternative ELBO objectives, yet the core problems still persist in multimodal VAEs [34, 35]. Due to these ongoing limitations, most multimodal studies have been constrained to low-dimensional datasets such as MNIST or PolyMNIST [15, 34]. Additionally, Daunhawer et al. [5] highlights the generative discrepancies in mixture-based models and raises concerns about the practicality of these models in real-world applications.

On the other hand, diffusion models have emerged as a powerful class of generative models, gaining substantial attention in recent years for their ability to produce high-quality samples [6, 9]. These models operate by gradually transforming a data distribution into a noise distribution, with the goal of learning to reverse this process. The forward process is a Markovian sequence that incrementally adds noise to the data distribution over several timesteps, while the reverse process, typically modeled by a neural network, learns to remove the noise and recon-

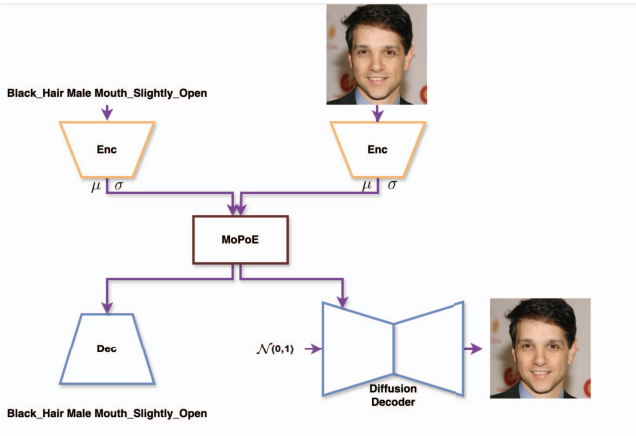


Figure 1. Architecture of our model where modality-specific encoders map the input data from different modalities into latent space, which are then fused using a mixture of PoE and passed to the respective decoders.

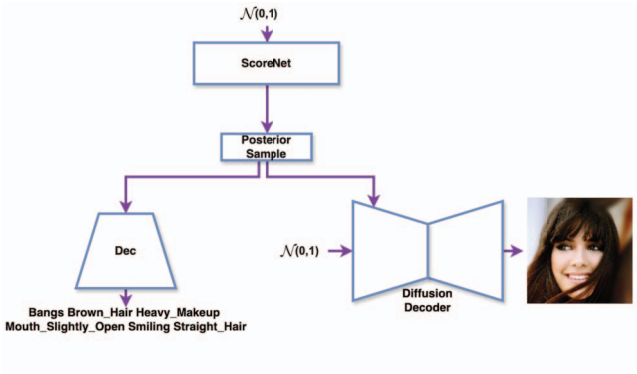


Figure 2. Unconditional sampling technique that leverages a score-based model to transform a Gaussian noise distribution into the PoE posterior distribution of all modalities.

struct the data [9]. A notable advantage of diffusion models is their capacity to generate exceptionally high-quality samples, achieving state-of-the-art performance in various generative tasks, including image synthesis and inpainting [20, 28]. Preechakul et al. [25] introduced diffusion autoencoder, where the encoder resembles that of a VAE, but the decoder is modeled using a diffusion process. The key difference between a diffusion decoder and a traditional feedforward decoder used in VAEs is that the diffusion decoder conditions the generation process on the latent representation rather than directly feeding it as input. This approach enables the diffusion autoencoder to learn meaningful representations as well as generate high-quality samples.

In this study, we introduce a novel multimodal generative model that combines the strengths of diffusion autoencoders and multimodal VAEs to learn a unified joint representation across multiple modalities. Traditional multimodal VAEs often face limitations in generating high-quality outputs from their joint representations, which may not capture all the details of all modalities [11, 22]. These challenges are further exacerbated by the low-performance decoders typically used in standard VAEs, resulting in suboptimal model performance. To address these issues, we propose the use of a flexible diffusion decoder, which allows the compressed joint representation to effectively generate high-quality outputs for complex modalities while leveraging feedforward decoders for simpler modalities.

To achieve this, we propose a multimodal ELBO that is integrated with diffusion decoders for modalities where feed-forward decoders have shown limitations, such as images, ensuring high-quality sample generation and improved representation capacity. Conversely, for modalities where standard decoders have proven effective, such as text or sparse outputs, we retain standard VAE decoders. This hybrid approach allows our model to harness the strengths of each generative model class while effectively addressing their respective limitations.

Our model achieves better coherence in conditional generation among the generated modalities while generating high-quality samples. We use a mixture of product of experts (MoPoE) to fuse the representations from different modalities [35]. This approach enables our model to capture the correlations and dependencies between modalities. For unconditional generation tasks, where the goal is to generate samples from all modalities simultaneously without conditioning on any specific modality, we propose the use of an auxiliary model. This auxiliary model transforms a Gaussian noise distribution into the approximate posterior product of experts distribution. By leveraging this technique, our model gains increased flexibility and can perform well in both conditional and unconditional generation settings.

Figure 1 illustrates the general design and architecture of our proposed model. The modality-specific encoders map the input data from different modalities into latent representations, which are then fused using the MoPoE mechanism. The fused representation is then passed to the respective decoders, which can be either diffusion decoder or a standard one, depending on the modality. Figure 2 shows how to perform unconditional sampling where the score-based model is employed to provide the necessary input distributions to the decoders.

Our proposed model demonstrates superior performance compared to previous variants of multimodal variational autoencoder models. Our key contributions can be summarized as follows:

- We introduce a novel generative multimodal autoencoder architecture that seamlessly integrates diffusion and standard decoders. This hybrid design allows our model to generate high-quality samples across modalities without compromising the coherence and consistency among the generated modalities. In doing so, our architecture strikes an optimal balance between sample quality and cross-modal coherence.
- To enhance the model's performance in unconditional generation settings, we propose to use a sampling technique that employs an auxiliary score-based model that transforms a Gaussian noise distribution into the approximate distribution of the PoE posterior to improve the model's ability to generate coherent and diverse samples across all modalities in an unconditional manner for multimodal VAEs.
- Our approach can be scaled to high-dimensional data modalities with good generative quality compared to previous approaches, which opens a new research direction toward multimodal VAEs.

Through these contributions, our model addresses the critical limitations of previous multimodal VAE approaches while retaining their desirable properties and extending their capabilities to handle high-dimensional data modalities effectively. Our approach paves the way for more powerful and flexible multimodal generative models capable of generating high-quality samples with enhanced coherence.

2. Related Works

Multimodal learning has emerged as a promising research direction in the deep learning field that has started from early works such as Ngiam et al. [21], Tsai et al. [38]. Ngiam et al. [21] demonstrated the potential of multimodal deep learning models in leveraging information from multiple modalities to improve performance on multiple tasks by learning combined features. As variational autoencoders (VAEs) gained traction for simultaneously learning representations and generation models within individual modalities, researchers soon explored extending the VAE framework to handle multiple modalities simultaneously, giving rise to multiple variants of multimodal VAEs.

Suzuki et al. [36] first presented a model that is trained jointly and can accept multiple modalities. They propose training additional encoders for each modality and minimizing the KL loss between the joint ones to handle missing modalities. Wu and Goodman [43] proposed a product of experts approach to encode the modalities jointly. They posterior is constructed as $q_\phi(\mathbf{z}|\mathbf{x}) = p(\mathbf{z})\prod_i^M q_\phi(\mathbf{z}|\mathbf{x_i})$. This approach allowed different formulations of the ELBO in future works because the product of experts can be calculated in a closed form, and it allows the generation of missing modalities by simply plugging 1 in the product for the missing modalities. The model proposed by Wu and Goodman [43], which is called MVAE, generates images comparative to a unimodal VAE but has low coherence amongst the generated modalities.

Shi et al. [30] proposed a mixture of experts posterior formulation that samples one modality at a time from the set containing all the modalities. Their model shows better coherence but, in general, has degraded quality, especially visible in image modalities. Sutter et al. [35] further expanded the subsets in the mixture by proposing a mixture of product of experts posterior. Daunhawer et al. [5] showed that models that subsample will have a loss in generative quality because of reconstructing all available modalities from a few of them. Other approaches by Lee and Pavlovic [18], Palumbo et al. [22], Sutter et al. [34] explored splitting the latent space into modality-specific and shared representations to improve the generative quality of multimodal VAEs.

In the work by Hwang et al. [11], the authors propose an evidence lower bound (ELBO) objective derived from an information-theoretic perspective that uses a product of experts approach. Their main objective is to learn a representation that minimizes the conditional total correlation between the joint conditional and the factorized conditionals. Their final objective is a convex combination of two terms: one based on the variational information bottleneck (VIB) and the other based on the conditional VIB.

Pandey et al. [24] proposed a two-stage training in which a unimodal VAE is trained in the first stage, and a diffusion model conditioned on the output of the VAE is trained in the second stage to refine the lower-quality output of a VAE. Similar approaches have also been proposed in multimodal works to refine the outputs of the image modalities of multimodal VAEs by Palumbo et al. [23], Wesego and Rooshenas [42], but the diffusion model is not directly integrated to the ELBO like our proposed work.

Alternatively, Bounoua et al. [3], Wesego and Rooshenas [42] proposed to use unimodal VAEs to represent each modality and use an additional neural network to fuse the modalities instead of a product/mixture of experts using a two-step training process. They both select a score-based network to join the modalities in the latent space and sample from the joint distribution. Rombach et al. [27] proposed text-to-image generation that employs a high-quality autoencoder to compress images into latent space, allowing for diffusion processes to occur in this reduced dimensionality. They integrated CLIP models using cross-attention to effectively represent and condition text inputs. Building upon this, Tang et al. [37], Xu et al. [44] proposed multi-modal latent diffusion models that can generate any modality from any given condition. They expand Rombach et al. [27] work to use multiple diffusion flows to achieve this. It's worth noting that these advanced models require training on extremely large-scale datasets, which requires

extensive computational resources.

Preechakul et al. [25] introduced diffusion autoencoder as a means for diffusion models to generate images and learn representations. They transformed the diffusion model into an autoencoder form where an encoder learns a latent space that will be used in the reconstruction process. Hudson et al. [10] modified the diffusion autoencoder to reconstruct multiple views in the diffusion decoder. Wang et al. [41] proposed a diffusion autoencoder architecture where the latent variable $\mathbf{z}$ is encouraged to learn disentangled meaningful representation by regularizing the objective with mutual information among $\mathbf{x}$ and $\mathbf{z}$.

In this work, we focus on improving multimodal VAEs by proposing an objective that flexibly combines both diffusion autoencoder and standard VAEs. Overall, Daunhawer et al. [5] illustrated that the current multimodal VAEs are very limited due to their lack of generation capability as the modalities become more complex. This is caused by how the multimodal ELBO objective is constructed and by the limitation in VAEs to generate high-quality images [24]. Our approach enhances the generative quality of multimodal VAEs while still preserving the fundamental framework and strengths of the multimodal VAE architecture in conditional and unconditional settings.

3. Methodology

We first present our method starting with diffusion autoencoders. Subsequently, we introduce the multimodal VAE framework that incorporates diffusion autoencoders. The section concludes by explaining how to perform both conditional and unconditional sampling techniques.

3.1. Diffusion Autoencoder

A diffusion model can be formulated as a latent variable model where the marginal distribution is expressed as $p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_{0:T})\mathrm{d}\mathbf{x}_{1:T}$ [9]. The combined probability distribution $p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is the reverse joint process where $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. One can sample any intermediate time step t in the forward process starting from the data point $\mathbf{x}_0$ using the Gaussian distribution $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$ to add noise to the data. A neural network is trained to reverse this process where $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$. Ho et al. [9] observed that using a simplified objective and predicting the noise is better and easier to implement. The objective is shown in eq. 1.

$$L(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{x}_t, t)\|^2\right] \tag{1}$$

where $\mathbf{x_t} = \sqrt{\bar{\alpha}_t}\mathbf{x_0} + \sqrt{1-\bar{\alpha}_t}\epsilon_t$ and $\bar{\alpha}_t$ are schedule hyperparameters.

A diffusion autoencoder combines the architectural setup of an autoencoder with a diffusion decoder [25]. An encoder $q(\mathbf{z}|\mathbf{x})$ projects the data into a latent representation $\mathbf{z}$. This latent representation is fed to the diffusion decoder which generates samples conditioned on $\mathbf{z}$. By training the model this way, the diffusion model can be used as a powerful representation where the encoder learns a meaningful latent representation of the data similar to an autoencoder [10]. The objective for a diffusion autoencoder is very similar to the diffusion objective, except we have additional conditioning from the encoder. It is trained by minimizing $\mathbb{E}_{\mathbf{x_0},\boldsymbol{\epsilon_t}}[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{z}, t)\|^2]$ where $\mathbf{z}$ comes from the encoder network $q(\mathbf{z}|\mathbf{x})$.

In addition, diffusion models can be trained to maximize the log-likelihood of training data by using a specific weight on the score-matching loss term [32]. By using this objective, which is upper bounded by the log-likelihood, diffusion models can achieve log-likelihoods that are comparable to the best models. Specifically, with a stochastic differential equation (SDE) of the form $d\mathbf{x} = f(\mathbf{x}, t)dt + g(t)d\mathbf{w}$ that diffuses data to noise with a drift coefficient f and a diffusion coefficient g where $\mathbf{w}$ is a standard Wiener process, we can use a weighting $\lambda(t) = g(t)^2$ for likelihood weighting. The denoising score-matching loss to train diffusion models for likelihood training is shown in eq. 2, where additional conditioning term $\mathbf{z}$ is added from the encoder to form the autoencoder setup and a Monte Carlo estimate is used for the expectation with t sampled from uniform distribution, $\mathbf{x}$ from the training data, and $s_\theta()$ denotes the score function [32].

$$L(\theta) = \mathbb{E}_{p(\mathbf{x})p(t)p(\mathbf{z}|\mathbf{x})p_t(\mathbf{x}_t|\mathbf{x},\mathbf{z})}\left[\frac{1}{2}\lambda(t)\right. \tag{2}$$
$$\left.\left\|\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t \mid \mathbf{x}, \mathbf{z}) - s_\theta(\mathbf{x}_t, \mathbf{z}, t)\right\|_2^2\right]$$

3.2. Multimodal VAE with Diffusion Decoder

Assuming that we have M modalities, the observed data is represented as $\mathbf{x} = \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)}$. Mixture-based multimodal models encode the posterior distribution using a mixture of experts of Gaussian distributions $q_\phi^S(\mathbf{z}|\mathbf{x}) = \sum_{A\in\mathcal{S}} \omega_A q_\phi(\mathbf{z}|\mathbf{x}^A)$ where $0 \leq \omega_A \leq 1$ and $\mathcal{S}$ is a selected set of modalities formed from the powerset of the combination of modalities. Using a mixture of experts allows easy inference when some of the modalities are not observed, and training can be performed by sampling one component from the mixture at a time. Daunhawer et al. [5] highlighted that mixture-based multimodal models face an inherent limitation compared to unimodal VAEs due to the sub-sampling of modalities in the encoder. In these models, a restricted subset of modalities is used to encode information for all modalities, requiring the decoders to reconstruct all modalities from this limited input. This challenge is particularly evident in multimodal VAEs, where the decoder

often struggles with the insufficient information provided by the encoder. However, incorporating a more powerful decoder can help address this issue effectively. For any subset $\mathcal{S}$ of modalities, a multimodal evidence lower bound (ELBO) can be derived for the log-likelihood of the multimodal data [5]:

$$\log p_\theta(\mathbf{x}) \geq \sum_{A \in \mathcal{S}} \omega_A \Bigg\{ \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^A)} \left[\log p_\theta(\mathbf{x} \mid \mathbf{z}) \right] - D_{\mathrm{KL}}\big(q_\phi(\mathbf{z} \mid \mathbf{x}^A) \,\|\, p(\mathbf{z})\big) \Bigg\}. \tag{3}$$

The full derivation is given in the supplementary material for reference. Daunhawer et al. [5] demonstrates that this bound generalizes various multimodal VAE objectives, such as MoPoE [35], MMVAE [30], and MVAE [43].

Assuming that the modalities are conditionally independent given the latent representation $\mathbf{z}$ (i.e. $\mathbf{x}^i \perp \mathbf{x}^{\setminus i} \mid \mathbf{z}$), the ELBO can be written the following:

$$\log p(\mathbf{x}) \geq \sum_{A \in \mathcal{S}} \omega_A \Bigg\{ \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_A)} \left[\sum_{i=1}^{M} \log p_\theta(\mathbf{x}^i \mid \mathbf{z}) \right] - D_{\mathrm{KL}}\big(q_\phi(\mathbf{z} \mid \mathbf{x}^A) \,\|\, p(\mathbf{z})\big) \Bigg\}. \tag{4}$$

These conditional independencies enable us to support modality-specific decoders – specifically, feed-forward and diffusion-based decoders. To distinguish between them, we denote the number of modalities with feed-forward decoders as M_F and those with diffusion-based decoders as M_D.

The objective defined in eq. 5 is a valid lower bound on the marginal likelihood of the data under the proposed model.

$$\begin{aligned} \log & p(\mathbf{x}) \geq \\ & \sum_{A \in \mathcal{S}} \omega_A \Bigg\{ \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_A)} \Bigg[\sum_{i=1}^{M_F} \log p_\theta(\mathbf{x}_i|\mathbf{z}) + \\ & \sum_{j=1}^{M_D} \mathbb{E}_{\mathbf{x}_{jt}} \frac{1}{2} \lambda(t) \Big\| \nabla_{\mathbf{x}_{jt}} \log p_t(\mathbf{x}_{jt} \mid \mathbf{x}_j, \mathbf{z}) - \\ & s_\theta(\mathbf{x}_{jt}, \mathbf{z}, t) \Big\|_2^2 \Bigg] - D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}|\mathbf{x}_A) \big\| p(\mathbf{z})\big) \Bigg\}. \end{aligned} \tag{5}$$

See Appendix for the proof.

The objective in the eq 5 can be decomposed into the normal multimodal VAE objective shown in eq. 4 if we use normal feed-forward decoders for each modality. If we use diffusion decoder models, we train the reconstruction terms $p_\theta(\mathbf{x}|\mathbf{z})$ using the denoising score-matching loss shown in 2 for likelihood training which is equivalent to maximizing the likelihood of the data under the diffusion process conditioned on $\mathbf{z}$.

3.3. Conditional Sampling

To perform conditional sampling (i.e. generate missing modalities given some modalities) utilizing our model, we can leverage the factorized nature of the posterior distribution. We first sample from the joint posterior distribution conditioned on the available modalities and then use the sampled latent variable to generate samples using the respective generative model for each modality. The posterior will be the product of the observed modalities $q_\phi(\mathbf{z}|\mathbf{x}_o) = \prod_{i \in o} q_\phi(\mathbf{z}_i|\mathbf{x}_i)$ where the subscript o denotes observed modalities. Then, we can generate an unobserved modality u from the posterior $q_\phi(\mathbf{z}|\mathbf{x}_o)$ by feeding it to the appropriate decoder $p(\mathbf{x}|\mathbf{z})$. By following this, we can generate any modality from any other given modality.

3.4. Unconditional Sampling

Unconditional generation occurs when no specific modality is provided as input. In this scenario, the goal is to produce samples that exhibit coherence across all modalities without relying on any initial conditions. Previous multimodal VAEs perform unconditional generation by first sampling from a $\mathcal{N}(0, \mathbf{I})$ and then feeding the sample to each decoder. Doing that may produce suboptimal results as there is a gap between the posterior and the prior [1]. To get better samples, we train an auxiliary score-based diffusion model to sample from the posterior. Figure 2 demonstrates how the score model facilitates unconditional sampling. This auxiliary score model is trained using the diffusion objective but on the latent space $\mathbf{z}$ to generate samples from the posterior. The objective is $\mathbb{E}_{t,\mathbf{z}_0,\boldsymbol{\epsilon}}[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{z}_t, t)\|^2$ where $\mathbf{z_0}$ is sampled from $q_\phi(\mathbf{z}|\mathbf{x}) = \prod_i q_\phi(\mathbf{z}_i|\mathbf{x}_i)$ where $\mathbf{x}_i$ consists of each modality in $\mathbf{x}$. The main objective of training this auxiliary model is to get high-quality samples from the posterior instead of using $\mathcal{N}(0, \mathbf{I})$ as the starting sample for unconditional generation. To generate samples unconditionally, we first sample $\mathbf{z}$ from the auxiliary score model and use the respective generative model $p_\theta(\mathbf{x}_i|\mathbf{z})$ of each modality to get the final samples.

4. Experiments

We perform experiments on two datasets. The first one is from Wah et al. [40], which contains images of birds with text captions describing the birds. We use actual images of the birds instead of ResNET features, unlike previous multimodal VAEs that struggle with image data [30]. The second dataset is a high-dimensional CelebAMask-HQ [17] dataset that consists of images, masks, and attributes of celebrities expressed in the three modalities. We compare different multimodal VAE baselines consisting of

a small sized bird that has multiple tones of grey all over

Figure 3. Conditional samples given text using Diff-MVAE (left), Diff-MVAE*(right)

MVTCAE [11], MoPoE [35], and MMVAE+ [22]. MMVAE+ is not included in the CUB dataset experiment as we use pre-trained VAE for the text captions, which will require changing the structure of the text VAE because of the modality-specific and shared representations in MMVAE+. Nonetheless, we added the results from their paper for the text-to-image results, which can be referenced from their work [22]. We also add additional baselines from Bounoua et al. [3], Wesego and Rooshenas [42] that use a latent score-based diffusion model to learn the joint latent space discussed in the next sections. We compare these baselines with our proposed model variants Diff-MVAE and Diff-MVAE*. Diff-MVAE* is trained using the same objective as eq. 5, but with $\lambda(t) = 1$. This choice introduces a mismatch between the ELBO and the score-matching objective, resulting in a violation of the ELBO in Eq. 5. However, empirical results indicate that setting $\lambda(t) = 1$ leads to higher-quality generated samples, a finding that has also been reported in previous works [9, 39].

4.1. Caltech Birds (CUB)

The CUB dataset consists of two modalities: image and text describing the image (caption). In order to get meaningful text outputs, we initialized the text VAE from the pre-trained weight of Li et al. [19] that uses a BERT encoder and GPT-2 decoder for all models. The images are resized to 64x64. We used a latent size of 768 for both the image and text. The encoder architecture for the image modality is similar to that of Daniel and Tamar [4] and the diffusion decoder uses a UNET architecture. The models were trained for 500 epochs using the Adam optimizer [13]. The baseline models are trained using different β values listed in the appendix, and the best model was selected by using the average of the conditional and unconditional FID. The unconditional auxiliary score model also uses a 1D UNET architecture and accepts input similar to the size of the latent dimension. DDIM sampling was used for 50 steps to generate samples for the image modality for Diff-MVAE* and for unconditional sampling [31]. We used Euler–Maruyama sampling for Diff-MVAE. More details about the training and inference are in the Appendix section.

We use FID, which is a widely adopted metric for evaluating the quality of images, as the evaluation metric to measure conditionally and unconditionally generated samples [8]. In addition to that, we use Clip-Score to measure the similarity of the image-text pairs [3, 7]. It's clear that a good multimodal VAE model will have coherent unconditional samples where the image and text represent the same concept expressed in its respective modality. In addition, conditionally generated images should follow the description in the text, and conditionally generated text should follow the given image.

Table 1. CUB FID Result

	Txt-to-Img	Unc
Diff-MVAE	57.4$_{(\pm 0.2)}$	70.4$_{(\pm 0.05)}$
Diff-MVAE*	**35.5**$_{(\pm 0.5)}$	**37.5**$_{(\pm 0.1)}$
MoPoE	290.6$_{(\pm 0.5)}$	199.6$_{(\pm 0.75)}$
MVTCAE	176.3$_{(\pm 0.05)}$	168.9$_{(\pm 0.2)}$
MLD	62.6	63.4
MMVAE+	164.9	-

this bird has wings that are brown and has a white belly

a small bird with a long bill and bright orange breast

Figure 4. Unconditional samples using Diff-MVAE (left), Diff-MVAE*(right)

We present the results in Table 1 and 2. In addition to the implemented baselines, we add results from previous works of MLD by Bounoua et al. [3] and MMVAE+ on text-to-image FID. The FID scores evaluation demonstrates that our model, Diff-MVAE, outperforms the baselines, generating high-quality images that exhibit greater coherence and alignment with the text modality. In contrast, the baselines struggle to generate meaningful images that are consistent with the textual information, supporting the findings of Daunhawer et al. [5] regarding the challenges faced by existing multimodal VAEs in difficult tasks. Table 2 also shows that Diff-MVAE generates more coherent conditional outputs that follow the given modality. We also added qualitative results from our model, which is the only multimodal VAE model generating good images in this dataset, are shown in figure 3 and 4 for conditional and unconditional generation respectively.

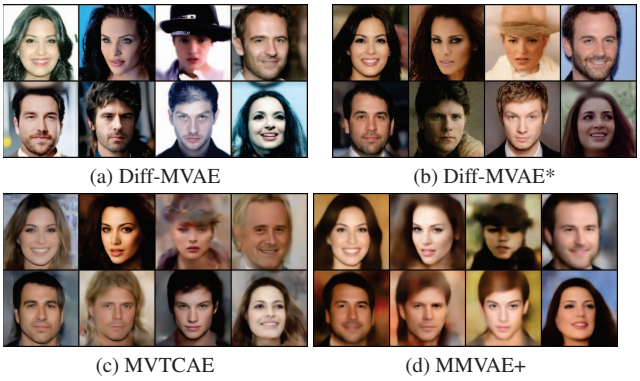
4.2. CelebAMask-HQ

The CelebAMask-HQ dataset provides multimodal representations of visual characteristics related to human faces. In this dataset, the images, masks, and attributes can be viewed as distinct yet complementary modalities that capture various aspects of an individual's appearance. We resize the images and masks to a resolution of 128 by 128 pixels. We combine multiple masks of parts of the face into a single 0/1 mask similar to Wesego and Rooshenas [42]. Specifically, all the provided masks in the CelebAMask-HQ dataset, except for the skin mask, are drawn on top of each other, resulting in a single composite mask. This composite mask encodes the presence or absence of various facial features. For the attribute modality, we follow the preprocessing approach described by Wesego and Rooshenas [42], Wu and Goodman [43]. Out of the 40 existing attributes in the CelebAMask-HQ dataset, we selectively choose 18 attributes that are most relevant and informative for representing the visual characteristics of a person's face. We used a latent size of 256. Similar to the previous dataset, the encoder and decoder architecture for the image modality are inspired by Daniel and Tamar [4] except for the diffusion decoder, which uses a UNET architecture. The unconditional auxiliary score-based model also works on the 256 dimension latent size resized to 16x16 utilizing a UNET architecture. We used a similar sampling strategy of the previous dataset for Diff-MVAE variants. More details about the experimental setup are located in the appendix section. The baselines here are MoPoE, MVTCAE, MMVAE+ plus SBM-VAE-C [42] and MLD [3], which use unimodal variational autoencoders (VAEs) or autoencoders (AEs) for each modality and fuse them using a score-based model for learning the joint distribution. We also add MoPoE*, which is trained with the same hyperparameters as Diff-MVAE but without diffusion decoders.

To assess the performance of the models in generating high-quality samples across different modalities, we employ two quantitative evaluation metrics. For the image modality, we utilize the Fréchet Inception Distance (FID) score. On the other hand, for the mask and attribute modalities, we employ the sample-average F1 score as the evaluation metric. The F1 score provides a comprehensive measure of the model's ability to accurately generate binary predictions in multi-label classifications.

Table 2. CUB Clip Score Result

	Img-to-Txt	Txt-to-Img	Unc
Diff-MVAE	27.7$_{(\pm 0.01)}$	**28.8**$_{(\pm 0.02)}$	**28.23**$_{(\pm 0.01)}$
Diff-MVAE*	**28.1**$_{(\pm 0.001)}$	28.52$_{(\pm 0.02)}$	28.12$_{(\pm 0.005)}$
MoPoE	26.8$_{(\pm 0.03)}$	22.0$_{(\pm 0.03)}$	24.8$_{(\pm 0.05)}$
MVTCAE	27.1$_{(\pm 0.03)}$	26.6$_{(\pm 0.004)}$	28.08$_{(\pm 0.02)}$



(a) Diff-MVAE (b) Diff-MVAE* (c) MVTCAE (d) MMVAE+

Figure 5. Conditional generated images for different models given the same masks and attribute for all the models

We present our results in table 3 which shows the performance of the models in both conditional and unconditional settings. Diff-MVAE generates high-quality images, achieving the best FID scores among all models in image generation. Not only that but also the performance of the model in the other modalities where Diff-MVAE outperforms the baselines in attribute prediction, even if the attribute modality architectures are the same for all models. Our models also have very close results to supervised prediction models that are trained to predict the attribute or mask directly from the image modality [42]. Unconditional results also show that the baselines do not generate good-quality images unconditionally. The effect of the auxiliary score-based model can be seen in this case, where Diff-MVAE can leverage it to generate coherent outputs along all the modalities. In general, the results indicate that our proposed Diff-MVAE model exhibits superior performance in both conditional and unconditional generation tasks, consistently producing higher-quality and more coherent outputs compared to the baselines. We show qualitative results in Figure 5. The figure shows conditionally generated images for Diff-MVAE and the baselines where Diff-MVAE generates higher-quality images that are more coherent with the given conditions.

4.3. Ablation: Effect of the Auxiliary Score Model

This section covers the importance of using the auxiliary score-based model. By default, multimodal VAEs sample from $\mathcal{N}(0, \mathbf{I})$ to perform unconditional generation. The problem with this is there could be considerable differences in distribution between the prior and the posterior, which will lead to low-quality samples. Previous unimodal VAE works improved sample quality by training an energy-based model on the latent space that will serve as a prior instead of the standard normal distribution and provide samples closer

Table 3. CelebAMask-HQ Result

	Attribute		Image				Mask	
GIVEN	Both	Img	Both	Mask	Attr	Unc	Both	Img
	F1	F1	FID	FID	FID	FID	F1	F1
Diff-MVAE	**0.76**(±0.001)	0.75(±0.001)	42.3(±0.2)	41.9(±0.06)	45.1(±0.45)	43.5(±0.2)	**0.90**(±0.001)	0.90(±0.001)
Diff-MVAE*	**0.76**(±0.001)	**0.76**(±0.001)	**28.3**(±0.02)	**28.5**(±0.1)	**32.3**(±0.35)	**28.4**(±0.3)	**0.90**(±0.001)	0.90(±0.001)
MoPoE*	0.74(±0.002)	0.75(±0.002)	104.8(±0.31)	105.8(±0.18)	182.3(±0.29)	139.8(±0.65)	**0.90**(±0.001)	0.90(±0.001)
MoPoE	0.68(±0.002)	0.71(±0.004)	114.9(±0.32)	101.1(±0.16)	186.7(±0.28)	164.8(±0.62)	0.85(±0.002)	**0.92**(±0.001)
MVTCAE	0.71(±0.001)	0.69(±0.004)	94(±0.45)	84.2(±0.32)	87.2(±0.08	162.2(±1.08)	0.89(±0.001)	0.89(±0.003)
MMVAE+	0.64(±0.003)	0.61(±0.002)	133(±14.28)	97.3(±0.04)	153(±0.49)	103.7(±0.61)	0.82(±0.03)	0.89(±0.002)
SBM-VAE-C	0.69(±0.005)	0.66(±0.001)	82.4(±0.1)	81.7(±0.29)	76.3(±0.7)	79.1(±±0.3)	0.84(±0.02)	0.84(±0.001)
MLD	0.71(±0.005)	0.67(±0.006)	81.7(±0.25)	82.4(±0.15)	80.29(±0.6)	82.8(±±0.08)	0.86(±0.001)	0.86(±0.001)
Supervised		0.79(±0.001)						0.94(±0.001)

to the posterior, thereby improving generated outputs [1]. Preechakul et al. [25] also used a score-based model in the unimodal diffusion autoencoder to generate samples unconditionally. In light of these, we train an auxiliary score-based model to generate prior samples for unconditional multi-modal generation. The auxiliary score-based model will transform the standard normal sample to a sample from the posterior $q_\phi(\mathbf{z}|\mathbf{x})$. Figure 6 compares outputs with and without the auxiliary score model, showing inferior results on the left and improved ones on the right.

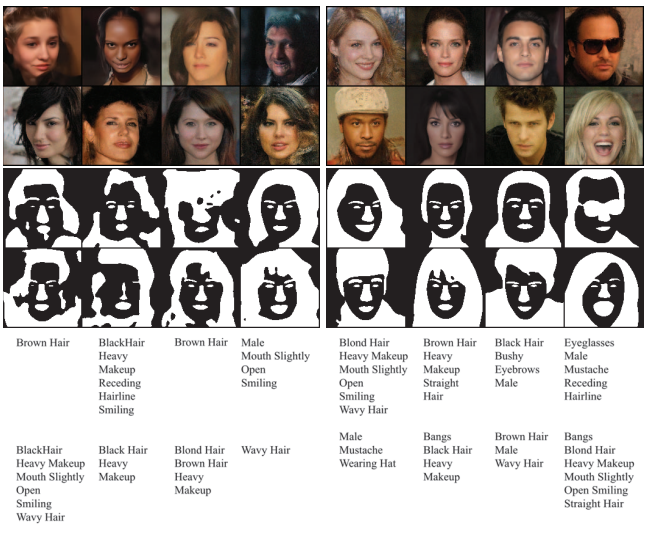


Figure 6. Unconditional generation using Diff-MVAE* when using no auxiliary score model (left) and when using auxiliary score model (right)

5. Conclusion and Discussion

In this paper, we introduce Diff-MVAE, a new multimodal VAE that enhances the capabilities of traditional multimodal VAEs by incorporating diffusion decoders, leading to superior performance. Additionally, we implement an auxiliary model to bridge the gap between the prior and posterior distributions, achieving high-quality and coherent results. A notable limitation of Diff-MVAE is its higher computational demand during sampling due to the iterative diffusion process. Nonetheless, our model demonstrates significant performance improvements across multiple challenging datasets, advancing the field of multimodal VAEs and positioning them as a preferred choice for multimodal applications. Future work will focus on optimizing the multimodal model to reduce computational overhead. Our contributions provide a robust framework for future research and development in multimodal VAE generative modeling.

References

[1] Jyoti Aneja, Alex Schwing, Jan Kautz, and Arash Vahdat. A contrastive learning approach for training variational autoencoder priors. In *Advances in Neural Information Processing Systems*, pages 480–493. Curran Associates, Inc., 2021. 5, 8

[2] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41: 423–443, 2019. 1

[3] Mustapha Bounoua, Giulio Franzese, and Pietro Michiardi. Multi-modal latent diffusion. *Entropy*, 26(4), 2024. 3, 6, 7

[4] Tal Daniel and Aviv Tamar. Soft-introvae: Analyzing and improving the introspective variational autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition (CVPR), pages 4391–4400, 2021. 6, 7, 5

[5] Imant Daunhawer, Thomas M. Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E Vogt. On the limitations of multimodal VAEs. In *International Conference on Learning Representations*, 2022. 1, 3, 4, 5, 6

[6] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021. 1

[7] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 6

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 6

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 1, 2, 4, 6

[10] Drew A. Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K. Lampinen, Andrew Jaegle, James L. McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning, 2023. 4

[11] HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Multi-view representation learning via total correlation objective. In *Advances in Neural Information Processing Systems*, pages 12194–12207. Curran Associates, Inc., 2021. 1, 2, 3, 6

[12] Peiguang Jing, Xianyi Liu, Lijuan Zhang, Yun Li, Yu Liu, and Yuting Su. Multimodal attentive representation learning for micro-video multi-label classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(6), 2024. 1

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6, 5

[14] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 1

[15] Y. LeCun and C. Cortes. Mnist handwritten digit database, 2010. 1

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015. 1

[17] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[18] Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal vae for learning of latent representations. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1692–1700, 2021. 3

[19] Chunyuan Li, Xiang Gao, Yuan Li, Xiujun Li, Baolin Peng, Yizhe Zhang, and Jianfeng Gao. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *EMNLP*, 2020. 6, 5

[20] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. 2022. 2

[21] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, page 689–696, Madison, WI, USA, 2011. Omnipress. 1, 3

[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*, 2023. 2, 3, 6

[23] Emanuele Palumbo, Laura Manduchi, Sonia Laguna, Daphné Chopard, and Julia E Vogt. Deep generative clustering with multimodal diffusion variational autoencoders. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[24] Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *CoRR*, abs/2201.00308, 2022. 3, 4

[25] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4, 8

[26] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1278–1286, Bejing, China, 2014. PMLR. 1

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, pages 36479–36494. Curran Associates, Inc., 2022. 2

[29] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015. 1

[30] Yuge Shi, Siddharth Narayanaswamy, Brooks Paige, and Philip H. S. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*

2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 15692–15703, 2019. 1, 3, 5

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 6

[32] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34: 1415–1428, 2021. 4, 3

[33] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, 15(84):2949–2980, 2014. 1

[34] Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. Multimodal generative learning utilizing jensen-shannon-divergence. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1, 3

[35] Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. Generalized multimodal ELBO. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1, 2, 3, 5, 6

[36] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. 1, 3

[37] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *CoRR*, abs/2305.11846, 2023. 3

[38] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 3

[39] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space, 2021. 6

[40] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5

[41] Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. Infodiffusion: Representation learning using information maximizing diffusion models. *CoRR*, abs/2306.08757, 2023. 4

[42] Daniel Wesego and Pedram Rooshenas. Score-based multimodal autoencoder. *Transactions on Machine Learning Research*, 2024. 1, 3, 6, 7, 4

[43] Mike Wu and Noah D. Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5580–5590, 2018. 1, 3, 5, 7

[44] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model, 2023. 3