MITIGATING THE LIMITATIONS OF MULTIMODAL VAEs WITH COORDINATION-BASED APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

One of the key challenges in multimodal variational autoencoders (VAEs) is inferring a joint representation from arbitrary subsets of modalities. The state-of-the-art approach to achieving this is to sub-sample the modality subsets and learn to generate all modalities from them. However, this sub-sampling in the mixture-based approach has been shown to degrade other important features of multimodal VAEs, such as quality of generation, and furthermore, this degradation is theoretically unavoidable. In this study, we focus on another approach to learning the joint representation by bringing unimodal inferences closer to joint inference from all modalities, which does not have the above limitation. Although there have been models that can be categorized under this approach, they were derived from different backgrounds; therefore, the relation and superiority between them were not clear. To take a unified view, we first categorize them as coordination-based multimodal VAEs and show that these can be derived from the same multimodal evidence lower bound (ELBO) and that the difference in their performance is related to whether they are more tightly lower bounded. Next, we point out that these existing coordination-based models perform poorly on cross-modal generation (or cross-coherence) because they do not learn to reconstruct modalities from unimodal inferences. Therefore, we propose a novel coordination-based model that incorporates these unimodal reconstructions, which avoids the limitations of both mixture and coordination-based models. Experiments with diverse and challenging datasets show that the proposed model mitigates the limitations in multimodal VAEs and performs well in both cross-coherence and generation quality.

1 INTRODUCTION

Deep generative models have recently shown high performance on multimodal data, including images and captions. In particular, multimodal learning with variational autoencoders (VAEs) (Kingma & Welling, 2013; Rezende et al., 2014) has attracted much attention to obtain a joint representation from modalities (Wu & Goodman, 2018; Shi et al., 2019; Sutter et al., 2021). Such representations can be used for predicting common concepts from these modalities or for generating other modalities.

An essential challenge of multimodal VAEs is inferring a joint representation from subsets of modalities. State-of-the-art models attempt to accomplish this by training to generate all modalities from a joint representation inferred from missing (or sub-sampled) modalities. These are called mixture-based multimodal VAEs (Daunhawer et al., 2021a), and examples include MMVAE (Shi et al., 2019) and MoPoE-VAE (Sutter et al., 2021) [1]. However, it has been pointed out that the quality of modality generation is lower than that of unimodal VAEs, and furthermore, cross-generation between modalities, or cross-coherence, can be degraded. This is an inherent limitation on mixture-based VAEs due to modality sub-sampling during training, and theoretical and empirical evidence shows that this limitation cannot be avoided (Daunhawer et al., 2021a).

To alleviate this issue, we focus on another approach in multimodal VAEs that brings the representation inferred from each modality closer to that inferred from all modalities. This approach avoids the

[1] Daunhawer et al. (2021a) also include a special case of MVAE (Wu & Goodman, 2018) in mixture-based models. However, we focus on the mixture-based nature of cross-generation of modalities from subsets as their shortcomings, so we exclude MVAE, which does not perform cross-generation.

limitation of degraded generation quality because its objective does not include generation from sub-sampled modalities. Models that fit this approach have been proposed, but their objectives are derived from different backgrounds. For example, MVTCAE (Hwang et al., 2021), which can be regarded as one of them, is derived its objective from the total correlation; therefore, its relationship to the multimodal VAEs' objective, multimodal evidence lower bound (ELBO), is not clear. Moreover, another model of this approach, MMJSD (Sutter et al., 2020), has been shown experimentally to perform worse than MVTCAE (Hwang et al., 2021), but the theoretical reason for this is unclear.

We first categorize them as *coordination-based multimodal VAEs* [2] and show that these models can be viewed as unified, i.e., all are derived from multimodal ELBO, and that the differences in models correspond to whether they are more tightly lower bounded.

Next, we point out that the existing coordination-based VAEs still have a shortcoming in cross-coherence since no generation from unimodal inferences is included in the training. We then propose a novel objective that introduces unimodal reconstruction terms. This objective can also be regarded as a multimodal ELBO, and since it does not include cross-generation, it also avoids the generation degradation issue of the mixture-based models while mitigating the issue of coordination-based models. Note that our model alleviates the issues by only changing the objective, unlike other methods for improving multimodal VAEs by changing the architecture, such as introducing additional latent variables (Tsai et al., 2018; Hsu & Glass, 2018; Sutter et al., 2020; Palumbo et al., 2022).

We conducted experiments on five diverse multimodal datasets, including challenging ones, e.g., those that cannot be adequately trained with existing models or require additional architecture for adequate training. We confirmed that the proposed method outperforms existing mixture- and coordination-based models in terms of cross-coherence and generation quality.

2 Preliminaries

2.1 Multimodal VAEs

Suppose that we have a set of multimodal examples, where each example is a set of M modalities $X = \{\mathbf{x}_m\}_{m=1}^M$. We assume that these examples are derived from data distribution $p_d(X)$ and that each example X has a corresponding common latent concept $\mathbf{z}$, i.e., a joint representation.

Given a training set $\{X^{(i)}\}_{i=1}^N$, our goals are to infer a joint representation from the subset of modalities $X_S \subseteq X$[3] and to generate another subset $X_{S'} \subseteq X$ via that representation. In other words, we aim to obtain an inference $p(\mathbf{z}|X)$ and a joint distribution $p(X, \mathbf{z}) = \prod_{\mathbf{x}_m \in X} p(\mathbf{x}_m|\mathbf{z})p(\mathbf{z})$.

To achieve this, it is a natural choice to use VAEs (Kingma & Welling, 2013), deep generative models that can learn inference to acquire representations from data in addition to the generation. The generative model $p_\theta(X|\mathbf{z}) = \prod_{m:\mathbf{x}_m \in X} p_\theta(\mathbf{x}_m|\mathbf{z})$ is parameterized by deep neural networks, and the prior of the latent variable is set as a standard Gaussian $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The objective of VAEs is to maximize the expected log-likelihood over data distribution Since it is tractable to optimize this likelihood directly, we introduce an approximated posterior $q_\phi(\mathbf{z}|X)$ and instead optimize the following evidence lower bound (ELBO) on the expected log-likelihood:

$$\mathcal{L}(\theta, \phi) \equiv \mathbb{E}_{p_d(X)}[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|X)||p(\mathbf{z}))] \leq \mathbb{E}_{p_d(X)}[p_\theta(X)]. \quad (1)$$

In this paper, the model group that maximizes this *multimodal ELBO* is collectively called *multimodal VAEs*. We aim to optimize Eq. 1 to obtain a joint representation and generate modalities from it given a modality subset X_S. However, Eq. 1 only includes the inference given all modalities; therefore, it does not learn to infer representations from subsets.

2.2 Mixture-based multimodal VAEs

An inference from subsets is commonly expressed as a combination of unimodal posteriors from each modality. Product of experts (PoE) (Hinton, 2002; Wu & Goodman, 2018) $q_\phi^{PoE}(\mathbf{z}|X_S) \equiv$

[2] The term *coordination* is taken from *coordinated representations*, a category of multimodal learning (Baltrušaitis et al., 2018), in which representations inferred from different modalities are learned close together.

[3] We denote a subset of the multimodal set X by X_S respectively, where S represents a subset of $\{1, .., M\}$.

$p(\mathbf{z})\prod_{m:\mathbf{x}_m\in X_S} q_\phi(\mathbf{z}|\mathbf{x}_m)$ and mixture of experts (MoE) $q_\phi^{MoE}(\mathbf{z}|X_S) \equiv \sum_{m:\mathbf{x}_m\in X_S} q_\phi(\mathbf{z}|\mathbf{x}_m)$ are known as ways to aggregate inferences. Moreover, the following mixture of products of experts (MoPoE) (Sutter et al., 2021) generalizes them:

$$q_\phi^{MoPoE}(\mathbf{z}|X) \equiv \sum_{S:X_S\in\mathcal{P}(X)} \omega_S \left(\prod_{m:\mathbf{x}_m\in X_S} p(\mathbf{z})q_\phi(\mathbf{z}|\mathbf{x}_m)\right) = \sum_{S:X_S\in\mathcal{P}(X)} \omega_S q_\phi^{PoE}(\mathbf{z}|X_S), \tag{2}$$

where $\sum_{S:X_S\in\mathcal{P}(X)} \omega_S = 1$, $\omega_S \in [0,1]$, and $\mathcal{P}(X)$ is a powerset of X. This MoPoE aggregation is equivalent to PoE when $X_S = X$ and MoE when $|X_S| = 1$ for all $X_S \in \mathcal{P}(X)$.

Substituting the MoPoE-aggregated joint posterior (Eq. 2) into Eq. 1, we obtain the following:

$$\mathcal{L}(\theta,\phi) \geq \mathbb{E}_{p_d(X)}\left[\sum_{S:X_S\in\mathcal{P}(X)} \omega_S \left[\mathbb{E}_{q_\phi^{PoE}(\mathbf{z}|X_S)}[\log p_\theta(X|\mathbf{z})] - D_{KL}(q_\phi^{PoE}(\mathbf{z}|X_S)||p(\mathbf{z}))\right]\right] \equiv \mathcal{L}^M(\theta,\phi). \tag{3}$$

Multimodal VAEs with Eq. 3 are called *mixture-based multimodal VAEs* (Daunhawer et al., 2021a). MMVAE (Shi et al., 2019) and MoPoE-VAE (Sutter et al., 2021) are members of this family.

Comparing with Eq. 1, we can see that the reconstruction term (the first term in the expectation) of Eq. 1 is designed to generate all modalities from the sub-sampled modality subset. In other words, mixture-based multimodal VAEs encourage cross-generation during training, aiming at inferring a consistent joint representation from subset modalities and generating modalities from it.

However, mixture-based VAEs have a severe issue in generating modalities, which is one of the essential goals of multimodal VAEs. Daunhawer et al. (2021a) show that the following inequality holds between Eq. 3 and the expected marginal log-likelihood:

$$\mathbb{E}_{p_d(X)}[\log p_\theta(X)] \geq \Delta(X) + \mathcal{L}^M(\theta,\phi;X), \tag{4}$$

where $\Delta(X) \equiv \sum_{S:X_S\in\mathcal{P}(X)} w_S H(X_{\{1,\ldots,M\}\setminus S}|X_S)$ is the discrepancy independent of the learning parameters. This inequality means that multimodal ELBO of mixture-based VAEs cannot be tighter than this discrepancy $\Delta(X)$ no matter how much it is maximized. Therefore, the quality of their generating modalities always degrades compared to unimodal VAEs. Moreover, this discrepancy, by definition, depends on how much different information one modality subset $X_{\{1,\ldots,M\}\setminus S}$ has compared to the rest of them X_S. Therefore, the more information specific to each modality, the larger this quantity and the worse the generation quality.

This also leads to degrading cross-coherence (Shi et al., 2019), or the ability to generate the same semantic meaning across modalities conditionally. Intuitively, mixture-based VAEs are required to take sub-sampled subsets as inputs to generate the rest, but these inputs have insufficient information for that, resulting in degraded cross-generation.

3 Methods

3.1 Avoiding the limitation of multimodal VAEs

What brings about this irreducible discrepancy $\Delta(X)$? According to the proof by Daunhawer et al. (2021a) (see B.4 in Daunhawer et al. (2021a)), it is due to the inclusion in the objective of a cross-generation term, i.e., a term that generates unobserved modalities from observed modalities. Therefore, one promising way to avoid this limitation is to *not include in the objective of the multimodal VAE the cross-generation term*[4].

Note that MVAE (Wu & Goodman, 2018), which first uses PoE for the joint posterior, uses the sum of ELBOs over all modality subsets as the objective, but does not include cross-generation and only includes *reconstruction* of subsets[5], thus enabling high-quality generation. However, MVAE cannot perform cross-generation between different modalities because its objective does not include terms to encourage it. Despite this significant drawback, determining how to work around the limitation of mixture-based models leads to our proposed model described later.

[4]Note that the reconstruction term $\mathbb{E}_{q(\mathbf{z}|X_S)}[\log p(X_{S'}|\mathbf{z})]$ (where $X_{S'} \in X_S$), i.e., all the modalities to be generated are subsets of the input, is not included in the cross-generation term discussed here.

[5]Wu & Goodman (2018) state that MVAE utilizes a sub-sampling training paradigm, but this means "ELBOs with sub-sampled modality sets as input", i.e., each ELBO does not include cross-generation term.

3.2 COORDINATION-BASED MULTIMODAL VAEs

In the study of multimodal VAEs, MVTCAE (Hwang et al., 2021) and MMJSD (Sutter et al., 2020) have been proposed to learn to infer joint representation from subset modalities by bringing joint and unimodal posteriors closer together. They omit sub-sampling in the objective unless the posterior is a mixture distribution and thus can learn the joint representation without being affected by the limitation of the mixture-based VAEs. However, they are not proposed in light of the above limitation nor derived from the same objective. Therefore, we first consider them as a unified *coordination-based multimodal VAEs* and analyze their relation [6].

3.2.1 MVTCAE

Hwang et al. (2021) derive the following objective as a lower bound for the total correlation:

$$\mathcal{L}^C(\theta,\phi) \equiv \mathbb{E}_{p_d(X)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})] - \sum_{m=0}^{M}\pi_m D_{KL}(q_\phi(\mathbf{z}|X)||q_\phi(\mathbf{z}|\mathbf{x}_m))\right], \tag{5}$$

where $\sum_m \pi_m = 1$ and $\pi_m \in [0,1]$. Note that we changed the coefficients and removed non-optimized terms from the original paper (Hwang et al., 2021). Moreover, for simplicity of notation, we use $q_{\phi_0}(\mathbf{z}|\mathbf{x}_0) \equiv p(\mathbf{z})$. The model trained with this objective is called MVTCAE. In particular, when $M = 2$, this model is equal to JMVAE (Suzuki et al., 2016). However, JMVAE uses a network for joint posterior separately from unimodal posteriors, while MVTCAE represents joint posteriors by aggregating unimodal posteriors in PoE.

3.2.2 MMJSD

Sutter et al. (2020) propose the following objective, called MMJSD, which replaces the KL divergence of the multimodal ELBO with the Jensen-Shannon divergence:

$$\mathcal{L}^{C'}(\theta,\phi) = \mathbb{E}_{p_d(X)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})] - \sum_{m=0}^{M}\pi_m D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}_m)||f_{\mathcal{M}}(\mathbf{z}|X))\right], \tag{6}$$

where function $f_{\mathcal{M}}$ defines the mixture distribution of $\{q_\phi(\mathbf{z}|\mathbf{x}_m)\}_{m=0}^M$. It is proved (in Lemma 2 in Sutter et al. (2020)) that Eq. 6 is a lower bound for multimodal ELBO with $f_{\mathcal{M}}(\mathbf{z}|X)$ replaced by the following dynamic prior: $p^f(\mathbf{z}|X) = f(\{q_\phi(\mathbf{z}|\mathbf{x}_m)\}_{m=0}^M)$, where f is a function that encompasses each unimodal posterior, and Sutter et al. (2020) choose the weighted PoE. With this replacement for the dynamic prior, MMJSD can be regarded as a member of multimodal VAEs[7]. Moreover, this replacement can be thought of as introducing the *prior knowledge* that all modalities have the same joint representation (Sutter et al., 2020).

Note that in Sutter et al. (2020) the posterior of MMJSD is assumed to be MoE, which is affected by the limitation of mixture-based VAEs. Since we would like to focus on the properties of MMJSD contributing to its coordination-based nature, we call *MMJSD-PoE* the model that replaces the posterior in the reconstruction term of Eq. 6 with PoE. Hereafter, MMJSD is positioned in this paper as a member of mixture-based VAEs, and MMJSD-PoE as coordination-based VAEs.

Compared to the objective of MVTCAE, we can see that the *direction* of KL divergence in Eq. 6 is reversed: MVTCAE uses forward KL divergence, while MMJSD(-PoE) uses reverse KL divergence. (Hwang et al., 2021) shows that MVTCAE has better performance than MMJSD by experiment, but the theoretical reason is still unclear.

3.3 RELATION BETWEEN COORDINATION-BASED MULTIMODAL VAEs

In this section, by using the dynamic prior, we show that: 1) the objective of MVTCAE is multimodal ELBO; 2) in special cases, the objective of MVTCAE is an upper bound on the objective of MMJSD.

[6] Note that the coordination-based VAEs limited to two modalities are not summarized here because they are not scalable. See Section 3.5 for a description of them.

[7] In Sutter et al. (2020), Eq. 6 replaced by the dynamic prior is called MMJSD, which we follow.

Lemma 1 *The objective of MVTCAE in Eq. 5 is equal to the multimodal ELBO in Eq. 1 with a weighted PoE-based dynamic prior as its prior.*

Proof 1 *Following MMJSD, we introduce a weighted PoE-based dynamic prior for Eq. 3 as follows:*

$$\mathbb{E}_{p_d(X)}[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|X)||p^f(\mathbf{z}|X))]]$$
$$= \mathbb{E}_{p_d(X)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})] - \mathbb{E}_{q_\phi(\mathbf{z}|X)}\left[\sum_{m=0}^{M}\pi_m q_\phi(\mathbf{z}|X) - \pi_m q_\phi(\mathbf{z}|\mathbf{x}_m)\right]\right] = \mathcal{L}^C(\theta,\phi).\square$$

Therefore, MVTCAE can be considered a member of multimodal VAEs. Furthermore, this obviously indicates that the objective of JMVAE (Suzuki et al., 2016) is also multimodal ELBO.

Lemma 2 *If the posterior is MoE-aggregated, the objective of MVTCAE in Eq. 5 is an upper bound on that of MMJSD in Eq. 6, i.e., the following inequality holds:* $\mathcal{L}^C(\theta,\phi) \geq \mathcal{L}^{C'}(\theta,\phi)$.

Proof 2 *Lemma 1 shows that the objective of MVTCAE is equal to a multimodal ELBO with a dynamic prior. In addition, Lemma 2 in Sutter et al. (2020) states that the objective of MMJSD is a lower bound on multimodal ELBO with a dynamic prior and MoE-aggregated posterior. Therefore, the objective of MMJSD is a lower bound on that of MVTCAE if the posterior is MoE-aggregated.* $\square$

Lemma 2 indicates that MVTCAE better approximates the true log-likelihood and thus promotes good generation, which provides one possible explanation as to why MVTCAE performs better than MMJSD. Note that Lemma 2 comes from the fact that the joint posterior of the KL divergence in MMJSD assumes MoE; therefore, it is still unclear whether this holds when the posterior is PoE.

3.4 Coordinated and unimodal Reconstructed Multimodal VAEs

Coordination-based multimodal VAEs are intended to keep inference from any given modalities consistent by bringing joint and unimodal posteriors closer. However, in practice, it is impossible to make them perfectly close; therefore, even if modality generation is appropriately learned from the joint posterior in the reconstruction term, it remains difficult to properly generate modalities from the unimodal posterior. On the other hand, including cross-generation in the objective conversely promotes generation degradation due to the limitation of mixture-based models.

Therefore, for the unimodal posterior to *experience* generation from its inference during learning while avoiding these limitations, including a term for *unimodal reconstruction* in the objective is an appropriate way. The reconstruction term in Eq. 5 can be rewritten as

$$\mathbb{E}_{p_d(X)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})]\right]$$
$$= \frac{1}{2}\mathbb{E}_{p_d(X)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})]\right] + \frac{1}{2}\sum_{m=1}^{M}\mathbb{E}_{p_d(\mathbf{x}_m)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_m)}[\log p_\theta(\mathbf{x}_m|\mathbf{z})]\right], \quad (7)$$

where we use $\int q_\phi(\mathbf{z}|X)p_d(X_{C/m}|\mathbf{x}_m)dX_{C/m} = q_\phi(\mathbf{z}|\mathbf{x}_m)$[8] for each $\mathbf{x}_m$ in the deformation of Eq. 7. Therefore, the objective of MVTCAE can be transformed to include all-modality and unimodal reconstructions, remaining a valid multimodal ELBO. This also mitigates the shortcoming of coordination-based VAEs while avoiding the limitation of mixture-based VAEs. We call this novel model *Coordinated and unimodal Reconstructed Multimodal VAE (CRMVAE)*.

3.5 Related works

There have been many studies of multimodal VAEs (Suzuki & Matsuo, 2022), and the early ones before mixture-based required additional inference models. Suzuki et al. (2016) proposed to approximate the unimodal inference models to explicitly approach the joint posterior of VAEs, called JMVAE. Vedantam et al. (2018) introduced TELBO, where the objective is composed of the sum of ELBOs that take each set of modalities as input. M^2VAE (Korthals et al., 2019) can be regarded as a combination of JMVAE and TELBO, and VAEVAE (Wu & Goodman, 2019) excludes a KL

[8]This deformation also used in Wu & Goodman (2019) to show that the objective is a valid ELBO.

divergence term between the joint inference and prior from M^2VAE. Except for TELBO, they can be regarded as precursors to the coordination-based approach; however, the number of inference models needed for these models grows exponentially with the number of modalities.

One important research direction in multimodal VAEs is the introduction of additional latent variables, which includes hierarchical latent variables (Sutter & Vogt, 2021; Wolff et al., 2021; Vasco et al., 2022) and modality-specific latent variables (Tsai et al., 2018; Hsu & Glass, 2018; Sutter et al., 2020; Lee & Pavlovic, 2020; Daunhawer et al., 2021b; Palumbo et al., 2022). This paper aims to improve the performance of multimodal VAEs in a different direction from them, i.e., whether limitations of existing mixture-based models can be mitigated by changing the objective while *maintaining the same architecture*. Note that our proposed model and the introduction of additional latent variables are orthogonal concepts and can be easily incorporated into the proposed model.

Recently, it has been proposed to translate between modalities by learning large-scale models such as transformers (Vaswani et al., 2017). Ramesh et al. (2022) used CLIP (Radford et al., 2021) for conditional generation between diverse modalities at higher performance. Bachmann et al. (2022) introduced MultiMAE, which uses vision transformers (Dosovitskiy et al., 2020) and allows cross-generation from any modality and fine-tuning for new tasks. However, these models do not aim to obtain a consistent joint representation like multimodal VAEs. Future directions should integrate multimodal studies using these large-scale models with multimodal VAEs.

4 Experiments

We use the following five diverse datasets for our experiments. **MNIST-SVHN-Text** (Sutter et al., 2020) is the MNIST-SVHN (Shi et al., 2019) with the addition of the text modality, which represents the English name of each number as text and has diversity. Different modalities of the same example have the same digits. **PolyMNIST** (Sutter et al., 2021) consists of a set of five different MNIST images with different backgrounds and handwriting styles, each of which is considered a different modality. **Translated-PolyMNIST** (Daunhawer et al., 2021a) is composed of the same backgrounds as PolyMNIST with the numbers downscaled[9] and randomly transformed into position in the images. This is intended to allow for more significant variation among modalities than PolyMNIST. Caltech Birds (**CUB**) Wah et al. (2011) is a dataset with bird images and their captions. We follow Daunhawer et al. (2021a), i.e., we treat as 64×64 real images without feature extraction. **Bimodal CelebA** (Sutter et al., 2020) is a dataset of face images from the CelebA (Liu et al., 2015) with additional text describing the corresponding attribute labels and changing the starting index to give a diversity. These datasets are commonly used in studies of multimodal VAEs, but to our knowledge, this is the first study that experiments have been conducted on *all* of these datasets. In particular, CUB and Translated-PolyMNIST were noted to be challenging datasets that could not be successfully trained in existing models (Daunhawer et al., 2021a). It also shows that Bimodal CelebA and CUB could only be learned in models with additional latent variables (Sutter et al., 2020; Palumbo et al., 2022).

To evaluate the models in terms of both 1) cross-coherence and 2) generation quality, we measure each of them in terms of 1) the test accuracy of a pre-trained classifier[10] that predicts labels from cross-generated images and 2) the Fréchet inception distance (FID) (Heusel et al., 2017) in cross-generated modalities. Note that we evaluated the generated results of the text modality in MNIST-SVHN-Text in terms of the Levenshtein distance from the true text (without space characters) since we found that the pre-trained classifier from Sutter et al. (2021) did not correctly evaluate the generated results (we discuss in detail in the experimental results.) Next, to assess the cross-coherence for CUB, we use the cosine similarity between the latent space of the two inferred modalities since there are no labels. We also used this evaluation of cross-coherence by cosine similarity for Bimodal CelebA.

For the network architectures, we follow Sutter et al. (2020); Daunhawer et al. (2021a), with 64-dimensional latent variables for Bimodal CelebA and CUB, 32 for MNIST-SVHN-Text, and 512 for PolyMNIST and Translated-PolyMNIST. The distributions for image modalities are set to Laplace, and text is set to categorical. We use Adam (Kingma & Ba, 2014) for optimization with a learning rate of 0.0005 and a mini-batch size of 256. We set β (the coefficient of the KL divergence term in the multimodal ELBO) to 1 for PolyMNIST and MNIST-SVHN-Text and 0.1 for Translated-PolyMNIST,

[9] In setting of this study, the handwritten numbers were downscaled to 70% of their original size.

[10] We used the pre-trained classifier in PolyMNIST and MNIST-SVHN-Text from Sutter et al. (2021).

Table 1: Cross-coherence by classification accuracy on MNIST-SVHN-Text. M, S, and T stand for MNIST, SVHN, and Text. M+S means both M and S. Bold represents the highest accuracy in each modality. The top line shows modalities generated, and the second line shows modalities conditioned.

		M			S			T		
	MODEL	S	T	S+T	M	T	M+T	M	S	M+S
	MVAE	0.29	0.11	0.30	0.36	0.29	0.67	0.13	0.20	0.22
Mixture-based	MMVAE	0.81	**0.99**	0.91	0.39	0.44	0.41	0.48	0.40	0.44
	MoPoE-VAE	**0.82**	**0.99**	0.94	0.35	0.37	0.36	0.49	0.42	0.51
	MMJSD	0.81	**0.99**	0.90	0.39	0.41	0.40	0.44	0.37	0.40
Coodination-based	MMJSD-PoE	0.12	0.78	0.80	0.56	0.73	0.86	0.71	0.14	0.74
	MVTCAE	0.23	0.87	0.90	0.59	0.78	0.80	0.76	0.27	0.79
	CRMVAE	0.60	0.97	**0.97**	**0.73**	**0.87**	**0.87**	**0.91**	**0.71**	**0.95**

Table 2: FID scores for cross-generation. Lower is better.

MODEL	S→M	M→S
MVAE	39.0	36.0
MMVAE	91.8	307.9
MoPoE-VAE	85.3	312.8
MMJSD	86.7	287.0
MMJSD-PoE	70.0	141.4
MVTCAE	41.6	**31.8**
CRMVAE	**24.1**	37.3

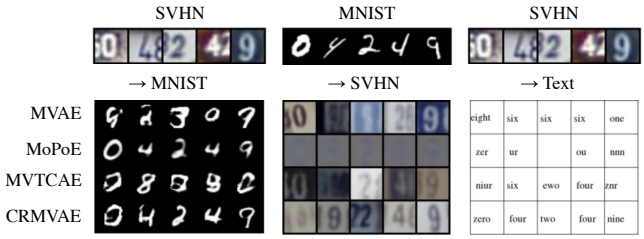

Figure 1: Visualization of cross-generation.

π_m to $\frac{1}{M+1}$, and ω_S to $\frac{1}{2^M}$. We train 100 epochs for MNIST-SVHN-Text, 150 for CUB, 200 for Bimodal CelebA, and 500 for PolyMNIST and Translated-PolyMNIST. All experimental results are the average of three runs with different seeds. We implement them using PyTorch (Paszke et al., 2019). Our experiments were performed using two machines with five NVIDIA Titan RTX GPUs. We compare CRMVAE with MVAE, mixture-based VAEs (MMVAE, MoPoE, and MMJSD), and coordination-based VAEs (MVTCAE and MMJSD-PoE), all with the same architecture settings.

4.1 MNIST-SVHN-TEXT

Table 1 shows the cross-coherence results, and Table 2 shows the FID scores of images on MNIST-SVHN-Text. First, we can confirm that MVAE has high generation quality but the lowest cross-coherence since it learns only the modality reconstruction. We next find that mixture-based models perform poorly in cross-generation of SVHN and Text. Note here that the results of Text differ significantly from the results of previous studies because the Levenshtein distance is used for the generated results, not for the pre-trained classifiers. According to the qualitative results shown in Figure 1, the cross-generation of Text on MoPoE-VAE is not successful; however, the pre-trained classifier used in the previous study evaluates it as a high performance. On the other hand, we find that the evaluation using the Levenshtein distance is consistent with the qualitative results, so we consider this more appropriate. While Table 1 shows that cross-generation of MNIST appears to perform well, the qualitative results in Figure 1 show that the generated images are average with no variation. Table 2 indicates that FID scores of mixture-based models on MNIST are significantly lower than these of MVAE, which corresponds to these results.

Existing coordination-based models mitigate this limitation and perform better on SVHN. Furthermore, FID scores show that the coordination-based models significantly outperform the mixture-based models in terms of generation quality and are comparable to MVAE. Comparing MMJSD-PoE and MVTCAE, MVTCAE performs better in terms of both cross-coherence and generation quality, confirming the effectiveness of MVTCAE experimentally. However, the performance of the coordination-based approach is degraded by cross-generation, especially from SVHN.

CRMVAE mitigates coordination-based issues and significantly improves cross-coherence performance. In particular, CRMVAE outperforms all existing methods, including mixture-based methods, for cross-coherence of SVHN and Text. We also confirm that CRMVAE retains the advantage of the

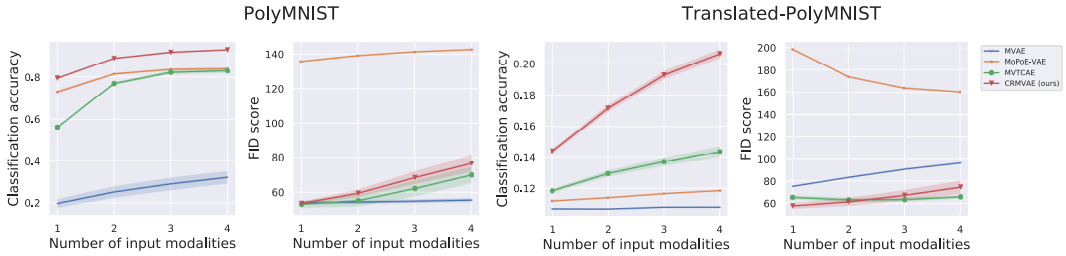

Figure 2: Performance evaluation on PolyMNIST and Translated-PolyMNIST. The left panel for each dataset shows the cross-coherence given subset of size corresponding to each number and the right panel shows the FID score given subset (lower is better). The markers are the means of the three runs, and the error bands are their standard deviations.

coordination-based models, i.e., high generation quality. In particular, we have confirmed the highest quality of cross-generation on MNIST among all models that include MVAE. The visualization results also confirm that CRMVAE achieves high coherence while maintaining high quality.

4.2 PolyMNIST and Translated-PolyMNIST

Figure 2 shows the results on PolyMNIST and Translated-PolyMNIST. Based on the MNIST-SVHN-Text results, we select MoPoE-VAE and MVTCAE as mixture and coordination-based models for comparison. First the results of the existing models on PolyMNIST show the same trend as on MNIST-SVHN-Text, i.e., there is a trade-off between cross-coherence and quality of generation. CRMVAE mitigates these limitations and achieves the highest cross-coherence while maintaining high generation quality. Note that it might seem strange that both MVTCAE and CRMVAE produce lower quality when the number of modalities increases. This is thought to be due to a decrease in the diversity of the generated data distribution as the number of modalities increases.

Translated-PolyMNIST is a slightly modified PolyMNIST, but Daunhawer et al. (2021a) point out that the generation quality and cross-coherence performance of mixed-based models is significantly degraded on this dataset. Our experiments also confirm that most existing models fail to cross-generate across modalities. CRMVAE, on the other hand, clearly improves cross-coherence, although not as much as PolyMNIST. Next, CRMVAE also has the highest generation quality, especially significantly outperforming the mixture-based models. See Appendix C.1 for visualization of the generated images on PolyMNIST and Translated-PolyMNIST.

4.3 CUB and Bimodal CelebA

Figures 3 and 4 show the quantitative and qualitative results for CUB and Bimodal CelebA, respectively. We examine the generation quality and cross-coherence performance as an ablation, when varying β. First, the FID score indicates that MVAEs have high generation quality, while MoPoE-VAEs have significantly lower quality. In most cases, the FID score does not change much or worsen when β is increased, while the FID score for MoPoE-VAE improves slightly. This is thought to be because increasing β in MoPoE-VAE corresponds to making the distribution closer to the standard Gaussian, which diminishes the effect of generation from sub-sampling. Furthermore, the performance of the coordination-based methods, including CRMVAE, is close to that of MVAE, indicating that they do not have a generation quality issue on these challenging datasets.

Next, evaluation of cosine similarity shows that MVAE and MoPoE-VAE do not change that much when β is increased, while the coordination-based models including CRMVAE improve when β is increased. This is because in the coordination-based models, the term adjusted by β corresponds to the term that brings the inference of all modalities closer to that of each modality. Compared to MVTCAE, CRMVAE performs better for Bimodal CelebA for all values of β and for CUB when β is large. Furthermore, the cross-coherence with pre-trained classifiers in Bimodal CelebA also shows that CRMVAE performs better than MVTCAE. Note that MoPoE-VAE has the highest

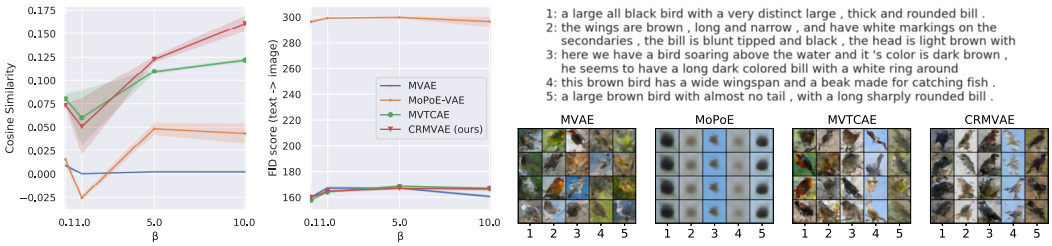

Figure 3: Left: Performance of the cosine similarity of the two modalities in latent space (higher is better, left) and the FID score of the image generated from text (right) for varying β in the CUB dataset. Right: Visualization of the CUB images generated by each model ($\beta = 10$, bottom) conditioned on each text (top). The number in each column of each model represents the number of the text being conditioned. Different rows correspond to different sample results.

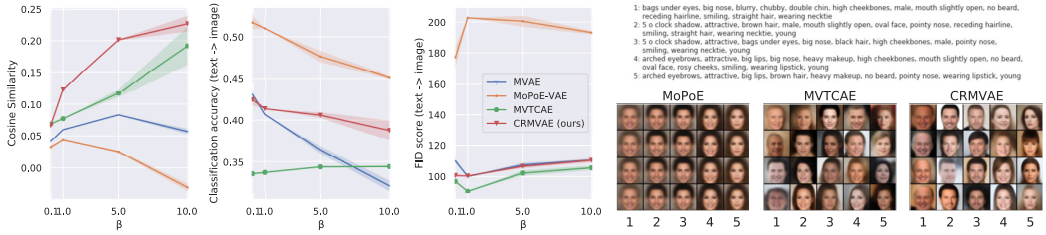

Figure 4: Left: Performance of the cosine similarity (left), classification accuracy (center), and FID score (right) for varying β in the Bimodal CelebA dataset. Right: Visualization of the Bimodal CelebA images generated by each model ($\beta = 5$, bottom) conditioned on each text (top).

cross-coherence value, but as the qualitative results show, this only improves classification accuracy at the expense of image quality, e.g., diversity (as is the case for MNIST in MNIST-SVHN-Text).

Qualitative results showed the following limitations: MVAE failed to produce images corresponding to the text, and MoPoE-VAE produced only average or blurred images. Comparing MVTCAE and CRMVAE, CRMVAE produces more appropriately conditioned images. For example, in Bimodal CelebA, Text 3 contains a male attribute and CRMVAE produces a corresponding face image, while MVTCAE produces some female face images.

5 Conclusion and Discussion

We focused on coordination-based multimodal VAEs to avoid the limitations of multimodal VAEs. We first discuss the new relation among coordination-based multimodal VAEs as multimodal VAEs, point out the issue of unimodal inference not learning from reconstruction in them, and then introduce a new objective involving unimodal reconstruction. Experiments show that the proposed model mitigates the limitations of existing models (both mixture-based and coordination-based) and shows overall superior performance on diverse datasets for both generation quality and cross-coherence.

One limitation of CRMVAE is that there still remains a trade-off between generation quality and cross-coherence. For example, in the PolyMNIST experiment, CRMVAE performed best in cross-coherence, but the FID score was worse than MVTCAE. This is also true for the other results, where the FID score is slightly worse than MVTCAE in exchange for the performance of cross-coherence. Also, although we achieved the performance improvement for our objective without changing the architecture, it is not obvious that the same improvement can be achieved for other architectures, such as those with modality-specific or hierarchical latent variables. In addition, although we evaluated the performance on several diverse datasets in this study, these datasets are much smaller than the various multimodal information obtained from the real world. Therefore, it is necessary to validate our model using such large multimodal information.

REFERENCES

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. *arXiv preprint arXiv:2204.01678*, 2022.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Imant Daunhawer, Thomas M Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E Vogt. On the limitations of multimodal vaes. *arXiv preprint arXiv:2110.04121*, 2021a.

Imant Daunhawer, Thomas M Sutter, Ričards Marcinkevičs, and Julia E Vogt. Self-supervised disentanglement of modality-specific and shared factors improves multimodal generative models. *Pattern Recognition*, 12544:459, 2021b.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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. *Advances in neural information processing systems*, 30, 2017.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Wei-Ning Hsu and James Glass. Disentangling by partitioning: A representation learning framework for multimodal sensory data. *arXiv preprint arXiv:1805.11264*, 2018.

HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Multi-view representation learning via total correlation objective. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12194–12207. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/65a99bb7a3115fdede20da98b08a370f-Paper.pdf.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Timo Korthals, Daniel Rudolph, Jürgen Leitner, Marc Hesse, and Ulrich Rückert. Multi-modal generative models for learning epistemic active sensing. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 3319–3325. IEEE, 2019.

Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal vae for learning of hybrid latent representations. *arXiv preprint arXiv:2012.13024*, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Emanuele Palumbo, Imant Daunhawer, and Julia E Vogt. MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022. URL https://openreview.net/forum?id=B42UJTVdDZ5.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and

R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Yuge Shi, Narayanaswamy Siddharth, Brooks Paige, and Philip Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems*, pp. 15718–15729, 2019.

Thomas M Sutter and Julia E Vogt. Multimodal relational vae. 2021.

Thomas M Sutter, Imant Daunhawer, and Julia E Vogt. Multimodal generative learning utilizing jensen-shannon-divergence. *arXiv preprint arXiv:2006.08242*, 2020.

Thomas M Sutter, Imant Daunhawer, and Julia E Vogt. Generalized multimodal elbo. *arXiv preprint arXiv:2105.02470*, 2021.

Masahiro Suzuki and Yutaka Matsuo. A survey of multimodal deep generative models. *Advanced Robotics*, 36(5-6):261–278, 2022.

Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

Miguel Vasco, Hang Yin, Francisco S Melo, and Ana Paiva. Leveraging hierarchy in multimodal generative models for effective cross-modality inference. *Neural Networks*, 146:238–255, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. In *International Conference on Learning Representations*, 2018.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Jannik Wolff, Rahul G Krishnan, Lukas Ruff, Jan Nikolas Morshuis, Tassilo Klein, Shinichi Nakajima, and Moin Nabi. Hierarchical multimodal variational autoencoders. 2021.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 5575–5585, 2018.

Mike Wu and Noah Goodman. Multimodal generative models for compositional representation learning. *arXiv preprint arXiv:1912.05075*, 2019.

A DETAILED DERIVATION OF PROOF 1

$$
\begin{aligned}
&\mathbb{E}_{p_d(X)}[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|X)||p^f(\mathbf{z}|X))] \\
&= \mathbb{E}_{p_d(X)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})] - \mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log q_\phi(\mathbf{z}|X) - \log p^f(\mathbf{z}|X)]\right] \\
&= \mathbb{E}_{p_d(X)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})] - \mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log q_\phi(\mathbf{z}|X) - \sum_{m=0}^{M}\pi_m \log q_\phi(\mathbf{z}|\mathbf{x}_m)]\right] \\
&\qquad\qquad \left(\because p^f(\mathbf{z}|X) = \sum_{m=0}^{M}\pi_m \log q_\phi(\mathbf{z}|\mathbf{x}_m)\right) \\
&= \mathbb{E}_{p_d(X)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})] - \mathbb{E}_{q_\phi(\mathbf{z}|X)}\left[\sum_{m=0}^{M}\pi_m \log q_\phi(\mathbf{z}|X) - \sum_{m=0}^{M}\pi_m \log q_\phi(\mathbf{z}|\mathbf{x}_m)\right]\right] \\
&\qquad\qquad \left(\because \log q_\phi(\mathbf{z}|X) = \sum_{m=0}^{M}\pi_m \log q_\phi(\mathbf{z}|X) \text{ and } \sum_{m=0}^{M}\pi_m = 1\right) \\
&= \mathbb{E}_{p_d(X)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})] - \sum_{m=0}^{M}\pi_m \mathbb{E}_{q_\phi(\mathbf{z}|X)}\left[\log q_\phi(\mathbf{z}|X) - \log q_\phi(\mathbf{z}|\mathbf{x}_m)\right]\right] \\
&= \mathbb{E}_{p_d(X)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})] - \sum_{m=0}^{M}\pi_m D_{KL}\left(q_\phi(\mathbf{z}|X)||q_\phi(\mathbf{z}|\mathbf{x}_m)\right)\right] \\
&= \mathcal{L}^C(\theta,\phi)\square
\end{aligned}
$$

B DETAILED DERIVATION OF EQ. (7)

$$
\begin{aligned}
&\mathbb{E}_{p_d(X)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})]\right] \\
&= \mathbb{E}_{p_d(X)}\left[\frac{1}{2}\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})] + \frac{1}{2}\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})]\right] \\
&= \mathbb{E}_{p_d(X)}\left[\frac{1}{2}\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})] + \frac{1}{2}\mathbb{E}_{q_\phi(\mathbf{z}|X)}\left[\sum_{m=1}^{M}\log p_\theta(\mathbf{x}_m|\mathbf{z})\right]\right] \\
&= \mathbb{E}_{p_d(X)}\left[\frac{1}{2}\mathbb{E}_{q_\phi(\mathbf{z}|X)}\left[\log p_\theta(X|\mathbf{z})\right]\right] + \frac{1}{2}\int\int q_\phi(\mathbf{z}|X)p_d(X)\sum_{m=1}^{M}\log p_\theta(\mathbf{x}_m|\mathbf{z})d\mathbf{z}dX \\
&= \mathbb{E}_{p_d(X)}\left[\frac{1}{2}\mathbb{E}_{q_\phi(\mathbf{z}|X)}\left[\log p_\theta(X|\mathbf{z})\right]\right] + \frac{1}{2}\sum_{m=1}^{M}\int\int q_\phi(\mathbf{z}|X)p_d(X)\log p_\theta(\mathbf{x}_m|\mathbf{z})d\mathbf{z}dX \\
&= \mathbb{E}_{p_d(X)}\left[\frac{1}{2}\mathbb{E}_{q_\phi(\mathbf{z}|X)}\left[\log p_\theta(X|\mathbf{z})\right]\right] + \frac{1}{2}\sum_{m=1}^{M}\int\int q_\phi(\mathbf{z}|X)p_d(X_{C/m}|\mathbf{x}_m)p_d(\mathbf{x}_m)\log p_\theta(\mathbf{x}_m|\mathbf{z})d\mathbf{z}dX \\
&= \mathbb{E}_{p_d(X)}\left[\frac{1}{2}\mathbb{E}_{q_\phi(\mathbf{z}|X)}\left[\log p_\theta(X|\mathbf{z})\right]\right] + \frac{1}{2}\sum_{m=1}^{M}\int\int\int q_\phi(\mathbf{z}|X)p_d(X_{C/m}|\mathbf{x}_m)dX_{C/m}p_d(\mathbf{x}_m)\log p_\theta(\mathbf{x}_m|\mathbf{z})d\mathbf{z}d\mathbf{x}_m \\
&= \mathbb{E}_{p_d(X)}\left[\frac{1}{2}\mathbb{E}_{q_\phi(\mathbf{z}|X)}\left[\log p_\theta(X|\mathbf{z})\right]\right] + \frac{1}{2}\sum_{m=1}^{M}\int\int q_\phi(\mathbf{z}|\mathbf{x}_m)p_d(\mathbf{x}_m)\log p_\theta(\mathbf{x}_m|\mathbf{z})d\mathbf{z}d\mathbf{x}_m \\
&\qquad\qquad \left(\because \int q_\phi(\mathbf{z}|X)p_d(X_{C/m}|\mathbf{x}_m)dX_{C/m} = q_\phi(\mathbf{z}|\mathbf{x}_m)\right) \\
&= \frac{1}{2}\mathbb{E}_{p_d(X)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X|\mathbf{z})]\right] + \frac{1}{2}\sum_{m=1}^{M}\mathbb{E}_{p_d(\mathbf{x}_m)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_m)}[\log p_\theta(\mathbf{x}_m|\mathbf{z})]\right] \qquad (8)
\end{aligned}
$$

Table 3: Cross-coherence on MNIST-SVHN-Text with standard deviations.

	M			S		
MODEL	S	T	S+T	M	T	M+T
MVAE	0.29± 0.04	0.11± 0.00	0.30± 0.04	0.36± 0.08	0.29± 0.08	0.67± 0.05
MMVAE	0.81± 0.00	0.99± 0.00	0.91± 0.00	0.39± 0.01	0.44± 0.02	0.41± 0.01
MoPoE-VAE	0.82± 0.00	0.99± 0.00	0.94± 0.01	0.35± 0.02	0.37± 0.02	0.36± 0.02
MMJSD	0.81± 0.01	0.99± 0.00	0.90± 0.00	0.39± 0.01	0.41± 0.01	0.40± 0.01
MMJSD-PoE	0.12± 0.01	0.78± 0.03	0.80± 0.03	0.56± 0.01	0.73± 0.03	0.86± 0.01
MVTCAE	0.23± 0.01	0.87± 0.02	0.90± 0.01	0.59± 0.01	0.78± 0.00	0.80± 0.01
CRMVAE	0.60± 0.01	0.97± 0.01	0.97± 0.01	0.73± 0.01	0.87± 0.01	0.87± 0.01

	T		
MODEL	M	S	M+S
MVAE	0.13± 0.01	0.20± 0.07	0.22± 0.08
MMVAE	0.48± 0.05	0.40± 0.04	0.44± 0.05
MoPoE-VAE	0.49± 0.04	0.42± 0.03	0.51± 0.04
MMJSD	0.44± 0.02	0.37± 0.01	0.40± 0.02
MMJSD-PoE	0.71± 0.02	0.14± 0.02	0.74± 0.04
MVTCAE	0.76± 0.02	0.27± 0.02	0.79± 0.02
CRMVAE	0.91± 0.02	0.71± 0.00	0.95± 0.01

Table 4: FID score of the generated SVHN images with standard deviations.

MODEL	S→M	M→S
MVAE	39.0± 4.73	36.0± 2.08
MMVAE	91.8± 1.46	307.9± 0.92
MoPoE-VAE	85.3± 2.51	312.8± 1.81
MMJSD	86.7± 1.94	287.0± 1.15
MMJSD-PoE	70.0± 1.17	141.4± 3.20
MVTCAE	41.6± 1.63	31.8± 0.51
CRMVAE	24.1± 0.71	37.3± 0.52

C RESULTS ON MNIST-SVHN-TEXT WITH STANDARD DEVIATIONS

Table 3 and Table 4 are Table 1 and Table 2 with their standard deviations of three trials explicitly added.

C.1 CONDITIONAL GENERATION ON POLYMNIST AND TRANSLATED-POLYMNIST

Figure 5 and Figure 6 visualize the conditionally generated images on the PolyMNIST and Translated-PolyMNIST datasets.

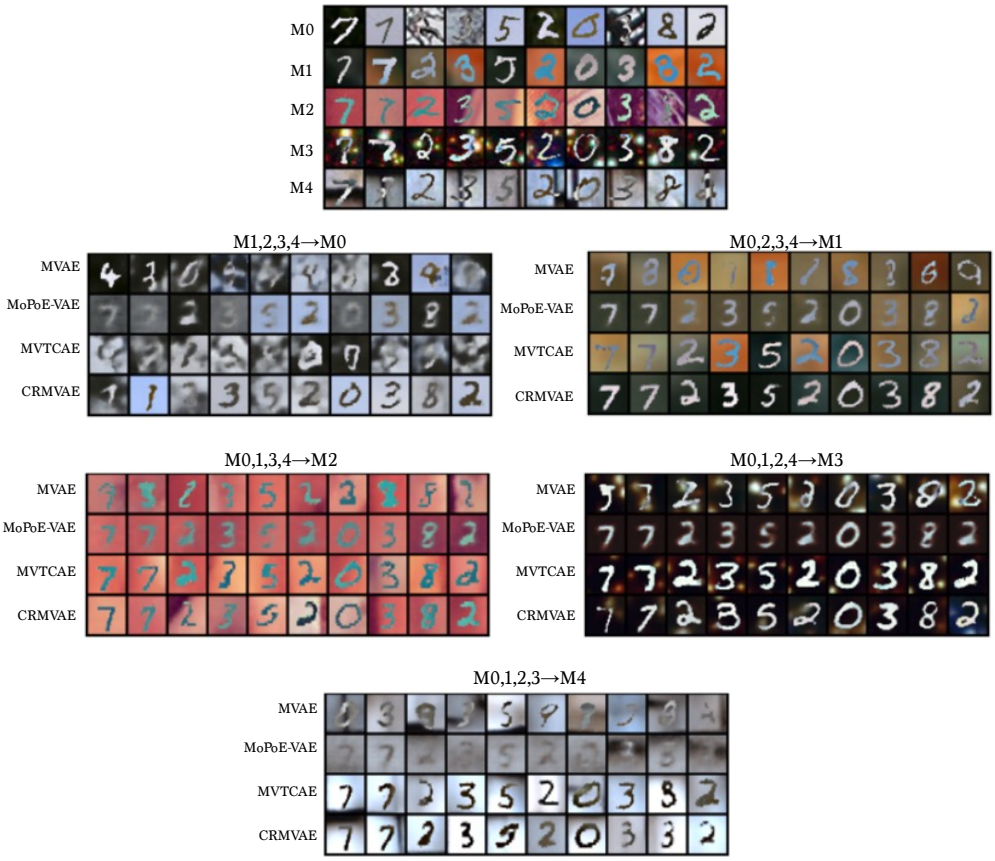

Figure 5: Cross-generation of PolyMNIST. The top panel shows examples of modalities M0 to M4 (we refer to the five different modalities as M0 to M4), and the bottom panels show the generation of a modality conditioned on remaining modalities by each model.

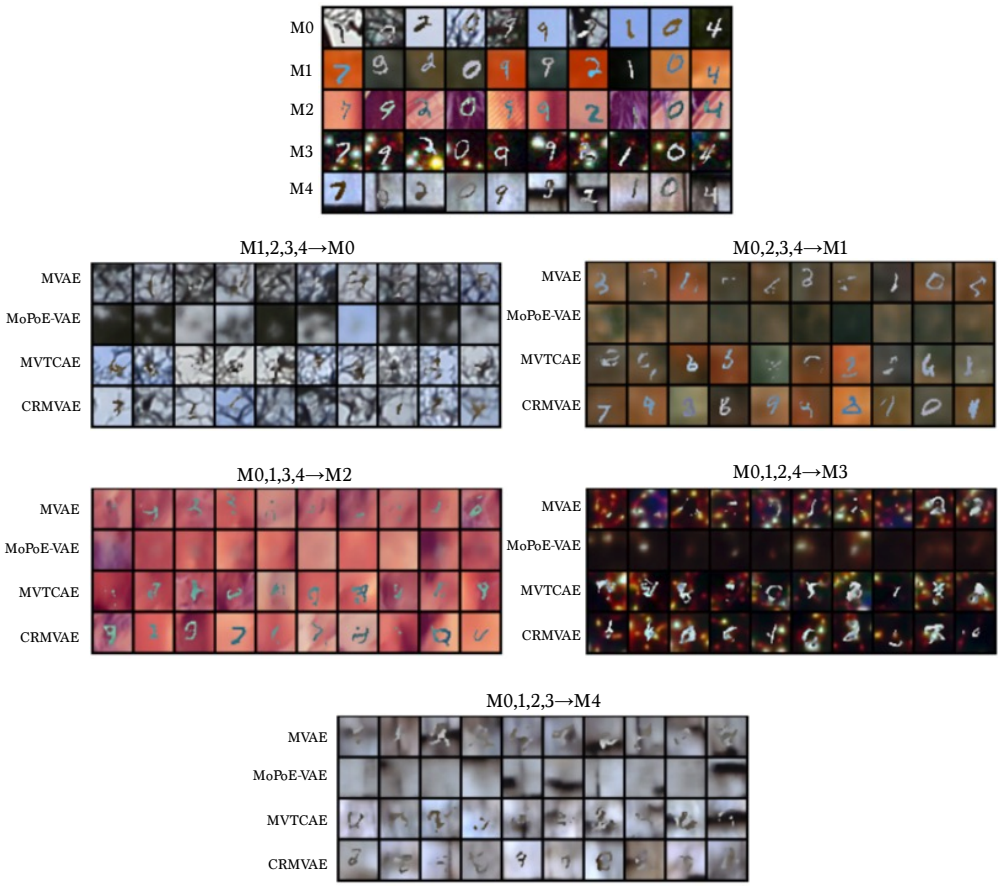

Figure 6: Cross-generation of Translated-PolyMNIST.