On the Limitations of Multimodal VAEs

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multimodal variational autoencoders (VAEs) have shown promise as efficient generative models for weakly-supervised data. Yet, despite their advantage of weak supervision, they exhibit a gap in generative quality compared to unimodal VAEs, which are completely unsupervised. In an attempt to explain this gap, we uncover a fundamental limitation that applies to a large family of mixture-based multimodal VAEs. We prove that the sub-sampling of modalities enforces an undesirable upper bound on the multimodal ELBO and thereby limits the generative quality of mixture-based multimodal VAEs. Empirically, we showcase the generative quality gap on both synthetic and real data and transparently present the tradeoffs between different variants of multimodal VAEs. We find that none of the existing approaches fulfills all desired criteria of an effective multimodal generative model when applied on more complex datasets than those used in previous benchmarks. In summary, we identify, formalize, and validate fundamental limitations of VAE-based approaches for modeling weakly-supervised data and discuss implications for real-world applications.

1 Introduction

In recent years, multimodal VAEs have shown great potential as efficient generative models for weakly-supervised data, such as pairs of images or paired images and captions. Previous works (Wu and Goodman, 2018; Shi et al., 2019; Sutter et al., 2020) demonstrate that multimodal VAEs leverage weak supervision to learn generalizable representations, useful for downstream tasks (Dorent et al., 2019; Minoura et al., 2021) and for the conditional generation of missing modalities (Lee and van der Schaar, 2021). However, despite the advantage of weak supervision, state-of-the-art multimodal VAEs consistently underperform when compared to simple unimodal VAEs in terms of generative quality.[1] This paradox serves as a starting point for our work, which aims to explain the observed lack of generative quality in terms of a fundamental limitation that underlies existing multimodal VAEs.

What is limiting the generative quality of multimodal VAEs? We find that the sub-sampling of modalities during training leads to a problem that affects all *mixture-based* multimodal VAEs—a family of models that subsumes the MMVAE (Shi et al., 2019), MoPoE-VAE (Sutter et al., 2021), and a special case of the MVAE (Wu and Goodman, 2018). We prove that modality sub-sampling enforces an upper bound on the multimodal ELBO and prevents a tight approximation of the joint distribution when there is modality-specific variation in the data. Our experiments demonstrate that modality sub-sampling can explain the gap in generative quality compared to unimodal VAEs and that the gap typically increases with each additional modality. Through extensive ablations on three different datasets, we validate the generative quality gap between unimodal and multimodal VAEs and transparently present the tradeoffs between different approaches.

[1] The lack of generative quality can even be recognized by visual inspection of the qualitative results from previous works; for instance, see the supplementaries of Sutter et al. (2021) or Shi et al. (2021).

Under review at ICBINB Workshop @ NeurIPS 2021. Do not distribute.

Table 1: Overview of mixture-based multimodal VAEs and their properties. Entries for generative quality and generative coherence denote properties that are observed empirically in previous works. The lightning symbol (↯) denotes properties for which our work presents contrary evidence.

Model	Definition of $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$	Modality sub-sampling	Generative quality	Generative coherence
MVAE (Wu and Goodman, 2018)	$\prod_{i=1}^{M} p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)$	✗	good	poor
MMVAE (Shi et al., 2019)	$\frac{1}{M} \sum_{i=1}^{M} p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)$	✓	limited	good ↯
MoPoE-VAE (Sutter et al., 2021)	$\frac{1}{\lvert\mathcal{P}(M)\rvert} \sum_{A \in \mathcal{P}(M)} \prod_{i \in A} p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)$	✓	limited	good ↯

Our results raise serious concerns about the utility of multimodal VAEs for real-world applications. We show that none of the existing approaches fulfills all desired criteria (Shi et al., 2019; Sutter et al., 2020) of an effective multimodal generative model when applied to slightly more complex datasets than used in previous benchmarks. In particular, we demonstrate that generative coherence (Shi et al., 2019) cannot be guaranteed for any of the existing approaches, if the information shared between modalities cannot be predicted in expectation across modalities. Our findings are particularly relevant for applications on datasets that exhibit a relatively high degree of modality-specific variation, which we argue to be a typical characteristic of many real-world datasets.

2 Related work

This work addresses multimodal VAEs, which belong to the broad class of multimodal generative models with encoder-decoder architectures (Baltrušaitis et al., 2019). Based on unimodal VAEs (Kingma and Welling, 2014), the first multimodal extensions (Suzuki et al., 2016; Hsu and Glass, 2018; Vedantam et al., 2018) implement separate inference networks for each subset of modalities, which quickly becomes intractable as the number of inference networks grows exponentially with the number of modalities. A more efficient alternative is provided by *mixture-based* multimodal VAEs—a family of models that subsumes the MMVAE (Shi et al., 2019), MoPoE-VAE (Sutter et al., 2021), and a special case of the MVAE (Wu and Goodman, 2018) without sub-sampling. The focus of our work lies on this family and the named special cases.

Table 1 provides an overview of the different variants of mixture-based multimodal VAEs and the properties that one can infer from empirical results in previous works (Shi et al., 2019, 2021; Sutter et al., 2021). Most importantly, there appears to be a tradeoff between generative quality and generative coherence (i.e., the ability to generate semantically related samples across modalities). Our work explains *why* the generative quality is worse for models that sub-sample modalities (Section 3) and shows that the joint distribution can be approximated better without sub-sampling, which manifests in an improved generative quality but not necessarily in a better generative coherence. Through systematic ablations, we validate the theoretical limitations and showcase the tradeoff between these two metrics (Section 4). Our experiments also reveal that generative coherence cannot be guaranteed on more complex datasets than those used in previous benchmarks, which challenges the utility of existing methods for real-world applications.

3 Modality sub-sampling limits the multimodal ELBO

Let $\boldsymbol{X} := \{X_1, \ldots, X_M\}$ be a set of random vectors describing M modalities and let $\boldsymbol{x} := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ be a sample from the joint distribution $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)$. For conciseness, denote subsets of modalities by subscripts; for example, $\boldsymbol{X}_{\{1,3\}}$ or $\boldsymbol{x}_{\{1,3\}}$ respectively for modalities 1 and 3.

Let us define the family of mixture-based multimodal VAEs, which subsumes the MMVAE, MoPoE-VAE, and a special case of the MVAE without sub-sampling (for details, please refer to Appendix A.1).

Definition 1. *The family of mixture-based multimodal VAEs is comprised of all models that optimize the multimodal ELBO (Wu and Goodman, 2018)*

$$\mathcal{L}_{ELBO}(\boldsymbol{x}; \theta, \phi) = \mathbb{E}_{p(\boldsymbol{x}) p_\theta(\boldsymbol{z} \mid \boldsymbol{x})}[\log q_\phi(\boldsymbol{x} \mid \boldsymbol{z})] - \beta \, \mathbb{E}_{p(\boldsymbol{x})}[D_{KL}(p_\theta(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, q(\boldsymbol{z}))] \tag{1}$$

using a mixture-based encoder $p_\theta^{\mathcal{S}}(\boldsymbol{z} \mid \boldsymbol{x}) = \sum_{A \in \mathcal{S}} \omega_A \, p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)$*, where* $\mathcal{S} = \{(A, \omega_A) \mid A \subseteq \{1, \ldots, M\},\ A \neq \emptyset,\ \omega_A \in [0,1]\}$ *denotes the model-specific set of subsets of modalities and corresponding mixture coefficients, which should sum to one.*

Theorem 1 states our main theoretical result. It describes a non-trivial limitation of multimodal VAEs, showing that mixture-based approaches that sub-sample modalities cannot provide a tight estimate of the joint evidence when there is modality-specific variation in the data. This limitation conflicts with the goal of modeling real-world multimodal data, which typically exhibits a high degree of modality-specific variation.

Theorem 1. *Any mixture-based multimodal VAE (Definition 1) approximates the expected log-evidence* $\mathbb{E}_{p(\boldsymbol{x})}[\log p(\boldsymbol{x})]$ *up to an irreducible error* $\Delta(\boldsymbol{X}, \mathcal{S})$ *that depends on the model-specific mixture distribution* $\mathcal{S}$*, as well as on the amount of modality-specific information in* $\boldsymbol{X}$*:*

$$\Delta(\boldsymbol{X}, \mathcal{S}) = \sum_{A \in \mathcal{S}} \omega_A \, H(\boldsymbol{X}_{\{1,\ldots,M\} \setminus A} \mid \boldsymbol{X}_A) \,. \tag{2}$$

From Theorem 1, we can derive two corollaries, which apply to relevant special cases.

Corollary 1. *Without modality sub-sampling, there is no irreducible error.*

Corollary 1 applies to the MVAE without sub-sampling and suggests that this model should yield a tighter approximation of the joint distribution and hence a better generative quality than mixture-based multimodal VAEs that sub-sample modalities.

Corollary 2. *For the MMVAE and MoPoE-VAE, the irreducible error increases with each additional modality, if the new modality is sufficiently diverse.*

Corollary 2 predicts an increased approximation error (and hence, a decline of generative quality) when we increase the number of modalities for the MMVAE and MoPoE-VAE. The notion of *diversity* requires a more formal treatment of the underlying information-theoretic quantities, which we defer to Appendix A.7. Intuitively, a new modality is sufficiently diverse, if it does not add too much redundant information with respect to the existing modalities.

In Appendix A we prove and further discuss the above theoretical results. In the following, we present empirical support for the limitations described by Theorem 1 and its corollaries.

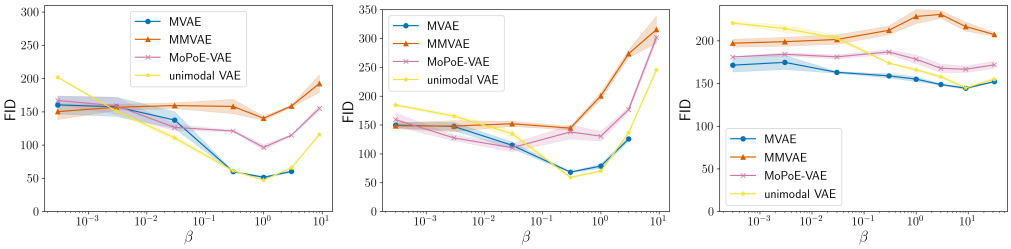
4 Experiments

Figure 1 presents the three considered datasets. PolyMNIST (Sutter et al., 2021) is a simple, synthetic dataset that allows us to conduct systematic ablations. Translated-PolyMNIST is a new dataset that introduces a small tweak—the downscaling and random translation of digits—to demonstrate the limitations of existing methods when shared information cannot be predicted in expectation across modalities. Caltech Birds (CUB; Wah et al., 2011; Shi et al., 2019) allows us to validate the limitations on a more realistic dataset. Please note that we use CUB with *real images* and not the simplified version based on ResNet-features that was used in Shi et al. (2019) and Shi et al. (2021).

Figure 2 compares the generative quality, measured by FID (Heusel et al., 2017), of unimodal VAEs and mixture-based multimodal VAEs. As predicted by Theorem 1, for multimodal VAEs that sub-sample modalities (i.e., MMVAE and MoPoE-VAE) we observe a pronounced generative quality

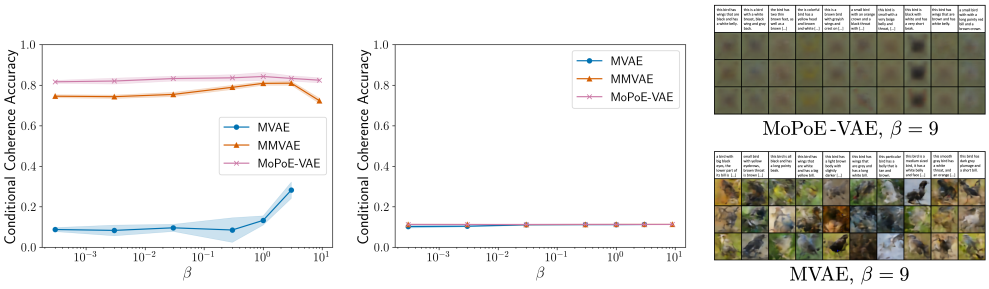
(a) PolyMNIST (b) Translated-PolyMNIST (c) Caltech Birds (CUB)

Figure 1: The three considered datasets. Each column shows a sample from the respective dataset. The two PolyMNIST datasets are conceptually similar in that the digit label is shared between five synthetic modalities. The Caltech Birds (CUB) dataset provides a more realistic application for which there is no annotation on what is shared between paired images and captions.



(a) PolyMNIST (b) Translated-PolyMNIST (c) Caltech Birds (CUB)

Figure 2: Generative quality for one modality over a range of β values. Points denote FID averaged over three seeds and bands show one standard deviation respectively. Due to numerical instabilities, MVAE could not be trained with larger β values. Figure 7 shows similar results for other modalities.



(a) PolyMNIST (b) Translated-PolyMNIST (c) Caltech Birds (CUB)

Figure 3: Generative coherence for the conditional generation across modalities. For PolyMNIST, we plot the average leave-one-out coherence (described in Appendix B.1). For CUB, we show qualitative results for the conditional generation of images given captions. Best viewed zoomed and in color.

gap when we compare the best models. In contrast, the MVAE (without sub-sampling) reaches the generative quality of unimodal VAEs, which is consistent with our theoretical result from Corollary 1. Further, Figure 4 in Appendix B supports our results from Corollary 2, showing that the generative quality of the MMVAE and MoPoE-VAE deteriorates when we increase the number of modalities.

In terms of generative coherence—another desired criterion (Shi et al., 2019) for multimodal generative models—Figure 3 reveals that the positive results from previous work (Shi et al., 2019; Sutter et al., 2021) do not apply to more complex datasets. On PolyMNIST, we successfully replicate the coherence results from Sutter et al. (2021) for a range of β values. However, on Translated-PolyMNIST, the stark decline of all models makes it evident that coherence cannot be guaranteed when shared information cannot be predicted in expectation across modalities. On CUB, we resort to a qualitative evaluation of coherence, because there is no ground truth annotation of shared factors (like for many real-world datasets). The qualitative results confirm that none of the existing approaches generates images that are both of sufficiently high quality and coherent with respect to the given caption. Overall, the negative results on Translated-PolyMNIST and CUB showcase the limitations of existing approaches when applied to more complex datasets than those used in previous benchmarks.

5 Conclusion

In this work, we have identified, formalized, and demonstrated several limitations of multimodal VAEs. Across different datasets, we revealed a significant gap in generative quality between unimodal and mixture-based multimodal VAEs that sub-sample modalities. We proved that modality sub-sampling can enforce an undesirable upper bound on the multimodal ELBO and can therefore limit the generative quality of the respective models. Finally, we studied two failure cases—Translated-PolyMNIST and CUB—that demonstrate the serious limitations of multimodal VAEs when applied to more complex datasets than those used in previous benchmarks. We urge future research to design more realistic benchmarks and to consider the tradeoff between generative quality and coherence.

References

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal Machine Learning: A survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Dorent, R., Joutard, S., Modat, M., Ourselin, S., and Vercauteren, T. (2019). Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *Medical Image Computing and Computer Assisted Intervention*, volume 11765, pages 74–82. Springer.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*.

Hsu, W.-N. and Glass, J. (2018). Disentangling by Partitioning: A Representation Learning Framework for Multimodal Sensory Data. *arXiv preprint arXiv:1805.11264*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *International Conference on Learning Representations*.

Lee, C. and van der Schaar, M. (2021). A variational information bottleneck approach to multi-omics data integration. In Banerjee, A. and Fukumizu, K., editors, *International Conference on Artificial Intelligence and Statistics*.

Minoura, K., Abe, K., Nam, H., Nishikawa, H., and Shimamura, T. (2021). A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Reports Methods*, page 100071.

Shi, Y., Paige, B., Torr, P., and N, S. (2021). Relating by contrasting: A data-efficient framework for multimodal generative models. In *International Conference on Learning Representations*.

Shi, Y., Siddharth, N., Paige, B., and Torr, P. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems*.

Sutter, T. M., Daunhawer, I., and Vogt, J. E. (2020). Multimodal generative learning utilizing jensen-shannon-divergence. In *Advances in Neural Information Processing Systems*.

Sutter, T. M., Daunhawer, I., and Vogt, J. E. (2021). Generalized multimodal ELBO. In *International Conference on Learning Representations*.

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

Vedantam, R., Fischer, I., Huang, J., and Murphy, K. (2018). Generative models of visually grounded imagination. In *International Conference on Learning Representations*.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wu, M. and Goodman, N. (2018). Multimodal Generative Models for Scalable Weakly-Supervised Learning. In *Advances in Neural Information Processing Systems*.

A Proofs

A.1 Definition 1 subsumes all considered multimodal VAEs

To show: Definition 1 subsumes the MMVAE, MoPoE-VAE, and a special case of the MVAE without sub-sampling.

Proof. It was already shown by previous work that all of the considered multimodal VAEs maximize the multimodal ELBO (Definition 1). Hence, it suffices to show that the mixture-based encoder $p_\theta^{\mathcal{S}}$ is equal to the encoder of the respective model, given a certain instance of $\mathcal{S}$.

Since all of the considered models use uniform mixture coefficients for all subsets, we can write

$$p_\theta^{\mathcal{S}}(\boldsymbol{z} \mid \boldsymbol{x}) = \sum_{A \in \mathcal{S}} \omega_A \, p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A) = \frac{1}{|\mathcal{S}|} \sum_{A \in \mathcal{S}} p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A) \; . \tag{3}$$

For the following instances of $\mathcal{S}$, we recover the encoders of the respective models (listed in Table 1):

$$p_\theta^{\mathcal{S}}(\boldsymbol{z} \mid \boldsymbol{x}) = \left\{ \begin{array}{ll} p_\theta(\boldsymbol{z} \mid \boldsymbol{x}), & \text{if } \mathcal{S} \text{ is comprised of the set } \{1, \ldots, M\} \\ \frac{1}{M} \sum_{i=1}^{M} p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i), & \text{if } \mathcal{S} \text{ is comprised of subsets } \{1\}, \ldots, \{M\} \\ \frac{1}{|\mathcal{P}(M)|} \sum_{A \in \mathcal{P}(M)} p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A), & \text{if } \mathcal{S} \text{ is comprised of the powerset } \mathcal{P}(M) \setminus \{\emptyset\} \end{array} \right\}$$

All of the above instances of $\mathcal{S}$ are consistent with the definition of $\mathcal{S}$ in Definition 1, since each set is comprised of non-empty subsets of modalities and all mixture-coefficients sum to one.

Note that for the definition of mixture-based multimodal VAEs it is possible but not necessary to further factorize $p(\boldsymbol{z} \mid \boldsymbol{x})$ or $p(\boldsymbol{z} \mid \boldsymbol{x}_A)$ as a product of experts (PoE). This implies that even if the data satisfies the assumptions of the PoE, the limitations from Theorem 1 and its corollaries apply to the MVAE and MoPoE-VAE, both of which use the PoE-factorization. □

A.2 All mixture-based multimodal VAEs maximize the multimodal ELBO

Lemma 1. *All mixture-based multimodal VAEs (Definition 1) maximize the following lower bound $\mathcal{L}_{ELBO}^{\mathcal{S}}$ on the multimodal ELBO $\mathcal{L}_{ELBO}$ from Equation* (1)*:*

$$\mathcal{L}_{ELBO}^{\mathcal{S}}(\boldsymbol{x}; \theta, \phi) = \sum_{A \in \mathcal{S}} \omega_A \left\{ \mathbb{E}_{p(\boldsymbol{x}) p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)} [\log q_\phi(\boldsymbol{x} \mid \boldsymbol{z})] - \mathbb{E}_{p(\boldsymbol{x})} \left[D_{KL} \left(p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A) \,||\, q(\boldsymbol{z}) \right) \right] \right\} \tag{4}$$

Proof. Recall from Equation (1) the multimodal ELBO

$$\mathcal{L}_{\text{ELBO}}(\boldsymbol{x}; \theta, \phi) = \mathbb{E}_{p(\boldsymbol{x}) p_\theta(\boldsymbol{z} \mid \boldsymbol{x})} [\log q_\phi(\boldsymbol{x} \mid \boldsymbol{z})] - \mathbb{E}_{p(\boldsymbol{x})} [D_{\text{KL}}(p_\theta(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, q(\boldsymbol{z}))] \; . \tag{5}$$

For the encoder $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$, plug in the mixture-based encoder $p_\theta^{\mathcal{S}}(\boldsymbol{z} \mid \boldsymbol{x}) = \sum_{A \in \mathcal{S}} \omega_A \, p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)$ from Definition 1 and re-write as follows:

$$\begin{aligned}
& \mathbb{E}_{p(\boldsymbol{x}) p_\theta^{\mathcal{S}}(\boldsymbol{z} \mid \boldsymbol{x})} [\log q_\phi(\boldsymbol{x} \mid \boldsymbol{z})] - \mathbb{E}_{p(\boldsymbol{x})} [D_{\text{KL}}(p_\theta^{\mathcal{S}}(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, q(\boldsymbol{z}))] \\
&= \mathbb{E}_{p(\boldsymbol{x}) \sum_{A \in \mathcal{S}} \omega_A \, p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)} [\log q_\phi(\boldsymbol{x} \mid \boldsymbol{z})] - \mathbb{E}_{p(\boldsymbol{x}) \sum_{A \in \mathcal{S}} \omega_A \, p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)} [\log p_\theta^{\mathcal{S}}(\boldsymbol{z} \mid \boldsymbol{x}) - \log q(\boldsymbol{z})] \\
&= \sum_{A \in \mathcal{S}} \omega_A \left\{ \mathbb{E}_{p(\boldsymbol{x}) p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)} [\log q_\phi(\boldsymbol{x} \mid \boldsymbol{z})] - \mathbb{E}_{p(\boldsymbol{x}) p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)} [\log p_\theta^{\mathcal{S}}(\boldsymbol{z} \mid \boldsymbol{x})] + \mathbb{E}_{p(\boldsymbol{x}) p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)} [\log q(\boldsymbol{z})] \right\} \\
&= \sum_{A \in \mathcal{S}} \omega_A \left\{ \mathbb{E}_{p(\boldsymbol{x}) p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)} [\log q_\phi(\boldsymbol{x} \mid \boldsymbol{z})] + \mathbb{E}_{p(\boldsymbol{x})} \left[CE \left(p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A), p_\theta^{\mathcal{S}}(\boldsymbol{z} \mid \boldsymbol{x}) \right) \right] - \mathbb{E}_{p(\boldsymbol{x})} \left[CE \left(p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A), q(\boldsymbol{z}) \right) \right] \right\} \\
&= \sum_{A \in \mathcal{S}} \omega_A \left\{ \mathbb{E}_{p(\boldsymbol{x}) p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)} [\log q_\phi(\boldsymbol{x} \mid \boldsymbol{z})] + \mathbb{E}_{p(\boldsymbol{x})} \left[D_{\text{KL}} \left(p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A) \,||\, p_\theta^{\mathcal{S}}(\boldsymbol{z} \mid \boldsymbol{x}) \right) \right] - \mathbb{E}_{p(\boldsymbol{x})} \left[D_{\text{KL}} \left(p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A) \,||\, q(\boldsymbol{z}) \right) \right] \right\} \\
&\geq \sum_{A \in \mathcal{S}} \omega_A \left\{ \mathbb{E}_{p(\boldsymbol{x}) p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)} [\log q_\phi(\boldsymbol{x} \mid \boldsymbol{z})] - \mathbb{E}_{p(\boldsymbol{x})} \left[D_{\text{KL}} \left(p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A) \,||\, q(\boldsymbol{z}) \right) \right] \right\} \\
&=: \mathcal{L}_{\text{ELBO}}^{\mathcal{S}}(\boldsymbol{x}; \theta, \phi)
\end{aligned}$$

where $CE(p, q)$ denotes the cross-entropy between distributions p and q, which can be decomposed into $H(p) + D_{\text{KL}}(p||q)$. This concludes the proof that any mixture-based multimodal VAE maximizes a lower bound on the multimodal ELBO from Equation (1). □

A.3 The mixture-based multimodal ELBO is special case of the VIB objective

Lemma 2. *All mixture-based multimodal VAEs (Definition 1) approximate the expected log-evidence by optimizing the variational information bottleneck (VIB) objective*

$$\min_{\psi} \sum_{A \in \mathcal{S}} \omega_A \left\{ H_\psi(\boldsymbol{X} \mid Z_A) + I_\psi(\boldsymbol{X}; Z_A) \right\} , \tag{6}$$

where the encoding $Z_A = f_\psi(\boldsymbol{X}_A)$ is a function of a subset $\boldsymbol{X}_A$, the terms $H_\psi(\boldsymbol{X} \mid Z_A)$ and $I_\psi(\boldsymbol{X}; Z_A)$ denote variational upper bounds of $H(\boldsymbol{X} \mid Z_A)$ and $I(\boldsymbol{X}; Z_A)$ respectively, and ψ summarizes the parameters of these variational estimators.

Proof. Let us start from the ELBO-objective optimized by all mixture-based multimodal VAEs and show that the ELBO can be seen as a special case of the more general VIB objective.

The ELBO optimized by all mixture-based multimodal VAEs (derived in Appendix A.2) is defined as

$$\mathcal{L}^{\mathcal{S}}_{\text{ELBO}}(\boldsymbol{x}; \theta, \phi) = \sum_{A \in \mathcal{S}} \omega_A \Big\{ \underbrace{\mathbb{E}_{p(\boldsymbol{x}) p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)} [\log q_\phi(\boldsymbol{x} \mid \boldsymbol{z})]}_{(i)} - \underbrace{\mathbb{E}_{p(\boldsymbol{x})} \left[D_{\text{KL}} \left(p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A) \,||\, q(\boldsymbol{z}) \right) \right]}_{(ii)} \Big\} .$$

Each term within the sum is comprised of two terms: (i) the log-likelihood estimation based on a variational decoder $q_\phi(\boldsymbol{x} \mid \boldsymbol{z})$; (ii) the regularization of the stochastic encoder $p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)$ with respect to a variational prior $q(\boldsymbol{z})$. The encoding sampled from $p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)$ can be viewed as a function $Z_A = f_\theta(\boldsymbol{X}_A)$ of a subset of modalities.

To see the relation to the underlying information terms $H(\boldsymbol{X} \mid Z_A)$ and $I(\boldsymbol{X}; Z_A)$, let us "undo" the variational approximation for (i) and (ii) by re-introducing the unobserved ground truth decoder $p(\boldsymbol{x} \mid \boldsymbol{z})$ and the ground truth prior $p(\boldsymbol{z})$.

For (i) we have

$$\begin{aligned}
\mathbb{E}_{p(\boldsymbol{x}) p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)} [\log q_\phi(\boldsymbol{x} \mid \boldsymbol{z})] &\leq \mathbb{E}_{p(\boldsymbol{x}) p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)} [\log q_\phi(\boldsymbol{x} \mid \boldsymbol{z})] + \mathbb{E}_{p(\boldsymbol{z})} [D_{\text{KL}}(p(\boldsymbol{x} \mid \boldsymbol{z}) \,||\, q_\phi(\boldsymbol{x} \mid \boldsymbol{z}))] \\
&= \mathbb{E}_{p(\boldsymbol{x}) p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A)} [\log p(\boldsymbol{x} \mid \boldsymbol{z})] \\
&= -H(\boldsymbol{X} \mid Z_A)
\end{aligned}$$

For (ii), we have

$$\begin{aligned}
\mathbb{E}_{p(\boldsymbol{x})} [D_{\text{KL}}(p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A) \,||\, q(\boldsymbol{z}))] &\geq \mathbb{E}_{p(\boldsymbol{x})} [D_{\text{KL}}(p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A) \,||\, q(\boldsymbol{z}))] - D_{\text{KL}}(p(\boldsymbol{z}) \,||\, q(\boldsymbol{z})) \\
&= \mathbb{E}_{p(\boldsymbol{x})} [D_{\text{KL}}(p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_A) \,||\, p(\boldsymbol{z}))] \\
&= I(\boldsymbol{X}; Z_A)
\end{aligned}$$

Since $\mathcal{L}^{\mathcal{S}}_{\text{ELBO}}(\boldsymbol{x}; \theta, \phi)$ is being maximized, (i) is being maximized, while (ii) is being minimized. The maximization of (i) is equal to the minimization of a variational upper bound on $H(\boldsymbol{X} \mid Z_A)$. Similarly, the minimization of (ii) is equal to the minimization of a variational upper bound on $I(\boldsymbol{X}; Z_A)$. Hence, we have established that $\mathcal{L}^{\mathcal{S}}_{\text{ELBO}}(\boldsymbol{x}; \theta, \phi)$ is a special case of the more general VIB objective (Equation (6)) where the information terms are estimated with a mixture-based multimodal VAE that is parameterized by $\psi = \{\theta, \phi\}$.

□

A.4 Proof of Theorem 1

Theorem 1. *Any mixture-based multimodal VAE (Definition 1) approximates the expected log-evidence $\mathbb{E}_{p(\boldsymbol{x})}[\log p(\boldsymbol{x})]$ up to an irreducible error $\Delta(\boldsymbol{X}, \mathcal{S})$ that depends on the model-specific mixture distribution $\mathcal{S}$, as well as on the amount of modality-specific information in $\boldsymbol{X}$:*

$$\Delta(\boldsymbol{X}, \mathcal{S}) = \sum_{A \in \mathcal{S}} \omega_A \, H(\boldsymbol{X}_{\{1,\ldots,M\} \setminus A} \mid \boldsymbol{X}_A) \, . \tag{2}$$

Proof. Lemma 2 shows that all mixture-based multimodal VAEs approximate the expected log-evidence via the more general VIB-objective

$$\min_{\psi} \sum_{A \in \mathcal{S}} \omega_A \left\{ H_\psi(\boldsymbol{X} \mid Z_A) + I_\psi(\boldsymbol{X}; Z_A) \right\} \tag{7}$$

where the encoding $Z_A = f_\psi(\boldsymbol{X}_A)$ is a function of a subset $\boldsymbol{X}_A$. The fact that Z_A is a function of a *subset* $\boldsymbol{X}_A \subseteq \boldsymbol{X}$, allows us to use the following decomposition (for details, see Appendix A.5):

$$H(\boldsymbol{X} \mid Z_A) + I(\boldsymbol{X}; Z_A) = H(\boldsymbol{X}_{\{1,\ldots,M\} \setminus A} \mid \boldsymbol{X}_A) + H(\boldsymbol{X}_A \mid Z_A) + I(\boldsymbol{X}_A; Z_A) \, . \tag{8}$$

Hence, for the optimization of the VIB-objective (Equation (7)), we have

$$\underbrace{\sum_{A \in \mathcal{S}} \omega_A \left\{ H(\boldsymbol{X}_{\{1,\ldots,M\} \setminus A} \mid \boldsymbol{X}_A) \right\}}_{\Delta(\boldsymbol{X}, \mathcal{S})} + \min_{\psi} \sum_{A \in \mathcal{S}} \omega_A \left\{ H_\psi(\boldsymbol{X}_A \mid Z_A) + I_\psi(\boldsymbol{X}_A; Z_A) \right\} \tag{9}$$

where the first term is data-dependent (i.e, independent of the optimization) and therefore represents an irreducible error $\Delta(\boldsymbol{X}, \mathcal{S})$ for the estimation of the expected log-evidence. $\Delta(\boldsymbol{X}, \mathcal{S})$ depends on the definition of the mixture distribution and it is larger than zero, if $\mathcal{S}$ contains strict subsets of modalities with non-negative mixture coefficients. Equation (9) applies to the multimodal ELBO, since it is a special case of the more general VIB-objective (see Lemma 2).

□

A.5 Explanation of Equation (8)

To show: If $Z_A = f(\boldsymbol{X}_A)$, for some function f, then the following equality holds:

$$H(\boldsymbol{X} \mid Z_A) + \beta I(\boldsymbol{X}; Z_A) = H(\boldsymbol{X}_{\{1,\ldots,M\} \setminus A} \mid \boldsymbol{X}_A) + H(\boldsymbol{X}_A \mid Z_A) + \beta I(\boldsymbol{X}_A; Z_A) \, . \tag{10}$$

Proof. When Z_A is a function of a subset $\boldsymbol{X}_A \subseteq \boldsymbol{X}$, we have the Markov chain $Z_A \leftarrow \boldsymbol{X}_A - \boldsymbol{X}_{\{1,\ldots,M\} \setminus A}$, since Z_A is a function of the (observed) subset of modalities and depends on the remaining (unobserved) modalities only through $\boldsymbol{X}_A$.

For the first term, we have

$$\begin{aligned}
H(\boldsymbol{X} \mid Z_A) &= H(\boldsymbol{X} \mid Z_A, \boldsymbol{X}_A) + I(\boldsymbol{X}; \boldsymbol{X}_A \mid Z_A) && (11)\\
&= H(\boldsymbol{X} \mid \boldsymbol{X}_A) + I(\boldsymbol{X}; \boldsymbol{X}_A \mid Z_A) && (12)\\
&= H(\boldsymbol{X}_{\{1,\ldots,M\} \setminus A} \mid \boldsymbol{X}_A) + I(\boldsymbol{X}; \boldsymbol{X}_A \mid Z_A) && (13)\\
&= H(\boldsymbol{X}_{\{1,\ldots,M\} \setminus A} \mid \boldsymbol{X}_A) + H(\boldsymbol{X}_A \mid Z_A) && (14)
\end{aligned}$$

Equation (11) applies the definition of the conditional mutual information. Equation (12) is based on the conditional independence $\boldsymbol{X} \perp\!\!\!\perp Z_A \mid \boldsymbol{X}_A$ implied by the Markov chain. Equation (13) removes the "known" information that we condition on. Finally, Equation (14) follows from $\boldsymbol{X}_A \subseteq \boldsymbol{X}$, which implies that $I(\boldsymbol{X}; \boldsymbol{X}_A) = H(\boldsymbol{X}_A)$ and $I(\boldsymbol{X}; \boldsymbol{X}_A \mid Z_A) = H(\boldsymbol{X}_A \mid Z_A)$.

For the second term, $I(\boldsymbol{X}; Z_A) = I(\boldsymbol{X}_A; Z_A)$ follows immediately from the Markov chain, since $\boldsymbol{X}$ is the concatenation of $\boldsymbol{X}_A$ and $\boldsymbol{X}_{\{1,\ldots,M\} \setminus A}$ and Z_A depends on $\boldsymbol{X}_{\{1,\ldots,M\} \setminus A}$ only through $\boldsymbol{X}_A$. □

A.6 Proof of Corollary 1

Corollary 1. *Without modality sub-sampling, there is no irreducible error.*

Proof. Without modality sub-sampling, $\mathcal{S}$ consists of a single subset $\{1, \dots, M\}$, $\boldsymbol{X}_A = \boldsymbol{X}$, and $\boldsymbol{X}_{\{1,\dots,M\}\setminus A} = \emptyset$. It follows that $\Delta(\boldsymbol{X}, \mathcal{S}) = H(\boldsymbol{X}_{\{1,\dots,M\}\setminus A} \mid \boldsymbol{X}_A) = H(\emptyset \mid \boldsymbol{X}) = 0$, since the conditional entropy of the empty set is zero. □

A.7 Proof of Corollary 2

Corollary 2. *For the MMVAE and MoPoE-VAE, the irreducible error increases given an additional modality X_{M+1}, if the new modality is sufficiently diverse in the following sense:*

$$\frac{1}{|\mathcal{S}^+||\mathcal{S}|} \sum_{A \in \mathcal{S}} H(\boldsymbol{X}_A | X_{M+1}) + \frac{1}{|\mathcal{S}^+|} \sum_{A \in \mathcal{S}} H(X_{M+1} | \boldsymbol{X}) > \left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|} \right) \sum_{A \in \mathcal{S}} I(\boldsymbol{X}_{\{1,\dots,M\}\setminus A}; X_{M+1} | \boldsymbol{X}_A)$$

where $\mathcal{S}$ denotes the model-specific mixture distribution over the set of subsets of modalities given modalities $X_1, \dots, X_M$ and $\mathcal{S}^+$ is the respective mixture distribution over the extended set of subsets of modalities given $X_1, \dots, X_{M+1}$.

Proof. Let X_{M+1} be the new modality, let $\boldsymbol{X}^+ := \{X_1, \dots, X_{M+1}\}$ denote the extended set of modalities, and let $\mathcal{S}^+$ denote the new mixture distribution over subsets given $\boldsymbol{X}^+$. Note that all subsets from $\mathcal{S}$ are still contained in $\mathcal{S}^+$, but that $\mathcal{S}^+$ contains new subsets in addition to those in $\mathcal{S}$. Further, due to the re-weighting of mixture coefficients, $\mathcal{S}^+$ can have different mixture coefficients for the subsets it shares with $\mathcal{S}$. Let us denote by $S^- := \{(A, \omega_A^+) \in \mathcal{S}^+ : A \notin \mathcal{S}\}$ the set of new subsets and let ω_A^+ denote the new mixture coefficients where typically $\omega_A \neq \omega_A^+$, due to the re-weighting.

We are interested in the change of the irreducible error, when we add modality X_{M+1}:

$$\Delta(\boldsymbol{X}^+, \mathcal{S}^+) - \Delta(\boldsymbol{X}, \mathcal{S}) = \sum_{B \in \mathcal{S}^+} \omega_B^+ \, H(\boldsymbol{X}_{\{1,\dots,M+1\}\setminus B} \mid \boldsymbol{X}_B) - \sum_{A \in \mathcal{S}} \omega_A \, H(\boldsymbol{X}_{\{1,\dots,M\}\setminus A} \mid \boldsymbol{X}_A) \; . \quad (15)$$

Let us re-write the right hand side in terms of subsets that are contained in both $\mathcal{S}$ and $\mathcal{S}^+$ and subsets that are only contained in $\mathcal{S}^+$. For this, we decompose the first term as follows

$$\sum_{B \in \mathcal{S}^+} \omega_B^+ \, H(\boldsymbol{X}_{\{1,\dots,M+1\}\setminus B} \mid \boldsymbol{X}_B) \quad (16)$$

$$= \sum_{A \in \mathcal{S}} \omega_A^+ \, H(\boldsymbol{X}_{\{1,\dots,M+1\}\setminus A} \mid \boldsymbol{X}_A) + \sum_{B \in \mathcal{S}^-} \omega_B^+ \, H(\boldsymbol{X}_{\{1,\dots,M+1\}\setminus B} \mid \boldsymbol{X}_B) \quad (17)$$

$$= \sum_{A \in \mathcal{S}} \omega_A^+ \, H(\boldsymbol{X}_{\{1,\dots,M\}\setminus A} \mid \boldsymbol{X}_A) + \sum_{A \in \mathcal{S}} \omega_A^+ \, H(X_{M+1} \mid \boldsymbol{X}) + \sum_{B \in \mathcal{S}^-} \omega_B^+ \, H(\boldsymbol{X}_{\{1,\dots,M+1\}\setminus B} \mid \boldsymbol{X}_B) \quad (18)$$

where the last equation follows from

$$H(\boldsymbol{X}_{\{1,\dots,M+1\}\setminus A} \mid \boldsymbol{X}_A) = H(\boldsymbol{X}_{\{1,\dots,M\}\setminus A} \mid \boldsymbol{X}_A) + H(X_{M+1} \mid \boldsymbol{X}_A, \boldsymbol{X}_{\{1,\dots,M\}\setminus A}) \quad (19)$$

$$= H(\boldsymbol{X}_{\{1,\dots,M\}\setminus A} \mid \boldsymbol{X}_A) + H(X_{M+1} \mid \boldsymbol{X}) \; . \quad (20)$$

We can use the decomposition from Equation (18) to re-write the right hand side of Equation (15) by collecting the corresponding terms for $H(\boldsymbol{X}_{\{1,\dots,M\}\setminus A} \mid \boldsymbol{X}_A)$:

$$\sum_{A \in \mathcal{S}} (\omega_A^+ - \omega_A) H(\boldsymbol{X}_{\{1,\dots,M\}\setminus A} \mid \boldsymbol{X}_A) + \sum_{A \in \mathcal{S}} \omega_A^+ \, H(X_{M+1} \mid \boldsymbol{X}) + \sum_{B \in \mathcal{S}^-} \omega_B^+ \, H(\boldsymbol{X}_{\{1,\dots,M+1\}\setminus B} \mid \boldsymbol{X}_B) \; . \quad (21)$$

Notice that in Equation (21) only the first term can be negative, due to the re-weighting of mixture coefficients for terms that do not contain X_{M+1}. Hence, in the general case, the irreducible error can only decrease, if the mixture coefficients change in such a way that the first term in Equation (21) dominates the other two terms.

For the relevant special case of uniform mixture weights, which applies to both the MMVAE and MoPoE-VAE, we can further decompose Equation (21) into (i) information shared between $\boldsymbol{X}$ and X_{M+1}, and (ii) information that is specific to $\boldsymbol{X}$ or X_{M+1}.

Using $\omega_A = \frac{1}{|\mathcal{S}|}$ and $\omega_A^+ = \frac{1}{|\mathcal{S}^+|}$ for all subsets, we can factor out the uniform mixture coefficients and re-write Equation (21) as follows:

$$\left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|}\right) \sum_{A \in \mathcal{S}} H(\boldsymbol{X}_{\{1,\dots,M\}\setminus A} \mid \boldsymbol{X}_A) + \frac{1}{|\mathcal{S}^+|} \sum_{A \in \mathcal{S}} H(X_{M+1} \mid \boldsymbol{X}) + \frac{1}{|\mathcal{S}^+|} \sum_{B \in \mathcal{S}^-} H(\boldsymbol{X}_{\{1,\dots,M+1\}\setminus B} \mid \boldsymbol{X}_B) \tag{22}$$

where the second term already denotes information that is specific to X_{M+1}. Hence, let us decompose the first and last terms corresponding to (i) and (ii).

For the first term from Equation (22), we have

$$\left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|}\right) \sum_{A \in \mathcal{S}} H(\boldsymbol{X}_{\{1,\dots,M\}\setminus A} \mid \boldsymbol{X}_A) \tag{23}$$

$$= \left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|}\right) \sum_{A \in \mathcal{S}} \Big\{ H(\boldsymbol{X}_{\{1,\dots,M\}\setminus A} \mid \boldsymbol{X}_A, X_{M+1}) + I(\boldsymbol{X}_{\{1,\dots,M\}\setminus A}; X_{M+1} \mid \boldsymbol{X}_A) \Big\} . \tag{24}$$

For the last term from Equation (22), we have

$$\frac{1}{|\mathcal{S}^+|} \sum_{B \in \mathcal{S}^-} H(\boldsymbol{X}_{\{1,\dots,M+1\}\setminus B} \mid \boldsymbol{X}_B) \tag{25}$$

$$= \frac{1}{|\mathcal{S}^+|} \Big\{ H(\boldsymbol{X} \mid X_{M+1}) + \sum_{A \in \mathcal{S}} \mathbf{1}_{\{(A \cup \{M+1\}) \in \mathcal{S}^-\}} H(\boldsymbol{X}_{\{1,\dots,M\}\setminus A} \mid \boldsymbol{X}_A, X_{M+1}) \Big\} \tag{26}$$

where we can further decompose

$$\frac{1}{|\mathcal{S}^+|} H(\boldsymbol{X} \mid X_{M+1}) = \frac{1}{|\mathcal{S}^+|} \Big\{ H(\boldsymbol{X} \mid \boldsymbol{X}_A, X_{M+1}) + I(\boldsymbol{X}; \boldsymbol{X}_A \mid X_{M+1}) \Big\} \tag{27}$$

$$= \frac{1}{|\mathcal{S}^+|} \Big\{ H(\boldsymbol{X} \mid \boldsymbol{X}_A, X_{M+1}) + H(\boldsymbol{X}_A \mid X_{M+1}) \Big\} \tag{28}$$

$$= \frac{1}{|\mathcal{S}^+||\mathcal{S}|} \sum_{A \in \mathcal{S}} \Big\{ H(\boldsymbol{X}_{\{1,\dots,M\}\setminus A} \mid \boldsymbol{X}_A, X_{M+1}) + H(\boldsymbol{X}_A \mid X_{M+1}) \Big\} \tag{29}$$

Collecting all corresponding terms, we can re-write Equation (22) as follows:

$$\left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|} + \frac{1}{|\mathcal{S}^+||\mathcal{S}|}\right) \sum_{A \in \mathcal{S}} H(\boldsymbol{X}_{\{1,\dots,M\}\setminus A} \mid \boldsymbol{X}_A, X_{M+1}) \tag{30}$$

$$+ \left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|}\right) \sum_{A \in \mathcal{S}} I(\boldsymbol{X}_{\{1,\dots,M\}\setminus A}; X_{M+1} \mid \boldsymbol{X}_A) \tag{31}$$

$$+ \frac{1}{|\mathcal{S}^+|} \sum_{A \in \mathcal{S}} \mathbf{1}_{\{(A \cup \{M+1\}) \in \mathcal{S}^-\}} H(\boldsymbol{X}_{\{1,\dots,M\}\setminus A} \mid \boldsymbol{X}_A, X_{M+1}) \tag{32}$$

$$+ \frac{1}{|\mathcal{S}^+||\mathcal{S}|} \sum_{A \in \mathcal{S}} H(\boldsymbol{X}_A \mid X_{M+1}) \tag{33}$$

$$+ \frac{1}{|\mathcal{S}^+|} \sum_{A \in \mathcal{S}} H(X_{M+1} \mid \boldsymbol{X}) \tag{34}$$

For both the MMVAE and MoPoE, the first and last terms cancel out, which one can see by plugging in the respective definitions of $\mathcal{S}$ (e.g., see Appendix A.1). Hence, we arrive at the following expression

$$\underbrace{\left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|}\right) \sum_{A \in \mathcal{S}} I(\boldsymbol{X}_{\{1,\dots,M\}\setminus A}; X_{M+1} \mid \boldsymbol{X}_A)}_{(i)} + \underbrace{\frac{1}{|\mathcal{S}^+||\mathcal{S}|} \sum_{A \in \mathcal{S}} H(\boldsymbol{X}_A \mid X_{M+1}) + \frac{1}{|\mathcal{S}^+|} \sum_{A \in \mathcal{S}} H(X_{M+1} \mid \boldsymbol{X})}_{(ii)}$$

where only the first term can be negative since $|\mathcal{S}^+| > |\mathcal{S}|$. This concludes the proof of Corollary 2, showing that $\Delta(\boldsymbol{X}^+, \mathcal{S}^+) - \Delta(\boldsymbol{X}, \mathcal{S}) \geq 0$, if X_{M+1} is sufficiently diverse in the sense that $(ii) > (i)$. □

B Experiments

B.1 Implementation details

Our experiments are based on the publicly available code from Sutter et al. (2021), which already provides an implementation of PolyMNIST. A notable difference in our implementation is that we employ ResNet architectures, because we found that the previously used convolutional neural networks did not have sufficient capacity for the more complex datasets we use. For internal consistency, we use ResNets for PolyMNIST as well; we found no significant difference compared to the results from Sutter et al. (2021). In total, more than 300 models were trained, requiring approximately 1.5 GPU years of compute on a single NVIDIA GeForce RTX 2080 Ti GPU.

Hyperparameters All models were trained using the Adam optimizer (Kingma and Welling, 2014) with learning rate 5e-4 and a batch size of 256. For image modalities we estimate likelihoods using Laplace distributions and for captions we employ one-hot categorical distributions. Models were trained for 500, 1000, and 150 epochs on PolyMNIST, Translated-PolyMNIST, and CUB respectively. Similar to previous work, we use Gaussian priors and a latent space with 512 dimensions for PolyMNIST and 64 dimensions for CUB. For a fair comparison, we reduce the latent dimensionality of unimodal VAEs proportionally (wrt. the number of modalities) to control for capacity. The β-ablations use $\beta \in \{$3e-4, 3e-3, 3e-1, 1, 3, 9$\}$ and, in addition, 32 for CUB.

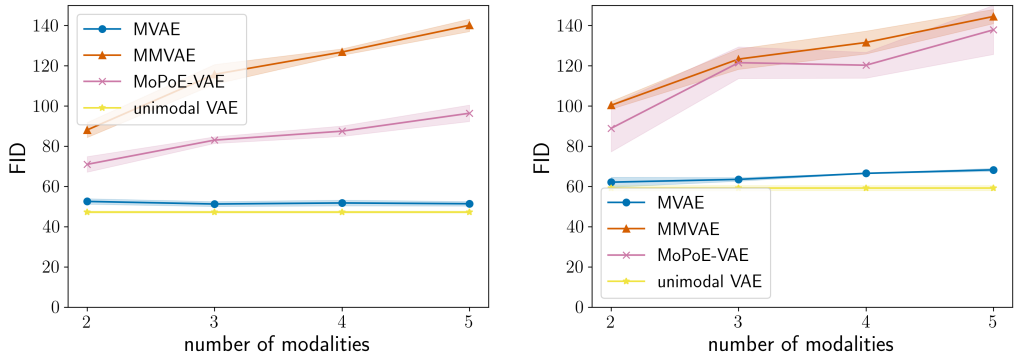
Metrics For the evaluation of *generative quality*, we use the Frechet inception distance (FID; Heusel et al., 2017)—a standard metric for evaluating the quality of generated images. In Appendix B.2, we also provide log-likelihoods and qualitative results for both images and captions. To compute *generative coherence*, it requires annotation on what is shared between modalities. For PolyMNIST and Translated-PolyMNIST, by design only the digit label is shared and since the ground truth is known, we can compute generative coherence as follows: For each modality, we first train a digit classifier on the respective training data; then, at test time, we run the conditionally generated images through the respective digit classifier and check, if the predicted class corresponds to the class of the input image. To be consistent with Sutter et al. (2021), Figure 3 shows the average *leave-one-out coherence*, for which the input consists of all modalities except the one that is being generated. Notably, for PolyMNIST, we use *same* digit classifiers that are provided in the publicly available code from (Sutter et al., 2021).

B.2 Additional experimental results

Figure 4 examines how generative quality is affected, if we vary the number of modalities. Notably, for mixture-based approaches that sub-sample modalities (MMVAE and MoPoE-VAE), the generative quality gap increases almost continuously with the number of modalities, which is consistent with our theoretical prediction from Corollary 2. Interestingly, for the MVAE, the generative gap on Translated-PolyMNIST also increases slightly, but the change is comparatively small.

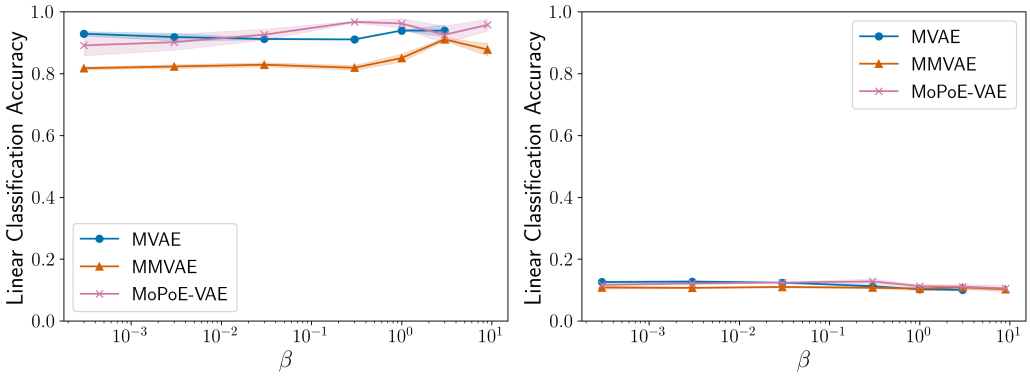
Figure 5 shows the linear classification accuracy on the learned representations. Shi et al. (2019) propose linear classification as a measure of latent factorization, to judge the quality of learned representations and to assess how well the information decomposes into shared and modality-specific features. The results suggest that not only does the generative coherence decline when we switch from PolyMNIST to translated-PolyMNIST, but also the quality of the learned representations.

Figure 6 shows the generative quality in terms of joint log-likelihoods. We observe a similar ranking of models as with FID, but we notice that the gap between MVAE and MoPoE-VAE appears less pronounced. The reason for this discrepancy is that, to be consistent with Sutter et al. (2021), we estimate joint log-likelihoods given *all* modalities—a procedure that resembles reconstruction more than it does unconditional generation. It can be of independent interest that log-likelihoods might overestimate the generative quality for unconditional generation for certain types of models. Qualitative results for unconditional generation (Figure 8) support the hypothesis that the presented log-likelihoods do not reflect the visible lack of generative quality for the MoPoE-VAE. Qualitative results for conditional generation (Figure 8) indicate a lack of diversity for both the MMVAE and MoPoE-VAE: even though we draw different samples from the posterior, the respective conditionally generated samples (i.e., the ten samples along each column) show little diversity in terms of backgrounds or writing styles.



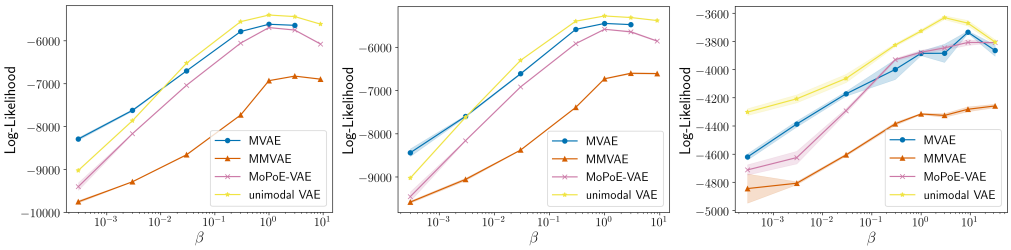
(a) PolyMNIST (b) translated-PolyMNIST

Figure 4: Generative quality as a function of the number of modalities. All models are trained with $\beta = 1$ on PolyMNIST and $\beta = 0.3$ on translated-PolyMNIST respectively. The results are averaged across three seeds and the bands show one standard deviation. For the unimodal VAE, the generative quality is plotted as a constant.



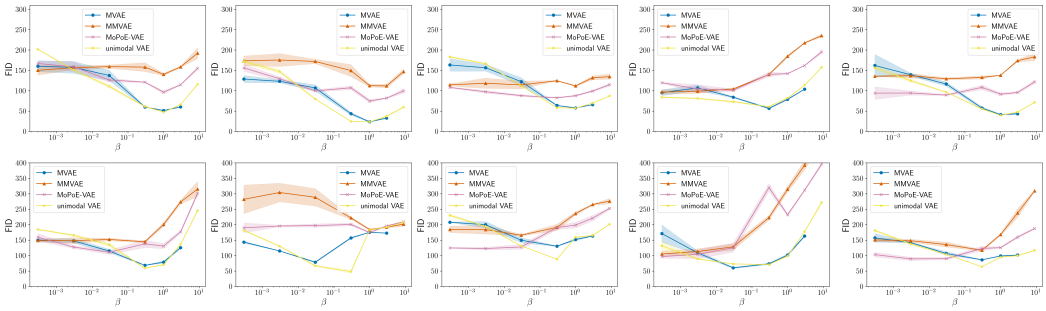
(a) PolyMNIST (b) Translated-PolyMNIST

Figure 5: Linear classification of latent representations. Points denotes the average digit classification accuracy of digit classifiers trained on the respective datasets. The results are averaged across three seeds and the bands show one standard deviation respectively. Note that due to numerical instabilities MVAE could not be trained with larger β values. For CUB, classification performance cannot be computed, because shared factors are not annotated.



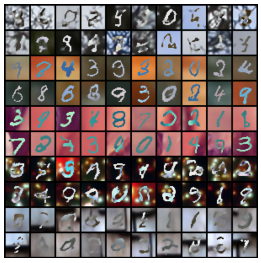
(a) PolyMNIST (b) Translated-PolyMNIST (c) Caltech Birds (CUB)

Figure 6: Joint log-likelihoods over a range of β values. Each point denotes the estimated joint log-likelihood averaged across three different seeds and the bands show one standard deviation respectively. Note that due to numerical instabilities MVAE could not be trained with larger β values.

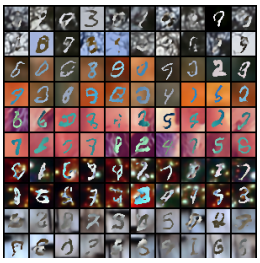


(a) X_1 (b) X_2 (c) X_3 (d) X_4 (e) X_5

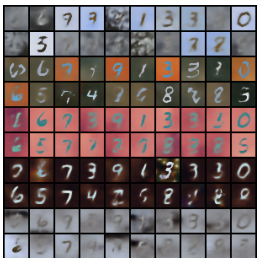
Figure 7: FID for modalities $X_1, \ldots, X_5$. Top row shows FID for PolyMNIST and bottom row for Translated-PolyMNIST. Points denote the FID averaged over three seeds and bands show one standard deviation respectively. Note that due to numerical instabilities MVAE could not be trained with larger β values.



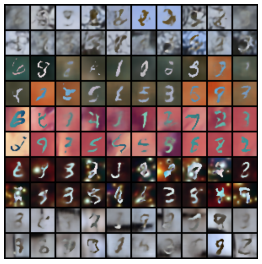
(a) unimodal VAE, $\beta = 1$



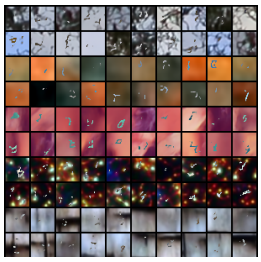
(b) MVAE, $\beta = 1$



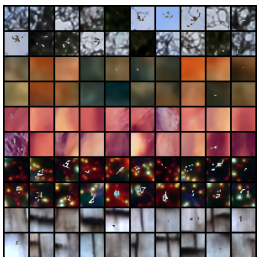
(c) MMVAE, $\beta = 1$



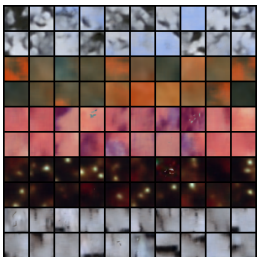
(d) MoPoE-VAE, $\beta = 1$



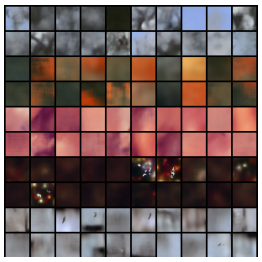
(e) unimodal VAE, $\beta = 0.3$



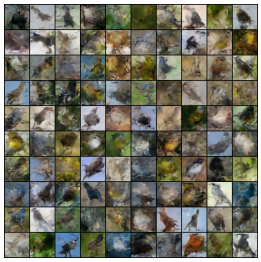
(f) MVAE, $\beta = 0.3$



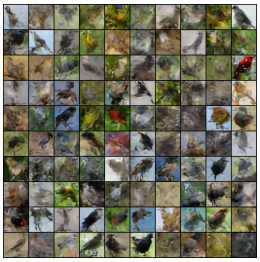
(g) MMVAE, $\beta = 0.3$



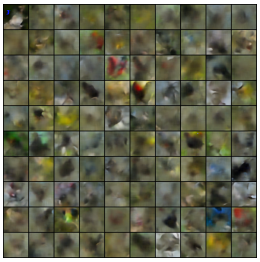
(h) MoPoE-VAE, $\beta = 0.3$



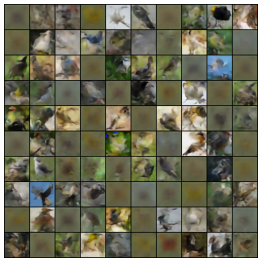
(i) unimodal VAE, $\beta = 9$



(j) MVAE, $\beta = 9$



(k) MMVAE, $\beta = 9$



(l) MoPoE-VAE, $\beta = 9$

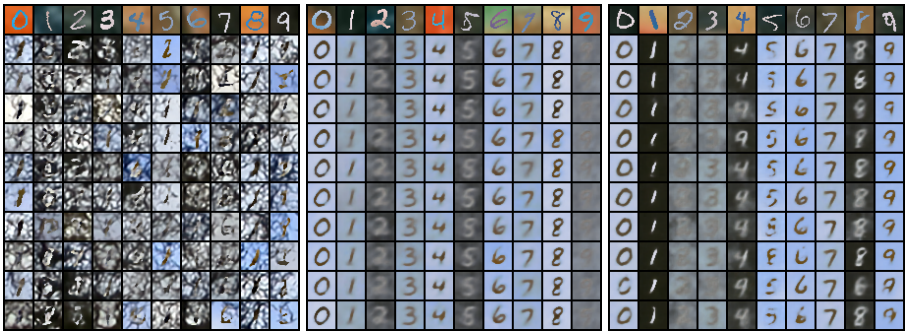
(m) unimodal VAE, $\beta = 9$

(n) MVAE, $\beta = 9$

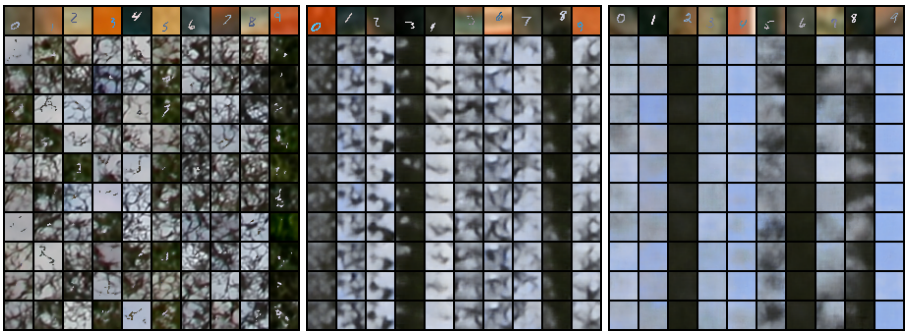
(o) MMVAE, $\beta = 9$

(p) MoPoE-VAE, $\beta = 9$

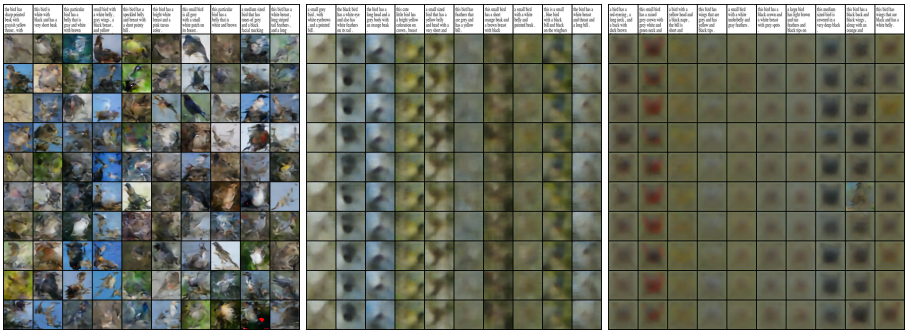
Figure 8: Qualitative results for the unconditional generation using prior samples. For PolyMNIST (Subfigures (a) to (d)) and Translated-PolyMNIST (Subfigures (e) to (h)), we show 20 samples for each modality. For CUB, we show 100 generated images (Subfigures (i) to (l)) and 100 generated captions (Subfigures (m) to (p)) respectively. Best viewed zoomed and in color.



(a) MVAE, $\beta = 1$ (b) MMVAE, $\beta = 1$ (c) MoPoE-VAE, $\beta = 1$



(d) MVAE, $\beta = 0.3$ (e) MMVAE, $\beta = 0.3$ (f) MoPoE-VAE, $\beta = 0.3$



(g) MVAE, $\beta = 9.0$ (h) MMVAE, $\beta = 9.0$ (i) MoPoE-VAE, $\beta = 9.0$

(j) MVAE, $\beta = 9.0$ (k) MMVAE, $\beta = 9.0$ (l) MoPoE-VAE, $\beta = 9.0$

Figure 9: Qualitative results for the conditional generation across modalities. For PolyMNIST (Subfigures (a) to (c)) and Translated-PolyMNIST (Subfigures (d) to (f)), we show 10 conditionally generated samples of modality X_1 given the sample from modality X_2 that is shown in the first row of the respective subfigure. For CUB, we show the generation of images given captions (Subfigures (g) to (i)), as well as the generation of captions given images (Subfigures (j) to (l)). Best viewed zoomed and in color.