Hölder++: Improving the Quality-Coherence Trade-off in Multimodal VAEs

Huyen Vo [1 2] **María Martínez-García** [1] **Isable Valera** [1 2]

Abstract

Existing approaches for multimodal variational autoencoders (VAEs) face a trade-off between generative quality and coherence—i.e., they struggle to generate realistic and diverse samples that, at the same time, are semantically consistent across modalities. A recent work shows that using a simple approximation to Hölder pooling as an aggregation method improves coherence over the SOTA MMVAE+, despite assuming a single shared representation across all modalities. Yet, it slightly compromises sample diversity. Inspired by this insight, we propose Hölder++, a novel multimodal VAE that improves the generative quality-coherence trade-off through: (i) the first implementation of *Hölder pooling without any approximation* for multimodal VAEs; (ii) an extended architecture that models *distinct shared and private* (i.e., modality-specific) representations (Hölder+); and (iii) *hierarchical inference* that further enhances the disentanglement between the shared and private representations (Hölder++). Our experiments corroborate that Hölder++ consistently improves the generative quality-coherence trade-off, yields more structured latent spaces, and learns shared representations that are informative for downstream tasks.

1. Introduction

Multimodal variational autoencoders (VAEs) are commonly designed and evaluated along two, often competing, criteria: *generative quality*, reflecting realism and sample diversity; and *generative coherence*, capturing semantic consistency across modalities. A key modeling design choice influencing the *generative quality-coherence trade-off* is the aggregation mechanism, which combines the representations produced by unimodal encoders into a shared representation across all modalities. The predominant approaches in the literature are Product-of-Experts (PoE) (Wu & Goodman, 2018) and Mixture-of-Experts (MoE) (Shi et al., 2019), which unfortunately suffer, respectively, from low coherence or reduced sample diversity (Daunhawer et al., 2022).

To address these limitations, Palumbo et al. (2023) introduced a new training objective that explicitly learns distinct shared and private (modality-specific) latent representations while avoiding *shortcuts*, resulting in MMVAE+. This was the first approach to achieve strong performance on both criteria, demonstrating that structuring the latent space is indeed crucial for improving the quality-coherence trade-off. A different line of work has focused on improving generative coherence by enforcing explicit disentanglement between private and shared subspaces through auxiliary loss terms, often grounded in information bottleneck or mutual information principles (Zhang et al., 2025; Gao et al., 2025; Lee & Pavlovic, 2020; Daunhawer et al., 2020).

Orthogonal to these works, Vo & Valera (2026) recently showed that the quality-coherence trade-off can also be improved through the choice of the aggregation method. In particular, they demonstrated that PoE and MoE can be viewed as special cases of Hölder pooling, a probabilistic opinion pooling framework that minimizes the α-divergence between the aggregated and individual densities. Focusing on the symmetric case $\alpha = 0.5$, they proposed a simple moment-matching approximation, Hellinger aggregation, which leads to an improved balance between generative quality and coherence. Notably, this approach achieves higher coherence than MMVAE+ despite relying on a single shared representation, at a slight deterioration of diversity.

Contributions. Motivated by these insights, we propose **Hölder++**, *a novel multimodal VAE that improves the state-of-the-art (SOTA) generative quality-coherence trade-off through symmetric Hölder pooling (i.e., $\alpha = 0.5$) and two key architectural choices*. Specifically, our contributions are threefold: (i) we provide the first implementation of symmetric Hölder pooling ($\alpha = 0.5$) as the aggregation mechanism for multimodal VAEs, without relying on approximations; (ii) we extend this framework with distinct shared and private latent subspaces, yielding the Hölder+ model; and (iii) we introduce top-down hierarchical inference to enforce disentanglement between shared and private representations by design, resulting in Hölder++. Experiments on three

[1]Department of Computer Science, Saarland University, DE [2]MPI-SWS, Saarland Informatics Campus, DE. Correspondence to: Huyen Vo <vothuckhanhhuyenvn@gmail.com>.

Preprint. May 1, 2026.

benchmark datasets (PolyMNIST, MNIST-SVHN, and CUBICC) demonstrate that Hölder++ consistently improves the quality-coherence trade-off, learns more structured and disentangled latent spaces, and infers shared representations that are highly informative for downstream tasks.

2. Background

2.1. Multimodal VAEs

Given data $\boldsymbol{X}$ consisting of M modalities, $\boldsymbol{X} := \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$, multimodal VAEs define a generative model with a shared latent variable $\boldsymbol{z}$. Under the assumption that modalities are conditionally independent given $\boldsymbol{z}$, the joint generative distribution factorizes as $p(\boldsymbol{X}, \boldsymbol{z}) = p(\boldsymbol{z}) \prod_{j=1}^{M} p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z})$, where $p(\boldsymbol{z})$ denotes the prior over the shared latent space and $p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z})$ is a modality-specific likelihood parameterized by a neural decoder with parameters $\boldsymbol{\theta}_j$. Multimodal VAEs are often trained by maximizing the Evidence Lower Bound (ELBO), given by

$$\text{ELBO} = \mathbb{E}_{q_\Phi(\boldsymbol{z}|\boldsymbol{X})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{X}|\boldsymbol{z})] - \text{KL}(q_\Phi(\boldsymbol{z}|\boldsymbol{X}) \| p(\boldsymbol{z})),$$

where the distribution $q_\Phi(\boldsymbol{z}|\boldsymbol{X})$ denotes a variational posterior with learnable parameters Φ. In this framework, the PoE (Wu & Goodman, 2018) and MoE (Shi et al., 2019) are standard approaches to obtain joint posterior approximations that scale with the number of modalities. The PoE approximates the joint posterior as $q_\Phi(\boldsymbol{z}|\boldsymbol{X}) = c \prod_{j=1}^{M} q_{\boldsymbol{\phi}_j}(\boldsymbol{z}|\boldsymbol{x}_j)$, where c is a normalization constant that ensures a valid probability distribution, whereas the MoE models it as an equally weighted mixture, given by $q_\Phi(\boldsymbol{z}|\boldsymbol{X}) = \frac{1}{M} \sum_{j=1}^{M} q_{\boldsymbol{\phi}_j}(\boldsymbol{z}|\boldsymbol{x}_j)$. In both cases, the joint posterior is constructed from modality-specific neural encoders with parameters $\boldsymbol{\phi}_j$, where $j \in \{1, 2, \ldots, M\}$.

Recent works have proposed novel aggregation methods for approximating $q_\Phi(\boldsymbol{z}|\boldsymbol{X})$ in this setting, including CoDE-VAE (Mancisidor et al., 2025), which leverages a consensus of dependent experts; WBVAE (Qiu et al., 2025), which adopts the 2-Wasserstein barycenter; and HELVAE (Vo & Valera, 2026), which introduces Hellinger aggregation, a Laplace approximation to Hölder pooling with $\alpha = 0.5$.

2.2. Hölder pooling

Probabilistic opinion pooling provides a principled approach for aggregating multiple probability density functions $\{q_j(\boldsymbol{z})\}_{j=1}^{M} \in \mathcal{P}^M$ into a single consensus (pooled) distribution as a weighted aggregation of individual densities, where the non-negative weights $\{\lambda_j\}_{j=1}^{M}$ satisfy $\sum_{j=1}^{M} \lambda_j = 1$. The pooling function is obtained by minimizing a weighted average of a chosen discrepancy measure between the aggregated and individual densities. Considering the family of α-divergences, this corresponds to $q(\boldsymbol{z}) = \arg\min_{\varphi \in \mathcal{P}} \sum_{j=1}^{M} \lambda_j \mathcal{D}_\alpha(q_j \| \varphi)$, which yields an α-parameterized family of Hölder pooling functions (Garg et al., 2004; Koliander et al., 2022). Recent work has studied PoE and MoE as special cases of Hölder pooling (Vo & Valera, 2026) and, based on this insight, proposed a novel aggregation method considering the symmetric case of the α-divergence family, which corresponds to $\alpha = 0.5$ (Hernandez-Lobato et al., 2016). The resulting aggregated distribution is given by

$$q(\boldsymbol{z}) = c \left(\sum_{j=1}^{M} \lambda_j^2 q_j(\boldsymbol{z}) + 2 \sum_{i=1}^{M} \sum_{j>i}^{M} \lambda_i \lambda_j \sqrt{q_i(\boldsymbol{z}) q_j(\boldsymbol{z})} \right), \tag{1}$$

where $c = 1 / \int \left(\sum_{j=1}^{M} \lambda_j \sqrt{q_j(\boldsymbol{z})} \right)^2 d\boldsymbol{z}$. Moreover, Vo & Valera (2026) derived a Laplace approximation to the symmetric Hölder pooling posterior via moment matching. The resulting method, referred to as Hellinger aggregation, significantly improves SOTA performance and achieves a better quality-coherence trade-off in multimodal VAEs that rely on a single latent space shared across all modalities.

2.3. Introducing shared and modality-specific subspaces

While earlier works explored shared and modality-specific latent subspaces in multimodal VAEs (Wang et al., 2016; Bouchacourt et al., 2018; Tsai et al., 2019), MM-VAE+ (Palumbo et al., 2023) was the first to achieve strong performance in both generative quality and generative coherence, and it remains a leading SOTA approach. In this framework, modality $\boldsymbol{x}_m$ is modeled with both a *private* (a.k.a. modality-specific) latent $\boldsymbol{w}_m$, and a *shared latent representation* $\boldsymbol{z}$ modeling the shared information across all modalities. The generative model assumes independence between the shared and private representations, factorizing as $p_\Theta(\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{W}) = p(\boldsymbol{z}) \prod_{j=1}^{M} p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z}, \boldsymbol{w}_j) p(\boldsymbol{w}_j)$, where $\boldsymbol{W} := \{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_M\}$, and the variational posterior is assumed to factorize accordingly as $q_\Phi(\boldsymbol{z}, \boldsymbol{W}|\boldsymbol{X}) = q_{\Phi_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}) \prod_{j=1}^{M} q_{\boldsymbol{\phi}_j}(\boldsymbol{w}_j|\boldsymbol{x}_j)$, where $q_{\Phi_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X})$ is approximated using MoE as aggregation method.

Wang et al. (2016), however, showed that such an approach can lead to *shortcuts*, where the modality-specific subspaces capture all the information, thus neglecting the shared representation. To avoid this behavior, Palumbo et al. (2023) introduced a modified objective (see Eq. (8), Appendix A.1) that distinguishes between self- and cross-reconstruction. For each modality, the likelihood (and thus, the reconstruction loss) is computed using a shared representation sampled from one of the unimodal encoders, $\boldsymbol{z} \sim q_{\boldsymbol{\phi}_{\boldsymbol{z}_j}}(\boldsymbol{z}|\boldsymbol{x}_j)$, together with the corresponding private latent sampled as:

$$\boldsymbol{w}_n \sim \begin{cases} q_{\boldsymbol{\phi}_{\boldsymbol{w}_n}}(\boldsymbol{w}_n \mid \boldsymbol{x}_n), & n = j \quad \textit{(self-term)}, \\ r_n(\boldsymbol{w}_n), & n \neq j \quad \textit{(cross-term)}, \end{cases} \tag{2}$$

where $\{r_n(\boldsymbol{w}_n)\}_{n=1}^M$ are (non-informative) auxiliary prior distributions on the private representations. This design forces the decoder to rely on the shared latent variable $\boldsymbol{z}$ when reconstructing unobserved modalities, thus preventing *shortcuts*. Moreover, CMVAE (Palumbo et al., 2024) extends this approach with a mixture prior over latent $\boldsymbol{z}$ to further enforce structure in the shared latent space.

To further improve generative coherence, several recent works enhance disentanglement between shared and private representations via auxiliary loss terms (Zhang et al., 2025; Gao et al., 2025; Lee & Pavlovic, 2020; Daunhawer et al., 2020). For instance, DCMEM (Gao et al., 2025) relies on a contrastive mutual-information loss, but is limited to bimodal settings. Similarly, DMVAE (Lee & Pavlovic, 2020) applies total-correlation regularization over the concatenated $[\boldsymbol{z}, \boldsymbol{w}]$ to enforce independence across latent dimensions. Both approaches therefore require nontrivial hyperparameter tuning. In contrast, we propose a variational factorization that is directly applicable to any number of modalities and encourages disentanglement by design.

3. Hölder++ VAE

In this section, we present the core components of our method. We first introduce exact symmetric Hölder pooling ($\alpha = 0.5$) as aggregation in multimodal VAEs (Section 3.1). We then incorporate a shared-private latent structure, yielding Hölder+ (Section 3.2). Finally, we introduce hierarchical inference to improve disentanglement between shared and private representations, yielding **Hölder++** (Section 3.3).

3.1. Exact (symmetric) Hölder pooling as aggregation

In multimodal VAEs with a single latent space shared across modalities, the Hellinger VAE (HELVAE) (Vo & Valera, 2026) has been shown empirically to improve the quality-coherence trade-off by using Hellinger aggregation, a Laplace approximation of Hölder pooling with $\alpha = 0.5$. In this paper, we *provide the first exact implementation of symmetric Hölder pooling ($\alpha = 0.5$) for multimodal VAEs*. The resulting pooled posterior admits a mixture representation consisting of unimodal and pairwise components, thus explicitly capturing the multimodal nature of the task.

More in detail, when applied to the joint posterior approximation, the Hölder pooling operator in Eq. (1) can be expressed as a mixture of Gaussians, comprising both unimodal and pairwise components, i.e.:

$$q(\boldsymbol{z}|\boldsymbol{X}) = \sum_{j=1}^{M} \pi_j q_{\boldsymbol{\phi}_{\boldsymbol{z}_j}}(\boldsymbol{z}|\boldsymbol{x}_j) + \sum_{i=1}^{M}\sum_{j>i}^{M} \pi_{ij} q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i, \boldsymbol{x}_j),$$

where π_j and π_{ij} denote the mixture weights of, respectively, the unimodal $q_{\boldsymbol{\phi}_{\boldsymbol{z}_j}}(\boldsymbol{z}|\boldsymbol{x}_j)$ and the pairwise $q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i, \boldsymbol{x}_j)$ components. For brevity, we omit the explicit dependence on the variational parameters and use the subscript ij for the pairwise components. We remark that the above formulation generalizes MMVAE (Shi et al., 2019), which uses MoE, by including additional pairwise components. Below, we provide the details on how each of these terms is computed; Appendix A.1 contains the full derivations.

The *unimodal mixture components* $\{q_{\boldsymbol{\phi}_{\boldsymbol{z}_j}}(\boldsymbol{z}|\boldsymbol{x}_j)\}_{j=1}^M$ are assumed to be Gaussian probability densities with diagonal covariance in $\mathbb{R}^D$, $\{\mathcal{N}(\boldsymbol{\mu}_j, \mathrm{diag}(\boldsymbol{\sigma}_j^2))\}_{j=1}^M$, with $\boldsymbol{z} \in \mathbb{R}^D$ being the latent representation (i.e., the latent variable), $\boldsymbol{\mu}_j = (\mu_{j,1}, \mu_{j,2}, \ldots, \mu_{j,D})^\top \in \mathbb{R}^D$ the mean vector, and $\boldsymbol{\sigma}_j^2 = (\sigma_{j,1}^2, \sigma_{j,2}^2, \ldots, \sigma_{j,D}^2)^\top \in \mathbb{R}^D$ the variance vector.

Each *pairwise mixture component* $q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i, \boldsymbol{x}_j)$ is obtained by normalizing the geometric mean of the corresponding unimodal posteriors $q_{\boldsymbol{\phi}_{\boldsymbol{z}_i}}(\boldsymbol{z}|\boldsymbol{x}_i)$ and $q_{\boldsymbol{\phi}_{\boldsymbol{z}_j}}(\boldsymbol{z}|\boldsymbol{x}_j)$. This results in a Gaussian distribution of the form $q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_{ij}^2)$. The parameters for each modality pair (i, j) with $1 \le i < j \le M$ and for each latent dimension $d \in \{1, 2, \ldots, D\}$ are given by:

$$\mu_{ij,d} = \frac{\mu_{i,d}\,\sigma_{j,d}^2 + \mu_{j,d}\,\sigma_{i,d}^2}{\sigma_{i,d}^2 + \sigma_{j,d}^2}, \quad \sigma_{ij,d}^2 = \frac{2\sigma_{i,d}^2\sigma_{j,d}^2}{\sigma_{i,d}^2 + \sigma_{j,d}^2}.$$

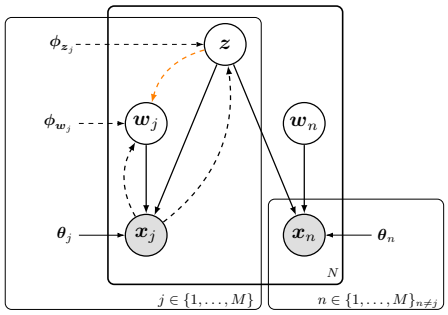
Finally, the *mixture weights* can be computed as $\pi_j = c$ and $\pi_{ij} = 2cS_{ij}$, where c denotes the normalization constant of the aggregated distribution in Eq. (1) and can be computed in closed form as

$$c = \left(M + 2\sum_{i=1}^{M}\sum_{j>i}^{M} S_{ij}\right)^{-1};$$

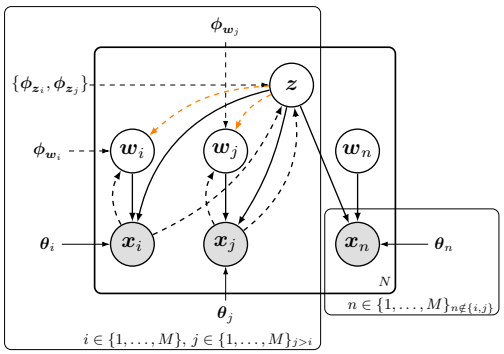
being S_{ij} the Bhattacharyya coefficient between the unimodal posteriors $q_{\phi_{\boldsymbol{z}_i}}$ and $q_{\phi_{\boldsymbol{z}_j}}$, i.e.,

$$S_{ij} = \prod_{d=1}^{D} \sqrt{\frac{2\sigma_{i,d}\sigma_{j,d}}{\sigma_{i,d}^2 + \sigma_{j,d}^2}} \exp\left(-\frac{(\mu_{i,d} - \mu_{j,d})^2}{4(\sigma_{i,d}^2 + \sigma_{j,d}^2)}\right).$$

Remark. Hölder VAE suffers from two limitations relative to HELVAE, both stemming from its mixture subsampling scheme. First, it substantially increases computational complexity by requiring sampling from a mixture with M^2 components (see Table 7 in Appendix C.1 for a computational comparison). Second, we expect the Hölder VAE to suffer from limited generative quality, like other mixture-based multimodal VAEs, such as MMVAE (Shi et al., 2019) and MoPoE (Sutter et al., 2021), that rely on mixture subsampling of a single shared representation (Daunhawer et al., 2022) (see results in Section 4.1). Fortunately, as shown in MMVAE+ (Palumbo et al., 2023), this effect can be mitigated by splitting the latent representation into shared and private subspaces. Thus, motivating our Hölder+ extension.



(a) Unimodal components in the Hölder++ objective



(b) Pairwise components in the Hölder++ objective

$$\mathcal{L}_{\text{Hölder++}}(\boldsymbol{x}_{1:M}) = \sum_{j=1}^{M} \pi_j \mathbb{E}_{\substack{q_{\boldsymbol{\phi}_{\boldsymbol{z}_j}}(\boldsymbol{z}|\boldsymbol{x}_j) \\ q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j,\boldsymbol{z}) \\ \{\tilde{\boldsymbol{w}}_n \sim r_n(\boldsymbol{w}_n)\}_{n\neq j}}} \log \left(\frac{p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z},\boldsymbol{w}_j) p(\boldsymbol{z}) p(\boldsymbol{w}_j)}{q_{\boldsymbol{\phi}_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}) q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j,\boldsymbol{z})} \prod_{n\neq j} p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z},\tilde{\boldsymbol{w}}_n) \right)$$

$$+ \sum_{i=1}^{M}\sum_{j>i}^{M} \pi_{ij} \mathbb{E}_{\substack{q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i,\boldsymbol{x}_j) \\ q_{\boldsymbol{\phi}_{\boldsymbol{w}_i}}(\boldsymbol{w}_i|\boldsymbol{x}_i,\boldsymbol{z}) \\ q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j,\boldsymbol{z}) \\ \{\tilde{\boldsymbol{w}}_n \sim r_n(\boldsymbol{w}_n)\}_{n\notin\{i,j\}}}} \log \left(\frac{p_{\boldsymbol{\theta}_i}(\boldsymbol{x}_i|\boldsymbol{z},\boldsymbol{w}_i) p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z},\boldsymbol{w}_j) p(\boldsymbol{z}) p(\boldsymbol{w}_i) p(\boldsymbol{w}_j)}{q_{\boldsymbol{\phi}_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}) q_{\boldsymbol{\phi}_{\boldsymbol{w}_i}}(\boldsymbol{w}_i|\boldsymbol{x}_i,\boldsymbol{z}) q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j,\boldsymbol{z})} \prod_{n\notin\{i,j\}} p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z},\tilde{\boldsymbol{w}}_n) \right).$$

(c) Hölder++ objective

Figure 1. A graphical-model view of the unimodal and pairwise components used by Hölder++ (top) and the resulting training objective (bottom). Gray circles denote observed variables, white circles denote latent variables, and non-circled symbols denote model parameters. Solid arrows indicate the generative process, while dashed arrows indicate amortized posterior inference. The objective is a weighted sum of unimodal and pairwise ELBO-style terms, where highlighted terms indicate the modifications introduced by hierarchical inference relative to the baseline architecture. We apply the shortcut-avoiding scheme in Eqs. (2) and (3) to the hierarchical modality-specific posteriors, i.e., $\{q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j,\boldsymbol{z})\}_{j=1}^{M}$, with $j \in \{1,2,\ldots,M\}$. For clarity, the graphical model explicitly shows that, in the pairwise component (i,j), the posterior over $\boldsymbol{z}$ is parameterized by the unimodal encoders $\boldsymbol{\phi}_{\boldsymbol{z}_i}$ and $\boldsymbol{\phi}_{\boldsymbol{z}_j}$.

3.2. Hölder+: Learning shared and private subspaces

Models that rely on a single latent space shared across modalities often empirically suffer from limited sample diversity, a phenomenon observed in novel methods such as HELVAE (Vo & Valera, 2026) or CoDEVAE (Mancisidor et al., 2025), despite the improvements in generative coherence. We address this limitation by factorizing the latent space into shared and modality-specific subspaces. To prevent *shortcuts*, where private latent representations capture shared semantic information, we adopt the scheme proposed in MMVAE+ and rely on auxiliary non-informative distributions to sample private features in the cross-modal reconstruction terms (Palumbo et al., 2023). As highlighted in the previous section, the latter design choice forces the decoder to rely only on the shared latent $\boldsymbol{z}$ when reconstructing unobserved modalities, thus preventing shortcuts.

Specifically, when $\boldsymbol{z}$ is sampled from a unimodal component j, i.e., $\boldsymbol{z} \sim q_{\boldsymbol{\phi}_j}(\boldsymbol{z}|\boldsymbol{x}_j)$, we compute the conditional log-likelihood terms $\log p(\boldsymbol{x}_n|\boldsymbol{z},\boldsymbol{w}_n)$ for all modalities $n \in \{1,2,\ldots,M\}$, where each $\boldsymbol{w}_n$ is sampled following Eq. (2). Analogously, when $\boldsymbol{z}$ is sampled from a pairwise component (i,j), i.e., $\boldsymbol{z} \sim q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i,\boldsymbol{x}_j)$, we evaluate the same reconstruction terms with $\boldsymbol{w}_n$ sampled as

$$\boldsymbol{w}_n \sim \begin{cases} q_{\boldsymbol{\phi}_{\boldsymbol{w}_n}}(\boldsymbol{w}_n|\boldsymbol{x}_n), & n \in \{i,j\}, \\ r_n(\boldsymbol{w}_n), & n \notin \{i,j\}, \end{cases} \tag{3}$$

where $\{r_n(\boldsymbol{w}_n)\}_{n=1}^{M}$ are auxiliary, non-informative priors for the private (i.e., modality-specific) latent representations. We refer to the resulting model as **Hölder+**, with the full objective presented in Appendix A.1, Eq. (9). We remark that Hölder+ optimizes a valid ELBO and is therefore a proper multimodal VAE, as proven in Appendix A.2.

Remark. We stress that, while HELVAE shows advantages over Hölder in a single shared representation setting, that is not the case when considering distinct shared-private subspaces. This is due to the fact that Hellinger aggregation does sample the shared representation $\boldsymbol{z}$ after aggregating all the modalities, thus failing to distinguish from self- and cross-reconstruction (and sampling) (see Eqs. (2) and (3)), which is essential to avoid shortcuts. Consequently, we expect Hölder+ to yield improved quality-coherence trade-offs relative to HELVAE, analogously to the improvement of MMVAE+ over MMVAE (see experiments in Section 4.1).

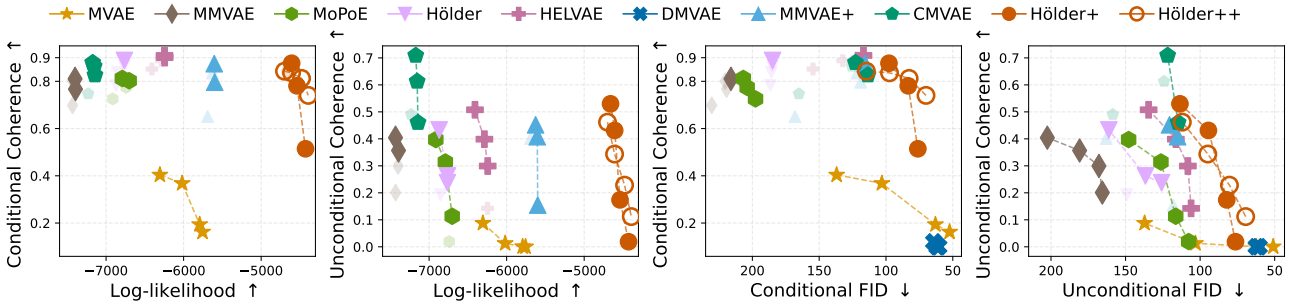


Figure 2. Trade-offs on PolyMNIST between generative coherence (↑) and log-likelihood estimation (↑), as well as between generative coherence (↑) and generative quality (↓), with $\beta \in \{1, 2.5, 5, 10\}$. For each model, the Pareto front (dashed line) connects the non-dominated points that achieve the best trade-offs. Optimal region: upper-right for all plots.

3.3. Hölder++: Disentangling private-share subspaces

Existing methods promote shared-private disentanglement by introducing auxiliary loss terms based on information-bottleneck or mutual-information principles (Zhang et al., 2025; Gao et al., 2025; Lee & Pavlovic, 2020; Daunhawer et al., 2020). In contrast, we propose a variational factorization that encourages *disentanglement by design*.

Multimodal VAEs often approximate the posterior distribution by assuming that shared and private representations are conditionally independent given the data, i.e., $q_\Phi(\boldsymbol{z}, \boldsymbol{W}|\boldsymbol{X}) = q_{\phi_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X})\, q_{\phi_{\boldsymbol{W}}}(\boldsymbol{W}|\boldsymbol{X})$. However, the true posterior generally does not factorize with respect to either the shared or the private representations, even if independence is assumed in the prior. To improve our posterior approximation, we instead adopt hierarchical inference by parameterizing the variational posterior as

$$q_\Phi(\boldsymbol{z}, \boldsymbol{W}|\boldsymbol{X}) = q_{\Phi_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}) \prod_{j=1}^{M} q_{\phi_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j, \boldsymbol{z}),$$

i.e., we first infer the shared latent representation $\boldsymbol{z}$ from all modalities (the *top level* of our hierarchy), and then infer each modality-specific latent $\boldsymbol{w}_j$ (the *bottom level*) conditioned on both the input data $\boldsymbol{x}_j$ and the shared latent $\boldsymbol{z}$. This design choice reflects the intuition that, in multimodal VAEs, generative coherence across modalities can only be achieved with an informative shared representation. Assuming that the shared and private representations act as information bottlenecks, and adopting a top-down factorization of the approximate posterior, we ensure that $\boldsymbol{w}_j$ does not capture all the information in data (Sønderby et al., 2016; Vahdat & Kautz, 2020; Havtorn et al., 2021). Instead, $\{\boldsymbol{w}_j\}_{j=1}^M$ models only modality-specific (residual) information not captured by $\boldsymbol{z}$, thereby avoiding undesirable shortcuts.

By applying hierarchical inference to Hölder+, we obtain **Hölder++**. Figure 1 shows the Hölder++ graphical model for the unimodal and pairwise components (panels (a) and (b)) and the corresponding training objective (panel (c)), highlighting the modifications relative to Hölder+.

Remark. Our approach is fundamentally different from the Hierarchical Multimodal VAE (HMVAE) (Wolff et al., 2022), which uses a top-down hierarchy for both inference and generation. In HMVAE, each modality-specific latent is conditioned on the shared latent, and only the private representations are passed to the unimodal decoders, potentially limiting the coherence across modalities if the modality-specific representations are expressive enough to capture all the information in the data. In our implementation of Hölder++, we instead assume that the private and shared representations are independent *a priori*, and use hierarchical inference to enhance disentanglement during inference. Yet, extensions to account for a top-down (from shared to private) generative model of Hölder++ are straightforward and could be of interest particularly in some applications, e.g., where the shared content affects the modality-specific style (Von Kügelgen et al., 2021; Daunhawer et al., 2023).

4. Experimental results

Datasets. We evaluate our approach on three standard benchmark datasets: PolyMNIST (Sutter et al., 2021), MNIST-SVHN (Shi et al., 2019), and CUBICC (Palumbo et al., 2024). PolyMNIST is a synthetic dataset with five modalities, where each example is generated by patching MNIST digits of the same class onto random crops from five background images. MNIST-SVHN pairs MNIST and Street View House Numbers (SVHN) digits with the same labels but different visual styles. CUBICC, a variant of the CUB image-caption dataset, contains bird images paired with textual descriptions, grouped into eight species categories, and we use it to evaluate downstream clustering.

Baselines. We compare against SOTA methods that use a single latent shared across modalities—MVAE (Wu & Goodman, 2018), MMVAE (Shi et al., 2019), MoPoE (Sutter et al., 2021), and HELVAE (Vo & Valera, 2026)—as well as approaches that use distinct shared and modality-specific latents, including DMVAE (Lee & Pavlovic, 2020), MMVAE+ (Palumbo et al., 2023), CMVAE (Palumbo et al., 2024), and DCMEM (Gao et al., 2025). For a fair compar-

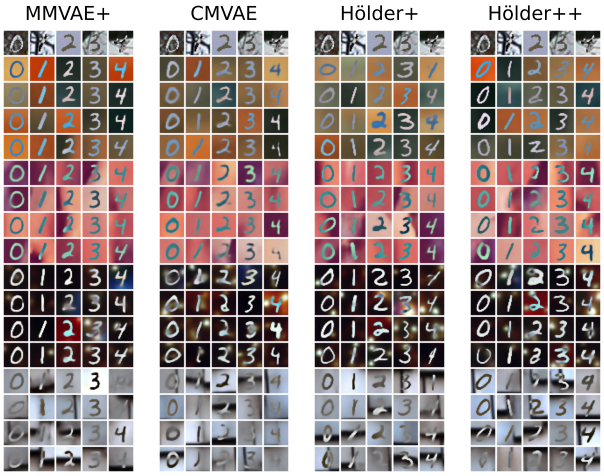


Figure 3. Qualitative results for conditional generation on PolyMNIST. The input example from the first modality is shown in the top row, and the rows below display four conditional samples for each of the remaining modalities.

ison with CMVAE on the downstream clustering task, we also apply the mixture prior on $\boldsymbol{z}$ to Hölder+ and Hölder++, yielding CHölder+ and CHölder++. All experiments report average performance over 3 random seeds, except on CUBICC, where we use 10 seeds. Further experimental details and results are provided in Appendices B and C.

Metrics. We assess the **quality-coherence trade-off** using Fréchet Inception Distance (FID) (Heusel et al., 2017) as a measure of generative quality and classification accuracy on the generated samples as coherence. Moreover, we evaluate the quality of the posterior approximation using the ELBO.

To assess the **disentanglement** between shared and private subspaces, we evaluate it both directly on the inferred representations and in the generated images. For the former, we follow prior work by training a linear classifier on the latent spaces and reporting accuracy. We expect high accuracy for the shared latent $\boldsymbol{z}$ and low accuracy for the private latent $\boldsymbol{w}$, which should not encode class information. For the latter, we introduce three disentanglement metrics: $\boldsymbol{z}$ *content stability* (↑), $\boldsymbol{z}$ *content accuracy* (↑), and $\boldsymbol{w}$ *content accuracy* (↓). For the first two, we fix $\boldsymbol{z}$, sample multiple $\boldsymbol{w} \sim p(\boldsymbol{w})$, decode, and classify the outputs. While $\boldsymbol{z}$ *content stability* measures agreement across samples, $\boldsymbol{z}$ *content accuracy* measures accuracy with respect to the ground-truth label. Higher values indicate that content is captured exclusively by $\boldsymbol{z}$, as fixing $\boldsymbol{z}$ yields outputs with accurate and invariant content despite variations in $\boldsymbol{w}$. For $\boldsymbol{w}$ *content accuracy*, we fix $\boldsymbol{w}$, sample multiple $\boldsymbol{z} \sim p(\boldsymbol{z})$, decode, classify the outputs, and measure classification accuracy. Lower values reflect stronger disentanglement. All three metrics are averaged over self- and cross-generation.

Finally, we assess representation quality via **downstream clustering** on the shared latent space using K-means and

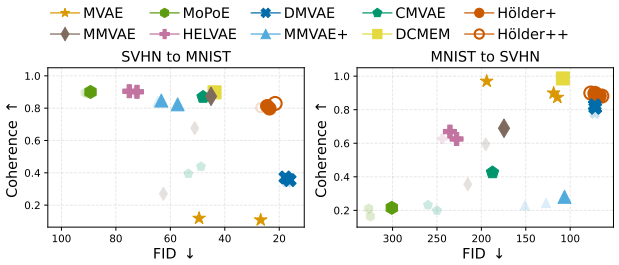


Figure 4. Trade-offs on MNIST-SVHN between generative coherence (↑) and generative quality (↓). For each model, non-dominated points are larger, and dominated points have low opacity. Optimal region: upper-right for both plots.

report accuracy (ACC), normalized mutual information (NMI), and adjusted rand index (ARI).

4.1. Generative quality and coherence trade-off

PolyMNIST. PolyMNIST has five modalities, so we omit DCMEM, as it is restricted to bimodal settings. We evaluate generative coherence and quality using samples drawn from both the joint posterior (conditional) and the prior (unconditional). Figure 2 summarizes these trade-offs across generative performance metrics for different β values (Higgins et al., 2017). Here, we first observe that HELVAE outperforms the rest of the *models with a single shared representation*, including Hölder. Yet, the latter significantly improves the achieved trade-offs compared to MMVAE and MoPoE, showing that the symmetric Hölder pooling improves both quality and coherence, even when limited by mixture subsampling. These results confirm our insights in Section 3.1 and thus argue against the use of mixture subsampling in shared-representation models. Consequently, we will not report further Hölder VAEs, but only HELVAE.

Moving to *models with distinct shared-private representations*, we first observe that DMVAE achieves the lowest FID but exhibits poor coherence; in particular, DMVAE's factorized representations are highly prone to shortcut solutions. Disregarding DMVAE, we observe that shared-private models, in general, outperform their single-representation counterparts in the achievable quality-coherence trade-offs, except for HELVAE, which remains overall competitive. Importantly, *our Hölder+ and Hölder++ yield the best Pareto trade-offs*, combining the highest log-likelihood with competitive FID while matching HELVAE in coherence. Although CMVAE achieves the highest unconditional coherence due to its mixture prior over $\boldsymbol{z}$, it significantly underperforms in log-likelihood. The large improvement of Hölder+ over MMVAE+ shows again the benefit of symmetric Hölder pooling. Moreover, Hölder++ closely matches Hölder+, indicating that hierarchical inference maintains Pareto optimality. Qualitative results in Figure 3 further confirm the generative quality of Hölder+ and Hölder++.

MNIST-SVHN. Figure 4 shows the coherence-FID trade-

Table 1. Digit classification accuracy on the latent representations for MNIST-SVHN, evaluated on the private $\boldsymbol{w}$, shared $\boldsymbol{z}$, and joint $[\boldsymbol{w}, \boldsymbol{z}]$ subspaces. The best and second-best results are highlighted in bold and underlined, respectively.

	MNIST Representation			SVHN Representation		
	Joint (↑)	Shared (↑)	Private (↓)	Joint (↑)	Shared (↑)	Private (↓)
DMVAE	0.970 ± 0.001	0.862 ± 0.010	0.673 ± 0.043	0.900 ± 0.010	0.898 ± 0.009	<u>0.113 ± 0.001</u>
MMVAE+	0.953 ± 0.002	0.857 ± 0.026	0.471 ± 0.056	0.910 ± 0.003	0.911 ± 0.003	0.118 ± 0.004
DCMEM	**0.989 ± 0.001**	**0.989 ± 0.001**	**0.241 ± 0.010**	0.911 ± 0.002	0.913 ± 0.004	0.129 ± 0.002
CMVAE	0.948 ± 0.001	0.861 ± 0.060	0.389 ± 0.135	0.913 ± 0.002	0.914 ± 0.002	0.116 ± 0.001
Hölder+	0.976 ± 0.001	0.966 ± 0.003	0.479 ± 0.007	**0.923 ± 0.004**	**0.922 ± 0.004**	0.114 ± 0.001
Hölder++	<u>0.977 ± 0.001</u>	<u>0.970 ± 0.001</u>	<u>0.387 ± 0.023</u>	<u>0.922 ± 0.001</u>	<u>0.922 ± 0.001</u>	**0.112 ± 0.001**

Table 2. Disentanglement metrics on MNIST-SVHN. The best and second-best results are marked bold and underlined, respectively.

	$\boldsymbol{w}$ content accuracy ↓	$\boldsymbol{z}$ content stability ↑	$\boldsymbol{z}$ content accuracy ↑
DMVAE	0.168 ± 0.006	0.523 ± 0.004	0.579 ± 0.005
MMVAE+	0.151 ± 0.017	0.712 ± 0.044	0.664 ± 0.016
CMVAE	0.139 ± 0.015	0.803 ± 0.053	0.665 ± 0.022
DCMEM	**0.102 ± 0.001**	**0.853 ± 0.006**	**0.883 ± 0.005**
Hölder+	0.121 ± 0.005	0.800 ± 0.008	0.838 ± 0.006
Hölder++	<u>0.119 ± 0.001</u>	<u>0.827 ± 0.010</u>	<u>0.857 ± 0.007</u>

Table 3. Disentanglement metrics on CUBICC. The best and second-best results are marked bold and underlined, respectively.

Method	$\boldsymbol{w}$ content accuracy ↓	$\boldsymbol{z}$ content stability ↑	$\boldsymbol{z}$ content accuracy ↑
MMVAE+	0.138 ± 0.013	0.846 ± 0.166	0.624 ± 0.049
DCMEM	0.321 ± 0.081	0.207 ± 0.110	0.204 ± 0.123
Hölder+	0.195 ± 0.048	0.442 ± 0.087	0.443 ± 0.051
Hölder++	<u>0.133 ± 0.004</u>	0.870 ± 0.097	0.619 ± 0.021
CMVAE	0.136 ± 0.011	<u>0.893 ± 0.145</u>	<u>0.641 ± 0.060</u>
CHölder+	0.142 ± 0.006	0.587 ± 0.045	0.535 ± 0.030
CHölder++	**0.132 ± 0.003**	**0.914 ± 0.066**	**0.642 ± 0.014**

off for conditional cross-modal generation (across β values). Several baselines exhibit a clear gap between directions: MMVAE+ and CMVAE perform well for SVHN→MNIST but degrade substantially for MNIST→SVHN, whereas MVAE and DMVAE show the opposite trend. In contrast, *Hölder+ and Hölder++ remain in the upper-right region in both directions*, combining consistently high coherence with among the overall lowest FID. Relative to DCMEM (varying α), Hölder-based models deliver higher sample quality (lower FID) while being competitive in coherence. Finally, as before, Hölder++ overlaps Hölder+.

Take-away. *While HELVAE remains the SOTA among single shared-representation models, Hölder+ and Hölder++ consistently achieve the best Pareto frontier* compared to all competing methods overall. Moreover, Hölder++ closely matches Hölder+, showing that hierarchical inference preserves the coherence-quality trade-off.

4.2. Disentanglement of shared and private subspaces

For PolyMNIST, separating (shared) digit identity from the (private) background is straightforward (Figure 8 in Appendix C); so we omit disentanglement metrics.

MNIST-SVHN. Moving to MNIST-SVHN, disentangling shared and private factors is more challenging due to SVHN's background clutter. Table 1 shows the *disentanglement measured directly on the latent representations*, indicating that Hölder++ notably reduces MNIST private-latent accuracy compared to Hölder+. Moreover, for Hölder++, performance on the joint $[\boldsymbol{w}, \boldsymbol{z}]$ and the shared $\boldsymbol{z}$ differs only slightly, showing that the drop in private-latent accuracy reflects reduced leakage into $\boldsymbol{w}$ rather than a degradation of the $\boldsymbol{z}$ representation. Our models also outperform DCMEM on the SVHN modality, which is more complex than MNIST.

The *disentanglement metrics measured on the generated images* are reported in Table 2. Here, Hölder++ improves over Hölder+ by lowering $\boldsymbol{w}$ content accuracy and increasing $\boldsymbol{z}$ content accuracy and stability, yielding more robust generations under changes in private factors. While DCMEM achieves the best disentanglement, consistent with its high coherence, it is limited to bimodal settings; in contrast, our approach scales beyond two modalities and remains competitive in generative quality and coherence. Qualitative results in Appendix C, Figure 9, further show that, for cluttered SVHN inputs containing multiple digits, Hölder++ is more stable than DCMEM under variations in the private latent $\boldsymbol{w}$, while maintaining accurate and high-quality generations.

CUBICC. Table 3 reports disentanglement metrics for cross-modal generation with images as the target modality. Across all three metrics, hierarchical inference yields consistent gains (Hölder++ over Hölder+ and CHölder++ over CHölder+), supporting our modeling assumptions. Notably, CHölder++ achieves the strongest disentanglement with low variance, outperforming the next-best CMVAE, which exhibits much higher variance. Table 9 in Appendix C reports the disentanglement analysis on the latent representations.

Take-away. These results confirm that *hierarchical inference effectively prevents class information from leaking into the modality-specific representations*, especially on complex datasets such as CUBICC, while preserving generative performance and downstream-task results (see Section 4.3). This behavior is consistent with our inference design, which conditions $\boldsymbol{w}$ on $\boldsymbol{z}$ to capture posterior dependencies without introducing them into the generative model.

Table 4. Clustering performance on CUBICC using shared latent representations. We partition models according to whether $\boldsymbol{z}$ follows a structured clustering prior. The best and second-best results are marked bold and underlined, respectively.

	Image Representation (↑)			Caption Representation (↑)			Joint Representation (↑)		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
MMVAE+	30.1 ± 2.1	16.2 ± 2.8	9.2 ± 1.9	22.9 ± 2.6	7.6 ± 2.8	3.3 ± 2.0	35.5 ± 4.4	25.4 ± 5.2	15.7 ± 4.8
DCMEM	34.4 ± 28.5	20.7 ± 30.4	17.4 ± 26.9	28.0 ± 18.6	13.5 ± 19.3	9.9 ± 15.3	32.8 ± 26.5	21.1 ± 31.1	17.4 ± 27.0
Hölder+	28.8 ± 3.2	14.0 ± 4.3	7.9 ± 2.8	20.7 ± 1.0	4.9 ± 1.0	1.7 ± 0.6	32.3 ± 3.4	20.7 ± 3.4	11.8 ± 2.3
Hölder++	28.3 ± 2.6	14.3 ± 2.4	7.7 ± 1.8	19.7 ± 1.1	3.9 ± 0.9	1.2 ± 0.5	31.8 ± 2.4	18.8 ± 3.0	10.9 ± 2.3
CMVAE	51.9 ± 12.8	42.2 ± 12.1	31.1 ± 14.0	44.6 ± 13.8	**33.7 ± 12.8**	**23.8 ± 13.1**	58.0 ± 13.8	51.3 ± 12.4	39.3 ± 14.9
CHölder+	**59.1 ± 6.1**	**48.8 ± 3.4**	**36.2 ± 4.8**	**45.3 ± 4.2**	32.3 ± 2.2	21.3 ± 2.8	61.3 ± 4.6	51.7 ± 2.8	38.6 ± 3.4
CHölder++	57.3 ± 3.7	46.4 ± 2.1	34.1 ± 2.3	44.0 ± 3.4	31.4 ± 2.4	20.5 ± 2.4	**65.3 ± 5.8**	**55.1 ± 3.5**	**43.2 ± 5.2**

Table 5. Effect of hierarchical inference on CUBICC. The best and second-best results are marked bold and underlined, respectively.

Method	$\boldsymbol{w}$ content accuracy ↓	$\boldsymbol{z}$ content stability ↑	$\boldsymbol{z}$ content accuracy ↑
MMVAE+	0.138 ± 0.013	0.846 ± 0.166	0.624 ± 0.049
MMVAE++	0.133 ± 0.002	**0.960 ± 0.008**	0.640 ± 0.010
Hölder+	0.195 ± 0.048	0.442 ± 0.087	0.443 ± 0.051
Hölder++	0.133 ± 0.004	0.870 ± 0.097	0.619 ± 0.021
CMVAE	0.136 ± 0.011	0.893 ± 0.145	0.641 ± 0.060
CMVAE++	**0.130 ± 0.003**	0.959 ± 0.008	**0.651 ± 0.011**
CHölder+	0.142 ± 0.006	0.587 ± 0.045	0.535 ± 0.030
CHölder++	0.132 ± 0.003	0.914 ± 0.066	0.642 ± 0.014

4.3. Downstream clustering task

CUBICC. We apply (bird) clustering to the latent representations on the CUBICC dataset (Palumbo et al., 2024). Table 4 shows that assuming a mixture model on the shared latent space improves the results. CHölder+ and CHölder++ rank first and second across all clustering metrics across modalities. While CMVAE is comparable to ours on the caption representation, it exhibits higher variance, similar to DCMEM, indicating sensitivity to training initialization. Refer to Figure 5 in Appendix C for the plots for 10 independent runs of every model. We run paired Hotelling's T^2 tests (Hotelling, 1931) across 10 seeds to compare CMVAE with CHölder+ and CHölder++. Our results show that, while our improvement on the caption representation over CMVAE is not statistically significant, CHölder+ and CHölder++ provide statistically ($p < 0.05$) and marginally ($p \approx 0.05$) significant improvements, respectively, in the image representation. Remarkably, both Hölder-based models significantly outperform CMVAE in the joint representation.

4.4. Effect of hierarchical inference across models

CUBICC. We investigate the effect of hierarchical inference across models by applying it to MMVAE+ and CMVAE, yielding MMVAE++ and CMVAE++. Table 5 shows consistent improvements across the three disentanglement metrics for all backbones under hierarchical inference. The gains are larger for Hölder+ and CHölder+ because shortcut prevention for pairwise components is activated in the $M > 2$ setting, as in Eq. (3). Overall, MMVAE++, CMVAE++, and CHölder++ achieve comparable disentanglement, while CHölder++ remains best in clustering (see Table 8 in Appendix C). These results further confirm that: (i) Hölder-based VAEs yield better trade-offs across generative performance metrics; and (ii) hierarchical inference is a robust and effective design choice for learning disentangled private-shared representations in multimodal VAEs.

5. Conclusion

Summary. First, we present the first exact implementation of symmetric Hölder pooling, whose pairwise components implicitly capture soft cross-modality dependencies by increasing regions of mutual support, unlike PoE and MoE, which assume modality independence. Second, we extend the Hölder VAE with explicit shared and modality-specific subspaces, further improving the quality-coherence trade-off, increasing sample diversity, and outperforming strong baselines such as MMVAE+. Finally, we introduce a hierarchical variational posterior that reduces shared-information leakage into modality-specific representations, thereby enhancing private-shared disentanglement and yielding useful representations for downstream tasks.

Future work. While our models obtain the best trade-offs among multimodal VAEs, they do not yet match SOTA image generators like Denoising Diffusion Probabilistic Models (DDPMs). Existing approaches close this gap via post-processing, e.g., using a DDPM to denoise VAE-generated images (Palumbo et al., 2024; Zhang et al., 2025). As future work, we will investigate how to improve generative performance within the VAE framework via latent diffusion (Wesego & Rooshenas, 2024a; Bounoua et al., 2024) or flexible priors (Wesego & Rooshenas, 2024b; Yuan et al., 2024; Oubari et al., 2024; Senellart & Allassonnière, 2025). Also, following Xiao & Bamler (2023), we will explore separate β values for $\boldsymbol{z}$ and $\boldsymbol{w}$ to regularize the shared and private latent spaces and control how information is allocated between them. Finally, future work includes a theoretical analysis of our results, showing that Hölder++ can isolate shared (content) from modality-specific (style) factors.

Impact Statement

This paper advances multimodal VAEs by improving the quality-coherence trade-off and by promoting disentanglement between shared (content) and modality-specific (style) latent factors. These advances may enhance the interpretability and controllability of multimodal representations. Improved multimodal generative models could benefit applications that require consistent cross-modal reasoning, such as medical image-report modeling. Nevertheless, in high-stakes domains such as healthcare, we advise practitioners to exercise caution when interpreting the inferred latent spaces, particularly with respect to causal claims, as the proposed models capture statistical dependencies between modalities rather than causal relationships. Moreover, when trained on sensitive data, practitioners should assess the model's privacy and robustness with respect to potential adversarial attacks aimed at extracting sensitive information. Beyond these considerations, we do not foresee negative societal consequences that are unique to this work, beyond those generally associated with the development and deployment of machine learning models.

References

Bouchacourt, D., Tomioka, R., and Nowozin, S. Multi-Level Variational Autoencoder: Learning Disentangled representations From Grouped Observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Bounoua, M., Franzese, G., and Michiardi, P. Multi-modal latent diffusion. *Entropy*, 26(4):320, 2024.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Daunhawer, I., Sutter, T. M., Marcinkevičs, R., and Vogt, J. E. Self-Supervised Disentanglement of Modality-Specific and Shared Factors Improves Multimodal Generative Models. In *DAGM German Conference on Pattern Recognition*, pp. 459–473. Springer, 2020.

Daunhawer, I., Sutter, T. M., Chin-Cheong, K., Palumbo, E., and Vogt, J. E. On the Limitations of Multimodal VAEs. In *International Conference on Learning Representations*, 2022.

Daunhawer, I., Bizeul, A., Palumbo, E., Marx, A., and Vogt, J. E. Identifiability Results for Multimodal Contrastive Learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Gao, L., Chen, W., Wang, D., Guo, F., and Liang, C. Disentangled Cross-Modal Representation Learning with Enhanced Mutual Supervision. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Garg, A., Jayram, T. S., Vaithyanathan, S., and Zhu, H. Generalized Opinion Pooling. *Annals of Mathematics and Artificial Intelligence*, 2004. URL https://api.semanticscholar.org/CorpusID:17872187.

Havtorn, J. D., Frellsen, J., Hauberg, S., and Maaløe, L. Hierarchical vaes know what they don't know. In *International Conference on Machine Learning*, pp. 4117–4128. PMLR, 2021.

Hernandez-Lobato, J., Li, Y., Rowland, M., Bui, T., Hernández-Lobato, D., and Turner, R. Black-Box Alpha Divergence Minimization. In *International Conference on Machine Learning*, pp. 1511–1520. PMLR, 2016.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Hotelling, H. The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, 1931.

Javaloy, A., Meghdadi, M., and Valera, I. Mitigating modality collapse in multimodal vaes via impartial optimization. In *International Conference on Machine Learning*, pp. 9938–9964. PMLR, 2022.

Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Koliander, G., El-Laham, Y., Djurić, P. M., and Hlawatsch, F. Fusion of Probability Density Functions. *Proceedings of the IEEE*, 110(4):404–453, 2022.

Lee, M. and Pavlovic, V. Private-Shared Disentangled Multimodal VAE for Learning of Hybrid Latent Representations. *arXiv preprint arXiv:2012.13024*, 2020.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov): 2579–2605, 2008.

Mancisidor, R. A., Jenssen, R., Yu, S., and Kampffmeyer, M. Aggregation of Dependent Expert Distributions in Multimodal Variational Autoencoders. *International Conference on Machine Learning*, 2025.

Oubari, F., El Baha, M., Meunier, R., Décatoire, R., and Mougeot, M. A markov random field multi-modal variational autoencoder. *CoRR*, 2024.

Palumbo, E., Daunhawer, I., and Vogt, J. E. MMVAE+: Enhancing the Generative Quality of Multimodal VAEs without Compromises. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Qiu, P., Zhu, W., Kumar, S., Chen, X., Yang, J., Sun, X., Razi, A., Wang, Y., and Sotiras, A. Multimodal Variational Autoencoder: A Barycentric View. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20060–20068, 2025.

Rainforth, T., Kosiorek, A., Le, T. A., Maddison, C., Igl, M., Wood, F., and Teh, Y. W. Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, pp. 4277–4285. PMLR, 2018.

Robert, C. P., Casella, G., and Casella, G. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

Roeder, G., Wu, Y., and Duvenaud, D. K. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. *Advances in Neural Information Processing Systems*, 30, 2017.

Senellart, A. and Allassonnière, S. Bridging the inference gap in mutimodal variational autoencoders. *arXiv preprint arXiv:2502.03952*, 2025.

Shi, Y., Paige, B., Torr, P., et al. Variational Mixture-of-Eexperts Autoencoders for Multi-Modal Deep Generative Models. *Advances in Neural Information Processing Systems*, 32, 2019.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. Ladder variational autoencoders. *Advances in Neural Information Processing Systems*, 29, 2016.

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

Tsai, Y.-H. H., Liang, P. P., Zadeh, A., Morency, L.-P., and Salakhutdinov, R. Learning Factorized Multimodal Representations. In *International Conference on Learning Representations*, 2019.

Tucker, G., Lawson, D., Gu, S., and Maddison, C. J. Doubly reparameterized gradient estimators for monte carlo objectives. In *International Conference on Learning Representations*, 2019.

Vahdat, A. and Kautz, J. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020.

Vo, H. K. and Valera, I. Hellinger Multimodal Variational Autoencoders. *International Conference on Artificial Intelligence and Statistics*, 2026.

Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-Supervised Learning With Data Augmentations Provably Isolates Content from Style. *Advances in Neural Information Processing Systems*, 34:16451–16467, 2021.

Wang, W., Yan, X., Lee, H., and Livescu, K. Deep Variational Canonical Correlation Analysis. *arXiv preprint arXiv:1610.03454*, 2016.

Wesego, D. and Rooshenas, P. Multimodal elbo with diffusion decoders. *arXiv preprint arXiv:2408.16883*, 2024a.

Wesego, D. and Rooshenas, P. Score-based multimodal autoencoder. *Transactions on Machine Learning Research*, 2024b.

Wolff, J., Krishnan, R. G., Ruff, L., Morshuis, J. N., Klein, T., Nakajima, S., and Nabi, M. Hierarchical multimodal variational autoencoders. 2022.

Wu, M. and Goodman, N. Multimodal Generative Models for Scalable Weakly-Supervised Learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Xiao, T. Z. and Bamler, R. Trading information between latents in hierarchical variational autoencoders. In *International Conference on Learning Representations*, 2023.

Yuan, S., Cui, J., Li, H., and Han, T. Learning multimodal latent generative models with energy-based prior. In *European Conference on Computer Vision*, pp. 86–100. Springer, 2024.

Zhang, Y., Shen, Y., and Wang, W. Disentanglement of Variations with Multimodal Generative Modeling. *arXiv preprint arXiv:2509.23548*, 2025.

Hölder++: Improving the Quality-Coherence Trade-off in Multimodal VAEs Supplementary Material

Table of Contents

The supplementary material is organized as follows. Section A provides detailed derivations of the Hölder, Hölder+, and Hölder++ objectives, and includes proofs that each objective corresponds to a valid ELBO. Section B provides additional experimental details, including descriptions of the datasets, evaluation criteria, and training setup (architectures and hyperparameters). Section C contains further empirical results, including runtime analysis, clustering consistency across models and random seeds, and additional qualitative and quantitative results on all datasets.

A. Proofs

A.1. Derivations of Hölder, Hölder+, and Hölder++

Hölder. Consider the Hölder pooling function with $\alpha = 0.5$, applied to the set of unimodal posteriors $\{q_{\boldsymbol{\phi}_j}(\boldsymbol{z}|\boldsymbol{x}_j)\}_{j=1}^M$. Each posterior is a multivariate Gaussian distributions with diagonal covariance matrix, $q_{\boldsymbol{\phi}_j}(\boldsymbol{z}|\boldsymbol{x}_j) = \mathcal{N}\left(\boldsymbol{z}; \boldsymbol{\mu}_j, \mathrm{diag}(\boldsymbol{\sigma}_j^2)\right)$, where $\boldsymbol{z} = (z_1, z_2, \ldots, z_D)^\top \in \mathbb{R}^D$ denotes the latent variable, $\boldsymbol{\mu}_j = (\mu_{j,1}, \mu_{j,2}, \ldots, \mu_{j,D})^\top \in \mathbb{R}^D$ the mean vector, and $\mathrm{diag}(\boldsymbol{\sigma}_j^2) = \mathrm{diag}(\sigma_{j,1}^2, \sigma_{j,2}^2, \ldots, \sigma_{j,D}^2)^\top \in \mathbb{R}^{D\times D}$ the covariance matrix. For simplicity, we denote the unimodal posterior by $q_j(\boldsymbol{z}) = q_{\boldsymbol{\phi}_j}(\boldsymbol{z}|\boldsymbol{x}_j)$, and the approximate joint posterior by $q(\boldsymbol{z}) = q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{X})$. The pooled density is then defined as

$$q(\boldsymbol{z}) = c\left(\sum_{j=1}^{M} \lambda_j \sqrt{q_j(\boldsymbol{z})}\right)^2,$$

where $c = 1/\int \left(\sum_{j=1}^{M} \lambda_j \sqrt{q_j(\boldsymbol{z})}\right)^2 d\boldsymbol{z}$. Since determining the weights λ_j in multimodal VAEs is generally nontrivial, we follow PoE and MoE and set them uniformly. We then obtain

$$q(\boldsymbol{z}) = c\left(\sum_{j=1}^{M} \sqrt{q_j(\boldsymbol{z})}\right)^2 = c\left(\sum_{j=1}^{M} q_j(\boldsymbol{z}) + 2\sum_{i=1}^{M}\sum_{j>i}^{M} \sqrt{q_i(\boldsymbol{z})q_j(\boldsymbol{z})}\right).$$

As $q_i(\boldsymbol{z})$ and $q_j(\boldsymbol{z})$ are multivariate Gaussian distributions with diagonal covariance matrices, they factorize across dimensions as $q_i(\boldsymbol{z}) = \prod_{d=1}^{D} q_{i,d}(z_d)$ and $q_j(\boldsymbol{z}) = \prod_{d=1}^{D} q_{j,d}(z_d)$, where $q_{i,d}(z_d)$ and $q_{j,d}(z_d)$ denote one-dimensional Gaussian distributions along coordinate z_d. Hence, we can express the geometric mean $\sqrt{q_i(\boldsymbol{z})q_j(\boldsymbol{z})}$ between them as

$$\sqrt{q_i(\boldsymbol{z})q_j(\boldsymbol{z})} = \sqrt{\prod_{d=1}^{D} q_{i,d}(z_d)q_{j,d}(z_d)} = \prod_{d=1}^{D} \sqrt{q_{i,d}(z_d)q_{j,d}(z_d)}.$$

Leveraging the derivation from Vo & Valera (2026), we obtain $\sqrt{q_{i,d}(z_d)q_{j,d}(z_d)}$ for each dimension $d \in \{1, 2, \ldots, D\}$ as

$$\sqrt{q_{i,d}(z_d)q_{j,d}(z_d)} = S_{ij}^d q_{ij,d}^{(1/2)}(z_d) = S_{ij}^d \mathcal{N}(z_d; \mu_{ij,d}, \sigma_{ij,d}^2), \quad \text{where}$$

$$S_{ij}^d = \sqrt{\frac{2\sigma_{i,d}\sigma_{j,d}}{\sigma_{i,d}^2 + \sigma_{j,d}^2}} \exp\left(-\frac{1}{4}\frac{(\mu_{i,d} - \mu_{j,d})^2}{\sigma_{i,d}^2 + \sigma_{j,d}^2}\right), \quad \mu_{ij,d} = \frac{\mu_{i,d}\sigma_{j,d}^2 + \mu_{j,d}\sigma_{i,d}^2}{\sigma_{i,d}^2 + \sigma_{j,d}^2}, \quad \sigma_{ij,d}^2 = \frac{2\sigma_{i,d}^2\sigma_{j,d}^2}{\sigma_{i,d}^2 + \sigma_{j,d}^2}.$$

Then, $\sqrt{q_i(\boldsymbol{z})q_j(\boldsymbol{z})}$ becomes

$$\sqrt{q_i(\boldsymbol{z})q_j(\boldsymbol{z})} = S_{ij} q_{ij}^{(1/2)}(\boldsymbol{z}) = S_{ij}\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_{ij}^2),$$

where $q_{ij}^{(1/2)}(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_{ij}^2), \boldsymbol{\mu}_{ij} = (\mu_{ij,1}, \mu_{ij,2}, \ldots, \mu_{ij,D})^\top, \boldsymbol{\sigma}_{ij}^2 = (\sigma_{ij,1}^2, \sigma_{ij,2}^2, \ldots, \sigma_{ij,D}^2)^\top$, and

$$S_{ij} = \prod_{d=1}^{D} S_{ij}^d = \prod_{d=1}^{D} \sqrt{\frac{2\sigma_{i,d}\sigma_{j,d}}{\sigma_{i,d}^2 + \sigma_{j,d}^2}} \exp\left(-\frac{1}{4}\frac{(\mu_{i,d} - \mu_{j,d})^2}{\sigma_{i,d}^2 + \sigma_{j,d}^2}\right).$$

Overall, we can see that the Hölder pooled density $q(\boldsymbol{z})$ is a weighted mixture of Gaussians

$$q(\boldsymbol{z}) = \sum_{j=1}^{M} \pi_j q_j(\boldsymbol{z}) + \sum_{i=1}^{M}\sum_{j>i}^{M} \pi_{ij} q_{ij}^{(1/2)}(\boldsymbol{z}), \tag{4}$$

where $\pi_j = c, \pi_{ij} = 2cS_{ij}$, and $c = \left(M + 2\sum_{i=1}^{M}\sum_{j>i}^{M} S_{ij}\right)^{-1}$.

Hölder objective. Given M modalities $\boldsymbol{X} := \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$, we consider the standard multimodal VAE generative model $p_\Theta(\boldsymbol{X}, \boldsymbol{z}) = p(\boldsymbol{z}) \prod_{j=1}^{M} p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z})$, and optimize a lower bound of the log-evidence $p_\Theta(\boldsymbol{X})$ by maximizing the following objective

$$\mathcal{L}_{\text{VAE}}(\boldsymbol{x}_{1:M}) = \mathbb{E}_{q_\Phi(\boldsymbol{z}|\boldsymbol{X})}\left[\log \frac{p_\Theta(\boldsymbol{X}, \boldsymbol{z})}{q_\Phi(\boldsymbol{z}|\boldsymbol{X})}\right], \tag{5}$$

As in MMVAE (Shi et al., 2019), we set $q_\Phi(\boldsymbol{z}|\boldsymbol{X})$ to the uniform mixture, $q_\Phi(\boldsymbol{z} \mid \boldsymbol{X}) = \frac{1}{M}\sum_{j=1}^{M} q_{\boldsymbol{\phi}_j}(\boldsymbol{z}|\boldsymbol{x}_j)$. Substituting into Eq. (5) and using linearity of expectation gives an equivalent decomposition into unimodal expectations, while keeping the denominator $q_\Phi(\boldsymbol{z}|\boldsymbol{X})$ as the pooled posterior:

$$\mathcal{L}_{\text{MMVAE}}(\boldsymbol{x}_{1:M}) = \frac{1}{M}\sum_{j=1}^{M} \mathbb{E}_{q_{\boldsymbol{\phi}_j}(\boldsymbol{z}|\boldsymbol{x}_j)}\left[\log \frac{p_\Theta(\boldsymbol{X}, \boldsymbol{z})}{q_\Phi(\boldsymbol{z}|\boldsymbol{X})}\right].$$

We instead define $q_\Phi(\boldsymbol{z}|\boldsymbol{X})$ via Hölder pooling in Eq. (4), which admits a mixture form

$$q_\Phi(\boldsymbol{z}|X) = \sum_{j=1}^{M} \pi_j \, q_{\boldsymbol{\phi}_j}(\boldsymbol{z}|\boldsymbol{x}_j) + \sum_{i=1}^{M}\sum_{j>i}^{M} \pi_{ij} \, q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{6}$$

with $\pi_j, \pi_{ij} \geq 0$ and $\sum_j \pi_j + \sum_{i=1}^{M}\sum_{j>i}^{M} \pi_{ij} = 1$. Substituting Eq. (6) into the ELBO 5 gives

$$\mathcal{L}_{\text{Hölder}}(\boldsymbol{x}_{1:M}) = \sum_{j=1}^{M} \pi_j \mathbb{E}_{q_{\boldsymbol{\phi}_j}(\boldsymbol{z}|\boldsymbol{x}_j)}\left[\log \frac{p_\Theta(\boldsymbol{X}, \boldsymbol{z})}{q_\Phi(\boldsymbol{z}|\boldsymbol{X})}\right] + \sum_{i=1}^{M}\sum_{j>i}^{M} \pi_{ij} \mathbb{E}_{q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i,\boldsymbol{x}_j)}\left[\log \frac{p_\Theta(\boldsymbol{X}, \boldsymbol{z})}{q_\Phi(\boldsymbol{z}|\boldsymbol{X})}\right], \tag{7}$$

which is a valid lower bound on the log-evidence $\log p_\Theta(\boldsymbol{X})$.

Hölder+. We extend the shared-latent model by introducing modality-specific latent variables $\boldsymbol{W} := \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M\}$, where each $\boldsymbol{w}_j$ captures private factors of modality $\boldsymbol{x}_j$. The generative model is

$$p_\Theta(\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{W}) = p(\boldsymbol{z}) \prod_{j=1}^{M} p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z}, \boldsymbol{w}_j)\, p(\boldsymbol{w}_j),$$

where we assume independent priors over the private latents $\{p(\boldsymbol{w}_j)\}_{j=1}^{M}$. For inference, we use a variational family that factorizes the joint posterior into a pooled encoder for the shared latent and unimodal encoders for the private latents

$$q_\Phi(\boldsymbol{z}, \boldsymbol{W}|\boldsymbol{X}) = q_{\Phi_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X})\, q_{\Phi_{\boldsymbol{W}}}(\boldsymbol{W}|\boldsymbol{X}) = q_{\Phi_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}) \prod_{j=1}^{M} q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j).$$

To avoid shortcut solutions in cross-modal generation, MMVAE+ (Palumbo et al., 2023) introduces auxiliary distributions over private latent variables for unobserved modalities when estimating cross-modal reconstruction terms. Concretely, when $\boldsymbol{z}$ is sampled from expert j, we draw $\boldsymbol{w}_j \sim q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j)$ for the observed modality and draw $\tilde{\boldsymbol{w}}_n \sim r_n(\boldsymbol{w}_n)$ for each modality $n \neq j$. The resulting objective can be written as

$$\mathcal{L}_{\text{MMVAE+}}(\boldsymbol{x}_{1:M}) = \frac{1}{M}\sum_{j=1}^{M} \mathbb{E}_{\substack{q_{\boldsymbol{\phi}_{\boldsymbol{z}_j}}(\boldsymbol{z}|\boldsymbol{x}_j) \\ q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j) \\ \{\tilde{\boldsymbol{w}}_n \sim r_n(\boldsymbol{w}_n)\}_{n\neq j}}} \log\left(\frac{p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z}, \boldsymbol{w}_j) p(\boldsymbol{z}) p(\boldsymbol{w}_j)}{q_{\boldsymbol{\phi}_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}) q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j)} \prod_{n\neq j} p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z}, \tilde{\boldsymbol{w}}_n)\right), \tag{8}$$

where $\{r_n(\boldsymbol{w}_n)\}_{n=1}^{M}$ are auxiliary distributions over modality-specific latents.

Following the same principle, we extend Hölder pooling to the shared-private setting and use auxiliary distributions for modality-specific latents that are *not* inferred from data in a given mixture component, thereby preventing shortcut learning when estimating reconstruction terms for unobserved modalities. Let $q_\Phi(\boldsymbol{z}|\boldsymbol{X})$ denote the Hölder-pooled posterior with unimodal weights π_j and pairwise weights π_{ij} over components $q_{\boldsymbol{\phi}_{\boldsymbol{z}_j}}(\boldsymbol{z}|\boldsymbol{x}_j)$ and $q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i, \boldsymbol{x}_j)$. The Hölder+ objective is

$$\begin{aligned}\mathcal{L}_{\text{Hölder+}}(\boldsymbol{x}_{1:M}) = &\sum_{j=1}^{M} \pi_j \mathbb{E}_{\substack{q_{\boldsymbol{\phi}_{\boldsymbol{z}_j}}(\boldsymbol{z}|\boldsymbol{x}_j) \\ q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j) \\ \{\tilde{\boldsymbol{w}}_n \sim r_n(\boldsymbol{w}_n)\}_{n\neq j}}} \log\left(\frac{p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z}, \boldsymbol{w}_j) p(\boldsymbol{z}) p(\boldsymbol{w}_j)}{q_{\boldsymbol{\phi}_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}) q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j)} \prod_{n\neq j} p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z}, \tilde{\boldsymbol{w}}_n)\right) \\ + \sum_{i=1}^{M}\sum_{j>i}^{M} \pi_{ij} \mathbb{E}_{\substack{q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i,\boldsymbol{x}_j) \\ q_{\boldsymbol{\phi}_{\boldsymbol{w}_i}}(\boldsymbol{w}_i|\boldsymbol{x}_i) \\ q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j) \\ \{\tilde{\boldsymbol{w}}_n \sim r_n(\boldsymbol{w}_n)\}_{n\notin\{i,j\}}}} &\log\left(\frac{p_{\boldsymbol{\theta}_i}(\boldsymbol{x}_i|\boldsymbol{z}, \boldsymbol{w}_i) p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z}, \boldsymbol{w}_j) p(\boldsymbol{z}) p(\boldsymbol{w}_i) p(\boldsymbol{w}_j)}{q_{\boldsymbol{\phi}_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}) q_{\boldsymbol{\phi}_{\boldsymbol{w}_i}}(\boldsymbol{w}_i|\boldsymbol{x}_i) q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j)} \prod_{n\notin\{i,j\}} p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z}, \tilde{\boldsymbol{w}}_n)\right).\end{aligned} \tag{9}$$

Hölder++. Since shared and modality-specific factors can be coupled in the posterior, we adopt a hierarchical variational posterior in which each private latent depends on both the observation and the shared latent, i.e., $q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j, \boldsymbol{z})$

$$q_\Phi(\boldsymbol{z}, \boldsymbol{W}|\boldsymbol{X}) = q_{\Phi_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X})\, q_{\Phi_{\boldsymbol{W}}}(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{z}) = q_{\Phi_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}) \prod_{j=1}^{M} q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j, \boldsymbol{z}). \tag{10}$$

From the Hölder+ objective in Eq. (9), substituting the hierarchical variational posterior in Eq. (10) into the ELBO yields the Hölder++ objective

$$\mathcal{L}_{\text{Hölder++}}(\boldsymbol{x}_{1:M}) = \sum_{j=1}^{M} \pi_j \mathbb{E}_{\substack{q_{\boldsymbol{\phi}_{\boldsymbol{z}_j}}(\boldsymbol{z}|\boldsymbol{x}_j) \\ q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j,\boldsymbol{z}) \\ \{\tilde{\boldsymbol{w}}_n \sim r_n(\boldsymbol{w}_n)\}_{n\neq j}}} \log \left(\frac{p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z},\boldsymbol{w}_j) p(\boldsymbol{z}) p(\boldsymbol{w}_j)}{q_{\boldsymbol{\phi}_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}) q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j,\boldsymbol{z})} \prod_{n\neq j} p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z},\tilde{\boldsymbol{w}}_n) \right)$$
$$+ \sum_{i=1}^{M}\sum_{j>i}^{M} \pi_{ij} \mathbb{E}_{\substack{q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i,\boldsymbol{x}_j) \\ q_{\boldsymbol{\phi}_{\boldsymbol{w}_i}}(\boldsymbol{w}_i|\boldsymbol{x}_i,\boldsymbol{z}) \\ q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j,\boldsymbol{z}) \\ \{\tilde{\boldsymbol{w}}_n \sim r_n(\boldsymbol{w}_n)\}_{n\notin\{i,j\}}}} \log \left(\frac{p_{\boldsymbol{\theta}_i}(\boldsymbol{x}_i|\boldsymbol{z},\boldsymbol{w}_i) p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z},\boldsymbol{w}_j) p(\boldsymbol{z}) p(\boldsymbol{w}_i) p(\boldsymbol{w}_j)}{q_{\boldsymbol{\phi}_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}) q_{\boldsymbol{\phi}_{\boldsymbol{w}_i}}(\boldsymbol{w}_i|\boldsymbol{x}_i,\boldsymbol{z}) q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j,\boldsymbol{z})} \prod_{n\notin\{i,j\}} p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z},\tilde{\boldsymbol{w}}_n) \right). \quad (11)$$

where the highlighted terms indicate the changes relative to Hölder+.

A.2. Lower-bound guarantee of the Hölder+ and Hölder++ objectives

Hölder+ objective. We recall the objective of Hölder from Eq. (7)

$$\mathcal{L}_{\text{Hölder}}(\boldsymbol{x}_{1:M}) = \sum_{j=1}^{M} \pi_j \mathbb{E}_{q_{\boldsymbol{\phi}_j}(\boldsymbol{z}|\boldsymbol{x}_j)} \left[\log \frac{p(\boldsymbol{z}) p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z}) \prod_{n\neq j} p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z})}{q_\Phi(\boldsymbol{z}|\boldsymbol{X})} \right]$$
$$+ \sum_{i=1}^{M}\sum_{j>i}^{M} \pi_{ij} \mathbb{E}_{q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i,\boldsymbol{x}_j)} \left[\log \frac{p(\boldsymbol{z}) p_{\boldsymbol{\theta}_i}(\boldsymbol{x}_i|\boldsymbol{z}) p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z}) \prod_{n\notin\{i,j\}} p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z})}{q_\Phi(\boldsymbol{z}|\boldsymbol{X})} \right]. \quad (12)$$

Inspired by the proof in MMVAE+ (Palumbo et al., 2023), for the first term corresponding to the unimodal component, we consider each term in the sum: when $\boldsymbol{z} \sim q_{\boldsymbol{\phi}_{\boldsymbol{z}_j}}(\boldsymbol{z}|\boldsymbol{x}_j)$ is sampled from the unimodal encoder, we use this $\boldsymbol{z}$ to compute the conditional likelihood of all M modalities, including the self-reconstruction term $\log p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z})$ and the cross-modal reconstruction terms $\log p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z})$ for $n \neq j$. Building on this, and using the modality-specific encoder $q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j)$, we derive the lower bound on the likelihood as

$$\log p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z}) \geq \mathbb{E}_{q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j)} \left[\log \frac{p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z},\boldsymbol{w}_j) p(\boldsymbol{w}_j)}{q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j)} \right], \quad \text{and} \quad (13)$$

$$\log p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z}) = \log \mathbb{E}_{\tilde{\boldsymbol{w}}_n \sim r_n(\boldsymbol{w}_n)} p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z},\tilde{\boldsymbol{w}}_n) \geq \mathbb{E}_{\tilde{\boldsymbol{w}}_n \sim r_n(\boldsymbol{w}_n)} \log p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z},\tilde{\boldsymbol{w}}_n), \quad (14)$$

where $r_n(\boldsymbol{w}_n)$ is an auxiliary prior distribution specific to each target modality $n \in \{1, 2, \ldots, M\}$, and the second step follows from Jensen's inequality.

The second term in Eq. (12) corresponding to a pairwise component (i, j) is analogously: when $\boldsymbol{z} \sim q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i, \boldsymbol{x}_j)$, we bound the reconstruction terms for modalities i and j using their modality-specific posteriors $q_{\boldsymbol{\phi}_{\boldsymbol{w}_i}}(\boldsymbol{w}_i|\boldsymbol{x}_i)$ and $q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j)$ as in Eq. (13), while for each modality $n \notin \{i, j\}$ we use $\tilde{\boldsymbol{w}}_n \sim r_n(\boldsymbol{w}_n)$ and apply Eq. (14). Plugging the expressions in Eqs. (13) and (14) into Eq. (12), we obtain the Hölder+ objective as

$$\mathcal{L}_{\text{Hölder+}}(\boldsymbol{x}_{1:M}) = \sum_{j=1}^{M} \pi_j \mathbb{E}_{\substack{q_{\boldsymbol{\phi}_{\boldsymbol{z}_j}}(\boldsymbol{z}|\boldsymbol{x}_j) \\ q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j) \\ \{\tilde{\boldsymbol{w}}_n \sim r_n(\boldsymbol{w}_n)\}_{n\neq j}}} \log \left(\frac{p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z},\boldsymbol{w}_j) p(\boldsymbol{z}) p(\boldsymbol{w}_j)}{q_{\boldsymbol{\phi}_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}) q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j)} \prod_{n\neq j} p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z},\tilde{\boldsymbol{w}}_n) \right)$$
$$+ \sum_{i=1}^{M}\sum_{j>i}^{M} \pi_{ij} \mathbb{E}_{\substack{q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i,\boldsymbol{x}_j) \\ q_{\boldsymbol{\phi}_{\boldsymbol{w}_i}}(\boldsymbol{w}_i|\boldsymbol{x}_i) \\ q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j) \\ \{\tilde{\boldsymbol{w}}_n \sim r_n(\boldsymbol{w}_n)\}_{n\notin\{i,j\}}}} \log \left(\frac{p_{\boldsymbol{\theta}_i}(\boldsymbol{x}_i|\boldsymbol{z},\boldsymbol{w}_i) p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z},\boldsymbol{w}_j) p(\boldsymbol{z}) p(\boldsymbol{w}_i) p(\boldsymbol{w}_j)}{q_{\boldsymbol{\phi}_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}) q_{\boldsymbol{\phi}_{\boldsymbol{w}_i}}(\boldsymbol{w}_i|\boldsymbol{x}_i) q_{\boldsymbol{\phi}_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j)} \prod_{n\notin\{i,j\}} p_{\boldsymbol{\theta}_n}(\boldsymbol{x}_n|\boldsymbol{z},\tilde{\boldsymbol{w}}_n) \right).$$

Thus, the Hölder+ objective is a valid ELBO. □

Hölder++ objective. Compared to Hölder+ (Eq. (9)), Hölder++ (Eq. (11)) uses a hierarchical inference model for the private latent, i.e., $q_{\phi_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j, \boldsymbol{z})$ instead of $q_{\phi_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j)$. This changes the lower bound for the self-reconstruction term

$$\begin{aligned}
\log p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z}) &= \log \int p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z}, \boldsymbol{w}_j) p(\boldsymbol{w}_j) d\boldsymbol{w}_j \\
&= \log \int q_{\phi_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j, \boldsymbol{z}) \frac{p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z}, \boldsymbol{w}_j) p(\boldsymbol{w}_j)}{q_{\phi_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j, \boldsymbol{z})} d\boldsymbol{w}_j \\
&= \log \mathbb{E}_{q_{\phi_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j, \boldsymbol{z})} \left[\frac{p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z}, \boldsymbol{w}_j) p(\boldsymbol{w}_j)}{q_{\phi_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j, \boldsymbol{z})} \right] \\
&\geq \mathbb{E}_{q_{\phi_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j, \boldsymbol{z})} \left[\log \frac{p_{\boldsymbol{\theta}_j}(\boldsymbol{x}_j|\boldsymbol{z}, \boldsymbol{w}_j) p(\boldsymbol{w}_j)}{q_{\phi_{\boldsymbol{w}_j}}(\boldsymbol{w}_j|\boldsymbol{x}_j, \boldsymbol{z})} \right],
\end{aligned}$$

where the last step follows from Jensen's inequality. Thus, the Hölder++ objective is a valid ELBO. □

A.3. Tighter variational lower bound via IWAE and DReG estimators

The importance weighted autoencoder (IWAE). IWAE (Burda et al., 2015) provides a *tighter* variational lower bound than the ELBO in Eq. (5) by using a properly weighted multi-sample importance estimator, given by

$$\mathcal{L}_{\text{IWAE}}(\boldsymbol{x}_{1:M}) = \mathbb{E}_{\boldsymbol{z}^{1:K} \sim q_\Phi(\boldsymbol{z}|\boldsymbol{x}_{1:M})} \left[\log \sum_{k=1}^{K} \frac{1}{K} \frac{p_\Theta(\boldsymbol{X}, \boldsymbol{z}^k)}{q_\Phi(\boldsymbol{z}^k|\boldsymbol{X})} \right], \tag{15}$$

with K is the number of samples. In multimodal VAEs, IWAE estimator is often preferred because it typically yields higher-entropy variational posteriors, which is beneficial multimodal settings where each unimodal encoder should allocate probability mass to latent regions that explain other modalities. In MMVAE (Shi et al., 2019), under a MoE joint posterior, they extend $\mathcal{L}_{\text{IWAE}}$ in Eq. (15) via stratified sampling (Robert et al., 1999) over the M modalities as follows

$$\mathcal{L}^{\text{MoE}}_{\text{IWAE}}(\boldsymbol{x}_{1:M}) = \frac{1}{M} \sum_{j=1}^{M} \mathbb{E}_{\boldsymbol{z}^{1:K} \sim q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j)} \left[\log \sum_{k=1}^{K} \frac{1}{K} \frac{p_\Theta(\boldsymbol{X}, \boldsymbol{z}^k)}{q_\Phi(\boldsymbol{z}^k|\boldsymbol{X})} \right],$$

which is a valid ELBO. In our case, under a Hölder mixture with $\alpha = 0.5$ in Eq. (6), we extend $\mathcal{L}_{\text{IWAE}}$ in Eq. (15) via stratified sampling over the M modalities to obtain $\mathcal{L}^{\text{Hölder}}_{\text{IWAE}}$ as follows

$$\begin{aligned}
\mathcal{L}^{\text{Hölder}}_{\text{IWAE}}(\boldsymbol{x}_{1:M}) &= \sum_{j=1}^{M} \pi_j \mathbb{E}_{\boldsymbol{z}^{1:K} \sim q_{\phi_j}(\boldsymbol{z}|\boldsymbol{x}_j)} \left[\log \sum_{k=1}^{K} \frac{1}{K} \frac{p_\Theta(\boldsymbol{X}, \boldsymbol{z}^k)}{q_\Phi(\boldsymbol{z}^k|\boldsymbol{X})} \right] \\
&+ \sum_{i=1}^{M} \sum_{j>i}^{M} \pi_{ij} \mathbb{E}_{\boldsymbol{z}^{1:K} \sim q_{ij}^{(1/2)}(\boldsymbol{z}|\boldsymbol{x}_i, \boldsymbol{x}_j)} \left[\log \sum_{k=1}^{K} \frac{1}{K} \frac{p_\Theta(\boldsymbol{X}, \boldsymbol{z}^k)}{q_\Phi(\boldsymbol{z}^k|\boldsymbol{X})} \right],
\end{aligned}$$

which remains a valid lower bound and is tighter than the ELBO.

The doubly reparametrised gradient estimator (DReG). As mentioned by Roeder et al. (2017); Rainforth et al. (2018), the standard IWAE gradient estimator can exhibit undesirably high variance. Tucker et al. (2019) address this by re-applying the reparameterization trick to obtain the doubly reparameterized gradient (DReG) estimator. Following prior multimodal VAE work that optimizes multi-sample objectives with DReG (Shi et al., 2019; Palumbo et al., 2023; 2024), we adopt the same estimator in our experiments for Hölder-based models.

B. Experimental details

B.1. Datasets

PolyMNIST. Sutter et al. (2021) introduced the dataset as an MNIST extension with diverse backgrounds. Each sample overlays an MNIST digit on a randomly selected 28×28 crop from each of five background images, yielding a five-modality benchmark. The digit label is the shared (semantic) factor, while background content and handwriting style are modality-specific. The dataset contains 60 000 training and 10 000 test images.

MNIST-SVHN. Shi et al. (2019) constructed a bimodal dataset by pairing MNIST and SVHN such that the two modalities in each pair depict the same digit class, aiming to separate conceptual complexity (digit) from perceptual complexity (e.g., color, style, size). To increase cross-domain correspondences, each instance in either dataset is randomly paired with 20 instances of the same class from the other dataset. The dataset contains $56\,068$ training and $10\,000$ test images.

CUBICC. Palumbo et al. (2024) introduced CUBICC, a variant of CUB Image-Captions in which each datapoint is a paired of bird image and caption. To obtain a realistic multimodal clustering benchmark, they merge fine-grained bird sub-species into coarser species-level classes, increasing intra-class variability while preserving shared semantic structure across modalities. The benchmark contains $13\,131$ image-caption pairs, split into $11\,834$ train, 638 validation, and 659 test samples, covering 22 sub-species grouped into 8 species (Blackbird, Gull, Jay, Oriole, Tanager, Tern, Warbler, Wren).

B.2. Evaluation criteria

Generative coherence. To evaluate how well our model imputes missing modalities, we use classifier-based coherence. We train a classifier for each modality on its training samples. For unconditional coherence, we generate full multimodal samples from the prior and compute the fraction with consistent predicted labels across modalities. For conditional coherence, we condition on every non-empty subset (excluding the target modality), generate the target, report the fraction whose predicted label matches the ground truth. We report coherence averaged over all subsets and target modalities.

Generative quality. We use the Fréchet Inception Distance (FID) (Heusel et al., 2017) to quantify sample quality. The metric extracts features with a pretrained Inception network and measures the distance between the real and generated feature distributions. Lower FID implies higher visual realism and closer match to the data distribution.

Disentanglement metrics. To assess disentanglement between the shared and private subspaces, we evaluate both (i) the inferred latent representations and (ii) the generated images. For (i), we follow prior work and train a linear classifier on the latent codes, reporting classification accuracy. We expect high accuracy for the shared latent $\boldsymbol{z}$ and low accuracy for the private latent $\boldsymbol{w}$, since $\boldsymbol{w}$ should not encode class information. For (ii), we introduce **three generation-based metrics**: $\boldsymbol{z}$ content stability (↑), $\boldsymbol{z}$ content accuracy (↑), and $\boldsymbol{w}$ content accuracy (↓). To compute the first two, we fix $\boldsymbol{z}$, sample multiple $\boldsymbol{w} \sim p(\boldsymbol{w})$, decode, and classify the outputs. $\boldsymbol{z}$ content stability measures *pairwise label agreement* across generated samples, while $\boldsymbol{z}$ content accuracy measures correctness with respect to the ground-truth label. High values indicate that content is encoded primarily in $\boldsymbol{z}$: fixing $\boldsymbol{z}$ yields outputs with correct and invariant content despite changes in $\boldsymbol{w}$. For $\boldsymbol{w}$ content accuracy, we fix $\boldsymbol{w}$, sample multiple $\boldsymbol{z} \sim p(\boldsymbol{z})$, decode, and compute classification accuracy; lower values indicate that $\boldsymbol{w}$ carries little content information. All three metrics are averaged over self- and cross-generation. *Note that for CUBICC we evaluate generation only when the target is an image (image → image and text → image), whereas for MNIST-SVHN we evaluate generation for both modalities as targets.*

Downstream clustering metrics. We evaluate clustering performance using three standard metrics: clustering accuracy (ACC), normalized mutual information (NMI), and the adjusted rand index (ARI). ACC reports the best label-matching accuracy after permuting cluster IDs via the Hungarian algorithm. NMI measures the shared information between predicted clusters and ground-truth labels, and ARI quantifies partition similarity based on pairwise assignments, corrected for chance. Higher values indicate better clustering for all three metrics.

B.3. Experimental setups

We train all models with the reparameterization trick (Kingma & Welling, 2013) and optimize using Adam (Kingma, 2014) on NVIDIA A100-PCIE-80GB GPUs. Following prior work, we weight the KL term in the ELBO by a coefficient β (Higgins et al., 2017), i.e., $\beta\mathrm{KL}(q_\Phi(\boldsymbol{z}|\boldsymbol{X})\|p(\boldsymbol{z}))$, and select β via cross-validation over $\{1.0, 2.5, 5.0, 10.0\}$ for PolyMNIST and $\{1.0, 2.5, 5.0\}$ for MNIST-SVHN and CUBICC. For DCMEM, we choose the method-specific parameter α over $\{0.1, 0.5, 1.0\}$. The best value for each dataset is reported in Table 6, which also includes the optimal α for DCMEM. Across datasets, we use an isotropic Gaussian prior and model unimodal posteriors as diagonal-covariance Gaussians. We weight modality likelihood terms by relative dimensionality: the dominant modality is set to 1.0 and the remaining modalities are scaled proportionally to their data dimensions (Shi et al., 2019; Sutter et al., 2021; Javaloy et al., 2022). We report mean $\pm$ standard deviation over 3 seeds, except over 10 seeds for CUBICC. For baselines, we use the original implementations: MMVAE+, CMVAE and open source MultiVAE (Senellart & Allassonnière, 2025). *Note that in figures and tables reporting a single value per model, we select the value at the best-performing β (and α for DCMEM). In contrast, the quality-coherence trade-off figures plot results for all β and α values listed above. The only exception is Figure 5, which reports multiple seeds at the best hyperparameters.*

Table 6. Optimal β for all models (except DCMEM) and optimal α for DCMEM on all three datasets.

	PolyMNIST	MNIST-SVHN	CUBICC
DMVAE		1.0	
DCMEM		1.0	1.0
MMVAE+	2.5	2.5	1.0
CMVAE	2.5	2.5	1.0
Hölder+	5.0	5.0	1.0
CHölder+			1.0
MMVAE++			1.0
CMVAE++			1.0
Hölder++	5.0	5.0	1.0
CHölder++			1.0

PolyMNIST. The encoder and decoder architectures follow Palumbo et al. (2023), with ResNet backbones for all image modalities. We model each modality with a Laplace likelihood and use diagonal Gaussian priors and posteriors; for MMVAE, MMVAE+, and CMVAE we adopt Laplace priors and posteriors, consistent with their original setups. All single-shared-latent baselines are trained for 500 epochs with latent dimensionality 512, and batch size 256. For MMVAE, we set the latent dimensionality to 160. For MMVAE+, CMVAE, Hölder+, and Hölder++, we use separate shared and modality-specific subspaces of 32 dimensions each. We train MMVAE, MMVAE+, CMVAE with $K = 1$ for 150, 150, 250 epochs, respectively and batch size 256, whereas Hölder+ and Hölder++ are trained with a multi-sample objective with $K = 10$ for 50 epochs and batch size 32. All models are trained with learning rate $5e^{-4}$.

MNIST-SVHN. The encoder and decoder architectures follow Gao et al. (2025), using ResNet backbones for all image modalities. We model each modality with a Laplace likelihood and use diagonal Gaussian priors and posteriors by default; for MMVAE, MMVAE+, and CMVAE, we instead adopt Laplace priors and posteriors. All single-shared-latent baselines are trained for 100 epochs with latent dimensionality 20. For DMVAE, MMVAE+, CMVAE, Hölder+, and Hölder++, we use separate shared and modality-specific latent subspaces of 10 and 4 dimensions, respectively. We train MMVAE+ and CMVAE with $K = 1$ for 30 epochs. MMVAE, Hölder+, and Hölder++ are trained with a multi-sample objective $K = 10$ for 10 epochs, DMVAE, which is not a mixture-based model, is also trained for 10 epochs. For DCMEM, we follow the original setup with 32-dimensional shared and private subspaces and batch size 64. Finally, all models are trained with batch size 100 and learning rate $5e^{-4}$. *Note that for this dataset we do not rescale likelihood terms by modality dimensionality; instead, we set the likelihood weights to* 50.0 *for MNIST and* 1.0 *for SVHN.*

CUBICC. The encoder and decoder architectures follow Palumbo et al. (2024), using a ResNet backbone for images and a convolutional network for text. We model each modality with a Laplace likelihood and use diagonal Gaussian priors and posteriors by default. All models, except DCMEM, are trained with a multi-sample objective with $K = 10$ for 300 epochs and batch size 32; DCMEM is trained for 200 epochs with batch size 64. All models use a learning rate of $1e^{-4}$ and separate shared and modality-specific latent subspaces of 64 and 32 dimensions, respectively. Moreover, for models that use a clustering prior on $\boldsymbol{z}$, we set the number of clusters to a predefined value of 35.

C. Additional results

C.1. Computational analysis: Runtime comparison across models

In Table 7, we report the average per-batch training time on PolyMNIST, which has 5 modalities ($M = 5$). As observed, HELVAE has the shortest per-batch time. Comparing MMVAE-based and Hölder-based models, Hölder-based models require more computation per batch, consistent with their complexity: our pooling rule includes pairwise components with $\mathcal{O}(M^2)$ terms, compared to $\mathcal{O}(M)$ for MoE-mixture baselines. However, our models typically require fewer epochs to converge, making the total training time comparable to other methods while providing a more expressive mixture without a substantial increase in overall cost. Finally, Hölder+ and Hölder++ have nearly identical per-batch times, indicating that introducing hierarchical inference incurs negligible additional computational cost.

C.2. Additional results: Clustering performance consistency across models

As shown in Table 4, we observe large variance in CMVAE and DCMEM. To visualize this, we plot 10 different seeds for each model under its best configuration. Figure 5 shows that CMVAE and DCMEM are widely spread in the plot, indicating

Table 7. Training batch time (s) on PolyMNIST with five modalities. We group models by latent structure: shared-only vs. shared-private.

Model	Training batch time	Batch size	Epochs
MVAE	0.1887	256	500
MoPoE	0.1216	256	500
HELVAE	0.0895	256	500
DMVAE	0.3071	256	500
MMVAE ($K = 1$)	0.2362	256	150
Hölder ($K = 1$)	0.4007	256	50
MMVAE+ ($K = 1$)	0.2884	256	150
CMVAE ($K = 1$)	0.2803	256	250
Hölder+ ($K = 1$)	0.4789	256	50
Hölder++ ($K = 1$)	0.4759	256	50

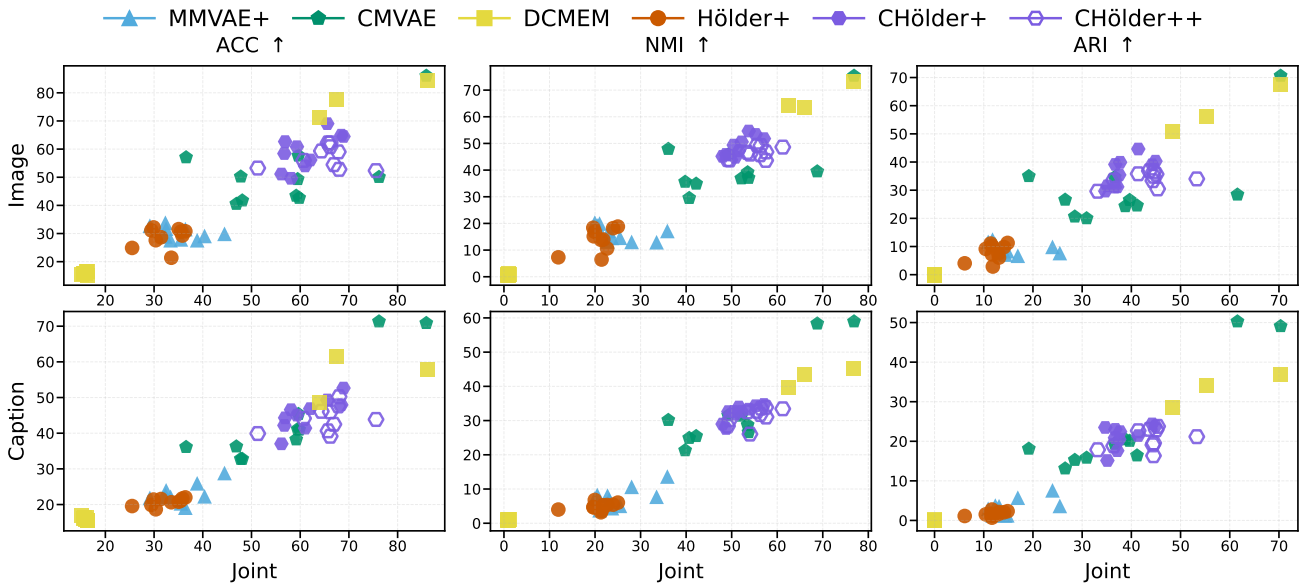


Figure 5. Clustering performance on CUBICC using latent representations, with each model evaluated at its best configuration. Per model, each point corresponds to a different seed (10 seeds total). The optimal region is the upper-right in both plots.

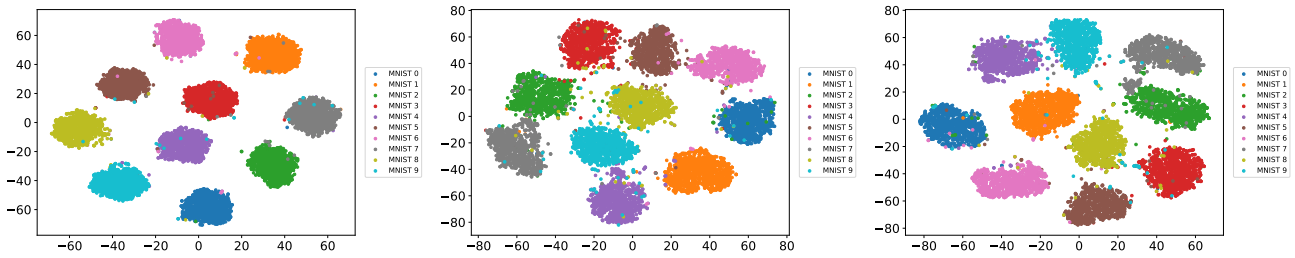
sensitivity to random seeds and thus producing inconsistent results. In contrast, our Hölder-based models are more robust to initialization and regularization, provide consistent results, and achieve the best overall clustering performance.

C.3. Additional results: PolyMNIST

Qualitative results. Figure 8 shows cross-modal generation from the first to the third modality, with five samples obtained by varying only the modality-specific latent variables. As expected on PolyMNIST, preserving shared semantic content is straightforward: most models keep the digit identity consistent across samples. However, MMVAE+ and CMVAE still misclassify the digit for challenging cases (4 and 9, respectively). In contrast, our Hölder+ and Hölder++ preserve the correct digit information while producing samples with different styles when changing the private factors $\boldsymbol{w}$.

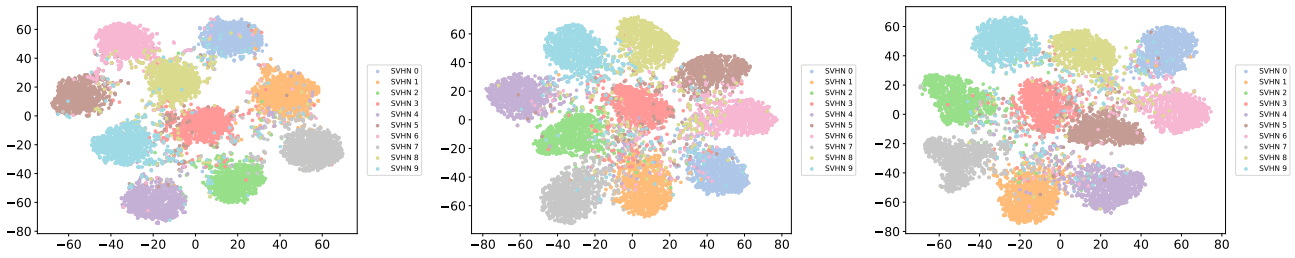
C.4. Additional results: MNIST-SVHN

Latent representation visualization. Figures 6 and 7 show t-SNE visualizations (Maaten & Hinton, 2008) of MNIST and SVHN latent representations on the test set. Thanks to hierarchical inference, Hölder++ yields more separable class clusters than Hölder+. Although our models are slightly less clearly separated than DCMEM on the SVHN representation, Hölder+ and Hölder++ still effectively separate classes in the latent space, comparably to DCMEM, without introducing additional regularization terms–purely through architectural design.



(a) DCMEM *(b)* Hölder+ *(c)* Hölder++

Figure 6. t-SNE visualization of MNIST latent representations for DCMEM, Hölder+, and Hölder++ on the test set of MNIST-SVHN.



(a) DCMEM *(b)* Hölder+ *(c)* Hölder++

Figure 7. t-SNE visualization of SVHN latent representations for DCMEM, Hölder+, and Hölder++ on the test set of MNIST-SVHN.

Qualitative results. Figure 9 shows cross-modal generation from SVHN to MNIST, with five samples obtained by varying only the modality-specific latent variables. DMVAE suffers from shortcut behavior as the generated digit changes when the private factor changes. MMVAE+ and CMVAE preserve the digit, but the resulting samples are less diverse and lower quality. For DCMEM, when the SVHN input has a cluttered background (multiple digits), it struggles to predict the correct label. In contrast, our Hölder-based models consistently predict the correct label, and the improvement from Hölder+ to Hölder++ suggests that hierarchical inference better isolates shared information, yielding generations with higher quality, more diversity, and better accuracy. Similarly, Figure 9 shows cross-modal generation from MNIST to SVHN, with five samples obtained by varying the private latents $\boldsymbol{w}$. MMVAE+ and CMVAE perform poorly in this setting, whereas DCMEM and our models (Hölder+ and Hölder++) preserve the digit as $\boldsymbol{w}$ changes. This qualitative behavior is consistent with the coherence-quality trade-off in Figure 4 and the disentanglement metrics in Table 2.

C.5. Additional results: CUBICC

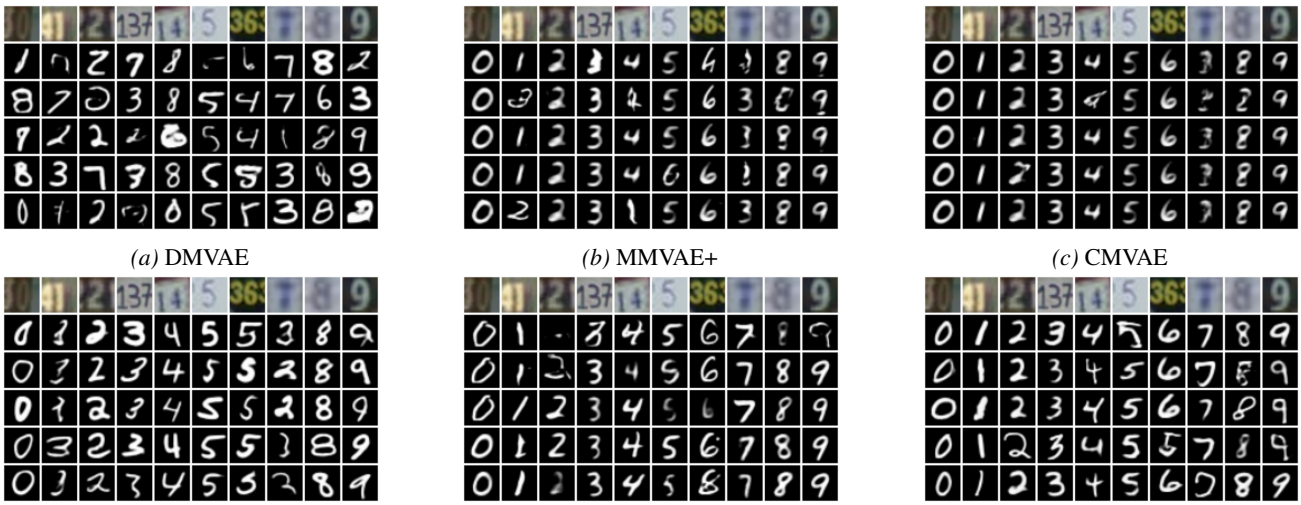
Effect of hierarchical inference on clustering performance across models. As shown in Table 5, hierarchical inference improves disentanglement metrics across models. Table 8 reports the corresponding downstream clustering performance in the same setting. For MMVAE+ and Hölder+, hierarchical inference does not affect clustering performance. While CMVAE is highly sensitive to random seeds under the original architecture, adding hierarchical inference (CMVAE++) substantially reduces variance across seeds, leading to more consistent performance. Overall, introducing a hierarchical posterior improves disentanglement between shared and private latent variables without hurting downstream performance, and CHölder+ and CHölder++ remain the best-performing models for clustering.

Disentanglement of shared and private subspaces in the latent representations. Table 9 shows classification accuracy using latent representations from the image and caption modalities. The performance of CMVAE, CHölder+, and CHölder++ is comparable, with each achieving the best or second-best results in this setting.

Qualitative results. Figure 11 shows the captions when conditioning on images. We see that Hölder+ improves caption quality over MMVAE+, producing more reasonable text relative to the image. Hölder++ also performs well in this setting.

(a) MMVAE+ *(b)* CMVAE *(c)* Hölder+

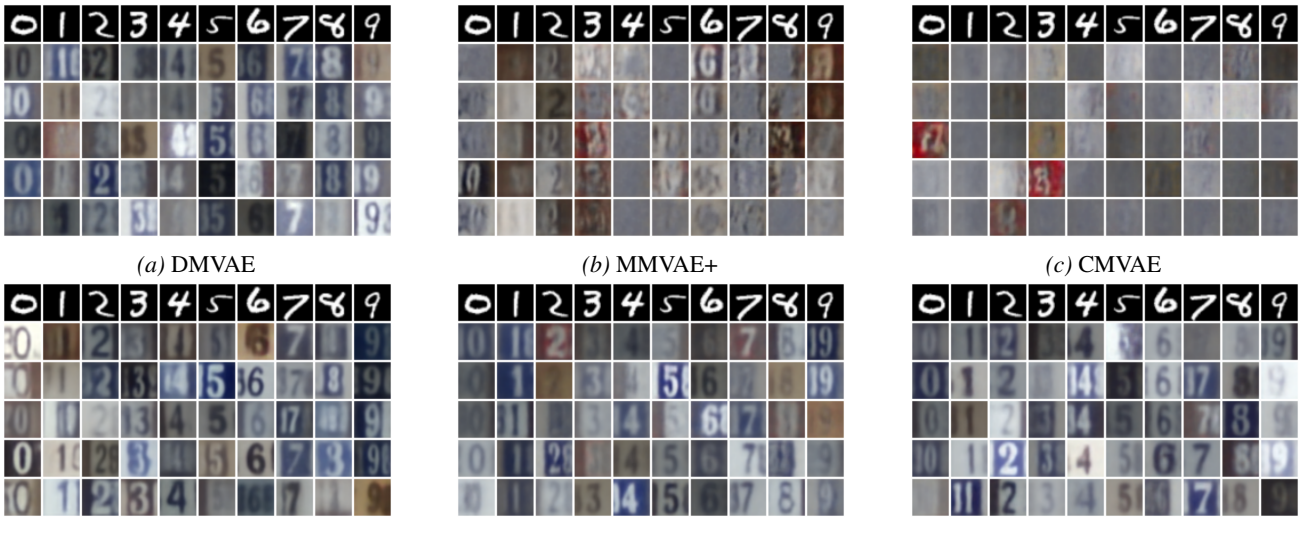
Figure 8. Five samples of the third modality conditioned on the first modality on PolyMNIST, generated by varying only the modality-specific latent variables. Each column corresponds to a ground-truth digit label from 0 to 9, so all samples within a column share the same digit information. As expected, all three models preserve the class label while changing the private factors.



(a) DMVAE *(b)* MMVAE+ *(c)* CMVAE

(d) DCMEM *(e)* Hölder+ *(f)* Hölder++

Figure 9. Five MNIST samples generated from SVHN on MNIST-SVHN by varying only the modality-specific latent variables. Each column corresponds to a ground-truth digit label from 0 to 9, so all samples within a column share the same digit information.



(a) DMVAE *(b)* MMVAE+ *(c)* CMVAE

(d) DCMEM *(e)* Hölder+ *(f)* Hölder++

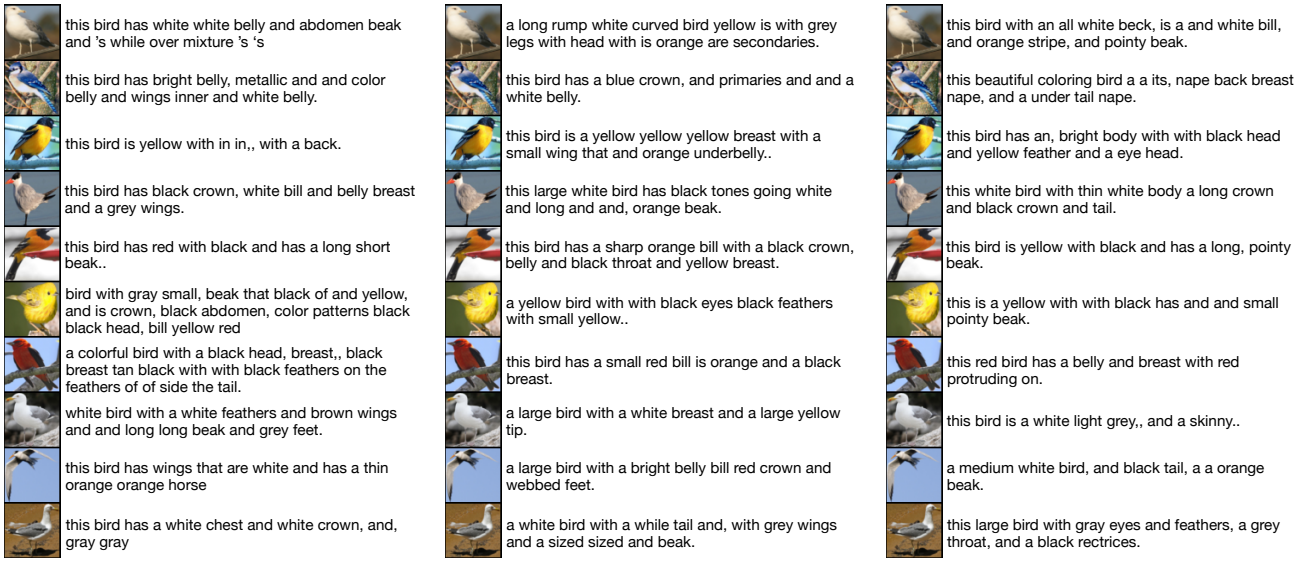
Figure 10. Five SVHN samples generated from MNIST on MNIST-SVHN by varying only the modality-specific latent variables. Each column corresponds to a ground-truth digit label from 0 to 9, so all samples within a column share the same digit information.

Table 8. Effect of hierarchical inference on clustering performance across models on CUBICC. We partition models according to whether $\boldsymbol{z}$ follows a structured clustering prior. The best and second-best results are marked bold and underlined, respectively.

	Image Representation (↑)			Caption Representation (↑)			Joint Representation (↑)		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
MMVAE+	30.1 ± 2.1	16.2 ± 2.8	9.2 ± 1.9	22.9 ± 2.6	7.6 ± 2.8	3.3 ± 2.0	35.5 ± 4.4	25.4 ± 5.2	15.7 ± 4.8
MMVAE++	29.4 ± 3.0	15.1 ± 2.4	8.3 ± 1.8	20.9 ± 2.0	4.8 ± 1.5	1.7 ± 0.9	36.1 ± 2.7	24.2 ± 2.2	15.1 ± 1.7
Hölder+	28.8 ± 3.2	14.0 ± 4.3	7.9 ± 2.8	20.7 ± 1.0	4.9 ± 1.0	1.7 ± 0.6	32.3 ± 3.4	20.7 ± 3.4	11.8 ± 2.3
Hölder++	28.3 ± 2.6	14.3 ± 2.4	7.7 ± 1.8	19.7 ± 1.1	3.9 ± 0.9	1.2 ± 0.5	31.8 ± 2.4	18.8 ± 3.0	10.9 ± 2.3
CMVAE	51.9 ± 12.8	42.2 ± 12.1	31.1 ± 14.0	44.6 ± 13.8	**33.7 ± 12.8**	**23.8 ± 13.1**	58.0 ± 13.8	51.3 ± 12.4	39.3 ± 14.9
CMVAE++	43.6 ± 3.0	33.7 ± 3.8	22.6 ± 2.7	37.7 ± 4.1	25.3 ± 3.9	15.9 ± 3.3	53.2 ± 6.3	45.0 ± 4.7	32.6 ± 4.8
CHölder+	**59.1 ± 6.1**	**48.8 ± 3.4**	**36.2 ± 4.8**	**45.3 ± 4.2**	32.3 ± 2.2	21.3 ± 2.8	61.3 ± 4.6	51.7 ± 2.8	38.6 ± 3.4
CHölder++	57.3 ± 3.7	46.4 ± 2.1	34.1 ± 2.3	44.0 ± 3.4	31.4 ± 2.4	20.5 ± 2.4	**65.3 ± 5.8**	**55.1 ± 3.5**	**43.2 ± 5.2**

Table 9. Bird-species classification accuracy on the latent representations for CUBICC, evaluated on the private $\boldsymbol{w}$, shared $\boldsymbol{z}$, and joint $[\boldsymbol{w}, \boldsymbol{z}]$ subspaces. The best and second-best results are highlighted in bold and underlined, respectively.

	Image Representation			Caption Representation		
	Joint (↑)	Shared (↑)	Private (↓)	Joint (↑)	Shared (↑)	Private (↓)
MMVAE+	0.798 ± 0.052	0.796 ± 0.045	0.169 ± 0.057	0.630 ± 0.076	0.612 ± 0.092	0.231 ± 0.050
DCMEM	0.571 ± 0.182	0.373 ± 0.309	0.446 ± 0.032	0.470 ± 0.114	0.289 ± 0.219	0.390 ± 0.042
Hölder+	**0.827 ± 0.058**	0.783 ± 0.087	0.311 ± 0.059	0.620 ± 0.077	0.592 ± 0.096	0.237 ± 0.047
Hölder++	0.768 ± 0.014	0.758 ± 0.017	0.173 ± 0.027	0.559 ± 0.020	0.546 ± 0.023	0.207 ± 0.014
CMVAE	0.811 ± 0.048	**0.810 ± 0.046**	**0.165 ± 0.039**	**0.635 ± 0.069**	**0.622 ± 0.080**	0.233 ± 0.041
CHölder+	0.820 ± 0.012	0.790 ± 0.017	0.234 ± 0.029	0.590 ± 0.018	0.561 ± 0.014	0.251 ± 0.022
CHölder++	0.801 ± 0.025	0.800 ± 0.023	**0.165 ± 0.024**	0.582 ± 0.011	0.574 ± 0.010	**0.200 ± 0.014**



(a) MMVAE+ *(b)* Hölder+ *(c)* Hölder++

Figure 11. Qualitative results for MMVAE+, Hölder+ and Hölder++ for image-to-caption generation on CUBICC.