Deep Generative Clustering with Multimodal Variational Autoencoders

Emanuele Palumbo [1 2] **Sonia Laguna** [2] **Daphné Chopard** [2] **Julia E. Vogt** [2]

Abstract

Multimodal VAEs have recently received significant attention as generative models for weakly-supervised learning with multiple heterogeneous modalities. In parallel, VAE-based methods have been explored as probabilistic approaches for clustering tasks. Our work lies at the intersection of these two research directions. We propose a novel multimodal VAE model, in which the latent space is extended to learn data clusters, leveraging shared information across modalities. Our experiments show that our proposed model improves generative performance over existing multimodal VAEs, particularly for unconditional generation. Furthermore, our method favorably compares to alternative clustering approaches, in weakly-supervised settings. Notably, we propose a post-hoc procedure that avoids the need for our method to have a priori knowledge of the true number of clusters, mitigating a critical limitation of previous clustering frameworks.

1. Introduction

Multimodal VAEs are powerful generative models for weakly-supervised learning with multiple modalities, with relevant applications in segmentation tasks or data integration (Lee & van der Schaar, 2021; Dorent et al., 2019) in the healthcare domain. They have the advantage of being able to handle numerous modalities efficiently, thanks to recently proposed scalable approaches (Wu & Goodman, 2018; Shi et al., 2019; Sutter et al., 2020; 2021; Hwang et al., 2021; Palumbo et al., 2023). Multimodal VAEs have been explored for generative tasks, succeeding in cross-modal generation (Hwang et al., 2021; Palumbo et al., 2023) despite less remarkable results in unconditional generation (Palumbo et al., 2023).

A parallel line of research focuses on VAE-based generative methods for clustering tasks (Jiang et al., 2016; Dilokthanakul et al., 2016). In particular deep variational clustering methods have been investigated in medical scenarios (Manduchi et al., 2022), being particularly suitable as they enable integration of domain knowledge from clinicians through prior probabilities (Manduchi et al., 2021). Other works use VAE-based methods to learn interpretable representations for clustering time-series (Fortuin et al., 2019).

With this work, we position ourselves at the intersection of these two lines of research by proposing a variational generative approach for clustering in a multimodal setting. In particular, we introduce a novel multimodal VAE model, called Clustering Multimodal VAE (CMVAE). The proposed model incorporates recent advancements for multimodal VAEs (Palumbo et al., 2023), and is designed to effectively model data clusters in the latent space. In our experiments, we show that our method represents an improvement over existing multimodal VAEs, particularly for unconditional generation, where existing methods struggle to achieve satisfactory results. Importantly, we evaluate the performance of CMVAE for clustering in weakly-supervised settings, highlighting its effectiveness where unimodal approaches fail to achieve adequate performance. In a realistic setting, our approach shows significant improvements compared to alternative scalable weakly-supervised methods. Furthermore, unlike most existing approaches whose performance relies on the knowledge of the true number of clusters at training time, CMVAE can effectively infer the number of clusters at test time.

2. Related work

Multimodal VAEs Multimodal VAEs extend the well-known VAE framework (Kingma & Welling, 2014) to handle data consisting of multiple modalities, leveraging the pairing across modalities as weak supervision. While early approaches (Suzuki et al., 2017; Vedantam et al., 2018) faced scalability challenges due to inference requiring a separate encoder network for each subset of modalities, more recent scalable approaches (Wu & Goodman, 2018; Shi et al., 2019; Sutter et al., 2020; 2021; Hwang et al., 2021) assume the joint encoder decomposes in terms of unimodal encoders. Despite promising applications (Lee & van der Schaar, 2021; Dorent et al., 2019), recent work (Daunhawer

[1]ETH AI Center, Zürich, Switzerland [2]Department of Computer Science, ETH Zürich, Switzerland. Correspondence to: Emanuele Palumbo <emanuele.palumbo@ai.ethz.ch>.

Accepted to ICML workshop on Structured Probabilistic Inference & Generative Modeling, Honolulu, Hawaii, USA. 2023.

et al., 2022) shows that the main formulations of multimodal VAEs suffer from important limitations, involving a trade-off between generative quality (the similarity of generated samples to real ones) and generative coherence (the semantic consistency across modalities). Previous attempts to enhance the performance of multimodal VAEs involved additional regularization terms (Sutter et al., 2020; Hwang et al., 2021), or mutual supervision (Joy et al., 2022), while recently Palumbo et al. (2023) proposed to incorporate the idea of separate shared and private subspaces from previous works (Sutter et al., 2020; Lee & Pavlovic, 2021; Wang et al., 2016) and design an ELBO that exploits auxiliary distributions to facilitate the estimation of cross-modal likelihood terms. The resulting MMVAE+ model achieves both high generative quality and high generative coherence.

Variational approaches for clustering Previous work proposes VAE-based unimodal clustering approaches such as GMVAE (Dilokthanakul et al., 2016) and VaDE (Jiang et al., 2016). Later, these models have been extended to condition on pair-wise prior constraints (Manduchi et al., 2021). However, they lack multi-modality support.

Clustering with weak- or self-supervision While clustering approaches for weakly-supervised data have been investigated, they are not designed to scale to a large number of modalities (Alwassel et al., 2020; Chen et al., 2021; Zhou & Shen, 2020). Therefore, to benchmark our method, we compare with an existing well-known multi-view approach, CMC (Tian et al., 2020). CMC maximizes mutual information between different views with a contrastive loss, and can be extended for clustering by training a K-means model on the learned representations. It is adapted to multi-modality by providing separate encoders for separate modality types. Note that the key feature for why this model is chosen for comparison is its ability to handle numerous modalities, not present for other existing multimodal approaches.

3. Method

3.1. A scalable VAE objective for modelling latent clusters in multimodal data

We assume data consisting of M modalities $\boldsymbol{X} := \boldsymbol{x}_1, \dots, \boldsymbol{x}_M$ is generated according to the following process. For each datapoint $\boldsymbol{x}_1^i, \dots, \boldsymbol{x}_M^i$ where $i \in \{1, \dots, N\}$ and N is the dataset size, a cluster assignment $\boldsymbol{c}^i$ is drawn from a categorical distribution $p_{\boldsymbol{\pi}}(\boldsymbol{c})$ with probabilities $\boldsymbol{\pi} = \pi_1, \dots, \pi_K$ where K is the number of clusters. Then the M modalities are drawn according to $\boldsymbol{x}_1^i, \dots, \boldsymbol{x}_M^i \sim p_{\theta_1}(\boldsymbol{x}_1|\boldsymbol{w}_1^i, \boldsymbol{z}^i), \dots, p_{\theta_M}(\boldsymbol{x}_M|\boldsymbol{w}_M^i, \boldsymbol{z}^i)$ where the shared encoding $\boldsymbol{z}^i \sim p(\boldsymbol{z}|\boldsymbol{c}^i)$ is generated conditioning on cluster assignment, while modality-speficic encodings $\boldsymbol{w}_1^i \sim p(\boldsymbol{w}_1), \dots, \boldsymbol{w}_M^i \sim p(\boldsymbol{w}_M)$ are drawn from prior distributions. The resulting generative model is $p_\Theta(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{z}, \boldsymbol{c}) = p_{\boldsymbol{\pi}}(\boldsymbol{c})p(\boldsymbol{z}|\boldsymbol{c}) \prod_{m=1}^M p_{\theta_m}(\boldsymbol{x}_m|\boldsymbol{w}_m, \boldsymbol{z})p(\boldsymbol{w}_m)$, where priors and likelihoods are assumed to belong to a specific family of distributions, e.g. Gaussian or Laplace, and likelihoods are parameterized by neural networks. Note that the shared encoding $\boldsymbol{z}$ and modality-specific encodings $\boldsymbol{w}_1, \dots \boldsymbol{w}_M =: \boldsymbol{W}$ are assumed to be independent.

To obtain a tractable objective, variational encoders $q_{\Phi_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}), q_{\phi_{\boldsymbol{w}_1}}(\boldsymbol{w}_1|\boldsymbol{x}_1), \dots, q_{\phi_{\boldsymbol{w}_M}}(\boldsymbol{w}_M|\boldsymbol{x}_M), q(\boldsymbol{c}|\boldsymbol{z}, \boldsymbol{X})$ are introduced to approximate posterior inference for each of the latent variables. In line with our generative assumptions, the shared and modality-specific encoders are assumed to be conditionally independent given the observed data. As in previous approaches (Shi et al., 2019; Palumbo et al., 2023), to achieve scalability in the number of modalities, we model the joint encoder as a mixture of experts $q_{\Phi_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}) = \frac{1}{M}\sum_{m=1}^M q_{\phi_{\boldsymbol{z}_m}}(\boldsymbol{z}|\boldsymbol{x}_m)$. In our objective, we incorporate two key ideas from previous related work. First, to accurately model both shared and modality-specific information in separate latent subspaces without conflicts, as in (Palumbo et al., 2023) we use auxiliary distributions for private features to estimate cross-modal reconstruction likelihoods, leading to our proposed CMVAE objective

$$\mathcal{L}(\boldsymbol{X}) = \frac{1}{M}\sum_{m=1}^M \mathbb{E}_{\substack{q(\boldsymbol{c}|\boldsymbol{z},\boldsymbol{X}) \\ q_{\phi_{\boldsymbol{z}_m}}(\boldsymbol{z}|\boldsymbol{x}_m) \\ q_{\phi_{\boldsymbol{w}_m}}(\boldsymbol{w}_m|\boldsymbol{x}_m)}} \left[G_{\boldsymbol{\pi},\Phi_{\boldsymbol{z}},\phi_{\boldsymbol{w}_m},\Theta}(\boldsymbol{X}, \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{w}_m) \right]$$

where

$$\begin{aligned} G_{\boldsymbol{\pi},\Phi_{\boldsymbol{z}},\phi_{\boldsymbol{w}_m},\Theta}(\boldsymbol{X}, \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{w}_m) &= \log p_{\theta_m}(\boldsymbol{x}_m|\boldsymbol{z}, \boldsymbol{w}_m) \\ &+ \sum_{n \neq m} \log \mathbb{E}_{\tilde{\boldsymbol{w}}_n \sim p(\boldsymbol{w}_n)}[p_{\theta_n}(\boldsymbol{x}_n|\boldsymbol{z}, \tilde{\boldsymbol{w}}_n)] \\ &+ \log \frac{p_\pi(\boldsymbol{c})p_\theta(\boldsymbol{z}|\boldsymbol{c})p(\boldsymbol{w}_m)}{q_{\Phi_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X})q_{\phi_{\boldsymbol{w}_m}}(\boldsymbol{w}_m|\boldsymbol{x}_m)q(\boldsymbol{c}|\boldsymbol{z}, \boldsymbol{X})} \end{aligned}$$

Furthermore, we adopt the formulation for the approximate posterior of cluster assignment given a latent code $\boldsymbol{z}$ proposed by (Jiang et al., 2016), and assume $q(\boldsymbol{c}|\boldsymbol{z}, \boldsymbol{X}) = p(\boldsymbol{c}|\boldsymbol{z}) = \frac{p(\boldsymbol{c})p(\boldsymbol{z}|\boldsymbol{c})}{\sum_{c'=1}^K p(\boldsymbol{c}')p(\boldsymbol{z}|\boldsymbol{c}')}$. Note that expectations with respect to $q_{\Phi_{\boldsymbol{z}}}(\boldsymbol{z}|\boldsymbol{X}), q_{\phi_{\boldsymbol{w}_1}}(\boldsymbol{w}_1|\boldsymbol{x}_1), \dots, q_{\phi_{\boldsymbol{w}_M}}(\boldsymbol{w}_M|\boldsymbol{x}_M)$ are approximated via sampling while the expectation with the respect to $q(\boldsymbol{c}|\boldsymbol{z}, \boldsymbol{X})$ can be computed exactly since $\boldsymbol{c}$ assumes a discrete finite set of values. Our objective can also be framed as the MMVAE+ ELBO objective from (Palumbo et al., 2023), with a mixture prior distribution and our specific choice for the approximate posterior $q(\boldsymbol{c}|\boldsymbol{z}, \boldsymbol{X})$. This in turn validates our proposed objective as an ELBO.

3.2. Entropy of posterior cluster assignment distribution for post-hoc learning of the number of clusters

A critical limitation of most existing methods for clustering is the reliance on prior knowledge of the number of clusters

Algorithm 1 Post-hoc selection of optimal latent clusters

Input: $p_{\boldsymbol{\pi}_K}(\boldsymbol{c}), p(\boldsymbol{z}|\boldsymbol{c}), q_{\phi_{\boldsymbol{z}_1}}(\boldsymbol{z}|\boldsymbol{x}_1) \ldots, q_{\phi_{\boldsymbol{z}_M}}(\boldsymbol{z}|\boldsymbol{x}_M)$ from trained CMVAE model with $K > \bar{K}$, data $\boldsymbol{X}^{1:N}$
Output: $p_{\boldsymbol{\pi}_{\hat{K}}}$, with $\pi_{\hat{K}}$ s.t. $\sum_{k=1}^{K} \mathbb{1}_{\pi_k \neq 0} = \hat{K}$
for $k = K$ **to** 2 **do**
  for $\boldsymbol{x}_1^i, \ldots, \boldsymbol{x}_M^i$ **in** $\boldsymbol{x}_1^{1:N}, \ldots, \boldsymbol{x}_M^{1:N}$ **do**
    $\boldsymbol{z}_1^i, \ldots, \boldsymbol{z}_M^i \sim q_{\phi_{\boldsymbol{z}_1}}(\boldsymbol{z}|\boldsymbol{x}_1^i), \ldots, q_{\phi_{\boldsymbol{z}_M}}(\boldsymbol{z}|\boldsymbol{x}_M^i)$
    for $m = 1$ **to** M **do**
      $p(\boldsymbol{c}|\boldsymbol{z}_m) = \frac{p_{\boldsymbol{\pi}_{\boldsymbol{k}}}(\boldsymbol{c}) p(\boldsymbol{z}_m|\boldsymbol{c})}{\sum_{\boldsymbol{c}'=1}^{k} p_{\boldsymbol{\pi}_{\boldsymbol{k}}}(\boldsymbol{c}') p(\boldsymbol{z}_m|\boldsymbol{c}')}$
    end for
    $\boldsymbol{c}_{as}^i = \text{assign}_{\text{c}}(p(\boldsymbol{c}|\boldsymbol{z}_1), \ldots, p(\boldsymbol{c}|\boldsymbol{z}_M))$
    $h_k^i = \frac{1}{M} \sum_{m=1}^{M} \bar{H}(p(\boldsymbol{c}|\boldsymbol{z}_m))$
  end for
  $h_k = \frac{1}{N} \sum_{n=1}^{N} h_k^n$
  $\boldsymbol{\pi}_{\boldsymbol{k-1}} = \text{compute}_{\pi}(\boldsymbol{c}_{as}^{1:N}, \boldsymbol{\pi}_{\boldsymbol{k}})$
end for
$p_{\boldsymbol{\pi}_{\hat{\boldsymbol{K}}}}$ where $\hat{K} = \text{argmin}_k(h_1, \ldots, h_k, \ldots, h_K)$

in the data. This can result in either highly complex model selection procedures, or in failure in settings where a proxy for this information cannot be obtained. Training CMVAE requires assuming a K value for the number of clusters modeled. However, K may differ from the true unknown number of clusters $\bar{K}$ of the true generating process. Specifying $K > \bar{K}$ leads to learning $p_{\boldsymbol{\pi}}(\boldsymbol{c})$ where multiple latent clusters, corresponding to the same true cluster, are allocated positive probability. Recovering the true number of clusters means obtaining $p_{\bar{\boldsymbol{\pi}}}(\boldsymbol{c})$ where $\sum_{k=1}^{K} \mathbb{1}_{\bar{\pi}_k \neq 0} = \bar{K}$, i.e. *exactly* $\bar{K}$ clusters have positive prior probability. To this end, we design a post-hoc procedure to learn the true number of clusters given a trained instance of our model with over-specified number of clusters, i.e. $K > \bar{K}$, described in the pseudocode in Algorithm 1.

In a nutshell, the procedure iterates over the dataset: clusters are ranked by counting the data points each one is assigned to, and the average normalized entropy of $p(\boldsymbol{c}|\boldsymbol{z})$, which is the posterior distribution of clusters assignments, is computed. Note that cluster assignments are determined by majority voting between modalities. Then the latent cluster with the least number of samples assigned is set null prior probability at the next iteration (other probabilities are recomputed accordingly to maintain a valid probability distribution). Iteratively, the latent clusters are effectively pruned, keeping the average normalized entropy of $p(\boldsymbol{c}|\boldsymbol{z})$ as a metric to select the optimal set of latent clusters to model the data. This procedure aims to find $\hat{K} = \bar{K}$, i.e. recover the true number of clusters. $\bar{H}$ in Algorithm 1 denotes normalized entropy. The procedure is fully unsupervised, and does not affect the unsupervised nature of our method.

4. Experiments

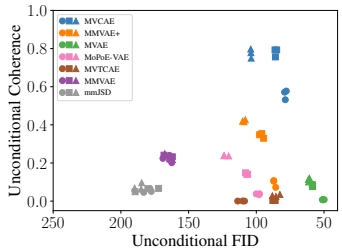
We test the performance of CMVAE, in comparison with alternative approaches, in two challenging experimental settings. The first setting is the PolyMNIST dataset (Sutter et al., 2021), which is a synthetic five-modality dataset depicting MNIST (LeCun et al., 2010) digits with modality-specific backgrounds. As a second experimental setting, we introduce a variation of the CUB Image-Captions dataset (Wah et al., 2011; Reed et al., 2016), conceived to test multimodal clustering methods in a realistic scenario, which we name CUBICC dataset. The original CUB Image-Captions dataset consists of images of birds paired with matching descriptive captions, and to adapt it to benchmark clustering approaches we group sub-species of birds in the original dataset in a single species. Details for the datasets are in Appendix A.1.

4.1. PolyMNIST: generative capabilities and multimodal clustering performance

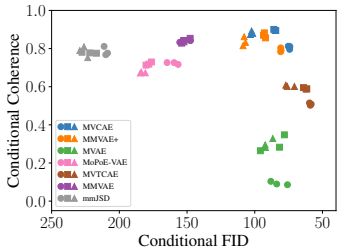
As a first experimental setting, we test the CMVAE on the PolyMNIST dataset. First, we compare its generative capabilities with alternative multimodal VAEs, then moving to benchmark its clustering performance.

Section 4.1 compares the performance of CMVAE with alternative multimodal VAEs, in terms of generative quality and generative coherence, for both unconditional and conditional generation. Results show that CMVAE outperforms all existing multimodal VAEs for both conditional and unconditional generation. While the advancement for conditional generation is moderate, CMVAE improves over other models by a significant margin in unconditional generation, particularly for unconditional coherence, which for existing methods represents a critical performance weakness.

Furthermore, we look at clustering results for our method, in this multimodal setting. As described in 3.2, we apply a post-hoc procedure to learn the optimal number of clusters, showing its effectiveness in Figure 2. Quantitative results in Table 1c show CMVAE can accurately model latent clusters in the data. In this setting, unimodal clustering approaches fail to achieve good performance: we find they rather model background features, which are prominent pixel-wise in the images, instead of the relevant digit content. To not restrict ourselves to only variational approaches for unimodal clustering, we include in the comparison the well-known DeepCluster (Caron et al., 2018). See more details about unimodal approaches and results in Appendix A.4. On the other hand, weak-supervision plays a fundamental role in this setting, as also confirmed by the results for CMC. Not surprisingly, CMVAE achieves comparable yet not superior performance to CMC in this setting, as it closely resembles a multi-view setting for which these approaches are conceived. However, for more heterogeneous modalities,



(a) Unconditional generation



(b) Conditional generation

PolyMNIST

	NMI	ARI
VaDE	0.41 (±0.04)	0.34 (±0.05)
DeepCluster	0.18 (±0.06)	0.12 (±0.03)
CMC	0.97 (±0.01)	0.97 (±0.01)
CMVAE	0.97 (±0.01)	0.97 (±0.00)

(c) Quantitative clustering results on PolyMNIST

Figure 1 (a),(b): Generative performance comparison of CMVAE with multimodal VAEs on PolyMNIST. Three independent runs for each model, with different symbols denoting different values of the β hyperparameter. In both plots, generative coherence is measured on the y-axis (higher is better), while on the x-axis generative quality is assessed via the FID-score (lower is better). Therefore optimal performance is at the top-right corner of each plot. Table 1 (c): Quantitative clustering performance comparison on PolyMNIST.

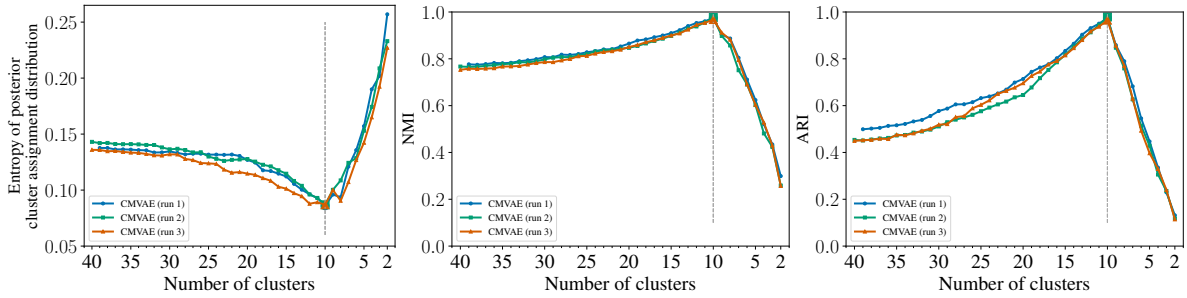


Figure 2: Post-hoc optimal latent cluster selection with the procedure in Algorithm 1. (Left) Evolution of the entropy term (y-axis) as clusters (x-axis) are pruned.The minimal entropy recovers the true number of clusters. The trends for test NMI (Center) and ARI (Center) support the validity of our procedure, with best results reached at minimal entropy value.

Figure 3: Qualitative results for CMVAE on the CUBICC dataset. For each row, three samples for both modalities are sampled from a different latent cluster.

the gap in performance between the two methods, is highly significant, as shown in the next section.

4.2. CUBICC

With this dataset we test the performance of CMVAE in a realistic multimodal scenario. Qualitative results in Figure 3 for unconditional generation show that CMVAE is able to exploit multiple modalities to model latent clusters also in a complex real-world setting. Note that this level of generative quality and semantic coherence for unconditional generation favorably compares with previous results for multimodal VAEs in this setting (see Palumbo et al. (2023)). Quantitative results for clustering in Table 2 validate the effectiveness of CMVAE in this multimodal challenging setting, where it outperforms both unimodal and alternative weakly-supervised approaches.

CUBICC

	NMI	ARI
VaDE	0.14 (±0.01)	0.07 (±0.01)
DeepCluster	0.16 (±0.03)	0.03 (±0.01)
CMC	0.33 (±0.02)	0.07 (±0.01)
CMVAE	0.53 (±0.04)	0.44 (±0.09)

Table 2: Quantitative clustering results for CUBICC.

5. Conclusion

In this work we introduce CMVAE, a novel multimodal VAE, which combines recent advances for multimodal VAEs with clustering capabilities. Our experiments show that CMVAE outperforms existing multimodal VAEs in generative performance, particularly for unconditional generation. Additionally, CMVAE favorably compares to both

unimodal and alternative weakly-supervised scalable approaches for clustering, outperforming these approaches in realistic datasets. Notably, we introduce and validate a post-hoc procedure to prune latent clusters based on the average normalized entropy of posterior cluster assignments.

References

Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., and Tran, D. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Chen, B., Rouditchenko, A., Duarte, K., Kuehne, H., Thomas, S., Boggust, A., Panda, R., Kingsbury, B., Feris, R., Harwath, D., et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8012–8021, 2021.

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.

Dilokthanakul, N., Mediano, P. A., Garnelo, M., Lee, M. C., Salimbeni, H., Arulkumaran, K., and Shanahan, M. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

Dorent, R., Joutard, S., Modat, M., Ourselin, S., and Vercauteren, T. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. *Medical Image Computing and Computer Assisted Intervention*, 2019.

Fortuin, V., Hüser, M., Locatello, F., Strathmann, H., and Rätsch, G. Som-vae: Interpretable discrete representation learning on time series. In *International Conference on Learning Representations*, 2019.

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

Jiang, Z., Zheng, Y., Tan, H., Tang, B., and Zhou, H. Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*, 2016.

Joy, T., Shi, Y., Torr, P., Rainforth, T., Schmon, S. M., and N, S. Learning multimodal VAEs through mutual supervision. In *International Conference on Learning Representations*, 2022.

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

LeCun, Y., Cortes, C., and Burges, C. MNIST handwritten digit database. *URL: http://yann.lecun.com/exdb/mnist*, 2010.

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

Lee, M. and Pavlovic, V. Private-shared disentangled multimodal vae for learning of latent representations. In *CVPR Workshop on Multimodal Learning and Applications*, 2021.

Manduchi, L., Chin-Cheong, K., Michel, H., Wellmann, S., and Vogt, J. E. Deep conditional Gaussian mixture model for constrained clustering. In *Advances in Neural Information Processing Systems*, 2021.

Manduchi, L., Marcinkevičs, R., Massi, M. C., Weikert, T., Sauter, A., Gotta, V., Müller, T., Vasella, F., Neidert, M. C., Pfister, M., Stieltjes, B., and Vogt, J. E. A deep variational approach to clustering survival data. In *International Conference on Learning Representations*, 2022.

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

Reed, S., Akata, Z., Lee, H., and Schiele, B. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 49–58, 2016.

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

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

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

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

Suzuki, M., Nakayama, K., and Matsuo, Y. Joint multimodal learning with deep generative models. 2017.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

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.

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

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

Wang, W., Yan, X., Lee, H., and Livescu, K. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454.*, 2016.

Wu, M. and Goodman, N. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, 2018.

Zhou, R. and Shen, Y.-D. End-to-end adversarial-attention network for multi-modal clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14619–14628, 2020.

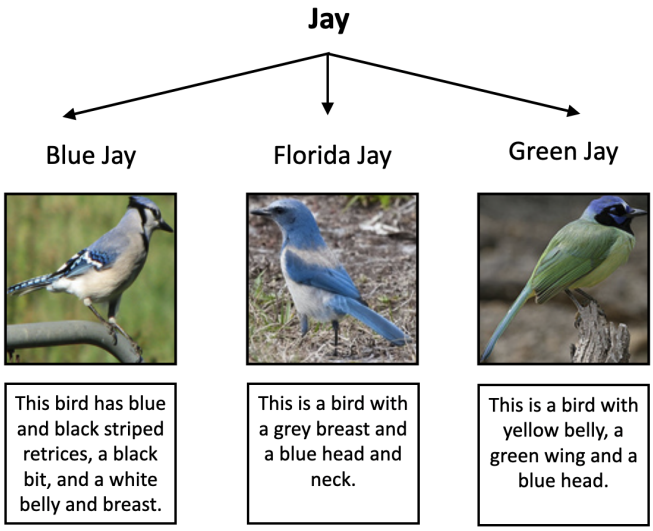
A. Experimental and implementation details

A.1. Datasets

In this section, we provide detailed information about the datasets used in this study. First, we describe the PolyMNIST dataset (Sutter et al., 2021), which is a synthetic and challenging dataset comprising five image modalities. Each modality depicts MNIST (LeCun et al., 2010) digits patched with random crops from five distinct background images, with one background image associated to each modality. Figure 4a visually illustrates examples of the dataset. Notably, the digit label is the only shared information across modalities, while the background features and the handwriting style of the digits are modality-specific for each data point.

As a second experimental setting, we introduce the CUBICC dataset, a variation of the CUB Image-Captions dataset (Wah et al., 2011; Reed et al., 2016). The CUB Image-Captions dataset consists of images of birds paired with corresponding descriptive captions. We modified this dataset to evaluate multimodal clustering methods in a realistic scenario. Our modified version, named CUBICC, involves grouping sub-species of birds into a single species category. As a result, the CUBICC dataset consists of nine classes, with each class representing a different bird species. Figure 4b provides an example illustration of the dataset. The grouping of sub-species into a single class introduces significant variability within each class, posing a considerable modeling challenge. It is worth noting that previous studies have evaluated multimodal VAEs on the original CUB Image-Captions dataset using pre-trained ResNet features (Shi et al., 2019; 2021), or generating directly in the data space. The latter more challenging setting (Daunhawer et al., 2022) is the one considered in this work. Only recent approaches have proved to be successful in this more demanding setting (Palumbo et al., 2023).

(a) PolyMNIST dataset: Each column is a single data point in this dataset, which consists of five image modalities with matching digit labels. Each row showcases samples from a given modality.



(b) CUBICC dataset: Example of the variability within a single bird class, jay. The image and its corresponding caption are arranged in a stacked manner, with the image positioned above the caption.

Figure 4: Illustrative samples for our considered experimental settings, PolyMNIST in (a) and CUBICC in (b).

A.2. Implementation details

To implement all multimodal VAEs (Wu & Goodman, 2018; Shi et al., 2019; Sutter et al., 2020; 2021; Hwang et al., 2021; Palumbo et al., 2023) included for comparison in terms of generative performance we follow the recommendations in Palumbo et al. (2023) for training these models on the PolyMNIST dataset. In particular, we use the same ResNet encoder and decoder networks for all compared models. CMVAE is trained for 250 epochs on this dataset, with $1e-3$ learning rate. As for the MMVAE+, CMVAE is set to have a shared latent space of 32 dimensions and modality-specific latent spaces of 32 dimensions in this setting. As for CMVAE on the CUBICC dataset, we use ResNet and convolutional encoder/decoder networks for the image and text modalities respectively, training the model with a 10-sample version of our objective and resorting to the DReG estimator for gradient (Tucker et al., 2019). We set 64 dimensions and 32 dimensions for shared and modality-specific latent spaces respectively.

For the methods reported as baseline comparisons for clustering tasks, we ensure compatibility with the encoder networks and shared latent space size adopted by CMVAE for a fair comparison. We follow best practices in their original work for training, but without resorting to any pre-training procedures, again for a fair comparison with CMVAE. For PolyMNIST CMVAE is trained with $K = 40$, before the post-hoc procedure shown in Figure 2 is applied. This shows that our method can recover the true number of clusters even with a largely overspecified K for training.

Experimental results across the paper are reported averaging for three independent runs and we report standard deviations.

A.3. Generative qualitative clustering results on PolyMNIST

Figure 5 showcases the generative and clustering capabilities of CMVAE, organized according to each modality. The results demonstrate that CMVAE effectively captures the true clusters present in the data by focusing on the meaningful digit content rather than the dominant background features. It successfully generates high-quality samples that exhibit semantic coherence within each cluster. Conversely, unimodal clustering approaches struggle to achieve satisfactory performance in this context.

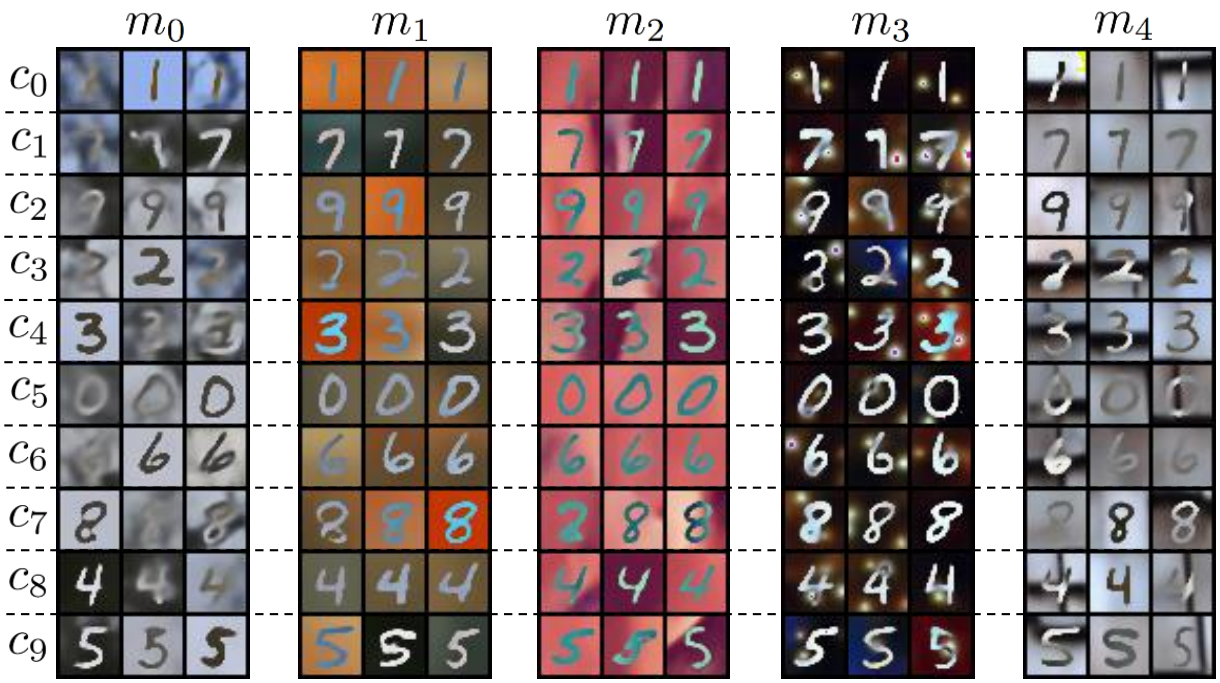


Figure 5: Generative qualitative examples of the clustering capabilities of CMVAE on PolyMNIST dataset. Each column corresponds to the generated samples of one of the five modalities, while each row reports three generation results for a different latent cluster.

A.4. Unimodal clustering approaches

Here we show additional results for the comparison of our proposed CMVAE with existing unimodal clustering approaches, namely VADE (Jiang et al., 2016) and DeepCluster (Caron et al., 2018). DeepCluster requires a prior selection of clusters for multi-class classification. Since previous research has shown that a larger number of clusters than the actual classes might yield better results, to replicate their experiments we consider 100, 35, and 10 clusters for the PolyMNIST dataset, and 100, 35, and 9 clusters for the CUB images. Additionally, we examine the performance of VADE using different values of the parameter β, specifically 1.0 and 2.5. In this section, we present the results for each individual modality in PolyMNIST, and for the image modality in the case of CUBICC, using these methods.

Overall, the performance varies significantly depending on the modality. Notably, in PolyMNIST, the prominence of the

background plays a crucial role, affecting the task's difficulty level. Multi-modality in CMVAE helps to overcome these challenges by learning a shared representation space. Regarding DeepCluster, a larger Normalized Mutual Information (NMI) is observed when using a higher number of clusters. However, the Adjusted Random Index (ARI) is higher when the true number of clusters is considered. VADE exhibits improved results with a β value of 2.5. For both PolyMNIST and CUBICC the unimodal clustering results reported in the main text are for K equal to the true number of clusters, and choosing the best-performing modality.

PolyMNIST						
		m0	m1	m2	m3	m4
VADE β=2.5	NMI	0.0381 ($\pm$0.0134)	0.1683 ($\pm$0.0308)	0.0725 ($\pm$0.0416)	0.4062 ($\pm$0.0444)	0.0122 ($\pm$0.0074)
	ARI	0.0174 ($\pm$0.0070)	0.0979 ($\pm$0.0242)	0.0411 ($\pm$0.0263)	0.3430 ($\pm$0.0516)	0.0072 ($\pm$0.0040)
DeepCluster K=10	NMI	0.0317 ($\pm$0.0054)	0.1690 ($\pm$0.0156)	0.0476 ($\pm$0.0064)	0.1848 ($\pm$0.0630)	0.0075 ($\pm$0.0028)
	ARI	0.0134 ($\pm$0.0029)	0.1025 ($\pm$0.0105)	0.0252 ($\pm$0.0014)	0.1160 ($\pm$0.0344)	0.0030 ($\pm$0.0015)

Table 3: Baseline results for unimodal clustering on PolyMNIST dataset.

PolyMNIST						
		m0	m1	m2	m3	m4
VADE β=1.0	NMI	0.0173 ($\pm$0.0018)	0.1342 ($\pm$0.0451)	0.0108 ($\pm$0.0100)	0.3921 ($\pm$0.0135)	0.0016 ($\pm$0.0008)
	ARI	0,0072 ($\pm$0.0008)	0.0824 ($\pm$0.0406)	0.0068 ($\pm$0.0061)	0.3262 ($\pm$0.0145)	0.0008 ($\pm$0.0006)
VADE β=2.5	NMI	0.0381 ($\pm$0.0134)	0.1683 ($\pm$0.0308)	0.0725 ($\pm$0.0416)	0.4062 ($\pm$0.0444)	0.0122 ($\pm$0.0074)
	ARI	0.0174 ($\pm$0.0070)	0.0979 ($\pm$0.0242)	0.0411 ($\pm$0.0263)	0.3430 ($\pm$0.0516)	0.0072 ($\pm$0.0040)
DeepCluster K=10	NMI	0.0317 ($\pm$0.0054)	0.1690 ($\pm$0.0156)	0.0476 ($\pm$0.0064)	0.1848 ($\pm$0.0630)	0.0075 ($\pm$0.0028)
	ARI	0.0134 ($\pm$0.0029)	0.1025 ($\pm$0.0105)	0.0252 ($\pm$0.0014)	0.1160 ($\pm$0.0344)	0.0030 ($\pm$0.0015)
DeepCluster K=35	NMI	0.0881 ($\pm$0.0091)	0.2369 ($\pm$0.0221)	0.1017 ($\pm$0.0103)	0.1520 ($\pm$0.0374)	0.0360 ($\pm$0.0073)
	ARI	0.0296 ($\pm$0.0054)	0.0981 ($\pm$0.0160)	0.0400 ($\pm$0.0055)	0.0554 ($\pm$0.0209)	0.0094 ($\pm$0.0030)
DeepCluster K=100	NMI	0.1075 ($\pm$0.0160)	0.2676 ($\pm$0.0308)	0.1295 ($\pm$0.0086)	0.2123 ($\pm$0.0201)	0.0444 ($\pm$0.0064)
	ARI	0.0204 ($\pm$0.0032)	0.2553 ($\pm$0.3531)	0.0219 ($\pm$0.0026)	0.0381 ($\pm$0.0072)	0.0046 ($\pm$0.0009)

Table 4: Baseline results for unimodal clustering on PolyMNIST dataset. Including ablation of beta and number of clusters.

CUB-Images		
VADE β=1.0	NMI	0.1361 ($\pm$0.0115)
	ARI	0.0704 ($\pm$0.0072)
DeepCluster K=9	NMI	0.1611 ($\pm$0.0335)
	ARI	0.0335 ($\pm$0.0109)

Table 5: Baseline results for unimodal clustering on images from CUBICC dataset.

CUB-Images		
DeepCluster K=9	NMI	0.1611 ($\pm$0.0335)
	ARI	0.0335 ($\pm$0.0109)
DeepCluster K=35	NMI	0.3207 ($\pm$0.0149)
	ARI	0.0205 ($\pm$0.0082)
DeepCluster K=100	NMI	0.4659 ($\pm$0.0143)
	ARI	0.0094 ($\pm$0.0085)

Table 6: Baseline results for unimodal clustering on images from CUBICC dataset. Including ablation of number of clusters.