Improving Multimodal Generative Models with Disentangled Latent Partitions

Imant Daunhawer, Thomas Sutter, Julia E. Vogt
Department of Computer Science
ETH Zurich
Universitaetstrasse 6, 8092 Zurich
{dimant,suttetho,jvogt}@inf.ethz.ch

Abstract

Multimodal generative models learn a joint distribution of data from different modalities—a task which arguably benefits from the disentanglement of modality-specific and modality-invariant information. We propose a factorized latent variable model that learns named disentanglement on multimodal data without additional supervision. We demonstrate the disentanglement capabilities on simulated data, and show that disentangled representations can improve the conditional generation of missing modalities without sacrificing unconditional generation.

1 Introduction

Learning a joint generative model of multimodal data is promising, because it enables the integration of unimodal beliefs into a richer joint representation, as well as the conditional generation of missing modalities [1].

Simpler alternatives to (joint) multimodal generative models include unimodal models with late fusion or with coordinated representations, as well as conditional models that translate between pairs of modalities. Yet, both alternatives have disadvantages compared to multimodal models: while unimodal models cannot handle missing modalities, conditional models only learn a mapping between sources, and neither learn to integrate beliefs from different modalities into a joint representation. In contrast, multimodal generative models approximate the joint distribution and thus implicitly provide the marginal (unimodal) and conditional distributions.

In trying to bridge the gap between marginal and conditional models, we propose a multimodal generative model that disentangles modality-specific and modality-invariant factors of variation. We argue that such disentanglement is a key component for improving the conditional generation, because it sifts out shared semantics from modality-specific variations and thus reinforces the composability of the learned representations. For example, this allows the generation of a missing modality given an aggregated representation of shared semantics from the present modalities and random samples from the modality-specific prior. We show that the disentanglement capabilities of our model emerge naturally under mild assumptions on the data generating process.

We support our claims by demonstrating (1) the disentanglement capabilities of our model and (2) an improved conditional generation in a controlled setting.

2 Related Work

Multimodal generative models Our model builds on the multimodal variational autoencoder [19]: a multimodal generative model that efficiently handles missing data. The original model, however, is limited to shared latent factors, and we extend it by modeling also modality-specific factors to improve

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

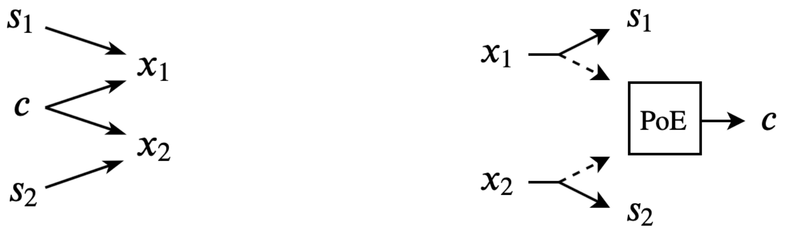


Figure 1: Graphical model and inference network for the special case of two modalities. *Left:* A sample x_m from modality m is assumed to be generated by modality-specific factors s_m and modality-invariant factors c. *Right:* Inference network with the aggregation of beliefs about modality-invariant factors through a product of experts (PoE). Dashed lines represent simulated missing modalities as used during training to incite disentanglement.

the conditional generation of missing modalities. Hsu and Glass [7] also use a partitioned latent space to model modality-specific and modality-invariant factors, but in contrast to their work we provide a scalable multimodal inference network which arises naturally from the same assumptions and handles any combination of missing modalities efficiently. Moreover, there are supervised approaches [20, 17] that are limited to labeled multimodal data.

Disentanglement Our goal is not the unsupervised disentanglement of all generative factors, which Locatello et al. [12] show to be theoretically impossible with a factorizing prior and claim to be impossible in general. Instead, we are only concerned with the disentanglement of modality-specific and modality-invariant sets of factors and argue that such disentanglement is possible given modalities that share a set of (unknown) generative factors. This is a form of implicit supervision that constitutes an inductive bias we exploit for the named disentanglement. Previous approaches to disentanglement through implicit supervision have used grouping information [2, 8] or temporal dependencies [11], but do not extend to a multimodal setting which poses further challenges such as modality-specific objectives.

3 Method

We consider a generative process with a partition into modality-specific and modality-invariant (shared) latent factors (Figure 1). A *multimodal sample* $\mathbf{x} = (x_1, \dots, x_M)$ with data from M modalities is assumed to be generated from a set of shared factors c and a set of modality-specific factors s_m. Consequently, samples from different modalities are assumed to be conditionally independent given c. In the following, we denote the set of all modality-specific factors of a multimodal sample as $\mathbf{s} = (s_1, \dots s_M)$.

Given a dataset $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$ of multimodal samples, our goal is to learn a generative model $p_\theta(\mathbf{x} \mid c, \mathbf{s})$ with a neural network parameterized by θ. From the above assumptions on the data generating process we get the following joint distribution

$$p(\mathbf{x}, \mathbf{s}, c) = p(c) \prod_{m=1}^{M} p(s_m) p(x_m \mid c, s_m) \tag{1}$$

which allows us to consider only the observed modalities for the computation of the marginal likelihood.

Since we use a decoder that is parameterized by a neural network, the computation of the exact likelihood is intractable. Therefore, we resort to amortized variational inference and instead maximize the following evidence lower bound (ELBO)

$$\begin{aligned}\mathcal{L}_{\text{ELBO}}(\mathbf{x}) = \sum_{m=1}^{M} \Big[&\mathbb{E}_{q_{\phi_c}(c \mid \mathbf{x}) q_{\phi_m}(s_m \mid x_m)} \left[\log p_{\theta_m}(x_m \mid c, s_m)\right] - \beta_m D_{\text{KL}}\left(q_{\phi_m}(s_m|x_m) \,||\, p(s_m)\right) \Big] \\ &- \beta_c D_{\text{KL}}\left(q_{\phi_c}(c|\, \mathbf{x}) \,||\, p(c)\right) \end{aligned} \tag{2}$$

which is an objective composed of m unimodal ELBOs for the marginal likelihoods and an additional KL-divergence for the multimodal encoder. The coefficients β_m and β_c can be used to balance the unimodal and multimodal KL-divergences which act as regularizers that constrain the capacity of the respective latent channel [5]. We use neural networks for the unimodal decoders $p_{\theta_m}(x_m \mid c, s_m)$, the unimodal encoders $q_{\phi_m}(s_m \mid x_m)$, as well as for the multimodal encoder $q_{\phi_c}(c \mid \mathbf{x})$ and denote the network parameters by the respective subscripts for decoder parameters θ and encoder parameters ϕ. Further, we follow the convention of using isotropic unit Gaussian priors, as well as Gaussian variational posteriors parameterized by the estimated means and variances from the respective encoder.

3.1 Multimodal inference network

A key aspect in the design of multimodal models should be the capability to handle missing modalities efficiently [1]. In our case, only the multimodal encoder $q_{\phi_c}(c| \mathbf{x})$, depends on all modalities and should ideally be able to cope with any combination of missing inputs, which would require 2^M multimodal inference networks in a naive implementation. A more efficient alternative is offered by Wu and Goodman [19] who use a product of experts [6] to handle missing modalities. Analogously, we can show that the multimodal posterior in our case is proportional to a product of unimodal posteriors

$$p(c \mid \mathbf{x}) \propto \frac{1}{p(c)^{M-1}} \prod_{m=1}^{M} p(c \mid x_m)\,. \tag{3}$$

which—for the special case of Gaussian variational posteriors—has an efficient closed-form solution (see Appendix A.4). Therefore, as in Wu and Goodman [19], M unimodal inference networks suffice to handle all 2^M combination of missing modalities. In contrast to previous work, however, our model is not restricted to shared latent factors, but extends to modality-specific factors.

3.2 Disentangling c and s_m

To disentangle modality-specific and modality-invariant factors, our key idea is to direct shared information through the multimodal encoder by computing reconstructions on $p(x_m \mid \tilde{c}, s_m)$ where $\tilde{c} \sim q_{\phi_c}(c \mid \tilde{\mathbf{x}})$ is computed from $\tilde{\mathbf{x}} \subseteq \mathbf{x} \setminus x_m$ (i.e., a subset of modalities that does not contain x_m). Intuitively, this should incite the shared representation to be useful across modalities.

We can efficiently compute $q_{\phi_c}(c \mid \tilde{\mathbf{x}})$ for any subset $\tilde{\mathbf{x}}$ by masking modalities at random during training—a process that is handled efficiently by the product of experts as discussed previously.

There exists a trivial solution that can prevent disentanglement when the model has sufficient capacity. For instance, if a modality-specific encoder $q_{\phi_m}(s_m|x_m)$ has high capacity, the model can direct all information—both shared and modality-specific—through the unimodal encoder (instead of the multimodal encoder) to compute the reconstructions. However, we can constrain the capacity of the unimodal encoders by increasing β_m which also allows us to control the degree of disentanglement for each modality.[1]

To measure the degree of disentanglement between the learned representations c and s_m we estimate the total correlation [18]

$$\begin{aligned} TC(c, s_m) &= D_{\mathrm{KL}}(q(c, s_m) \mid\mid q(c)q(s_m)) && (4)\\ &= \mathbb{E}_{q(c,s_m)}\left[\log \frac{q(c, s_m)}{q(c)q(s_m)}\right] && (5)\\ &\approx \mathbb{E}_{q(c,s_m)}\left[\log \frac{D(c, s_m)}{1 - D(c, s_m)}\right] && (6) \end{aligned}$$

using the density-ratio trick [14, 16] where a discriminator D is trained to distinguish between samples from the joint distribution $q(c, s_m)$ and samples from the product of marginals.[2] This

[1] The trivial solution also applies without the use of $\tilde{\mathbf{x}}$ for reconstructions, but with it the multimodal encoder is regularized more strongly.

[2] To sample from the joint distribution, we use the aggregate posterior $q(c, s_m) = \frac{1}{N} \sum_{\mathbf{x}} q_\phi(c, s_m \mid \mathbf{x})$, while for the marginals we use random pairs, e.g., through batch-wise shuffling of samples from the joint.

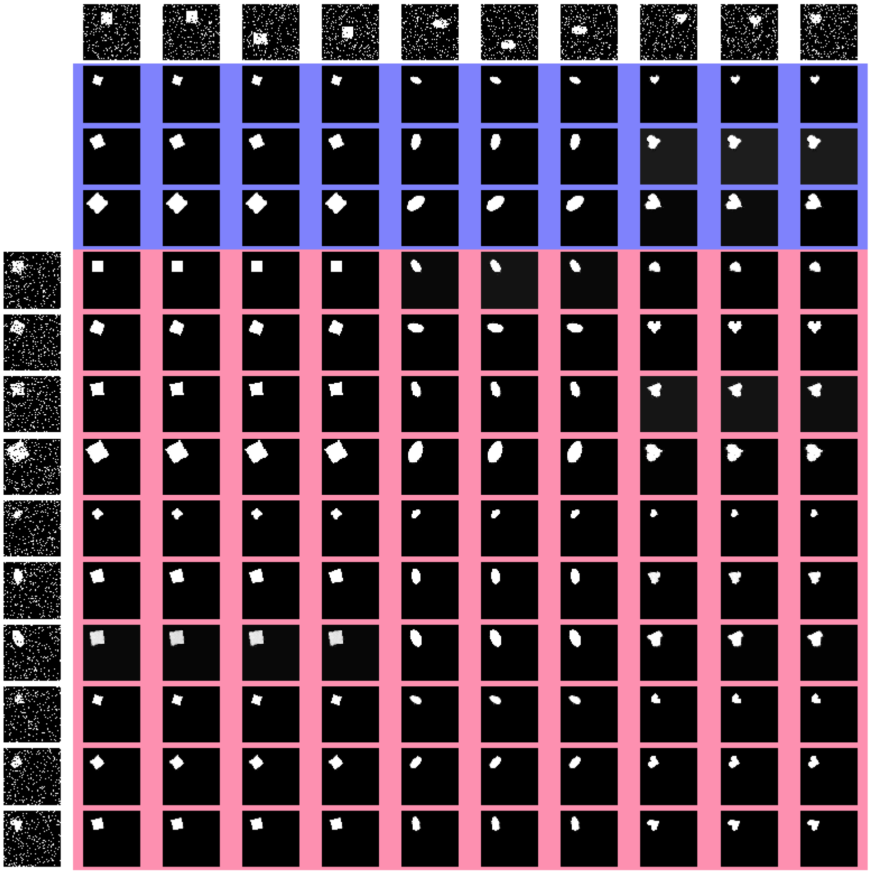


Figure 2: Conditional generation of the second modality given the first. The first column/row shows samples from the first/second modality. Modality-specific factors are fixed along the x-axis and modality-invariant factors are fixed along the y-axis. In blue rows, the modality-specific representation is drawn from the Gaussian prior; in red rows, it is computed from the respective image of the other modality (same index, first row).

procedure is very similar to the one used by Kim and Mnih [9] with the important difference that we do not estimate the total correlation between all elements in a single latent representation, but between partitions c, s_m of the latent space, of which c is shared between modalities.

4 Experiments

To investigate the disentanglement capabilities we use a modified version of the dSprites dataset [13] adapted to a multimodal setting. We create pairs of images (sprites) that share a common factor, but not any of the other generative factors. In particular, we share the shape information between modalities, so that one modality consists of sprites of varying shape, size, and orientation, while the other modality has sprites of varying shape and x/y-position. This allows us to control modality-specific and modality-invariant factors, and to measure how well the model disentangles them. To make the problem more difficult, we flip pixels randomly with a probability of 0.1. The resulting dataset consists of 737280 pairs of sprites and is partitioned into 80/20 percent of training/test data; only the test set is used for the qualitative and quantitative evaluation. Unless otherwise stated, we fix $\beta_m = \beta_c = 1$ and use 5 latent dimensions for both c and s_m in all of the following experiments. The implementation details are described in Appendix C.

Figure 2 presents the quality of the conditional generation. Given a sample x_1 from the first modality (first column), we take its shared representation (i.e., the predicted means from $q_{\phi_1}(c \mid x_1)$) and

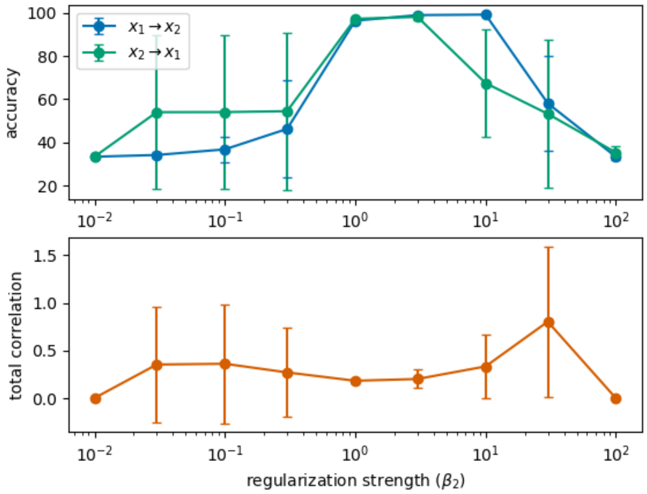


Figure 3: *Top*: Shape-classification accuracy on generated images. The blue curve shows the results from the conditional generation of the second modality given the first, and vice versa for the green line. *Bottom*: Estimated total correlation between learned representations c and s_2. For both plots, points/error-bars represent the means/standard-deviations across 3 runs with different seeds. $\beta_2 \in [0.01, 0.03, 0.1, \ldots, 100]$ are shown on the x-axis which is shared across both subplots and displayed on a log-scale.

generate the opposite modality through the decoder $p_{\theta_2}(x_2 \,|\, c, s_2)$. If s_2 is sampled from the Gaussian prior, this process corresponds to the conditional generation of a missing modality, whereas, if s_2 is computed from an image of the opposite modality (first row), the process resembles an image-to-image translation. The results are indicative of an effective disentanglement, visible in the row-wise consistency of modality-specific factors (scale and orientation) and the column-wise consistency of modality-invariant factors (shape).

Figure 3 investigates whether stronger disentanglement coincides with an improved generative performance. For this experiment, we keep $\beta_1 = \beta_c = 1$ fixed and control the disentanglement between c and s_2 by varying the regularization strength through β_2 only. We measure the generative performance with a separate image classifier (pre-trained on the images from the training set) that predicts the shape information from the conditionally generated images. The results indicate that the conditional generation is best when the total correlation is low (i.e., when the disentanglement is strong). The notable exception is a trivial solution, that can be observed when the regularization gets too strong and the KL-divergence collapses, exhibiting low total correlation only because s_2 carries no information about x_2. Thus, the results suggest that an improved conditional generation coincides with a stronger disentanglement of modality-specific and modality-invariant factors.

The implications of above results are limited to synthetic images and to the special case of two modalities. The total correlation estimate has shown significant variability, as observed in previous studies [15]. The improvement in conditional generation over a non-partitioned latent space (which corresponds to the special case of large β_m for all modalities) is further investigated qualitatively in Appendix B.

5 Conclusion

We have introduced a new latent-variable model that learns a joint distribution of multimodal data and aims to disentangle modality-specific and modality-invariant latent factors. In a controlled setting, we demonstrated the conditional generation of missing modalities and found that an improved generative performance coincides with stronger disentanglement. Our analysis was limited to synthetic images from two modalities. We plan to further investigate the applicability of our method on real multimodal data and on a larger scale.

References

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

[2] Bouchacourt, D., Tomioka, R., and Nowozin, S. (2018). Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *32nd AAAI Conference on Artificial Intelligence*.

[3] Cao, Y. and Fleet, D. J. (2014). Generalized product of experts for automatic and principled fusion of Gaussian process predictions. *arXiv preprint arXiv:1410.7827*.

[4] Doersch, C. (2016). Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*.

[5] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations*.

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

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

[8] Ilse, M., Tomczak, J. M., Louizos, C., and Welling, M. (2019). DIVA: Domain Invariant Variational Autoencoders. *arXiv preprint arXiv:1905.10427*.

[9] Kim, H. and Mnih, A. (2018). Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*.

[10] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic gradient descent. *3rd International Conference on Learning Representations*.

[11] Li, Y. and Mandt, S. (2018). Disentangled sequential autoencoder. In *Proceedings of the 35th International Conference on Machine Learning*.

[12] Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning*.

[13] Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. (2017). dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/.

[14] Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861.

[15] Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. (2019). On Variational Bounds of Mutual Information. In *Proceedings of the 36th International Conference on Machine Learning*.

[16] Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044.

[17] Tsai, Y.-H. H., Liang, P. P., Zadeh, A., Morency, L.-P., and Salakhutdinov, R. (2019). Learning Factorized Multimodal Representations. In *7th International Conference on Learning Representations*.

[18] Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82.

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

[20] Yang, X., Ramesh, P., Chitta, R., Madhvanath, S., Bernal, E. A., and Luo, J. (2017). Deep multimodal representation learning from temporal data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

A Proofs

A.1 Joint probability distribution

$$
\begin{aligned}
p(\mathbf{x}, \mathbf{s}, c) &= \prod_{m=1}^{M} p(c, s_m) p(x_m \mid c, s_m) && (7) \\
&= p(c) \prod_{m=1}^{M} p(s_m) p(x_m \mid c, s_m) && (8)
\end{aligned}
$$

which follows directly from the assumptions on the data generating process (Figure 1).

A.2 Derivation of $\mathcal{L}_{\text{ELBO}}$

Starting from the usual definition of the ELBO (e.g., [4]):

$$
\begin{aligned}
&\mathbb{E}_{q_\phi(c,\mathbf{s} \mid \mathbf{x})} \left[\log p_\theta(\mathbf{x} \mid c, \mathbf{s})\right] - D_{\mathrm{KL}} \left(q_\phi(c, \mathbf{s} \,|\, \mathbf{x}) \,||\, p(c, \mathbf{s})\right) \\
&= \sum_{m=1}^{M} \mathbb{E}_{q_\phi(c,\mathbf{s} \mid \mathbf{x})} \left[\log p_{\theta_m}(x_m \mid c, s_m)\right] - D_{\mathrm{KL}} \left(q_\phi(c, \mathbf{s} \,|\, \mathbf{x}) \,||\, p(c, \mathbf{s})\right) && (9) \\
&= \sum_{m=1}^{M} \mathbb{E}_{q_{\phi_c}(c \mid \mathbf{x}) q_{\phi_m}(s_m \mid x_m)} \left[\log p_{\theta_m}(x_m \mid c, s_m)\right] - \sum_{m=1}^{M} D_{\mathrm{KL}} \left(q_{\phi_m}(s_m | x_m) \,||\, p(s_m)\right) \\
&\quad - D_{\mathrm{KL}} \left(q_{\phi_c}(c | \mathbf{x}) \,||\, p(c)\right) && (10)
\end{aligned}
$$

where the first equality follows from the conditional independence assumption $(x_i \perp x_j) \,|\, c$ for $i \neq j$, and the second equality from the assumption of a factorizing posterior $q(\mathbf{s}, c \mid \mathbf{x}) = q(\mathbf{s} \mid \mathbf{x}) q(c \mid \mathbf{x})$.

A.3 Proportionality to a product of experts

Using Bayes' rule as well as the assumptions on the data generating process (Figure 1):

$$
\begin{aligned}
p(c \mid \mathbf{x}) &= \frac{p(c) p(\mathbf{x} \mid c)}{p(\mathbf{x})} && (11) \\
&= \frac{p(c)}{p(\mathbf{x})} \prod_{m=1}^{M} p(x_m \mid c) && (12) \\
&= \frac{p(c)}{p(\mathbf{x})} \prod_{m=1}^{M} \frac{p(c \mid x_m) p(x_m)}{p(c)} && (13) \\
&= \frac{\prod_{m=1}^{M} p(x_m)}{p(\mathbf{x}) p(c)^{M-1}} \prod_{m=1}^{M} p(c \mid x_m) && (14) \\
&\propto \frac{1}{p(c)^{M-1}} \prod_{m=1}^{M} p(c \mid x_m)\,. && (15)
\end{aligned}
$$

Similar to Wu and Goodman [19] we use a product $p(c) \prod_{m=1}^{M} q_{\phi_m}(c \mid x_m)$ of variational posteriors approximating $p(c \mid x_m) p(c)$ and one "prior expert" $p(c)$.

A.4 Product of Gaussians

A product of multi-dimensional Gaussians is also a Gaussian with means μ and covariance matrix V that can be computed in closed form [3, 19]:

$$
\begin{aligned}
V(\mathbf{x}; \phi_c) &= \left(\sum_{m=1}^{M} V^{-1}(x_m; \phi_c)\right)^{-1} && (16) \\
\mu(\mathbf{x}; \phi_c) &= \mu(x_m; \phi_c) V^{-1}(x_m; \phi_c) V(\mathbf{x}; \phi_c) && (17)
\end{aligned}
$$

where $\mu(x_m; \phi_c)$ is the vector of means and $V(x_m; \phi_c)$ the covariance matrix that are output by the unimodal branches of the multimodal encoder (before fusion).

B Additional qualitative results

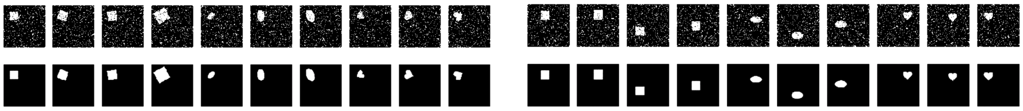


Figure 4: Samples from the first/second modality and their reconstructions.

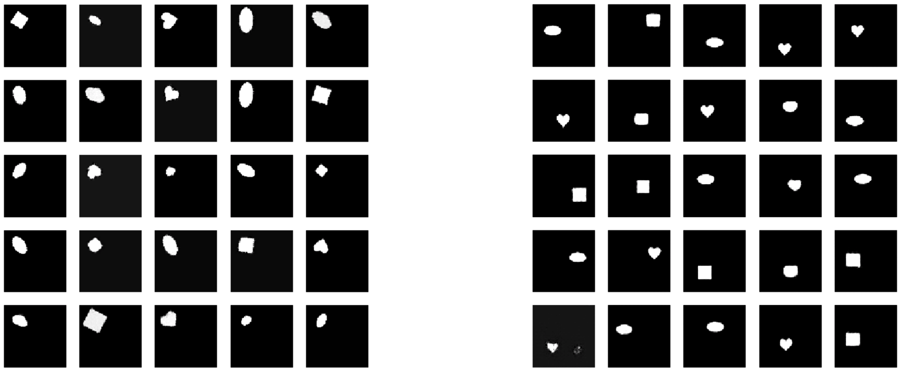
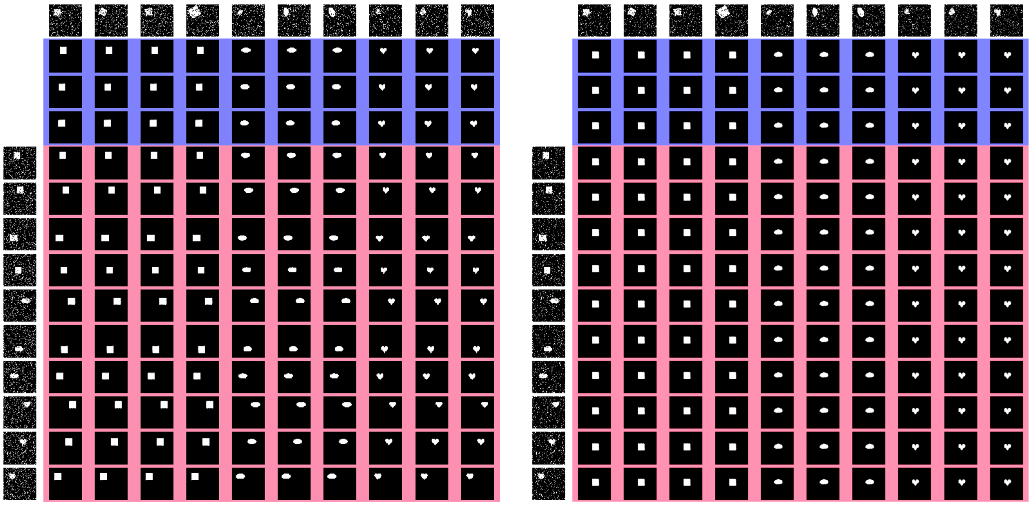


Figure 5: Unconditional generation of 25 samples from each modality.



(a) Ours

(b) Re-implementation of [19]

Figure 6: Conditional generation of the first modality given the second. For our model (left subfigure) we use the same procedure as in Figure 2, but in reverse; for the re-implementation of [19] (right subfigure) conditional samples are drawn from the aggregate posterior $q_{\phi_c}(c \mid \mathbf{x}_2)$. To make the comparison fair, we use 10 latent dimensions for c in the re-implementation, while in our model we use 5 latent dimensions for c and s_m respectively—in total, both models have the same number of parameters. While both models achieve a high shape-classification accuracy for the conditional generation (i.e., we observe column-wise consistency of shapes), only our model generates diverse samples that preserve modality-specific information.

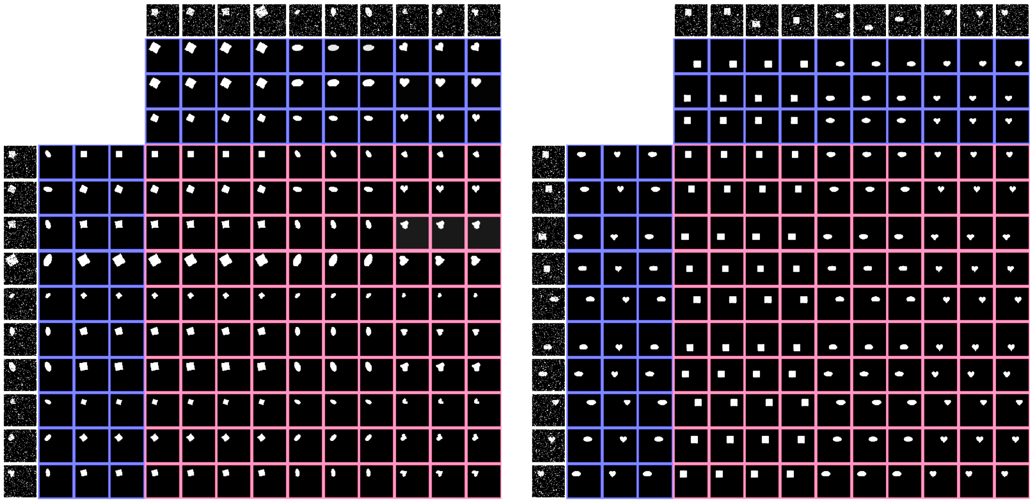


Figure 7: Within-modality swapping of c and s_m for the first modality (left subfigure) and second modality (right subfigure). In blue rows of ten images, s_m is drawn from the Gaussian prior, while in blue columns of ten images, c is drawn from the Gaussian prior. Again, one can observe row-wise consistency of modality-specific factors (scale and orientation, or x/y-position respectively) and column-wise consistency of modality-invariant factors (shape).

C Implementation Details

The network architectures are based on Kim and Mnih [9] and adapted to a multimodal setting. The encoders $p_{\phi_m}(s_m \mid x_m)$ and $p_{\phi_c}(c \mid x_m)$ share most of their weights; only the topmost layers are separate linear fully connected layers that map from 256 to 5 latent dimensions for both the estimated means and log-variances (which are propagated through by using the reparameterization trick).

In each experiment, the model is trained for 100 epochs on the multimodal dSprites dataset and optimized with Adam [10] using learning rate 0.001 and betas (0.9, 0.999). Unless otherwise noted, we use $\beta_m = \beta_c = 1$ for all modalities.

Table 1: Architecture for the shared backbone of the encoders $q_{\phi_m}(s_m \mid x_m)$ and $q_{\phi_c}(c \mid x_m)$. As input, the network takes a 64×64 dSprites image. The parameters for Conv2d are output channels, kernel size, and stride.

Block	Details
1	Conv2d(32, 4, 2), ReLU
2	Conv2d(32, 4, 2), ReLU
3	Conv2d(64, 4, 2), ReLU
4	Conv2d(64, 4, 2), ReLU
5	Conv2d(256, 4, 1), ReLU

Table 2: Architecture for the decoder $p_{\theta_m}(x_m \,|\, c, s_m)$. As input, the decoder takes the 10-dimensional concatenation (c, s_m). The parameters for ConvTranspose2d are output channels, kernel size, and stride.

Block	Details
1	Linear(256), ReLU
2	ConvTranspose2d(64, 4, 1), ReLU
3	ConvTranspose2d(64, 4, 2), ReLU
4	ConvTranspose2d(32, 4, 2), ReLU
5	ConvTranspose2d(32, 4, 2), ReLU
6	ConvTranspose2d(1, 4, 2), ReLU

Table 3: Architecture for the total correlation estimator. As input, the network takes the 10-dimensional concatenation (c, s_m). The network parameters are optimized to discriminate between pairs (c, s_m) and random pairs (c', s_m) (batch-wise permutations) by minimizing the cross-entropy loss as in Kim and Mnih [9].

Block	Details
1	Linear(1000), LeakyReLU(0.2)
2	Linear(1000), LeakyReLU(0.2)
3	Linear(1000), LeakyReLU(0.2)
4	Linear(1000), LeakyReLU(0.2)
5	Linear(1000), LeakyReLU(0.2)
6	Linear(2)

Table 4: Architecture for the shape classifier. The network takes as input original dSprites samples of size 64×64 and it is trained by minimizing the cross-entropy loss on an output of size 3, because there are three types of shapes (square, ellipse, heart). The parameters for Conv2d are output channels, kernel size, and stride.

Block	Details
1	Conv2d(32, 4, 2), ReLU
2	Conv2d(32, 4, 2), ReLU
3	Conv2d(64, 4, 2), ReLU
4	Conv2d(64, 4, 2), ReLU
5	Conv2d(256, 4, 1), ReLU
6	Linear(3)