Learning Disentangled Representations with Identifiable Diffusion Models

Term Paper
Advanced Machine Learning Seminar 2023/24

Constantin Kühne and Till Zemann

Hasso Plattner Institute for Digital Engineering
{constantin.kuehne, till.zemann}@student.hpi.uni-potsdam.de

Abstract. This work introduces a novel framework that combines the Denoising Diffusion Probabilistic Model (DDPM) with an identifiable Variational Autoencoder (iVAE) for disentangled representation learning. Our approach leverages the strengths of diffusion models in capturing data distributions across scales and the iVAE's capability to learn identifiable latent representations. Through extensive quantitative experimentation using various disentanglement metrics to compare our model to baseline models, and a qualitative evaluation on the Shapes3d dataset, we demonstrate our method's ability to generate disentangled representations.

Keywords: Disentanglement · Identifiable VAE · DDPM · Nonlinear ICA

1. Introduction

In recent years, disentangled representation learning has emerged as a pivotal area of research. This approach aims to decompose high-dimensional data such as images into their underlying low-dimensional generative factors, which increases the interpretability [19] and generalizability [2], as well as the predictive accuracy for downstream tasks, such as classification [12] and segmentation [6].

We propose a new model for disentangled representation learning that merges the Denoising Diffusion Probabilistic Model (DDPM) [9] with the identifiable VAE (iVAE) [13]. The iVAE conditions the latent representation prior $p_\psi(z)$ on additional observed variables to ensure identifiability up to invertible component-wise transformations [10]. In the context of disentangled representation learning, identifiability means that the model not only learns the marginal distribution of the data but also learns the true joint distribution of the hidden generative factors and observed data.

We hypothesize that incorporating diffusion processes within an iVAE architecture will improve the disentanglement of latent representations. By extending the architecture with a score-based DDPM, we tap into the ability of diffusion models to methodically learn features – initially learning broad, global variations in the

early stages of diffusion, followed by refining and capturing more fine-grained variations in the later diffusion steps. This systematic approach allows for a nuanced representation of data variance across various scales.

1.1. Definition of Disentanglement

In a disentangled representation [8] created by a model, the different generative factors of data, such as the color or shape of an object in an image, correspond to distinct latent dimensions and remain invariant to changes in all other latent variables.

More formally, disentanglement describes the concept where observational data are transcribed into a lower-dimensional space in which there is a one-to-one correspondence between latent variables and generative factors.

1.2. Measuring Disentanglement

Discrete MIG: The Discrete Mutual Information Gap (MIG) [5] that is computed over factors is one of many quantitative metrics designed to assess the level of disentanglement in the learned representations of a model. Formally, the MIG is defined for a given set of latent representations, μ, and a set of ground truth generative factors, y, as follows:

$$\text{MIG} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{H(y_k)} \left(\max_i I(\mu_i, y_k) - \max_{j \neq i} I(\mu_j, y_k) \right) \tag{1}$$

where $I(\mu_i y_k)$ represents the mutual information between the discretized i-th latent dimension and the k-th generative factor. Furthermore, $H(y_k)$ is the entropy of the k-th generative factor, and K is the total number of generative factors. The MIG formula captures the difference between the highest mutual information of a latent variable with a given factor and the second-highest mutual information of any other latent variable with that factor. By normalizing this difference with the entropy of the generative factor, $H(y_k)$, we account for the inherent uncertainty in the factor itself, thereby ensuring that the score reflects the disentanglement relative to the informativeness of the factor.

The MIG score inherently favors scenarios where each latent dimension correlates with a unique generative factor. Thus, models with high MIG scores are interpreted as having achieved disentanglement, whereas low scores suggest entanglement – where factors are mixed within dimensions.

1.3. Related Work

Our work is mainly based on two papers. Firstly, Abstreiter et al. [1] introduce diffusion-based representation learning (DRL), a new framework for unsupervised representation learning using diffusion-based generative models with the denoising score-matching objective and an additional encoder. The approach shows significant

improvements in downstream tasks such as semi-supervised image classification and offers a new method for representation learning without supervision that competes with traditional autoencoders and contrastive learning methods.

Secondly, Khemakhem et al. [13] present a novel approach to independent component analysis (ICA) using a variant of the Variational Autoencoder (VAE) termed the identifiable VAE (iVAE). The paper's core contribution is the theoretical framework that allows for the identifiability of latent variables in the nonlinear ICA setting.

2. Methodology

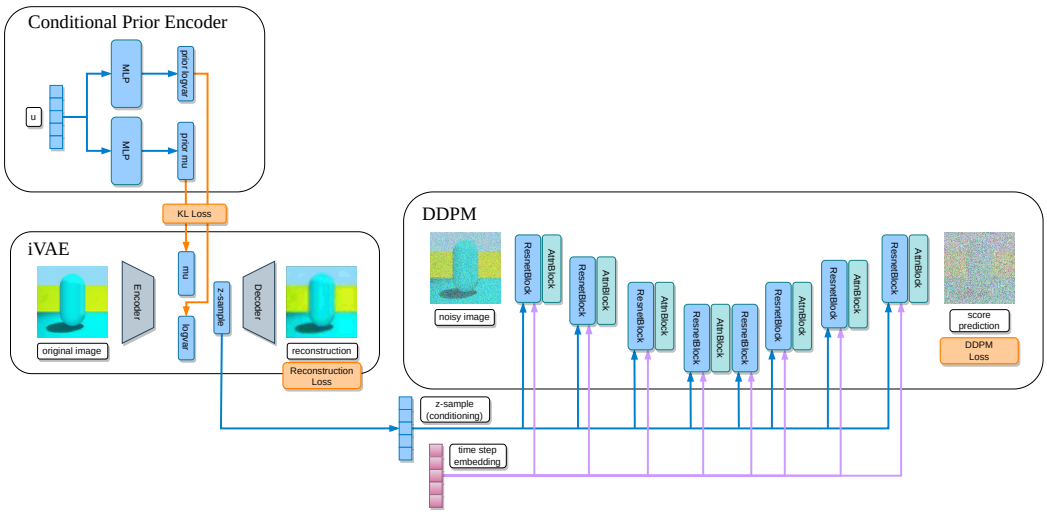

Figure 1: Model architecture and overview where the losses (illustrated in orange) are applied.

2.1. Architecture

The foundation of our model is an identifiable Variational Autoencoder (iVAE) [13], which includes an encoder that learns $p_\theta(z|x)$ and a decoder that learns $p_\omega(\hat{x}|z)$. The encoder maps the original image to a latent space representation, while the decoder reconstructs the image from this latent space. What sets the iVAE apart from regular VAEs is an additional encoder that learns a prior $p_\psi(z|u)$ conditioned on a vector u with additional information, consisting of the variance and optionally the mean, of the label y. If the condition is met that the dimensions of u modulate the variance of the generative factors, then using the prior as a regularization target for z enables the iVAE to learn identifiable latent representations. These representa-

tions can reconstruct the generative factors by applying invertible component-wise operations (e.g. permutations and rotations). The iVAE model provides theoretical assurances that the model is able to approximately solve non-linear independent component analysis (ICA), meaning that it can learn the true joint distribution of the hidden generative factors and observed data.

Building upon the iVAE, we integrate a score-based DDPM [18] (see Figure 1). Unlike using VAEs as generative models that solely rely on a latent space for generation, DDPMs gradually denoise data by approximately solving a reverse stochastic differential equation (SDE) to model the data distribution $p_\omega(x|z)$. For this work, we use a variance-preserving SDE (VP SDE) that generally produces good results for images [16].

By integrating a DDPM with the iVAE, we benefit from the DDPMs ability to capture the complex data distribution at various scales of detail. The DDPM part of our architecture consists of several stacked residual blocks within a U-Net [15] that are conditioned on the latent vector obtained from the iVAE encoder, providing a conditioned diffusion process that is capable of generating samples with higher fidelity compared to the iVAE decoder.

2.2. Loss

Our framework's optimization objective is composed of three weighted loss functions: the score-based DDPM loss (3), a KL-Loss between the conditional prior and the encoded samples, and a reconstruction loss between the iVAE reconstruction $\hat{x}$ and original sample x. The combined loss is given by:

$$\mathcal{L} = \mathcal{L}_{\text{DDPM}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} \tag{2}$$

The hyperparameters λ_{rec} and λ_{DDPM} weight the contributions of the iVAE in comparison to the DDPM part of the model. Increasing both λ_{rec} and λ_{DDPM} will place a greater emphasis on the iVAE losses, lowering the contribution of the DDPM loss.

The first term of the combined objective consists of the weighted score-matching DDPM loss [17, 18]:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t,x_0,x_t}\left[\lambda(t)\left|\left|\nabla_x \log p_t(x) - s_\phi(x_t, z, t)\right|\right|_2^2\right] \tag{3}$$

where we train the predicted score $s_\phi(x_t, z, t) = \text{DDPM}_\phi(x_t, z, t)$ to match the actual score $\nabla_x \log p_t(x)$, with $\lambda(t) = \sigma^2(t)$ being a time-dependent weighting factor of the score-matching loss at time t to account for increasing variance in the diffusion process. Using $\sigma^2(t)$ as the weighting function results in the KL-divergence objective presented by Song et al. [17].

Next, we have two weighted losses that control the weight put on iVAE reconstructions and the KL divergence between the mean and standard deviation priors and latent code. The reconstruction loss is given by $\mathcal{L}_{\text{rec}} = \mathbb{E}\left[||x - \hat{x}||_2^2\right]$. Lastly, we have the KL loss that is tasked with pulling the posterior towards the prior: $\mathcal{L}_{\text{KL}} = \mathbb{E}\left[KL\left(q_\theta(z|x)||p_\psi(z)\right)\right]$.

3. Evaluation

3.1. Dataset

We use the Shapes3d dataset [4] for our experiments and results. This dataset consists of 480,000 RGB images with a size of 64 x 64 pixels. They are images of various objects that stand on a floor and before a background. Each image contains six generative factors. These describe the image's floor color, background color, and background orientation as well as the object's color, scale/size, and shape (see Figure 2). Some of the generative factors are of categorical nature, for example, there are four kinds of shapes (see Figure 3). Other factors, such as the color of the floor, object, and background are categorical because of the dataset's finite size but can also be interpreted as continuous dimensions.

3.2. Experiments

We conduct a range of quantitative and qualitative experiments, including hyperparameter tuning through grid search and Bayesian optimization (see Table 1). Additionally, we assess the replicability of the best-performing hyperparameters by training models with different seeds. Our experiments also involve fixing the weights of the iVAE and using the pretrained weights of a bad-performing iVAE and our best iVAE to test whether the conditioning of the DDPM on the iVAE latent vector z is working as expected. Lastly, we qualitatively validate that the generations of the iVAE decoder align with the generations of the DDPM.

Using the described hyperparameter tuning protocols, we first rank trained models based on the discrete MIG score (1) and qualitatively evaluate the disentanglement of image generations by the iVAE decoder and DDPM when interpolating each dimension of the iVAE's latent space. Following this procedure, the best hyperparameters for the loss coefficients are $\lambda_{\text{KL}} = 0.0005$ and $\lambda_{\text{rec}} = 0.1$. Additionally, not changing the mean of the conditioning vector u yields better results.

Tuned Parameters		
Parameter	**Search space**	**Best value**
λ_{KL}	[1e-5, 0.1] (log uniform)	5e-4
λ_{rec}	[0.01, 5] (uniform)	0.1
dataset_modulate_mean	{true, false}	false

Fixed Parameters	
Parameter	**Value**
batch size	32
learning rate	2e-4
number of epochs	10

Table 1: Parameters and search space used for hyperparameter tuning.

4. Results

To test our hypothesis that integrating the diffusion process improves the disentanglement of latent representations, we conduct a comprehensive empirical study (see Table 2) that demonstrates our method's ability to learn disentangled representations, using various disentanglement metrics on the Shapes3d dataset [4]. Additionally, we validate the results by comparing our model's qualitative generations (see Figures 4 to 6) to baselines (see Figures 7 and 8) when interpolating individual dimensions of the latent space.

The experimental results indicate that our approach can, in certain instances, outperform or match traditional VAEs and iVAEs in terms of disentanglement, as seen in the comparative metrics over the hyperparameter sweep, which is provided in Table 2. This is particularly notable in the context of the Spearman Mean Correlation Coefficient (MCC) with a mean of 0.23 and a maximum of 0.51 and Disentanglement metric with a mean of 0.13 and a maximum of 0.39, where our iVAE+DDPM framework demonstrates the ability to separate the latent factors associated with the different generative aspects of the dataset better than the other models.

Despite the VAE and iVAE sometimes showing superior performance to the iVAE+DDPM in the quantitative results, the addition of the DDPM has enhanced the model's ability to generate more nuanced and detailed interpolation. This synergy between the iVAE's structured latent space and the DDPM's generative capabilities has led to more refined and coherent image generations, as showcased in Figure 4 in comparison to Figures 7 and 8.

Additionally, from the interpolations and latent representations of our best-performing model (see Figures 4 to 6) we observe that it almost perfectly learns three latent variables, namely the object color, the background color as well as the floor color. We can also see that the model encodes the object shape which it does not learn as well as the other factors, as can be seen in the metrics with a rather low Discrete MIG score of only 0.059 in comparison to the other factors (see Dim 1 in Figure 6), and also visually in the interpolations in Figure 4. We therefore conclude that learning the color is much easier for this model than the other latent variables. In particular, it missed two factors, the object size, and the background orientation completely. Lastly, when looking at Figure 6, one can count the modes that indicate how many discrete values of the different latent variables the model has learned. From the visual inspection, we infer that the model probably did not find all of the distinct values for every factor.

Another important finding from the sweep over multiple random seeds to test whether the best-performing hyperparameters of the iVAE+DDPM model can consistently achieve the same performance (see Table 3) is that the iVAE+DDPM does not reach the same results on average, indicating that the proposed model is not

very robust yet.

When using and fixing the pretrained weights of the iVAE of the best- and a bad-performing iVAE+DDPM model to train new models, we can observe that the interpolations of the newly trained models are of the same quality. Which in return means that the conditioning of the DDPM on the latent vector z of the iVAE is working. This is also further solidified by qualitatively comparing the interpolations of the iVAE and DDPM which match for good models (e.g. compare Figure 4 and Figure 5).

Metrics	**VAE**	**iVAE**	**iVAE + DDPM**
Discrete MIG	$0.06 \pm 0.04/0.14$	$\mathbf{0.19 \pm 0.07}/0.34$	$0.16 \pm 0.10/\mathbf{0.39}$
Spearman MCC	$0.20 \pm 0.13/0.49$	$0.19 \pm 0.09/0.39$	$\mathbf{0.23 \pm 0.14/0.51}$
Disentanglement	$0.09 \pm 0.09/0.26$	$0.11 \pm 0.10/0.30$	$\mathbf{0.13 \pm 0.11/0.39}$
Modularity	$\mathbf{0.51 \pm 0.25/0.87}$	$0.41 \pm 0.12/0.63$	$0.45 \pm 0.23/0.77$

Table 2: Disentanglemnt metrics on the Shapes3d dataset showing the mean, standard deviation, and max per metric (in that order) for a hyperparameter grid search to compare the robustness across hyperparameters of our model with the VAE and iVAE baselines.

5. Limitations and Future Work

Our proposed framework has demonstrated some promising results in learning disentangled representations. However, it is not yet very robust and does not consistently achieve disentangled results. We identified some notable limitations that might contribute to this and could be addressed in future work:

Modelling Mismatch for Discrete Generative Factors. The iVAE in our framework utilizes Gaussian distributions for the latent space, which is a common approach for continuous representations. However, some of the generative factors in the Shapes3d dataset are inherently discrete. Thus, using Gaussian latents might be a poor modelling choice, especially for the object shapes. Future work could explore the incorporation of a discrete iVAE, that uses Gumbel-Softmax [11] or straight-through gradients [3] to learn categorical distributions. This adjustment could make it easier for the network to capture the discrete generative factors, potentially avoiding the learning collapse to only a few of the factors.

Prior Learning. Our current model regularizes the latent representations to align with a conditioned prior. However, the training process could be further improved by also encouraging the prior to be more predictable by the latent representation.

This idea, known as KL balancing [7], makes use of the fact that the KL divergence is not symmetric. Therefore, one can use $KL\left(q_\theta(z|x)||p_\psi(z)\right)$ and $KL\left(p_\psi(z)||q_\theta(z|x)\right)$ as losses with different weight coefficients to minimize the KL divergence more aggressively with respect to the prior than the representations. This improves the prior and avoids over-regularization of the representations towards an undertrained prior. Implementing KL balancing in our framework could potentially lead to more stable, disentangled training runs.

Validation on Other Datasets. While our current evaluation on the Shapes3d dataset has shown that the conditioning on the latent vectors works and the DDPM outputs align with the iVAE decoder for the best runs, it is crucial to assess the generalizability and robustness of our proposed method across different datasets. Future work should aim to validate the best hyperparameters and model configurations on other disentanglement benchmarks, such as the dSprites dataset [14].

6. Conclusion

Our research presents a promising direction in disentangled representation learning by integrating diffusion models into the iVAE architecture. Despite still facing challenges related to model robustness, our method shows potential to improve the disentanglement of the learned factors. We furthermore lay a path for future work that can address and potentially mitigate the current limitations through model adjustments to further validate and refine our approach.

References

[1] K. Abstreiter, S. Mittal, S. Bauer, B. Schölkopf, and A. Mehrjou. *Diffusion-Based Representation Learning*. 2022. arXiv: 2105.14257 [cs.LG].

[2] Y. Bengio, A. C. Courville, and P. Vincent. "Unsupervised Feature Learning and Deep Learning: A Review and New Perspectives". In: *CoRR* abs/1206.5538 (2012). arXiv: 1206.5538.

[3] Y. Bengio, N. Léonard, and A. Courville. *Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation*. 2013. arXiv: 1308.3432 [cs.LG].

[4] C. Burgess and H. Kim. *3D Shapes Dataset*. https://github.com/google-deepmind/3d-shapes/. 2018.

[5] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud. *Isolating Sources of Disentanglement in Variational Autoencoders*. 2019. arXiv: 1802.04942 [cs.LG].

[6] S. Chu, D. Kim, and B. Han. *Learning Debiased and Disentangled Representations for Semantic Segmentation*. 2021. arXiv: 2111.00531 [cs.CV].

[7] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba. *Mastering Atari with Discrete World Models*. 2022. arXiv: 2010.02193 [cs.LG].

[8] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner. *Towards a Definition of Disentangled Representations*. 2018. arXiv: 1812.02230 [cs.LG].

[9] J. Ho, A. Jain, and P. Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG].

[10] A. Hyvarinen, I. Khemakhem, and H. Morioka. *Nonlinear Independent Component Analysis for Principled Disentanglement in Unsupervised Deep Learning*. 2023. arXiv: 2303.16535 [cs.LG].

[11] E. Jang, S. Gu, and B. Poole. *Categorical Reparameterization with Gumbel-Softmax*. 2017. arXiv: 1611.01144 [stat.ML].

[12] J. Jia, F. He, N. Gao, X. Chen, and K. Huang. *Learning Disentangled Label Representations for Multi-label Classification*. 2022. arXiv: 2212.01461 [cs.CV].

[13] I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen. *Variational Autoencoders and Nonlinear ICA: A Unifying Framework*. 2020. arXiv: 1907.04809 [stat.ML].

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

[15] O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[16] Y. Song. *Generative Modeling by Estimating Gradients of the Data Distribution*. https://yang-song.net/blog/2021/score/. May 2021.

[17] Y. Song, C. Durkan, I. Murray, and S. Ermon. *Maximum Likelihood Training of Score-Based Diffusion Models*. 2021. arXiv: 2101.09258 [stat.ML].

[18] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *CoRR* abs/2011.13456 (2020). arXiv: 2011.13456.

[19] X. Zhu, C. Xu, and D. Tao. *Where and What? Examining Interpretable Disentangled Representations*. 2021. arXiv: 2104.05622 [cs.CV].

A. Appendix

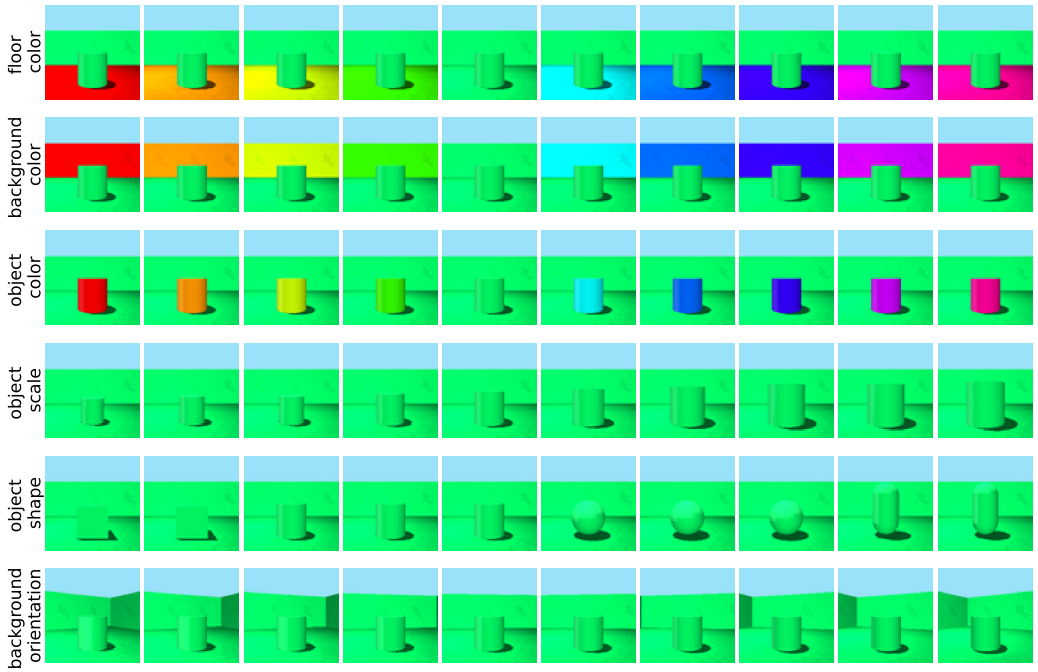

Figure 2: Example images from the Shapes3d dataset visualizing interpolation across generative factors.

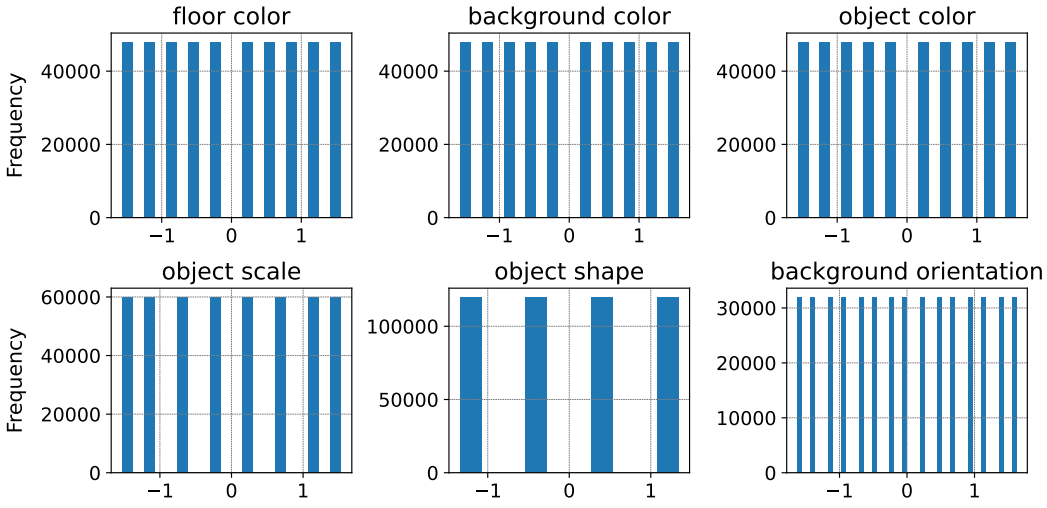

Figure 3: Distributions of the Shapes3d dataset generative factors.

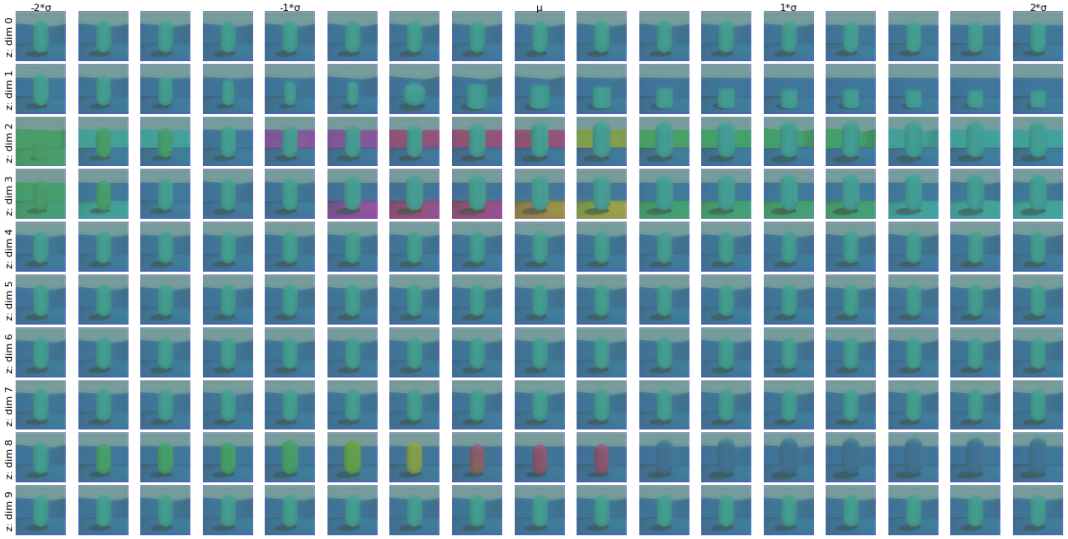

Figure 4: Conditional DDPM generations (using the best iVAE+DDPM model) for feature-space interpolations of one latent dimension at a time, showcasing disentangled representations.

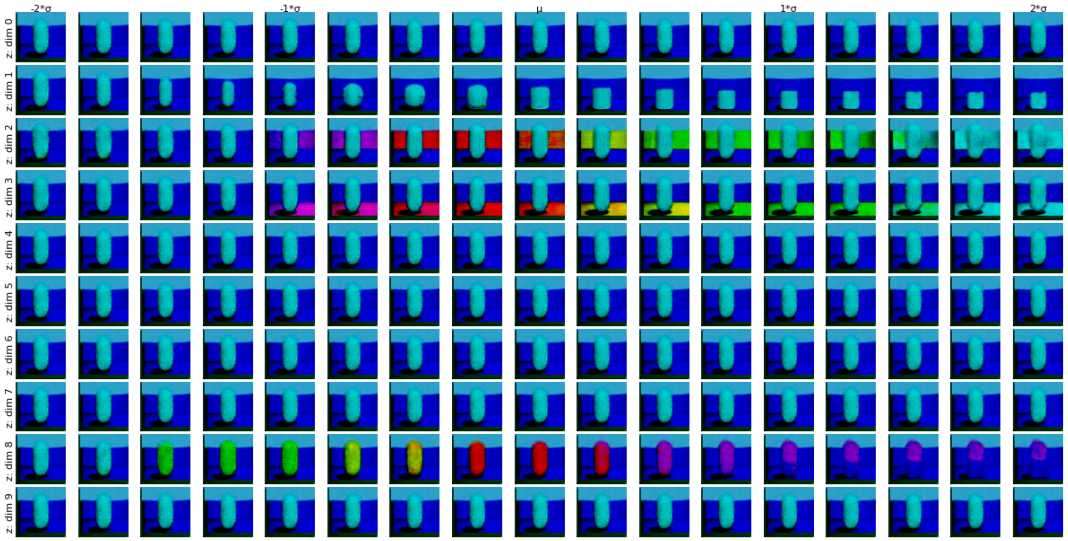

Figure 5: Conditional iVAE generations (using the best iVAE+DDPM model) for feature-space interpolations of one latent dimension at a time, showcasing disentangled representations.

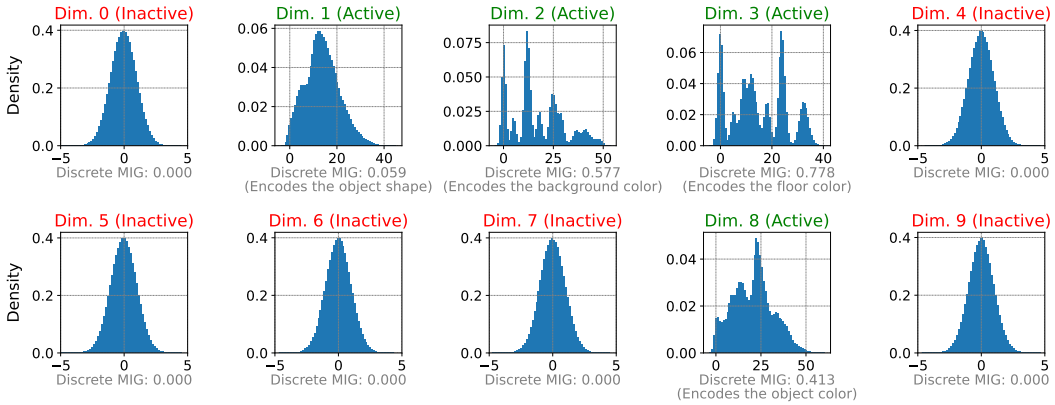

Figure 6: Distributions of the learned Shapes3d z encodings obtained by passing every image in the Shapes3d dataset through the iVAE encoder and sampling a latent vector. The model has learned 4 out of 6 ground truth factors, encoded in dimensions 1,2,3, and 8. All other latent dimensions do not encode any information and thus show the normally distributed samples of z from the iVAE encoder.

Metrics	**Best qualitative iVAE+DDPM**	**Different Seeds (N=5)**
Discrete MIG	0.39	$0.26 \pm 0.09/0.34$
Spearman MCC	0.21	$0.11 \pm 0.08/0.21$
Disentanglement	0.21	$0.13 \pm 0.11/0.25$
Modularity	0.39	$0.35 \pm 0.11/0.47$

Table 3: Comparison of the best-performing iVAE+DDPM model (in terms of qualitative generations) and aggregate values over different random seeds using the same hyperparameters to test the model's robustness. For the sweep over random seeds, the mean, standard deviation, and maximum values are reported for each metric.

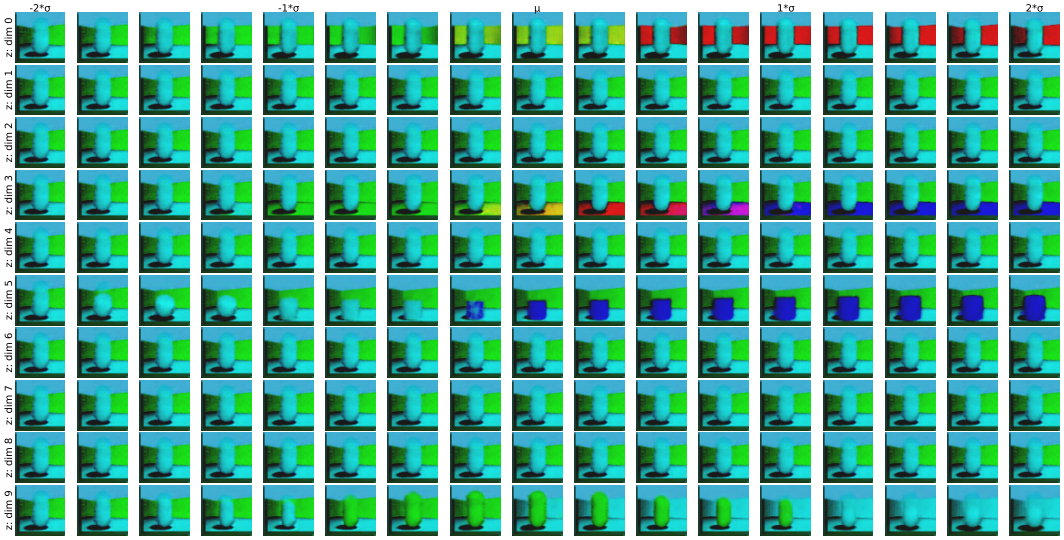

Figure 7: Conditional generations using the best iVAE baseline model for feature-space interpolations of one latent dimension at a time, showcasing disentangled representations.

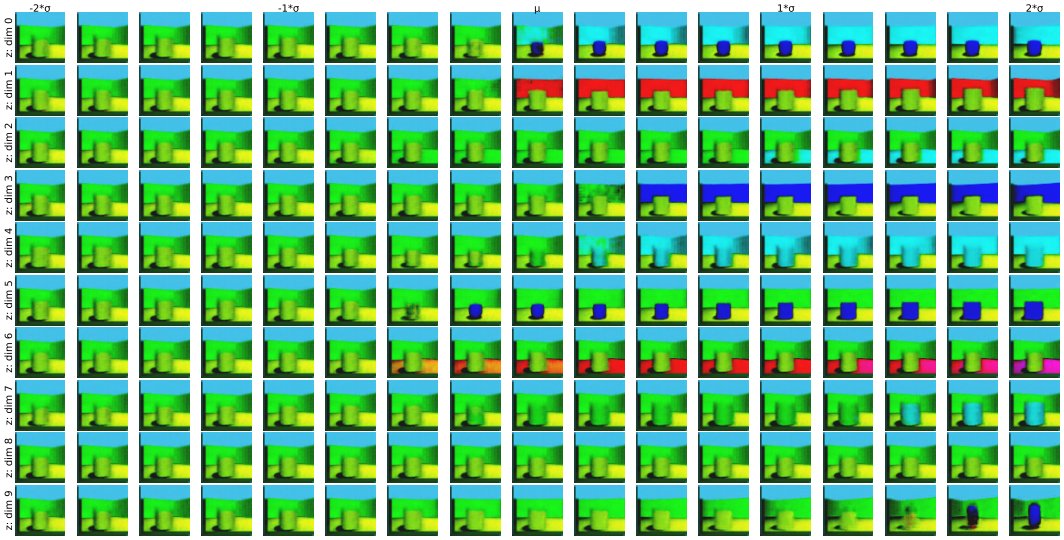

Figure 8: Conditional generations using the best VAE baseline model for feature-space interpolations of one latent dimension at a time, showcasing disentangled representations.