A. Proof of Proposition 1

Let $f^* : \mathcal{C} \to \mathcal{Z}$ be a diffeomorphism that transforms $p^*(\mathbf{c})$ (the true prior distribution of the factors of variation) into the fixed model prior $p(\mathbf{z})$:

$$p^{*(\mathbf{c})}(\mathbf{c}) = p^{(\mathbf{z})}(f^*(\mathbf{c}))\,|\det \nabla_{\mathbf{c}} f^*(\mathbf{c})| \ . \tag{5}$$

Superscripts are included here to clarify which random variable a distribution is defined over, but will be often omitted. The existence of f^* implies that the true generative model $p^*(\mathbf{x} \,|\, \mathbf{z})$ can be expressed as a composition of a deterministic transformation, $(f^*)^{-1}$, and a stochastic one, $p^*(\mathbf{x} \,|\, \mathbf{c})$.

Similarly, let $f_\theta : \mathcal{C} \to \mathcal{Z}$ be a diffeomorphism, parameterized by θ, that defines a distribution $p_\theta(\mathbf{c})$ (in general different from $p^*(\mathbf{c})$) with support $\mathcal{C}$:

$$p_\theta^{(\mathbf{c})}(\mathbf{c}) = p^{(\mathbf{z})}(f_\theta(\mathbf{c}))\,|\det \nabla_{\mathbf{c}} f_\theta(\mathbf{c})| \ . \tag{6}$$

We will assume that the learned generative model $p_\theta(\mathbf{x} \,|\, \mathbf{z})$ can be expressed as a composition of the learned deterministic transformation f_θ^{-1} and the true $p^*(\mathbf{x} \,|\, \mathbf{c})$:

$$\begin{aligned} p_\theta(\mathbf{x}) &= \int_{\mathbf{z}} p_\theta(\mathbf{x} \,|\, \mathbf{z}) p(\mathbf{z}) d\mathbf{z} \\ &= \int_{\mathbf{c}} p_\theta(\mathbf{x} \,|\, f_\theta(\mathbf{c})) p_\theta(\mathbf{c}) d\mathbf{c} \\ &= \int_{\mathbf{c}} p^*(\mathbf{x} \,|\, \mathbf{c}) p_\theta(\mathbf{c}) d\mathbf{c} \end{aligned} \tag{7}$$

where $\mathbf{c} = f_\theta^{-1}(\mathbf{z})$. Note that the distribution $p_\theta(\mathbf{c})$ is implicitly learned by learning f_θ, and it represents the learned prior distribution over the true factors of variation.

The expected log likelihood we wish to maximize is

$$\mathbb{E}_{p^*(\mathbf{x})}[\log p_\theta(\mathbf{x})] = -\mathcal{H}(p^*(\mathbf{x})) - D_{\mathrm{KL}}(p^*(\mathbf{x}) \| p_\theta(\mathbf{x}))$$

where the differential entropy $\mathcal{H}(p^*(\mathbf{x}))$ is constant with respect to the model parameters, and can therefore be ignored. The KL term can be rewritten as

$$\begin{aligned} D_{\mathrm{KL}}(p^*(\mathbf{x}) \| p_\theta(\mathbf{x})) &= \int_{\mathbf{x}} p^*(\mathbf{x}) \log \frac{p^*(\mathbf{x})}{p_\theta(\mathbf{x})} d\mathbf{x} \\ &= \int_{\mathbf{x}} \int_{\mathbf{c}} p^*(\mathbf{x} \,|\, \mathbf{c}) p^*(\mathbf{c}) \log \frac{p^*(\mathbf{c}) \frac{p^*(\mathbf{x} \,|\, \mathbf{c})}{p^*(\mathbf{c} \,|\, \mathbf{x})}}{p_\theta(\mathbf{c}) \frac{p^*(\mathbf{x} \,|\, \mathbf{c})}{p^*(\mathbf{c} \,|\, \mathbf{x})}} d\mathbf{c}\, d\mathbf{x} \\ &= \int_{\mathbf{x}} \int_{\mathbf{c}} p^*(\mathbf{x} \,|\, \mathbf{c}) p^*(\mathbf{c}) \log \frac{p^*(\mathbf{c})}{p_\theta(\mathbf{c})} d\mathbf{c}\, d\mathbf{x} \\ &= \int_{\mathbf{c}} p^*(\mathbf{c}) \log \frac{p^*(\mathbf{c})}{p_\theta(\mathbf{c})} \int_{\mathbf{x}} p^*(\mathbf{x} \,|\, \mathbf{c}) d\mathbf{x}\, d\mathbf{c} \\ &= D_{\mathrm{KL}}(p^*(\mathbf{c}) \| p_\theta(\mathbf{c})) \ . \end{aligned} \tag{8}$$

Note that, since the KL divergence is always non-negative, the maximum likelihood corresponds to $D_{\mathrm{KL}}(p^*(\mathbf{x}) \| p_\theta(\mathbf{x})) = 0$.

Let a matrix be σ-diagonal if there exists a permutation σ that makes it diagonal. Since by assumption $p^*(\mathbf{c})$ does not factorize while $p(\mathbf{z})$ does, it follows that $\nabla_{\mathbf{c}} f^*(\mathbf{c})$ (the Jacobian of f^*) is not σ-diagonal.[6] However, if the representations $\mathbf{z}$ are disentangled w.r.t. the true factors $\mathbf{c}$ then the Jacobian of f_θ is σ-diagonal. Thus, $f_\theta(\mathbf{c})$ cannot be equal to $f^*(\mathbf{c})$ almost everywhere. This in turn means that $D_{\mathrm{KL}}(p^*(\mathbf{c}) \| p_\theta(\mathbf{c})) > 0$, hence $\mathbb{E}_{p^*(\mathbf{x})}[\log p_\theta(\mathbf{x})] < \mathbb{E}_{p^*(\mathbf{x})}[\log p^*(\mathbf{x})]$. This proves that if the generative model is disentangled w.r.t. the true factors then its expected likelihood is less than the optimal likelihood.

On the other hand, in the general case in which the representations are not necessarily disentangled, we can choose θ such that $f_\theta(\mathbf{c}) = f^*(\mathbf{c})$ almost everywhere, which implies that $D_{\mathrm{KL}}(p^*(\mathbf{c}) \| p_\theta(\mathbf{c})) = 0$. Thus, there exists an entangled model that has optimal likelihood.

We have proved that (i) if the generative model is constrained to be disentangled then the optimal likelihood cannot be achieved, and (ii) if it is *not* constrained to be disentangled then the optimal likelihood *can* be achieved. Equivalently, the optimal likelihood can be attained if and only if the generative model is entangled w.r.t. the true generative factors.

B. Implementation details

Unsupervised Disentanglement methods. For the sake of comparison, the considered disentanglement methods in this work cover the full collection of state-of-the-art approaches in `disentanglement_lib` from Locatello et al. (2019b) based on representations learned by VAEs. The set contains six different methods that enforce disentanglement of the representation by equipping the loss with different regularizers that aim at enforcing the special structure of the posterior aggregate encoder distribution. A detailed description of the regularizer forms used in this work, specifically β-VAE (Higgins et al., 2017a), FactorVAE (Kim & Mnih, 2018), AnnealedVAE (Burgess et al., 2018), DIP-VAE-I, DIP-VAE-II (Kumar et al., 2018) and β-TC-VAE (Chen et al., 2018) is provided in Locatello et al. (2019b). We use the same encoder and decoder architecture with 10 latent dimensions for every model.

Joint distributions of correlated factors in datasets. In Fig. 6 we show the joint probability distributions of the correlated pair of FoV for all datasets and correlation strengths considered in this study. Dataset A, B and C were designed with correlated factors of variation that are ordinal for a

[6] If $p(\mathbf{z})$ factorizes and $\nabla_{\mathbf{c}} f^*(\mathbf{c})$ is σ-diagonal, then $p^*(\mathbf{c})$ also factorizes. Thus, since by assumption $p^*(\mathbf{c})$ does *not* factorize, either $\nabla_{\mathbf{c}} f^*(\mathbf{c})$ is not σ-diagonal or $p(\mathbf{z})$ does not factorize. Because the latter is false by assumption, it must be that $\nabla_{\mathbf{c}} f^*(\mathbf{c})$ (the Jacobian of f^*) is *not* σ-diagonal.

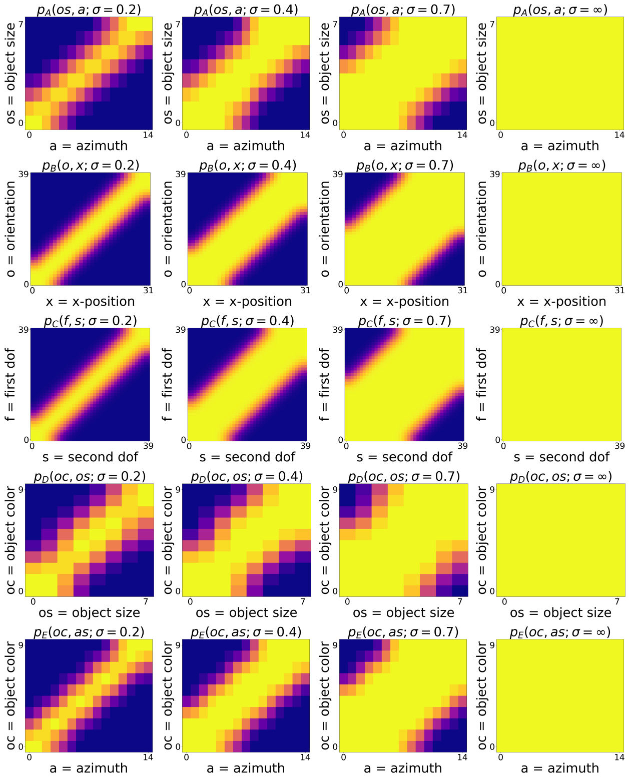

Figure 6. Probability distributions for sampling training data in the correlated pair of FoVs in the respective datasets (A, B, C, D, E) considering correlation strengths of $\sigma = 0.2$, $\sigma = 0.4$, $\sigma = 0.7$ and $\sigma = \infty$, the uncorrelated limit (from left to right).

natural visual interpretation of the traversals. In contrast, datasets D and E contain a correlated factor of variation that has no such natural ordering.

Pairwise entanglement metric. We base the computation of the pairwise entanglement metric on a procedure developed in Locatello et al. (2020a): Starting from the GBT feature importance matrix between encoded FoV and latent codes, a bipartite graph between latent codes and FoVs is constructed such that the edge weights are the corresponding matrix elements. After deleting all edges that have weight smaller than a given threshold, one counts the number of disconnected sub graphs having more than one vertex and how many FoV are still connected with at least one latent. When repeating this computation for different thresholds, one can track at which particular threshold a given pair of FoV is merged with each other in this bipartite graph, resulting in the metric we are reporting. A pair of FoV that are being merged at a higher threshold are statistically more related to each other via shared latent code dimensions. This computation can be not only based on the GBT feature importance matrix but likewise on weight matrices inferred from the mutual information.

corr. strength		$\sigma = 0.2$	$\sigma = 0.4$	$\sigma = 0.7$	$\sigma = \infty$ (uc)
Shapes3D (A)	object size - azimuth	0.38 (0.28)	0.26 (0.25)	0.13 (0.2)	0.08 (0.17)
	median other pairs	0.09 (0.2)	0.09 (0.2)	0.09 (0.19)	0.08 (0.18)
dSprites (B)	orientation - position x	0.17 (0.34)	0.16 (0.31)	0.14 (0.24)	0.11 (0.14)
	median other pairs	0.13 (0.16)	0.13 (0.18)	0.13 (0.19)	0.13 (0.15)
MPI3D (C)	First DOF - Second DOF	0.2 (0.54)	0.19 (0.52)	0.17 (0.5)	0.16 (0.49)
	median other pairs	0.16 (0.25)	0.16 (0.25)	0.15 (0.26)	0.15 (0.25)
Shapes3D (D)	object color - object size	0.29 (0.38)	0.28 (0.31)	-	-
	median uncorrelated pairs	0.07 (0.11)	0.07 (0.11)	-	-
Shapes3D (E)	object color - azimuth	0.25 (0.43)	0.23 (0.3)	-	-
	median uncorrelated pairs	0.1 (0.15)	0.09 (0.15)	-	-

Table 4. Pairwise entanglement scores help to uncover still existent correlations in the latent representation. Mean of the pairwise entanglement scores for the correlated pair (red) and the median of the uncorrelated pairs. We see that stronger correlation leads to statistically more entanglement latents across all datasets studied compared to their baseline pairwise entanglement where the data exhibits no correlations (blue). Each score is the mean across 180 models for each dataset and correlation strength. Scores are based on GBT feature importance; scores in brackets are based on the Mutual Information.

Unfairness between a pair of FoV. The scores reported are based on a notion of demographic parity for predicting a target variable y given a protected and sensitive variable s. Both y and s can be associated with a factor of variation here. Rather than using the global total variation average as defined in Locatello et al. (2019a), we report the individual demographic parities for the correlated factors specifically.

Disentanglement metrics. To measure disentanglement of a learned representation, various metrics have been proposed, each requiring access to the ground truth labels. The BetaVAE score is based upon the prediction of a fixed factor from the disentangled representation using a linear classifier (Higgins et al., 2017a). The FactorVAE score is intended to correct for some failures of the former by utilizing majority vote classifiers based on a normalized variance of each latent dimension (Kim & Mnih, 2018). The SAP score represents the mean distance between the classification errors of the two latent dimensions that are most predictable (Kumar et al., 2018). For the MIG score, one computes the mutual information between the latent representation and the ground truth factors and calculates the final score using a normalized gap between the two highest MI entries for each factor. Finally, a disentanglement score proposed by Eastwood & Williams (2018), often referred to as DCI score, is calculated from a dimension-wise entropy reflecting the usefulness of the dimension to predict a single factor of variation.

C. Additional Results Section 4

C.1. Section 4.2

Latent structure and pairwise entanglement. Our hypothesis that the latent representations are less correlated if the correlation strength is weaker is shown for a model

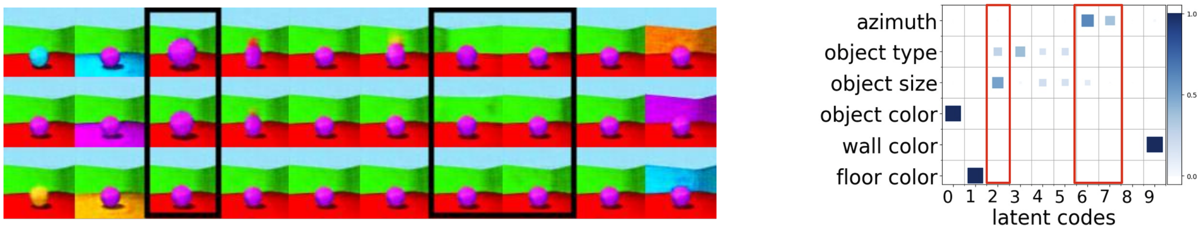

Figure 7. We show latent traversals (left) of the best DCI score model among all 180 trained models with weak correlation ($\sigma = 0.7$) in object size and azimuth. The traversals in latent codes 2, 6, and 7 (highlighted in black), suggest that these dimensions encode no mixture of azimuth and object size compared to the models with stronger correlation. This is supported by the GBT feature importance matrix of this model (right).

(a) $\beta-$VAE

corr. strength		0.2	0.4	0.7	∞ (uc)
Shapes3D (A)	object size - azimuth	0.38	0.25	0.14	0.08
	median uncorrelated pairs	0.08	0.09	0.09	0.07
dSprites (B)	orientation - position x	0.18	0.16	0.14	0.12
	median uncorrelated pairs	0.14	0.14	0.14	0.12
MPI3D (C)	First DOF - Second DOF	0.22	0.19	0.18	0.16
	median uncorrelated pairs	0.16	0.15	0.15	0.14
Shapes3D (D)	object color - object size	0.28	0.3	-	-
	median uncorrelated pairs	0.08	0.07	-	-
Shapes3D (E)	object color - azimuth	0.24	0.25	-	-
	median uncorrelated pairs	0.11	0.09	-	-

(b) Factor-VAE

corr. strength		0.2	0.4	0.7	∞ (uc)
Shapes3D (A)	object size - azimuth	0.48	0.3	0.1	0.03
	median uncorrelated pairs	0.07	0.06	0.07	0.04
dSprites (B)	orientation - position x	0.23	0.2	0.16	0.12
	median uncorrelated pairs	0.14	0.15	0.14	0.13
MPI3D (C)	First DOF - Second DOF	0.23	0.22	0.19	0.18
	median uncorrelated pairs	0.15	0.15	0.14	0.15
Shapes3D (D)	object color - object size	0.36	0.33	-	-
	median uncorrelated pairs	0.02	0.03	-	-
Shapes3D (E)	object color - azimuth	0.3	0.28	-	-
	median uncorrelated pairs	0.1	0.08	-	-

(c) Annealed-VAE

corr. strength		0.2	0.4	0.7	∞ (uc)
Shapes3D (A)	object size - azimuth	0.32	0.25	0.13	0.11
	median uncorrelated pairs	0.09	0.09	0.11	0.1
dSprites (B)	orientation - position x	0.17	0.16	0.14	0.1
	median uncorrelated pairs	0.14	0.15	0.14	0.15
MPI3D (C)	First DOF - Second DOF	0.17	0.17	0.15	0.15
	median uncorrelated pairs	0.15	0.15	0.15	0.14
Shapes3D (D)	object color - object size	0.33	0.28	-	-
	median uncorrelated pairs	0.07	0.08	-	-
Shapes3D (E)	object color - azimuth	0.25	0.19	-	-
	median uncorrelated pairs	0.1	0.1	-	-

(d) β-TC-VAE

corr. strength		0.2	0.4	0.7	∞ (uc)
Shapes3D (A)	object size - azimuth	0.41	0.26	0.09	0.05
	median uncorrelated pairs	0.07	0.09	0.06	0.05
dSprites (B)	orientation - position x	0.18	0.15	0.13	0.11
	median uncorrelated pairs	0.14	0.14	0.13	0.12
MPI3D (C)	First DOF - Second DOF	0.24	0.22	0.19	0.17
	median uncorrelated pairs	0.18	0.17	0.15	0.15
Shapes3D (D)	object color - object size	0.3	0.29	-	-
	median uncorrelated pairs	0.05	0.06	-	
Shapes3D (E)	object color - azimuth	0.23	0.23	-	-
	median uncorrelated pairs	0.09	0.07	-	

(e) Dip-VAE-I

corr. strength		0.2	0.4	0.7	∞ (uc)
Shapes3D (A)	object size - azimuth	0.38	0.24	0.14	0.07
	median uncorrelated pairs	0.1	0.1	0.11	0.09
dSprites (B)	orientation - position x	0.13	0.13	0.12	0.11
	median uncorrelated pairs	0.11	0.11	0.11	0.11
MPI3D (C)	First DOF - Second DOF	0.16	0.15	0.14	0.14
	median uncorrelated pairs	0.13	0.13	0.13	0.13
Shapes3D (D)	object color - object size	0.27	0.25	-	-
	median uncorrelated pairs	0.06	0.06	-	-
Shapes3D (E)	object color - azimuth	0.22	0.22	-	-
	median uncorrelated pairs	0.1	0.1	-	-

(f) Dip-VAE-II

corr. strength		0.2	0.4	0.7	∞ (uc)
Shapes3D (A)	object size - azimuth	0.28	0.23	0.19	0.14
	median uncorrelated pairs	0.12	0.11	0.11	0.11
dSprites (B)	orientation - position x	0.14	0.14	0.13	0.12
	median uncorrelated pairs	0.14	0.14	0.13	0.13
MPI3D (C)	First DOF - Second DOF	0.22	0.21	0.18	0.17
	median uncorrelated pairs	0.15	0.14	0.15	0.15
Shapes3D (D)	object color - object size	0.23	0.2	-	-
	median uncorrelated pairs	0.13	0.12	-	-
Shapes3D (E)	object color - azimuth	0.23	0.19	-	-
	median uncorrelated pairs	0.12	0.13	-	-

Table 5. Pairwise entanglement scores from Table 4 separated along each disentanglement regularizer. Mean of the pairwise entanglement scores for the correlated pair (red) and the median of the uncorrelated pairs. We see that stronger correlation leads to statistically more entanglement latents across all datasets and regularizers studied compared to their baseline pairwise entanglement where the data exhibits no correlations. Each pairwise score is the mean across 30 models for each dataset and correlation strength. Scores are based on GBT feature importance.

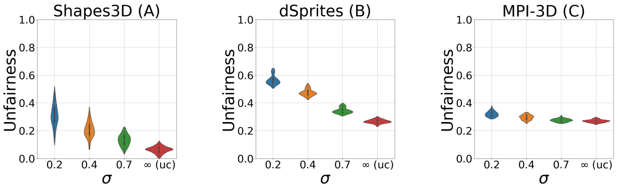

Figure 8. Disentangled representations trained on correlated data are anti-correlated with higher fairness properties. The plots show the mean unfairness scores between the correlated factors with decreasing correlation strength for Shapes3D (A), dSprites (B) and MPI3D-real (C).

on Shapes3D (A) with weak correlation in Fig. 7. Here the latent traversals do not mirror the major and minor axis of the correlated joint distribution.

To make our conclusion more sound we perform an empirical analysis of the pairwise entanglement metrics for the correlated pair vs. the median of all other pairs across the entire unsupervised study on all datasets and models trained. Table 4 shows the results of this analysis, aggregated across all disentanglement models. To avoid missing any particular disentanglement regularizer that might disentangle the correlated pair but is hidden among the combined aggregation, we also separately report the thresholds for each of the six disentanglement models in Table 5. We can clearly see that the correlated pair has a much higher entanglement than the rest of the pairs in the trained models across the full board, thus confirming our conclusion that inductive bias of current SOTA unsupervised disentanglement learners is insufficient. Another pairwise metric that tracks the correlation strength in our scenario is the unfairness score between the correlated pair of factors that is being shown for datasets A, B and C in Fig. 8.

Shortcomings of existing metrics. Following recent studies, we evaluate the trained models with the help of a broad range of disentanglement metrics that aim at quantifying overall success by a single scalar measure. Perhaps surprisingly, Fig. 9 and Fig. 10 show no clear trend among all implemented disentanglement scores w.r.t. correlation strength. The metrics have been evaluated by both, either sampling from the correlated data distribution or from the uncorrelated distribution. Given our extensive analysis of latent entanglements of the correlated FoV pair from above, we thus argue that common disentanglement metrics are limited in revealing those when correlations are introduced into the training data and we partly account this to the averaging procedures across many FoV with these pairs. Note that regarding BetaVAE and FactorVAE this observed trend is to some degree expected as they would yield perfect disentanglement scores even if we would take the correlated ground truth factors or a linear transformation in the case of BetaVAE as the representation.

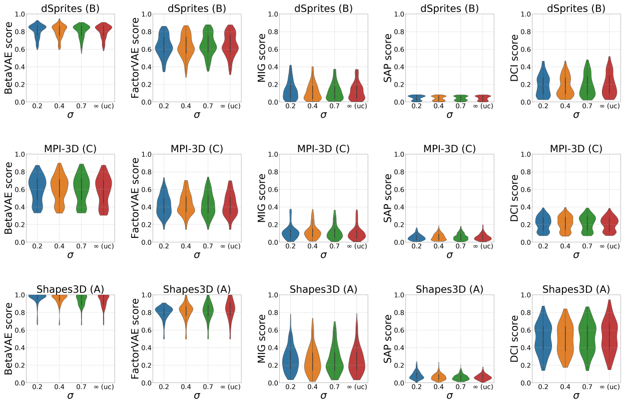

Figure 9. Standard global disentanglement metrics evaluated on the correlated training set showing no clear trend for different correlation strengths.

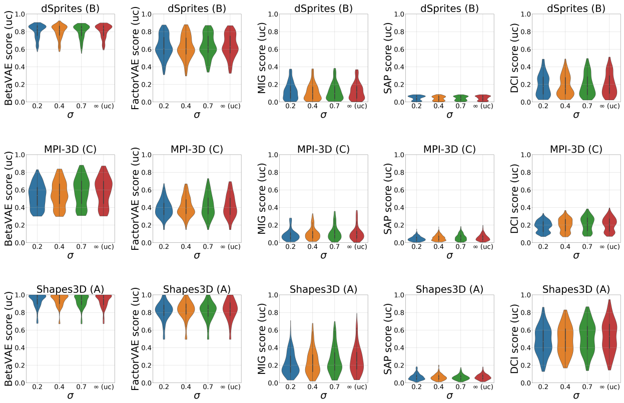

Figure 10. Standard global disentanglement metrics evaluated on the uncorrelated (uc) dataset set showing no clear trend for different correlation strengths.

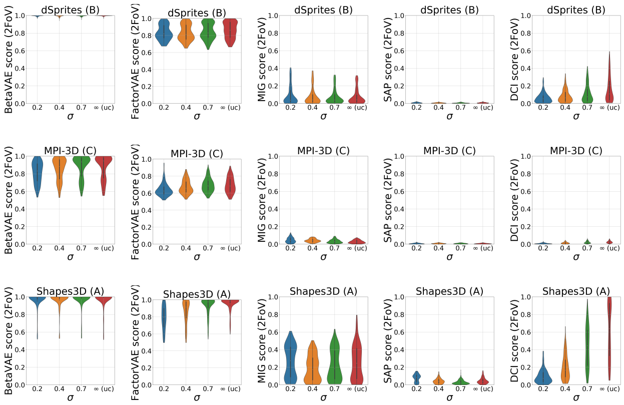

Figure 11. When disentanglement metrics are only evaluated regarding the 2 correlated FoV we can observe the still persisting entanglement in the latents using DCI score.

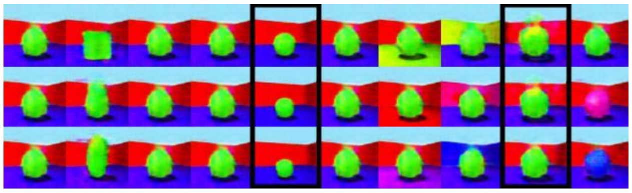

Figure 12. Generalization capabilities towards out-of-distribution test data. Latent traversals from an observations the model has never seen during training. The starting point corresponds to a factor configuration in point number 2 from Fig. 4 Shown are the results of the model with highest DCI score among all 180 trained models on Shapes3d (A) with a very restricted correlation strength $\sigma = 0.2$ in object size and azimuth

As these metrics are defined with respect to the whole set of underlying ground truth FoV and employ various averaging techniques to form a single scalar measure we want to investigate how much the observed latent entanglements are hidden though imperfect disentanglement of other factors. Thus, we evaluate the same metrics but only on the two correlated FoV excluding all other remaining factors. Indeed, as can be seen from Fig. 11, only DCI tracks the entangled latents under this reduced disentanglement score, while the others show no or only weak trends. We refer the interested reader to Locatello et al. (2020a), where a detailed discussion is provided why MIG is not tracking the latent entanglement we observed.

C.2. Section 4.3

Generalization Properties In order to support our conclusion that disentanglement methods can generalize towards unseen FoV configurations we show in Fig. 12 latent traversals originating from OOD point number 2 with smallest object size and largest azimuth. We observe that changes in the remaining factors reliably yield the expected reconstructions.

Emphasizing the generalization results from the main paper, we are visualizing the latent spaces with similar extrapolation and generalization capabilities of four additional models from the two strongest correlation dataset variants of Shapes3d (D) and Shapes3d (E) in Fig. 13. These latent spaces further support that OOD examples are meaningfully encoded into the existing structure of the latent space and that the decoder is equally capable of generating observations from such unseen representations.

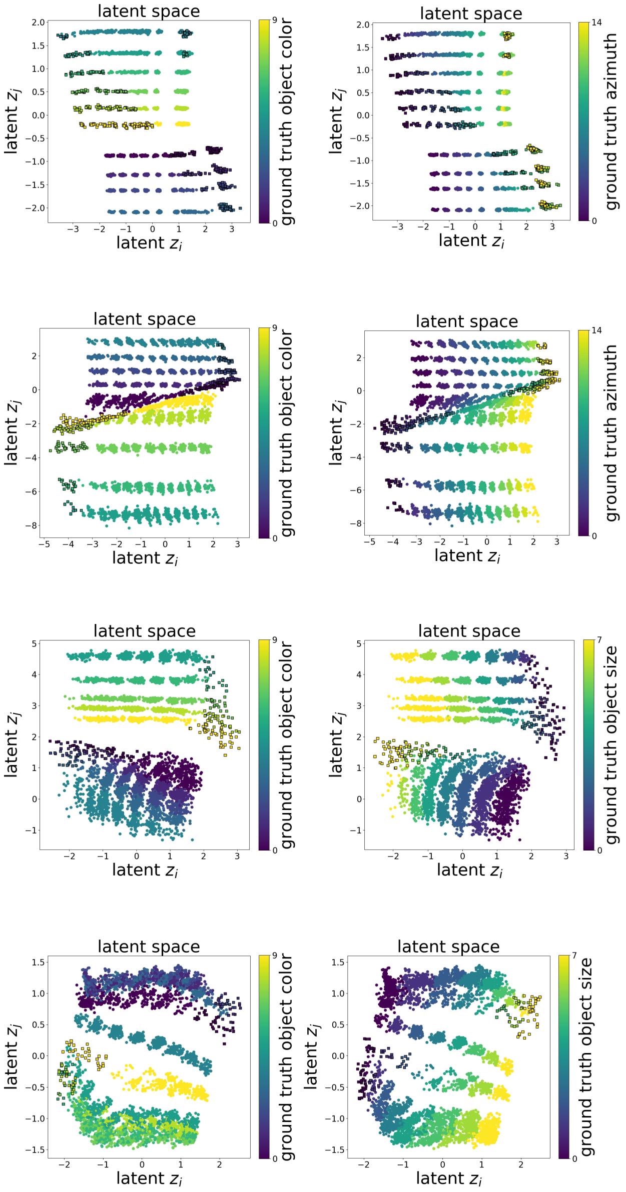

Figure 13. Latent space distribution of the two entangled dimensions of the best DCI model in Shapes3d (E) with $\sigma = 0.2$ (top row), in Shapes3d (E) with $\sigma = 0.4$ (second row), in Shapes3d (D) with $\sigma = 0.2$ (third row) and in Shapes3d (D) with $\sigma = 0.4$ (bottom row). Latent codes sampled from correlated observations (circle without edge) and (2) latent codes sampled with an object size-azimuth configuration not encountered during training(squares with black edge). Each column shows the ground truth values of the two correlated factors by color.

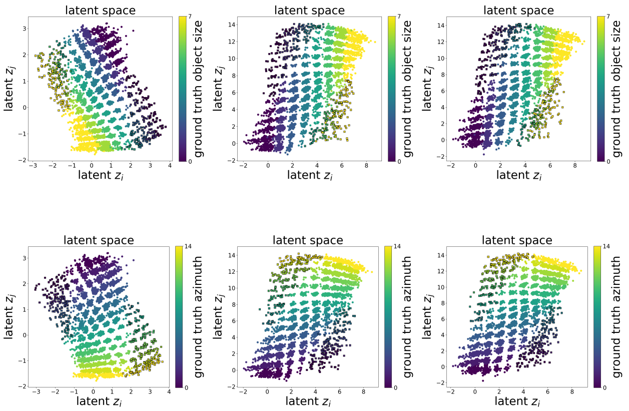

Figure 14. Latent space distribution of the two entangled dimensions of the best DCI model in Shapes3D (A). Latent codes sampled from correlated observations (circle without edge) and (2) latent codes sampled with an object size-azimuth configuration not encountered during training (squares with black edge). Left column shows the latent space of the two correlated factors by color. Middle and right column show the fast adapted space using linear regression and 100 or 1000 labels respectively.

D. Additional Results Section 5

D.1. Post-hoc alignment correction with few labels

In Fig. 14, we see the axis alignment of the correlated latent space after fast adaptation using linear regression on a model trained on Shapes3D (A). Fast adaptation with linear regression substitution fails in some settings: when no two latent dimensions encode the applied correlation isolated from the other latent codes, or when the correlated variables do not have a unique natural ordering (e.g. color or categorical variables). Additionally, the functional form of the latent manifolds beyond the training distribution is unknown and in general expected to be nonlinear. We test the possibility of fast adaptation in this case using as substitution function a one-hidden layer MLP classifier of size 100 on the correlated Shapes3D variants. Under this method, we sample the few labels from a uniform independent distribution. A small number of such samples could practically be labeled manually. Using only 1000 labeled data points for our fast adaptation method shows a significant reduction in entanglement thresholds for the correlated pair (Table 6).

D.2. Alignment during training using weak supervision

Using the studied weakly supervision Ada-GVAE method with $k = 1$ from Locatello et al. (2020b), we showed that weak supervision can provide a strong inductive bias capable of finding the right factorization and resolving spurious correlations for datasets of unknown degree of correlation. Besides the results shown on Shapes3D (A) in the main paper, representative latent space visualizations that show strong axis alignment across all three correlation variants in Shapes3D (A, D, E) are shown in Fig. 15. This study contains a total of 360 trained models.

dataset	labels	0	1000
A ($\sigma = 0.2$)	object size - azimuth	0.37	0.26
	median uncorrelated pairs	0.09	0.1
D ($\sigma = 0.2$)	object color - object size	0.3	0.16
	median uncorrelated pairs	0.07	0.07
E ($\sigma = 0.2$)	object color - azimuth	0.25	0.2
	median uncorrelated pairs	0.1	0.11

Table 6. Mean of the pairwise entanglement scores for the correlated pair (red) and the median of the uncorrelated pairs (based on GBT feature importance) for all pairs of variables in Shapes3D (D) (top), Shapes3D (E) (middle) and Shapes3D (A) (bottom) all with correlation strength $\sigma = 0.2$. Each pairwise score is the mean across 180 models for each dataset and correlation strength. First column is the unsupervised baseline without any fast adaptation and the second column shows that fast adaption using a one-hidden layer MLP reduces these correlations with as little as 1000 labels.

In addition to the experiment from the main paper where pairs are constructed solely from the correlated observational data, we want to study two scenarios where we have some intervention capabilities on the FoV to generate training pairs. The resulting distribution of FoVs (still exhibiting correlations) in these pairs depends on whether the correlation between two pairs is due to a causal link or due to a common confounder.

Scenario I-1: We assume there is a confounder (which is not among the observable factors in the data) causing a spurious correlation between the pair of correlated factors. Then, the correlation is broken whenever our interventional sampling procedure yields a pair where the changing FoV is one of the correlated ones. In that case, the value of the changing variable is sampled uniformly in the second observation of the pair. Note that this still means that the vast majority of sampled pairs exhibit correlated FoV as in most cases the changing factor will be one of the other independent uncorrelated FoV.

As under the default scenario from the main paper, we consistently observe high disentanglement models, often achieving perfect DCI score irrespective of correlations in the data set. This is depicted together with some selected latent space visualizations that show strong axis alignment in Fig. 16. The latent spaces of the correlated FoV in the train data tend to strongly align their coordinates with the ground truth label axis. We chose 10 random seeds per configuration in this study, yielding 720 models in total.

Scenario I-2: Let us assume c_1 causes c_2 in our examples, which manifests as the studied linear correlation. If we intervene on (or "fix") all factors except for the effect c_2, we cannot sample uniformly in c_2 as it is causally affected by c_1. Intervening on all factors but c_1, however, allows us to sample any value in c_1 as it is not causally affected by

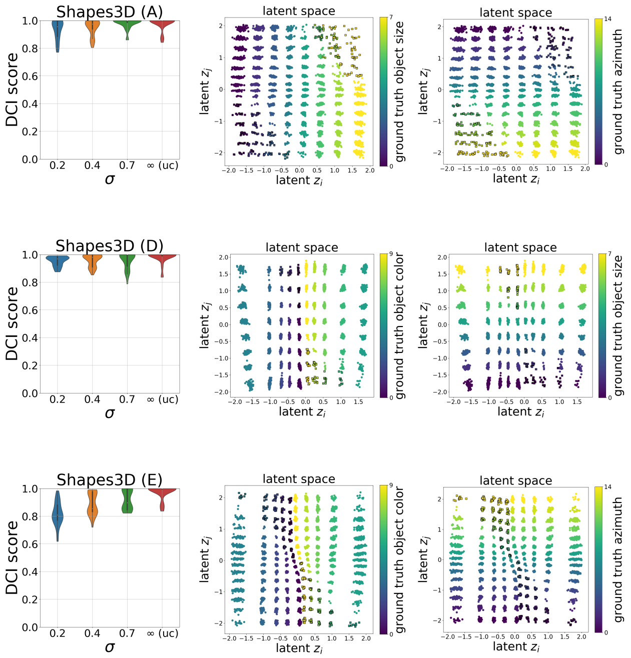

Figure 15. Left: For the weakly supervised scenario using correlated observational data trained models on Shapes3D (A), (D) and (E) correlating object color and azimuth learn consistently improved, often perfect, disentangled representation across all correlation strengths. Right: Latent dimensions of a best DCI model trained on strongly correlated observational data. Representations are strongly axis-aligned with respect to both of the correlated variables ground truth values (right).

c_2. To test the hypothesis that this constraint also allows for disentangling the correlation, we trained on Shapes3D (A) and sampled pairs consistent with this causal model. Besides observing visually disentangled factors in the latent traversals, we show a summary of our results in Fig. 17 with the same significant improvements regarding disentangling the correlated FoVs. Besides the correlation strengths used throughout the paper, we additionally trained the same models using a very strong correlation of $\sigma = 0.1$. The study of scenario I-2 thus comprises 300 models.

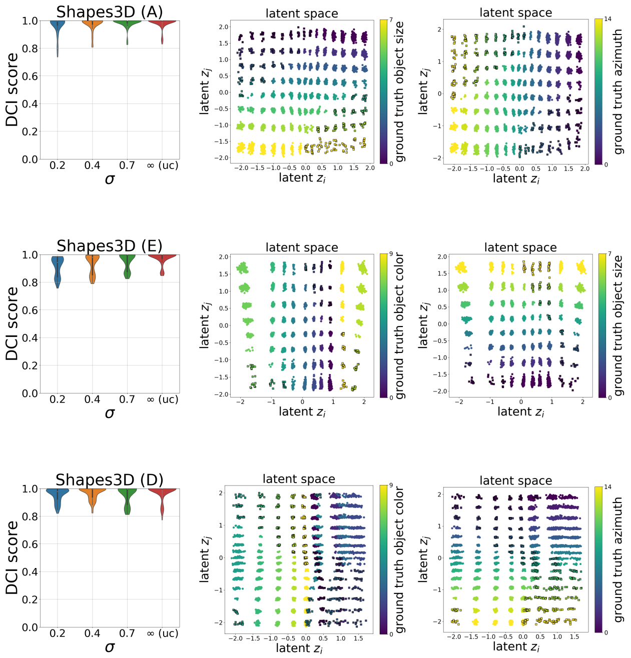

Figure 16. Left: For the weakly supervised scenario with intervening capabilities (Scenario I-1) trained models on Shapes3D (A), (D) and (E) correlating object color and azimuth learn consistently improved, often perfect, disentangled representation across all correlation strengths. Right: Latent dimensions of a best DCI model with strong correlation ($\sigma = 0.2$). Representations are strongly axis-aligned with respect to both of the correlated variables ground truth values.

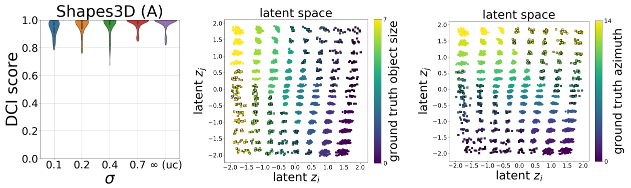

Figure 17. DCI scores and latent spaces show strong disentanglement using weak supervision with intervening capabilities (Scenario I-2) - even under the stronger assumption that sampling of observation pairs follow its causal generative model. We show the learned latent space encoding of the two correlated factors of variation for a model on Shapes3D with $\sigma = 0.1$.