A THEORY

We first present the properties of the exponential family distribution. Then we provide the proof for Proposition 1.

A.1 EXPONENTIAL FAMILY DISTRIBUTIONS

Definition 2 (Exponential family) *A probability distribution belongs to the exponential family if its density has the following form*:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{Q(\mathbf{x})e^{<\mathbf{T}(\mathbf{x}),\boldsymbol{\lambda}(\boldsymbol{\theta})>}}{Z(\boldsymbol{\theta})} \tag{15}$$

where $\mathbf{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \Theta$, $\mathbf{T} : \mathcal{X} \to \mathbb{R}^k$ is the sufficient statistic, $\boldsymbol{\lambda} : \Theta \to \mathbb{R}^k$ is the corresponding parameter, $Q : \mathcal{X} \to \mathbb{R}$ is the base measure (Lebesgue measure), $Z(\boldsymbol{\theta})$ is the normalizing constant and $< \cdot, \cdot >$ stands for the dot product. More precisely, we can refer to equation (15) as conditional exponential family (Pacchiardi & Dutta, 2022).

Theorem 1 (Universal approximation capability) *Let $p(\mathbf{x}|\boldsymbol{\theta})$ be a conditional probability density function. Assume that $\mathcal{X}$ and Θ are compact Hausdorff spaces, and that $p(\mathbf{x}|\boldsymbol{\theta}) > 0$ almost surely $\forall(\mathbf{x}, \boldsymbol{\theta}) \in \mathcal{X} \times \Theta$. Then for each $\epsilon > 0$, there exists $(\boldsymbol{\phi}, k) \in \Phi \times \mathbb{N}, \boldsymbol{\phi} = (\mathbf{T}, \boldsymbol{\lambda})$, where k is the dimension of the feature extractor, such that $\sup_{(\mathbf{x},\boldsymbol{\theta})\in\mathcal{X}\times\Theta} |p_{\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta}) - p(\mathbf{x}|\boldsymbol{\theta})| < \epsilon$.*

Authors in Khemakhem et al. (2020b) provide the above Theorem that shows the universal approximation capability of conditional exponential family. Specifically, they show that for compact Hausdorff spaces $\mathcal{X}$ and Θ, any conditional probability density $p(\mathbf{x}|\boldsymbol{\theta})$ can be approximated arbitrarily well. Thus, through the consideration of a freely varying k, and general $\mathbf{T}$ and $\boldsymbol{\lambda}$, the conditional exponential family is endowed with universal approximation capability over the set of conditional probability densities. In practice, it is possible to achieve almost perfect approximation by choosing k greater than the input dimension. However, this result does not take into account the practicality of fitting the approximation family to data, and increasing k may make it more difficult to fit the data distribution in real-world scenarios.

A.2 PROOF OF PROPOSITION 1

Proposition 1 *Under the assumptions of infinite capacity for E and $\mathbf{f}$, the solution $(\boldsymbol{\phi}^*, \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*) \in \text{argmin}_{\boldsymbol{\phi},\boldsymbol{\gamma},\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\theta})$ of the loss function (14) guarantees that $\bar{E}_{\boldsymbol{\phi}^*,\boldsymbol{\gamma}^*}(\mathbf{x})$ is disentangled with respect to ξ, as defined in Definition 1.*

Proof Following the approach in Shen et al. (2022), we give a proof of this proposition. We suppose that $d = m$. For $\forall\, i = 1, 2, ..., m$, we consider two cases separately. In the first case, $\mathcal{L}_{sup,i}(\boldsymbol{\phi}, \boldsymbol{\gamma})$ is cross-entropy loss:

$$\begin{aligned} \mathcal{L}_{sup,i}(\boldsymbol{\phi}, \boldsymbol{\gamma}) &= \mathbb{E}_{(\mathbf{x},\mathbf{y},\mathbf{u})}\left[-\boldsymbol{y}^i \log\sigma(\bar{E}(\mathbf{x}, \mathbf{u})^i) - (1 - \boldsymbol{y}^i)\log(1 - \sigma(\bar{E}(\mathbf{x}, \mathbf{u})^i))\right] \\ &= -\int q(\mathbf{x}, \mathbf{u})p(\boldsymbol{y}^i|\mathbf{x}, \mathbf{u})[\boldsymbol{y}^i \log\sigma(\bar{E}(\mathbf{x}, \mathbf{u})^i) \\ &\quad +(1 - \boldsymbol{y}^i)\log(1 - \sigma(\bar{E}(\mathbf{x}, \mathbf{u})^i))]d\mathbf{x}d\mathbf{u}d\boldsymbol{y}^i \end{aligned} \tag{16}$$

where $P(\boldsymbol{y}^i = 1|\mathbf{x}, \mathbf{u}) = \mathbb{E}(\boldsymbol{y}^i|\mathbf{x}, \mathbf{u})$, $P(\boldsymbol{y}^i = 0|\mathbf{x}, \mathbf{u}) = 1 - \mathbb{E}(\boldsymbol{y}^i|\mathbf{x}, \mathbf{u})$. Let $\frac{\partial\mathcal{L}_{sup,i}(\boldsymbol{\phi},\boldsymbol{\gamma})}{\partial\sigma(\bar{E}(\mathbf{x},\mathbf{u})^i)} = 0$, we know that $\bar{E}^*(\mathbf{x}, \mathbf{u})^i = \sigma^{-1}(\mathbb{E}(\boldsymbol{y}^i|\mathbf{x}, \mathbf{u})) = \sigma^{-1}(\boldsymbol{\xi}^i)$ can minimize $\mathcal{L}_{sup,i}(\boldsymbol{\phi}, \boldsymbol{\gamma})$.

In the second case, $\mathcal{L}_{sup,i}(\boldsymbol{\phi}, \boldsymbol{\gamma})$ is Mean Squared Error (MSE):

$$\mathcal{L}_{sup,i}(\boldsymbol{\phi}, \boldsymbol{\gamma}) = \mathbb{E}_{(\mathbf{x},\mathbf{y},\mathbf{u})}\left[(\boldsymbol{y}^i - \bar{E}(\mathbf{x}, \mathbf{u})^i)^2\right] = \int q(\mathbf{x}, \mathbf{u})p(\boldsymbol{y}^i|\mathbf{x}, \mathbf{u})(\boldsymbol{y}^i - \bar{E}(\mathbf{x}, \mathbf{u})^i)^2 d\mathbf{x}d\mathbf{u}d\boldsymbol{y}^i \tag{17}$$

Let $\frac{\partial\mathcal{L}_{sup,i}(\boldsymbol{\phi},\boldsymbol{\gamma})}{\partial\sigma(\bar{E}(\mathbf{x},\mathbf{u})^i)} = 0$, we know that $\bar{E}^*(\mathbf{x}, \mathbf{u})^i = \mathbb{E}(\boldsymbol{y}^i|\mathbf{x}, \mathbf{u}) = \boldsymbol{\xi}^i$ can minimize the $\mathcal{L}_{sup,i}(\boldsymbol{\phi}, \boldsymbol{\gamma})$.

Next, because of the infinite capacity of $\mathbf{f}$ and Theorem 1, we know that $q_{\boldsymbol{\phi}^*,\boldsymbol{\gamma}^*}(\mathbf{x}, \widetilde{\mathbf{z}}|\mathbf{u})$ is contained within the distribution family of $p_{\boldsymbol{\theta}}(\mathbf{x}, \widetilde{\mathbf{z}}|\mathbf{u})$. Then by minimizing the loss in (14) over $\boldsymbol{\theta}$, we can find $\boldsymbol{\theta}^*$ such that $p_{\boldsymbol{\theta}^*}(\mathbf{x}, \widetilde{\mathbf{z}}|\mathbf{u})$ matches $q_{\boldsymbol{\phi}^*,\boldsymbol{\gamma}^*}(\mathbf{x}, \widetilde{\mathbf{z}}|\mathbf{u})$ and thus

$$-\text{ELBO}(\boldsymbol{\phi^*}, \boldsymbol{\gamma^*}, \boldsymbol{\theta^*}) = D_{\text{KL}}(q_{\boldsymbol{\phi}^*,\boldsymbol{\gamma}^*}(\mathbf{x}, \widetilde{\mathbf{z}}|\mathbf{u})||p_{\boldsymbol{\theta}^*}(\mathbf{x}, \widetilde{\mathbf{z}}|\mathbf{u})) + \text{constant} \tag{18}$$

reaches minimum.

Therefore, by minimizing the loss function (14) of CauF-VAE, we can get the solution, that is, $\bar{E}^{*}(\mathbf{x}, \mathbf{u})^{i} = r_i(\boldsymbol{\xi}^i)$ with $r_i(\boldsymbol{\xi}^i) = \sigma^{-1}(\boldsymbol{\xi}^i)$ if cross-entropy loss is used, and $r_i(\boldsymbol{\xi}^i) = \boldsymbol{\xi}^i$ if MSE is used. Here, $\mathbf{x}$ denotes random variable.

To sum up, $\bar{E}_{\boldsymbol{\phi}^*, \boldsymbol{\gamma}^*}(\mathbf{x}, \mathbf{u})$ is disentangled with respect to ξ, as defined in Definition 1.

B EXPERIMENTAL DETAILS

We present the main settings used in our experiments. Our experiments on Pendulum utilize one NVIDIA GeForce RTX 2080ti GPU, while experiments on CelebA use one NVIDIA GeForce RTX 3080 GPU. To train DEAR, we use two NVIDIA GeForce RTX 2080ti GPUs. Due to blind review requirements, code is currently available only in the supplementary material.

B.1 DATA PREPROCESSING

Pendulum We generate the pendulum dataset using the synthetic simulators mentioned in Shen et al. (2022) and Yang et al. (2021). For detailed generation procedures, please refer to Appendix F in Shen et al. (2022) or Appendix C.1.1 in Yang et al. (2021). The pendulum images are resized to 64×64×3 resolution, and the training and testing sets consist of 5847 and 1461 samples, respectively.

CelebA We employ the default training and testing sets in CelebA, with 162770 and 19962 samples, respectively. We also resize the images to 64×64×3 resolution. We set the values of features that are -1 to 0.

B.2 EXPERIMENTAL SETUP

Causal Flows Implementation In CauF-VAE, we incorporate causal flows to enhance the model's ability to capture causal underlying factors. As discussed in Section 2.2, complex composite functions can be obtained by composing multiple simple transformations, which is the main mechanism adopted by flow models. However, in order to preserve the causal relationships between variables in the output representation and to verify the ability of the causal flows, we only use a single layer of causal flow in our experiments. As Proposition 1 shows, the $mean$ of the output latent variables from the causal flow can achieve disentanglement. In practice, finding the exact value of the $mean$ can be challenging, so we use random sampling to approximate it by drawing N samples. In our experiments, we set N to be 1. In causal flows, besides the affine transformation introduced in Section 3, other implementation methods such as integration-based transformers (Wehenkel & Louppe, 2019) can also be used, but the computational cost for sampling or density estimation needs to be taken into consideration when modeling.

Conditional prior Implementation Regarding the setting of conditional prior, since it is generally difficult to directly fit the exponential family of distributions, we use a special form of exponential distribution, namely the Gaussian distribution in our experiments. For simplicity, we adopt a factorial distribution, as described in Yang et al. (2021) and Khemakhem et al. (2020a). However, unlike them, we set the $mean$ and $variance$ as learnable parameters for training, which enhances the flexibility of the prior distribution.

Supervised loss Implementation For Pendulum, we resize the factors' labels to $[-1, 1]$ since they are continuous, and we use MSE as the loss function $\mathcal{L}_{sup}$. For CelebA, as the factors' labels are binary, we convert -1 to 0 and use cross-entropy loss.

Architecture and Hyperparameters The network structures of encoder and decoder are presented in Table 3. The encoder's output, i.e., $mean$ and $\log variance$, share parameters except for the final layer. A single-layer causal flow implemented by an MLP implementation is used for fitting the posterior distribution, and a causal weight matrix A is added in a manner inspired by Wehenkel & Louppe (2021). The decoder's output is resized to generate pixel values for a three-channel color image. β_{sup} is roughly tuned during our hyperparameter selection process. We train the model using the Adam optimizer, and all training parameters are shown in Table 4. Although we cannot guarantee finding the optimal solution in the experiments, the results still demonstrate the excellent performance of our model.

Table 3: Architecture for the Encoder and Decoder in CauF-VAE ($d = 4$ for Pendulum and $d = 100$ for CelebA).

Encoder	**Decoder**
-	Input $\widetilde{\mathbf{z}} \in \mathbb{R}^d$
3×3 conv, MaxPool, 8 SELU, stride 1	FC, 256 SELU
3×3 conv, MaxPool, 16 SELU, stride 1	FC, 16×12×12 SELU
3×3 conv, MaxPool, 32 SELU, stride 1	2×2 conv, 32 SELU, stride 1
3×3 conv, MaxPool, 64 SELU, stride 1	2×2 conv, 64 SELU, stride 1
3×3 conv, MaxPool, 8 SELU, stride 1	2×2 conv, 128 SELU, stride 1
FC 256×2	FC, 3×64×64 Tanh

Table 4: Hyperparameters of CauF-VAE.

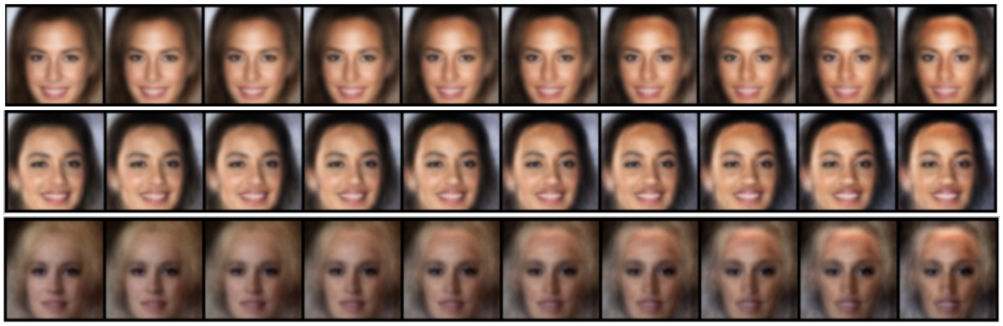
Parameters	**Values (Pendulum)**	**Values (CelebA)**
Batch size	128	128
Epoch	801	101
Latent dimension	4	100
σ	0.1667	0.1667
β_{sup}	8	5
β_1	0.2	0.2
β_2	0.999	0.999
ϵ	1e−8	1e−8
Learning rate of Encoder	5e−5	3e−4
Learning rate of Causal Flow	5e−5	3e−4
Learning rate of A	1e−3	1e−3
Learning rate of Conditional prior	5e−5	3e−4
Learning rate of Decoder	5e−5	3e−4

Experimental setup for baseline models We compare our method with several representative baseline models for disentanglement in VAE (Locatello et al., 2019a), including vanilla VAE (Kingma & Welling, 2013), β-VAE (Higgins et al., 2017), β-TCVAE (Chen et al., 2018), and DEAR (Shen et al., 2022). We use the same conditional prior and loss term with labeled data for each of these methods as in CauF-VAE, except that DEAR's prior is SCM prior. Furthermore, apart from models that specifically choose the architecture of encoder and decoder, we employ identical encoder and decoder structures for the baselines. The implementations of β-TCVAE and DEAR are separately referred to publicly available source codes `https://github.com/AntixK/PyTorch-VAE` and `http://jmlr.org/papers/v23/21-0080.html`. The training optimizer of baseline models is Adam except DEAR, and the parameters are the default parameters. For DEAR, except for the β_1 and β_2 parameters in Adam, which are consistent with our model, all other parameters are the same as in the original paper. We train DEAR for 400 epochs on Pendulum, 100 epochs on CelebA (smile), and 70 epochs on CelebA (Attractive). All other baseline models are trained for 100 epochs.

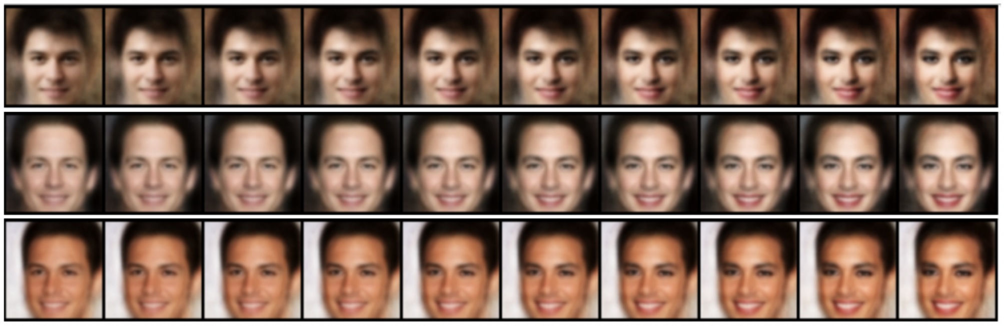
We did not compare our results with CausalVAE (Yang et al., 2021) due to several reasons. Firstly, the latent variable dimension in CausalVAE is equivalent to the number of underlying factors of interest. However, when applied to real-world datasets like CelebA, it fails to consider all generative factors of the images. Although there is consistency correlation between latent variables and factors in the prior, it does not achieve true disentanglement as defined in Definition 1. Secondly, since the decoder in CausalVAE contains a Mask layer, it is impossible to observe the changes of a single factor in the reconstructed images when traversing each dimension of the learned representation. Finally, in the experimental aspect, the authors used a multi-dimensional latent variable vector to approximate each concept to ensure good performance of the model. Therefore, the dimensionality setting of our model's latent variables cannot be unified with that of CausalVAE.

C ADDITIONAL RESULTS

Samples from interventional distributions In section 6.2, we describe the capability of our model to perform interventions by generating new images that do not exist in the dataset. Specifically, our model utilizes causal flows to sample from the interventional distributions, even though the model is trained on observational data. The steps for intervening on one factor are explained in section 6.2, and the same applies to intervening on multiple factors. As depicted in Figure 8(a), we intervene on the values of two factors by fixing gender as female and gradually adjusting the value of receding hairline. This produces a series of images showing women with a gradually receding hairline. Furthermore, as shown in Figure 8(b), we intervene on gender and makeup, generating a series of images of men with gradually applied makeup. These images are not commonly found or even don't exist in the training data which highlights the ability of our model to sample from the interventional distributions.



(a) Female gradually with receding hairline

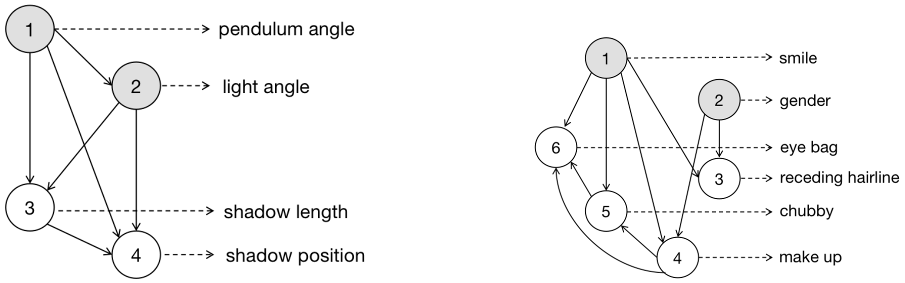


(b) Male gradually with make up

Figure 8: Sample from interventional distributions.

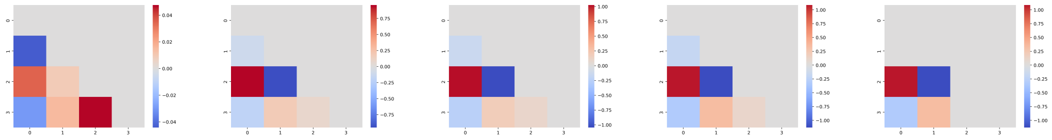
Learning of causal structure CauF-VAE has the potential to learn causal structure between underlying factors, even without using a SCM. This is helped by the supervised loss term. Here we present the learning process of the adjacency weight matrix A, whose super-graph is shown in Figure 9. Figure 9(b) shows the learning process of CelebA (Attractive). If we set a threshold of 0.25, i.e., considering edges in the causal graph smaller than the threshold as non-existing, we can obtain Figure 11(e). We observe an almost accurate graph, suggesting that there is potential for further enhancement in causal discovery through future design.

Examining Causally Disentangled Representation To verify whether our model has obtained causally disentangled representations, we consider two types of intervention operations: the "traverse" operation and the "intervention" operation introduced in Section 6.2. We present the experimental results of CauF-VAE on CelebA(Attractive) in Figure 12 and the results of baseline models on three datasets in Figure 13, 14, and 15.



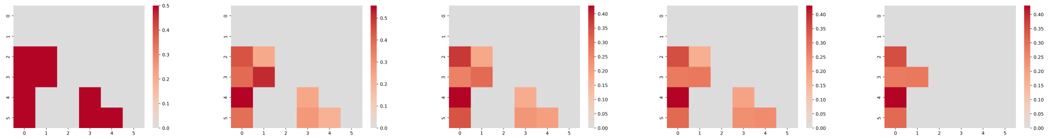
(a) Pendulum (b) CelebA(Attractive)

Figure 9: Super-graph of Pendulum and CelebA(Attractive).



(a) Original A (b) After 20 epochs (c) After 50 epochs (d) After 80 epochs (e) Final A

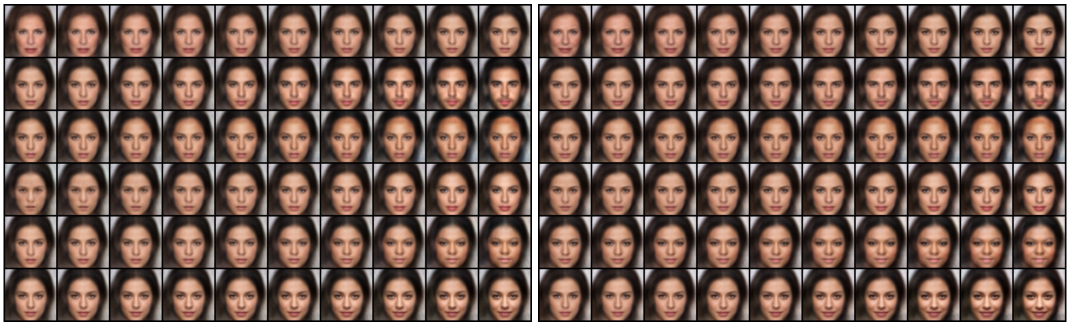
Figure 10: The learned weighted adjacency matrix A given a super-graph on Pendulum. (a)-(d) illustrate the changes in A as the training progresses. (e) represents A after edge pruning.



(a) Original A (b) After 5 epochs (c) After 20 epochs (d) After 120 epochs (e) Final A

Figure 11: The learned weighted adjacency matrix A given a super-graph on CelebA(Attractive). (a)-(d) illustrate the changes in A as the training progresses. (e) represents A after edge pruning.

We also use MIC (Maximal Information Coefficient) and TIC (Total Information Coefficient) (Yang et al., 2021; Kinney & Atwal, 2014) to measure the strength of association between learned representations and ground-truth factors. We use all testing data to calculate MIC and TIC. As shown in Table 5, bolded values indicate optimal results, and underlined values indicate sub-optimal results. We can see that CauF-VAE achieves superior performance on the pendulum, while both CauF-VAE and β-TCVAE show comparable results on CelebA. For CelebA, although β-TCVAE performs slightly better than CauF-VAE, the improvements over CauF-VAE are marginal (2.6% and 3.01% for "Attractive" and 1.31% and 0.61% for "Smile" in terms of MIC and TIC, respectively). In contrast, CauF-VAE outperforms the current state-of-the-art causal disentanglement model DEAR by 22.03% and 22.68% for "Attractive" and 11.29% and 10.67% for "Smile" in terms of MIC and TIC. Additionally, compared to β-VAE and VAE, our model's improvements are more substantial. Furthermore, our downstream experiments in Table 1 and 2, and qualitative results in Appendix C all corroborate our model's superiority. Hence, this reveals that β-TCVAE's higher MIC and TIC scores on CelebA lies in its one latent variable containing information from multiple factors, resulting in high correlations with multiple factors simultaneously, thus cannot achieve causal disentangled representation learning. This confirms that our learned representations exhibit stronger correlation with the ground-truth factors than most baseline models, further validating the effectiveness of our model.



(a) Traverse of CauF-VAE on CelebA(Attractive) (b) Intervention of CauF-VAE on CelebA(Attractive)

Figure 12: Results of the CauF-VAE model under two types of interventions on CelebA(Atractive). Each row corresponds to one factor, in the same order as in Figure 2(b). We observe that our model achieves disentanglement, and when intervening on the causal variables, it affects the effect variables, while the opposite is not true.

Table 5: MIC and TIC between ground-truth factors and latent variable representations obtained by different models on three datasets.

	Pendulum		**CelebA(Attractive)**		**CelebA(Smile)**	
Model	MIC	TIC	MIC	TIC	MIC	TIC
CauF-VAE	**93.27**	**89.90**	<u>42.38</u>	<u>42.09</u>	<u>58.53</u>	<u>58.60</u>
DEAR	32.90	30.55	34.73	34.31	52.59	52.95
β-VAE	41.57	35.22	15.61	15.81	33.80	33.65
β-TCVAE	<u>86.18</u>	<u>82.18</u>	**43.47**	**43.36**	**59.30**	**58.96**
VAE	60.78	55.09	27.40	27.33	35.79	35.70

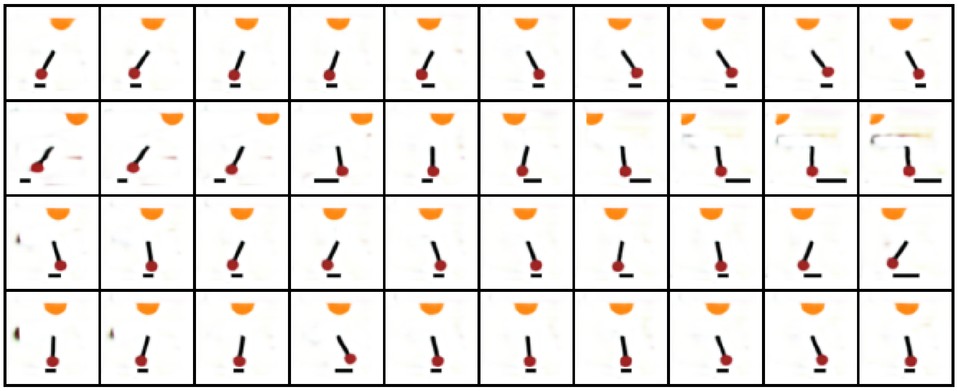
D REFLECTIONS ON DISENTANGLEMENT METRICS

Numerous disentanglement studies introduce their own metrics for assessing disentanglement, including the β-VAE metric (Higgins et al., 2017), the FactorVAE metric (Kim & Mnih, 2018), the Mutual Information Gap (MIG) (Chen et al., 2018), DCI (Eastwood & Williams, 2018) and more. For a comprehensive overview and discussion of these metrics, we direct readers to Locatello et al. (2019a).

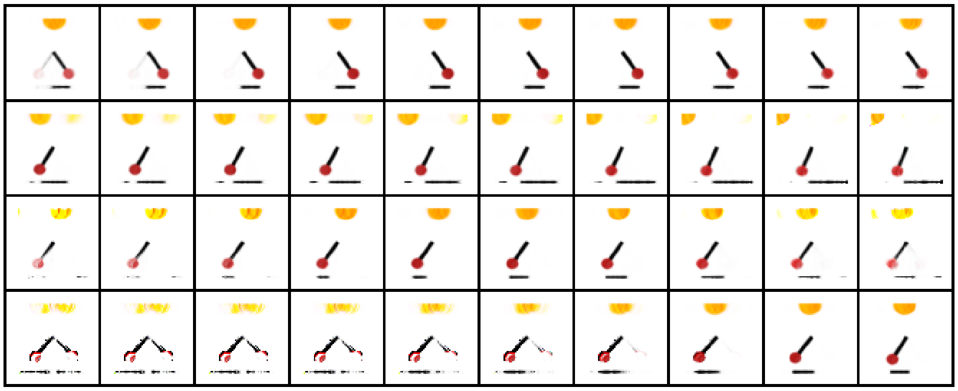
Nevertheless, all these metrics are applicable solely in scenarios where the underlying generative factors are mutually independent; they do not extend to cases where factors exhibit correlation. As an illustration (Shen et al., 2022), the MIG score gauges the normalized gap in mutual information for each factor between the highest and second highest coordinates in $\bar{E}(\mathbf{x}, \mathbf{u})$. Consider a situation where factor ξ_1 is correlated with ξ_2, and a disentangled representation $\bar{E}(\mathbf{x}, \mathbf{u})$ is such that it accommodates one-to-one functions, r_1 and r_2, leading to $\bar{E}_1(\mathbf{x}, \mathbf{u}) = r_1(\xi_1)$ and $\bar{E}_2(\mathbf{x}, \mathbf{u}) = r_1(\xi_2)$. Consequently, both the mutual information of ξ_1 with the highest coordinate $\bar{E}_1(\mathbf{x}, \mathbf{u})$ and the second highest coordinate $\bar{E}_2(\mathbf{x}, \mathbf{u})$ would be substantial, resulting in a minimal difference between them. As a consequence, a disentangled representation in this context would not yield the anticipated high MIG score.

So far, the literature on metrics for causal disentanglement is limited and each has its shortcomings. Shen et al. (2022) proposed a metric based on the FactorVAE metric. Compared to metrics like IRS (Suter et al., 2019) and (UC and GC) (Reddy et al., 2022) that apply only under the assumption of conditional independence of generative factors, this metric is the only suitable one for our model. However, Kim et al. (2019) showed this metric does not work well. The FactorVAE metric may only be partially indicative of the underlying disentanglement: all models attain a perfect scores, which is also confirmed in our experiments. The results revealed in our supplementary experiments show that the FactorVAE scores (Shen et al., 2022) obtained on CauF-VAE, VAE, β-TCVAE, β-VAE, and DEAR were 0.50, 0.50, 0.50, 0.50, and 0.28, respectively. Such outcomes indicate that, at the very least, our model is superior to the latest models DEAR in the causal disentangled representation learning. However, it's evident that the shortcomings of this metric.

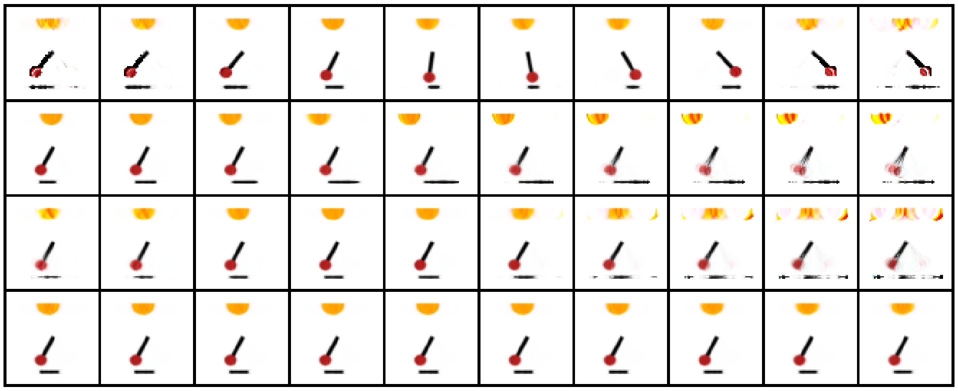
Therefore, for quantitative evaluation, we chose to conduct experiments solely on downstream tasks and measured the performance using the MIC and TIC metrics, thereby demonstrating the superiority of our model.



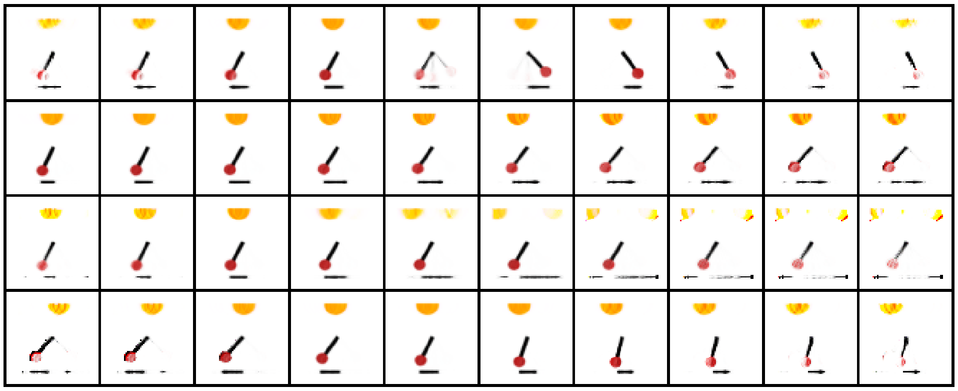
(a) DEAR



(b) β-VAE

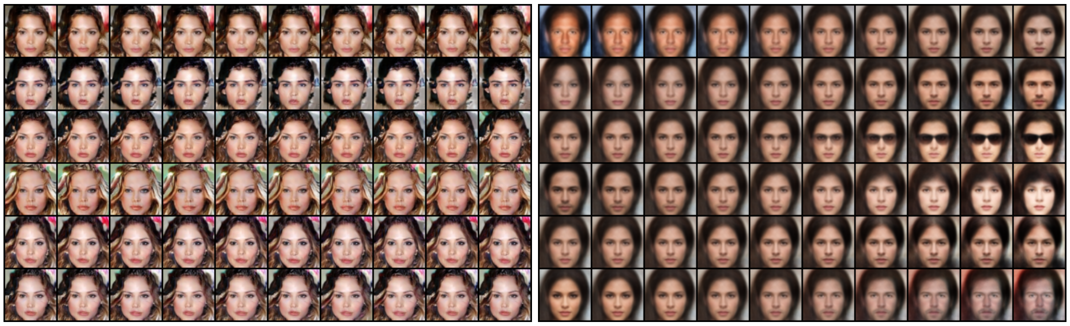


(c) β-TCVAE

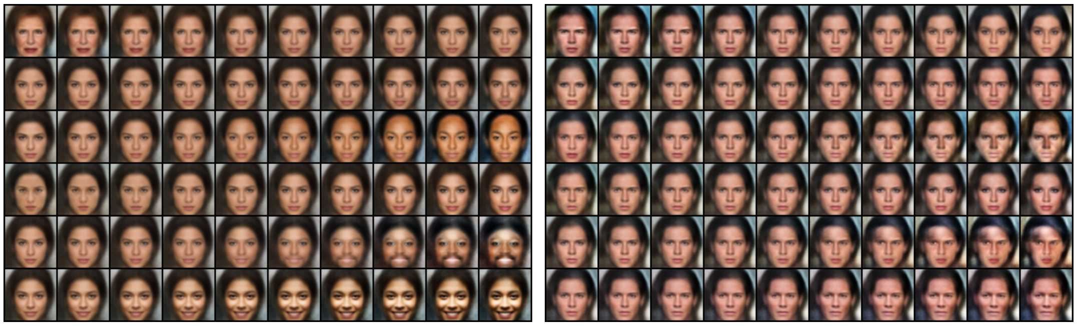


(d) VAE

Figure 13: Traverse results of four baseline models on Pendulum. We observe that changing one factor may result in changes in multiple factors, or no changes in any factor, such as the shadow length in the β-TCVAE. Therefore, their representations are all entangled on Pendulum.

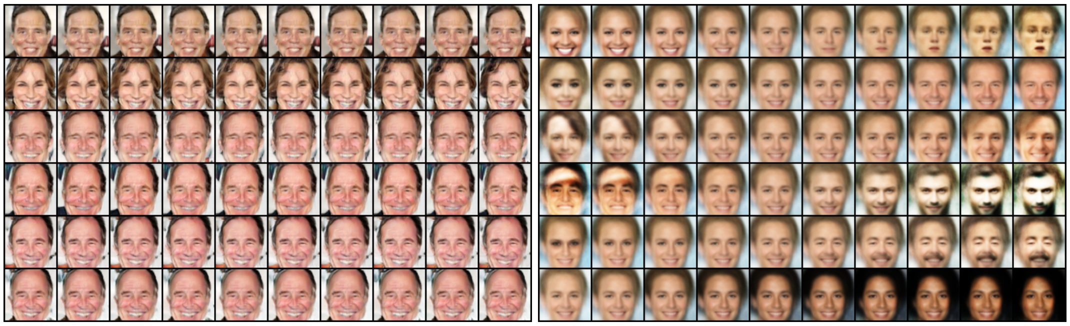


(a) DEAR (b) β-VAE

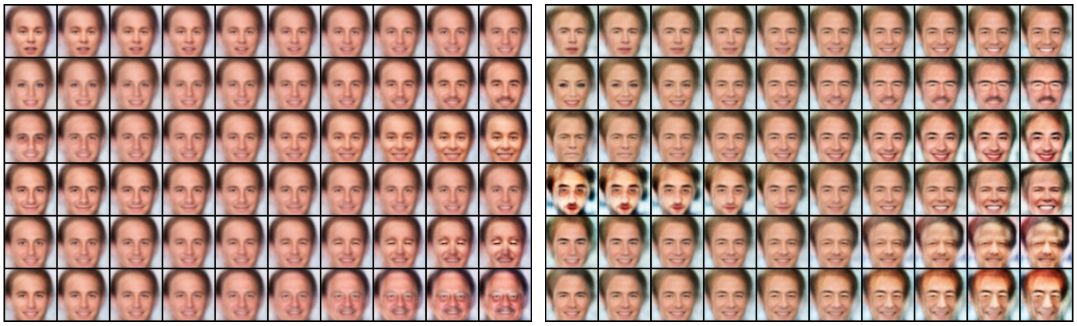


(c) β-TCVAE (d) VAE

Figure 14: Traverse results of four baseline models on CelebA(Attractive). We observe that the representations learned by the four models are still entangled, and some latent variables may not even capture the corresponding factor, as there is no change in the corresponding factor when traversing its value, such as the smile in the β-TCVAE.



(a) DEAR (b) β-VAE



(c) β-TCVAE (d) VAE

Figure 15: Traverse results of four baseline models on CelebA(Smile). The representations learned by the four models are entangled. When the causal variable is changed, not only itself changes, but also the effect variable changes, such as smile and mouth open in these four figures.