A Appendix

A.1 Objective Functions and their Derivations

Here, $\mathbb{E}_{\mathbf{x}}$ and $\mathbb{E}_{\mathbf{x},\mathbf{u}}$ indicate expectations for the marginal and joint distribution of the dataset, respectively.

Objective function of vanilla VAE.

$$\begin{aligned}\min_{\theta,\phi,f} \quad & \mathbb{E}_{\mathbf{x}}\mathbb{E}_{q(\boldsymbol{\epsilon}|\mathbf{x};\phi)}\left[\frac{1}{2}\parallel \mathbf{x} - D(F(\boldsymbol{\epsilon};f,B);\theta)\parallel^2\right] + \beta\cdot\mathbb{E}_{\mathbf{x}}[\mathcal{KL}(q(\boldsymbol{\epsilon}|\mathbf{x};\phi)\|p(\boldsymbol{\epsilon}))] \\ + \quad & \lambda\cdot\mathbb{E}_{\mathbf{x},\mathbf{u}}[\ell(F(\mu(\mathbf{x};\phi);f,B),\mathbf{u})].\end{aligned}$$

Objective function of InfoMax VAE.

$$\begin{aligned}\min_{\theta,\phi,f} \quad & \mathbb{E}_{\mathbf{x}}\mathbb{E}_{q(\boldsymbol{\epsilon}|\mathbf{x};\phi)}\left[\frac{1}{2}\parallel \mathbf{x} - D(F(\boldsymbol{\epsilon};f,B);\theta)\parallel^2\right] + \beta\cdot\mathbb{E}_{\mathbf{x}}[\mathcal{KL}(q(\boldsymbol{\epsilon}|\mathbf{x};\phi)\|p(\boldsymbol{\epsilon}))] \\ + \quad & \lambda\cdot\mathbb{E}_{\mathbf{x},\mathbf{u}}[\ell(F(\mu(\mathbf{x};\phi);f,B),\mathbf{u})] \\ - \quad & \gamma\cdot\Big(\mathbb{E}_{\mathbf{x}}\mathbb{E}_{q(\boldsymbol{\epsilon}|\mathbf{x};\phi)}[t(\mathbf{x},\boldsymbol{\epsilon})] - \mathbb{E}_{\mathbf{x}}\mathbb{E}_{q(\boldsymbol{\epsilon};\phi)}[\exp(t(\mathbf{x},\boldsymbol{\epsilon})-1)]\Big),\end{aligned}$$

where $q(\boldsymbol{\epsilon};\phi)$ is aggregated posterior distribution and $t:\mathbb{R}^m\times\mathbb{R}^d\mapsto\mathbb{R}$ is parameterized with the neural network.

ELBO derivation of CDG-VAE. By Jensen's inequality, the lower bound of log-likelihood is written as

$$\mathbb{E}_{\mathbf{x}}[\log p(\mathbf{x};\theta,\beta)] \geq \mathbb{E}_{\mathbf{x}}\mathbb{E}_{q(\mathbf{z}|\mathbf{x};\phi)}[\log p(\mathbf{x}|\mathbf{z};\theta,\beta)] - \mathbb{E}_{\mathbf{x}}[\mathcal{KL}(q(\mathbf{z}|\mathbf{x};\phi)\|p(\mathbf{z}))].$$

Due to the nonlinear function f, the prior and posterior distribution of $\mathbf{z} = f\left((I-B^\top)^{-1}\boldsymbol{\epsilon}\right)$ is not defined well, and consequently, the KL-divergence term is intractable.

However, since f is an element-wise 1-to-1 function, we can derive ELBO based on $f^{-1}(\mathbf{z})$ instead of $\mathbf{z}$. By (1), the prior and posterior distribution of $f^{-1}(\mathbf{z})$ are

$$\begin{aligned} f^{-1}(\mathbf{z}) \quad &\sim \quad N\Big(0,(I-B^\top)^{-1}(I-B^\top)^{-\top}\Big), \\ f^{-1}(\mathbf{z})|\mathbf{x} \quad &\sim \quad N\Big((I-B^\top)^{-1}\mu(\mathbf{x};\phi),(I-B^\top)^{-1}diag(\sigma^2(\mathbf{x};\phi))(I-B^\top)^{-\top}\Big),\end{aligned}$$

and let $p(f^{-1}(\mathbf{z}))$ and $q(f^{-1}(\mathbf{z})|\mathbf{x};\phi)$ be the prior and posterior distribution of $f^{-1}(\mathbf{z})$, respectively. For the simplicity of notaton, denote $\Omega_1 \coloneqq (I-B^\top)^{-1}(I-B^\top)^{-\top}$ and $\Omega_2 \coloneqq (I-B^\top)^{-1}diag(\sigma^2(\mathbf{x};\phi))(I-B^\top)^{-\top}$.

We assume that the prior and posterior distributions of the latent variable share the same causal adjacency matrix B, which determines the covariance structure. Therefore, as demonstrated by the equations presented below, we can show that the KL-divergence between the prior and posterior distributions of $f^{-1}(\mathbf{z})$ is equivalent to that of the exogenous variable $\boldsymbol{\epsilon}$.

$$\begin{aligned} & \mathcal{KL}(q(f^{-1}(\mathbf{z})|\mathbf{x};\phi)\|p(f^{-1}(\mathbf{z}))) \\ = \quad & \frac{1}{2}\left(\log\frac{|\Omega_1|}{|\Omega_2|} - d + tr(\Omega_1^{-1}\Omega_2) + ((I-B^\top)^{-1}\mu(\mathbf{x};\phi))^\top\Omega_1^{-1}((I-B^\top)^{-1}\mu(\mathbf{x};\phi))\right) \\ = \quad & \frac{1}{2}\left(\log\frac{1}{|diag(\sigma^2(\mathbf{x};\phi))|} - d + tr(diag(\sigma^2(\mathbf{x};\phi))) + \mu(\mathbf{x};\phi)^\top\mu(\mathbf{x};\phi)\right) \\ = \quad & \frac{1}{2}\left(-\sum_{j=1}^{d}\log\sigma^2(\mathbf{x};\phi)_j - d + \sum_{j=1}^{d}\sigma^2(\mathbf{x};\phi)_j + \|\mu(\mathbf{x};\phi)\|^2\right) \\ = \quad & \mathcal{KL}(q(\boldsymbol{\epsilon}|\mathbf{x};\phi)\|p(\boldsymbol{\epsilon})).\end{aligned}$$

Based on the above results, the ELBO is derived as

$$\begin{aligned}\mathbb{E}_{\mathbf{x}}[\log p(\mathbf{x};\theta,\beta)] \quad \geq \quad & \mathbb{E}_{\mathbf{x}}\mathbb{E}_{q(\mathbf{z}|\mathbf{x};\phi)}[\log p(\mathbf{x}_\sigma|\mathbf{z};\theta,\beta)] - \mathbb{E}_{\mathbf{x}}[\mathcal{KL}(q(\mathbf{z}|\mathbf{x};\phi)\|p(\mathbf{z}))] \\ = \quad & \mathbb{E}_{\mathbf{x}}\mathbb{E}_{q(f^{-1}(\mathbf{z})|\mathbf{x};\phi)}[\log p(\mathbf{x}_\sigma|\mathbf{z};\theta,\beta)] - \mathbb{E}_{\mathbf{x}}[\mathcal{KL}(q(f^{-1}(\mathbf{z})|\mathbf{x};\phi)\|p(f^{-1}(\mathbf{z}))] \\ = \quad & \mathbb{E}_{\mathbf{x}}\mathbb{E}_{q(\boldsymbol{\epsilon}|\mathbf{x};\phi)}[\log p(\mathbf{x}_\sigma|F(\boldsymbol{\epsilon};f,B);\theta,\beta)] - \mathbb{E}_{\mathbf{x}}[\mathcal{KL}(q(\boldsymbol{\epsilon}|\mathbf{x};\phi)\|p(\boldsymbol{\epsilon}))] \\ = \quad & -\frac{1}{\beta}\Big(\mathbb{E}_{\mathbf{x}}\mathbb{E}_{q(\boldsymbol{\epsilon}|\mathbf{x};\phi)}\left[\frac{1}{2}\parallel \mathbf{x}_\sigma - D(F(\boldsymbol{\epsilon};f,B);\theta)_\sigma\parallel^2\right] + \beta\cdot\mathbb{E}_{\mathbf{x}}[\mathcal{KL}(q(\boldsymbol{\epsilon}|\mathbf{x};\phi)\|p(\boldsymbol{\epsilon}))] + \beta\cdot\frac{m}{2}\log 2\pi\beta\Big) \\ = \quad & -\frac{1}{\beta}\cdot\mathbb{E}_{\mathbf{x}}[\mathcal{L}(\mathbf{x};\theta,\phi,f)] - \frac{m}{2}\log 2\pi\beta.\end{aligned}$$

A.2 Proof of Proposition 8

Proof. Let $D(\cdot;\theta):\mathbb{R}^d\mapsto\mathbb{R}^p$ be a decoder mean vector of a generative model where the input is denoted as $\mathbf{z}\in\mathbb{R}^d$. Suppose that the generative model has a disentangled representation (Definition 4) and satisfies CDG (Definition 5). Then, for all $s=1,\cdots,K$, the blockwise causal relationships from $\mathbf{z}_\pi$ to $D(\mathbf{z};\theta)_\sigma$ can be represented as the DAG structure of Figure 6. Consider an arbitrary $s\in\{1,\cdots,K\}$.

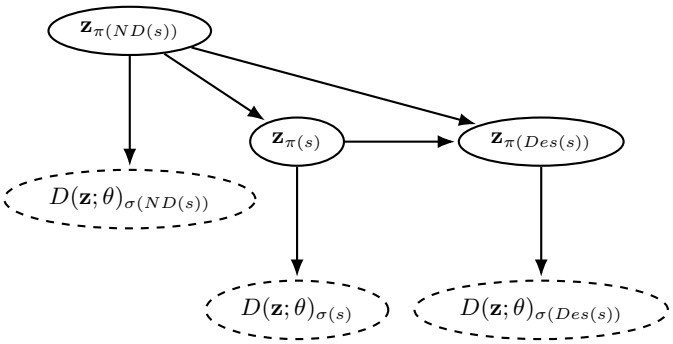


Figure 6: For $s = 1, \cdots, K$, DAG structure of the latent variable $\mathbf{z}_\pi$ and $D(\mathbf{z};\theta)_\sigma$, based on GM3 of Figure 1. Deterministic nodes are denoted as a dashed line.

1. For all $i \in \pi(s)$ and $j \in \sigma(ND(s))$, there is no directed path from $\mathbf{z}_i$ to $D(\mathbf{z};\theta)_j$ (see Figure 6). Therefore, there is no total causal effect from $\mathbf{z}_i$ to $D(\mathbf{z};\theta)_j$ (graphical criteria for total causal effects [21]).

2. Assume that there is a directed path from $\mathbf{z}_i$ to $\mathbf{z}_j$ for some $i \in \pi(s)$ and $j \in \pi(l)$ where $l \in \{s\} \cup Des(s)$. Then, by the blockwise causal relationship of Figure 6, there is a directed path from $\mathbf{z}_i$ to $D(\mathbf{z};\theta)_k$, for all $k \in \sigma(l)$. Therefore, there is a total causal effect from $\mathbf{z}_i$ to $D(\mathbf{z};\theta)_k$ by Assumption 7 (graphical criteria for total causal effects [21]).

Since the above arguments hold true for all $s = 1, \cdots, K$, therefore the proof is complete. □

A.3 Proof of Proposition 10

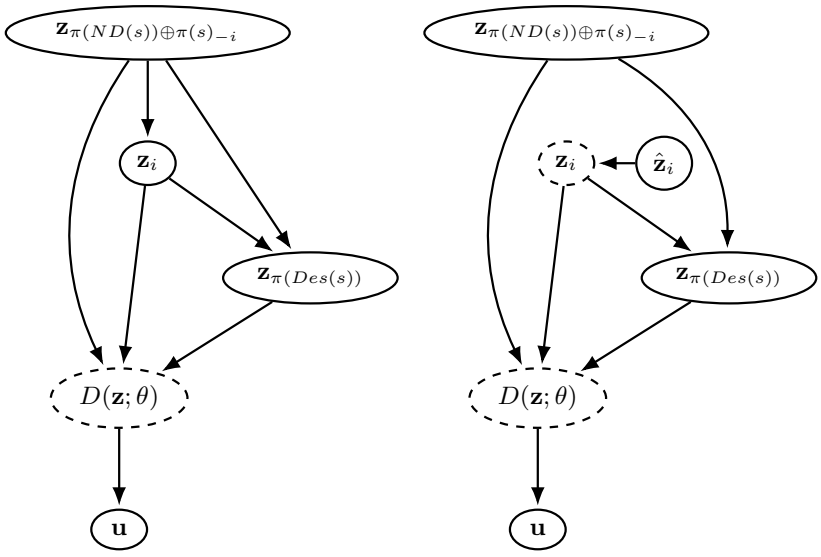


Figure 7: DAG structures based on GM3 of Figure 1. Deterministic nodes are denoted as a dashed line. (a) DAG structure of the latent variable $\mathbf{z}_\pi$, $D(\mathbf{z};\theta)$, and the annotation vector $\mathbf{u}$. (b) The augmented graph of (a) where $\hat{\mathbf{z}}_i$ is a decision variable of $\mathbf{z}_i$. Note that $\mathbf{z}_i$ is now a deterministic node.

A.3.1 Deterministic CPD

$D(\mathbf{z};\theta)$ is a deterministic node which is the deterministic function of its parent $\mathbf{z}$. Denote $\hat{\mathbf{x}} \coloneqq D(\mathbf{z};\theta)$. It implies that the conditional probability distribution (CPD) of $\hat{\mathbf{x}}$ given z is

$$p(\hat{\mathbf{x}}|z) = \delta(\hat{\mathbf{x}} - D(z;\theta)) \tag{4}$$

where δ is the Dirac delta function such that

$$\delta(v) = \begin{cases} +\infty, & \text{if } v = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Using above CPD, $\mathbb{E}[\mathbf{u}|z]$ can be re-written as

$$\begin{aligned} \mathbb{E}[\mathbf{u}|z] &= \int \mathbb{E}[\mathbf{u}|\hat{\mathbf{x}}, z] \cdot p(\hat{\mathbf{x}}|z) d\hat{\mathbf{x}} \\ &= \int \mathbb{E}[\mathbf{u}|\hat{\mathbf{x}}, z] \cdot \delta(\hat{\mathbf{x}} - D(z;\theta)) d\hat{\mathbf{x}} \\ &= \int \mathbb{E}[\mathbf{u}|\hat{\mathbf{x}}] \cdot \delta(\hat{\mathbf{x}} - D(z;\theta)) d\hat{\mathbf{x}} \\ &= \mathbb{E}[\mathbf{u}|D(z;\theta)], \end{aligned} \tag{6}$$

where the third equality holds true by the causal structure of Figure 7(a). Consequently, $\mathbb{E}[\mathbf{u}_c|\mathbf{z}] = \mathbb{E}[\mathbf{u}_c|D(\mathbf{z};\theta)]$, for $c = 1, \cdots, d$.

A.3.2 *Identification of Causal Effects*

For $i \in \pi(s), s = 1, \cdots, K$, assume that arbitrary x and $z_{\pi(ND(s))\oplus\pi(s)_{-i}}$ are given, where we denote $\pi(s)_{-i}$ as the partition tuple $\pi(s)$ without an index i.

By the query simplification rules [21],

$$\mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i)] = \mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, \mathbf{z}_i = z_i], \tag{7}$$

since $\hat{\mathbf{z}}_i \perp\!\!\!\perp \mathbf{u}|\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}, \mathbf{z}_i$ in the graph of Figure 7(b).

$z_{\pi(Des(s))}(z_i, x)$ denotes the deterministic value of $\mathbf{z}_{\pi(Des(s))}$ defined in (1) under intervention $do(\mathbf{z}_i := z_i)$ given x and $z_{\pi(ND(s))\oplus\pi(s)_{-i}}$. It implies that the CPD of $\mathbf{z}_{\pi(Des(s))}$ given x, $z_{\pi(ND(s))\oplus\pi(s)_{-i}}$, and $do(\mathbf{z}_i := z_i)$ is

$$q(\mathbf{z}_{\pi(Des(s))}|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i), x) = \delta\Big(\mathbf{z}_{\pi(Des(s))} - z_{\pi(Des(s))}(z_i, x)\Big), \tag{8}$$

where δ is the Dirac delta function such that

$$\delta(v) = \begin{cases} +\infty, & \text{if } v = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Using the above results,

$$\begin{aligned} & \mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i)] \\ = & \int \mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i), \mathbf{z}_{\pi(Des(s))}] \cdot p(\mathbf{z}_{\pi(Des(s))}|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i)) d\mathbf{z}_{\pi(Des(s))} \\ = & \int \mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i), \mathbf{z}_{\pi(Des(s))}] \cdot q(\mathbf{z}_{\pi(Des(s))}|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i), \mathbf{x}) \\ & \times p(\mathbf{x}|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i)) d\mathbf{z}_{\pi(Des(s))} d\mathbf{x} \\ = & \int \mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i), \mathbf{z}_{\pi(Des(s))}] \cdot \delta(\mathbf{z}_{\pi(Des(s))} - z_{\pi(Des(s))}(z_i, \mathbf{x})) \cdot p(\mathbf{x}) d\mathbf{z}_{\pi(Des(s))} d\mathbf{x} \\ = & \int \mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, \mathbf{z}_i = z_i, \mathbf{z}_{\pi(Des(s))}] \cdot \delta(\mathbf{z}_{\pi(Des(s))} - z_{\pi(Des(s))}(z_i, \mathbf{x})) \cdot p(\mathbf{x}) d\mathbf{z}_{\pi(Des(s))} d\mathbf{x} \\ = & \int \mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, \mathbf{z}_i = z_i, z_{\pi(Des(s))}(z_i, \mathbf{x})] \cdot p(\mathbf{x}) d\mathbf{x} \\ = & \mathbb{E}_{p(\mathbf{x})}\Big[\mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i, z_{\pi(Des(s))}(z_i, \mathbf{x})]\Big], \end{aligned} \tag{10}$$

where $p(\mathbf{x})$ is the probability density function of $\mathbf{x}$ and the third to last equality holds true by (7).

A.3.3 *Main Proof Part*

Proof. Since we assume the causal relationship between $D(\mathbf{z};\theta)_\sigma$ and $\mathbf{u}_\pi$ as $D(\mathbf{z};\theta)_{\sigma(j)} \to \mathbf{u}_{\pi(j)}$, it is implied that $\mathbb{E}[\mathbf{u}_i|D(\mathbf{z};\theta)_{\sigma(j)}] = \mathbb{E}[\mathbf{u}_i|D(\mathbf{z};\theta)]$, for $i \in \pi(j)$, $j = 1, \cdots, K$. Denote $z^{(1)}, z^{(2)}$ as the vector of maximum and minimum values of latent variables given the observed dataset. For $i \in \pi(s), s = 1, \cdots, K$, assume that arbitrary x, $z_{\pi(ND(s))\oplus\pi(s)_{-i}}$ are given. $z^{(j)}_{(i, z_{\pi(ND(s))}, x)} := (z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(j)}, z_{\pi(Des(s))}(z_i^{(j)}, x))$ denotes $\mathbf{z}$ defined in (1) under intervention $do(\mathbf{z}_i := z_i^{(j)})$ given x and $z_{\pi(ND(s))\oplus\pi(s)_{-i}}$,

for $j = 1, 2$. Let $D(\cdot;\theta) : \mathbb{R}^d \mapsto \mathbb{R}^p$ be the decoder of a generative model which satisfies CDG and has a disentangled representation. And the input of the decoder is denoted as $\mathbf{z} \in \mathbb{R}^d$.

1.

Consider $c \in \pi(ND(s))$. Since $z^{(1)}_{(i,z_{\pi(ND(s))},\mathbf{x})}$ and $z^{(2)}_{(i,z_{\pi(ND(s))},\mathbf{x})}$ have the same values for the subvector corresponding to $\pi(ND(s))$ and $D(z^{(j)};\theta)_{\sigma(ND(s))}$ only depends on $z_{\pi(ND(s))}$ by Definition 5,

$$D(z^{(1)}_{(i,z_{\pi(ND(s))},\mathbf{x})};\theta)_{\sigma(ND(s))} = D(z^{(2)}_{(i,z_{\pi(ND(s))},\mathbf{x})};\theta)_{\sigma(ND(s))}. \tag{11}$$

By the causal relationship of GM3 of Figure 1, (6), and (11),

$$\begin{aligned}
\mathbb{E}[\mathbf{u}_c|z^{(1)}_{(i,z_{\pi(ND(s))},\mathbf{x})}] &= \mathbb{E}[\mathbf{u}_c|D(z^{(1)}_{(i,z_{\pi(ND(s))},\mathbf{x})};\theta)] \\
&= \mathbb{E}[\mathbf{u}_c|D(z^{(1)}_{(i,z_{\pi(ND(s))},\mathbf{x})};\theta)_{\sigma(ND(s))}] \\
&= \mathbb{E}[\mathbf{u}_c|D(z^{(2)}_{(i,z_{\pi(ND(s))},\mathbf{x})};\theta)_{\sigma(ND(s))}] \\
&= \mathbb{E}[\mathbf{u}_c|D(z^{(2)}_{(i,z_{\pi(ND(s))},\mathbf{x})};\theta)] \\
&= \mathbb{E}[\mathbf{u}_c|z^{(2)}_{(i,z_{\pi(ND(s))},\mathbf{x})}],
\end{aligned} \tag{12}$$

where $c \in \pi(ND(s))$.

Since (12) holds for all $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
0 &\leq ACE(\mathbf{u}_c, \mathbf{z}_i, \mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}} = z_{\pi(ND(s))\oplus\pi(s)_{-i}}) \\
&= \left|\mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i^{(1)})] - \mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i^{(2)})]\right| \\
&= \left|\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(1)}, z_{\pi(Des(s))}(z_i^{(1)}, \mathbf{x})]] - \mathbb{E}_{p(\mathbf{x})}[\mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(2)}, z_{\pi(Des(s))}(z_i^{(2)}, \mathbf{x})]]\right| \\
&= \left|\mathbb{E}_{p(\mathbf{x})}\left[\mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(1)}, z_{\pi(Des(s))}(z_i^{(1)}, \mathbf{x})] - \mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(2)}, z_{\pi(Des(s))}(z_i^{(2)}, \mathbf{x})]\right]\right| \\
&\leq \mathbb{E}_{p(\mathbf{x})}\left[\left|\mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(1)}, z_{\pi(Des(s))}(z_i^{(1)}, \mathbf{x})] - \mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(2)}, z_{\pi(Des(s))}(z_i^{(2)}, \mathbf{x})]\right|\right] \\
&= \mathbb{E}_{p(\mathbf{x})}\left[\left|\mathbb{E}[\mathbf{u}_c|z^{(1)}_{(i,z_{\pi(ND(s))},\mathbf{x})}] - \mathbb{E}[\mathbf{u}_c|z^{(2)}_{(i,z_{\pi(ND(s))},\mathbf{x})}]\right|\right] \\
&= 0,
\end{aligned}$$

by the identification of (10) and Jensen's inequality. Therefore, $ACE(\mathbf{u}_c, \mathbf{z}_i, \mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}} = z_{\pi(ND(s))\oplus\pi(s)_{-i}}) = 0$.

From Figure 6, for $s = 1, \cdots, K$, there is no directed path from $\mathbf{z}_i$ to $D(\mathbf{z};\theta)_j$, for all $i \in \pi(s)$ and $j \in \sigma(ND(s))$. Since we assume the blockwise causal relationship between $D(\mathbf{z};\theta)_{\sigma(ND(s))}$ and $\mathbf{u}_{\pi(ND(s))}$, there is also no directed path from $\mathbf{z}_i$ to $\mathbf{u}_j$, for all $i \in \pi(s)$ and $j \in \pi(ND(s))$. Consequently, $ACE(\mathbf{u}_c, \mathbf{z}_i, \mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}} = z_{\pi(ND(s))\oplus\pi(s)_{-i}}) = 0$ is consistent with the fact that there is no total causal effect from $\mathbf{z}_i$ to $\mathbf{u}_j$, for all $i \in \pi(s)$ and $j \in \pi(ND(s))$ because there is no directed path (graphical criteria for total causal effects [21]).

2.

Assume that there is a directed path from $\mathbf{z}_i$ to $\mathbf{z}_c$ where $c \in \pi(l)$ and $l \in \{s\} \cup Des(s)$. Then, by the blockwise causal relationship of Figure 6, there is a directed path from $\mathbf{z}_i$ to $D(\mathbf{z};\theta)_k$, for all $k \in \sigma(l)$. Since we assume the blockwise causal relationship between $D(\mathbf{z};\theta)_{\sigma(l)}$ and $\mathbf{u}_{\pi(l)}$, there is also a directed path from $\mathbf{z}_i$ to $\mathbf{u}_c$ since $c \in \pi(l)$. Therefore, by faithfulness assumption (Assumption 7), there is a total causal effect from $\mathbf{z}_i$ to $\mathbf{u}_c$ because there is a directed path (graphical criteria for total causal effects [21]). Consequently, it implies that $ACE(\mathbf{u}_c, \mathbf{z}_i, \mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}} = z_{\pi(ND(s))\oplus\pi(s)_{-i}}) \neq 0$.

By Definition 5, $D(z^{(j)};\theta)_{\sigma(s)\oplus\sigma(Des(s))}$ only depends on $z_{\pi(s)\oplus\pi(Des(s))}$. Due to different intervention values $z_i^{(1)} \neq z_i^{(2)}$, $z^{(1)}_{(i,z_{\pi(ND(s))},\mathbf{x})}$ and $z^{(2)}_{(i,z_{\pi(ND(s))},\mathbf{x})}$ have the different values for the subvector corresponding to $\pi(s) \oplus \pi(Des(s))$. Therefore,

$$D(z^{(1)}_{(i,z_{\pi(ND(s))},\mathbf{x})};\theta)_{\sigma(s)\oplus\sigma(Des(s))} \neq D(z^{(2)}_{(i,z_{\pi(ND(s))},\mathbf{x})};\theta)_{\sigma(s)\oplus\sigma(Des(s))}. \tag{13}$$

By the causal relationship of GM3 of Figure 1, (6), and (13),

$$\begin{aligned}
\mathbb{E}[\mathbf{u}_c|z^{(1)}_{(i,z_{\pi(ND(s))},\mathbf{x})}] &= \mathbb{E}[\mathbf{u}_c|D(z^{(1)}_{(i,z_{\pi(ND(s))},\mathbf{x})};\theta)] \\
&= \mathbb{E}[\mathbf{u}_c|D(z^{(1)}_{(i,z_{\pi(ND(s))},\mathbf{x})};\theta)_{\sigma(s)\oplus\sigma(Des(s))}] \\
&\neq \mathbb{E}[\mathbf{u}_c|D(z^{(2)}_{(i,z_{\pi(ND(s))},\mathbf{x})};\theta)_{\sigma(s)\oplus\sigma(Des(s))}] \\
&= \mathbb{E}[\mathbf{u}_c|D(z^{(2)}_{(i,z_{\pi(ND(s))},\mathbf{x})};\theta)] \\
&= \mathbb{E}[\mathbf{u}_c|z^{(2)}_{(i,z_{\pi(ND(s))},\mathbf{x})}],
\end{aligned} \tag{14}$$

where $c \notin \pi(ND(s))$.

Since (14) holds for all $\mathbf{x} \in \mathcal{X}$,

$$
\begin{aligned}
0 &< ACE(\mathbf{u}_c, \mathbf{z}_i, \mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}} = z_{\pi(ND(s))\oplus\pi(s)_{-i}}) \\
&= \left|\mathbb{E}[\mathbf{u}_c | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i^{(1)})] - \mathbb{E}[\mathbf{u}_c | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i^{(2)})]\right| \\
&= \left|\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}[\mathbf{u}_c | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(1)}, z_{\pi(Des(s))}(z_i^{(1)}, \mathbf{x})]] - \mathbb{E}_{p(\mathbf{x})}[\mathbb{E}[\mathbf{u}_c | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(2)}, z_{\pi(Des(s))}(z_i^{(2)}, \mathbf{x})]]\right| \\
&= \left|\mathbb{E}_{p(\mathbf{x})}\left[\mathbb{E}[\mathbf{u}_c | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(1)}, z_{\pi(Des(s))}(z_i^{(1)}, \mathbf{x})] - \mathbb{E}[\mathbf{u}_c | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(2)}, z_{\pi(Des(s))}(z_i^{(2)}, \mathbf{x})]\right]\right| \\
&\leq \mathbb{E}_{p(\mathbf{x})}\left[\left|\mathbb{E}[\mathbf{u}_c | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(1)}, z_{\pi(Des(s))}(z_i^{(1)}, \mathbf{x})] - \mathbb{E}[\mathbf{u}_c | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(2)}, z_{\pi(Des(s))}(z_i^{(2)}, \mathbf{x})]\right|\right] \\
&= \mathbb{E}_{p(\mathbf{x})}\left[\left|\mathbb{E}[\mathbf{u}_c | z^{(1)}_{(i, z_{\pi(ND(s))}, \mathbf{x})}] - \mathbb{E}[\mathbf{u}_c | z^{(2)}_{(i, z_{\pi(ND(s))}, \mathbf{x})}]\right|\right] \\
&\neq 0,
\end{aligned}
$$

by the identification of (10) and Jensen's inequality. □

A.3.4 *Average Causal Effect and Potential Outcome*

Furthermore, if we make additional common assumptions of the potential outcome framework, we can interpret the average causal effect defined in Definition 9 in terms of potential outcomes. Consider $i \in \pi(s), s \in \{1, \cdots, K\}$ and $c \in \{1, \cdots, d\}$. For $j = 1, 2$, we denote $\mathbf{u}_c(z_i^{(j)})$ as potential outcome of $\mathbf{u}_c$ under $\mathbf{z}_i = z_i^{(j)}$. And we assume the followings:

1. conditional exchangeability: $\mathbf{u}_c(z_i^{(j)}) \perp\!\!\!\perp \mathbf{z}_i | \mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}$
2. consistency: $\mathbf{z}_i = z_i^{(j)} \Rightarrow \mathbf{u}_c = \mathbf{u}_c(z_i^{(j)})$

Then, for $j = 1, 2$,

$$
\begin{aligned}
&\mathbb{E}[\mathbf{u}_c | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i^{(j)})] \\
=&\mathbb{E}[\mathbf{u}_c | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, \mathbf{z}_i = z_i^{(j)}] && (\because (7)) \\
=&\mathbb{E}[\mathbf{u}_c(z_i^{(j)}) | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, \mathbf{z}_i = z_i^{(j)}] && (\because \text{consistency}) \\
=&\mathbb{E}[\mathbf{u}_c(z_i^{(j)}) | z_{\pi(ND(s))\oplus\pi(s)_{-i}}]. && (\because \text{conditional exchangeability})
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& ACE(\mathbf{u}_c, \mathbf{z}_i, \mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}} = z_{\pi(ND(s))\oplus\pi(s)_{-i}}) \\
= & \left|\mathbb{E}[\mathbf{u}_c | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i^{(1)})] - \mathbb{E}[\mathbf{u}_c | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i^{(2)})]\right| \\
= & \left|\mathbb{E}[\mathbf{u}_c(z_i^{(1)}) | z_{\pi(ND(s))\oplus\pi(s)_{-i}}] - \mathbb{E}[\mathbf{u}_c(z_i^{(2)}) | z_{\pi(ND(s))\oplus\pi(s)_{-i}}]\right|,
\end{aligned}
$$

which means that the average causal effect defined in Definition 9 is equivalent to the difference in the expected potential outcomes.

A.4 *Identification of Causal Disetanglement Metric (CDM)*

For $i \in \pi(s), s = 1, \cdots, K$, denote $\pi(s)_{-i}$ as the tuple $\pi(s)$ without an index i. Since all latent dimensions are active, the posterior variance of $\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}$ is close to zero [1]. It implies that the conditional distribution of $\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}$ given $\mathbf{x}$ defined in (1) can be approximated as

$$
q(\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}} | \mathbf{x}) = \delta(\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}} - z_{\pi(ND(s))\oplus\pi(s)_{-i}}(\mathbf{x})), \tag{15}
$$

where $z_{\pi(ND(s))\oplus\pi(s)_{-i}}(\mathbf{x})$ is the deterministic component value of posterior distribution given $\mathbf{x}$, and δ is the Dirac delta function such that

$$
\delta(v) = \begin{cases} +\infty, & \text{if } v = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{16}
$$

Based on the results of Appendix A.3, for $c = 1, \cdots, d$, $CDM(c,i)$ is identified as

$$
\begin{aligned}
& CDM(c,i) \\
= \; & \mathbb{E}_{p(\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}})}\Big[\Big|\mathbb{E}[\mathbf{u}_c|\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i^{(1)})] - \mathbb{E}[\mathbf{u}_c|\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i^{(2)})]\Big|\Big] \\
= \; & \mathbb{E}_{p(\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}})}\Big[\Big|\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}[\mathbf{u}_c|\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(1)}, \mathbf{z}_{\pi(Des(s))}(z_i^{(1)}, \mathbf{x})]] \\
& -\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}[\mathbf{u}_c|\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(2)}, \mathbf{z}_{\pi(Des(s))}(z_i^{(2)}, \mathbf{x})]]\Big|\Big] \\
= \; & \mathbb{E}_{p(\mathbf{x})\cdot q(\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}|\mathbf{x})}\Big[\Big|\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}[\mathbf{u}_c|\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(1)}, \mathbf{z}_{\pi(Des(s))}(z_i^{(1)}, \mathbf{x})]] \\
& -\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}[\mathbf{u}_c|\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(2)}, \mathbf{z}_{\pi(Des(s))}(z_i^{(2)}, \mathbf{x})]]\Big|\Big] \\
= \; & \mathbb{E}_{p(\mathbf{x})}\int \delta(\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}} - z_{\pi(ND(s))\oplus\pi(s)_{-i}}(\mathbf{x}))\Big[\Big|\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}[\mathbf{u}_c|\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(1)}, \mathbf{z}_{\pi(Des(s))}(z_i^{(1)}, \mathbf{x})]] \\
& -\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}, z_i^{(2)}, \mathbf{z}_{\pi(Des(s))}(z_i^{(2)}, \mathbf{x})]]\Big|\Big] d\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}} \\
= \; & \mathbb{E}_{p(\mathbf{x})}\Big[\Big|\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}(\mathbf{x}), z_i^{(1)}, z_{\pi(Des(s))}(z_i^{(1)}, \mathbf{x})]] \\
& -\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}(\mathbf{x}), z_i^{(2)}, z_{\pi(Des(s))}(z_i^{(2)}, \mathbf{x})]]\Big|\Big],
\end{aligned}
$$

where $p(\mathbf{x})$ is the probability density function of $\mathbf{x}$ and $p(\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}) = \int p(\mathbf{x}) \cdot q(\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}|\mathbf{x})d\mathbf{x}$.

Finally, we can obtain the lower and upper bounds of CDM as

$$CDM_L(c,i) \leq CDM(c,i) \leq CDM_U(c,i)$$

where

$$
\begin{aligned}
CDM_L(c,i) \; \coloneqq \; & \Big|\mathbb{E}_{p(\mathbf{x})}\Big[\mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}(\mathbf{x}), z_i^{(1)}, z_{\pi(Des(s))}(z_i^{(1)}, \mathbf{x})] \\
& -\mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}(\mathbf{x}), z_i^{(2)}, z_{\pi(Des(s))}(z_i^{(2)}, \mathbf{x})]\Big]\Big| \\
= \; & \Big|\mathbb{E}_{p(\mathbf{x})}\Big[\mathbb{E}[\mathbf{u}_c|D\big(z_{\pi(ND(s))\oplus\pi(s)_{-i}}(\mathbf{x}), z_i^{(1)}, z_{\pi(Des(s))}(z_i^{(1)}, \mathbf{x}); \theta\big)] \\
& -\mathbb{E}[\mathbf{u}_c|D\big(z_{\pi(ND(s))\oplus\pi(s)_{-i}}(\mathbf{x}), z_i^{(2)}, z_{\pi(Des(s))}(z_i^{(2)}, \mathbf{x}); \theta\big)]\Big]\Big| \\
CDM_U(c,i) \; \coloneqq \; & \mathbb{E}_{p(\mathbf{x})}\Big[\Big|\mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}(\mathbf{x}), z_i^{(1)}, z_{\pi(Des(s))}(z_i^{(1)}, \mathbf{x})] \\
& -\mathbb{E}[\mathbf{u}_c|z_{\pi(ND(s))\oplus\pi(s)_{-i}}(\mathbf{x}), z_i^{(2)}, z_{\pi(Des(s))}(z_i^{(2)}, \mathbf{x})]\Big|\Big] \\
= \; & \mathbb{E}_{p(\mathbf{x})}\Big[\Big|\mathbb{E}[\mathbf{u}_c|D\big(z_{\pi(ND(s))\oplus\pi(s)_{-i}}(\mathbf{x}), z_i^{(1)}, z_{\pi(Des(s))}(z_i^{(1)}, \mathbf{x}); \theta\big)] \\
& -\mathbb{E}[\mathbf{u}_c|D\big(z_{\pi(ND(s))\oplus\pi(s)_{-i}}(\mathbf{x}), z_i^{(2)}, z_{\pi(Des(s))}(z_i^{(2)}, \mathbf{x}); \theta\big)]\Big|\Big],
\end{aligned}
$$

by (6) and Jensen's inequality.

A.4.1 Metric Computation Details

Consider a disentangled representation (Definition 4) map $w(\cdot;\eta)$ where $w : \mathbb{R}^p \mapsto \mathbb{R}^d$ is a neural network parameterized with η. By Definition 4,

$$\begin{aligned} w(\mathbf{x};\eta)_c &= h_c^{-1}(\mathbb{E}[\mathbf{u}_c|\mathbf{x}]) \\ h_c(w(D(\mathbf{z};\theta);\eta)_c) &= \mathbb{E}[\mathbf{u}_c|D(\mathbf{z};\theta)], \end{aligned} \tag{17}$$

for $c = 1, \cdots, d$.

By (17), $\mathbb{E}[\mathbf{u}_c|D(\mathbf{z};\theta)]$ in the lower and upper bound of CDM ($CDM_L(c,i), CDM_U(c,i)$) is replaced with $h_c(w(D(\mathbf{z};\theta);\eta)_c)$. In order to calculate metrics from different models and make fair comparisons, the same architecture of w and parameter η are used for all models. Since $\mathbf{u} \in [0,1]^d$, the sigmoid function σ is used for the function h_c for all $c = 1, \cdots, d$.

For $s = 1, \cdots, K$, the blockwise causal relationship from $D(\mathbf{z};\theta)_\sigma$ to $\mathbf{u}_\pi$ can be written as

$$\mathbb{E}[\mathbf{u}_{\pi(s)}|D(\mathbf{z};\theta)] = \mathbb{E}[\mathbf{u}_{\pi(s)}|D(\mathbf{z};\theta)_{\sigma(s)}], \tag{18}$$

where the rearranged vector of $\mathbf{u}$ with π is denoted as $\mathbf{u}_\pi$.

To ensure that (18) is satisfied with $w(\cdot;\eta)$, we construct w as

$$w(\mathbf{x};\eta)_{\pi(s)} \coloneqq w_s(\mathbf{x}_{\sigma(s)};\eta_s)$$

where $\mathbf{x} \in \mathbb{R}^p$, $w_s : \mathbb{R}^{|\sigma(s)|} \mapsto \mathbb{R}^{|\pi(s)|}$ is a neural network parameterized with η_s, $|\pi(s)|$ and $|\sigma(s)|$ are cardinalities of partition tuples, and $\eta = (\eta_1, \cdots, \eta_K)$. Finally, we estimate the parameter η using the following objective function:

$$\max_\eta \frac{1}{|I_L|} \sum_{i=1}^{|I_L|} \sum_{j=1}^{d} u_j^{(i)} \cdot \log \sigma(w(x^{(i)};\eta)_j) + (1 - u_j^{(i)}) \cdot \log(1 - \sigma(w(x^{(i)};\eta)_j)), \tag{19}$$

where $(x^{(i)}, u^{(i)})$ is the ith observation of the annotated dataset I_L, $|I_L|$ is cardinality of I_L, and subscript j indicates its jth element.

A.5 Pendulum Dataset Details

The ground-truth factors.

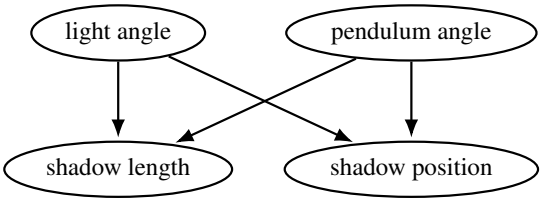


Figure 8: DAG structure of the ground-truth factors of the pendulum dataset.

- light angle: angle of light which determines the position in x-axis (the position in y-axis is fixed)
- pendulum angle: angle of the pendulum (amplitude)
- shadow length: length of the pendulum shadow
- shadow position: position of the pendulum shadow center

The ground-truth SCM.

$$
\begin{aligned}
\mathit{light\_angle} &\coloneqq U\left(\frac{\pi}{4}, \frac{\pi}{2}\right) \\
\mathit{pendulum\_angle} &\coloneqq U\left(0, \frac{\pi}{4}\right) \\
\mathit{shadow\_length} &\coloneqq \left(x + \ell \cdot \sin(\mathit{pendulum\_angle}) - \frac{y - \ell \cdot \cos(\mathit{pendulum\_angle}) + 0.5}{\tan(\mathit{light\_angle})}\right) \\
&\quad - \left(x - \frac{y + 0.5}{\tan(\mathit{light\_angle})}\right) + \delta_1 \\
\mathit{shadow\_position} &\coloneqq 0.5 \cdot \left(x + \ell \cdot \sin(\mathit{pendulum\_angle}) - \frac{y - \ell \cdot \cos(\mathit{pendulum\_angle}) + 0.5}{\tan(\mathit{light\_angle})} + x \right. \\
&\quad \left. - \frac{y + 0.5}{\tan(\mathit{light\_angle})}\right) + \delta_2,
\end{aligned}
$$

where (x, y) is a coordinate of frictionless pivot, ℓ is length of pendulum, and $\delta_1, \delta_2 \sim N(0, (0.1)^2)$ are independent measurement errors.

A.6 Experimental Details

We run all experiments using Geforce RTX 3090 GPU, and our experimental codes are all available with `pytorch`. First of all, we scaled each column of B with the in-degree of the node for stability.

A.6.1 Pendulum Dataset

- The images were scaled from -1 to 1. Both width and height are size 64.
- We consider the 4-dimensional latent space. In other words, the dimension size of the latent variable is the same as the number of ground-truth factors. The encoder returns the Gaussian distribution parameters, mean vector, and diagonal covariance elements. Thus, the encoder maps $\mathcal{X}$ to $(\mathbb{R}^4 \times \mathbb{R}_+^4)$.
- We used the planar flow with the single transformation [25] for the nonlinear function f in structural functions (1) and its inverse is computed via fixed-point iteration [2].
- Downstream task labeling: We generated the target label $\mathbf{y}$ as

$$
\begin{aligned}
\mathit{logit}(\mathbf{g}) &= \beta^\top \mathbf{g} + 2\sin(\beta^\top \mathbf{g}) \\
\mathbf{y}|\mathbf{g} &\sim \mathit{Ber}\Big(1/(1 + \exp(-\mathit{logit}(\mathbf{g})))\Big)
\end{aligned}
$$

where $\beta^\top = [1, -1, 0.5, -0.5]$.

Table 5: Model descriptions of the encoder and decoder. Here, K is the number of partition blocks. The encoder is shared for VAE, InfoMax VAE, and CDG-VAE. VAE and InfoMax VAE use the decoder in the middle column; however, CDG-VAE uses the decoder in the right column. $|\sigma(j)|$ denotes the cardinality of partition tuple $\sigma(j)$.

encoder	**decoder**	**decoder of CDG-VAE**
(64, 64, 3) image	d-dimension latent	$\|\pi(j)\|$-dimensional latent, $j = 1, \cdots, K$
Flatten	Dense(300), ELU	$K\times$ (Dense(300), ELU)
Dense(300), ELU	Dense(300), ELU	$K\times$ (Dense(300), ELU)
Dense(300), ELU	Dense($64 \times 64 \times 3$), tanh	Dense($\|\sigma(j)\|$), tanh, $j = 1, \cdots, K$
$2 \times$ Dense(d), Linear	Reshape(64, 64, 3)	Concatenate, Reshape(64, 64, 3)

Table 6: Hyper-parameter settings for the pendulum dataset experiments. For CausalVAE and DEAR, we used default settings in each paper.

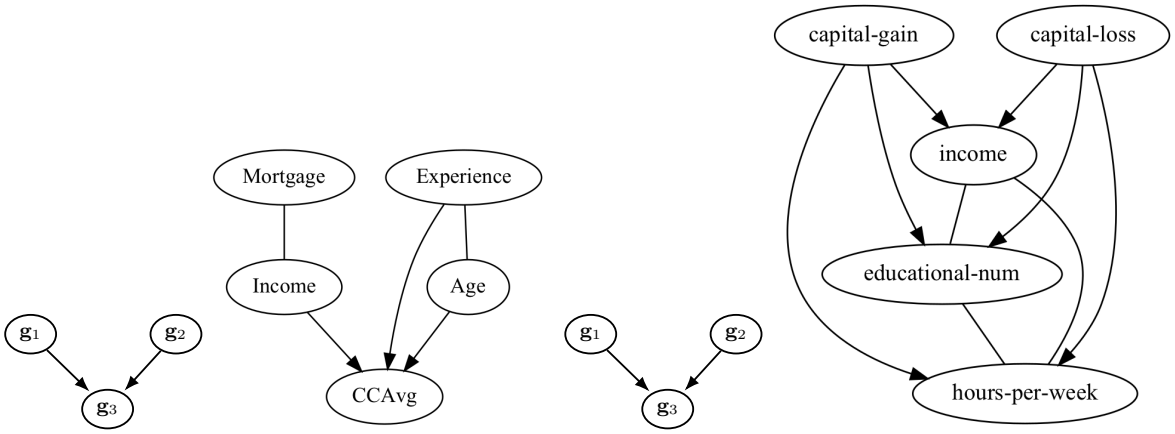
Model	epochs	batch size	learning rate	β	λ	γ
VAE	100	128	0.001	0.1	5	-
InfoMax	100	128	0.001	0.1	5	5
CDG-VAE	100	128	0.001	0.1	5	-

A.6.2 Tabular Datasets

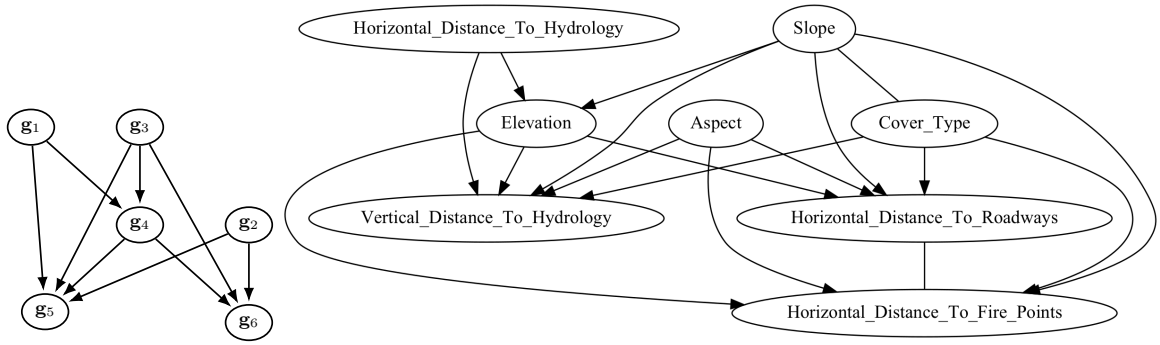
- `loan` dataset:
  - Target variable: `CCAvg` (regression problem)
  - Train/Test split: 4k/1k
  - Chain components of $\mathbf{x}_\sigma$ are $\mathbf{x}_{\sigma(1)} = (\text{Mortgage, Income})$, $\mathbf{x}_{\sigma(2)} = (\text{Experiences, Age})$, and $\mathbf{x}_{\sigma(3)} = (\text{CCAvg})$.
  - https://www.kaggle.com/datasets/teertha/personal-loan-modeling
- `adult` dataset:
  - Target variable: `income` (binary classification problem)
  - Train/Test split: 40k/5k
  - Chain components of $\mathbf{x}_\sigma$ are $\mathbf{x}_{\sigma(1)} = (\text{capital-gain})$, $\mathbf{x}_{\sigma(2)} = (\text{capital-loss})$, and $\mathbf{x}_{\sigma(3)} = (\text{income, educational-num, hours-per-week})$.
  - https://archive.ics.uci.edu/ml/datasets/Adult
- `covertype` dataset:
  - Target variable: `Cover_Type` (multi-class classification problem)
  - Train/Test split: 13k/2k
  - Chain components of $\mathbf{x}_\sigma$ are $\mathbf{x}_{\sigma(1)} = (\text{Horizontal\_Distance\_To\_Hydrology})$, $\mathbf{x}_{\sigma(2)} = (\text{Aspect})$, $\mathbf{x}_{\sigma(3)} = (\text{Slope, Cover\_Type})$, $\mathbf{x}_{\sigma(4)} = (\text{Elevation})$, $\mathbf{x}_{\sigma(5)} = (\text{Vertical\_Distance\_To\_Hydrology})$, and $\mathbf{x}_{\sigma(6)} = (\text{Horizontal\_Distance\_To\_Roadways, Horizontal\_Distance\_To\_Fire\_Points})$.
  - https://www.kaggle.com/competitions/forest-cover-type-prediction/data?select=train.csv

Table 7: Classifier and regressor used in the evaluation of synthetic data quality. The names of all parameters used in the description are consistent with those defined in corresponding packages.

Tasks	Model	Description
`Regression`	Linear Regression	Package: `sklearn.linear_model.LinearRegression`, setting: defaulted values
	Random Forest	Package: `sklearn.ensemble.RandomForestRegressor`, setting: random_state=0, and defaulted values
	Gradient Boosting	Package: `sklearn.ensemble.GradientBoostingRegressor`, setting: random_state=0, and defaulted values
`Classification`	Logistic Regression	Package: `sklearn.linear_model.LogisticRegression`, setting: multi_class='ovr', and defaulted values
	Random Forest	Package: `sklearn.ensemble.RandomForestClassifier`, setting: random_state=0, and defaulted values
	Gradient Boosting	Package: `sklearn.ensemble.GradientBoostingClassifier`, setting: random_state=0, and defaulted values



(a) `loan` dataset (b) `adult` dataset



(c) `covertype` dataset

Figure 9: DAG structures of ground-truth factors and chain graphs of real tabular datasets.

Table 8: Hyper-parameter settings for tabular dataset experiments.

Dataset	Model	epochs	batch size	learning rate	latent dimension	β (range)	λ	γ
`loan`	VAE	200	256	0.01	3	0.01	1	-
	InfoMax	200	256	0.01	3	0.01	1	1
	TVAE	200	256	0.005	3	[0.01, 0.1]	-	-
	CTAB-GAN	150	500	0.0002	3	-	-	-
	CDG-VAE	200	256	0.01	3	0.01	5	-
	CDG-TVAE	300	256	0.001	3	[0.01, 0.1]	5	-
`adult`	VAE	200	256	0.01	3	0.01	10	-
	InfoMax	200	256	0.01	3	0.01	5	1
	TVAE	200	256	0.005	3	[0.01, 0.1]	-	-
	CTAB-GAN	150	500	0.0002	3	-	-	-
	CDG-VAE	200	256	0.01	3	0.01	5	-
	CDG-TVAE	300	256	0.001	3	[0.1, 1]	5	-
`covertype`	VAE	200	256	0.01	6	0.01	10	-
	InfoMax	200	256	0.01	6	0.01	10	1
	TVAE	200	256	0.005	6	[0.01, 0.1]	-	-
	CTAB-GAN	150	500	0.0002	6	-	-	-
	CDG-VAE	200	256	0.01	6	0.01	10	-
	CDG-TVAE	300	256	0.001	6	[0.005, 0.01]	10	-

A.7 Toy Example

A.7.1 Distributional Robustness of Causal Representation

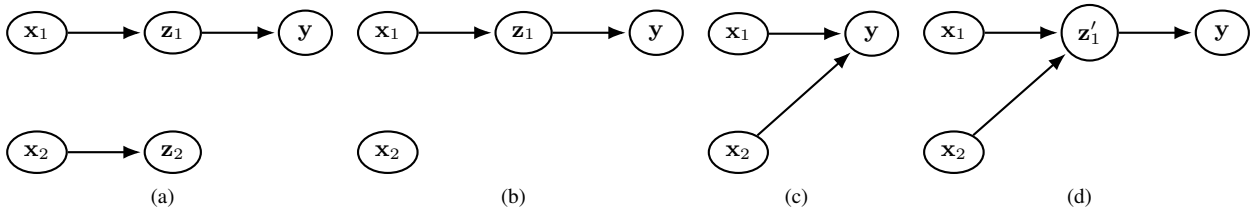


Figure 10: (a) The ground-truth DAG structure. (b) DAG structure of causally disentangled VAE. (c) DAG structure of ERM. (d) DAG structure of entangled VAE.

Let $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2]^\top$ to be an observation and $\mathbf{y}$ is the target label. Suppose that $\mathbf{x}_1$ and $\mathbf{x}_2$ are disentangled block partitions and denote their representations as $\mathbf{z}_1, \mathbf{z}_2$, respectively. And we assume that the target label $\mathbf{y}$ is causally generated by $\mathbf{z}_1$. It implies that $\mathbf{x}_2$ is the spurious feature. Therefore, the representation of interest is $\mathbf{z}_1$, and our goal is learning the causally disentangled VAE model, which can extract the representation $\mathbf{z}_1$, which is the causal representation of target label $\mathbf{y}$. To impose a distributional shift, we manipulated $\mathbf{x}_2$ to correlate strongly with the training dataset's target label and not with the test dataset's target label. In our experiment setting, $Corr(\mathbf{x}_2, \mathbf{y}) = 0.715$ in the training dataset and $Corr(\mathbf{x}_2, \mathbf{y}) = 0.002$ in the test dataset.

Here, we compared the downstream task's performance in distributional robustness. First, in the causally disentangled VAE case, we trained the classifier upon the learned disentangled representation $\mathbf{z}_1$. Next, in ERM (Empirical Risk Minimization) case, we trained the classifier directly upon $\mathbf{x}_1$ and $\mathbf{x}_2$. Lastly, in the entangled VAE case, the encoder of entangled VAE learns the entangled representation $\mathbf{z}_1'$ in which information of $\mathbf{x}_1$, and $\mathbf{x}_2$ are mixed (note that $\mathbf{z}_1 \neq \mathbf{z}_1'$). The causal DAGs of these cases are visualized in Figure 10.

The classification accuracy is shown in Table 9. Table 9 indicates that the causally disentangled VAE shows the best test accuracy. This is because the disentangled encoder only captures the information of $\mathbf{x}_1$, so it does not affect by the distribution shift of $\mathbf{x}_2$. However, models fitted with ERM and the entangled VAE are vulnerable to the dataset's distribution shift because directed paths exist from $\mathbf{x}_2$ to $\mathbf{y}$. Therefore, the disentangled representation is invariant to distribution shifts (distributional robustness property).

Table 9: Train and test dataset classification accuracies of a toy example.

Model	Train Accuracy(%)	Test Accuracy(%)
ERM	89.56	64.99
Entangled VAE	**89.59**	64.91
Causally disentangled VAE	77.05	**78.23**

A.7.2 Necessary of Disentangled Decoder

Suppose two random variables exist $\mathbf{x}_1, \mathbf{x}_2$, and the ground-truth causal relationship is $\mathbf{x}_1 \to \mathbf{x}_2$. Let the disentangled latent variables $\mathbf{z}_1, \mathbf{z}_2$ satisfying $\mathbf{x}_1 \to \mathbf{z}_1$ and $\mathbf{x}_2 \to \mathbf{z}_2$. And suppose that we have the disentangled encoder where $\mathbf{z}_1 = \boldsymbol{\epsilon}_1$, $\mathbf{z}_2 = \alpha \cdot \mathbf{z}_1 + \boldsymbol{\epsilon}_2$. We consider two cases below that can be obtained from the empirical risk minimization:

1. $\mathbf{x}_1 = \beta \cdot \mathbf{z}_1$, $\mathbf{x}_2 = \gamma \cdot \mathbf{z}_2$, and
2. $\mathbf{x}_1 = \beta/\alpha \cdot \mathbf{z}_2$, $\mathbf{x}_2 = \gamma \cdot \mathbf{z}_2$.

The first case satisfies the desired causal disentanglement in the decoder. But the second case does not. In the second case, if we intervene on $\mathbf{z}_2$ (the child of $\mathbf{z}_1$), $\mathbf{x}_1$ is affected by the intervention, which is not desired. This is a simple example where the disentangled decoder is required. Figure 11 in Appendix A.8 shows the example that the causally disentangled image generation can not be obtained where the encoder of the VAE model is causally disentangled, but the generated image from the decoder is not.

A.8 Additional Experiment Results

(a) Cross-entropy loss (b) Counterfactual images (c) Posterior variances

Figure 11: Visualizations of the cross-entropy loss values for each latent dimension, generated counterfactual images under do-interventions, and posterior variances from the fitted VAE model with linear f.

The left panel of Figure 11 shows that all latent dimensions have low cross-entropy loss values, where the cross-entropy loss plays the role of regularizing the encoder. The left panel implies that the VAE model has a disentangled representation. In the middle panel of Figure 11, from top to bottom, intervened dimensions are light angle, pendulum angle, shadow length, and shadow position. Note that generated counterfactual images in the bottom row are the same when the intervened dimension is shadow position. The right panel of Figure 11 indicates that the latent dimension of the shadow position is non-active since its posterior variance is much larger than zero. Therefore, Figure 11 indicates that the causally plausible counterfactual samples can not be generated when the non-active latent dimension is intervened.

Table 10: Metrics overview of **Interventional Robustness**: The pairs of numbers are the lower and upper bounds of CDM. 'L' and 'NL' denote the model with linear and nonlinear f, and '*' denotes the semi-supervised learned model. Mean and standard deviation values are obtained from 10 repeated experiments. 'pos' denotes shadow position.

Model	$CDM(light, length)$	$CDM(angle, length)$	$CDM(light, pos)$	$CDM(angle, pos)$
VAE(L)	$(0.44, 0.44)_{\pm(0.35,0.35)}$	$(0.33, 0.33)_{\pm(0.33,0.33)}$	$(0.38, 0.38)_{\pm(0.33,0.32)}$	$(0.28, 0.28)_{\pm(0.30,0.31)}$
VAE(NL)	$(0.38, 0.40)_{\pm(0.28,0.27)}$	$(0.30, 0.32)_{\pm(0.27,0.26)}$	$(0.40, 0.42)_{\pm(0.28,0.26)}$	$(0.27, 0.33)_{\pm(0.25,0.24)}$
InfoMax(L)	$(0.42, 0.43)_{\pm(0.39,0.38)}$	$(0.23, 0.24)_{\pm(0.27,0.27)}$	$(0.35, 0.37)_{\pm(0.27,0.25)}$	$(0.38, 0.38)_{\pm(0.34,0.34)}$
InfoMax(NL)	$(0.37, 0.39)_{\pm(0.32,0.30)}$	$(0.21, 0.24)_{\pm(0.23,0.22)}$	$(0.40, 0.42)_{\pm(0.26,0.23)}$	$(0.26, 0.33)_{\pm(0.28,0.25)}$
CausalVAE	$(0.28, 0.28)_{\pm(0.11,0.10)}$	$(0.21, 0.22)_{\pm(0.10,0.09)}$	$(0.33, 0.33)_{\pm(0.06,0.06)}$	$(0.17, 0.17)_{\pm(0.09,0.08)}$
DEAR	$(0.21, 0.23)_{\pm(0.16,0.15)}$	$(0.27, 0.28)_{\pm(0.19,0.19)}$	$(0.27, 0.30)_{\pm(0.22,0.20)}$	$(0.26, 0.29)_{\pm(0.25,0.24)}$
CDG-VAE(L)	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$
CDG-VAE(NL)	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$
CDG-VAE(L)*	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$
CDG-VAE(NL)*	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$

Table 11: Metrics overview of **Counterfactural Generativeness**: The pairs of numbers are lower and upper bounds of CDM. 'L' and 'NL' denote the model with linear and nonlinear f, and '*' denotes the semi-supervised learned model. Mean and standard deviation values are obtained from 10 repeated experiments. 'pos' denotes shadow position.

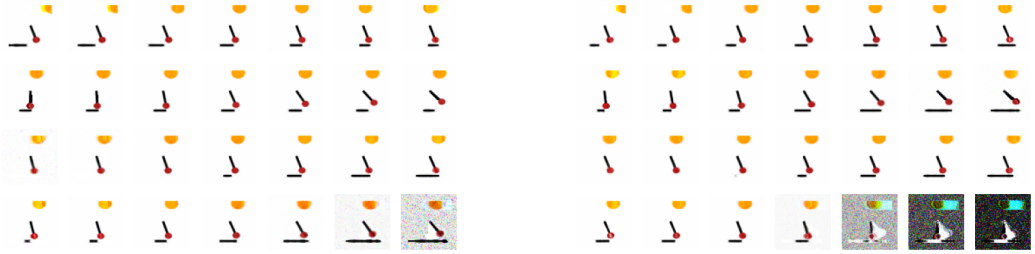
Model	$CDM(light, light)$	$CDM(length, light)$	$CDM(pos, light)$
VAE(L)	$(0.65, 0.65)_{\pm(0.27,0.27)}$	$(0.49, 0.49)_{\pm(0.23,0.23)}$	$(0.51, 0.52)_{\pm(0.27,0.26)}$
VAE(NL)	$(0.68, 0.68)_{\pm(0.22,0.22)}$	$(0.42, 0.43)_{\pm(0.27,0.25)}$	**(0.52, 0.53)**$_{\pm(0.22,0.21)}$
InfoMax(L)	$(0.61, 0.62)_{\pm(0.27,0.25)}$	**(0.56, 0.56)**$_{\pm(0.28,0.27)}$	$(0.34, 0.36)_{\pm(0.25,0.21)}$
InfoMax(NL)	$(0.63, 0.64)_{\pm(0.24,0.22)}$	$(0.48, 0.49)_{\pm(0.28,0.27)}$	$(0.43, 0.44)_{\pm(0.22,0.19)}$
CausalVAE	$(0.29, 0.29)_{\pm(0.12,0.12)}$	$(0.15, 0.15)_{\pm(0.04,0.04)}$	$(0.18, 0.18)_{\pm(0.08,0.08)}$
DEAR	$(0.41, 0.42)_{\pm(0.29,0.28)}$	$(0.21, 0.23)_{\pm(0.13,0.13)}$	$(0.23, 0.25)_{\pm(0.19,0.18)}$
CDG-VAE(L)	$(0.85, 0.85)_{\pm(0.02,0.02)}$	$(0.29, 0.29)_{\pm(0.13,0.13)}$	$(0.35, 0.36)_{\pm(0.14,0.13)}$
CDG-VAE(NL)	**(0.86, 0.86)**$_{\pm(0.02,0.02)}$	$(0.37, 0.37)_{\pm(0.19,0.19)}$	$(0.35, 0.35)_{\pm(0.12,0.11)}$
CDG-VAE(L)*	$(0.84, 0.84)_{\pm(0.01,0.01)}$	$(0.26, 0.26)_{\pm(0.10,0.10)}$	$(0.38, 0.38)_{\pm(0.14,0.13)}$
CDG-VAE(NL)*	**(0.86, 0.86)**$_{\pm(0.02,0.02)}$	$(0.30, 0.30)_{\pm(0.09,0.09)}$	$(0.36, 0.36)_{\pm(0.11,0.10)}$

Model	$CDM(angle, angle)$	$CDM(length, angle)$	$CDM(pos, angle)$
VAE(L)	$(0.76, 0.76)_{\pm(0.31,0.31)}$	$(0.31, 0.32)_{\pm(0.16,0.15)}$	$(0.33, 0.33)_{\pm(0.21,0.20)}$
VAE(NL)	$(0.78, 0.78)_{\pm(0.23,0.23)}$	$(0.33, 0.34)_{\pm(0.12,0.12)}$	$(0.29, 0.29)_{\pm(0.15,0.15)}$
InfoMax(L)	$(0.69, 0.69)_{\pm(0.32,0.32)}$	$(0.40, 0.40)_{\pm(0.26,0.25)}$	**(0.38, 0.38)**$_{\pm(0.17,0.17)}$
InfoMax(NL)	$(0.75, 0.75)_{\pm(0.25,0.25)}$	**(0.44, 0.44)**$_{\pm(0.21,0.21)}$	$(0.34, 0.34)_{\pm(0.13,0.13)}$
CausalVAE	$(0.04, 0.05)_{\pm(0.04,0.04)}$	$(0.10, 0.10)_{\pm(0.04,0.04)}$	$(0.13, 0.14)_{\pm(0.05,0.05)}$
DEAR	$(0.45, 0.46)_{\pm(0.30,0.29)}$	$(0.23, 0.25)_{\pm(0.23,0.23)}$	$(0.14, 0.21)_{\pm(0.17,0.17)}$
CDG-VAE(L)	**(0.95, 0.95)**$_{\pm(0.01,0.01)}$	$(0.24, 0.25)_{\pm(0.10,0.09)}$	$(0.32, 0.33)_{\pm(0.11,0.11)}$
CDG-VAE(NL)	$(0.92, 0.92)_{\pm(0.10,0.10)}$	$(0.35, 0.36)_{\pm(0.16,0.15)}$	$(0.34, 0.34)_{\pm(0.10,0.09)}$
CDG-VAE(L)*	$(0.94, 0.94)_{\pm(0.01,0.01)}$	$(0.21, 0.22)_{\pm(0.09,0.07)}$	$(0.33, 0.34)_{\pm(0.10,0.10)}$
CDG-VAE(NL)*	$(0.92, 0.92)_{\pm(0.09,0.09)}$	$(0.29, 0.30)_{\pm(0.12,0.11)}$	$(0.35, 0.35)_{\pm(0.09,0.09)}$

Model	$CDM(length, length)$	$CDM(pos, pos)$
VAE(L)	$(0.43, 0.43)_{\pm(0.36,0.36)}$	$(0.27, 0.28)_{\pm(0.25,0.24)}$
VAE(NL)	$(0.56, 0.56)_{\pm(0.31,0.31)}$	$(0.31, 0.34)_{\pm(0.21,0.20)}$
InfoMax(L)	$(0.35, 0.35)_{\pm(0.30,0.30)}$	$(0.29, 0.31)_{\pm(0.22,0.20)}$
InfoMax(NL)	$(0.52, 0.52)_{\pm(0.30,0.29)}$	$(0.31, 0.34)_{\pm(0.19,0.16)}$
CausalVAE	$(0.18, 0.18)_{\pm(0.03,0.03)}$	$(0.29, 0.29)_{\pm(0.09,0.09)}$
DEAR	$(0.22, 0.25)_{\pm(0.19,0.18)}$	$(0.16, 0.20)_{\pm(0.18,0.16)}$
CDG-VAE(L)	$(0.77, 0.77)_{\pm(0.25,0.25)}$	$(0.69, 0.69)_{\pm(0.25,0.25)}$
CDG-VAE(NL)	**(0.83, 0.83)**$_{\pm(0.19,0.19)}$	**(0.78, 0.78)**$_{\pm(0.24,0.24)}$
CDG-VAE(L)*	$(0.83, 0.83)_{\pm(0.14,0.14)}$	$(0.66, 0.66)_{\pm(0.22,0.22)}$
CDG-VAE(NL)*	**(0.86, 0.86)**$_{\pm(0.10,0.10)}$	**(0.79, 0.79)**$_{\pm(0.21,0.21)}$

(a) VAE(L)

(b) VAE(NL)

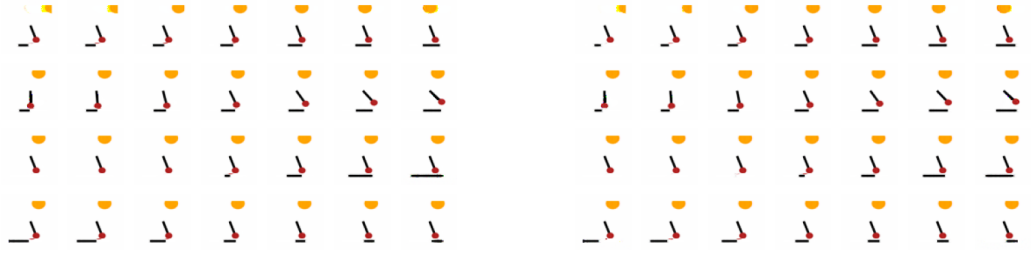


(c) InfoMax VAE(L)

(d) InfoMax VAE(NL)

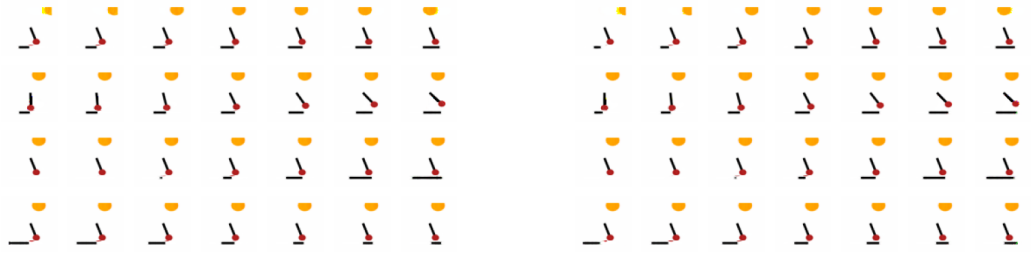
(e) CausalVAE

(f) DEAR



(g) CDG-VAE(L)

(h) CDG-VAE(NL)



(i) CDG-VAE(L)*

(j) CDG-VAE(NL)*

Figure 12: Visualizations of generated counterfactual images under *do*-interventions from various models. From top to bottom, intervened dimensions are light angle, pendulum angle, shadow length, and shadow position. 'L' and 'NL' denote the model with linear and nonlinear f, and '*' denotes the semi-supervised learned model.