A Introduction of do calculus.

Do-calculus consists of three rules that help with identifying causal effects.

Rule A.1 (Insertion/deletion of observations)**.**

$$P(y|do(x), z, w) = P(y|do(x), w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}} \tag{13}$$

Rule A.2 (Action/observation exchange)**.**

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}} \tag{14}$$

Rule A.3 (Insertion/deletion of actions)**.**

$$P(y|do(x), do(z), w) = P(y|do(x), w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\,\overline{Z(W)}}} \tag{15}$$

where $G_{\overline{X}}$ is the graph with all incoming edges to X being removed, $G_{\underline{W}}$ is the graph with all outcoming edges to W being removed, and $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node.

Intuitively, Rule A.1 states when an observant can be omitted in estimating the interventional distribution, Rule A.2 illustrates under what condition, the interventional distribution can be estimated using the observational dataset, and Rule A.3 decides when we can ignore an intervention.

B Proofs

B.1 Proof of Proposition 2.1

Proposition B.1. *Let Z_1 and Z_2 be two random variables, $\mathbf{C}^*$ be the ground truth confounder set. If $\mathbf{C}$ is a superset of or is equivalent to $\mathbf{C}^*$, i.e., $\mathbf{C}^* \subseteq \mathbf{C}$, with c being a realization of $\mathbf{C}$, we have*

$$P(Z_2|do(Z_1)) = \sum_{c \in \mathbf{C}} P(Z_2|Z_1, \mathbf{C} = c)P(\mathbf{C} = c) \tag{16}$$

if no $C \in \mathbf{C}$ is a descendent of $\mathbf{Z}$.

Proof.

$$\begin{aligned} P(Z_2|do(Z_1)) &= P(Z_2|do(Z_1), \mathbf{C})P(\mathbf{C}|do(Z_1)) \\ P(Z_2|do(Z_1), \mathbf{C}) &\overset{\text{Rule A.2}}{=\!=\!=} P(Z_2|Z_1, \mathbf{C}) \\ P(\mathbf{C}|do(Z_1)) &\overset{\text{Rule A.3}}{=\!=\!=} P(\mathbf{C}) \\ P(Z_2|do(Z_1)) &= \sum_{c \in \mathbf{C}} P(Z_2|Z_1, \mathbf{C} = c)P(\mathbf{C} = c) \end{aligned}$$

□

B.2 Proof of Theorem 4.1

Theorem B.2. *Suppose that the latent variable $\mathbf{Z}$ on dataset $\mathbf{X}$ given $\mathbf{C} = c$ is Gaussian $\mathcal{N}(\mu^c(\mathbf{X}), \Sigma^c(\mathbf{X}))$. Specifically,*

$$P(\mathbf{Z}|\mathbf{C} = c, \mathbf{X}) = (2\pi)^{-D/2} \det(\Sigma^c)^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{Z} - \mu^c)^\mathsf{T}(\Sigma^c)^{-1}(\mathbf{Z} - \mu^c)\right),$$

where $\mathbf{Z} \in \mathbb{R}^D$. If $\Sigma^c(\mathbf{X})$ is diagonal for all c, we have

$$l_c = \sum_{i=1}^{D} [\mathbb{E}(Z_i|do^c(Z_{-i}), \mathbf{X}) - \mathbb{E}(Z_i|\mathbf{X})] = 0. \tag{17}$$

Proof. We suppose that

$$P(\mathbf{Z}|\mathbf{C}=c,\mathbf{X}) = (2\pi)^{-D/2}\det(\Sigma^c)^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{Z}-\mu^c)^{\mathsf{T}}(\Sigma^c)^{-1}(\mathbf{Z}-\mu^c)\right) \tag{18}$$

where we omit $\mathbf{X}$ for simplicity and D is the dimension of $\mathbf{Z}$ for any given c. By definition of l_c (Equation (9)) and proposition 2.1,

$$l_c = \sum_{i=1}^{D} d\left(\mathbb{E}[Z_i|do^c(Z_{-i}),\mathbf{X}] - \mathbb{E}[Z_i|\mathbf{X}]\right) \tag{19}$$

$$= \sum_{i}^{D} d(E[Z_i|Z_{-i},\mathbf{X},C=c], E[Z_i|\mathbf{X},C=c]) \tag{20}$$

$$= \sum_{i}^{D} d(E[Z_i^c|Z_{-i}^c], E[Z_i^c]) \tag{21}$$

where we denote $\mathbf{Z}^c = [\mathbf{Z}|\mathbf{X},C=c]$ for simplicity. Notice that $\mathbf{Z}^c \sim \mathcal{N}(\mu^c,\Sigma^c) \in \mathbb{R}^D$, we therefore know that the conditional distribution of any subset vector Z_k^c, given the complement vector Z_j^c, is also a multivariate Gaussian distribution [22]

$$Z_k^c|Z_j^c \sim \mathcal{N}(\mu_{k|j}^c, \Sigma_{k|j}^c) \tag{22}$$

where

$$\mu_{k|j}^c = \mu_k^c + \Sigma_{k,j}^c(\Sigma_{j,j}^c)^{-1}(Z_j^c - \mu_j^c), \quad \Sigma_{k|j}^c = \Sigma_{k,k}^c - \Sigma_{k,j}^c(\Sigma_{j,j}^c)^{-1}\Sigma_{j,k}^c, \tag{23}$$

given that $\Sigma_{j,j}^c$ is nonsingular.

Hence we know that the first expectation in Equation (21) becomes

$$E[Z_i^c|Z_{-i}^c] = \mu_i^c + \Sigma_{i,-i}^c(\Sigma_{-i,-i}^c)^{-1}(Z_{-i}^c - \mu_{-i}^c) \tag{24}$$

assuming that $\Sigma_{-i,-i}^c$ is nonsingular. Since $\mathbb{E}[Z_i^c] = \mu_i^c$, the loss l_c can be written as

$$l_c = \sum_{i}^{D} d(\mu_i^c + \Sigma_{i,-i}^c(\Sigma_{-i,-i}^c)^{-1}(Z_{-i}^c - \mu_{-i}^c), \mu_i^c). \tag{25}$$

We assume further that Σ^c is a diagonal matrix. Therefore $\Sigma_{-i,-i}^c = \mathbf{0}$ is a zero row vector. Then

$$l_c = \sum_{i}^{D} d(\mu_i^c, \mu_i^c) = 0 \tag{26}$$

□

C Related Work

Disentangled Representations The pursuit for disentangled representation can be dated to the surge of representation learning and is always closely associated with the generative process in modern machine learning, following the intuition that each dimension should encode different features. [6] attempts to control the underlying factors by maximizing the mutual information between the images and the latent representations. [8] propose a quantitative metric with the information theory. They evaluate the disentanglement, completeness, and informativeness by fitting linear models and measuring the deviation from the ideal mapping. [9, 11, 5, 18] encourage statistical independence by penalizing the Kullback-Leibler divergence (KL) term in the VAE objective. However, the non-causal definitions of disentanglement fail to consider the cases where correlated features in the observational dataset can be disentangled in the generative process. Such a challenge is well-approached through a line of research from the causal perspective.

Causal Generative Process. Causal methods are widely used for eliminating spurious features in various domains and improving understandable modelling behaviours[25, 26, 14]. It is not until [23] that it was introduced for a strict characterization of the generative process. [23] first provided a rigorous definition of a causal generative process and the definition of disentangled causal representation as the non-existence of causal relationships between two variables, i.e., the intervention on one variable does not alter the distribution of the others. The authors further introduce *interventional robustness* as an evaluation metric and show its advantage on multiple benchmarks. [21] follow the path of [23] and further propose two evaluation metrics and the Candle dataset. The confounded assumption allows for correlation in the latent space without tempering with the disentanglement in the data generative. Despite effective evaluation tools, there is still a missing piece on how to infer a set of causally disentangled features. Using the proposed evaluation metric as regulation, the model implicitly assumes unconfoundedness and it falls back to finding statistical independence in the latent space. The problem of unrealistic unconfoundedness assumption is identified by [24]. They assume that confounders exist but they are unobservable. They further propose an evaluation metric considering the existence of confounders, that causally disentangled latent variables have independent support measured by the IOSS score. Similar to the evaluation metrics introduced in [23, 21], IOSS is also a necessary condition of the causal disentanglement. More importantly, as in previous work focusing on obtaining statistical independence, such a regulation suffers from the identifiability issue.

Weak Supervision for Inductive Bias. The identifiability issue in unsupervised disentangled representation learning is first identified in [16]. Specifically, they show from the theory that such a learning task is impossible without inductive biases on both the models and the data. Naturally, a series of weak-supervised or semi-supervised methods [4, 1, 2] are proposed with a learning objective of statistical independence or alignment. In this paper, we take a step further for the confounding assumption, assuming that the confounders are observable with proper inductive bias so that the latent representation can be better identified. We, similarly, adopt partial labels of the dataset as the supervision signal. We treat the labels as a source of possible confounders and allow the learning of correlated but causally disentangled latent generative factors to be learned.

D Experimental Details

D.1 Experimental Details

The experiments are conducted on 4 NVIDIA GeForce RTX 2080Ti. In each experiment, we repeat 5 times with different seeds and report the averaged results. In all experiments, only partial information on the ground truth confounder is provided. Specifically, for example, the 3dshape dataset, we first make some predefined rules, such as " 70% cubes are red". Then we generate 700 red cubes and 300 cubes in other colors. The generation process naturally divides the dataset into different subgroups, and we can thus explicitly control how inductive bias is provided, i.e., the grouping. In the celebA dataset, since we do not have access to the ground truth generative factors, so we assume any label sets only contain partial information.

D.2 Ablation study

We investigate how the choice of $\mathbf{C}$ affect the model performance and how to adapt C-Disentanglement to the existing method aiming for statistical independence, as shown in appendix D.

Table 3: Performance under different $\mathbf{C}$ on shape classification on 3dshape dataset, shift severity=0.5.

Choice of $\mathbf{C}$	Acc - T $\uparrow$	IRS $\uparrow$
$\mathbf{C} = \emptyset$	79.2	0.82
$\mathbf{C} = \mathbf{C}^*$	88.2	0.89
partial $\mathbf{C}^*$	84.5	0.87

Table 4: Adapt cdVAE to existing methods. The IOSS and Reconstruction loss are measured based on image generation task and the performance drop is measured on shape classification on the target domain under shift severity=0.5.

Methods	IOSS $\downarrow$	Recon $\downarrow$	Acc-T $\uparrow$
IOSS	0.14	0.12	81.8
cdVAE + IOSS	0.12	0.08	84.5

C-Disentanglement improves the learning of ground truth generative factors under a reasonable choice of label set. To understand how the choice of $\mathbf{C}$ affects the performance, we repeat the shape classification task with different choices of $\mathbf{C}$ under 50% shift severity. When with $\mathbf{C}$, we assume the generative process is unconfounded, and cdVAE degrades to the vanilla VAE model. With partial $\mathbf{C}$, we partition the data according to only 2 values of the shifting variables instead of 4.

With full $\mathbf{C}$, we provide the full confounders. As shown in Table 3, even partial information of the confounders improves the model performance in OOD generalization and obtains more robust latent representations.

Adapting C-Disentanglement to existing works further improve their performance. We compare the performance between regulation through IOSS[24] and cdVAE + IOSS in image generation and classification tasks on 3dshape dataset. In the cdVAE + IOSS, we apply additional regularization terms based on the $\mathbf{Z}$. The results show that C-Disentanglement framework could further improve the performance with desired level of inductive bias given.

D.3 Pseudo-code

Algorithm 1 Train a VAE such that the latent representation is causally disentangled

Input: Number of labels N_C, training data $\mathbf{X}$ with labels c, ratio of each categories/confounders $P(\mathbf{C} = c)$ in the training set, dimension of latent space D

1: **for** $x \in \mathbf{X}$ **do**

2: **for** $c \in \mathbf{C}$ **do**

3: Define $\mathbf{Z}^c = [\mathbf{Z}|x, C = c]$, and obtain from encoder $\mathbf{Z}^c \sim \mathcal{N}(\mu^c(x), \Sigma^c(x))$ for each c, assuming $\Sigma^c(x)$ to be diagonal matrix:

$$\Theta^c_{enc}(x) = [\mu^c, \mathtt{diag}(\Sigma^c)] \in \mathbb{R}^{2d}, \quad \mu^c \in \mathbb{R}^d, \quad \mathtt{diag}(\Sigma^c) \in \mathbb{R}^d \tag{27}$$

4: Sample from $\mathbf{Z}^c \sim \mathcal{N}(\mu^c(x), \Sigma^c(x))$:

$$\mathbf{Z}^c = \mu^c + (\Sigma^c)^{\frac{1}{2}} \epsilon^c, \epsilon^c \sim \mathcal{N}(\mathbf{0}, I) \tag{28}$$

5: Parametrize $\pi^c \sim \mathcal{N}(\mu_{\pi^c}(x), \sigma_{\pi^c}(x)) \in \mathbb{R}$ with neural network.

6: Regulate the covariance matrix to be identity matrix with KL divergence

$$D^c_{KL} = \frac{1}{2}\left[\log \frac{1}{\det \Sigma^c} - D + \mathrm{tr}(\Sigma^c)\right] \tag{29}$$

7: **end for**

8: Normalize $\Pi_C = (\pi^{c_1}, ..., \pi^{c_{N_C}})$ such that $\|\Pi_C\|^2 = 1$.

9: Compute classification loss between Π_C with label c:

$$\mathcal{L}_{cls} = H(\Pi_C, c) \tag{30}$$

10: Let $\mathbf{Z}(x) = \sum_{c \in \mathbf{C}} \pi^c \mathbf{Z}^c(x)$, and obtain the reconstructed sample from decoder: $x' = \Theta_{dec}(\mathbf{Z}(x))$. Compute reconstruction loss for $\mathbf{Z}(x)$:

$$\mathcal{L}_{rec} = \mathtt{mse}(x', x) \tag{31}$$

11: Compute total loss

$$\mathcal{L}_{total}(x) = \mathcal{L}_{rec} + L_{cls} + \sum_{c \in \mathbf{C}} D^c_{KL} \tag{32}$$

and update encoders and decoders.

12: **end for**

D.4 Additional Experimental Results

The classification accuracy on both the source and the target distribution with variance is given in the table below.

Table 5: **Compare cdVAE with β-Vae, CAUSAL-REP on classification under distribution shift. T represents accuracy on the target data, S represents the performance on the target domain when the classifier trained on the source data is directly tested on the target data.**

	shift = 0.4		**shift = 0.5**		**severity = 0.6**	
Methods	**Acc-S**	**Acc-T**	**Acc-S**	**Acc-T**	**Acc-S**	**Acc-T**
CAUSAL-REP	94.1±0.04	82.1±0.08	94.3±0.03	81.9±0.07	94.3±0.02	81.8±0.11
β-VAE	93.4±0.07	80.7±0.12	93.6±0.05	80.7±0.03	93.4±0.04	80.3±0.09
cdVAE	94.6±0.02	84.5±0.05	94.6±0.04	84.4±0.05	94.5±0.03	84.4±0.04