A Proof for Theorem 3.1

Proof. To estimate $p\big(y|\boldsymbol{x}, do(\boldsymbol{z}_s)\big)$, we introduce variables $\boldsymbol{Z}_c$:

$$p\big(y|\boldsymbol{x}, do(\boldsymbol{z}_s)\big) = \int_{\boldsymbol{z}_c} p\big(y|\boldsymbol{z}_c, \boldsymbol{x}, do(\boldsymbol{z}_s)\big) p\big(\boldsymbol{z}_c|\boldsymbol{x}, do(\boldsymbol{z}_s)\big)\, \mathrm{d}\boldsymbol{z}_c \tag{7}$$

We can simplify the calculation of $p(y|\boldsymbol{z}_c, \boldsymbol{x}, do(\boldsymbol{z}_s))$ using Pearl's Do-Calculus Rules.

$$\begin{aligned} p\big(y|\boldsymbol{z}_c, \boldsymbol{x}, do(\boldsymbol{z}_s)\big) =& p\big(y|\boldsymbol{z}_c, do(\boldsymbol{z}_s)\big) \quad (Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{Z}_c, \boldsymbol{Z}_s)_{\mathcal{G}_{\overline{\boldsymbol{Z}_s}}} \quad \text{According to Rule 1} \\ =& p(y|\boldsymbol{z}_c) \quad (Y \perp\!\!\!\perp \boldsymbol{Z}_s|\boldsymbol{Z}_c)_{\mathcal{G}_{\overline{\boldsymbol{Z}_s}}} \quad \text{According to Rule 3} \end{aligned} \tag{8}$$

We employ the Backdoor Adjustment Theorem and Pearl's Do-Calculus Rule 2 to estimate $p\big(\boldsymbol{z}_c|\boldsymbol{x}, do(\boldsymbol{z}_s)\big)$,

$$\begin{aligned} p(\boldsymbol{z}_c|\boldsymbol{x}, do(\boldsymbol{z}_s)) =& \frac{p\big(\boldsymbol{z}_c|do(\boldsymbol{z}_s)\big) p\big(\boldsymbol{x}|\boldsymbol{z}_s, do(\boldsymbol{z}_s)\big)}{p\big(\boldsymbol{x}|do(\boldsymbol{z}_s)\big)} \quad \text{Bayes's Theorem} \\ =& \frac{p\big(\boldsymbol{z}_s|do(\boldsymbol{z}_s)\big) p\big(\boldsymbol{x}|\boldsymbol{z}_c, \boldsymbol{z}_s\big)}{p\big(\boldsymbol{x}|do(\boldsymbol{z}_s)\big)} \quad (\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Z}_s|\boldsymbol{Z}_c)_{\mathcal{G}_{\underline{\boldsymbol{Z}_s}}} \quad \text{According to Rule 2} \end{aligned} \tag{9}$$

Between $\boldsymbol{Z}_s, \boldsymbol{Z}_c$, there is a valid backdoor path from $\boldsymbol{Z}_c \leftarrow Y \leftarrow \boldsymbol{U}_{xy} \rightarrow \boldsymbol{Z}_s$, we can directly apply the Backdoor Adjustment Theorem with a valid adjusting set $\{Y\}$:

$$\begin{aligned} p\big(\boldsymbol{z}_c|do(\boldsymbol{z}_s)\big) &= \sum_y p(\boldsymbol{z}_c|y, \boldsymbol{z}_s)p(y) \\ &= \sum_y p(\boldsymbol{z}_c|y)p(y) \quad (\boldsymbol{Z}_c \perp\!\!\!\perp \boldsymbol{Z}_s|Y)_{\mathcal{G}} \\ &= p(\boldsymbol{z}_c) \end{aligned} \tag{10}$$

Between $\boldsymbol{Z}_s, \boldsymbol{X}$, there is a valid backdoor path from $\boldsymbol{X} \leftarrow \boldsymbol{Z}_c \leftarrow Y \leftarrow \boldsymbol{U}_{xy} \rightarrow \boldsymbol{Z}_s$. We are able to adjust on $\{Y\}$, $\{\boldsymbol{Z}_c\}$ or $\{Y, \boldsymbol{Z}_c\}$[3]. In our case, we choose to adjust on $\{\boldsymbol{Z}_c\}$:

$$p\big(\boldsymbol{x}|do(\boldsymbol{z}_s)\big) \overset{\text{Adjust on } \boldsymbol{Z}_c}{=} \int_{\boldsymbol{z}_c} p(\boldsymbol{x}|\boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_c)\, \mathrm{d}\boldsymbol{z}_c \tag{11}$$

Substitute Eq. equation 10, equation 11 into Eq. equation 9, we obtain:

$$p\big(\boldsymbol{z}_c|\boldsymbol{x}, do(\boldsymbol{z}_s)\big) = \frac{p(\boldsymbol{x}|\boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_c)}{\int_{\boldsymbol{z}_c} p(\boldsymbol{x}|\boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_c)\, \mathrm{d}\boldsymbol{z}_c} \tag{12}$$

Substitute Eq. equation 12 and equation 8 into Eq. equation 7, we obtain Eq. equation 1 in **Theorem 3.1**:

$$\begin{aligned} p\big(y|\boldsymbol{x}, do(\boldsymbol{z}_s)\big) &= \int_{\boldsymbol{z}_c} p\big(y|\boldsymbol{z}_c, \boldsymbol{x}, do(\boldsymbol{z}_s)\big) p\big(\boldsymbol{z}_c|\boldsymbol{x}, do(\boldsymbol{z}_s)\big)\, \mathrm{d}\boldsymbol{z}_c \\ &= \int_{\boldsymbol{z}_c} p(y|\boldsymbol{z}_c) \frac{p(\boldsymbol{x}|\boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_c)}{\int_{\boldsymbol{z}_c} p(\boldsymbol{x}|\boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_c)\, \mathrm{d}\boldsymbol{z}_c}\, \mathrm{d}\boldsymbol{z}_c \end{aligned} \tag{13}$$

□

[3]The equations for conditioning on those three different adjusting sets are the same.

B DERIVATIONS

B.1 THE DERIVATION IN EQ. (2)

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{z}_s|\boldsymbol{x})}[p(y|\boldsymbol{x}, do(\boldsymbol{z}_s))] =&\mathbb{E}_{p(\boldsymbol{z}_s|\boldsymbol{x})}[\frac{\int_{\boldsymbol{z}_c} p(y|\boldsymbol{z}_c)p(\boldsymbol{x}|\boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_c)\,\mathrm{d}\boldsymbol{z}_c}{\int_{\boldsymbol{z}_c} p(\boldsymbol{x}|\boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_c)\,\mathrm{d}\boldsymbol{z}_c}] \\ \approx&\frac{1}{N}\sum_{n=1}^{N}[\sum_{l=1}^{L} p(y|\boldsymbol{z}_{c,l})\frac{p(\boldsymbol{x}|\boldsymbol{z}_{c,l}, \boldsymbol{z}_{s,n})}{\sum_{l'=1}^{L} p(\boldsymbol{x}|\boldsymbol{z}_{c,l'}, \boldsymbol{z}_{s,n})}], \boldsymbol{z}_{c,l} \sim p(\boldsymbol{z}_c), \boldsymbol{z}_{s,n} \sim p(\boldsymbol{z}_s|\boldsymbol{x}) \\ =&\frac{1}{N}\sum_{n=1}^{N}[\sum_{l=1}^{L} p(y|\boldsymbol{z}_{c,l})\boldsymbol{\omega}(\boldsymbol{z}_{c,l}, \boldsymbol{z}_{s,n})], \text{where } \boldsymbol{\omega}(\boldsymbol{z}_{c,l}, \boldsymbol{z}_{s,n}) = \frac{p(\boldsymbol{x}|\boldsymbol{z}_{c,l}, \boldsymbol{z}_{s,n})}{\sum_{l'=1}^{L} p(\boldsymbol{x}|\boldsymbol{z}_{c,l'}, \boldsymbol{z}_{s,n})}\end{aligned} \tag{14}$$

B.2 THE DERIVATION OF THE MARGINAL LIKELIHOOD IN EQ. (4)

We character the joint likelihood over the variables in the proposed SCM. However, there are 6 variables of interest and only the observations for 2 of them are available. Therefore, we start from the marginal likelihood $p(\boldsymbol{x}, y)$.

$$\begin{aligned}&\log p(\boldsymbol{x}, y) \\ =&\log \int_{\boldsymbol{z}_c}\int_{\boldsymbol{z}_s}\sum_{u_x}\sum_{u_{xy}} p(u_x, u_{xy}, \boldsymbol{z}_s, \boldsymbol{z}_c, \boldsymbol{x}, y)\,\mathrm{d}\boldsymbol{z}_c\,\mathrm{d}\boldsymbol{z}_s \\ =&\log \int_{\boldsymbol{z}_c}\int_{\boldsymbol{z}_s}\sum_{u_x}\sum_{u_{xy}} p(u_x)p(u_{xy})p(\boldsymbol{z}_s|u_x, u_{xy})p(y|u_{xy})p(\boldsymbol{z}_c|y)p(\boldsymbol{x}|\boldsymbol{z}_s, \boldsymbol{z}_c)\,\mathrm{d}\boldsymbol{z}_c\,\mathrm{d}\boldsymbol{z}_s \quad \text{Bayesian Network Chain Rule} \\ =&\log \frac{1}{p(y)}\int_{\boldsymbol{z}_c}\int_{\boldsymbol{z}_s}\sum_{u_x}\sum_{u_{xy}} p(u_x)p(u_{xy})p(\boldsymbol{z}_s|u_x, u_{xy})p(y|u_{xy})p(y|\boldsymbol{z}_c)p(\boldsymbol{z}_c)p(\boldsymbol{x}|\boldsymbol{z}_s, \boldsymbol{z}_c)\,\mathrm{d}\boldsymbol{z}_c\,\mathrm{d}\boldsymbol{z}_s \quad \text{Bayes Theorem} \\ =&\log \frac{1}{p(y)} + \log \int_{\boldsymbol{z}_c}\int_{\boldsymbol{z}_s}\sum_{u_x}\sum_{u_{xy}} p(u_x)p(u_{xy})p(\boldsymbol{z}_s|u_x, u_{xy})p(y|u_{xy})p(y|\boldsymbol{z}_c)p(\boldsymbol{z}_c)p(\boldsymbol{x}|\boldsymbol{z}_s, \boldsymbol{z}_c)\,\mathrm{d}\boldsymbol{z}_c\,\mathrm{d}\boldsymbol{z}_s \\ \geq&\log \int_{\boldsymbol{z}_c}\int_{\boldsymbol{z}_s}\big[\sum_{u_x}\sum_{u_{xy}} p(u_x)p(u_{xy})p(\boldsymbol{z}_s|u_x, u_{xy})p(y|u_{xy})\big]p(y|\boldsymbol{z}_c)p(\boldsymbol{z}_c)p(\boldsymbol{x}|\boldsymbol{z}_s, \boldsymbol{z}_c)\,\mathrm{d}\boldsymbol{z}_c\,\mathrm{d}\boldsymbol{z}_s \\ =&\log \int_{\boldsymbol{z}_c}\int_{\boldsymbol{z}_s}\big[\sum_{u_x}\sum_{u_{xy}} p(u_x|u_{xy})p(u_{xy})p(\boldsymbol{z}_s|u_x, u_{xy})p(y|u_{xy})\big]p(y|\boldsymbol{z}_c)p(\boldsymbol{z}_c)p(\boldsymbol{x}|\boldsymbol{z}_s, \boldsymbol{z}_c)\,\mathrm{d}\boldsymbol{z}_c\,\mathrm{d}\boldsymbol{z}_s \quad U_x \perp\!\!\!\perp U_{xy} \\ =&\log \int_{\boldsymbol{z}_c}\int_{\boldsymbol{z}_s}\frac{\big[\sum_{u_{xy}} p(u_{xy})p(\boldsymbol{z}_s|u_{xy})p(y|u_{xy})\big]p(y|\boldsymbol{z}_c)p(\boldsymbol{z}_c)p(\boldsymbol{x}|\boldsymbol{z}_s, \boldsymbol{z}_c)}{q(\boldsymbol{z}_s, \boldsymbol{z}_c|\boldsymbol{x})}q(\boldsymbol{z}_s, \boldsymbol{z}_c|\boldsymbol{x})\,\mathrm{d}\boldsymbol{z}_c\,\mathrm{d}\boldsymbol{z}_s \\ =&\log \mathbb{E}_{q(\boldsymbol{z}_s, \boldsymbol{z}_c|\boldsymbol{x})}\Big[\frac{\big[\sum_{u_{xy}} p(u_{xy})p(\boldsymbol{z}_s|u_{xy})p(y|u_{xy})\big]p(y|\boldsymbol{z}_c)p(\boldsymbol{z}_c)p(\boldsymbol{x}|\boldsymbol{z}_s, \boldsymbol{z}_c)}{q(\boldsymbol{z}_s, \boldsymbol{z}_c|\boldsymbol{x})}\Big] \\ =&\log \mathbb{E}_{q(\boldsymbol{z}_s, \boldsymbol{z}_c|\boldsymbol{x})}\Big[\frac{\big[\sum_{u_{xy}} p(u_{xy})p(\boldsymbol{z}_s|u_{xy})p(y|u_{xy})\big]p(y|\boldsymbol{z}_c)p(\boldsymbol{z}_c)p(\boldsymbol{x}|\boldsymbol{z}_s, \boldsymbol{z}_c)}{q(\boldsymbol{z}_s|\boldsymbol{x})q(\boldsymbol{z}_c|\boldsymbol{x})}\Big] \quad \text{Assume } q(\boldsymbol{z}_s, \boldsymbol{z}_c|\boldsymbol{x}) = q(\boldsymbol{z}_s|\boldsymbol{x})q(\boldsymbol{z}_c|\boldsymbol{x}) \\ \geq&\mathbb{E}_{q(\boldsymbol{z}_s, \boldsymbol{z}_c|\boldsymbol{x})}\log\Big[\frac{\big[\sum_{u_{xy}} p(u_{xy})p(\boldsymbol{z}_s|u_{xy})p(y|u_{xy})\big]p(\boldsymbol{z}_c)}{q(\boldsymbol{z}_s|\boldsymbol{x})q(\boldsymbol{z}_c|\boldsymbol{x})}p(y|\boldsymbol{z}_c)p(\boldsymbol{x}|\boldsymbol{z}_s, \boldsymbol{z}_c)\Big] \quad \text{Jensen's inequality} \\ =&\mathbb{E}_{q(\boldsymbol{z}_s, \boldsymbol{z}_c|\boldsymbol{x})}\Big[\log\frac{\big[\sum_{u_{xy}} p(u_{xy})p(\boldsymbol{z}_s|u_{xy})p(y|u_{xy})\big]}{q(\boldsymbol{z}_s|\boldsymbol{x})} + \log\frac{p(\boldsymbol{z}_c)}{q(\boldsymbol{z}_c|\boldsymbol{x})} + \log p(y|\boldsymbol{z}_c) + \log p(\boldsymbol{x}|\boldsymbol{z}_s, \boldsymbol{z}_c)\Big] \\ =&\mathbb{E}_{q(\boldsymbol{z}_s|\boldsymbol{x})}\log\frac{\sum_{u_{xy}} p(u_{xy})p(\boldsymbol{z}_s|u_{xy})p(y|u_{xy})}{q(\boldsymbol{z}_s|\boldsymbol{x})} + \mathbb{E}_{q(\boldsymbol{z}_c|\boldsymbol{x})}\log\frac{p(\boldsymbol{z}_c)}{q(\boldsymbol{z}_c|x)} + \mathbb{E}_{q(\boldsymbol{z}_c|\boldsymbol{x})}\log p(y|\boldsymbol{z}_c) + \mathbb{E}_{q(\boldsymbol{z}_c, \boldsymbol{z}_s|\boldsymbol{x})}\log p(\boldsymbol{x}|\boldsymbol{z}_c, \boldsymbol{z}_s) \\ =&\mathbb{E}_{q(\boldsymbol{z}_s|\boldsymbol{x})}\log\frac{\sum_{u_{xy}} p(u_{xy})p(\boldsymbol{z}_s|u_{xy})p(y|u_{xy})}{q(\boldsymbol{z}_s|\boldsymbol{x})} - KL\big(q(\boldsymbol{z}_c|\boldsymbol{x})||p(\boldsymbol{z}_c)\big) + \mathbb{E}_{q(\boldsymbol{z}_c|\boldsymbol{x})}\log p(y|\boldsymbol{z}_c) + \mathbb{E}_{q(\boldsymbol{z}_c, \boldsymbol{z}_s|\boldsymbol{x})}\log p(\boldsymbol{x}|\boldsymbol{z}_c, \boldsymbol{z}_s)\end{aligned} \tag{15}$$

We further simplify the first term as follows,

$$
\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{z}_s|\boldsymbol{x})} \log \frac{\left[\sum_{u_{xy}} p(u_{xy})p(\boldsymbol{z}_s|u_{xy})p(y|u_{xy})\right]}{q(\boldsymbol{z}_s|\boldsymbol{x})} \\
=&\mathbb{E}_{q(\boldsymbol{z}_s|\boldsymbol{x})} \log \Big[\sum_{u_{xy}} \frac{p(\boldsymbol{z}_s|u_{xy})p(y|u_{xy})}{q(\boldsymbol{z}_s|\boldsymbol{x})} p(u_{xy})\Big] \\
=&\mathbb{E}_{q(\boldsymbol{z}_s|\boldsymbol{x})} \log \mathbb{E}_{p(u_{xy})} \frac{p(\boldsymbol{z}_s|u_{xy})p(y|u_{xy})}{q(\boldsymbol{z}_s|\boldsymbol{x})} \\
\geq&\mathbb{E}_{q(\boldsymbol{z}_s|\boldsymbol{x})}\mathbb{E}_{p(u_{xy})} \log \Big[\frac{p(\boldsymbol{z}_s|u_{xy})p(y|u_{xy})}{q(\boldsymbol{z}_s|\boldsymbol{x})}\Big] \quad \text{Jensen's inequality} \\
=&\mathbb{E}_{q(\boldsymbol{z}_s|\boldsymbol{x})}\mathbb{E}_{p(u_{xy})} \Big[\log \frac{p(\boldsymbol{z}_s|u_{xy})}{q(\boldsymbol{z}_s|\boldsymbol{x})} + \log p(y|u_{xy})\Big] \\
=&\mathbb{E}_{p(u_{xy})}\mathbb{E}_{q(\boldsymbol{z}_s|\boldsymbol{x})} \Big[\log \frac{p(\boldsymbol{z}_s|u_{xy})}{q(\boldsymbol{z}_s|\boldsymbol{x})}\Big] + \mathbb{E}_{p(u_{xy})} \log p(y|u_{xy}) \\
=&\mathbb{E}_{p(u_{xy})}\mathbb{E}_{q(\boldsymbol{z}_s|\boldsymbol{x})} \Big[\log \frac{\sum_{u_x} p(\boldsymbol{z}_s|u_x, u_{xy})p(u_x|u_{xy})}{q(\boldsymbol{z}_s|\boldsymbol{x})}\Big] + \mathbb{E}_{p(u_{xy})} \log p(y|u_{xy}) \text{ Re-introduce } U_x \\
=&\mathbb{E}_{p(u_{xy})}\mathbb{E}_{q(\boldsymbol{z}_s|\boldsymbol{x})} \Big[\log \sum_{u_x} \frac{p(\boldsymbol{z}_s|u_x, u_{xy})}{q(\boldsymbol{z}_s|\boldsymbol{x})} p(u_x)\Big] + \mathbb{E}_{p(u_{xy})} \log p(y|u_{xy}) \quad U_x \perp\!\!\!\perp U_{xy} \\
=&\mathbb{E}_{p(u_{xy})}\mathbb{E}_{q(\boldsymbol{z}_s|\boldsymbol{x})} \Big[\log \mathbb{E}_{p(u_x)} \frac{p(\boldsymbol{z}_s|u_x, u_{xy})}{q(\boldsymbol{z}_s|\boldsymbol{x})}\Big] + \mathbb{E}_{p(u_{xy})} \log p(y|u_{xy}) \\
\geq&\mathbb{E}_{p(u_{xy})}\mathbb{E}_{p(u_x)}\mathbb{E}_{q(\boldsymbol{z}_s|\boldsymbol{x})} \Big[\log \frac{p(\boldsymbol{z}_s|u_x, u_{xy})}{q(\boldsymbol{z}_s|\boldsymbol{x})}\Big] + \mathbb{E}_{p(u_{xy})} \log p(y|u_{xy}) \quad \text{Jensen's inequality} \\
=&-\mathbb{E}_{p(u_x)}\mathbb{E}_{p(u_{xy})} KL\big(q(\boldsymbol{z}_s|\boldsymbol{x})||p(\boldsymbol{z}_s|u_x, u_{xy})\big) + \mathbb{E}_{p(u_{xy})} \log p(y|u_{xy}) \\
=&-\mathbb{E}_{p(u_x, u_{xy})} KL\big(q(\boldsymbol{z}_s|\boldsymbol{x})||p(\boldsymbol{z}_s|u_x, u_{xy})\big) + \mathbb{E}_{p(u_{xy})} \log p(y|u_{xy})
\end{aligned}
\tag{16}
$$

We parameterize the encoder distributions using parameter $\Theta = \{\Theta_s, \Theta_c\}$, denoted as $q_{\Theta_s}(\boldsymbol{z}_s|\boldsymbol{x})$ and $q_{\Theta_c}(\boldsymbol{z}_c|\boldsymbol{x})$, the decoder distribution with parameter Φ as $p_\Phi(\boldsymbol{x}|\boldsymbol{z}_c, \boldsymbol{z}_s)$, and the classifier distribution with parameter Ψ as $p_\Psi(y|\boldsymbol{z}_c)$. During training, we optimize these defined parameters to construct the corresponding distributions. Additionally, we make assumptions or estimations about the prior distributions, specifically $p(\boldsymbol{z}_c)$ and $p(\boldsymbol{z}_s|u_x, u_{xy})$, to help regularize the learning of representations. Notably, we do not parameterize over $p(y|u_{xy})$, as it is not necessary for either obtaining representations or computing interventional distributions. As a result, we omit the term $\mathbb{E}_{p(u_{xy})} \log p(y|u_{xy})$, as it is independent of the parameters for optimization. By combining Eq.(15) with Eq.(16), we propose the following training objective for the causal representation learning procedure:

$$
\begin{aligned}
\mathcal{L}_{obj}(\boldsymbol{x}, y, \Theta, \Phi, \Psi) = &-\mathbb{E}_{p(u_x, u_{xy})} KL\big(q_{\Theta_s}(\boldsymbol{z}_s|\boldsymbol{x})||p(\boldsymbol{z}_s|u_x, u_{xy})\big) - KL(q_{\Theta_c}(\boldsymbol{z}_c|\boldsymbol{x})||p(\boldsymbol{z}_c)) \\
&+ \mathbb{E}_{q_{\Theta_c}(\boldsymbol{z}_c|\boldsymbol{x})} \log p_\Psi(y|\boldsymbol{z}_c) + \mathbb{E}_{q_\Theta(\boldsymbol{z}_c, \boldsymbol{z}_s|\boldsymbol{x})} \log p_\Phi(\boldsymbol{x}|\boldsymbol{z}_c, \boldsymbol{z}_s)
\end{aligned}
\tag{17}
$$

C Empirical Ablation Study

In this section, we perform three ablation studies: 1) We show the sensitivity of our proposed CRLII method with respect to the value of $|U|$ that we specify during SCM parameterization and learning. 2) We demonstrate the necessity of our intervention inference approach due to the imperfect disentanglement between $\boldsymbol{z}_s$ and $\boldsymbol{z}_c$. 3) We provide a detailed approach to choose the number of $\boldsymbol{z}_{c,l}$ that we need to obtain to perform interventional inference.

C.1 THE NUMBER OF DOMAINS

In this section, we explore how varying the number of domains $|U|$ affects out-of-distribution (OOD) prediction performance. We incrementally increase the values of $|U|$ from 1 to 5 and present the corresponding prediction performance in Figure 5.

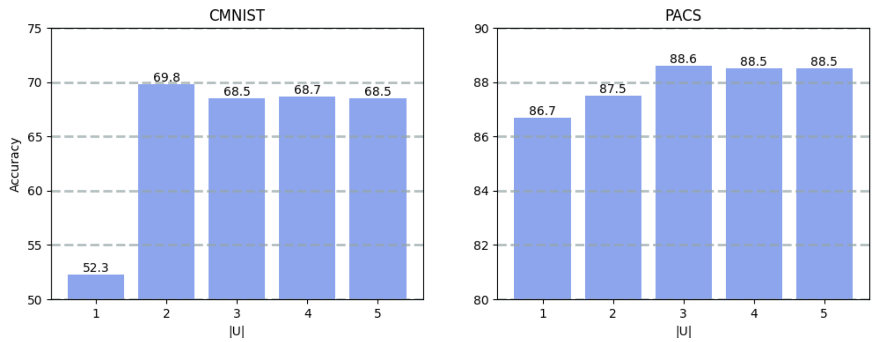

Figure 5: Ablation study on the influence from the number of domains $|U|$. The results on PACS dataset are averaged over four domains.

The findings from Figure 5 reveal that our method exhibits its poorest performance when the number of domains, denoted as $|U|$, is set to 1. In such a scenario, the assumption is made that $\boldsymbol{z}_s$ follows a Gaussian distribution, and its prior distribution mirrors that of $\boldsymbol{z}_c$. This results in a compromise in the asymmetry regularization between the two types of representations, leading to suboptimal disentanglement. The restoration of this asymmetry occurs when we set $|U| \geq 2$. However, performance results appear comparable for cases where $|U|$ exceeds 2. The optimal selection of $|U|$ hinges on the dissimilarities between the true data distributions across each domain. In situations involving observational datasets lacking domain-specific information, setting $|U| = 2$ can still yield a reasonably well-disentangled set of representations.

C.2 INFLUENCE OF INTERVENTIONAL INFERENCE

In Section 4.1, we provide theoretical justification for our choice of interventional inference, driven by the partial disentanglement observed between $\boldsymbol{z}_c$ and $\boldsymbol{z}_s$ during the SCM learning process. In Table 4, our empirical results demonstrate that the $\boldsymbol{z}_c$ representation we obtain still retains information from $\boldsymbol{z}_s$. Consequently, interventional inference proves effective in further enhancing OOD performance when compared to direct prediction using $p(y|\boldsymbol{z}_c^t)$, where $\boldsymbol{z}_c^t = \arg\max_{\boldsymbol{z}_c} q(\boldsymbol{z}_c|\boldsymbol{x}^t)$.

Table 4: Comparison between prediction and interventional inference.

Datasets	Accuracy (%)	
	Prediction with $\boldsymbol{z}_c$	Interventional Inference
CMNIST	52.6	**69.8**
PACS	86.7	**88.6**
VLCS	76.3	**79.1**
OfficeHome	67.7	**69.5**

C.3 THE SELECTION OF L

For the purpose of inference, we generated a set of $\boldsymbol{z}_c$ samples from the training inputs and computed their weighted sum for $p(y|\boldsymbol{z}_c)$. However, this process proved to be time-consuming and inefficient, especially when dealing with a large number of training inputs. We observed significant variation in the magnitudes of weights assigned to different samples of $\boldsymbol{z}_c$. Let's denote the obtained samples as $\boldsymbol{z}_{c,1}, \boldsymbol{z}_{c,2}, \cdots, \boldsymbol{z}_{c,l}$, with $\boldsymbol{\omega}(\boldsymbol{z}_{c,1}, \boldsymbol{z}_{s,n}) \geq \boldsymbol{\omega}(\boldsymbol{z}_{c,2}, \boldsymbol{z}_{s,n}) \geq \cdots \geq \boldsymbol{\omega}(\boldsymbol{z}_{c,L}, \boldsymbol{z}_{s,n})$. Notably, when $L > 5$, the ratio $\frac{\boldsymbol{\omega}(\boldsymbol{z}_{c,1}, \boldsymbol{z}_{s,n})}{\boldsymbol{\omega}(\boldsymbol{z}_{c,L}, \boldsymbol{z}_{s,n})}$ exceeds 10, and when $L > 10$, it surpasses 100. In cases where $\boldsymbol{\omega}(\boldsymbol{z}_{c,1}, \boldsymbol{z}_{s,n}) > 100\boldsymbol{\omega}(\boldsymbol{z}_{c,L}, \boldsymbol{z}_{s,n})$, the contribution of $\boldsymbol{\omega}(\boldsymbol{z}_{c,L}, \boldsymbol{z}_{s,n})$ to the weighted sum becomes negligible. Consequently, it becomes unnecessary to consider values of L greater than 10.