A. Proofs

A.1. Proofs of results in section 3 *framework*

A.1.1. Generalized MIB objective

We generalized the original MIB structural variational learning objective in equation 8. We show that by choosing $\mathbb{C}_1 = D_{\mathrm{KL}}(q_{\boldsymbol{\phi}} \parallel p_{\boldsymbol{\theta}}),\ T = 1$ and $\mathcal{G}^1 = \mathcal{G}^{\emptyset},\ K = 1$, we can recover the original MIB objective equation 5.

Proposition 1. *Let* $\mathrm{X} \sim P(\mathrm{X})$*, and let* $\mathcal{G}^{\emptyset}$ *be an empty Bayesian network over* X*. Then*

$$\mathbb{D}(p \parallel \mathcal{G}^{\emptyset}) = \min_{q \models \mathcal{G}} D_{\mathrm{KL}}(p \parallel q) = \mathcal{I}_p(\mathrm{X}) - \mathcal{I}_p^{\mathcal{G}^{\emptyset}}(\mathrm{X}) = \mathcal{I}_p(\mathrm{X}) \tag{19}$$

Proof. By definition, we have $\mathcal{I}_p^{\mathcal{G}^{\emptyset}}(\mathrm{X}) = 0$. □

Then we can see that our objective is equivalent to the original MIB objective equation 5 when $\alpha_1 = 1, \beta_1 = \gamma$.

$$\mathcal{L} = \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{str\_reg}} = \alpha_1 D_{\mathrm{KL}}(q_{\boldsymbol{\phi}} \parallel p_{\boldsymbol{\theta}}) + \beta_1 \mathbb{D}(q_{\boldsymbol{\phi}} \parallel \mathcal{G}^{\emptyset}) = \alpha_1 D_{\mathrm{KL}}(q_{\boldsymbol{\phi}} \parallel p_{\boldsymbol{\theta}}) + \beta_1 \mathcal{I}_q^{\mathcal{G}^q} \tag{20}$$

A.1.2. Derivation of equation 12

$$\begin{aligned} q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}^{\mathbb{S}}) &\propto p_{\boldsymbol{\theta}}(\mathbf{z}) \prod_{i \in \mathbb{S}} \frac{q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)}{p_{\boldsymbol{\theta}}(\mathbf{z})} \\ &= p_{\boldsymbol{\theta}}(\mathbf{z}) \prod_{i \in \mathbb{S}} \prod_{j=1}^{M} \left(\tilde{q}_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i)\right)^{\mathbf{m}_{ij}^q} \\ &= \prod_{j=1}^{M} \left(p_{\boldsymbol{\theta}}(\mathbf{z}_j) \prod_{i \in \mathbb{S}} \left(\tilde{q}_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i)\right)^{\mathbf{m}_{ij}^q} \right) \end{aligned} \tag{21}$$

A.1.3. Full table 2

We show the full Table 2 in Table 5.

A.2. Proof of results in section 4.1 *single-modal generative mode*

A.2.1. Unifying disentangled generative models

β-VAE For β-vae we have

$$\begin{aligned} \mathcal{L} =& \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{str\_reg}} \\ =& C_1 + (\beta - 1)C_3 + (\beta - 1)\mathcal{L}_{\text{str\_reg}}(\mathcal{G}^{\emptyset}) \\ =& C_1 + (\beta - 1)C_3 + (\beta - 1)\mathbb{D}(q_{\boldsymbol{\phi}} \parallel \mathcal{G}^{\emptyset}) \\ =& \mathbb{E}_{q_{\boldsymbol{\phi}}} \log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{u}) + \mathbb{E}_{q_{\boldsymbol{\phi}}} D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{u} \mid \mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{u})) + (\beta - 1) D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{u}) \parallel p_{\boldsymbol{\theta}}(\mathbf{u})) + (\beta - 1)\mathcal{I}_q(\mathbf{x}\ ;\ \mathbf{u}) \\ =& \mathbb{E}_{q_{\boldsymbol{\phi}}} \log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{u}) + (1 + \beta - 1) D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{u}) \parallel p_{\boldsymbol{\theta}}(\mathbf{u})) + (1 + \beta - 1)\mathcal{I}_q(\mathbf{x}\ ;\ \mathbf{u}) \\ =& \mathbb{E}_{q_{\boldsymbol{\phi}}} \log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{u}) + \beta \mathbb{E}_{q_{\boldsymbol{\phi}}} D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{u} \mid \mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{u})) \\ \equiv& \mathcal{L}_{\beta-\text{vae}} \end{aligned} \tag{22}$$

where we include the structural regularization $\mathcal{L}_{\text{str\_reg}}$ using an empty Bayesian network $\mathcal{G}^{\beta-\text{vae}} \equiv \mathcal{G}^{\emptyset}$. Thus we show that the β-vae objective is equivalent to imposing another empty Bayesian network structure in the latent space which implies the independent latent factors.

TCVAE (Chen et al., 2018) We further show that how we can unify other total-correlation based disentangled representation learning models (Chen et al., 2018; Esmaeili et al., 2019; Kim & Mnih, 2018) by explicitly imposing Bayesian structure $\mathcal{G}^p$

Table 5. A unified view of {single/multi}-{modal/domain/view} models

MODELS	N	①	②	$\mathcal{G}^q$	$\mathcal{G}^p$	$\mathcal{L}_{\text{dist}}$	$\mathcal{L}_{\text{str\_reg}}$
VAE	1	×	×	$\left[\mathcal{G}^q_{\text{single}}\right]$	$\left[\mathcal{G}^p_{\text{single}}\right]$	$[1, C_1]$	$[]$
ICA	1	×	×	$\left[\mathcal{G}^q_{\text{single}}\right]$	$[]$	$[]$	$[\beta, \mathcal{G}^p_{\text{single}}]$
GAN	1	×	×	$[]$	$\mathcal{G}^p_{\text{single}}$	$[1, C_2]$	$[]$
INFOGAN	1	×	×	$[]$	$\mathcal{G}^p_{\text{single}}$	$[1, C_2]$	$[1, \mathcal{G}^{\text{InfoGAN}}]$
β-VAE	1	×	×	$\left[\mathcal{G}^q_{\text{single}}\right]$	$\left[\mathcal{G}^p_{\text{single}}\right]$	$[1, C_1], [\beta - 1, C_3]$	$[\beta - 1, \mathcal{G}^{\emptyset}]$
β-TCVAE	1	×	×	$\left[\mathcal{G}^q_{\text{single}}\right]$	$\left[\mathcal{G}^p_{\text{single}}\right]$	$[1, C_1], [\alpha_2, C_3]$	$[\beta, \mathcal{G}^p]$
BIVCCA	2	×	×	$\left[\mathcal{G}^q_{\text{marginal}}\right]$	$[\mathcal{G}^p_{\text{joint}}]$	$[\alpha_i, C_4(\mathbf{x}_i, \mathbf{z})]]$	$[]$
JMVAE	2	×	×	$[\mathcal{G}^q_{\text{joint}}]$	$[\mathcal{G}^p_{\text{joint}}]$	$[1, C_1]$	$[\beta_i, \mathcal{G}^{\text{str}}_{\text{cross}}(\mathbf{x}_i)]$
TELBO	2	×	×	$\left[\mathcal{G}^q_{\text{joint}}, \mathcal{G}^q_{\text{marginal}}\right]$	$[\mathcal{G}^p_{\text{joint}}]$	$[1, C_1]$	$[\beta_i, \mathcal{G}^{\text{str}}_{\text{marginal}}(\mathbf{x}_i)]$
MVAE	N	×	×	$\left[\mathcal{G}^q_{\text{joint}}, \mathcal{G}^q_{\text{marginal}}\right]$	$[\mathcal{G}^p_{\text{joint}}]$	$[1, C_1]$	$[\beta_i, \mathcal{G}^{\text{str}}_{\text{marginal}}(\mathbf{x}_i)]$
WYNER	2	✓	×	$\left[\mathcal{G}^q_{\text{joint}}, \mathcal{G}^q_{\text{marginal}}\right]$	$[\mathcal{G}^p_{\text{joint}}]$	$[1, C_1]$	$[\beta_i, \mathcal{G}^{\text{str}}_{\text{cross}}(\mathbf{x}_i)], [\beta_i, \mathcal{G}^{\text{str}}_{\text{private}}(\mathbf{x}_i)]$
DIVA	3	✓	×	$\left[\mathcal{G}^q_{\text{marginal}}\right]$	$[\mathcal{G}^p_{\text{joint}}]$	$[1, C_1]$	$[\beta_i, \mathcal{G}^{\text{str}}_{\text{private}}(\mathbf{x}_i)]$
OURS-MM	N	✓	✓	$[\mathcal{G}^q_{\text{full}}]$	$[\mathcal{G}^p_{\text{full}}]$	$[1, C_0]$	$[\beta_i, \mathcal{G}^{\text{str}}_{\text{cross}}(\{\mathbf{x}_i\})]$

as structural regularization, where a factorized prior distribution is assumed.

$$
\begin{aligned}
&\mathcal{L} = C_1 + \alpha_2 C_3 + \beta \mathcal{L}_{\text{str\_reg}} \\
&\mathcal{L}_{\text{str\_reg}} = \mathbb{D}(q_{\boldsymbol{\phi}} \parallel \mathcal{G}^p) = \mathcal{I}_q^{\mathcal{G}^q} - \mathcal{I}_q^{\mathcal{G}^p} = \sum_j^M \mathcal{I}_q(\mathbf{x}\ ;\ \mathbf{u}_j) - \mathcal{I}_q(\mathbf{x}\ ;\ \mathbf{u}) = \mathcal{I}_q(\mathbf{u}) - \mathcal{I}_q(\mathbf{u} \mid \mathbf{x}) = \mathcal{I}_q(\mathbf{u}) \equiv TC(\mathbf{u})
\end{aligned}
\tag{23}
$$

Since we assume a factorized posterior distribution $q_{\boldsymbol{\phi}}(\mathbf{u} \mid \mathbf{x})$, we have $\mathcal{I}_q(\mathbf{u} \mid \mathbf{x}) = 0$ in the last line of above objective. Thus the total-correlation minimization term emerges as a structural regularization term naturally in our framework.

A.3. Proof of results in section 4.2 *multi-modal/domain/view generative model*

A.3.1. UNIFYING MULTI-MODAL/DOMAIN/VIEW GENERATIVE MODELS

We show that we can obtain several representative multi-modal generative models as special cases of our proposed framework here.

JMVAE (Suzuki et al., 2017) We can see that the objective of JMVAE is a speacial case of our proposed objective when $N = 2$.

Wyner-VAE (Ryu et al., 2020) By using structural regularization $\mathbb{D}(q_{\boldsymbol{\phi}} \parallel \mathcal{G}^{\text{str}}_{\text{cross}}(\mathbf{x}_i))$, we show that we can obtain the mutual information regularization term appeared in the learning objective of Wyner-VAE (Ryu et al., 2020)

$$
\begin{aligned}
\mathcal{L}_{\text{str\_reg}} &= \mathbb{D}(q_{\boldsymbol{\phi}} \parallel \mathcal{G}^{\text{str}}_{\text{cross}}(\mathbf{x}_i)) = \mathcal{I}_q^{\mathcal{G}^q} - \mathcal{I}_q^{\mathcal{G}^{\text{str}}_{\text{cross}}(\mathbf{x})} \\
&= \mathcal{I}_q(\mathbf{x}_1\ ;\ \mathbf{u}_1) + \mathcal{I}_q(\mathbf{x}_2\ ;\ \mathbf{u}_2) + \mathcal{I}_q(\mathbf{x}_1, \mathbf{x}_2\ ;\ \mathbf{z}) - \mathcal{I}_q(\mathbf{x}_1\ ;\ \mathbf{u}_1) - \mathcal{I}_q(\mathbf{x}_2\ ;\ \mathbf{u}_2) = \mathcal{I}_q(\mathbf{x}_1, \mathbf{x}_2\ ;\ \mathbf{z}) \\
\mathcal{L} &= \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{str\_reg}} = \beta \mathcal{I}_q(\mathbf{x}_1, \mathbf{x}_2\ ;\ \mathbf{z}) + \mathcal{L}_{\text{dist}} \equiv \mathcal{L}_{\text{wyner-vae}}
\end{aligned}
\tag{24}
$$

CorEx (Steeg & Galstyan, 2014a) One of the most interesting model with similar goal to decorrelate observed variables is

CorEx (Steeg & Galstyan, 2014a;b; 2016; Gao et al., 2019), whose objective is

$$\max_{G_j, q_{\phi}\left(\mathbf{z}_j \mid \mathbf{x}_{G_j}\right)} \mathcal{L}_{CorEx} = \sum_{j=1}^{M} TC\left(\mathbf{x}_{G_j}\right) - TC\left(\mathbf{x}_{G_j} \mid \mathbf{z}_j\right) \quad \text{s.t.} \quad G_j \cap G_{j' \neq j} = \emptyset \tag{25}$$

For each $1 \leq j \leq M$, CorEx objective aims to search for a latent variable Z_j to achieve maximum total-correlation reduction $TC\left(\mathbf{x}_{G_j}\right) - TC\left(\mathbf{x}_{G_j} \mid \mathbf{z}_j\right)$ of a group of observed variables X_{G_j}. We use $\mathrm{M}^q_{:,j}$ and $\mathrm{M}^p_{i,:}$ to represent G_j equivalently, then our objective is

$$\begin{aligned}
&\mathcal{L}_{\text{dist}} = \mathbb{D}(q_{\phi} \parallel \mathcal{G}^p) = \mathcal{I}_q^{\mathcal{G}^q} - \mathcal{I}_q^{\mathcal{G}^p} = \sum_{j=1}^{M} \mathcal{I}_q(\mathbf{z}_j \ ; \ \mathbf{x}^{\mathbf{m}_j^q}) - \sum_{i=1}^{N} \mathcal{I}_q(\mathbf{z}^{\mathbf{m}_i^p} \ ; \ \mathbf{x}_i) \\
&= \sum_{j=1}^{M} \sum_{i=1}^{N} \mathbf{m}_{ij}^q \mathcal{I}_q(\mathbf{z}_j \ ; \ \mathbf{x}_i) + \sum_{j=1}^{M} \left[\mathcal{I}_q(\mathbf{x}^{\mathbf{m}_j^q} \mid \mathbf{z}_j) - \mathcal{I}_q(\mathbf{x}^{\mathbf{m}_j^q}) \right] - \sum_{i=1}^{N} \sum_{j=1}^{M} \mathbf{m}_{ij}^p \mathcal{I}_q(\mathbf{z}_j \ ; \ \mathbf{x}_i) - \sum_{i=1}^{N} \left[\mathcal{I}_q(\mathbf{z}^{\mathbf{m}_i^p} \mid \mathbf{x}_i) - \mathcal{I}_q(\mathbf{z}^{\mathbf{m}_i^p}) \right] \\
&\leq \sum_{j=1}^{M} \left[\mathcal{I}_q(\mathbf{x}^{\mathbf{m}_j^q} \mid \mathbf{z}_j) - \mathcal{I}_q(\mathbf{x}^{\mathbf{m}_j^q}) \right] + \sum_{i=1}^{N} I_q(\mathbf{z}^{\mathbf{m}_i^p}) \\
&\equiv -\mathcal{L}_{CorEx} + \sum_{i=1}^{N} \mathcal{I}_q(\mathbf{z}^{\mathbf{m}_i^p})
\end{aligned} \tag{26}$$

Thus with structural regularization $\mathcal{G}^p$ we obtained an objective coincides with CorEx-based variational autoencoder (Gao et al., 2019), which is also upper-bound of original CorEx objective(Steeg & Galstyan, 2014a) with additional disentanglement regularization over latent variables.

A.3.2. Derivation of objective equation 16

We show the detailed derivation of the learning objective of our multi-domain generative model here. As introduced in 4.2, we impose N structural regularization for each individual $\mathrm{X}^{\mathbb{S}} = \{\mathrm{X}_i\}$ as $\mathbb{D}(q_{\phi} \parallel \mathcal{G}^{\text{str}}_{\text{cross}}(\{\mathbf{x}_i\}))$. First we hvae

Proposition 2. *We have following upper-bound*

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{D}(q_{\phi} \parallel \mathcal{G}^{\text{str}}_{\text{cross}}(\{\mathbf{x}_i\})) \leq \mathcal{L}_{\mathbf{u}} + \sum_{i=1}^{N} \mathbb{E}_{q_{\phi}} D_{\text{KL}}(q_{\phi}(\mathbf{z} \mid \mathbf{x}) \parallel q_{\phi}(\mathbf{z} \mid \mathbf{x}_i)) \tag{27}$$

Proof.

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\mathbb{D}(q_{\boldsymbol{\phi}} \parallel \mathcal{G}^{\text{str}}_{\text{cross}}(\{\mathbf{x}_i\})) = \mathcal{I}_q(\mathbf{u}\,;\,\mathbf{x}) + \frac{1}{N}\sum_{i=1}^{N}\left[\mathcal{I}_q(\mathbf{z}\,;\,\mathbf{x}) - \mathcal{I}_q(\mathbf{z}\,;\,\mathbf{x}_i) - \sum_{k\neq i}^{N}\mathcal{I}_q(\mathbf{z}\,;\,\mathbf{x}_k)\right] \\
&= \mathcal{I}_q(\mathbf{u}\,;\,\mathbf{x}) + \frac{1}{N}\sum_{i=1}^{N}\mathcal{I}_q(\mathbf{z}\,;\,\mathbf{x}) + \frac{1}{N}\sum_{i=1}^{N}\left[-\mathcal{I}_q(\mathbf{z}\,;\,\mathbf{x}_i) + \sum_{k\neq i}^{N}\mathcal{I}_q(\mathbf{z}\,;\,\mathbf{x}_k)\right] \\
&= \mathcal{I}_q(\mathbf{u}\,;\,\mathbf{x}) + \mathcal{I}_q(\mathbf{z}\,;\,\mathbf{x}) - \sum_{i=1}^{N}\mathcal{I}_q(\mathbf{z}\,;\,\mathbf{x}_i) \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}} D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{u} \mid \mathbf{x}) \parallel q^{\text{mg}}_{\boldsymbol{\phi}}(\mathbf{u})) + \sum_{i=1}^{N}\mathbb{E}_{q_{\boldsymbol{\phi}}} D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel q^{\text{mg}}_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)) \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}} D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{u} \mid \mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{u})) + \sum_{i=1}^{N}\mathbb{E}_{q_{\boldsymbol{\phi}}} D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)) \\
&- \mathbb{E}_{q_{\boldsymbol{\phi}}} D_{\text{KL}}(q^{\text{mg}}_{\boldsymbol{\phi}}(\mathbf{u}) \parallel p_{\boldsymbol{\theta}}(\mathbf{u})) - \sum_{i=1}^{N}\mathbb{E}_{q_{\boldsymbol{\phi}}} D_{\text{KL}}(q^{\text{mg}}_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)) \\
&\leq \mathbb{E}_{q_{\boldsymbol{\phi}}} D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{u} \mid \mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{u})) + \sum_{i=1}^{N}\mathbb{E}_{q_{\boldsymbol{\phi}}} D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)) \\
&= \mathcal{L}_{\mathbf{u}} + \sum_{i=1}^{N}\mathbb{E}_{q_{\boldsymbol{\phi}}} D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i))
\end{aligned}
$$

□

where $q^{\text{mg}}_{\boldsymbol{\phi}}(\mathbf{u}) \equiv \mathbb{E}_{q_{\boldsymbol{\phi}}} q_{\boldsymbol{\phi}}(\mathbf{u} \mid \mathbf{x})$ and $q^{\text{mg}}_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x} \mid \mathbf{x}_i)} q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ denote the induced marginalization of $q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{u}, \mathbf{z})$. Note that by using the above upper-bound, the inference network distribution $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)$ introduced in 3.4 is trained to approximate the true marginalization $q^{\text{mg}}_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$. Thus we have following full objective

$$
\begin{aligned}
\mathcal{L} = \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{str\_reg}} &= D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z}, \mathbf{u}) \parallel p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}, \mathbf{u})) + \frac{1}{N}\sum_{i=1}^{N}\mathbb{D}(q_{\boldsymbol{\phi}} \parallel \mathcal{G}^{\text{str}}_{\text{cross}}(\{\mathbf{x}_i\})) \\
&= -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}, \mathbf{u} \mid \mathbf{x})} \log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}, \mathbf{u}) && (\mathcal{L}_{\mathbf{x}}) \\
&+ \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x})} D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{u} \mid \mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{u})) && (\mathcal{L}_{\mathbf{u}}) \\
&+ \sum_{i=0}^{N}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x})} D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)) && (\mathcal{L}_{\mathbf{z}}) \\
&\equiv \mathcal{L}_{\mathbf{x}} + \mathcal{L}_{\mathbf{u}} + \mathcal{L}_{\mathbf{z}} && (28)
\end{aligned}
$$

We use $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_0) \equiv p_{\boldsymbol{\theta}}(\mathbf{z})$ for the simplicity of notations. We further show that $\mathcal{L}_{\mathbf{z}}$ can be viewed as a generalized JS-divergence for the reverse KL-divergence (Nielsen, 2019). We decompose $\mathcal{L}_{\mathbf{z}}$ regarding each latent variable Z_j,

$$
\begin{aligned}
&\mathcal{L}_{\mathbf{z}} = \sum_{j=1}^{M}\mathcal{L}_{\mathbf{z}_j}, \quad q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}) \propto \prod_{i=0}^{N} q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i)^{\gamma_{ij}} \\
&\mathcal{L}_{\mathbf{z}_j} = D^{\text{KL}^*}_{\text{JS}}(q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_0), q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_1), \ldots, q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_N)) \\
&\sum_{i=0}^{N}\gamma_i = 1, \gamma_{0j} = 1 - \sum_{i=1}^{N}\mathbf{m}^{q}_{ij}, \quad \gamma_{ij} = \mathbf{m}^{q}_{ij}\ i > 0
\end{aligned}
\qquad (29)
$$

where we use KL^* to denote the reverse KLD and following the same notation in (Nielsen, 2019) for the generalized JSD.

A.4. Proof of results in section 5 *case study: fair representation learning*

We show the detailed derivation of the learning objective 17 here.

$$
\begin{aligned}
\mathcal{L} &= \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{str\_reg}} = D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z},\mathbf{u}) \parallel p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z},\mathbf{u})) + \beta_1 \mathbb{D}(q_{\boldsymbol{\phi}} \parallel \mathcal{G}^{\text{str}}_{\text{informative}}) + \beta_2 \mathbb{D}(q_{\boldsymbol{\phi}} \parallel \mathcal{G}^{\text{str}}_{\text{invariant}}) \\
&= D_{\text{KL}}(q_{\boldsymbol{\phi}} \parallel p_{\boldsymbol{\theta}}) + \beta_1 I_q(\mathbf{x};\mathbf{a} \mid \mathbf{z}) + \beta_2 \mathcal{I}_q(\mathbf{z}\ ;\ \mathbf{u}) + const \\
&\leq -\mathbb{E}_{q_{\boldsymbol{\phi}}} \log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{a} \mid \mathbf{z},\mathbf{u}) + \beta_2 \mathcal{I}_q(\mathbf{z}\ ;\ \mathbf{u}) + (1+\beta_1)\mathbb{E}_{q_{\boldsymbol{\phi}}} D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x},\mathbf{a}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})) + const
\end{aligned} \tag{30}
$$

We can interpret this derived learning objective as first seeking for a succinct latent representation Z that captures the sufficient correlation between X and A, then Z is served as a proxy variable to learn an informative representation U with all information relevant to A eliminated by minimizing $\mathcal{I}_q(\mathbf{z}\ ;\ \mathbf{u})$.

A.5. Details of section 6 *case study: invariant risk minimization*

We show that the idea in (Arjovsky et al., 2019) can be directly integrated into our proposed framework by imposing stable M^p structure as constraints across environments, measured by gradient-penalty term shown below

$$
\mathcal{L}_{\text{gp}} = \mathcal{L}_{\text{dist}} + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{e})} \left\| \nabla_{\mathrm{M}^p} \mathcal{L}_{\text{score}} \right\| \tag{31}
$$

B. Experiments

B.1. Generative modeling

Datasets Following the same evaluation protocol proposed by previous works (Ryu et al., 2020; Wu & Goodman, 2018), we construct the bi-modal datasets MNIST-Label by using the digit label as a second modality, MNIST-SVHN by pairing each image sample in MNIST with another random SVHN image sharing the same digit label and a bi-view dataset MNIST-MNIST-Plus-1 by pairing each MNIST sample X_1 with another random sample X_2 correlated as $\mathrm{label}(\mathrm{X}_1)+1 = \mathrm{label}(\mathrm{X}_2)$. We illustrate the data generating process using Bayesian networks in Figure 6.

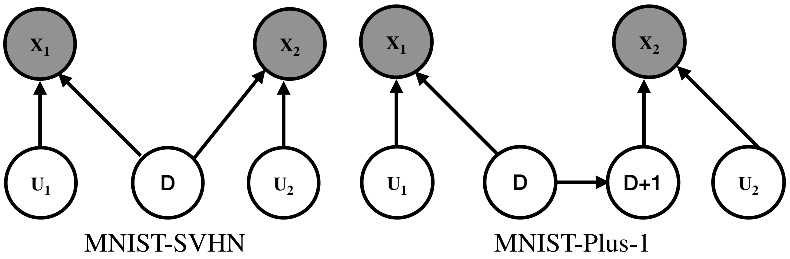

Figure 6. Bayesian networks for illustrating the data generating process of MNIST-SVHN dataset and MNIST-PLUS-1 dataset.

Training details and hyper-parameters For MNIST-Label dataset, we use MLPs with 2 hidden layers for both encoders and decoders, following the same neural network architecture in (Wu & Goodman, 2018). The dimension of Z modeling the shared information is 2. The dimension of U_1 modeling MNIST image is 20. We don't include U_2 in this setting and set the dimension of U_2 to 0. For MNIST-SVHN dataset, the dimension of Z is 2, the dimension of U_1 for MNIST is 20 and the dimension of U_2 for SVHN is 20. For MNIST-MNIST-Plus-1 dataset, the dimension of Z is 2, and the dimension of U_1 for MNIST is 20. We train the model using the Adam optimizer with a learning rate starting from 0.001, and decay the learning rate by a factor 0.1 whenever a validation loss plateau is found during training. We train the model up to 1000 epochs for all datasets. We learn the structural variable M with $steps\_dist = 1$ and $steps\_str = 3$ in all experiments. We use the same neural network architectures for encoder and decoder as (Ryu et al., 2020) in MNIST-SVHN and MNIST-MNIST-Plus-1 datasets.

Qualitative results of MNIST-Label Due to the space limit constraint, we include the qualitative results of MNIST-Label experiment here. We show the conditionally generated samples in figure 7.

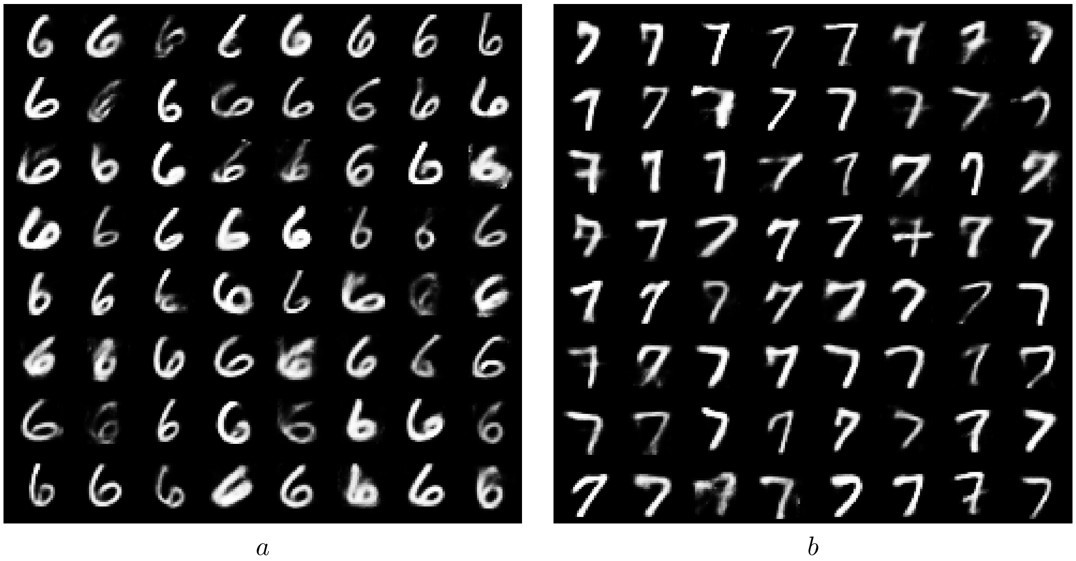

Figure 7. Conditionally generated samples when (a) $label = 6$ and (b) $label = 7$.

B.2. Fairness

Training details and hyperparameter sensitivity We follow the same neural network architecture design and evaluation process in (Song et al., 2019). The dimension of U is 10 for German and Adult datasets, the dimension of Z is 5. We find that the experimental result is not sensitive to the dimension of Z when it's in range 2 to 10. We train the model up to 10000 epochs using Adam optimizer with leraning rate 0.001, and decay the learning rate by a factor 0.1 when loss plateau is detected. We don't train the structural variables in this experiment. We re-scale the likelihood in objective to make the loss terms balance for the consideration of training stability. Numbers in table 3 are evaluated with 10 random runs with different random seeds.

B.3. Out-of-Distribution Generalization

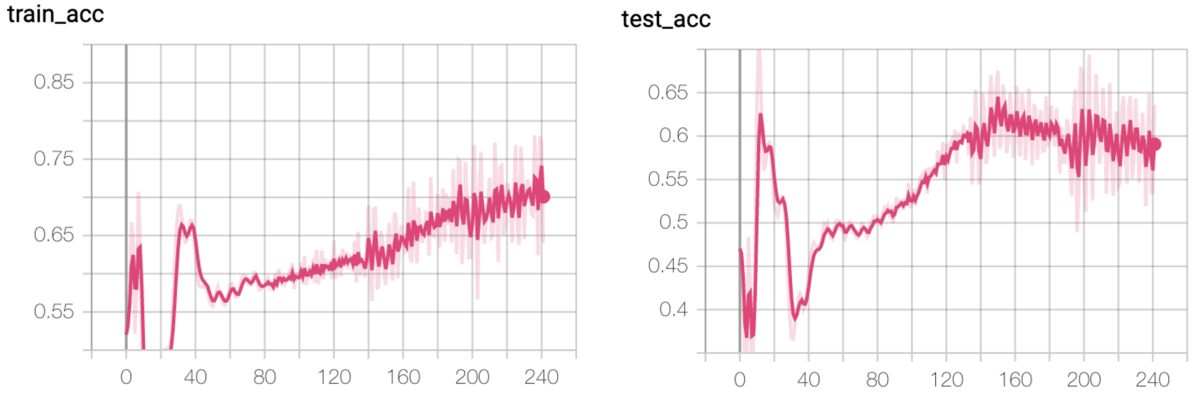

Figure 8. Training environment accuracy (Left) and testing environment accuracy (Right) on Colored-MNIST dataset

Colored MNIST *Colored MNIST* is an experiment that was used in (Arjovsky et al., 2019), in which the goal is to predict the label of a given digit in the presence of varying exterior factor e. The dataset for this experiment is derived from MNIST. Each member of the *Colored MNIST* dataset is constructed from an image-label pair (x, y) in MNIST, as follows.

1. Generate a binary label $\hat{y}_{obs}$ from y with the following rule: $\hat{y}_{obs} = 0$ if $y \in \{0 \sim 4\}$ and $\hat{y}_{obs} = 1$ otherwise.
2. Produce y_{obs} by flipping $\hat{y}_{obs}$ with a fixed probability p.
3. Let x_{fig} be the binary image corresponding to y.

4. Put $y_{obs} = \hat{x}_{ch1}$, and construct x_{ch1} from $\hat{x}_{ch}$ by flipping $\hat{x}_{ch1}$ with probability p_e.
5. Construct $x_{obs} = x_{fig} \times [x_{ch0}, (1 - x_{ch0}), 0]$.(that is, make the image red if $x_{ch1} = 1$ and green if $x_{ch1} = 0$.) Indeed, x_{obs} has exactly same information as the pair (x_{fig}, x_{ch1}).

The goal of this experiment is to use the dataset with p_e values in small compact range (training dataset) to train a model that can perform well on all ranges of p_e. In particular, we use the dataset with $p_e \in \{0.1, 0.2\}$ and evaluate the model on the dataset with $p_e = 0.9$. For more details of Colored MNIST experiment, please consult the original article.

Training details We follow the same neural network architecture design of encoder and evaluation process in (Arjovsky et al., 2019). The decoder is 1-layer MLP. We re-scale the likelihood terms to make the gradient norm of each one stays in the same magnitude. We train the model in a full-batch training manner, that the batch size is 50000. For semi-supervised training, we randomly partitioned the dataset into two halfs and alternate between training $(\mathrm{X}, \mathrm{E}, \mathrm{Y})$ and (X, E). The dimension of Z is 4. Following the same practice in (Arjovsky et al., 2019), we use early-stopping on validation set as regularization. Numbers in table 4 are evaluated with 10 random runs with different random seeds. We illustrate the training dynamics of our model by plotting the accuracy progression in both training environments and testing environment in Figure 8.