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 $C_1 = D_{KL}(q_{\phi} \parallel p_{\theta})$, $T = 1$ and $G^1 = G^0$, $K = 1$, we can recover the original MIB objective equation 5.

**Proposition 1.** Let $X \sim P(X)$, and let $G^0$ be an empty Bayesian network over $X$. Then

$$\mathbb{D}(p \parallel G^0) = \min_{q_{\phi} \parallel G^0} D_{KL}(p \parallel q) = I_p(X) - I_{p^G}(X) = I_p(X)$$

(19)

**Proof.** By definition, we have $I_{p^G}(X) = 0$. 

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

$$\mathcal{L} = \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{str.reg}} = \alpha_1 D_{KL}(q_{\phi} \parallel p_{\theta}) + \beta_1 \mathbb{D}(q_{\phi} \parallel G^0) = \alpha_1 D_{KL}(q_{\phi} \parallel p_{\theta}) + \beta_1 I_{p^G}$$

(20)

A.1.2. Derivation of equation 12

$$q_{\phi}(z \mid x^S) \propto p_{\theta}(z) \prod_{i \in S} q_{\phi}(z \mid x_i)$$

$$\frac{\prod_{i \in S} q_{\phi}(z \mid x_i)}{p_{\theta}(z)}$$

$$= p_{\theta}(z) \prod_{i \in S} \prod_{j=1}^{M} (q_{\phi}(z_j \mid x_i))^{m_{ij}}$$

$$= \prod_{j=1}^{M} \left( p_{\theta}(z_j) \prod_{i \in S} (q_{\phi}(z_j \mid x_i))^{m_{ij}} \right)$$

(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

$\beta$-VAE For $\beta$-vae we have

$$\mathcal{L} = \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{str.reg}}$$

$$= C_1 + (\beta - 1)C_3 + (\beta - 1) \mathcal{L}_{\text{str.reg}}(G^0)$$

$$= C_1 + (\beta - 1)C_3 + (\beta - 1) \mathbb{D}(q_{\phi} \parallel G^0)$$

$$= E_{q_{\phi}} \log p_{\theta}(x \mid u) + E_{q_{\phi}} D_{KL}(q_{\phi}(u \mid x) \parallel p_{\theta}(u)) + (\beta - 1)D_{KL}(q_{\phi}(u \mid x) \parallel p_{\theta}(u)) + (\beta - 1)I_q(x ; u)$$

$$= E_{q_{\phi}} \log p_{\theta}(x \mid u) + (1 + \beta - 1)D_{KL}(q_{\phi}(u \mid x) \parallel p_{\theta}(u)) + (1 + \beta - 1)I_q(x ; u)$$

$$= E_{q_{\phi}} \log p_{\theta}(x \mid u) + \beta E_{q_{\phi}} D_{KL}(q_{\phi}(u \mid x) \parallel p_{\theta}(u))$$

$$= \mathcal{L}_{\beta-\text{vae}}$$

(22)

where we include the structural regularization $\mathcal{L}_{\text{str.reg}}$ using an empty Bayesian network $G^\beta-\text{vae} \equiv G^0$. Thus we show that the $\beta$-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 $G^P$
We show that we can obtain several representative multi-modal generative models as special cases of our proposed framework. 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

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

$$\mathcal{L} = C_1 + \alpha_2 C_3 + \beta \mathcal{L}_{\text{str}_reg}$$

$$\mathcal{L}_{\text{str}_reg} = \mathcal{D}(\phi \| \mathcal{G}^p) = \mathcal{I}^{G_q} - \mathcal{I}^{G_p} = \sum_{j} M \mathcal{I}_q(x; u_j) - \mathcal{I}_q(x; u) = \mathcal{I}_q(u) - \mathcal{I}_q(u \mid x) = \mathcal{I}_q(u) \equiv TC(u) \quad (23)$$

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

### A.3.1. 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 special case of our proposed objective when $N = 2$.

#### Wyner-VAE (Ryu et al., 2020)

By using structural regularization $\mathcal{D}(\phi \| \mathcal{G}_{\text{str}}(x_i))$, we show that we can obtain the mutual information regularization term appeared in the learning objective of Wyner-VAE (Ryu et al., 2020)

$$\mathcal{L}_{\text{str}_reg} = \mathcal{D}(\phi \| \mathcal{G}_{\text{str}}(x_i)) = \mathcal{I}^{G_q} - \mathcal{I}^{G_{\text{str}}(x)}$$

$$= \mathcal{I}_q(x_1 \mid u_1) + \mathcal{I}_q(x_2 \mid u_2) + \mathcal{I}_q(x_1, x_2 \mid z) - \mathcal{I}_q(x_1 \mid u_1) - \mathcal{I}_q(x_2 \mid u_2) = \mathcal{I}_q(x_1, x_2 \mid z) \quad (24)$$

$$\mathcal{L} = \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{str}_reg} = \beta \mathcal{I}_q(x_1, x_2 \mid z) + \mathcal{L}_{\text{dist}} \equiv \mathcal{L}_{\text{wyner-va}}$$

#### CorEx (Steeg & Galstyan, 2014a)

One of the most interesting model with similar goal to decorrelate observed variables is
Learning Structured Latent Factors from Dependent Data

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

$$\max_{G_j, q_\phi(x_j | x_{G_j})} \mathcal{L}_{CorEx} = \sum_{j=1}^{M} TC(x_{G_j}) - TC(x_{G_j} | z_j)$$

s.t. $G_j \cap G_{j' \neq j} = \emptyset$ (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 $TC(x_{G_j}) - TC(x_{G_j} | z_j)$ of a group of observed variables $X_{G_j}$. We use $M^q_{:,j}$ and $M^p_{:,i}$ to represent $G_j$ equivalently, then our objective is

$$\mathcal{L}_{dist} = D(q_\phi \parallel G^p_{str}^{cross}(\{x_i\})) = \sum_{j=1}^{M} I_q(z_j ; x_i) - \sum_{j=1}^{N} I_q(z^m_j ; x_i)$$

$$= \sum_{j=1}^{M} \sum_{i=1}^{N} m^q_{ij} I_q(z_j ; x_i) + \sum_{j=1}^{M} \left[I_q(x_{G_j} | z_j) - I_q(x_{G_j})\right] - \sum_{i=1}^{N} \sum_{j=1}^{M} m^p_{ij} I_q(z_j ; x_i) - \sum_{i=1}^{N} \left[I_q(z^m_i | x_i) - I_q(z^m_i)\right]$$

$$\leq \sum_{j=1}^{M} \left[I_q(x_{G_j} | z_j) - I_q(x_{G_j})\right] + \sum_{i=1}^{N} I_q(z^m_i)$$

$$= -\mathcal{L}_{CorEx} + \sum_{i=1}^{N} I_q(z^m_i)$$

Thus with structural regularization $G^p$ 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 disentangle-ment 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 $X_i = \{x_i\}$ as $\mathbb{D}(q_\phi \parallel G^{str}_{cross}(\{x_i\}))$. First we have

Proposition 2. We have following upper-bound

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{D}(q_\phi \parallel G^{str}_{cross}(\{x_i\})) \leq \mathcal{L}_{u} + \sum_{i=1}^{N} E_{q_\phi}D_{KL}(q_\phi(z | x) \parallel q_\phi(z | x_i))$$

(27)
Proof.

\[ \frac{1}{N} \sum_{i=1}^{N} D(q_{\phi} \parallel G_{cross}^\text{str}(\{x_i\})) = I_{q}(u \mid x) + \frac{1}{N} \sum_{i=1}^{N} \left[ I_{q}(z \mid x) - I_{q}(z \mid x_i) - \sum_{k \neq i} I_{q}(z \mid x_k) \right] \]

\[ = I_{q}(u \mid x) + \frac{1}{N} \sum_{i=1}^{N} I_{q}(z \mid x) + \frac{1}{N} \sum_{i=1}^{N} \left[-I_{q}(z \mid x_i) + \sum_{k \neq i} I_{q}(z \mid x_k) \right] \]

\[ = I_{q}(u \mid x) + I_{q}(z \mid x) - \sum_{i=1}^{N} I_{q}(z \mid x_i) \]

\[ = E_{q_{\phi}} D_{KL}(q_{\phi}(u \mid x) \parallel q_{\phi}^{mg}(u)) + \sum_{i=1}^{N} E_{q_{\phi}} D_{KL}(q_{\phi}(z \mid x) \parallel q_{\phi}(z \mid x_i)) \]

\[ = E_{q_{\phi}} D_{KL}(q_{\phi}(u \mid x) \parallel p_{\theta}(u)) + \sum_{i=1}^{N} E_{q_{\phi}} D_{KL}(q_{\phi}(z \mid x) \parallel q_{\phi}(z \mid x_i)) \]

\[ - E_{q_{\phi}} D_{KL}(q_{\phi}^{mg}(u) \parallel p_{\theta}(u)) - \sum_{i=1}^{N} E_{q_{\phi}} D_{KL}(q_{\phi}^{mg}(z \mid x) \parallel q_{\phi}(z \mid x_i)) \]

\[ \leq E_{q_{\phi}} D_{KL}(q_{\phi}(u \mid x) \parallel p_{\theta}(u)) + \sum_{i=1}^{N} E_{q_{\phi}} D_{KL}(q_{\phi}(z \mid x) \parallel q_{\phi}(z \mid x_i)) \]

\[ = L_u + \sum_{i=1}^{N} E_{q_{\phi}} D_{KL}(q_{\phi}(z \mid x) \parallel q_{\phi}(z \mid x_i)) \]

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

\[ L = L_{\text{dist}} + L_{\text{str-reg}} = D_{KL}(q_{\phi}(x, z, u) \parallel p_{\theta}(x, z, u)) + \frac{1}{N} \sum_{i=1}^{N} D(q_{\phi} \parallel G_{cross}^\text{str}(\{x_i\})) \]

\[ = -E_{q_{\phi}(x,z,u)} \log p_{\theta}(x \mid z, u) \]
\[ + E_{q_{\phi}(x)} D_{KL}(q_{\phi}(u \mid x) \parallel p_{\theta}(u)) \]

\[ + \sum_{i=0}^{N} E_{q_{\phi}(x)} D_{KL}(q_{\phi}(z \mid x) \parallel q_{\phi}(z \mid x_i)) \]

\[ \equiv L_x + L_u + L_z \]

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

\[ L_z = \sum_{j=1}^{M} L_{z_j}, \quad q_{\phi}(z_j \mid x) \propto \prod_{i=0}^{N} q_{\phi}(z_j \mid x_i)^{\gamma_{ij}} \]

\[ L_{z_j} = D_{KL^*}^{JS}(q_{\phi}(z_j \mid x_0), q_{\phi}(z_j \mid x_1), \ldots, q_{\phi}(z_j \mid x_N)) \]

\[ \sum_{i=0}^{N} \gamma_i = 1, \gamma_{0j} = 1 - \sum_{i=1}^{N} m_{ij}^*, \quad \gamma_{ij} = m_{ij}^*, \quad i > 0 \]

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.

\[
\mathcal{L} = \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{str-reg}} = D_{\text{KL}}(q_{\phi}(x, z, u) \parallel p_{\theta}(x, z, u)) + \beta_1 D(q_{\phi} \parallel \mathcal{G}_{\text{instructed}}) + \beta_2 D(q_{\phi} \parallel \mathcal{G}_{\text{invariant}}) \\
= D_{\text{KL}}(q_{\phi} \parallel p_{\theta}) + \beta_1 I_{q_{\phi}(x, a \mid z, u)} + \beta_2 I_{q_{\phi}(z \mid x, a)} + \text{const} \\
\leq -\mathbb{E}_{q_{\phi}} \log p_{\theta}(x, a \mid z, u) + \beta_1 I_{q_{\phi}(z \mid x, a)} + (1 + \beta_1) D_{\text{KL}}(q_{\phi}(z \mid x, a) \parallel p_{\theta}(z)) + \text{const} 
\]

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 \(I_{q_{\phi}}(z \mid x, a)\).

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

\[
\mathcal{L}_{GP} = \mathcal{L}_{\text{dist}} + \mathbb{E}_{q_{phi}(e)} \| \nabla_{M^p} \mathcal{L}_{\text{score}} \| 
\]

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 \(\text{label}(X_1) + 1 = \text{label}(X_2)\).

We illustrate the data generating process using Bayesian networks in Figure 6.

![Bayesian networks for illustrating the data generating process of MNIST-SVHN dataset and MNIST-PLUS-1 dataset.](image)

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 \(\text{steps\_dist} = 1\) and \(\text{steps\_str} = 3\) 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.
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 learning 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

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 $y$. 

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

*Figure 8. Training environment accuracy (Left) and testing environment accuracy (Right) on Colored-MNIST dataset.*
4. Put $y_{\text{obs}} = \hat{x}_{\text{ch}1}$, and construct $x_{\text{ch}1}$ from $\hat{x}_{\text{ch}}$ by flipping $\hat{x}_{\text{ch}1}$ with probability $p_e$.
5. Construct $x_{\text{obs}} = x_{\text{fig}} \times [x_{\text{ch}0}, (1 - x_{\text{ch}0}), 0]$. (that is, make the image red if $x_{\text{ch}1} = 1$ and green if $x_{\text{ch}1} = 0$.) Indeed, $x_{\text{obs}}$ has exactly same information as the pair $(x_{\text{fig}}, x_{\text{ch}1})$.

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 halves and alternate between training $(X, E, Y)$ 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.