Supplementary Materials

A Collective decision making through multi-agent settings

Proposition 1. *Consider the problem of mixture representation learning in a multi-agent VAE framework with $A \geq 2$ agents using type-preserving data augmentation, where the accuracy of categorical assignment for a single agent is imperfect. The confidence of the correct assignment for the multi-agent VAE is higher than that of the single agent VAE. Moreover, there exists an A such that the correct category receives the highest confidence score in the A-agent framework, independent of the categorical prior.*

Proof. Following Definition 1 of the multi-agent framework, each agent is represented by

$$p(\mathbf{x}_a|c_a) \propto \frac{p(c_a|\mathbf{x}_a)}{p(c_a)} \,. \tag{1}$$

In this framework, using a type-preserving data augmentation, each agent receives a noisy copy $\mathbf{x}_a$ of given sample $\mathbf{x}$. Without loss of generality, let $\mathbf{x} \sim p(\mathbf{x}|m)$, where $m \in \{1, \ldots, K\}$ denotes the true categorical assignment, respectively. Accordingly, $\mathbf{x}_a \sim p(\mathbf{x}_a|m)$ also belong to the same category-conditioned distribution $p(\mathbf{x}|m)$, $\forall m \in \{1, \ldots, K\}$. Considering the joint categorical assignment as c, where $c = c_1 = \cdots = c_A$. Defining $\mathbf{X} := \{\mathbf{x}_a\}_{1:A}$, the confidence measure for category k for samples from category m is expressed as,

$$\begin{aligned}
\mathcal{C}_m^A(k) &= \mathbb{E}_{p(\mathbf{X}|m)} \left[\log p(c = c_1 = \cdots = c_A = k|\mathbf{X}) \right], \qquad (2) \\
&= \mathbb{E}_{(\mathbf{X}|m)} \left[\log \frac{p(\mathbf{x}_1|\mathbf{X} \setminus \{\mathbf{x}_1\}, c = k) p(\mathbf{x}_2|\mathbf{X} \setminus \{\mathbf{x}_1, \mathbf{x}_2\}, c = k) \ldots p(\mathbf{x}_A|c = k) p(c = k)}{p(\mathbf{X})} \right], \\
&= \mathbb{E}_{p(\mathbf{X}|m)} \left[\log \frac{p(p(\mathbf{x}_1|\mathbf{X} \setminus \{\mathbf{x}_1\}, c = k) \ldots p(\mathbf{x}_A|c = k)}{p(\mathbf{X})} \right] + \log p(c = k)
\end{aligned}$$

where we use $p(c) = p(c = c_1 = \cdots = c_A)$ to simplify the notation. Using the type-preserving data augmentation, since all samples independently generated from the same class-conditioned distribution, the confidence measure can be simplified as,

$$\begin{aligned}
\mathcal{C}_m^A(k) &= \mathbb{E}_{p(\mathbf{x}_a|m)} \left[\prod_{a=1}^{A} \log \frac{p(\mathbf{x}_a|c = k)}{p(\mathbf{x}_a)} \right] + \log p(c = k) \qquad (3) \\
&= \sum_{a=1}^{A} \mathbb{E}_{p(\mathbf{x}_a|m)} \left[\log \frac{p(\mathbf{x}_a|c = k)}{p(\mathbf{x}_a)} \right] + \log p(c = k) \qquad (4)
\end{aligned}$$

Since all of the augmented data are sampled from the same distribution, log-likelihood values of the augmented samples are equal on expectation, as follows:

$$\sum_{a=1}^{A} \mathbb{E}_{p(\mathbf{x}_a|m)} \left[\log p(\mathbf{x}_a|c) \right] = A \mathbb{E}_{\mathbf{x}} \left[\log p(\mathbf{x}|c) \right] \tag{5}$$

Therefore, the confidence over category k in an A-agent framework is defined as,

$$\mathcal{C}_m^A(k) = A \mathbb{E}_{p(\mathbf{x}|m)} \left[\log \frac{p(\mathbf{x}|c = k)}{p(\mathbf{x})} \right] + \log p(c = k). \tag{6}$$

According to Eq. 6, for a single agent (1-agent) and A-agent frameworks, the confidence values for the true categorical assignment, i.e. category m, are formulated as follows.

$$\begin{aligned}
\mathcal{C}_m^1(m) &= \mathbb{E}_{p(\mathbf{x}|m)} \left[\log \frac{p(\mathbf{x}|c = m)}{p(\mathbf{x})} \right] + \log p(c = m), \qquad (7) \\
&= D_{KL} \left(p(\mathbf{x}|c = m) \| p(\mathbf{x}) \right) + \log p(c = m)
\end{aligned}$$

$$\mathcal{C}_m^A(m) = A \, D_{KL} \left(p(\mathbf{x}|c = m) \| p(\mathbf{x}) \right) + \log p(c = m), \tag{8}$$

Since $D_{KL}\left(p(\mathbf{x}|c=m)\|p(\mathbf{x})\right)>0$ for $K>1$, for $A>1$, $\mathcal{C}_m^A(m)>\mathcal{C}_m^1(m)$.

Moreover, assuming $p(c=m|\mathbf{x})\neq p(c=n|\mathbf{x})$, $\forall\, m,n\in\{1,\ldots,K\}, n\neq m$, the correct categorical assignment receives the highest confidence for the 1-agent case, i.e. $\mathcal{C}_m^1(m)>\mathcal{C}_m^1(n)$, if and only if,

$$\mathbb{E}_{p(\mathbf{x}|m)}\left[\log\frac{p(\mathbf{x}|c=m)}{p(\mathbf{x}|c=n)}\right] \quad > \quad \log\frac{p(c=n)}{p(c=m)},\ \ \forall n\neq m, \tag{9}$$

which is a function of categorical distributions and is not always satisfied for any arbitrary prior distribution. When there are A agents receiving type-preserving noisy copies of the given sample $\mathbf{x}$,

$$\mathcal{C}_m^A(k) \quad = \quad A\mathbb{E}_{\mathbf{x}}\left[\log p(\mathbf{x}|c=k)\right]+\log p(c=k)-A\mathbb{E}_{\mathbf{x}}\left[\log p(\mathbf{x})\right], \tag{10}$$

Therefore, $\mathcal{C}_m^A(m)>\mathcal{C}_m^A(n),\ \ \forall n\neq m$, if and only if,

$$A\mathbb{E}_{\mathbf{x}}\left[\log\frac{p(\mathbf{x}|c=m)}{p(\mathbf{x}|c=n)}\right]>\log\frac{p(c=n)}{p(c=m)},\ \ \forall n\neq m. \tag{11}$$

Thus, when the number of agents, A, satisfies

$$A \quad > \quad \max_m\{\max\left(\rho(m)D^{-1}(m),1\right)\}, \tag{12}$$

where $\rho(m)=\max_{n\neq m}\log\dfrac{p(c=n)}{p(c=m)}$ and $D(m)=\min_{n\neq m}D_{KL}(p(\mathbf{x}|m)\|p(\mathbf{x}|n)))$, we have

$$\mathcal{C}_m^A(m)>\mathcal{C}_m^A(n),\ \ \forall n\neq m\ . \tag{13}$$

□

Corollary 1. *For a uniform prior on the discrete factors, one pair of VAE agents ($A=2$) is sufficient to increase the confidence of correct categorical assignment.*

Proof. For uniformly distributed clusters, $\rho(k)=0, \forall k\in\{1,\ldots,K\}$. According to Eq. 12, for any $A\geq 2$, the confidence increase criteria is satisfied. □

Remark 1. *When the augmentation is type-preserving, by definition, $p(\mathbf{x}_a|\mathbf{x}_b,c=k)=p(\mathbf{x}_a|c=k)$, where $\mathbf{x}_b$ could be either the given training sample or another noisy copy. If the augmented samples concentrate around $\mathbf{x}_b$, i.e. the augmenter under-explores the category-conditioned distribution, the above proof should be adapted by keeping the conditioning on $\mathbf{x}_b$ explicit. Conditionally independent terms used in Eq. 3 should be replaced by $p(\mathbf{x}_a|\mathbf{x}_b,c=k)$ as follows.*

Eq. 2 should read as

$$\begin{aligned}\mathcal{C}_m^A(k) \quad &= \quad \mathbb{E}_{p(\mathbf{X}|m)}\left[\log p(c=c_1=\cdots=c_A=k|\mathbf{X})\right], \\ &= \quad \mathbb{E}_{p(\mathbf{X}|m)}\left[\log\frac{p(\mathbf{x}_1|\mathbf{x}_b,\mathbf{X}\setminus\{\mathbf{x}_1,\mathbf{x}_b\},c=k)\ldots p(\mathbf{x}_A|\mathbf{x}_b,c=k)p(\mathbf{x}_b|c=k)p(c=k)}{p(\mathbf{X})}\right]\end{aligned} \tag{14}$$

Since all augmented samples are generated from sample $\mathbf{x}_b$, the conditional probability distribution can be simplified as follows.

$$p(\mathbf{x}_a|\mathbf{x}_b,\mathbf{X}\setminus\{\mathbf{x}_a,\mathbf{x}_b\},c)=p(\mathbf{x}_a|\mathbf{x}_b,c),\ \ \textit{for } a\neq b \tag{15}$$

Accordingly, Eq. 14 can be simplified as,

$$\begin{aligned}\mathcal{C}_m^A(k) \quad &= \quad \mathbb{E}_{p(\mathbf{x}_a|\mathbf{x}_b,m)}\left[\log\prod_{a=1}^{A-1}\frac{p(\mathbf{x}_a|\mathbf{x}_b,c=k)}{p(\mathbf{x}_a|\mathbf{x}_b)}\right]+\mathbb{E}_{p(\mathbf{x}_b|m)}\left[\log\frac{p(\mathbf{x}_b|c=k)}{p(\mathbf{x}_b)}\right]+\log p(c=k) \\ &= \quad (A-1)\mathbb{E}_{p(\mathbf{x}_a|\mathbf{x}_b,m)}\left[\log\frac{p(\mathbf{x}_a|\mathbf{x}_b,c=k)}{p(\mathbf{x}_a|\mathbf{x}_b)}\right]+\mathcal{C}_m^1(k)\end{aligned} \tag{16}$$

Based on Eq. 16, if the data augmenter only regenerates given sample $\mathbf{x}$, the confidence value is equal to the confidence of the single framework. Then, Eq. 8 should read as

$$\mathcal{C}_m^A(m) \quad = \quad (A-1)D_{KL}\left(p(\mathbf{x}_a|\mathbf{x}_b,c=m)\|p(\mathbf{x}_a|\mathbf{x}_b)\right)+\mathcal{C}_m^1(m)\ . \tag{17}$$

Accordingly, $\forall\mathbf{x}_a$, if $\mathbf{x}_a=\mathbf{x}_b$, then $\mathcal{C}_m^A(m)=\mathcal{C}_m^1(m)$.

B Variational lower bound for conditional single mix-VAE

For completeness, we first derive the evidence lower bound (ELBO) for an observation $\mathbf{x}$ described by one categorical random variable (RV), $\mathbf{c}$, and one continuous RV, $\mathbf{s}$, without assuming conditional independence of $\mathbf{c}$ and $\mathbf{s}$ given $\mathbf{x}$. The variational approach to choosing the latent variables corresponds to solving the optimization equation

$$q^*(\mathbf{s},\mathbf{c}|\mathbf{x}) = \arg\min\nolimits_{q(\mathbf{s},\mathbf{c}|\mathbf{x})\in\mathcal{D}} D_{\mathrm{KL}}\left(q(\mathbf{s},\mathbf{c}|\mathbf{x})\|p(\mathbf{s},\mathbf{c}|\mathbf{x})\right) \ , \tag{18}$$

where $\mathcal{D}$ is a family of density functions over the latent variables. However, evaluating the objective function requires knowledge of $p(\mathbf{x})$, which is usually unknown. Therefore, we rewrite the divergence term as

$$\begin{aligned}
D_{\mathrm{KL}}\left(q(\mathbf{s},\mathbf{c}|\mathbf{x})||p(\mathbf{s},\mathbf{c}|\mathbf{x})\right) \quad &= \quad \int_{\mathbf{s}}\sum_{\mathbf{c}} q(\mathbf{s}|\mathbf{c},\mathbf{x})q(\mathbf{c}|\mathbf{x}) \log \frac{q(\mathbf{s}|\mathbf{c},\mathbf{x})q(\mathbf{c}|\mathbf{x})}{\frac{p(\mathbf{x}|\mathbf{s},\mathbf{c})p(\mathbf{s}|\mathbf{c})p(\mathbf{c})}{p(\mathbf{x})}}\, d\mathbf{s} \\
&= \quad \int_{\mathbf{s}}\sum_{\mathbf{c}} q(\mathbf{s}|\mathbf{c},\mathbf{x})q(\mathbf{c}|\mathbf{x}) \log \frac{q(\mathbf{s}|\mathbf{c},\mathbf{x})}{p(\mathbf{s}|\mathbf{c})}\, d\mathbf{s} \\
&\quad\quad + \int_{\mathbf{s}}\sum_{\mathbf{c}} q(\mathbf{s}|\mathbf{c},\mathbf{x})q(\mathbf{c}|\mathbf{x}) \log \frac{q(\mathbf{c}|\mathbf{x})}{p(\mathbf{c})}\, d\mathbf{s} \\
&\quad\quad + \int_{\mathbf{s}}\sum_{\mathbf{c}} q(\mathbf{s}|\mathbf{c},\mathbf{x})q(\mathbf{c}|\mathbf{x}) \log p(\mathbf{x})\, d\mathbf{s} \\
&\quad\quad - \int_{\mathbf{s}}\sum_{\mathbf{c}} q(\mathbf{s}|\mathbf{c},\mathbf{x})q(\mathbf{c}|x) \log p(\mathbf{x}|\mathbf{s},\mathbf{c})\, d\mathbf{s} \\
&= \quad \log p(\mathbf{x}) - \mathbb{E}_{q(\mathbf{c}|\mathbf{x})}\left[\mathbb{E}_{(q(\mathbf{s}|\mathbf{c},\mathbf{x}))}\left[\log p(\mathbf{x}|\mathbf{s},\mathbf{c})\right]\right] \\
&\quad\quad + \mathbb{E}_{q(\mathbf{c}|\mathbf{x})}\left[D_{KL}\left(q(\mathbf{s}|\mathbf{c},\mathbf{x})\|p(\mathbf{s}|\mathbf{c})\right)\right] + \mathbb{E}_{q(\mathbf{s}|\mathbf{c},\mathbf{x})}\left[D_{KL}\left(q(\mathbf{c}|\mathbf{x})\|p(\mathbf{c})\right)\right] && (19) \\
&= \quad \log p(\mathbf{x}) - \mathcal{L}_{\mathbf{s}} && (20)
\end{aligned}$$

Since $\log p(\mathbf{x})$ is unknown, instead of minimizing Eq. 19, the variational lower bound

$$\mathcal{L}_{\mathbf{s}} = \mathbb{E}_{q(\mathbf{c}|\mathbf{x})}\left[\mathbb{E}_{(q(\mathbf{s}|\mathbf{c},\mathbf{x}))}\left[\log p(\mathbf{x}|\mathbf{s},\mathbf{c})\right]\right] - \mathbb{E}_{q(\mathbf{c}|\mathbf{x})}\left[D_{KL}\left(q(\mathbf{s}|\mathbf{c},\mathbf{x})\|p(\mathbf{s}|\mathbf{c})\right)\right] - \mathbb{E}_{q(\mathbf{s}|\mathbf{c},\mathbf{x})}\left[D_{KL}\left(q(\mathbf{c}|\mathbf{x})\|p(\mathbf{c})\right)\right] \tag{21}$$

can be maximized. We choose $q(\mathbf{s}|\mathbf{c},\mathbf{x})$ to be a factorized Gaussian, parametrized using the reparametrization trick, and assume that the corresponding prior distribution is also a factorized Gaussian, $\mathbf{s}|\mathbf{c} \sim \mathcal{N}(0,\mathbf{I})$. Similarly, for the categorical variable, we assume a uniform prior, $\mathbf{c} \sim U(K)$.

C Variational inference for multi-agent autoencoding networks

As discussed in the main text, the collective decision making for an A-agent VAE network can be formulated as an equality constrained optimization as follows.

$$\begin{gathered}
\max \quad \mathcal{L}(\boldsymbol{\phi}_1,\boldsymbol{\theta}_1,\mathbf{x}_1,\mathbf{s}_1,\mathbf{c}_1) + \cdots + \mathcal{L}(\boldsymbol{\phi}_A,\boldsymbol{\theta}_A,\mathbf{x}_A,\mathbf{s}_A,\mathbf{c}_A) \\
\text{s.t. } \mathbf{c}_1 = \cdots = \mathbf{c}_A
\end{gathered} \tag{22}$$

Without loss of generality, the optimization in Eq. 22 can be rephrased as follows.

$$\begin{gathered}
\max \quad \mathcal{L}(\boldsymbol{\phi}_1,\boldsymbol{\theta}_1,\mathbf{s}_1,\mathbf{c}_1) + \mathcal{L}(\boldsymbol{\phi}_2,\boldsymbol{\theta}_2,\mathbf{s}_2,\mathbf{c}_2) + \cdots + \mathcal{L}(\boldsymbol{\phi}_A,\boldsymbol{\theta}_A,\mathbf{s}_A,\mathbf{c}_A) \\
\text{s.t. } \mathbf{c}_1 = \mathbf{c}_2 \\
\mathbf{c}_1 = \mathbf{c}_3 \\
\cdots \\
\mathbf{c}_1 = \mathbf{c}_A \\
\cdots \\
\mathbf{c}_{A-1} = \mathbf{c}_A
\end{gathered} \tag{23}$$

where the equality constraint is represented as $\binom{A}{2}$ pairs of categorical agreements. Multiplying the objective term in Eq. 22 by a constant value, $A-1$, we obtain,

$$\begin{gathered}
\max \quad (A-1)\left(\mathcal{L}\left(\boldsymbol{\phi}_1,\boldsymbol{\theta}_1,\mathbf{s}_1,\mathbf{c}_1\right) + \mathcal{L}\left(\boldsymbol{\phi}_2,\boldsymbol{\theta}_2,\mathbf{s}_2,\mathbf{c}_2\right) + \cdots + \mathcal{L}\left(\boldsymbol{\phi}_A,\boldsymbol{\theta}_A,\mathbf{s}_A,\mathbf{c}_A\right)\right) \\
\text{s.t. } \ \mathbf{c}_a = \mathbf{c}_b \quad \forall a,b \in [1,A], a < b
\end{gathered} \tag{24}$$

Consider one pair of $\mathcal{L}$ objectives for two agents a and b:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi}_a, \boldsymbol{\theta}_a, \mathbf{s}_a, \mathbf{c}_a) + \mathcal{L}(\boldsymbol{\phi}_b, \boldsymbol{\theta}_b, \mathbf{s}_b, \mathbf{c}_b) = {} & \mathbb{E}_{q_{\boldsymbol{\phi}_a}(\mathbf{s}_a, \mathbf{c}_a|\mathbf{x}_a)}\left[\log p_{\boldsymbol{\theta}_a}(\mathbf{x}_a|\mathbf{s}_a, \mathbf{c}_a)\right] + \mathbb{E}_{q_{\boldsymbol{\phi}_b}(\mathbf{s}_b, \mathbf{c}_b|\mathbf{x}_b)}\left[\log p_{\boldsymbol{\theta}_b}(\mathbf{x}_b|\mathbf{s}_b, \mathbf{c}_b)\right] \\ & - \mathbb{E}_{q_{\boldsymbol{\phi}_a}(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q_{\boldsymbol{\phi}_a}(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a)\right)\right] - \mathbb{E}_{q_{\boldsymbol{\phi}_b}(\mathbf{c}_b|\mathbf{x}_b)}\left[D_{KL}\left(q_{\boldsymbol{\phi}_b}(\mathbf{s}_b|\mathbf{c}_b, \mathbf{x}_b)\|p(\mathbf{s}_b|\mathbf{c}_b)\right)\right] \\ & - \mathbb{E}_{q_{\boldsymbol{\phi}_a}(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)}\left[D_{KL}\left(q_{\boldsymbol{\phi}_a}(\mathbf{c}_a|\mathbf{x}_a)\|p(\mathbf{c}_a)\right)\right] - \mathbb{E}_{q_{\boldsymbol{\phi}_b}(\mathbf{s}_b|\mathbf{c}_b, \mathbf{x}_b)}\left[D_{KL}\left(q_{\boldsymbol{\phi}_b}(\mathbf{c}_b|\mathbf{x}_b)\|p(\mathbf{c}_b)\right)\right] \qquad (25)\end{aligned}$$

Since all agents receive augmented samples from the same original distribution, we have $p(\mathbf{c}_a) = p(\mathbf{c}_b) = p(\mathbf{c})$. Using a simplified notation, $q_a = q_{\boldsymbol{\phi}_a}(\mathbf{c}_a|\mathbf{x}_a)$, the last two KL divergence terms can be expressed as,

$$\begin{aligned}D_{KL}\left(q_a\|p(\mathbf{c})\right) + D_{KL}\left(q_b\|p(\mathbf{c})\right) &= \sum_{\mathbf{c}_a} q_a \log\frac{q_a}{p(\mathbf{c})} + \sum_{\mathbf{c}_b} q_b \log\frac{q_b}{p(\mathbf{c})} \\ &= \sum_{\mathbf{c}_a}\sum_{\mathbf{c}_b} q_a q_b \log\frac{q_a}{p(\mathbf{c})} + \sum_{\mathbf{c}_a}\sum_{\mathbf{c}_b} q_a q_b \log\frac{q_b}{p(\mathbf{c})} \\ &= \sum_{\mathbf{c}_a}\sum_{\mathbf{c}_b} q_a q_b \log\frac{q_a q_b}{p(\mathbf{c})}\end{aligned} \tag{26}$$

Now, if we marginalize $p(\mathbf{c})$ over the joint distribution $p(\mathbf{c}_a, \mathbf{c}_b)$, we can represent the categorical prior distribution as follows.

$$p(\mathbf{c}) = \sum_{\mathbf{c}_a, \mathbf{c}_b} p(\mathbf{c}|\mathbf{c}_a, \mathbf{c}_b) p(\mathbf{c}_a, \mathbf{c}_b) \tag{27}$$

Since there is a categorical agreement condition i.e., $\mathbf{c}_a = \mathbf{c}_b$, $p(\mathbf{c})$ can be expressed as,

$$p(\mathbf{c}) = \sum_{\mathbf{m}} p(\mathbf{c}|\mathbf{c}_a = \mathbf{c}_b = \mathbf{m}) p(\mathbf{c}_a = \mathbf{c}_b = \mathbf{m}) \tag{28}$$

where

$$p(\mathbf{c}|\mathbf{c}_a = \mathbf{c}_b = \mathbf{m}) = \begin{cases} 1 & \mathbf{m} = \mathbf{c} \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

Accordingly, under the $\mathbf{c}_a = \mathbf{c}_b$ constraint, we merge those KL divergence terms as follows:

$$\begin{aligned}D_{KL}\left(q_a\|p(\mathbf{c})\right) + D_{KL}\left(q_b\|p(\mathbf{c})\right) &= \sum_{\mathbf{c}_a}\sum_{\mathbf{c}_b} q_a q_b \log\frac{q_a q_b}{p(\mathbf{c}_a, \mathbf{c}_b)} \\ &= D_{KL}(q_a q_b\|p(\mathbf{c}_a, \mathbf{c}_b))\end{aligned} \tag{30}$$

Finally, the optimization in Eq. 24 can be expressed as

$$\begin{aligned}\max \quad & \sum_{a=1}^{A} (A-1)\left(\mathbb{E}_{q(\mathbf{s}_a, \mathbf{c}_a|\mathbf{x}_a)}\left[\log p(\mathbf{x}_a|\mathbf{s}_a, \mathbf{c}_a)\right] - \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a)\right)\right]\right) - \\ & \qquad \sum_{a<b} \mathbb{E}_{q(\mathbf{s}_a, \mathbf{s}_b|\mathbf{c}_a, \mathbf{c}_b, \mathbf{x}_a, \mathbf{x}_b)}\left[D_{KL}\left(q(\mathbf{c}_a|\mathbf{x}_a) q(\mathbf{c}_b|\mathbf{x}_b)\|p(\mathbf{c}_a, \mathbf{c}_b)\right)\right] \\ & \qquad \text{s.t. } \mathbf{c}_a = \mathbf{c}_b \quad \forall a, b \in [1, A], a < b\end{aligned} \tag{31}$$

D VARIATIONAL LOWER BOUND FOR CPL-MIXVAE

In this section, using a pair of VAE agents, first we generalize the loss function for the single mix-VAE i.e., $\mathcal{L}_{\mathbf{s}}$ in Eq. 21, to the multi-agent case, and show that we can achieve the same objective function in Eq. 31. Then, we derive a relaxation for the equality constrained optimization.

Given input data $\mathbf{x}_a$, an agent approximates two models $q(\mathbf{c}_a|\mathbf{x}_a)$ and $q(\mathbf{s}_a|\mathbf{x}_a, \mathbf{c}_a)$. If we use pairwise coupling to allow interactions between the agents, then, for a pair of VAE agents, a and b, the variational lower bound obtained from the KL divergence in Equation (19) can be generalized as

$$\begin{aligned}\Delta(a,b) \quad &\triangleq \quad D_{\mathrm{KL}}\left(q(\mathbf{s}_a, \mathbf{s}_b, \mathbf{c}_a, \mathbf{c}_b|\mathbf{x}_a, \mathbf{x}_b)\|p(\mathbf{s}_a, \mathbf{s}_b, \mathbf{c}_a, \mathbf{c}_b|\mathbf{x}_a, \mathbf{x}_b)\right) \\ &= \quad \int_{\mathbf{s}_a}\int_{\mathbf{s}_b}\sum_{\mathbf{c}_a}\sum_{\mathbf{c}_b} q(\mathbf{s}_a, \mathbf{s}_b|\mathbf{c}_a, \mathbf{c}_b, \mathbf{x}_a, \mathbf{x}_b) q(\mathbf{c}_a, \mathbf{c}_b|\mathbf{x}_a, \mathbf{x}_b) \\ &\qquad \times \log\frac{q(\mathbf{s}_a, \mathbf{s}_b|\mathbf{c}_a, \mathbf{c}_b, \mathbf{x}_a, \mathbf{x}_b) q(\mathbf{c}_a, \mathbf{c}_b|\mathbf{x}_a, \mathbf{x}_b)}{\left(\dfrac{p(\mathbf{x}_a, \mathbf{x}_b|\mathbf{s}_a, \mathbf{s}_b, \mathbf{c}_a, \mathbf{c}_b) p(\mathbf{s}_a, \mathbf{s}_b|\mathbf{c}_a, \mathbf{c}_b) p(\mathbf{c}_a, \mathbf{c}_b)}{p(\mathbf{x}_a, \mathbf{x}_b)}\right)} d\mathbf{s}_a d\mathbf{s}_b\end{aligned} \tag{32}$$

When each agent learns the continuous factor independent of other agents, we have $q(\mathbf{s}_a, \mathbf{s}_b|\mathbf{c}_a, \mathbf{c}_b, \mathbf{x}_a, \mathbf{x}_b) = q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)q(\mathbf{s}_b|\mathbf{c}_b, \mathbf{x}_b)$. Equivalently, for independent samples $\mathbf{x}_a$ and $\mathbf{x}_b$, we have $q(\mathbf{c}_a, \mathbf{c}_b|\mathbf{x}_a, \mathbf{x}_b) = q(\mathbf{c}_a|\mathbf{x}_a)q(\mathbf{c}_b|\mathbf{x}_b)$. Hence,

$$\begin{aligned}
\Delta(a,b) = \log p(\mathbf{x}_a, \mathbf{x}_b) &+ \int_{\mathbf{s}_a} \sum_{\mathbf{c}_a} q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)q(\mathbf{c}_a|\mathbf{x}_a) \log \frac{q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)}{p(\mathbf{s}_a|\mathbf{c}_a)}\, d\mathbf{s}_a \\
&+ \int_{\mathbf{s}_b} \sum_{\mathbf{c}_b} q(\mathbf{s}_b|\mathbf{c}_b, \mathbf{x}_b)q(\mathbf{c}_b|\mathbf{x}_b) \log \frac{q(\mathbf{s}_b|\mathbf{c}_b, \mathbf{x}_b)}{p(\mathbf{s}_b|\mathbf{c}_b)}\, d\mathbf{s}_b \\
&- \int_{\mathbf{s}_a} \sum_{\mathbf{c}_a} q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)q(\mathbf{c}_a|\mathbf{x}_a) \log p(\mathbf{x}_a|\mathbf{s}_a, \mathbf{c}_a)\, d\mathbf{s}_a \\
&- \int_{\mathbf{s}_b} \sum_{\mathbf{c}_b} q(\mathbf{s}_b|\mathbf{c}_b, \mathbf{x}_b)q(\mathbf{c}_b|\mathbf{x}_b) \log p(\mathbf{x}_b|\mathbf{s}_b, \mathbf{c}_b)\, d\mathbf{s}_b \\
&+ \int_{\mathbf{s}_a} \int_{\mathbf{s}_b} \sum_{\mathbf{c}_a} \sum_{\mathbf{c}_b} q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)q(\mathbf{s}_b|\mathbf{c}_b, \mathbf{x}_b)q(\mathbf{c}_a|\mathbf{x}_a)q(\mathbf{c}_b|\mathbf{x}_b) \log \frac{q(\mathbf{c}_a|\mathbf{x}_a)q(\mathbf{c}_b|\mathbf{x}_b)}{p(\mathbf{c}_a, \mathbf{c}_b)}\, d\mathbf{s}_a d\mathbf{s}_b
\end{aligned} \tag{33}$$

$$\begin{aligned}
\Delta(a,b) = &-\mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[\mathbb{E}_{q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)}\left[\log p(\mathbf{x}_a|\mathbf{s}_a, \mathbf{c}_a)\right]\right] - \mathbb{E}_{q(\mathbf{c}_b|\mathbf{x}_b)}\left[\mathbb{E}_{q(\mathbf{s}_b|\mathbf{c}_b, \mathbf{x}_b)}\left[\log p(\mathbf{x}_b|\mathbf{s}_b, \mathbf{c}_b)\right]\right] \\
&+\mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a)\right)\right] + \mathbb{E}_{q(\mathbf{c}_b|\mathbf{x}_b)}\left[D_{KL}\left(q(\mathbf{s}_b|\mathbf{c}_b, \mathbf{x}_b)\|p(\mathbf{s}_b|\mathbf{c}_b)\right)\right] \\
&+\mathbb{E}_{q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)}\left[\mathbb{E}_{q(\mathbf{s}_b|\mathbf{c}_b, \mathbf{x}_b)}\left[D_{KL}\left(q(\mathbf{c}_a|\mathbf{x}_a)q(\mathbf{c}_b|\mathbf{x}_b)\|p(\mathbf{c}_a, \mathbf{c}_b)\right)\right]\right] + \log p(\mathbf{x}_a, \mathbf{x}_b)
\end{aligned} \tag{34}$$

Therefore, the variational lower bound for a pair of coupled VAE agents can be expressed as,

$$\begin{aligned}
\mathcal{L}_{\text{pair}}(a,b) \quad = \quad & \mathbb{E}_{q(\mathbf{s}_a, \mathbf{c}_a|\mathbf{x}_a)}\left[\log p(\mathbf{x}_a|\mathbf{s}_a, \mathbf{c}_a)\right] + \mathbb{E}_{q(\mathbf{s}_b, \mathbf{c}_b|\mathbf{x}_b)}\left[\log p(\mathbf{x}_b|\mathbf{s}_b, \mathbf{c}_b)\right] \\
&-\mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a)\right)\right] - \mathbb{E}_{q(\mathbf{c}_b|\mathbf{x}_b)}\left[D_{KL}\left(q(\mathbf{s}_b|\mathbf{c}_b, \mathbf{x}_b)\|p(\mathbf{s}_b|\mathbf{c}_b)\right)\right] \\
&-\mathbb{E}_{q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)}\left[\mathbb{E}_{q(\mathbf{s}_b|\mathbf{c}_b, \mathbf{x}_b)}\left[D_{KL}\left(q(\mathbf{c}_a|\mathbf{x}_a)q(\mathbf{c}_b|\mathbf{x}_b)\|p(\mathbf{c}_a, \mathbf{c}_b)\right)\right]\right]
\end{aligned} \tag{35}$$

which is equivalent to the loss function in Eq. 31, for $A = 2$.

To compute the joint distribution $p(\mathbf{c}_a, \mathbf{c}_b)$, here, we define an auxiliary continuous random variable e representing the mismatch (error) between $\mathbf{c}_a$ and $\mathbf{c}_b$ such that $\forall \mathbf{c}_a, \mathbf{c}_b \in \mathcal{S}^K$, and $0 < \epsilon \ll 1$,

$$p(\mathbf{c}_a, \mathbf{c}_b|e) = \begin{cases} 1 & |e - d^2(\mathbf{c}_a, \mathbf{c}_b)| < \epsilon/2 \\ 0 & \text{otherwise} \end{cases} \tag{36}$$

Here, $d(\mathbf{c}_a, \mathbf{c}_b)$ denotes the distance between $\mathbf{c}_a$ and $\mathbf{c}_b$ in the simplex $\mathcal{S}^K$, as a measure of mismatch between categorical variables. The random variable e is distributed according to an exponential probability density function with parameter λ i.e., $\forall e \geq 0$, $f(e, \lambda) = \lambda \exp(-\lambda e)$, where $\lambda > 0$. Accordingly, the joint categorical distribution can be represented as,

$$\begin{aligned}
p(\mathbf{c}_a, \mathbf{c}_b) \quad &= \quad \int p(\mathbf{c}_a, \mathbf{c}_b|e)p(e)de && (37) \\
&= \quad \int_{-\epsilon/2 + d^2(\mathbf{c}_a, \mathbf{c}_b)}^{\epsilon/2 + d^2(\mathbf{c}_a, \mathbf{c}_b)} f(e, \lambda)de = \epsilon f\left(d^2\left(\mathbf{c}_a, \mathbf{c}_b\right), \lambda\right) + E && (38)
\end{aligned}$$

where E is the error bound of the Midpoint integral rule. For given exponential function $f(e, \lambda)$, since $|f''(e, \lambda)| \leq \lambda^3$, $\forall e > 0$, the Midpoint approximation error is bounded by,

$$|E| \leq \frac{(\lambda\epsilon)^3}{24}. \tag{39}$$

Subsequently, the joint probability distribution is equivalent to:

$$p(\mathbf{c}_a, \mathbf{c}_b) = \epsilon\lambda \exp\left(-\lambda d^2\left(\mathbf{c}_a, \mathbf{c}_b\right)\right) + E \tag{40}$$

where ϵ and λ are arbitrarily constant values. We can approximate the joint distribution as follows.

$$p(\mathbf{c}_a, \mathbf{c}_b) \approx \epsilon\lambda \exp\left(-\lambda d^2\left(\mathbf{c}_a, \mathbf{c}_b\right)\right) \tag{41}$$

Thus, the last KL divergence in Eq. 35 can be approximated as,

$$\begin{aligned}
D_{\mathrm{KL}}\left(q(\mathbf{c}_a|\mathbf{x}_a)q(\mathbf{c}_a|\mathbf{x}_b)\|p(\mathbf{c}_a,\mathbf{c}_b)\right) = & \sum_{\mathbf{c}_a}\sum_{\mathbf{c}_b} q(\mathbf{c}_a|\mathbf{x}_a)q(\mathbf{c}_b|\mathbf{x}_b)\log\frac{q(\mathbf{c}_a|\mathbf{x}_a)q(\mathbf{c}_b|\mathbf{x}_b)}{p(\mathbf{c}_a,\mathbf{c}_b)} \qquad (42)\\
= & \; -H(\mathbf{c}_a|\mathbf{x}_a) - H(\mathbf{c}_b|\mathbf{x}_b) - \sum_{\mathbf{c}_a}\sum_{\mathbf{c}_b} q(\mathbf{c}_a|\mathbf{x}_a)q(\mathbf{c}_b|\mathbf{x}_b)\log p(\mathbf{c}_a,\mathbf{c}_b)\\
\approx & \; -H(\mathbf{c}_a|\mathbf{x}_a) - H(\mathbf{c}_b|\mathbf{x}_b) + \lambda\mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\mathbb{E}_{q(\mathbf{c}_b|\mathbf{x}_b)}\left[d^2\left(\mathbf{c}_a,\mathbf{c}_b\right)\right] - \log\epsilon\lambda, \; (43)
\end{aligned}$$

Therefore, the approximated variational cost for a pair of VAE agents can be written as follows:

$$\begin{aligned}
\mathcal{L}_{\mathrm{pair}}(a,b) = & \; \mathbb{E}_{q(\mathbf{s}_a,\mathbf{c}_a|\mathbf{x}_a)}\left[\log p(\mathbf{x}_a|\mathbf{s}_a,\mathbf{c}_a)\right] + \mathbb{E}_{q(\mathbf{s}_b,\mathbf{c}_b|\mathbf{x}_b)}\left[\log p(\mathbf{x}_b|\mathbf{s}_b,\mathbf{c}_b)\right]\\
& -\mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a)\right)\right] - \mathbb{E}_{q(\mathbf{c}_b|\mathbf{x}_b)}\left[D_{KL}\left(q(\mathbf{s}_b|\mathbf{c}_b,\mathbf{x}_b)\|p(\mathbf{s}_b|\mathbf{c}_b)\right)\right]\\
& +H(\mathbf{c}_a|\mathbf{x}_a) + H(\mathbf{c}_b|\mathbf{x}_b) - \lambda\mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\mathbb{E}_{q(\mathbf{c}_b|\mathbf{x}_b)}\left[d^2\left(\mathbf{c}_a,\mathbf{c}_b\right)\right] \qquad (44)
\end{aligned}$$

Now, by extending $\mathcal{L}_{pair}$ from two agents to A agents, in which there are $\binom{A}{2}$ paired networks, the total loss function for A agents can be written as

$$\begin{aligned}
\mathcal{L}_{\mathrm{cpl}} = & \sum_{a=1}^{A-1}\sum_{b=a+1}^{A}\mathcal{L}_{\mathrm{pair}}(a,b)\\
= & \sum_{a=1}^{A}(A-1)\mathbb{E}_{q(\mathbf{s}_a,\mathbf{c}_a|\mathbf{x}_a)}\left[\log p(\mathbf{x}_a|\mathbf{s}_a,\mathbf{c}_a)\right] - (A-1)\mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a)\right)\right]\\
& +\sum_{a<b} H(\mathbf{c}_a|\mathbf{x}_a) + H(\mathbf{c}_b|\mathbf{x}_b) - \lambda\mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\mathbb{E}_{q(\mathbf{c}_b|\mathbf{x}_b)}\left[d^2\left(\mathbf{c}_a,\mathbf{c}_b\right)\right]. \qquad (45)
\end{aligned}$$

E Proof of Proposition 2

In this section, we first briefly review some critical definitions in *Aitchison geometry*. Then, to support the proof of Proposition 2, here we introduce Lemma 1 and Propositions 3 and 4.

According to Aitchison geometry, a simplex of K parts can be considered as a vector space $(\mathcal{S}^K,\oplus,\otimes)$, in which $\oplus$ and $\otimes$ corresponds to *perturbation* and *power* operations, respectively, as follows.

$$Perturbation:\ \forall\mathbf{x},\mathbf{y}\in\mathcal{S}^K,\ \mathbf{x}\oplus\mathbf{y} = \mathcal{C}\left(x_1y_1,\ldots,x_Ky_K\right)$$

$$Power:\ \forall\mathbf{x}\in\mathcal{S}^K\ and\ \forall\alpha\in\mathbb{R},\ \alpha\otimes\mathbf{x} = \mathcal{C}\left(x_1^{\alpha},\ldots,x_K^{\alpha}\right)$$

where $\mathcal{C}$ denotes the closure operation as follows.

$$\mathcal{C}(\mathbf{x}) = \left(\frac{cx_1}{\sum_{k=1}^{K}x_k},\ldots,\frac{cx_K}{\sum_{k=1}^{K}x_k}\right).$$

In the simplex vector space, for any $\mathbf{x},\mathbf{y}\in\mathcal{S}^K$, the distance is defined as,

$$d_{\mathcal{S}^K}\left(\mathbf{x},\mathbf{y}\right) = \left(\frac{1}{K}\sum_{i<j}\left(\log\frac{x_i}{x_j}-\log\frac{y_i}{y_j}\right)^2\right)^{1/2}. \qquad (46)$$

Furthermore, Aitchison has introduced *centered-logratio* transformation (CLR), which is an isometric transformation from a simplex to a $K-$dimensional real space, $clr(\mathbf{x})\in\mathbb{R}^K$. The CLR transformation involves the logratio of each x_k over geometric means in the simplex as follows.

$$clr(\mathbf{x}) = \left(\log\frac{x_1}{g(\mathbf{x})},\ldots,\log\frac{x_K}{g(\mathbf{x})}\right). \qquad (47)$$

where $g(\mathbf{x}) = \left(\prod_{k=1}^{K} x_k\right)^{1/K}$ and $\sum_{k=1}^{K} \log \frac{x_k}{g(\mathbf{x})} = 0$.

Since CLR is an isometric transformation, we have

$$\begin{aligned} d_{\mathcal{S}^K}(\mathbf{x}, \mathbf{y}) &= d_{\mathbb{R}^K}(clr(\mathbf{x}), clr(\mathbf{y})) \\ &= \|clr(\mathbf{x}) - clr(\mathbf{y})\|_2 \end{aligned}$$

The algebraic-geometric definition of $\mathcal{S}^K$ satisfies standard properties, such as

$$d_{\mathcal{S}^K}(\mathbf{x} \oplus \boldsymbol{u}, \mathbf{y} \oplus \boldsymbol{u}) = d_{\mathcal{S}^K}(\mathbf{x} \ominus \boldsymbol{u}, \mathbf{y} \ominus \boldsymbol{u}) = d_{\mathcal{S}^K}(\mathbf{x}, \mathbf{y}) \tag{48}$$

where $\boldsymbol{u} \in \mathcal{S}^K$ could be any arbitrary vector in the simplex.

Lemma 1. *Given a set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in \mathcal{S}^K$ where $\mathcal{S}^K$ is a simplex of K parts, then*

$$clr(\mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \cdots \oplus \mathbf{x}_N) = clr(\mathbf{x}_1) + clr(\mathbf{x}_2) + \cdots + clr(\mathbf{x}_N).$$

Proof. According to Aitchison geometry, addition of vectors in the simplex is defined as,

$$\mathbf{x}_1 \oplus \cdots \oplus \mathbf{x}_N = \left(\frac{\prod_{n=1}^{N} x_{n_1}}{c_N}, \ldots, \frac{\prod_{n=1}^{N} x_{n_K}}{c_N} \right) \tag{49}$$

where $c_N = \sum_{k=1}^{K} \prod_{n=1}^{N} x_{n_k}$.

By applying the centered-logratio transformation, we have

$$clr(\mathbf{x}_1 \oplus \cdots \oplus \mathbf{x}_N) = \left(\log \frac{\prod_{n=1}^{N} x_{n_1}}{\delta_{K,N}}, \ldots, \frac{\prod_{n=1}^{N} x_{n_K}}{\delta_{K,N}} \right) \tag{50}$$

where $\delta_{K,N} = c_N \left(\prod_{k=1}^{K} \frac{\prod_{n=1}^{N} x_{n_k}}{c_N} \right)^{1/K} = \left(\prod_{k=1}^{K} \prod_{n=1}^{N} x_{n_k} \right)^{1/K}$.

Now, we can rewrite Eq. 50 as,

$$\begin{aligned} clr(\mathbf{x}_1 \oplus \cdots \oplus \mathbf{x}_N) &= \left(\log \frac{x_{1_1} \ldots x_{N_1}}{\left(\prod_k x_{1_k}\right)^{1/K} \cdots \left(\prod_k x_{N_k}\right)^{1/K}}, \ldots, \log \frac{x_{1_K} \ldots x_{N_K}}{\left(\prod_k x_{1_k}\right)^{1/K} \cdots \left(\prod_k x_{N_k}\right)^{1/K}} \right) \\ &= \left(\sum_n \log \frac{x_{n_1}}{\left(\prod_k x_{n_k}\right)^{1/K}}, \ldots, \sum_n \log \frac{x_{n_K}}{\left(\prod_k x_{n_k}\right)^{1/K}} \right) \\ &= clr(x_1) + \cdots + clr(x_N) \end{aligned} \tag{51}$$

□

Proposition 3. *Given vectors $\mathbf{x}, \mathbf{y}, \mathbf{u}_x, \mathbf{u}_y \in \mathcal{S}^K$ where $\mathcal{S}^K$ is a simplex of K parts, then*

$$d^2_{\mathcal{S}^K}\left(\mathbf{x} \oplus \mathbf{u}_x, \mathbf{y} \oplus \mathbf{u}_y\right) - d^2_{\mathcal{S}^K}\left(\mathbf{x}, \mathbf{y}\right) \leq K\tau^2 - \frac{\Delta^2}{K}$$

where $\tau = \max_k \{\log u_{x_k} - \log u_{y_k}\}$ and $\Delta = \sum_k (\log u_{x_k} - \log u_{y_k})$.

Proof. According to Aitchison geometry, the distance between two vectors $\mathbf{x}, \mathbf{y} \in \mathcal{S}^K$ is defined as,

$$d^2_{\mathcal{S}^K}\left(\mathbf{x}, \mathbf{y}\right) = \|clr(\mathbf{x}) - clr(\mathbf{y})\|_2^2$$

Then,

$$d^2_{\mathcal{S}^K}\left(\mathbf{x} \oplus \mathbf{u}_x, \mathbf{y} \oplus \mathbf{u}_y\right) = \|clr(\mathbf{x} \oplus \mathbf{u}_x) - clr(\mathbf{y} \oplus \mathbf{u}_y)\|_2^2$$

According to Lemma 1,

$$\begin{aligned}
d^2_{\mathcal{S}^K}\left(\mathbf{x} \oplus \mathbf{u}_x, \mathbf{y} \oplus \mathbf{u}_y\right) &= \|\left(clr\left(\mathbf{x}\right) - clr\left(\mathbf{y}\right)\right) + \left(clr\left(\mathbf{u}_x\right) - clr\left(\mathbf{u}_y\right)\right)\|_2^2 \\
&= \|clr\left(\mathbf{x}\right) - clr\left(\mathbf{y}\right)\|_2^2 + \|clr\left(\mathbf{u}_x\right) - clr\left(\mathbf{u}_y\right)\|_2^2 + \\
&\quad \left(clr\left(\mathbf{x}\right) - clr\left(\mathbf{y}\right)\right)^T \left(clr\left(\mathbf{u}_x\right) - clr\left(\mathbf{u}_y\right)\right) + \left(clr\left(\mathbf{u}_x\right) - clr\left(\mathbf{u}_y\right)\right)^T \left(clr\left(\mathbf{x}\right) - clr\left(\mathbf{y}\right)\right) \\
&= d^2_{\mathcal{S}^K}\left(\mathbf{x}, \mathbf{y}\right) + d^2_{\mathcal{S}^K}\left(\mathbf{u}_x, \mathbf{u}_y\right) + 2\sum_{k=1}^{K}\left(\log\frac{x_k}{g(\mathbf{x})} - \log\frac{y_k}{g(\mathbf{y})}\right)\left(\log\frac{u_{x_k}}{g(\mathbf{u}_x)} - \log\frac{u_{y_k}}{g(\mathbf{u}_y)}\right)
\end{aligned} \tag{52}$$

For simplicity, let's define $d_1^2 := d^2_{\mathcal{S}^K}\left(\mathbf{x}, \mathbf{y}\right)$ and $d_2^2 := d^2_{\mathcal{S}^K}\left(\mathbf{x} \oplus \mathbf{u}_x, \mathbf{y} \oplus \mathbf{u}_y\right)$, then

$$\begin{aligned}
d_2^2 - d_1^2 &= d^2_{\mathcal{S}^K}\left(\mathbf{u}_x, \mathbf{u}_y\right) + 2\sum_{k=1}^{K}\left(\log\frac{x_k}{g(\mathbf{x})} - \log\frac{y_k}{g(\mathbf{y})}\right)\left(\log\frac{u_{x_k}}{g(\mathbf{u}_x)} - \log\frac{u_{y_k}}{g(\mathbf{u}_y)}\right) \\
&= d^2_{\mathcal{S}^K}\left(\mathbf{u}_x, \mathbf{u}_y\right) + 2\sum_{k=1}^{K}\log\frac{x_k}{g(\mathbf{x})}\left(\log\frac{u_{x_k}}{u_{y_k}} - \log\frac{g(\mathbf{u}_x)}{g(\mathbf{u}_y)}\right) - \\
&\quad 2\sum_{k=1}^{K}\log\frac{y_k}{g(\mathbf{y})}\left(\log\frac{u_{x_k}}{u_{y_k}} - \log\frac{g(\mathbf{u}_x)}{g(\mathbf{u}_y)}\right)
\end{aligned} \tag{53}$$

Moreover, we know that $\log\frac{u_{x_k}}{u_{y_k}} \leq \tau$ and $\log\frac{g(\mathbf{u}_x)}{g(\mathbf{u}_y)} = \log\frac{\left(\prod_k \mathbf{u}_{x_k}\right)^{1/K}}{\left(\prod_k \mathbf{u}_{y_k}\right)^{1/K}} = \frac{1}{K}\sum_k \log\frac{u_{x_k}}{u_{y_k}} = \frac{\Delta}{K}$, then

$$\begin{aligned}
d_2^2 - d_1^2 &= d^2_{\mathcal{S}^K}\left(\mathbf{u}_x, \mathbf{u}_y\right) + 2\sum_{k=1}^{K}\log\frac{u_{x_k}}{u_{y_k}}\left(\log\frac{x_k}{g(\mathbf{x})} - \log\frac{y_k}{g(\mathbf{y})}\right) - \frac{2\Delta}{K}\sum_{k=1}^{K}\left(\log\frac{x_k}{g(\mathbf{x})} - \log\frac{y_k}{g(\mathbf{y})}\right) \\
&\leq d^2_{\mathcal{S}^K}\left(\mathbf{u}_x, \mathbf{u}_y\right) + 2\left(\tau - \frac{\Delta}{K}\right)\left(\sum_k \log\frac{x_k}{g(\mathbf{x})} - \sum_k \log\frac{y_k}{g(\mathbf{y})}\right)
\end{aligned} \tag{54}$$

Since CLR is a zero-mean transformation, $\sum_k \log\frac{x_k}{g(\mathbf{x})} = 0$ and $\sum_k \log\frac{y_k}{g(\mathbf{y})} = 0$. Therefore,

$$d_2^2 - d_1^2 \quad \leq d^2_{\mathcal{S}^K}\left(\mathbf{u}_x, \mathbf{u}_y\right) \tag{55}$$

In addition, we have

$$\begin{aligned}
d^2_{\mathcal{S}^K}\left(\mathbf{u}_x, \mathbf{u}_y\right) &= \sum_{k=1}^{K}\left(\log\frac{u_{x_k}}{u_{y_k}} - \log\frac{g(\mathbf{u}_x)}{g(\mathbf{u}_y)}\right)^2 \\
&= \sum_{k=1}^{K}\left(\log\frac{u_{x_k}}{u_{y_k}}\right)^2 + \sum_{k=1}^{K}\left(\log\frac{g(\mathbf{u}_x)}{g(\mathbf{u}_y)}\right)^2 - 2\log\frac{g(\mathbf{u}_x)}{g(\mathbf{u}_y)}\sum_{k=1}^{K}\log\frac{u_{x_k}}{u_{y_k}} \\
&\leq K\tau^2 - \frac{\Delta^2}{K}
\end{aligned} \tag{56}$$

By inserting Eq. 56 in Eq. 55, we will have

$$d_2^2 - d_1^2 \quad \leq K\tau^2 - \frac{\Delta^2}{K} \tag{57}$$

□

Proposition 4. *Given samples $\mathbf{x}, \mathbf{y} \in \mathcal{S}^K$, where $\mathcal{S}^K$ is a simplex of K parts, we have*

$$0 \leq d_{\mathbf{u}}^2\left(\mathbf{x}, \mathbf{y}\right) - d_{\mathcal{S}^K}^2\left(\mathbf{x} \oplus \mathbf{u}_x, \mathbf{y} \oplus \mathbf{u}_y\right) \leq \frac{1}{K}(\tau_1 + \tau_2)^2$$

where $d_{\mathbf{u}}^2\left(\mathbf{x}, \mathbf{y}\right) = \sum\limits_k \left(\log x_k u_{x_k} - \log y_k u_{y_k}\right)^2$, $\tau_1 = \max\limits_k\{\log u_{x_k} - \log u_{y_k}\}$, and $\tau_2 = \max\limits_k\{\log x_k - \log y_k\}$.

Proof.

$$\begin{aligned} d_{\mathcal{S}^K}^2\left(\mathbf{x} \oplus \mathbf{u}_x, \mathbf{y} \oplus \mathbf{u}_y\right) \quad &= \sum_{k=1}^{K}\left(\log x_k u_{x_k} - \log y_k u_{y_k} - \frac{1}{K}\log\prod_k \frac{x_k u_{x_k}}{y_k u_{y_k}}\right)^2 \\ &= \sum_{k=1}^{K}\left(\log x_k u_{x_k} - \log y_k u_{y_k} - \frac{1}{K}\sum_k \log \frac{x_k u_{x_k}}{y_k u_{y_k}}\right)^2 \\ &= \sum_{k=1}^{K}\left(\log x_k u_{x_k} - \log y_k u_{y_k} - D\right)^2 \end{aligned} \tag{58}$$

where $D = \frac{1}{K}\sum\limits_k \left(\log x_k u_{x_k} - \log y_k u_{y_k}\right)$. Hence,

$$\begin{aligned} d_{\mathcal{S}^K}^2\left(\mathbf{x} \oplus \mathbf{u}_x, \mathbf{y} \oplus \mathbf{u}_y\right) \quad &= \sum_{k=1}^{K}\left(\log x_k u_{x_k} - \log y_k u_{x_k}\right)^2 + KD^2 - 2D\sum_{k=1}^{K}\left(\log x_k u_{x_k} - \log y_k u_{y_k}\right) \\ &= d_{\mathbf{u}}^2\left(\mathbf{x}, \mathbf{y}\right) - KD^2 \end{aligned}$$

$$d_{\mathbf{u}}^2\left(\mathbf{x}, \mathbf{y}\right) = d_{\mathcal{S}^K}^2\left(\mathbf{x} \oplus \mathbf{u}_x, \mathbf{y} \oplus \mathbf{u}_y\right) + KD^2 \tag{59}$$

Since $KD^2 \geq 0$, we have $d_{\mathbf{u}}^2\left(\mathbf{x}, \mathbf{y}\right) \geq d_{\mathcal{S}^K}^2\left(\mathbf{x} \oplus \mathbf{u}_x, \mathbf{y} \oplus \mathbf{u}_y\right)$.

Moreover we know that $\forall k,\ \log\frac{u_{x_k}}{u_{y_k}} \leq \tau_1$ and $\log\frac{x_k}{y_k} \leq \tau_2$, then

$$\begin{aligned} d_{\mathbf{u}}^2\left(\mathbf{x}, \mathbf{y}\right) - d_{\mathcal{S}^K}^2\left(\mathbf{x} \oplus \mathbf{u}_x, \mathbf{y} \oplus \mathbf{u}_y\right) \quad &= \frac{1}{K}\left(\sum_k\left(\log\frac{x_k}{y_k} + \log\frac{u_{x_k}}{u_{y_k}}\right)\right)^2 \\ &\leq \frac{1}{K}(\tau_1 + \tau_2)^2 \end{aligned} \tag{60}$$

□

Proposition 2. *Suppose $\mathbf{c}_{a_n}, \mathbf{c}_{b_n} \in \mathcal{S}^K$, where $\mathcal{S}^K$ is a simplex of K parts and n is the sample index. If $d\left(\mathbf{c}_{a_n}, \mathbf{c}_{b_n}\right)$ denotes the Aitchison distance, then*

$$d_\sigma^2\left(\mathbf{c}_{a_n}, \mathbf{c}_{b_n}\right) - d^2\left(\mathbf{c}_{a_n}, \mathbf{c}_{b_n}\right) \leq \frac{1}{K}\left(\left(\tau_{\mathbf{c}} + \tau_\sigma\right)^2 + K^2\tau_\sigma^2 - \Delta_\sigma^2\right)$$

where $\tau_{\mathbf{c}} = \max\limits_k\{\log c_{a_{n_k}} - \log c_{b_{n_k}}\}$, $\tau_\sigma = \max\limits_k\{(\sigma_{a_k}^{-1} - 1)\log c_{a_{n_k}} - (\sigma_{b_k}^{-1} - 1)\log c_{b_{n_k}}\}$, and $\Delta_\sigma = \sum\limits_k(\sigma_{a_k}^{-1} - 1)\log c_{a_{n_k}} - (\sigma_{b_k}^{-1} - 1)\log c_{b_{n_k}}$.

Proof. In Propositions 3 and 4, by considering $\mathbf{x} = \mathbf{c}_{a_n}$, $\mathbf{y} = \mathbf{c}_{b_n}$, $\mathbf{u}_x = \mathbf{u}_a = \left(\frac{c_{a_{n_1}}^{(\sigma_{a_1}^{-1}-1)}}{\gamma_a}, \ldots, \frac{c_{a_{n_K}}^{(\sigma_{a_K}^{-1}-1)}}{\gamma_a} \right)$, and $\mathbf{u}_y = \mathbf{u}_b = \left(\frac{c_{b_{n_1}}^{(\sigma_{b_1}^{-1}-1)}}{\gamma_b}, \ldots, \frac{c_{b_{n_K}}^{(\sigma_{b_K}^{-1}-1)}}{\gamma_b} \right)$, where $\gamma_a = \sum_k c_{a_{n_k}}^{(\sigma_{a_k}^{-1}-1)}$ and $\gamma_b = \sum_k c_{b_{n_k}}^{(\sigma_{b_k}^{-1}-1)}$, we have

$$d^2_{\mathcal{S}^K}\left(\mathbf{c}_a \oplus \mathbf{u}_a, \mathbf{c}_b \oplus \mathbf{u}_b\right) \quad = \sum_{k=1}^{K} \left(\sigma_{a_k}^{-1} \log c_{a_{n_k}} - \sigma_{b_k}^{-1} \log c_{b_{n_k}} - D \right)^2 \tag{61}$$

where $D = \frac{1}{K} \sum_k \left(\sigma_{a_k}^{-1} \log c_{a_{n_k}} - \sigma_{b_k}^{-1} \log c_{b_{n_k}} \right)$. Hence,

$$d^2_{\mathcal{S}^K}\left(\mathbf{c}_a \oplus \mathbf{u}_a, \mathbf{c}_b \oplus \mathbf{u}_b\right) - d^2_{\mathcal{S}^K}\left(\mathbf{c}_a, \mathbf{c}_b\right) \quad \leq K\tau_\sigma^2 - \frac{\Delta_\sigma^2}{K} \tag{62}$$

and

$$0 \leq d^2_\sigma\left(\mathbf{c}_a, \mathbf{c}_b\right) - d^2_{\mathcal{S}^K}\left(\mathbf{c}_a \oplus \mathbf{u}_a, \mathbf{c}_b \oplus \mathbf{u}_b\right) \quad \leq \frac{1}{K}(\tau_c + \tau_\sigma)^2 \tag{63}$$

Therefore,

$$d^2_\sigma\left(\mathbf{c}_a, \mathbf{c}_b\right) - d^2_{\mathcal{S}^K}\left(\mathbf{c}_a, \mathbf{c}_b\right) \quad \leq \frac{1}{K} \left((\tau_{\mathbf{c}} + \tau_\sigma)^2 + K^2 \tau_\sigma^2 - \Delta_\sigma^2 \right) \tag{64}$$

□

F MNIST DATASET ANALYSIS

A common assumption in "disentangling" the continuous and discrete factors of variability is the independence of the categorical and continuous latent variables, conditioned on data. Fig. 1 demonstrates that this assumption can be significantly violated for two commonly used, interpretable style variables, "angle" and "width," in the MNIST dataset.

Calculation of angle and width: We first calculate the inertia matrix for each sample by treating the image as a solid object with a mass distribution given by pixel brightness values. Then, we compute the principal axis of the image based on the inertia matrix. We report the angle between this vector and the vertical axis using the $[-\pi/2, \pi/2)$ range. To calculate the width, we project the image to the horizontal axis after aligning the principal axis with the vertical axis using the computed angle value. We report the support of this projected signal, normalized by the horizontal size of the image (here 28 pixels).

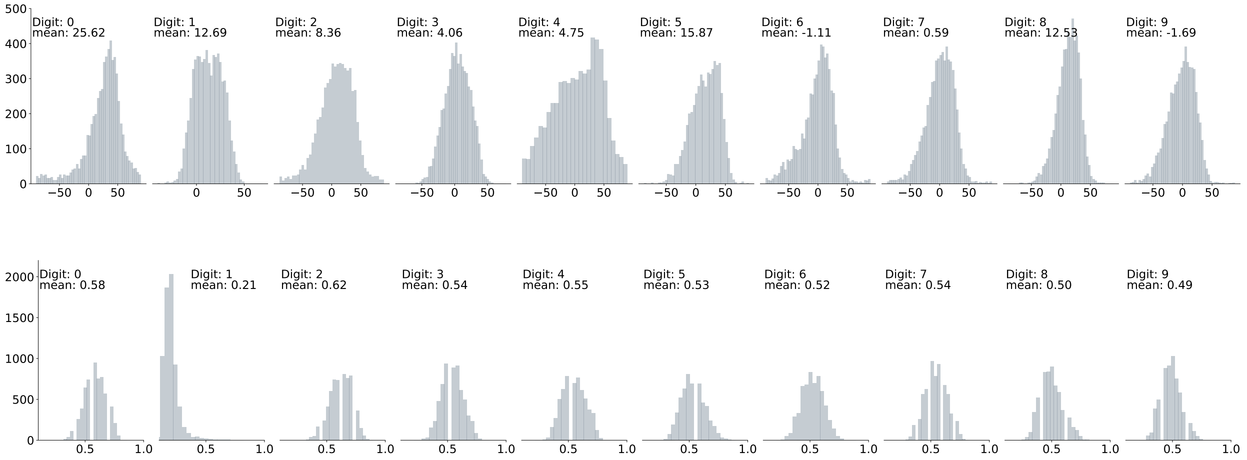


Figure 1: Histograms of angle and width for all digits in MNIST dataset. The empirical distributions of rotation (top) angle and character width (bottom) are illustrated. Comparing the reported mean values and the shape of the histograms demonstrates the dependency of the state variable on the digit type.

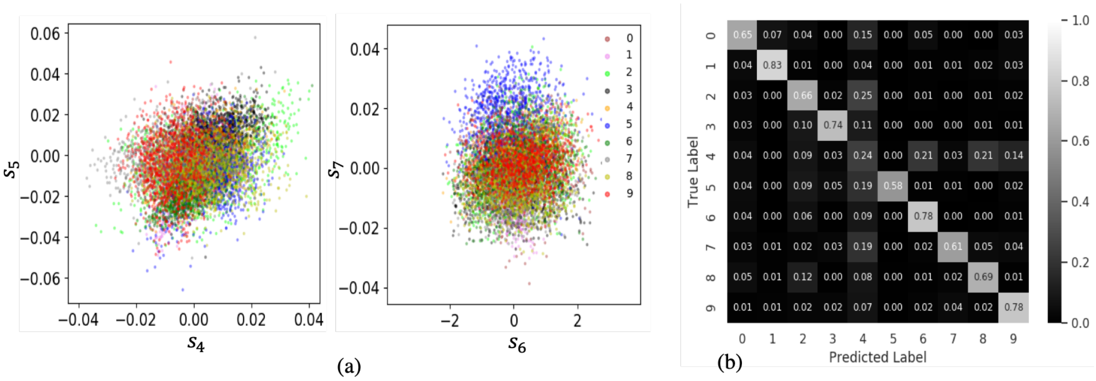


Figure 2: (a) 2-dimensional projections of the continuous variable obtained by JointVAE. Each dot represents a sample of the MNIST dataset and colors represent different digits. (b) Confusion matrix for MNIST digit clustering via GMM using only the continuous latent variable learned by JointVAE.

G Dependence of state and class label in JointVAE

We analyzed the effects of the dependency between the continuous and discrete latent factors on the results obtained by state-of-the-art methods for joint representation learning, e.g. JointVAE or CascadeVAE. These methods formulate the continuous and discrete variability as two independent factors such that the discrete factor is expected to determine the cluster to which each sample belongs, while the continuous factor represents the *class-independent* variability. In many applications, however, the assumption of a discrete-continuous dichotomy may not be satisfied. (Section F analyzes this assumption for the MNIST dataset.)

Fig. 2a illustrates four dimensions of the continuous latent variable $\mathbf{s}$ obtained by the JointVAE model for the MNIST dataset. Here, colors represent the digit type of each $\mathbf{s}$ sample. While the prior distribution is assumed to be Gaussian, the dependency of $\mathbf{s}|\mathbf{x}$ on the digit type, $\mathbf{c}$, is visible. To quantify this observation, we applied an unsupervised clustering method, i.e. Gaussian mixture model (GMM) with 10 clusters, to the continuous RV samples obtained from a JointVAE network trained for 150000 iterations. This unsupervised model achieved an overall clustering accuracy of 84%. Fig. 2b shows the results for individual digits, e.g. 83% for digit “1” (Fig. 2). Together, these results demonstrate the violation of the independence assumption for $q(\mathbf{s}|\mathbf{x})$ and $q(\mathbf{c}|\mathbf{x})$.

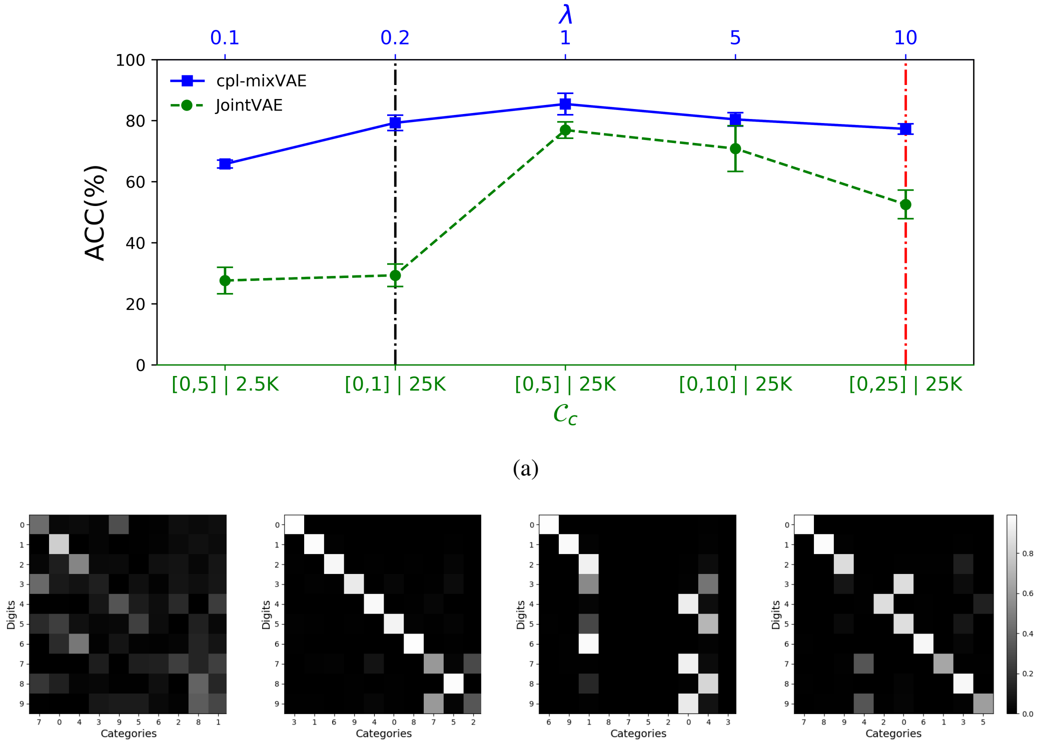
H Sensitivity of representation learning to the hyperparameters

The cpl-mixVAE framework, similar to other deep neural network approaches, has a regularization hyperparameter λ which controls coupling among a pair of autoencoder agents. In this section, we have conducted a series of experiments to assess the sensitivity of the cpl-mixVAE's performance to its coupling factor, in comparison with JointVAE which has four critical hyperparameters, two for the discrete and two for the continuous variables. Fig. 3 shows how the mixture representation performance changes for both JointVAE and cpl-mixVAE by changing their hyperparameters. For JointVAE, here, we only consider the channel capacity for the discrete variable, i.e. $\mathcal{C}_c$, which requires adjustment over training iterations.

Fig. 3a shows changes of the categorical assignment accuracy as a function of λ (for cpl-mixVAE) and $\mathcal{C}_c$ (for JointVAE). While cpl-mixVAE's performance is adequate for different values of the coupling factor, JointVAE is susceptible to the changes of the channel capacity factor. Although encoding channel capacity (as an estimation for mutual information) for each dataset with different latent space dimension and training iterations is computationally expensive, a main problem of using these hyperparameters happens when the learning of the model is highly sensitive to the channel capacity. For instance, Fig. 3b illustrates the categorical variables learned by JointVAE, when we reduced the maximum capacity from 5 to 1. Likewise, Fig. 3d shows a similar learning issue for JointVAE, when increasing the maximum channel from 5 to 25. In case of cpl-mixVAE, we can see that although obtaining the best performance requires parameter tuning, the model acceptably works with any empirical choice of $\lambda \in [0.1, 10]$.

I Data augmentation for scRNA-seq dataset

Generating augmented samples with the same class identity in the absence of within-class invariance is fairly challenging. In case of image datasets, e.g. MNIST, since there exist some intuitions about the identities of discrete and continuous



(a)

(b) JointVAE$([0,1]|25K)$ (c) cpl-mixVAE(0.2) (d) JointVAE$([0,25]|25K)$ (e) cpl-mixVAE(10)

Figure 3: (a) Effect of the coupling factor (λ) in cpl-mixVAE and the channel capacity ($\mathcal{C}_s$) in the JointVAE models. Reported values present the average accuracy of categorical assignment for 3 randomly initialized runs, over $15K$ training iteration, for the MNIST dataset. (b-c) Confusion matrices for JointVAE and cpl-mixVAE models, respectively corresponding to the hyperparameters marked by the dash-dotted black line. (d-e) Confusion matrices for JointVAE and cpl-mixVAE models, respectively corresponding to the hyperparameters marked by the dash-dotted red line.

variational factors, we can explicitly define a set of transformation such as rotation, translation, scaling, flipping, etc. that can be used as type-preserving augmentation. However, for non-image datasets, e.g. the single cell RNA-seq dataset, suggested alternative methods may fail to represent the class-conditioned variation in an unsupervised manner. Moreover, in case of biological datasets, learning an augmentation transformation is rather challenging due to the limited number of samples. Accordingly, in this section, we study the performance of the proposed data augmentation to investigate the extent to which our method is successful in realistic generation of the single-cell RNA-seq samples.

Fig. 4 illustrates a two-dimensional demonstrations for both original and augmented single cells samples. For two-dimensional visualizations, here, we used a regular autoencoder for non-linear dimension reduction. First, the autoencoder has been trained on the original cell samples. After learning a two-dimensional coordinate system for the original samples (left panel), we used the autoencoder to visualize the augmented samples (right panel). Comparing the visualizations demonstrates that the representations are qualitatively similar and all groups of cells sharing the same type (same color) are placed in similar locations. Additionally, in Fig. 5, we show the expression profiles of a subset of genes for an inhibitory cell. Again the qualitative comparison of the expression profiles reveals a similar variability across genes. Since the single cell RNA-seq data is heavily unbalanced, we additionally reported the data augmenter's performance at the single gene expression level. Fig. 6 illustrates the expression distribution of a subset of known genes for augmented cell samples (colorful histograms) compared with the original expressions (gray histograms).

J Architectures of the networks

Fig. 7 shows the network architecture for the 2-coupled mixVAE model applied on the benchmark datasets, e.g. MNIST and the scRNA-seq dataset, respectively. In this architecture, each VAE agent received an augmented copy of the original sample generated by the type-preserving augmentation. Fig. 8 illustrates the network design for type-preserving data augmentation for image datasets. For the scRNA-seq dataset, we used the similar design that is used for a single VAE agent, without mixture representation (only a continuous variable, with $|\mathbf{z}| = 10$).

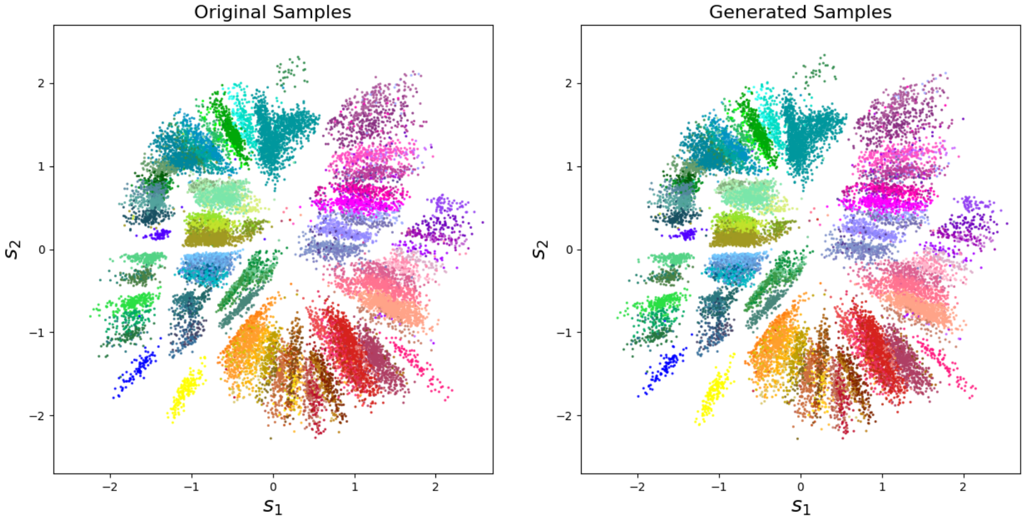


Figure 4: Evaluation of the generated sample by the proposed type-preserving data augmenter. Both figures represent a low dimensional visualization of single cells samples that are obtained from a regular autoencoder that is only used for non-linear dimension reduction. Left panel shows the original scRNA-seq dataset in a two-dimensional space by means of two coordinates. Right panel visualizes the generated cell samples by the augmenter in the same coordinate system. Both visualizations are obtained for 22, 000 cell samples with 5, 000 genes, and 115 neuron types. The color code is assigned according to the proposed taxonomy in Tasic et al., 2018.

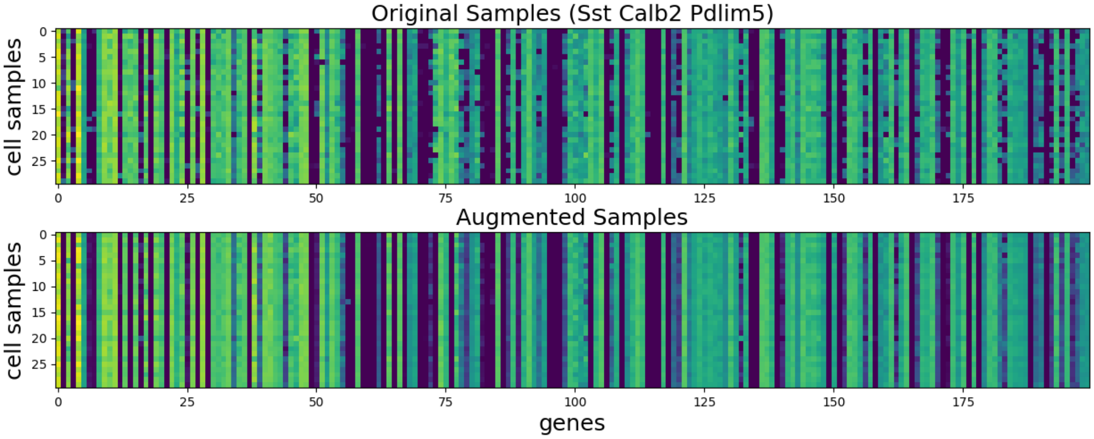


Figure 5: Qualitative comparison across the original and augmented gene expression profiles for an inhibitory Sst type cell.

For all dataset, To enhance the training process, we also applied 20% and 10% random dropout of the input sample and the state variable, respectively.

JointVAE† uses the same network architecture as a single agent of cpl-mixVAE. That is, it still uses the same loss function and learning procedure as JointVAE, but its convolutional layers are replaced by fully-connected layers, to demonstrate that these architecture choices do not explain the improvement achieved by cpl-mixVAE.

J.1 Training parameters for the MNIST dataset

Training details used for the MNIST dataset are listed as follows. For JointVAE† and JointVAE‡ model, we used the same training parameters as reported in (Dupont, 2018).

cpl-mixVAE

- Continuous and categorical variational factors: $\mathbf{s} \in \mathbb{R}^{10}$, $|\mathbf{c}| = 10$

- Batch size: 256
- Training epochs: 600
- Temperature for sampling from Gumbel-softmax distribution: 0.67
- Coupling weight, λ: 1
- Optimizer: Adam with learning rate 1e-4

JointVAE†, **JointVAE**‡

- Continuous and categorical variational factors: $\mathbf{s} \in \mathbb{R}^{10}$, $|\mathbf{c}| = 10$
- Batch size: 64
- Training epochs: 160
- Temperature for sampling from Gumbel-softmax distribution: 0.67
- $\gamma_{\mathbf{s}}$, $\gamma_{\mathbf{c}}$: 30
- $C_{\mathbf{s}}$, $C_{\mathbf{c}}$: Increased linearly from 0 to 5 in 25000 iterations
- Optimizer: Adam with learning rate 1e-4

J.2 Training parameters for the dSprites dataset

Training details used for the dSprites dataset are listed as follows.

cpl-mixVAE

- Continuous and categorical variational factors: $\mathbf{s} \in \mathbb{R}^{6}$, $|\mathbf{c}| = 3$
- Batch size: 256
- Training epochs: 600
- Temperature for sampling from Gumbel-softmax distribution: 0.67
- Coupling weight, λ: 10
- Optimizer: Adam with learning rate 1e-4

J.3 Training parameters for the scRNA-seq dataset

Training details used for the scRNA-seq dataset are listed as follows. For the JointVAE† model, we tried to set the parameters according to the reported numbers in (Dupont, 2018).

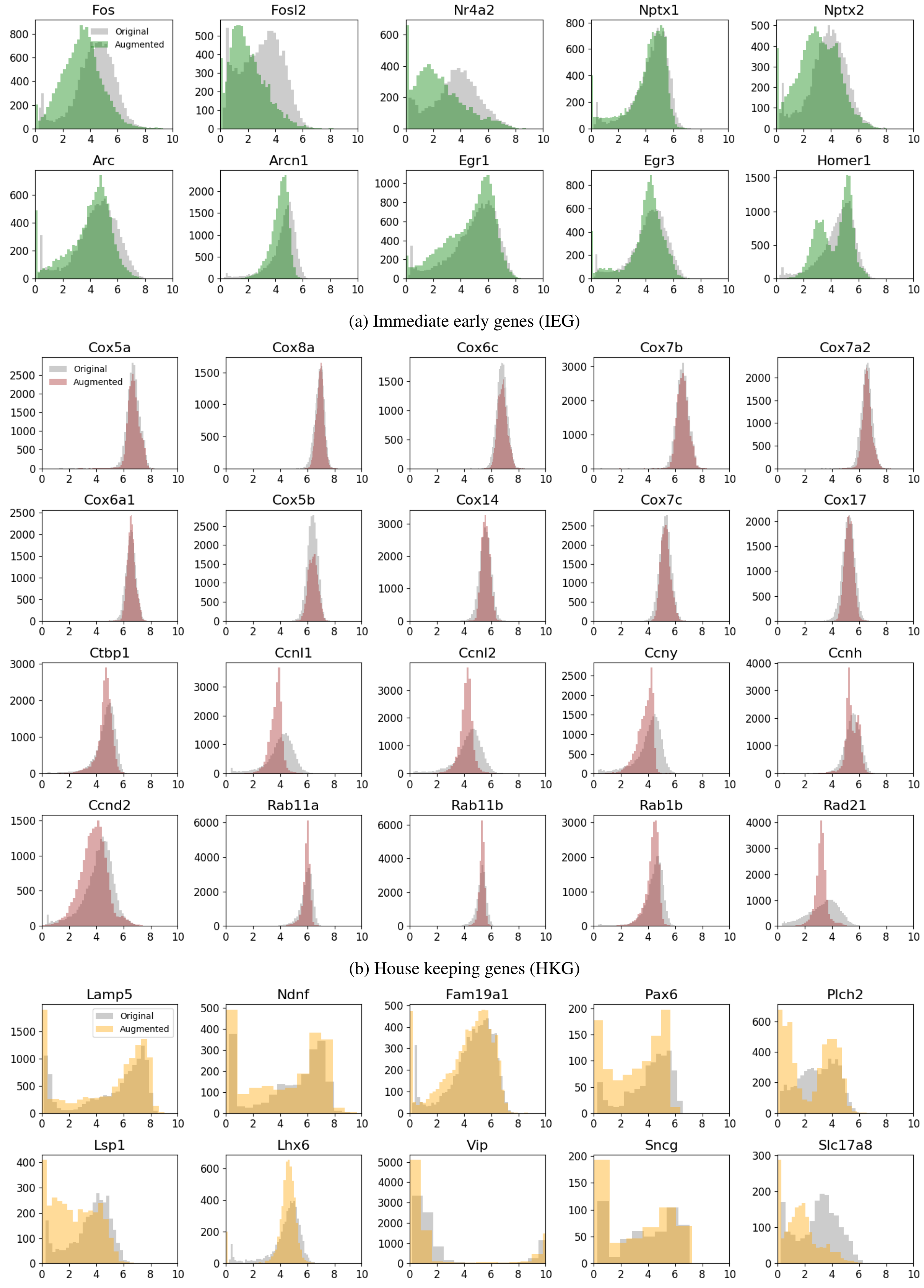
cpl-mixVAE

- Continuous and categorical variational factors: $\mathbf{s} \in \mathbb{R}^{2}$, $|\mathbf{c}| = 115$
- Batch size: 1000
- size of the last hidden layer, D: 10
- Training epochs: 10000
- Temperature for sampling from Gumbel-softmax distribution: 1
- Temperature for softmax function on $q(\mathbf{c}|\mathbf{x})$: 0.005 ($\approx$ 1/—z—)
- Coupling weight, λ: 1
- Optimizer: Adam with learning rate 1e-3

JointVAE†

- Continuous and categorical variational factors: $\mathbf{s} \in \mathbb{R}^{2}$, $|\mathbf{c}| = 115$
- Batch size: 1000
- size of the last hidden layer, D: 10

- Training epochs: 10000
- Temperature for sampling from Gumbel-softmax distribution: 0.005
- $\gamma_{\mathbf{s}}$, $\gamma_{\mathbf{c}}$: 100
- $C_{\mathbf{s}}$, $C_{\mathbf{c}}$: Increased linearly from 0 to 10 in 100000 iterations
- Optimizer: Adam with learning rate 1e-3

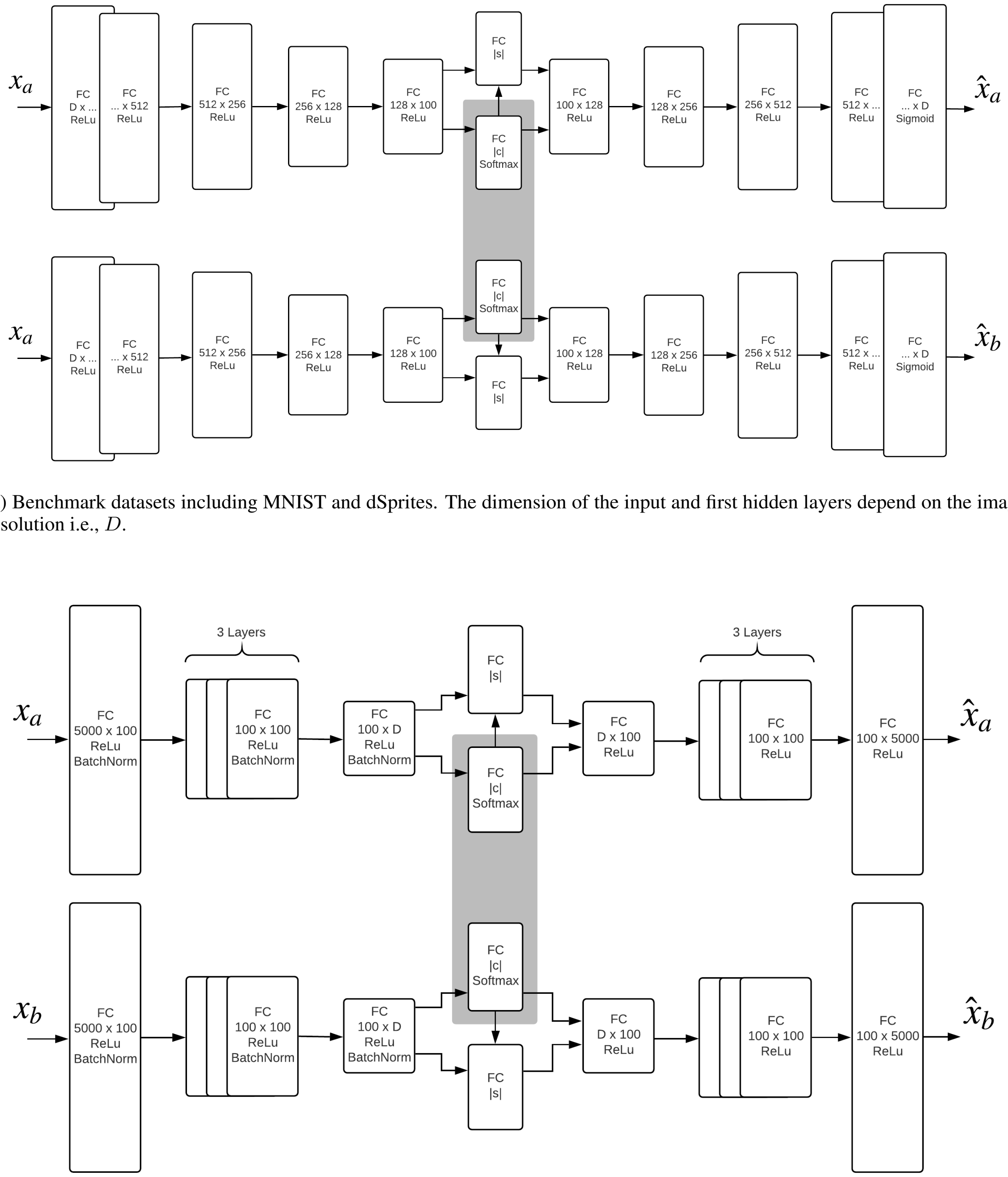


(a) Immediate early genes (IEG)

(b) House keeping genes (HKG)

(c) Marker Genes

Figure 6: Comparison between the distribution of genes in the original cell sample (gray color in all figures) and augmented samples for some biologically important subset of genes including (a) immediate early genes (green), (b) house keeping genes (brown), and (c) marker genes (yellow).



(a) Benchmark datasets including MNIST and dSprites. The dimension of the input and first hidden layers depend on the image resolution i.e., D.

(b) scRNA-seq dataset

Figure 7: cpl-mixVAE architectures including 2 agents.

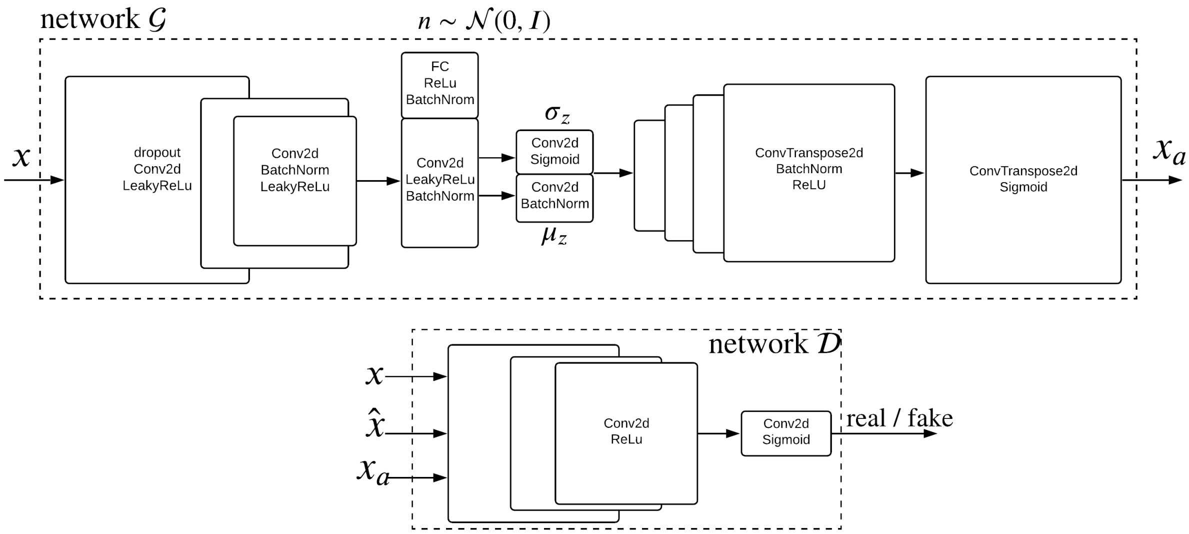


Figure 8: Network architecture for the proposed type-preserving data augmentation for image datasets.