Covariate-informed Representation Learning to Prevent Posterior Collapse of iVAE

Young-geun Kim
Department of Psychiatry and
Department of Biostatistics
Columbia University

Ying Liu
Department of Psychiatry and
Department of Biostatistics
Columbia University

Xuexin Wei
Department of Neuroscience and
Department of Psychology
The University of Texas at Austin

Abstract

The recently proposed identifiable variational autoencoder (iVAE) framework provides a promising approach for learning latent independent components (ICs). iVAEs use auxiliary covariates to build an identifiable generation structure from covariates to ICs to observations, and the posterior network approximates ICs given observations and covariates. Though the identifiability is appealing, we show that iVAEs could have local minimum solution where observations and the approximated ICs are independent given covariates. – a phenomenon we referred to as the posterior collapse problem of iVAEs. To overcome this problem, we develop a new approach, covariate-informed iVAE (CI-iVAE) by considering a mixture of encoder and posterior distributions in the objective function. In doing so, the objective function prevents the posterior collapse, resulting latent representations that contain more information of the observations. Furthermore, CI-iVAEs extend the original iVAE objective function to a larger class and finds the optimal one among them, thus having tighter evidence lower bounds than the original iVAE. Experiments on simulation datasets, EMNIST, Fashion-MNIST, and a large-scale brain imaging dataset demonstrate the effectiveness of our new method.

1 Introduction

Representation learning aims to identify low-dimensional latent representation that can be used to infer the structure of data generating process, to cluster observations by semantic meaning, and to detect anomalous patterns (Bengio et al., 2009, 2013). Recent progress on computer vision has shown that deep neural networks are effective in learning representations from rich high-dimensional data (LeCun et al., 2015). Now there is emerging interest in utilizing deep neural networks in scientific exploratory analysis to learn the representation of high-dimensional genetic/brain imaging data associated with phenotypes (Huang et al., 2017; Pinaya et al., 2019; Han et al., 2019; Qiang et al., 2020; Kim et al., 2021; Lu et al., 2021). In these scientific applications, autoencoder (AE, Bengio et al. 2007) and variational autoencoder (VAE, Kingma and Welling 2014) are popular representation learning methods. Compared to VAEs, the recently proposed identifiable VAE (iVAE, Khemakhem et al. 2020) provides appealing properties for scientific data analysis (Zhou and Wei, 2020; Schneider et al., 2022): (i) the identifiability of the learned representation; (ii) the representations are assumed to be associated with observed covariates. For example, in human health researches, (i) and (ii) are essential to identify latent independent components and learn representations associated with gender, age, and ethnicity, respectively. However, we find that when applying iVAEs to various datasets including a human brain imaging dataset, iVAEs sometimes converge to a bad local optimum where representations depend only on covariates (e.g., age, gender, and disease type), thus it is not a good representation for the observations (e.g., genetics, brain imaging data). Therefore, it is necessary to modify the objective function of iVAEs to learn better representations for scientific applications.

VAE consists of two networks: (i) the decoder network maps the prior latent independent components (ICs) to generate observations; (ii) the encoder network approximates the distribution of ICs given observations. iVAEs extend VAEs by introducing auxiliary covariates into the encoder and prior distributions, and construct identifiable data generation processes from covariates to ICs to observations. Inspired by the iVAE framework, recently, some identifiable generative models have been proposed. Kong et al. (2022) and Wang et al. (2022) founded identifiable generative models for domain-adaptation and causal inference, respectively. Zhou and Wei (2020) extended the data generation struc-

Preliminary work. Correspondence to: Ying Liu (Ying.Liu@nyspi.columbia.edu).

ture of iVAEs from continuous observation noise cases to Poisson. Sorrenson et al. (2020) explicitly imposed the inverse relation between encoders and decoders via general incompressible-flow networks (GIN), volume-preserving invertible neural networks. In this work, we focus on a problem of the current iVAEs implementation that could converge to a bad local optimum that lead to uninformative representations, and we propose a new approach that solves this problem by modifying objective functions.

Though the identifiability is appealing, we observe that representations from iVAEs ignore observations in many cases with experiments. The main reason is the Kullback-Leibler (KL) term in the evidence lower bounds (ELBOs, Bishop 2006) enforcing the posterior distribution of iVAEs to be prior distributions. This phenomenon is similar to the posterior collapse problem of VAEs where estimated ICs by encoders are independent of observations (Bowman et al., 2016; Lucas et al., 2019; He et al., 2019; Dai et al., 2020). A detailed review on the posterior collapse problem is provided in Section 2.2. We extend the notion of posterior collapse problem of VAEs to formulate this undesirable property of iVAEs, and coin it as the posterior collapse problem of iVAEs. With the formulation, we theoretically derive that iVAEs occur this problem under some conditions.

To overcome the limitation of iVAEs, we have developed a new method, the Covariate-Informed Identifiable VAE (CI-iVAE). Our new method leverages encoders in addition to the original posterior distribution considered in the previous iVAE to derive a new family of objective functions (ELBOs) for model fitting.[1] Crucially, in doing so, our objective function prevents the posterior collapse by modifying the KL term. CI-iVAEs extend the iVAE objective function to a larger class and finds the samplewise optimal one among them.

We demonstrate that our method can more reliably learn features of various synthetic datasets, two benchmark image datasets, EMNIST (Cohen et al., 2017) and Fashion-MNIST (Xiao et al., 2017), and a large-scale brain imaging dataset for adolescent mental health research. Especially, we apply our method and iVAEs to a brain imaging dataset, Adolescent Brain Cognitive Development (ABCD) study (Jernigan et al., 2018).[2] Our real data analysis on the ABCD dataset is the first application of identifiable neural networks in human brain imaging. Our method successfully learns representations from brain imaging data associated with the characteristics of subjects while the iVAE will learn non-informative representation due to posterior collapse problem on this real-data application.

Our contributions can be summarized as follows:

- We formulate the posterior collapse problem of iVAEs and derive that iVAEs may learn collapsed posteriors.
- We propose CI-iVAEs to learn better representations than iVAE by modifying the ELBO to prevent the posterior collapse.
- Experiments demonstrate that our method out performed iVAEs by preventing the posterior collapse problem.
- Our work is the first to learn ICs of human brain imaging with identifiable generative models.

All proofs of theoretical results are provided in Appendix A. Implementation details are provided in Appendix B.

2 Related Prior Work

2.1 Generative Autoencoders

Generative autoencoders are one of the prominent directions for representation learning (Higgins et al., 2017). They usually describe data generation processes with joint distributions of latent ICs and observations (Kingma et al., 2014), and optimize reconstruction error with penalty terms (Kingma and Welling, 2014). The reconstruction error is a distance between observations and their reconstruction results by encoders and decoders. For example, ELBO, the objective function of VAEs is a summation of the reconstruction probability (negative reconstruction error) and the KL divergence between encoder and prior distributions.

In the cases where auxiliary covariates are available, many conditional generative models have been proposed to incorporate covariates in generators in addition to latent variables. In conditional VAEs (Sohn et al., 2015) and conditional adversarial AEs (Makhzani et al., 2015), covariates are feed-forwarded by both encoder and decoder. Auxiliary classifiers for covariates are often applied to learn representations that can generate better results (Kameoka et al., 2018). The aforementioned methods have shown prominent results, but their models can learn the distribution of observations with many different prior distributions of latent variables, i.e., they are not *identifiable* (Khemakhem et al., 2020).

2.2 Posterior Collapse

The representations by VAEs are often poor due to the posterior collapse, which has been pointed as a practical drawback of VAEs and their variations (Kingma et al., 2016; Yang et al., 2017; Dieng et al., 2019). The posterior collapse refers to the phenomenon that the posterior converges to the prior. In this case, approximated ICs are independent of observations, i.e., representations lose the information in observations. One reason for this is that the KL divergence

[1] We distinguish encoders $q_\phi(z|x)$ and posterior $q_\phi(z|x,u)$ to avoid confusion where z, x, u, and ϕ indicate representations, observations, covariates, and network parameters, respectively. The posterior networks in iVAEs are different from encoders in usual VAEs.

[2] The ABCD dataset can be found at `https://abcdstudy.org`, held in the NIMH Data Archive (NDA).

term in the ELBO enforces the encoders to be close to prior distributions (Huang et al., 2018; Razavi et al., 2019).

A line of works has focused on posterior distributions and the KL term to alleviate this issue. He et al. (2019) aggressively optimized the encoder whenever the decoder is updated, Kim et al. (2018) introduced stochastic variational inference (Hoffman et al., 2013) to utilize instance-specific parameters for encoders, and Fu et al. (2019) monotonically increased the coefficient of the KL term from zero to ensure that posteriors are not collapsed in early stages. However, Dai et al. (2020) derived that VAEs sometimes have lower values of ELBOs than posterior collapse cases. It means that the surface of ELBOs naturally results in a bad local optima, posterior collapse cases. We will extend this theory, and show that iVAEs may result in local optima with collapsed posteriors, in which case the representations are independent to observations given covariates in Section 3.2. Recently, Wang et al. (2021) prevented the posterior collapse of VAEs with the latent variable identifiability, which refers to that distributions identify *latent variables* for given model parameters. It is important to note that latent variable identifiability is different from the model identifiability as discussed in the iVAE framework. The latter refers to that distributions identify *model parameters*. For brevity, we will use the term identifiability to refer to the model identifiability in this paper.

3 Proposed Method

3.1 Preliminaries

3.1.1 Basic Notations and Assumptions

We denote observations, covariates, and latent variables by $X \in \mathbb{R}^{d_X}$, $U \in \mathbb{R}^{d_U}$, and $Z \in \mathbb{R}^{d_Z}$, respectively. The dimension of latent variables is lower than that of observations, i.e., $d_Z < d_X$. For a given random variable (e.g., Z), its realization and probability density function (p.d.f.) is denoted by lower case (e.g., z) and p (e.g., $p(z)$), respectively. We distinguish encoder and posterior distributions and denote them by $q_\phi(z|x)$ and $q_\phi(z|x,u)$, respectively.

3.1.2 Identifiable Variational Autoencoders

The identifiability is an essential property to recover the true data generation structure and to conduct correct inference (Lehmann et al., 2005; Casella and Berger, 2021). A generative model is called identifiable if the distribution of generation results identifies parameters (Rothenberg, 1971; Koller and Friedman, 2009). The iVAE framework provides an appealing approach for learning latent ICs. The iVAE assumes the following data generation structure:

$$\begin{cases} Z|U \sim p_{T_0,\lambda_0}(z|u) \\ X = f_0(Z) + \epsilon \end{cases} \quad (1)$$

where Z denotes the IC (or *source*) and f_0 denotes the nonlinear mixing function. Here, p_{T_0,λ_0} is a conditionally factorial exponential family distribution with sufficient statistics T_0 and natural parameters λ_0, and ϵ is an observation noise. The iVAE models the mixing function with neural networks, a flexible nonlinear model, and its generation process is identifiable under certain conditions. Key components of iVAEs include label prior, decoder, and posterior networks. The label prior and decoder, respectively, models the conditional distribution of latent variables given covariates, $p_{T_0,\lambda_0}(z|u)$, and observations given latent variables, $p_{f_0}(x|z)$. The posterior networks $q_\phi(z|x,u)$ approximates the posterior distribution of the ICs, $p_{f_0,T_0,\lambda_0}(z|x,u)$. Again, we distinguish encoders $q_\phi(z|x)$ and posteriors $q_\phi(z|x,u)$ to avoid confusion. The posterior networks in iVAEs approximate distributions of ICs given observations and covariates, which is different from encoders in usual VAEs.

An essential condition on the label prior to ensure the identifiability is the conditionally factorial exponential family distribution assumption. The label prior network is denoted by $p_{T,\lambda}(z|u)$ and can be expressed as $p_{T,\lambda}(z|u) = \prod_{i=1}^{d_Z} p_{T_i,\lambda_i}(z_i|u)$ where $p_{T_i,\lambda_i}(z_i|u) = \exp\left(\lambda_i(u) \cdot T_i(z_i) - A(u) + B(z_i)\right)$ is the exponential family distribution with parameters $\lambda_i(u)$, sufficient statistics $T_i(z_i)$, and known functions A and B. We denote $T := (T_1, \ldots, T_{d_Z})$ and $\lambda := (\lambda_1, \ldots, \lambda_{d_Z})$. The decoder network is denoted by $p_f(x|z) := p(\epsilon = x - f(z))$ where f is the modeled mixing function and $p(\epsilon)$ is the p.d.f. of noise variables $\mathcal{E}$. With label prior and decoder networks, the data generation process can be expressed as $p_\theta(x,z|u) := p_f(x|z)p_{T,\lambda}(z|u)$ where $\theta = (f,T,\lambda)$ is all the parameters for the data generation process. The posterior network is denoted by $q_\phi(z|x,u)$. The encoder can be estimated by $q_\phi(z|x) = \int q_\phi(z|x,u)p(u|x)du$ or separately modeled. In the implementation, iVAEs model $p_{T,\lambda}(z|u)$ and $q_\phi(z|x,u)$ with Gaussian distributions.

Khemakhem et al. (2020) defined the model identifiability of the generation process by $p_\theta(x,z|u)$.

Definition 1. *(Identifiability, Khemakhem et al. 2020) The* $p_\theta(x,z|u)$ *is called identifiable if the following holds: for any* $\theta = (f,T,\lambda)$ *and* $\tilde{\theta} = (\tilde{f},\tilde{T},\tilde{\lambda})$, $p_\theta(x|u) = p_{\tilde{\theta}}(x|u)$ *implies* $\theta \sim \tilde{\theta}$. *Here,* $\theta \sim \tilde{\theta}$ *is defined as* $T(f^{-1}(x)) = \tilde{T}(\tilde{f}^{-1}(x))$ *up to a invertible affine transformation.*

With the identifiability, finding the maximum likelihood estimators (MLEs) implies learning the true mixing function and ICs in (1). Khemakhem et al. (2020) showed that the identifiability holds and the affine transformation is component-wise one-to-one transformations as in ICA if λ can make invertible matrix $(\lambda(u_1) - \lambda(u_0), \ldots, \lambda(u_{nk}) - \lambda(u_0))$ with some $nk+1$ distinct $u_0, \ldots, u_{nk}$, and some mild conditions hold.

The objective function of iVAEs, ELBO is with respect to (w.r.t.) the conditional log-likelihood of observations given covariates $\log p_\theta(x|u)$. The ELBO can be expressed as

$$\mathbb{E}_{q_\phi(z|x,u)} \log p_f(x|z) - \mathcal{D}_{\mathrm{KL}}(q_\phi(z|x,u)||p_{T,\lambda}(z|u)) \quad (2)$$

which equals to $\log p_\theta(x|u) - \mathcal{D}_{\text{KL}}(q_\phi(z|x,u)||p_\theta(z|x,u))$. When the space of $q_\phi(z|x,u)$ includes $p_\theta(z|x,u)$ for any θ, (2) can approximate $\log p_\theta(x|u)$, which justifies that iVAEs learn the ground-truth data generation structure (Khemakhem et al., 2020). However, we show in the next section that Equation (3) has a bad local optimum due to the KL term and iVAEs sometimes yield representations depending only on covariates by converging to this local optimum.

3.2 Motivation

In this section, we describe how the objective function used in previous iVAEs produces bad local optimums for posterior $q_\phi(z|x,u)$. We further show how modifying the objective function with encoders $q_\phi(z|x)$ can alleviate this problem.

We first describe the posterior collapse problem of VAEs, and then formulate the bad local optimums of (2). In the usual VAEs using only observations, the objective function is $\mathbb{E}_{q_\phi(z|x)} \log p_f(x|z) - \mathcal{D}_{\text{KL}}(q_\phi(z|x)||p(z))$ where $p(z)$ is the prior distribution. The posterior collapse of VAEs can be expressed as $q_\phi(z|x) = p(z)$, and one reason of this phenomenon is the KL term enforcing $q_\phi(z|x)$ be close to $p(z)$. Similarly, the KL term in (2) enforces posterior distributions $q_\phi(z|x,u)$ to be close to label prior $p_{T,\lambda}(z|u)$. We extend the notion of the posterior collapse of VAEs to formulate a bad local solution of (2), and coin it as the posterior collapse problem of iVAEs.

Definition 2. *(Posterior collapse of iVAEs) For a given dataset $\{(x_i,u_i)\}_{i=1}^n$, we call the posterior $q_\phi(z|x,u)$ in iVAEs is collapsed if $q_\phi(z|x_i,u_i) = p_{T,\lambda}(z|u_i)$ holds for all $i = 1,\dots,n$.*

In the following, we use the term posterior collapse to refer the posterior collapse of iVAEs. Under the posterior collapse, approximated ICs are independent of observations given covariates, i.e., we lose all the information in observations independent of covariates. Furthermore, we derive that the posterior collapse is a local optimum of (2) under some conditions, which is consistent with that the posterior collapse problem of VAEs is a local optimum of the objective function of VAEs (Dai et al., 2020). In the following theorem, we assume two conditions formulated by Dai et al. (2020): $(C1)$ the derivative of reconstruction error is Lipschitz continuous and $(C2)$ the reconstruction error is an increasing function w.r.t. the uncertainty of latent variables. A detailed formulation of $(C1)$ and $(C2)$ is provided in Appendix A.

Theorem 1. *Let $D = \{(x_i,u_i)\}_{i=1}^n$ be samples from (1) when $\epsilon \sim N(0,\gamma I)$ and the loss be the negative expectation of (2) over D. For any iVAEs satisfying $(C1)$ and $(C2)$, there is a posterior collapse case whose loss value is lower than that of the iVAEs when γ is sufficiently large.*

That is, the objective function of existing iVAEs evaluates posterior collapse cases as better solutions than iVAEs. Roughly speaking, when the noise of the observation, γ, is large, the KL term in (2) dominates the first term.

Our method modifies ELBOs to alleviate the posterior collapse problem of iVAEs. We note that the KL term $\mathcal{D}_{\text{KL}}(q_\phi(z|x,u)||p_{T,\lambda}(z|u))$ is a main reason of the posterior collapse and change $q_\phi(z|x,u)$ to linear mixtures of posterior $q_\phi(z|x,u)$ and encoder $q_\phi(z|x)$. For a motivating example, we can consider an alternative objective function (which can be shown to be an ELBO):

$$\mathbb{E}_{q_\phi(z|x)} \log p_f(x|z) - \mathcal{D}_{\text{KL}}(q_\phi(z|x)||p_{T,\lambda}(z|u)). \quad (3)$$

It is equivalent to $\log p_\theta(x|u) - \mathcal{D}_{\text{KL}}(q_\phi(z|x)||p_\theta(z|x,u))$. It can be viewed as using encoder $q_\phi(z|x)$ to approximate the posterior of label prior distributions $p_\theta(z|x,u)$. We name (2) and (3) by ELBOs with $q_\phi(z|x,u)$ and with $q_\phi(z|x)$, respectively. We derive that (3) prevents the posterior collapse problem of iVAEs. We say a label prior $p_{T,\lambda}$ non-trivial if $p_{T,\lambda}(z|u) \neq \int p_{T,\lambda}(z|u)p(u)du$ holds with positive probability w.r.t. $p(u)$, i.e., the label prior does not ignore covariates. All the label prior of iVAEs are non-trivial since λ should make invertible matrix $(\lambda(u_1) - \lambda(u_0), \dots, \lambda(u_{nk}) - \lambda(u_0))$ with some $nk+1$ distinct $u_0, \dots, u_{nk}$.

Proposition 1. *For any non-trivial label prior $p_{T,\lambda}$, $q_\phi(z|x) \neq p_{T,\lambda}(z|u)$ holds with positive probability w.r.t. $p(x,u)$.*

That is, the $q_\phi(z|x)$ can not collapse to non-trivial $p_{T,\lambda}(z|u)$ since it uses only observations. However, we derive that (3) can not approximate $\log p_\theta(x|u)$.

Proposition 2. *We assume that, for any θ, there is ϕ satisfying $q_\phi(z|x) = p_\theta(z|x)$ with probability 1 w.r.t. $p(x)$. For any θ forming non-trivial label prior,*

$$\max_\phi \mathbb{E}_{p(x,u)}\left(\mathbb{E}_{q_\phi(z|x)} \log p_f(x|z) - \mathcal{D}_{KL}(q_\phi(z|x)||p_{T,\lambda}(z|u))\right) < \mathbb{E}_{p(x,u)} \log p_\theta(x|u).$$

Thus, although the ELBO with $q_\phi(z|x)$ prevents the posterior collapse problem, it is strictly smaller than $\log p_\theta(x|u)$ even with the global optimum of ϕ. In contrast, the ELBO with $q_\phi(z|x,u)$ can approximate the log-likelihood if we find the global optimum based on the data, but it may converge to the posterior collapse cases. Thus, in practice, it is desirable to find a good balance between these two considerations. In the next section, we develop a new method by modifying the ELBOs to achieve such a goal.

3.3 Covariate-informed Identifiable VAE

In this section, we provide our method, CI-iVAE. We use the same network architecture of iVAEs described in Section 3.1.2 to inherit the identifiability of the likelihood model. Our key innovation is the development of a new class of objective functions (ELBOs) to prevent the posterior collapse.

Algorithm 1 Training CI-iVAEs

```
Input: Training samples $\{(x_i, u_i)\}_{i=1}^n$ and batch size $B$
Output: CI-iVAEs with ($f^*$, $T^*$, $\lambda^*$, $\phi^*$).
Initialize $(f, T, \lambda, \phi)$
While $(f, T, \lambda, \phi)$ did not converge Do
    Sample $\{(x_{i(b)}, u_{i(b)})\}_{b=1}^B$ from training samples
    Calculate samplewise optimal $\alpha^*(x_{i(b)}, u_{i(b)})$
    Update $(f, T, \lambda, \phi)$ by ascending
        $B^{-1} \sum_{b=1}^{B} \text{ELBO}_{\theta,\phi}(\alpha^*(x_{i(b)}, u_{i(b)}); x_{i(b)}, u_{i(b)})$
$(f^*, T^*, \lambda^*, \phi^*) \leftarrow (f, T, \lambda, \phi)$
```

We consider mixtures of distributions by encoders $q_\phi(z|x)$ and the posterior in the original iVAE, $q_\phi(z|x, u)$,

$$\{\alpha(x, u)q_\phi(z|x)+(1-\alpha(x, u))q_\phi(z|x, u)|\alpha(x, u) \in [0, 1]\}, \tag{4}$$

to derive ELBOs avoiding posterior collapse while approximating log-likelihoods. For simplicity, we use α to refer $\alpha(x, u)$, when there is no confusion. Any element in (4) can provide a lower bound of log-likelihood. We first formulate a set of ELBOs using (4) and then provide our method.

Proposition 3. *For any sample* (x, u), $\theta = (f, T, \lambda)$, ϕ, *and* $\alpha \in [0, 1]$, *ELBO*$_{\theta,\phi}(\alpha; x, u)$ *defined as*

$$\begin{aligned} &\mathbb{E}_{\alpha q_\phi(z|x)+(1-\alpha)q_\phi(z|x,u)} \log p_f(x|z) \\ &- \mathcal{D}_{KL}(\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x, u)||p_{T,\lambda}(z|u)) \end{aligned}$$

is a lower bound of $\log p_\theta(x|u)$.

Proposition 3 can be derived by performing variational inference with $\alpha(x, u)q_\phi(z|x) + (1 - \alpha(x, u))q_\phi(z|x, u)$. A set of ELBOs, $\{\text{ELBO}_{\theta,\phi}(\alpha; x, u)|\alpha \in [0, 1]\}$ is a continuum of ELBOs whose endpoints are ELBOs with $q_\phi(z|x, u)$ and with $q_\phi(z|x)$.

We next present our CI-iVAE method. For a given identifiable generative model, CI-iVAE uses covariates to find the samplewise optimal elements $\alpha^*(x, u) := \arg\max_{\alpha\in[0,1]}\text{ELBO}_{\theta,\phi}(\alpha; x, u)$ in (4) and utilizes it to maximize the tightest ELBOs. We refer to $\alpha^*(x, u)q_\phi(z|x) + (1 - \alpha^*(x, u))q_\phi(z|x, u)$ as the samplewise optimal posterior distributions. The objective function of the CI-iVAE method is given by

$$\text{ELBO}_{\theta,\phi}(\alpha^*(x, u); x, u), \tag{5}$$

where the KL term in (2) is changed to $\mathcal{D}_{\text{KL}}(\alpha^*(x, u)q_\phi(z|x) + (1 - \alpha^*(x, u))q_\phi(z|x, u)||p_{T,\lambda}(z|u))$ whose bad local solutions are $\alpha^*(x, u)q_\phi(z|x) + (1 - \alpha^*(x, u))q_\phi(z|x, u) = p_{T,\lambda}(z|u)$. We derive that the posterior collapse problem does not occur at this local solution.

Theorem 2. *For any* θ *forming non-trivial label prior and* ϕ, *if* $\alpha^*(x, u) > 0$ *and* $\alpha^*(x, u)q_\phi(z|x) + (1 - \alpha^*(x, u))q_\phi(z|x, u) = p_{T,\lambda}(z|u)$, *then* $q_\phi(z|x, u) \neq p_{T,\lambda}(z|u)$ *holds with positive probability.*

That is, even in the worst case where the KL term dominates (5), the $q_\phi(z|x, u)$ does not collapse to $p_{T,\lambda}(z|u)$.

Furthermore, we derive that our ELBO can approximate the log-likelihood.

Proposition 4. *We assume that, for any* θ, *there is* ϕ *satisfying* $q_\phi(z|x, u) = p_\theta(z|x, u)$ *with probability* 1 *w.r.t.* $p(x, u)$.[3] *Then,* $\max_\phi \mathbb{E}_{p(x,u)}$*ELBO*$_{\theta,\phi}(\alpha^*(x, u); x, u) = \mathbb{E}_{p(x,u)} \log p_\theta(x|u)$.

Proposition 4 implies that maximizing our ELBO is equivalent to learning MLEs with tighter lower bounds than (2). We also derive that the difference between our optimal ELBO and the ELBO of existing iVAE is significant under some conditions in Theorem 3 in Appendix A.

Though the CI-iVAE allows learning better representations, numerical approximation for $\alpha^*(x, u)$ in (5) may lead to burdensome computational costs. To overcome this technical obstacle, we provide an alternative expression of our ELBO that allows us to approximate $\alpha^*(x, u)$ within twice the computation time of the ELBO of the existing iVAE.

Proposition 5. *For any sample* (x, u), $\theta = (f, T, \lambda)$, ϕ, *and* $\alpha \in [0, 1]$, *ELBO*$_{\theta,\phi}(\alpha; x, u)$ *can be expressed as*

$$\begin{aligned} &\alpha \textit{ELBO}_{\theta,\phi}(1; x, u) + (1-\alpha)\textit{ELBO}_{\theta,\phi}(0; x, u) \\ &+ \alpha\mathcal{D}^{1-\alpha}_{\textit{Skew}}(q_\phi(z|x)||q_\phi(z|x, u)) + (1-\alpha)\mathcal{D}^{\alpha}_{\textit{Skew}}(q_\phi(z|x, u)||q_\phi(z|x)), \end{aligned}$$

where $\mathcal{D}^\alpha_{\textit{Skew}}$ *is the skew divergence defined as* $\mathcal{D}^\alpha_{\textit{Skew}}(p||q) := \mathcal{D}_{KL}(p||(1-\alpha)p + \alpha q)$ *(Lin, 1991).*

Details on using samplewise optimal posterior distributions to maximize (5) is described in Algorithm 1. The main bottleneck is on computing $\text{ELBO}_{\theta,\phi}(0; x, u)$, $\text{ELBO}_{\theta,\phi}(1; x, u)$, and conditional means and standard deviations of the label prior and encoder distributions, which requires roughly twice the computation time of the ELBO of iVAE. With Proposition 5 and reparametrization trick, for any α, the remaining process to compute $\text{ELBO}_{\theta,\phi}(\alpha; x, u)$ is completed in a short time.

4 Experiments

We have validated and applied our method on synthetic, EMNIST, Fashion-MNIST, and ABCD datasets. As in Zhou and Wei (2020), we model the posterior distribution by $q_\phi(z|x, u) \propto q_\phi(z|x)p_{T,\lambda}(z|u)$. The label prior $p_{T,\lambda}(z|u)$ and encoder $q_\phi(z|x)$ is modeled as Gaussian distributions.

[3] Again, Khemakhem et al. (2020) assumed this condition to justify iVAEs using $q_\phi(z|x, u)$.

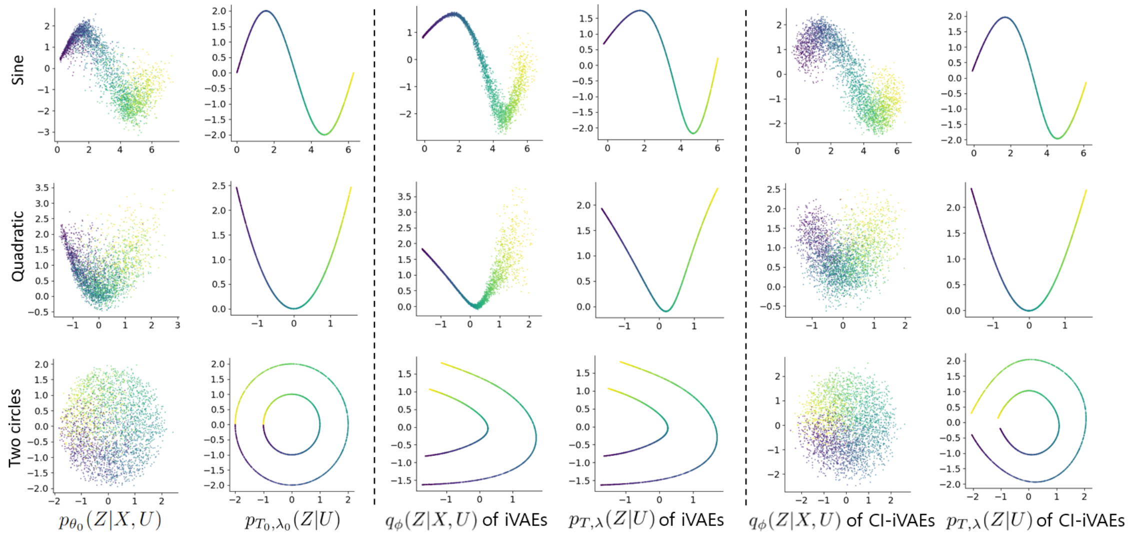

Figure 1: Visualization of latent variables from simulation datasets. In all datasets, the posterior $q_\phi(z|x,u)$ of iVAEs are close to the label prior $p_{T,\lambda}(z|u)$, and tend to underestimate the variability by observations. In contrast, by addressing samplewise optimal posteriors, CI-iVAEs learn posterior and prior distributions closer to GT.

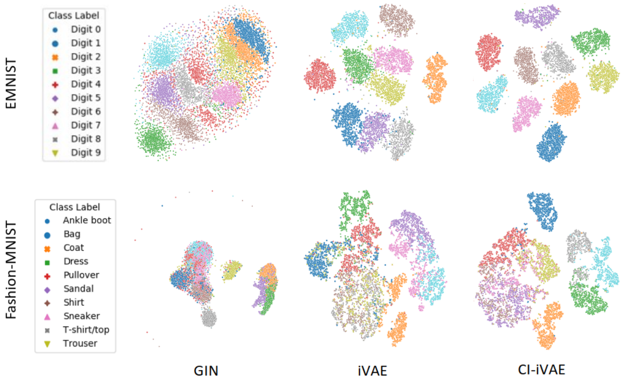

Figure 2: Visualization of the t-SNE embeddings of encoders $q_\phi(z|x)$ from various methods on EMNIST and Fashion-MNIST datasets.

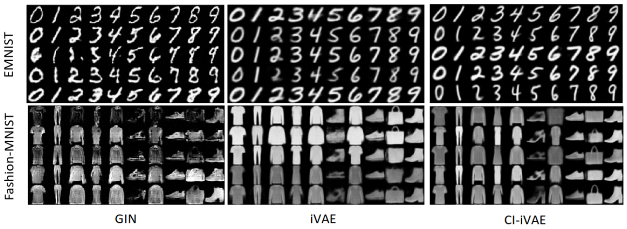

Figure 3: Generation results from various methods on EMNIST and Fashion-MNIST datasets. We generate five synthetic images in each class. Both iVAEs and CI-iVAEs produce clearer images than GIN.

Implementation details, including data descriptions and network architectures, are provided in Appendix B.1.[4]

4.1 Simulation Study

We first examine the effectiveness of samplewise optimal posteriors by comparing iVAEs and CI-iVAEs on synthetic datasets. We use three data generation schemes and named them by shapes of distributions of latent variables given covariates: (i) sine, (ii) quadratic, and (iii) two circles.

Table 1 and Figure 1 show the results from the synthetic datasets. We consider coefficient of determination (COD, Schneider et al. 2022), mean correlation coefficient (MCC, Khemakhem et al. 2020), and log-likelihood as evaluation metrics. Higher values of CODs and MCCs indicate closer to the GT. The log-likelihood is used to evaluate the whole data generation scheme. All methods use the same network architectures to use the same family of density functions. The log-likelihood is calculated by $\log p_\theta(x|u) = \log \int p_f(x|z)p_{T,\lambda}(z|u)dz$ with Monte Carlo approximation and is averaged over samples.

According to Table 1, iVAEs are worse than CI-iVAEs in all datasets and all evaluation metrics. When the iVAE is compared with the CI-iVAE, most of p-values for all datasets and all evaluation metrics are less .001. The only exception is the MCCs on $q_\phi(z|x)$ on the two circles dataset whose p-value is .22. That is, our sharper ELBO with samplewise optimal posteriors enhances performances on learning GT ICs and MLEs for mixing functions. We also observe that encoder distributions of CI-iVAEs yield performances

[4] The implementation code is provided in the supplementary file.

Table 1: Means of evaluation metrics with standard errors from iVAEs and CI-iVAEs on various latent structures. For all metrics, higher values are better. CI-iVAEs outperform iVAEs in all datasets and all metrics. The number of repeats is 20.

Latent Structure	Method	Evaluation Metric				
		COD ($q_\phi(z\|x,u)$)	COD ($q_\phi(z\|x)$)	MCC ($q_\phi(z\|x,u)$)	MCC ($q_\phi(z\|x)$)	Log-likelihood
Sine	iVAE	.9285 (.0009)	.9692 (.0004)	.7364 (.0286)	.7531 (.0266)	-131.8387 (1.7467)
	CI-iVAE	**.9823** (.0002)	**.9782** (.0003)	**.8898** (.0114)	**.8716** (.0117)	**-111.2823** (0.5576)
Quadratic	iVAE	.5238 (.0059)	.8256 (.0112)	.5467 (.0086)	.6966 (.0097)	-105.4009 (0.1260)
	CI-iVAE	**.9177** (.0006)	**.8755** (.0013)	**.8463** (.0192)	**.8424** (.0192)	**-101.0430** (0.2420)
Two Circles	iVAE	.2835 (.0119)	.7753 (.0186)	.6233 (.0108)	.8171 (.0106)	-114.9876 (1.2964)
	CI-iVAE	**.9156** (.0007)	**.9440** (.0008)	**.8278** (.0194)	**.8347** (.0196)	**-102.9267** (0.3198)

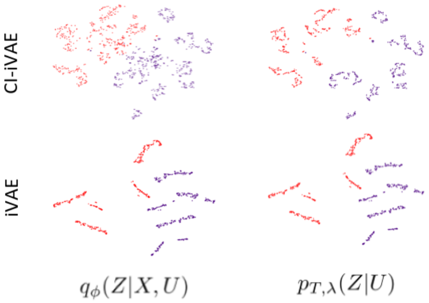

Figure 4: Visualization of the t-SNE embeddings of latent variables from iVAEs and CI-iVAEs on the ABCD dataset. Red and blue points indicate female and male, respectively. The posterior of iVAEs collapses to the label prior while that of CI-iVAEs learn variations by observations.

Table 2: Means of evaluation metrics with standard errors from various methods on EMNIST and Fashion-MNIST datasets. The number of repeats is 20.

Dataset	Method	SSW/SST (↓)
EMNIST	GIN	.6130 (.0075)
	iVAE	.5486 (.0037)
	CI-iVAE	**.4117** (.0032)
Fashion-MNIST	GIN	.8503 (.0026)
	iVAE	.6157 (.0046)
	CI-iVAE	**.4926** (.0024)

comparable to posterior distributions.

Visualization of latent variables and results is presented in Figure 1. In the left panel, the GT posterior and prior distributions are presented. In the middle and right panels, results from iVAEs and CI-iVAEs are presented, respectively. Points indicate conditional expectations of each distributions and are colored by covariates. For the two circles structure, we color datapoints by angles, U_1. In all latent structures, iVAEs learn $q_\phi(z|x,u)$ closer to $p_{T,\lambda}(z|u)$. The GT posterior distribution of ICs has variations by observations, and the $q_\phi(z|x,u)$ of iVAEs tends to underestimate these variations. In contrast, CI-iVAEs alleviate this phenomenon, consequently better recovering the GT posterior and prior distributions than iVAEs. Especially in the two circles dataset, we sample the angle U_1 from $[-\pi,\pi]$ to check whether each model can recover the connection at $U_1=-\pi$ and $U_1=\pi$. The points at $U_1=-\pi$ and π by CI-iVAEs locate closer to each other than those by iVAEs.

4.2 Applications to Real Datasets

4.2.1 EMNIST and Fashion-MNIST

We compare the proposed CI-iVAE with GIN and iVAE on two benchmark image datasets: EMNIST and Fashion-MNIST. GIN is a state-of-the-art identifiable deep image generative model using convolutional coupling blocks (Sorrenson et al., 2020). GIN does not incorporate covariates, so we compare representations from encoders $q_\phi(z|x)$ from various methods. In EMNIST and Fashion-MNIST, we use labels of digits and fashion-items as covariates.

The experimental results are presented in Table 2 and Figure 2. Further generation results are provided in Figures 1 and 2 in Appendix B.2. We consider the ratio of the within-cluster sum of squares (SSW) over the total sum of squares (SST) as evaluation metrics. The SSW/SST measures how well representations are clustered by covariates.

Results presented in Table 2 show that our method is better than both iVAEs and GIN in all datasets. When the iVAE is compared with the CI-iVAE, p-values are 2.3×10^{-27} for EMNIST and 8.8×10^{-25} for Fashion-MNIST datasets. When GIN is compared with the CI-iVAE, p-values are 2.1×10^{-25} for EMNIST and 3.1×10^{-48} for Fashion-MNIST datasets. That is, the CI-iVAE is better than iVAE and GIN for extracting covariates-related information from observations.

A visualization of latent variables using t-SNE (Van der Maaten and Hinton, 2008) embeddings is presented in Figure 2. In all datasets, CI-iVAEs yield more separable representations of the covariates than GIN and iVAEs.

We present generation results in Figure 3. iVAEs and CI-iVAEs tend to generate more plausible images than GIN.

Table 3: Means of prediction performances with standard errors from 5-fold cross-validation. We train support vector regression for age and CBCL scores, and support vector machine for sex and puberty. The representation from CI-iVAEs outperform that from AEs and iVAEs baselines. Prediction performances with our representations are comparable to or better than those with raw observations.

Input	Covariates			
	Age (MSE ↓)	Sex (Error rate ↓)	Puberty (F1 score ↑)	CBCL scores (MSE ↓)
x	.165 (.012)	.105 (.007)	.531 (.023)	**.072** (.006)
Representations from AE	.336 (.005)	.368 (.006)	.188 (.009)	.135 (.006)
Representations from iVAE	.361 (.005)	.479 (.010)	.046 (.005)	.148 (.006)
Representations from CI-iVAE	**.153** (.017)	**.086** (.009)	**.563** (.053)	.073 (.008)

In the third row in EMNIST, GIN fails to produce some digits. When we compare iVAEs and CI-iVAEs, iVAEs tend to generate more blurry results. In the 6th column in Fashion-MNIST, iVAEs fail to generate the sandal image in the second row and generate blurry results in other rows.

4.2.2 Application to Brain Imaging Data

Here we present the application of the CI-iVAE on the ABCD study dataset, which is the largest single-cohort prospective longitudinal study of neurodevelopment and children's mental health in the United States. The ABCD dataset provides resources to address a central scientific question in Psychiatry: can we find the brain imaging representations that are associated with phenotypes. For this purpose, the proposed CI-iVAE is appealing and more efficient in that it incorporates the information in demographics, symptoms, which, by domain knowledge, are associated with brain imaging. In this application, observations are the MRI mean thickness and functional connectivity data from Gorden Atlas (Gordon et al., 2016), and the covariates are interview age, gender, puberty level, and total Child Behavior Checklist (CBCL) scores. To find the brain imaging representations that can be interpreted with covariates, we train iVAEs and evaluate representations from encoders $q_\phi(z|x)$ using only test brain imaging. With $q_\phi(z|x)$, we can extract information contained only in brain imaging. We consider vanilla AEs and iVAEs as baselines.

The experimental results are presented in Figure 4 and Table 3. We quantify prediction performance using representations from various methods to evaluate how much covariate-related information in observations is extracted from brain measures. We use conditional expectations of encoders, $q_\phi(z|x)$, as representations. For the puberty level, we oversample minority classes to balance classes.

Visualization of t-SNE embeddings of latent variables is presented in Figure 4. The result from iVAEs demonstrates that $q_\phi(z|x, u)$ collapses to $p_{T,\lambda}(z|u)$ and iVAEs tend to learn less diverse representations than CI-iVAEs, which is consistent to previous results.

Table 3 shows that, for all covariates, representations from CI-iVAEs outperform those from AEs and iVAEs and are comparable to raw observations. The CI-iVAE is significantly better than the iVAE, p-values for age, sex, puberty level, and CBCL scores are 1.1×10^{-6}, 8.4×10^{-10}, 4.6×10^{-6}, and 3.1×10^{-5}, respectively. Due to the collapsed posterior, representations from encoders in iVAEs do not extract much information from brain imaging. In contrast, the encoder of CI-iVAEs extracts representations which are very informative of the covariates. Thus, our representations extracted from brain measures preserve more covariates-related information than the other methods.

5 Discussion

We proposed a new representation learning approach, CI-iVAEs, to overcome the limitations of iVAEs. Our objective function uses samplewise optimal posterior distributions to prevent the posterior collapse problem. Representations from our methods on various synthetic and real datasets were better than those from existing methods by extracting covariates-associated information in observations. Our work is the first to adapt identifiable generative models to human brain imaging. We used pre-processed ROI level summary measures as observations. An interesting future direction is to extract interpretable features from minimally-processed images. Another direction is to apply/develop interpretable machine learning tools creating feature importance scores (Ribeiro et al., 2019; Molnar, 2020; Guidotti et al., 2018) to reveal scientific insights from our representations.

6 Acknowledgement

This work is supported by NIMH grant R01 MH124106.

A Details on Theoretical Results

A.1 Further Theoretical Results

Our method introduces both $q_\phi(z|x)$ and $q_\phi(z|x,u)$ in consisting posterior distributions to conduct variational inference. We derive that the proposed ELBOs are concave, are sharper lower bounds than the ELBO of the existing iVAE, and prevent the posterior collapse issue in the existing iVAE.

We first show the concavity of the proposed ELBO and a necessary and sufficient condition for the concavity.

Proposition 6. *For any sample (x,u) and $\alpha \in [0,1]$, ELBO$_{\theta,\phi}(\alpha;x,u)$ is concave w.r.t. α. It is strictly concave if and only if $q_\phi(z|x) \neq q_\phi(z|x,u)$ for some z.*

Note that the first two terms in Proposition 5 are linear w.r.t. α, so the concavity comes from the last two terms. That is, the difference between $q_\phi(z|x)$ and $q_\phi(z|x,u)$ induces the concavity.

Next, we show that our method uses strictly sharper lower bounds than the existing iVAE. Let $\Delta_{1-0}(x,u) :=$ ELBO$_{\theta,\phi}(1;x,u) -$ ELBO$_{\theta,\phi}(0;x,u)$ and

$$\mathrm{SNR}(g) := \frac{(\mathbb{E}_{q_\phi(z|x)}g(z) - \mathbb{E}_{q_\phi(z|x,u)}g(z))^2}{\max(Var_{q_\phi(z|x)}g(z), Var_{q_\phi(z|x,u)}g(z))}. \tag{6}$$

Here, SNR(g) is the signal-to-noise ratio between $q_\phi(z|x)$ and $q_\phi(z|x,u)$ w.r.t. g quantifying the discrepancy between the two distributions. The proof of Theorem 3 is provided in Appendix A.3.

Theorem 3. *For any sample (x,u), $\theta = (f,T,\lambda)$, ϕ, and $\epsilon > 0$, if there is a function $g : \mathcal{Z} \to \mathbb{R}$ satisfying SNR$(g) \geq 1/\epsilon$ and $|\Delta_{1-0}(x,u)| \leq -\log\epsilon + O(\epsilon\log\epsilon)$, then*

$$\begin{aligned}&\mathit{ELBO}_{\theta,\phi}(\alpha^*(x,u);x,u) - \mathit{ELBO}_{\theta,\phi}(0;x,u)\\ &\geq \frac{-1+\sqrt{1+4\epsilon}}{2}|\Delta_{1-0}(x,u)| + o(|\Delta_{1-0}(x,u)|) + O(\epsilon\log\epsilon)\\ &\quad \mathit{as} \quad |\Delta_{1-0}(x,u)| \to \infty \ \mathit{and}\ \epsilon \to 0^+.\end{aligned}$$

That is, if $q_\phi(z|x)$ and $q_\phi(z|x,u)$ are different so that SNR(g) is large enough for some g, then $\alpha^*(x,u) > 0$ and our bound is sharper than that of iVAE with positive margins. In the simulation study in Table 2 in Appendix B.2, $\alpha^*(x,u) > 0$ holds with positive probability and $\alpha^*(x,u)$ can be approximated by a formula based on our theory. Under the same conditions in Theorem 3, we derived that

$$\alpha^*_{\text{approx}}(\epsilon, \Delta_{1-0}(x,u)) := \frac{1-\sqrt{1+4\epsilon}}{2} + \frac{\sqrt{1+4\epsilon}}{1+e^{-\sqrt{1+4\epsilon}\Delta_{1-0}(x,u)}} \tag{7}$$

is in $[0,1]$ and ELBO$_{\theta,\phi}(\alpha^*_{\text{approx}}(\epsilon,\Delta_{1-0}(x,u));x,u) -$ ELBO$_{\theta,\phi}(0;x,u)$ is greater than or equal to the positive margin in Theorem 3. Details are provided in Lemma 5 in Appendix A.3. The calculated values by this formula were similar to numerically approximated α^*, which supports the validity of our theory.

A.2 Proofs of Propositions

A.2.1 Proof of Proposition 1

We provide a proof by contradiction. We assume that $q_\phi(z|x,u) \neq p_{T,\lambda}(z|u)$ holds w.p. 0, which is equivalent to assume that the posterior collapse occurs, i.e., $q_\phi(z|x,u) = p_{T,\lambda}(z|u)$ holds w.p. 1. Since $q_\phi(z|x,u) = q_\phi(z|x)$, it implies that $q_\phi(z|x) = p_{T,\lambda}(z|u)$ holds w.p. 1. Now, $p_{T,\lambda}(z|u) = q_\phi(z|x) = \int q_\phi(z|x)p(u)du = \int p_{T,\lambda}(z|u)p(u)du$ contradicts to that $p_{T,\lambda}$ is non-trivial.

A.2.2 Proof of Proposition 2

We provide a proof by contradiction. Since $\mathbb{E}_{q_\phi(z|x)}\log p_f(x|z) - \mathcal{D}_{\text{KL}}(q_\phi(z|x)||p_{T,\lambda}(z|u))$ is equal to $\log p_\theta(x|u) - \mathcal{D}_{\text{KL}}(q_\phi(z|x)||p_\theta(z|x,u))$, $\max_\phi \mathbb{E}_{p(x,u)}$ELBO$_{\theta,\phi}(1;x,u)$ is equal to $\mathbb{E}_{p(x,u)}\log p_\theta(x|u)$ if and only if $\min_\phi \mathcal{D}_{\text{KL}}(q_\phi(z|x)||p_\theta(z|x,u)) = 0$. It implies $p_\theta(z|x,u) = p_\theta(z|x)$. By Bayes' theorem, $p_\theta(z|x,u) = p_\theta(x|z,u)p_{T,\lambda}(z|u)/p_\theta(x|u) = p_f(x|z)p_{T,\lambda}(z|u)/p_\theta(x|u)$ and $p_\theta(z|x) = p_f(x|z)p_\theta(z)/p_\theta(x)$. Thus, $p_{T,\lambda}(z|u)p_\theta(x) = p_\theta(z)p_\theta(x|u)$. Now, $p_{T,\lambda}(z|u) = \int p_{T,\lambda}(z|u)p_\theta(x)dx = \int p_\theta(z)p_\theta(x|u)dx = p_\theta(z)$ contradicts to that the label prior is non-trivial.

A.2.3 Proof of Proposition 3

For any sample (x,u), $\theta = (f,T,\lambda)$, ϕ, and $\alpha \in [0,1]$,

$$\begin{aligned}&\log p_\theta(x|u)\\ &= \log\Big(\int \frac{p_\theta(x,z|u)}{\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)}\Big(\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)\Big)dz\Big)\\ &\geq \int\Big(\log\frac{p_\theta(x,z|u)}{\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)}\Big)\Big(\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)\Big)dz\\ &= \mathbb{E}_{\alpha q_\phi(z|x)+(1-\alpha)q_\phi(z|x,u)}[\log p_\theta(x,z|u) - p_{T,\lambda}(z|u)]\\ &\quad - \mathcal{D}_{\text{KL}}(\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)||p_{T,\lambda}(z|u))\end{aligned}$$

holds by Jensen's inequality. Now, $p_\theta(x,z|u) = p_f(x|z)p_{T,\lambda}(z|u)$ concludes the proof.

A.2.4 Proof of Proposition 4

By the definition of $\alpha^*(x,u)$ and the existence of ϕ satisfying $q_\phi(z|x,u) = p_\theta(z|x,u)$ with probability (w.p.) 1 w.r.t. $p(x,u)$,

$$\begin{aligned}&\max_\phi \mathbb{E}_{p(x,u)}\mathrm{ELBO}_{\theta,\phi}(\alpha^*(x,u);x,u)\\ &\geq \max_\phi \mathbb{E}_{p(x,u)}\mathrm{ELBO}_{\theta,\phi}(0;x,u)\\ &= \mathbb{E}_{p(x,u)}\log p_\theta(x|u) - \min_\phi \mathcal{D}_{\text{KL}}(q_\phi(z|x,u)||p_\theta(z|x,u))\\ &= \mathbb{E}_{p(x,u)}\log p_\theta(x|u).\end{aligned}$$

Since ELBO$_{\theta,\phi}(\alpha^*(x,u);x,u)$ is a lower bound of $\log p_\theta(x|u)$, the proof is concluded.

A.2.5 Proof of Proposition 5

For any sample (x, u), $\theta = (f, T, \lambda)$, ϕ, and $\alpha \in [0, 1]$, $\text{ELBO}_{\theta,\phi}(\alpha; x, u)$ can be expressed as

$$\begin{aligned}
&\alpha\mathbb{E}_{q_\phi(z|x)} \log p_f(x|z) + (1-\alpha)\mathbb{E}_{q_\phi(z|x,u)} \log p_f(x|z) \\
&- \int \Big(\log \frac{\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)}{p_{T,\lambda}(z|u)}\Big)(\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u))dz \\
&= \alpha\Big(\text{ELBO}_{\theta,\phi}(1; x, u) + \int \Big(\log \frac{q_\phi(z|x)}{p_{T,\lambda}(z|u)}\Big)q_\phi(z|x)dz\Big) \\
&\quad + (1-\alpha)\Big(\text{ELBO}_{\theta,\phi}(0; x, u) + \int \Big(\log \frac{q_\phi(z|x,u)}{p_{T,\lambda}(z|u)}\Big)q_\phi(z|x,u)dz\Big) \\
&\quad - \alpha \int \Big(\log \frac{\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)}{p_{T,\lambda}(z|u)}\Big)q_\phi(z|x)dz \\
&\quad - (1-\alpha) \int \Big(\log \frac{\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)}{p_{T,\lambda}(z|u)}\Big)q_\phi(z|x,u)dz \\
&= \alpha\text{ELBO}_{\theta,\phi}(1; x, u) + (1-\alpha)\text{ELBO}_{\theta,\phi}(0; x, u) \\
&\quad + \alpha\mathcal{D}_{\text{KL}}(q_\phi(z|x)||\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)) \\
&\quad + (1-\alpha)\mathcal{D}_{\text{KL}}(q_\phi(z|x,u)||\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)) \\
&= \alpha\text{ELBO}_{\theta,\phi}(1; x, u) + (1-\alpha)\text{ELBO}_{\theta,\phi}(0; x, u) \\
&\quad + \alpha\mathcal{D}^{1-\alpha}_{\text{skew}}(q_\phi(z|x)||q_\phi(z|x,u)) + (1-\alpha)\mathcal{D}^{\alpha}_{\text{skew}}(q_\phi(z|x,u)||q_\phi(z|x)).
\end{aligned}$$

A.2.6 Proof of Proposition 6

For any sample (x, u) and $\alpha \in [0, 1]$, the first derivative of $\text{ELBO}_{\theta,\phi}(\alpha; x, u)$ can be expressed as

$$\begin{aligned}
&\text{ELBO}^{(1)}_{\theta,\phi}(\alpha; x, u) \\
&= \mathbb{E}_{q_\phi(z|x)} \log p_f(x|z) - \mathbb{E}_{q_\phi(z|x,u)} \log p_f(x|z) \\
&\quad - \frac{d}{d\alpha}\int \Big(\log \frac{\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)}{p_{T,\lambda}(z|u)}\Big)(\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u))dz \\
&= \Big(\text{ELBO}_{\theta,\phi}(1; x, u) + \mathcal{D}_{\text{KL}}(q_\phi(z|x)||p_{T,\lambda}(z|u))\Big) \\
&\quad - \Big(\text{ELBO}_{\theta,\phi}(0; x, u) + \mathcal{D}_{\text{KL}}(q_\phi(z|x,u)||p_{T,\lambda}(z|u))\Big) \\
&\quad - \Big[\int \frac{q_\phi(z|x) - q_\phi(z|x,u)}{\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)}(\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u))dz \\
&\quad + \int \Big(\log \frac{\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)}{p_{T,\lambda}(z|u)}\Big)(q_\phi(z|x) - q_\phi(z|x,u))dz\Big] \\
&= \text{ELBO}_{\theta,\phi}(1; x, u) - \text{ELBO}_{\theta,\phi}(0; x, u) \\
&\quad + \mathcal{D}_{\text{KL}}(q_\phi(z|x)||\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)) \\
&\quad - \mathcal{D}_{\text{KL}}(q_\phi(z|x,u)||\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)).
\end{aligned}$$

With this, the second derivative of $\text{ELBO}_{\theta,\phi}(\alpha; x, u)$ can be expressed as

$$\begin{aligned}
&\text{ELBO}^{(2)}_{\theta,\phi}(\alpha; x, u) \\
&= -\frac{d}{d\alpha}\int \Big(\log \frac{\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)}{p_{T,\lambda}(z|u)}\Big)(q_\phi(z|x) - q_\phi(z|x,u))dz \\
&= -\int \Big(\log \frac{q_\phi(z|x) - q_\phi(z|x,u)}{\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)}\Big)(q_\phi(z|x) - q_\phi(z|x,u))dz \\
&= -\int \frac{(q_\phi(z|x) - q_\phi(z|x,u))^2}{\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)}dz.
\end{aligned}$$

Thus, $\text{ELBO}^{(2)}_{\theta,\phi}(\alpha; x, u) \leq 0$ for any $\alpha \in [0, 1]$. The equality holds if and only if $q_\phi(z|x) = q_\phi(z|x, u)$ for all z, which concludes the proof.

A.3 Proofs of Theorems

A.3.1 Proof of Theorem 1

The proof is extended from the proof of Proposition 2 in Dai et al. (2020). We first provide detailed formulations of conditions for Theorem 1, and then derive Theorem 1. We denote the parameter in decoder networks by ψ. The reconstruction error at x with posterior $N(\mu, \sigma)$ can be expressed as $\mathbb{E}_{Z\sim N(\mu,\sigma^2)}||x - f(Z; \psi)||_2^2$. We formulate conditions $(C1)$ and $(C2)$ (Dai et al., 2020).

$(C1)$ $\frac{\partial}{\partial\mu}\mathbb{E}_{Z\sim N(\mu,\sigma^2)}||x - f(Z; \psi)||_2^2$ and $\frac{\partial}{\partial\sigma}\mathbb{E}_{Z\sim N(\mu,\sigma^2)}||x - f(Z; \psi)||_2^2$ are L-Lipschitz continuous.
$(C2)$ $\frac{\partial}{\partial\sigma}\mathbb{E}_{Z\sim N(\mu,\sigma^2)}||x - f(Z; \psi)||_2^2 \geq c$ for some $c > 0$.

The $(C1)$ means that the reconstruction error is sufficiently smooth and its partial derivatives have bounded slopes w.r.t. (μ, σ). The $(C2)$ means that the decoder increases the reconstruction error as the uncertainty of latent variables increases. The $(C2)$ does not hold when the decoder is degenerated, i.e., $f(z; \psi)$ is constant.

Now, we derive Theorem 1. We show that for any iVAEs satisfying $(C1)$ and $(C2)$, there is a posterior collapse case having larger value of ELBO. As in Dai et al. (2020), we consider that the observation noise ϵ follows a Gaussian distribution $N(0, \gamma I_{d_X})$. We reparametrize the decoder network with a scale parameter $w \in [0, 1]$ and denote the parameter for the decoder by $\psi = (w, \psi\backslash w)$. The output of the decoder with ψ can be expressed as $f(z; \psi) = f(wz; \psi\backslash w)$. We denote means and standard deviations of posterior distributions by

$$\begin{aligned}
&m_{Z|X,U}(\{(x_i, u_i)\}_{i=1}^n; \phi) \\
&:= (\mu_{Z|X,U}(x_1, u_1; \phi), \ldots, \mu_{Z|X,U}(x_n, u_n; \phi))
\end{aligned}$$

and

$$\begin{aligned}
&s_{Z|X,U}(\{(x_i, u_i)\}_{i=1}^n; \phi) \\
&:= (\sigma_{Z|X,U}(x_1, u_1; \phi), \ldots, \sigma_{Z|X,U}(x_n, u_n; \phi)),
\end{aligned}$$

respectively, where $\mu_{Z|X,U}(x_i, u_i; \phi)$ and $\sigma_{Z|X,U}(x_i, u_i; \phi)$ denote the mean and standard deviations of posterior distributions at i-th datum, respectively.

For any iVAEs with decoder parameter $\tilde{\psi}$ satisfying $(C1)$ and $(C2)$, posterior parameter $\tilde{\phi}$, and label prior parameter $(\tilde{T}, \tilde{\lambda})$, values evaluated at $\tilde{\phi}$ are denoted by $\tilde{m}_{Z|X,U}$ and $\tilde{s}_{Z|X,U}$. For simplicity, we use $m_{Z|X,U}$, $s_{Z|X,U}$, $\tilde{m}_{Z|X,U}$, and $\tilde{s}_{Z|X,U}$ if there is no confusion and use $m_{Z|X,U,i}$, $s_{Z|X,U,i}$, $\tilde{m}_{Z|X,U,i}$, and $\tilde{s}_{Z|X,U,i}$, respectively, to indicate their i-th components. In a similar manner, means and standard deviations by label prior networks are denoted by $m_{Z|U}$ and $s_{Z|U}$, respectively. We denote $(m^{\text{Scale}}_{Z|X,U,i})_j = ((m_{Z|X,U,i})_j - (\tilde{m}_{Z|U,i})_j)/(\tilde{s}_{Z|U,i})_j$, $(s^{\text{Scale}}_{Z|X,U,i})_j = (s_{Z|X,U,i})_j/(\tilde{s}_{Z|U,i})_j$, $(\tilde{m}^{\text{Scale}}_{Z|X,U,i})_j = ((\tilde{m}_{Z|X,U,i})_j - (\tilde{m}_{Z|U,i})_j)/(\tilde{s}_{Z|U,i})_j$, and $(\tilde{s}^{\text{Scale}}_{Z|X,U,i})_j = (\tilde{s}_{Z|X,U,i})_j/(\tilde{s}_{Z|U,i})_j$. With these terms, we can express the reconstruction error at the i-th datum as

$$
\begin{aligned}
&r(w(m^{\text{Scale}}_{Z|X,U,i} + \tilde{m}_{Z|U,i}/\tilde{s}_{Z|U,i}), ws^{\text{Scale}}_{Z|X,U,i}, \tilde{\psi}\backslash\tilde{w}, x_i) \\
&= \mathbb{E}_{\epsilon\sim N(0,I_d)}\|x_i - f\big(\tilde{s}_{Z|U}(w(m^{\text{Scale}}_{Z|X,U,i} + \tilde{m}_{Z|U,i}/\tilde{s}_{Z|U,i})) \\
&\quad + \tilde{s}_{Z|U,i}(ws^{\text{Scale}}_{Z|X,U,i})\epsilon; \tilde{\psi}\backslash\tilde{w}\big)\|_2^2
\end{aligned}
$$

and the average reconstruction error by $\bar{r}(w(m^{\text{Scale}}_{Z|X,U} + \tilde{m}_{Z|U}/\tilde{s}_{Z|U}), ws^{\text{Scale}}_{Z|X,U})$. Here, all the parameters in decoder and label prior but w are fixed. Then, the average of the negative ELBO of iVAE evaluated with $\{(x_i, u_i)\}_{i=1}^n$ is the same as

$$
\begin{aligned}
&h(m_{Z|X,U}, s_{Z|X,U}, w) \\
&:= \gamma^{-1}\bar{r}(w(m^{\text{Scale}}_{Z|X,U} + \tilde{m}_{Z|U}/\tilde{s}_{Z|U}), ws^{\text{Scale}}_{Z|X,U}) + d_X \log\gamma \\
&\quad + n^{-1}\sum_{i=1}^{n} 2\mathcal{D}_{\text{KL}}(N(m^{\text{Scale}}_{Z|X,U,i}, (s^{\text{Scale}}_{Z|X,U,i})^2)||N(0,1))
\end{aligned}
$$

up to constant addition and multiplication since $\mathcal{D}_{\text{KL}}(N(m_{Z|X,U,i}, s^2_{Z|X,U,i})||N(\tilde{m}_{Z|U,i}, \tilde{s}^2_{Z|U,i})) = \mathcal{D}_{\text{KL}}(N(m^{\text{Scale}}_{Z|X,U,i}, (s^{\text{Scale}}_{Z|X,U,i})^2)||N(0,1))$. We define h^{appr}, an approximation of an upper bound of h based on the Taylor series of $\bar{r}$ in Lemma 1. The h^{appr} is equal to h when $(m_{Z|X,U}, s_{Z|X,U}, w) = (\tilde{m}_{Z|X,U}, \tilde{s}_{Z|X,U}, \tilde{w})$ and is an upper bound of h when $w(s_{Z|X,U,i})_j \in \{0, \tilde{w}(\tilde{s}_{Z|X,U,i})_j\}$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, d_Z$.

Lemma 1. *For any $\{(x_i, u_i)\}_{i=1}^n$, we define*

$$
\begin{aligned}
&h^{\mathit{appr}}(m_{Z|X,U}, s_{Z|X,U}, w; \tilde{m}_{Z|X,U}, \tilde{s}_{Z|X,U}, \tilde{w}) \\
&:= \gamma^{-1}\bar{r}^{\mathit{appr}}(w(m^{\mathit{Scale}}_{Z|X,U} + \tilde{m}_{Z|U}/\tilde{s}_{Z|U}), ws^{\mathit{Scale}}_{Z|X,U}; \\
&\quad \tilde{w}(\tilde{m}^{\mathit{Scale}}_{Z|X,U} + \tilde{m}_{Z|U}/\tilde{s}_{Z|U}), \tilde{w}\tilde{s}^{\mathit{Scale}}_{Z|X,U}) + d_X\log\gamma \\
&\quad + n^{-1}\sum_{i=1}^{n} 2\mathcal{D}_{\mathit{KL}}(N(m^{\mathit{Scale}}_{Z|X,U,i}, (s^{\mathit{Scale}}_{Z|X,U,i})^2)||N(0,1))
\end{aligned}
$$

where $\bar{r}^{\mathit{appr}}(u, v; \tilde{u}, \tilde{v}) := \bar{r}(\tilde{u}, \tilde{v}) + (u - \tilde{u})^T\frac{\partial}{\partial u}\bar{r}(\tilde{u}, \tilde{v}) + \frac{L}{2}||u - \tilde{u}||_2^2 + \sum_{j=1}^{D} g^{\mathit{appr}}\Big(v_j, \tilde{v}_j, \frac{\partial}{\partial v_j}\bar{r}(\tilde{u}, \tilde{v})\Big)$ for any u, v, $\tilde{u}$ and $\tilde{v} \in \mathbb{R}^D$ and $g^{\mathit{appr}} : \mathbb{R}^3 \to \mathbb{R}$ is defined as follows:

$$
\begin{aligned}
&g^{\mathit{appr}}(v, \tilde{v}, \delta) \\
&= \begin{cases} -\frac{\delta^2}{2L} + \frac{\delta^2}{2L\tilde{v}^2}v^2 & \text{if } v \geq \tilde{v} - \frac{\delta}{L} \text{ and } \{v, \tilde{v}, \delta\} \geq 0, \\ \big(\frac{L\tilde{v}^2}{2} - \delta\tilde{v}\big) + \big(\frac{\delta}{\tilde{v}} - \frac{L}{2}\big)v^2 & \text{if } v < \tilde{v} - \frac{\delta}{L} \text{ and } \{v, \tilde{v}, \delta\} \geq 0, \\ \infty & \text{otherwise.} \end{cases}
\end{aligned}
\tag{8}
$$

Then,

$$
\begin{aligned}
&h^{\mathit{appr}}(\tilde{m}_{Z|X,U}, \tilde{s}_{Z|X,U}, \tilde{w}; \tilde{m}_{Z|X,U}, \tilde{s}_{Z|X,U}, \tilde{w}) \\
&\quad = h(\tilde{m}_{Z|X,U}, \tilde{s}_{Z|X,U}, \tilde{w})
\end{aligned}
$$

and

$$
\begin{aligned}
&h^{\mathit{appr}}(m_{Z|X,U}, s_{Z|X,U}, w; \tilde{m}_{Z|X,U}, \tilde{s}_{Z|X,U}, \tilde{w}) \\
&\quad \geq h(m_{Z|X,U}, s_{Z|X,U}, w)
\end{aligned}
$$

if $w(s_{Z|X,U,i})_j \in \{0, \tilde{w}(\tilde{s}_{Z|X,U,i})_j\}$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, d_Z$.

Proof of Lemma 1. Section 3 in the supplementary file of Dai et al. (2020) showed that $\bar{r}^{\text{appr}}(\tilde{u}, \tilde{v}; \tilde{u}, \tilde{v}) = \bar{r}(\tilde{u}, \tilde{v})$ and $\bar{r}^{\text{appr}}(u, v; \tilde{u}, \tilde{v}) \geq \bar{r}(u, v)$ if $v_j \in \{0, \tilde{v}_j\}$ for all j. It concludes the proof since the difference of h^{appr} from h is changing $\bar{r}$ to $\bar{r}^{\text{appr}}$. □

Let $c_{i,j}$ be coefficients of v^2 in (8) determined by $(v, \tilde{v}, \delta) = (w(s^{\text{Scale}}_{Z|X,U,i})_j, \tilde{w}(\tilde{s}^{\text{Scale}}_{Z|X,U,i})_j, \tilde{\delta}_{i,j})$ where $\tilde{\delta}_{i,j} := \nabla\bar{r}(\tilde{w}(\tilde{m}^{\text{Scale}}_{Z|X,U} + \tilde{m}_{Z|U}/\tilde{s}_{Z|U}), \tilde{w}\tilde{s}^{\text{Scale}}_{Z|X,U})_{(n+i-1)d_Z+j}$. The $c_{i,j}$ is positive and finite since $w(s^{\text{Scale}}_{Z|X,U,i})_j \geq 0$, $\tilde{w}(\tilde{s}^{\text{Scale}}_{Z|X,U,i})_j \geq 0$, and $0 < \tilde{\delta}_{i,j} \leq L$ by $(C1)$ and $(C2)$.

We denote the minimizer of h^{appr} by $(m^*_{Z|X,U}(w), s^*_{Z|X,U}(w))$. Since

$$
\begin{aligned}
&h^{\text{appr}}(m_{Z|X,U}, s_{Z|X,U}, w; \tilde{m}_{Z|X,U}, \tilde{s}_{Z|X,U}, \tilde{w}) \\
&= \text{Const.} + n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{d_Z}\bigg(\gamma^{-1}\Big(w(m^{\text{Scale}}_{Z|X,U,i} + \tilde{m}_{Z|U,i}/\tilde{s}_{Z|U,i})_j\tilde{\delta}_{i,j} \\
&\quad + \frac{L}{2}\big(w^2(m^{\text{Scale}}_{Z|X,U,i} + \tilde{m}_{Z|U,i}/\tilde{s}_{Z|U,i})_j^2 \\
&\quad - 2w(m^{\text{Scale}}_{Z|X,U,i} + \tilde{m}_{Z|U,i}/\tilde{s}_{Z|U,i})_j\tilde{w}(\tilde{m}^{\text{Scale}}_{Z|X,U,i} + \tilde{m}_{Z|U,i}/\tilde{s}_{Z|U,i})_j\big) \\
&\quad + c_{i,j}w^2(s^{\text{Scale}}_{Z|X,U,i})_j^2\Big) + (m^{\text{Scale}}_{Z|X,U,i})_j^2 + (s^{\text{Scale}}_{Z|X,U,i})_j^2 - \log\Big(s^{\text{Scale}}_{Z|X,U,i}\Big)_j^2\bigg),
\end{aligned}
$$

h^{appr} is a quadratic function w.r.t. $(m^{\text{Scale}}_{Z|X,U,i} + \tilde{m}_{Z|U,i}/\tilde{s}_{Z|U,i})_j$ and coefficients of second-order and first-order terms are $n^{-1}(\gamma^{-1}Lw^2/2+1)$ and $n^{-1}\big(\gamma^{-1}w(\tilde{\delta}_{i,j} - L\tilde{w}(\tilde{m}_{Z|X,U,i})_j) - 2\tilde{m}_{Z|U,i}/\tilde{s}_{Z|U,i}\big)$, respectively. This implies

$$
\begin{aligned}
&(m^*_{Z|X,U,i}(w))_j \\
&= \Big(w(\tilde{s}_{Z|U,i})_j\big(L\tilde{w}(\tilde{m}_{Z|X,U,i})_j - \tilde{\delta}_{i,j}\big) + 2\gamma(\tilde{m}_{Z|U,i})_j\Big)/(2\gamma + Lw^2).
\end{aligned}
$$

For $(s_{Z|X,i})_j$, $\partial h^{\text{appr}}/\partial(s^{\text{Scale}}_{Z|X,i})_j^2 = n^{-1}(\gamma^{-1}c_{i,j}w^2 + 1 - 1/(s^{\text{Scale}}_{Z|X,i})_j^2)$ and $\partial^2 h^{\text{appr}}/\partial\big((s^{\text{Scale}}_{Z|X,i})_j^2\big)^2 = n^{-1}/(s^{\text{Scale}}_{Z|X,i})_j^4 > 0$ imply $(s^*_{Z|X,U,i}(w))_j^2 = (\tilde{s}_{Z|U,i})_j^2(\gamma^{-1}c_{i,j}w^2 + 1)^{-1}$. By substituting $(m^*_{Z|X,U}(w), s^*_{Z|X,U}(w))$, we have

$$
\begin{aligned}
&\partial h^{\text{appr}}(m^*_{Z|X}(w), s^*_{Z|X}(w), w; \tilde{m}_{Z|X}, \tilde{s}_{Z|X}, \tilde{w})/\partial w^2 \\
&\quad = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{d_Z}\bigg(\frac{c_{i,j}}{\gamma + c_{i,j}w^2} + O(\gamma^{-2})\bigg).
\end{aligned}
$$

Since $c_{i,j}$ is positive, this partial derivative is positive for all $w \in [0, 1]$ when γ is sufficiently large. In this case, the optimal w is zero and $\big((m^*_{Z|X,U,i})_j(0), (s^*_{Z|X,U,i})_j(0)\big) = ((\tilde{m}_{Z|U,i})_j, (\tilde{s}_{Z|U,i})_j)$, i.e., posterior collapse cases. By Lemma 1, $h(\tilde{m}_{Z|U}, \tilde{s}_{Z|U}, 0) \leq h^{\text{appr}}(\tilde{m}_{Z|U}, \tilde{s}_{Z|U}, 0; \tilde{m}_{Z|X,U}, \tilde{s}_{Z|X,U}, \tilde{w})$ and $h^{\text{appr}}(\tilde{m}_{Z|X,U}, \tilde{s}_{Z|X,U}, \tilde{w}; \tilde{m}_{Z|X,U}, \tilde{s}_{Z|X,U}, \tilde{w}) = h(\tilde{m}_{Z|X,U}, \tilde{s}_{Z|X,U}, \tilde{w})$. Since $(\tilde{m}_{Z|U}, \tilde{s}_{Z|U}, 0)$ is the global optima of h^{appr}, $h(\tilde{m}_{Z|U}, \tilde{s}_{Z|U}, 0) < h(\tilde{m}_{Z|X,U}, \tilde{s}_{Z|X,U}, \tilde{w})$. That is, there is a posterior collapse case whose value of ELBO is better than current networks. Thus, the iVAEs are worse than the posterior collapse case.

A.3.2 Proof of Theorem 2

We provide a proof by contradiction. Let $q_\phi(z|x,u) = p_{T,\lambda}(z|u)$ holds w.p. 1. Since $\alpha^*(x,u) > 0$ and $\alpha^*(x,u)q_\phi(z|x) + (1-\alpha^*(z|x,u)q_\phi(z|x,u) = p_{T,\lambda}(z|u)$ w.p. 1, we have $q_\phi(z|x) = p_{T,\lambda}(z|u)$, which contradicts to that $p_{T,\lambda}(z|u)$ is non-trivial by Proposition 1.

A.3.3 Proof of Theorem 3

We first provide lemmas with proofs, and then derive the theorem.

Lemma 2. *(Equation (18) in Nishiyama (2019)) For any $t \in [0,1]$, real-valued function g, and probability density functions $p(z)$ and $q(z)$, $\partial\mathcal{D}_{KL}(p(z)||(1-t)p(z) + tq(z))/\partial t$ is greater than or equal to $t(\mathbb{E}_{q(z)}[g(z)] - \mathbb{E}_{p(z)}[g(z)])^2/\big(t(1-t)(\mathbb{E}_{q(z)}[g(z)] - \mathbb{E}_{p(z)}[g(z)])^2 + (1-t)Var_{p(z)}[g(z)] + tVar_{q(z)}[g(z)]\big)$.*

Lemma 3. *For any datum (x,u) and positive number ϵ, if there is a function $g : \mathcal{Z} \to \mathbb{R}$ satisfying SNR$(g) \geq 1/\epsilon$, then*

$$\textit{ELBO}_{\theta,\phi}(\alpha;x,u) - \textit{ELBO}_{\theta,\phi}(0;x,u) \geq \textit{LB}(\alpha,\epsilon,\Delta_{1-0}(x,u))$$

where LB$(\alpha,\epsilon,\Delta_{1-0}(x,u)) := \alpha\Delta_{1-0}(x,u) + \alpha\int_\alpha^1 (1-t)/(t(1-t)+\epsilon)dt + (1-\alpha)\int_0^\alpha t/(t(1-t)+\epsilon)dt.$

Proof of Lemma 3. By Proposition 5,

$$\begin{aligned}&\mathrm{ELBO}_{\theta,\phi}(\alpha;x,u) - \mathrm{ELBO}_{\theta,\phi}(0;x,u)\\ &= \alpha\Delta_{1-0}(x,u) + \alpha\mathcal{D}_{\mathrm{KL}}(q_\phi(z|x)||\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u))\\ &\quad + (1-\alpha)\mathcal{D}_{\mathrm{KL}}(q_\phi(z|x,u)||\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)).\end{aligned} \tag{9}$$

By substituting $q_\phi(z|x)$ and $q_\phi(z|x,u)$ to q and p in Lemma 2, respectively, and integrating both sides from 0 to α, we have

$$\begin{aligned}&\mathcal{D}_{\mathrm{KL}}(q_\phi(z|x,u)||\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u))\\ &\geq \int_0^\alpha \Big(\big(t(\mathbb{E}_{q_\phi(z|x,u)}[g(z)] - \mathbb{E}_{q_\phi(z|x)}[g(z)])^2\big)/\big(t(1-t)(\mathbb{E}_{q_\phi(z|x,u)}[g(z)] - \mathbb{E}_{q_\phi(z|x)}[g(z)])^2\\ &\quad + (1-t)Var_{q_\phi(z|x)}[g(z)] + tVar_{q_\phi(z|x,u)}[g(z)]\big)\Big)dt.\end{aligned}$$

Since SNR$(g) \geq 1/\epsilon$,

$$\mathcal{D}_{\mathrm{KL}}(q_\phi(z|x,u)||\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)) \geq \int_0^\alpha \frac{t}{t(1-t)+\epsilon}dt. \tag{10}$$

In a similar way, we can derive

$$\mathcal{D}_{\mathrm{KL}}(q_\phi(z|x)||\alpha q_\phi(z|x) + (1-\alpha)q_\phi(z|x,u)) \geq \int_\alpha^1 \frac{1-t}{t(1-t)+\epsilon}dt. \tag{11}$$

By (9), (10), and (11), the proof is concluded. □

Lemma 4. *The first and second partial derivatives of LB$(\alpha,\epsilon,\Delta_{1-0}(x,u))$ w.r.t. α, respectively, can be expressed as*

$$\frac{\partial \textit{LB}(\alpha,\epsilon,\Delta_{1-0}(x,u))}{\partial\alpha} = \Delta_{1-0}(x,u) + \frac{1}{\sqrt{1+4\epsilon}}\log\left|\frac{\alpha - \frac{1+\sqrt{1+4\epsilon}}{2}}{\alpha - \frac{1-\sqrt{1+4\epsilon}}{2}}\right|$$

and ∂^2LB$(\alpha,\epsilon,\Delta_{1-0}(x,u))/\partial\alpha^2 = -1/\big(\alpha(1-\alpha)+\epsilon\big)$. Thus, LB$(\alpha,\epsilon,\Delta_{1-0}(x,u))$ is strictly concave w.r.t. α when $\alpha \in [0,1]$.

Proof of Lemma 4. The $\partial\mathrm{LB}(\alpha,\epsilon,\Delta_{1-0}(x,u))/\partial\alpha$ can be expressed as $\Delta_{1-0}(x,u) + \int_\alpha^1 (1-t)/(t(1-t)+\epsilon)dt - \int_0^\alpha t/(t(1-t)+\epsilon)dt$. Let $t_+ = \frac{1+\sqrt{1+4\epsilon}}{2}$ and $t_- = \frac{1-\sqrt{1+4\epsilon}}{2}$. Since $\int \frac{t}{t(1-t)+\epsilon}dt = -\frac{t_+}{t_+-t_-}\log|t-t_+| + \frac{t_-}{t_+-t_-}\log|t-t_-| + C$ where C is the constant of integration and $\int(1-t)/(t(1-t)+\epsilon)dt = \int t/(t(1-t)+\epsilon)dt$, we can derive

$$\frac{\partial\mathrm{LB}(\alpha,\epsilon,\Delta_{1-0}(x,u))}{\partial\alpha} = \Delta_{1-0}(x,u) + \frac{1}{\sqrt{1+4\epsilon}}\log\left|\frac{\alpha - \frac{1+\sqrt{1+4\epsilon}}{2}}{\alpha - \frac{1-\sqrt{1+4\epsilon}}{2}}\right|.$$

By differentiating the first derivative w.r.t. α again, we have $\partial^2\mathrm{LB}(\alpha,\epsilon,\Delta_{1-0}(x,u))/\partial\alpha^2 = -1/\big(\alpha(1-\alpha)+\epsilon\big)$. □

Lemma 5. *The maximizer of LB$(\alpha,\epsilon,\Delta_{1-0}(x,u))$ over $\alpha \in [0,1]$ is $\alpha^*_{approx}(\epsilon,\Delta_{1-0}(x,u)) := \frac{1-\sqrt{1+4\epsilon}}{2} + \frac{\sqrt{1+4\epsilon}}{1+e^{-\sqrt{1+4\epsilon}\Delta_{1-0}(x,u)}}$ if and only if $\big|\Delta_{1-0}(x,u)\big| \leq \frac{1}{\sqrt{1+4\epsilon}}\log\frac{(\sqrt{1+4\epsilon}+1)^2}{4\epsilon} = -\log\epsilon + O(\epsilon\log\epsilon)$ as $\epsilon \to 0^+$.*

Proof of Lemma 5. By Lemma 4, the first partial derivative of $\mathrm{LB}(\alpha,\epsilon,\Delta_{1-0}(x,u))$ w.r.t. α is zero if and only if $\alpha = \alpha^*_{\mathrm{approx}}(\epsilon,\Delta_{1-0}(x,u))$. The solution of $\alpha^*_{\mathrm{approx}}(\epsilon,\Delta_{1-0}(x,u)) \in [0,1]$ can be expressed as $\big|\Delta_{1-0}(x,u)\big| \leq \frac{1}{\sqrt{1+4\epsilon}}\log\frac{(\sqrt{1+4\epsilon}+1)^2}{4\epsilon}$. Next, we prove $\frac{1}{\sqrt{1+4\epsilon}}\log\frac{(\sqrt{1+4\epsilon}+1)^2}{4\epsilon} = -\log\epsilon + O(\epsilon\log\epsilon)$ as $\epsilon \to 0^+$. We have

$$\begin{aligned}&\Big(\frac{1}{\sqrt{1+4\epsilon}}\log\frac{(\sqrt{1+4\epsilon}+1)^2}{4\epsilon} - (-\log\epsilon)\Big)\frac{1}{\epsilon\log\epsilon}\\ &= \frac{4}{\sqrt{1+4\epsilon}(\sqrt{1+4\epsilon}+1)} - \frac{2}{\sqrt{1+4\epsilon}}\frac{\log 2/(\sqrt{1+4\epsilon}+1)}{\epsilon\log\epsilon}.\end{aligned}$$

Here, the first term in RHS converges to 2 as $\epsilon \to 0^+$ and, by L'Hospital's rule, the limit of the second term is $\lim_{\epsilon\to 0^+}\frac{\log 2/(\sqrt{1+4\epsilon}+1)}{\epsilon\log\epsilon} = \lim_{\epsilon\to 0^+}\frac{-2/(\sqrt{1+4\epsilon}+1)\sqrt{1+4\epsilon}}{\log\epsilon+1} = 0$, which concludes the proof. □

Now, we prove Theorem 3. Let $t_+ = \frac{1+\sqrt{1+4\epsilon}}{2}$ and $t_- = \frac{1-\sqrt{1+4\epsilon}}{2}$. Since $\int\frac{t}{t(1-t)+\epsilon}dt = -\frac{t_+}{t_+-t_-}\log|t-t_+| + \frac{t_-}{t_+-t_-}\log|t-t_-| + C$ where C is the constant of integration and $\int(1-t)/(t(1-t)+\epsilon)dt = \int t/(t(1-t)+\epsilon)dt$, we can derive

$$\begin{aligned}&\mathrm{LB}(\alpha,\epsilon,\Delta_{1-0}(x,u))\\ &= \alpha\Delta_{1-0}(x,u) - \frac{\alpha t_+}{t_+-t_-}\log|\alpha-1+t_+| + \frac{\alpha t_-}{t_+-t_-}\log|\alpha-1+t_-|\\ &\quad - \frac{(1-\alpha)t_+}{t_+-t_-}\log|\alpha-t_+| + \frac{(1-\alpha)t_-}{t_+-t_-}\log|\alpha-t_-|\\ &\quad + \frac{t_+}{t_+-t_-}\log|t_+| - \frac{t_-}{t_+-t_-}\log|t_-|\\ &= \alpha\Delta_{1-0}(x,u) + \frac{1}{t_+-t_-}(\alpha-t_+)\log|\alpha-t_+| - \frac{1}{t_+-t_-}(\alpha-t_-)\log|\alpha-t_-|\\ &\quad + \frac{t_+}{t_+-t_-}\log|t_+| - \frac{t_-}{t_+-t_-}\log|t_-|.\end{aligned}$$

Here, the last equality is derived by using $t_+ + t_- = 1$. By Lemma 5, the maximizer is $\alpha^*_{\text{approx}}(\epsilon, \Delta_{1-0}(x,u)) = t_- + \sqrt{1+4\epsilon}\sigma(\sqrt{1+4\epsilon}\Delta_{1-0}(x,u))$ where $\sigma(x) := 1/(1+e^{-x})$ is the sigmoid function, so $\alpha^*_{\text{approx}}(\epsilon, \Delta_{1-0}(x,u)) - t_+ = -(t_+ - t_-)(1 - \sigma((t_+ - t_-)\Delta_{1-0}(x,u)))$ and $\alpha^*_{\text{approx}}(\epsilon, \Delta_{1-0}(x,u)) - t_- = (t_+ - t_-)\sigma((t_+ - t_-)\Delta_{1-0}(x,u))$. Now, substituting these equations and $(t_+ - t_-)\Delta_{1-0}(x,u) = \log\sigma((t_+ - t_-)\Delta_{1-0}(x,u))/(1 - \sigma((t_+ - t_-)\Delta_{1-0}(x,u)))$ gives

$$
\begin{aligned}
&\text{LB}(\alpha^*_{\text{approx}}(\epsilon, \Delta_{1-0}(x,u)), \epsilon, \Delta_{1-0}(x,u)) \\
&= t_-\Delta_{1-0}(x,u) + \sigma((t_+ - t_-)\Delta_{1-0}(x,u)) \log \frac{\sigma((t_+ - t_-)\Delta_{1-0}(x,u))}{1 - \sigma((t_+ - t_-)\Delta_{1-0}(x,u))} \\
&\quad - \sigma((t_+ - t_-)\Delta_{1-0}(x,u)) \log \Big((t_+ - t_-)\sigma((t_+ - t_-)\Delta_{1-0}(x,u)) \Big) \\
&\quad - (1 - \sigma((t_+ - t_-)\Delta_{1-0}(x,u))) \log \Big((t_+ - t_-)(1 - \sigma((t_+ - t_-)\Delta_{1-0}(x,u))) \Big) \\
&\quad + \frac{t_+}{t_+ - t_-} \log|t_+| - \frac{t_-}{t_+ - t_-} \log|t_-| \\
&= t_-\Delta_{1-0}(x,u) - \log\big(1 - \sigma((t_+ - t_-)\Delta_{1-0}(x,u))\big) \\
&\quad + \Big(-\log|t_+ - t_-| + \frac{t_+}{t_+ - t_-}\log|t_+| - \frac{t_-}{t_+ - t_-}\log|t_-| \Big) \\
&= \frac{1+\sqrt{1+4\epsilon}}{2}\Delta_{1-0}(x,u) - \log\sigma(\sqrt{1+4\epsilon}\Delta_{1-0}(x,u))) \\
&\quad + \Big(-\log\Big(\sqrt{1+4\epsilon}\Big) + \frac{1}{2\sqrt{1+4\epsilon}} \log \frac{(\sqrt{1+4\epsilon}+1)^2}{4\epsilon} + \frac{1}{2}\log\epsilon \Big).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\sup_{\alpha\in[0,1]} \text{ELBO}_{\theta,\phi}(\alpha; x, u) - \text{ELBO}_{\theta,\phi}(0; x, u) \\
&\geq \text{LB}(\alpha^*_{\text{approx}}(\epsilon, \Delta_{1-0}(x,u)), \epsilon, \Delta_{1-0}(x,u)) \\
&= \frac{1+\sqrt{1+4\epsilon}}{2}\Delta_{1-0}(x,u) - \log\sigma(\sqrt{1+4\epsilon}\Delta_{1-0}(x,u))) + O(\epsilon\log\epsilon)
\end{aligned}
$$

as $\epsilon \to 0^+$. Now, $-\log\sigma(\sqrt{1+4\epsilon}\Delta_{1-0}(x,u))) = o(\Delta_{1-0}(x,u))$ as $\Delta_{1-0}(x,u) \to \infty$ and $-\log\sigma(\sqrt{1+4\epsilon}\Delta_{1-0}(x,u))) = -\sqrt{1+4\epsilon}\Delta_{1-0}(x,u) + o(\Delta_{1-0}(x,u))$ as $\Delta_{1-0}(x,u) \to -\infty$ conclude the proof.

B Details on Experiments

B.1 Implementation Details

B.1.1 Dataset Description and Experimental Setting

We present in Table 1 the three data generation schemes used in the simulation study, which include: 1) distributions of covariates (U), 2) conditional distributions of latent variables given covariates ($Z|U$), and 3) conditional distributions of observations given latent variables ($X|Z$). Here, uniform and categorical distributions are denoted by Unif and Cat, respectively, and RealNVP (Dinh et al., 2017) is a flexible and invertible neural network mapping low-dimensional latent variables to high-dimensional observations. As in Zhou and Wei (2020), we use randomly initialized RealNVP networks as ground-truth mixing functions. The sample size is $30,000$ and the proportion of training, validation, and test samples are 80%, 10%, and 10%, respectively. The dimension of observations is 100. The number of repeats is 20, and for all datasets and methods, we train five models with different initial weights. All reported results are from models yielding the minimum validation loss and evaluated on the test dataset.

We provide descriptions on real datasets with implementation details.

EMNIST: An image dataset consisting of handwritten digits whose data format is the same as MNIST (LeCun, 1998) and has six split types. We use EMNIST split by digits to use images as observations (X) and digit labels as covariates (U). The official training dataset contains 240,000 images of digits from 0 to 9 in 28×28 gray-scale, and the test dataset contains 40,000 images. We randomly split the official training images by 200,000 and 40,000 images to make training and validation datasets for our experiments, and the number of repeats is 20.

Fashion-MNIST: An image dataset consisting of fashion-item images with item labels. There are ten classes such as ankle boot, bag, and coat, and we use images as observations (X) and item labels as covariates (U). The official training and test datasets contain 60,000 and 10,000 images in 28×28 gray-scale, respectively, and we randomly split the official training images by 50,000 and 10,000 images to make training and validation datasets. The number of repeats is 20.

ABCD: The ABCD study recruited 11,880 children aged 9–10 years (and their parents/guardians) were across 22 sites with 10-year-follow-up. For this analysis, we are using the baseline measures. After list-wise deletion for missing values, the sample size is 5,053, and the dimension of observations is 1,178. We conduct 5-fold cross-validation. For all data splits and methods, we train four models with different initial weights. All reported results are from the model yielding the minimum loss on the validation fold and evaluated on the test fold.

B.1.2 Network Architectures

In all experiments, iVAEs and CI-iVAEs use the same architectures of the label prior, encoder, and decoder networks for the purpose of fair comparison.

In the simulation study, we modify the official implementation code of pi-VAE. The architectures of label prior and encoder networks are Dense(60)-Tanh-Dense(60)-Tanh-Dense(2) for the sine latent structure and Dense(60)-Tanh-Dense(60)-Tanh-Dense(60)-Tanh-Dense(2) for quadratic and two circles latent structures. As in pi-VAE, we assume $q(z|x,u) \propto q_\phi(z|x)p_{T,\lambda}(z|u)$. The $q(z|x,u)$ is a Gaussian distribution since both label prior and encoder are Gaussian. The means and variances of $q(z|x,u)$ can be computed with those of label prior and encoder. The architecture of the decoder is the same as the modified GIN used in pi-VAE to guarantee injectivity. We use Adam optimizer (Kingma and Ba, 2014). The number of epochs, batch size, and the learning rate is 100, 300, and 5×10^{-4}, respectively.

In experiments on EMNIST and Fashion-MNIST datasets,

Table 4: A summary of distributions of variables in the simulation study. Examples can be founded at the first column in Figure 1 in the manuscript.

LATENT STRUCTURE	VARIABLES		
	COVARIATES (U)	LATENT VARIABLES GIVEN COVARIATES ($Z\|U$)	OBSERVATIONS GIVEN LATENT VARIABLES ($X\|Z$)
SINE	$\text{UNIF}(0, 2\pi)$	$N\big((U, 2\text{SIN}U)^T, (U/4\pi)I_2\big)$	$N(\text{REALNVP}(Z), I_{100})$
QUADRATIC	$\text{UNIF}(-\pi/2, \pi/2)$	$N\big((U, U^2)^T, (2U+\pi)/4\pi I_2\big)$	$N(\text{REALNVP}(Z), I_{100})$
TWO CIRCLES	$\text{UNIF}(-\pi, \pi) \times \text{CAT}_2(0.5, 0.5)$	$N\big((U_2\text{COS}U_1, U_2\text{SIN}U_1)^T, (-\|U_1\| + \pi)/10\pi I_2\big)$	$N(\text{REALNVP}(Z), I_{100})$

Table 5: Contingency tables to display the number of data by their α^* computed by grid search (column) and formula (row) on sine latent structure. Correlation coefficients between α^* by grid search and by formula are presented at the top-left corner.

LATENT STRUCTURE = SINE				
		FORMULA		
	CORRELATION COEFFICIENT: 0.99 (0.00)	0	IN-BETWEEN 0 AND 1	1
GRID	0	26.85% (0.18%)	1.88% (0.06%)	0.00% (0.00%)
	IN-BETWEEN 0 AND 1	0.19% (0.01%)	0.14% (0.02%)	0.01% (0.00%)
	1	0.00% (0.00%)	1.99% (0.05%)	68.95% (0.19%)
LATENT STRUCTURE = QUADRATIC				
		FORMULA		
	CORRELATION COEFFICIENT: 0.99 (0.00)	0	IN-BETWEEN 0 AND 1	1
GRID	0	30.95% (0.19%)	2.61% (0.06%)	0.00% (0.00%)
	IN-BETWEEN 0 AND 1	0.30% (0.02%)	0.28% (0.02%)	0.04% (0.01%)
	1	0.00% (0.00%)	2.68% (0.08%)	63.14% (0.26%)
LATENT STRUCTURE = TWO CIRCLES				
		FORMULA		
	CORRELATION COEFFICIENT: 0.99 (0.00)	0	IN-BETWEEN 0 AND 1	1
GRID	0	36.05% (0.19%)	3.44% (0.08%)	0.00% (0.00%)
	IN-BETWEEN 0 AND 1	0.39% (0.02%)	0.29% (0.02%)	0.02% (0.01%)
	1	0.00% (0.00%)	3.46% (0.08%)	56.36% (0.27%)

we modify the official implementation code of GIN. For GIN, we use the same architecture used in the GIN paper. For iVAEs and CI-iVAEs, the architecture of encoders is Conv(32, 3, 1, 1)-BN-LReLU-Conv(64, 4, 2, 1)-BN-LReLU-Conv(128, 4, 2, 1)-BN-LReLU-Conv(128, 7, 1, 0)-BN-LReLU-Dense(64), that of decoders is ConvTrans(128, 1, 1, 0)-BN-LReLU-ConvTrans(128, 7, 1, 0)-BN-LReLU-ConvTrans(64, 4, 2, 1)-BN-LReLU-ConvTrans(32, 4, 2, 1)-BN-LReLU-ConvTrans(1, 3, 1, 1)-Sigmoid, and that of the label prior is Dense(256)-LReLU-Dense(256)-LReLU-Dense(64). Here, Conv(f, k, s, p) and ConvTrans(f, k, s, p) denote the convolution layer and transposed convolution layer (Zeiler et al., 2010), respectively, where f, k, s, and p are the number of output channel, kernel size, stride, and padding, respectively. BN denotes the batch normalization layer (Ioffe and Szegedy, 2015), and LReLU denotes the Leaky ReLU activation layer (Xu et al., 2015). The initialized decoders are not injective, but our objective functions encourage them to be injective by enforcing the inverse relation between encoders and decoders. The number of learnable parameters of GIN architectures is 2,620,192, and that of iVAEs and CI-iVAEs is 2,062,209. We use Adam optimizer. The number of epochs and batch size is 100 and 240, respectively. The learning rate is 3×10^{-4} for the first 50 epochs and is 3×10^{-5} for the remaining epochs.

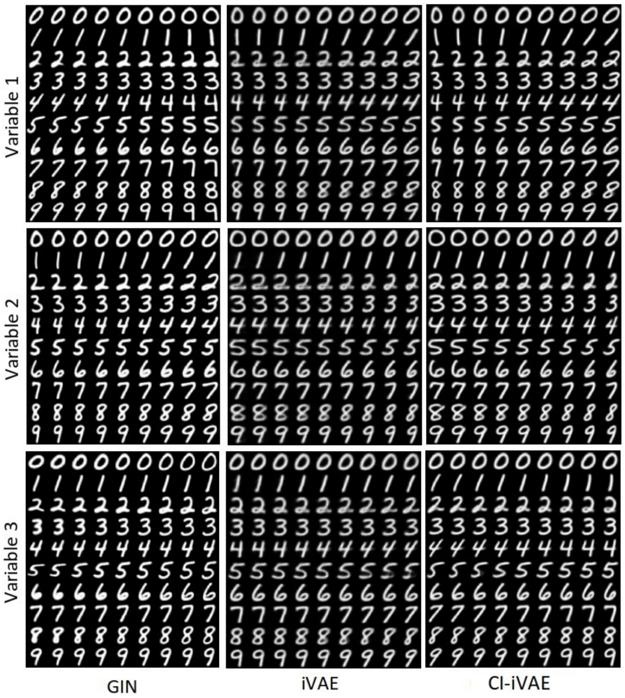

Figure 5: Generation results on EMNIST by varying top three latent attributes having the largest standard deviations. We calculate mean vector of latent variables and controlling the selected attribute from -2 to $+2$ standard deviations.

In the experiment on the ABCD dataset, the architecture of the label prior is Dense(256)-LReLU-Dense(256)-LReLU-Dense(128), that of encoders is Dense(4096)-BN-LReLU-Dense(4096)-BN-LReLU-Dense(4096)-BN-LReLU-Dense(4096)-BN-LReLU-Dense(128)-BN-LReLU-Dense(128), and that of decoders is Dense(4096)-LReLU-Dense(4096)-LReLU-Dense(4096)-LReLU-Dense(4096)-LReLU-Dense(128)-LReLU-Dense(128). We use Adam optimizer. The number of epochs, batch size, and the learning rate is 100 and 64, and 2×10^{-4}, respectively.

B.2 Further Experimental Results

We present further experimental results on the simulation study in Table 2.

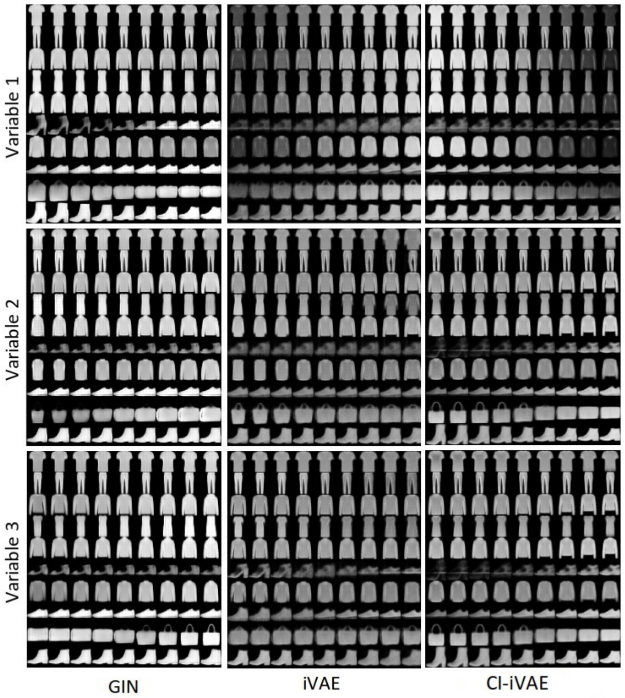

Figure 6: Generation results on Fashion-MNIST by varying top three latent attributes having the largest standard deviations. We calculate mean vector of latent variables and controlling the selected attribute from -2 to $+2$ standard deviations.

We present contingency tables for samplewise optimal α computed by grid search and by using approximating formula (Equation (7)) in Table 2. For grid search, we calculate $\mathrm{ELBO}_{\theta,\phi}(\alpha; x, u)$ for $\alpha \in \{0, 0.001, ..., 0.999, 1\}$ and pick the maximizer. For formula, we approximate ϵ with $f(\mathbf{z}) = z_j$ and $f(\mathbf{z}) = z_j^2$ for $j = 1, ..., d_Z$ and calculate α^*_{approx}. For all three settings, the correlation coefficients are high, which indicates the consistency of α^* from the proposed algorithm with theoretical approximation. Moreover, α^* does not degenerate at 0 or 1, so the proposed ELBO using samplewise optimal posteriors is different from the two ablation cases, ELBOs with $q_\phi(z|x, u)$ and with $q_\phi(z|x)$.

Generation results according to attributes having the largest standard deviations are provided in Figures 1 and 2. For all methods, the generated result changes as the value of attributes are changed. In Fashion-MNIST, iVAE-based methods change the contrast of fashion items while GIN does not.

References

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Bengio, Y. et al. (2009). Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127.

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, pages 153–160.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2016). Generating sentences from a continuous space. In *In Proceedings of Conference on Computational Natural Language Learning*.

Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. (2017). Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE.

Dai, B., Wang, Z., and Wipf, D. (2020). The usual suspects? reassessing blame for vae posterior collapse. In *International Conference on Machine Learning*, pages 2313–2322. PMLR.

Dieng, A. B., Kim, Y., Rush, A. M., and Blei, D. M. (2019). Avoiding latent variable collapse with generative skip models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2397–2405. PMLR.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). Density estimation using real nvp. In *International Conference on Learning Representations*.

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., and Carin, L. (2019). Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *In Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics*.

Gordon, E. M., Laumann, T. O., Adeyemo, B., Huckins, J. F., Kelley, W. M., and Petersen, S. E. (2016). Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral cortex*, 26(1):288–303.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Han, K., Wen, H., Shi, J., Lu, K.-H., Zhang, Y., Fu, D., and Liu, Z. (2019). Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, 198:125–136.

He, J., Spokoyny, D., Neubig, G., and Berg-Kirkpatrick, T. (2019). Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*.

Huang, C.-W., Tan, S., Lacoste, A., and Courville, A. C. (2018). Improving explorability in variational inference with annealed variational objectives. *Advances in Neural Information Processing Systems*, 31.

Huang, H., Hu, X., Zhao, Y., Makkie, M., Dong, Q., Zhao, S., Guo, L., and Liu, T. (2017). Modeling task fmri data via deep convolutional autoencoder. *IEEE transactions on medical imaging*, 37(7):1551–1561.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR.

Jernigan, T. L., Brown, S. A., and Dowling, G. J. (2018). The adolescent brain cognitive development study. *Journal of research on adolescence: the official journal of the Society for Research on Adolescence*, 28(1):154.

Kameoka, H., Kaneko, T., Tanaka, K., and Hojo, N. (2018). Acvae-vc: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder. *arXiv preprint arXiv:1808.05092*.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR.

Kim, J.-H., Zhang, Y., Han, K., Wen, Z., Choi, M., and Liu, Z. (2021). Representation learning of resting state fmri with variational autoencoder. *NeuroImage*, 241:118423.

Kim, Y., Wiseman, S., Miller, A., Sontag, D., and Rush, A. (2018). Semi-amortized variational autoencoders. In *International Conference on Machine Learning*, pages 2678–2687. PMLR.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., Mohamed, S., Jimenez Rezende, D., and Welling, M. (2014). Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kong, L., Xie, S., Yao, W., Zheng, Y., Chen, G., Stojanov, P., Akinwande, V., and Zhang, K. (2022). Partial disentanglement for domain adaptation. In *International Conference on Machine Learning*, pages 11455–11472. PMLR.

LeCun, Y. (1998). The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lehmann, E. L., Romano, J. P., and Casella, G. (2005). *Testing statistical hypotheses*, volume 3. Springer.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Lu, H., Liu, S., Wei, H., Chen, C., and Geng, X. (2021). Deep multi-kernel auto-encoder network for clustering brain functional connectivity data. *Neural Networks*, 135:148–157.

Lucas, J., Tucker, G., Grosse, R. B., and Norouzi, M. (2019). Don't blame the elbo! a linear vae perspective on posterior collapse. *Advances in Neural Information Processing Systems*, 32.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. (2015). Adversarial autoencoders. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Nishiyama, T. (2019). A new lower bound for kullback-leibler divergence based on hammersley-chapman-robbins bound. *arXiv preprint arXiv:1907.00288*.

Pinaya, W. H., Mechelli, A., and Sato, J. R. (2019). Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study. *Human brain mapping*, 40(3):944–954.

Qiang, N., Dong, Q., Sun, Y., Ge, B., and Liu, T. (2020). deep variational autoencoder for modeling functional brain networks and adhd identification. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 554–557. IEEE.

Razavi, A., van den Oord, A., Poole, B., and Vinyals, O. (2019). Preventing posterior collapse with delta-vaes. In *International Conference on Learning Representations*.

Ribeiro, M., Singh, S., and Guestrin, C. (2019). Local interpretable model-agnostic explanations (lime): An introduction.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica: Journal of the Econometric Society*, pages 577–591.

Schneider, S., Lee, J. H., and Mathis, M. W. (2022). Learnable latent embeddings for joint behavioral and neural analysis. *arXiv preprint arXiv:2204.00673*.

Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491.

Sorrenson, P., Rother, C., and Köthe, U. (2020). Disentanglement by nonlinear ica with general incompressible-flow networks (gin). *arXiv preprint arXiv:2001.04872*.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Wang, X., Saxon, M., Li, J., Zhang, H., Zhang, K., and Wang, W. Y. (2022). Causal balancing for domain generalization. *arXiv preprint arXiv:2206.05263*.

Wang, Y., Blei, D., and Cunningham, J. P. (2021). Posterior collapse and latent variable non-identifiability. *Advances in Neural Information Processing Systems*, 34:5443–5455.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Yang, Z., Hu, Z., Salakhutdinov, R., and Berg-Kirkpatrick, T. (2017). Improved variational autoencoders for text modeling using dilated convolutions. In *International conference on machine learning*, pages 3881–3890. PMLR.

Zeiler, M. D., Krishnan, D., Taylor, G. W., and Fergus, R. (2010). Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition*, pages 2528–2535. IEEE.

Zhou, D. and Wei, X.-X. (2020). Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-vae. In *NeurIPS*.