VARIATIONAL LEARNING OF DISENTANGLED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Disentangled representations allow models to separate factors shared across conditions from those that are condition-specific. This separation is crucial in domains such as biomedicine, where generalization to new treatments, patients, or species requires isolating stable biological signals from context-dependent effects. While several VAE-based extensions aim to achieve this, they often exhibit leakage between latent variables, limiting generalization. We introduce DisCoVR, a variational framework that explicitly separates invariant and condition-specific factors through: (i) a dual-latent architecture, (ii) parallel reconstructions to keep both representations informative, and (iii) a max–min objective that enforces separation without handcrafted priors. We show this objective maximizes data likelihood, promotes disentanglement, and admits a unique equilibrium. Empirically, DisCoVR achieves stronger disentanglement on synthetic data, natural images, and single-cell RNA-seq datasets, establishing it as a principled approach for multi-condition representation learning.

1 INTRODUCTION

Neural network–based models excel at learning rich representations of complex data and are increasingly applied in settings where each data point $x \in \mathcal{X} \subseteq \mathbb{R}^d$ is associated with a condition label $y \in 1, \ldots, K$. In biology, for example, conditions may represent treatments, patients, or species. Such representations are valuable for tasks like domain adaptation and transfer learning (Pan et al., 2010), where models must generalize from observed to novel conditions. Achieving this requires disentangled representations that separate factors shared across conditions from those specific to each y.

Generative models provide a natural framework for uncovering latent structure, with prominent examples including Generative Adversarial Networks (GANs) (Goodfellow et al., 2020), Variational Autoencoders (VAEs) (Kingma & Welling, 2014), and diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020). Among these, VAEs and their extensions are particularly well-suited to transfer learning and domain adaptation (Akrami et al., 2020; Lovrić et al., 2021; Godinez et al., 2022; Zhang et al., 2023), thanks to their probabilistic foundation and ability to capture uncertainty.

Thus, several VAE-based methods have been proposed to integrate data across multiple conditions or sources (Xu et al., 2021; Lotfollahi et al., 2019; Boyeau et al., 2022), but only a few explicitly disentangle shared and condition-specific components (Sohn et al., 2015; Klys et al., 2018; Joy et al., 2020). While these approaches improve separation to some degree, they often depend on handcrafted priors that are difficult to design in high-dimensional domains like single-cell genomics, and they frequently exhibit residual leakage between latent varaibles, limiting generalization across conditions.

In this work, we introduce a framework for learning *disentangled representations in multi-condition datasets*. Our main contributions are: (i) a method combining two distinct reconstruction objectives with adversarial learning, reducing reliance on restrictive priors or handcrafted components; (ii) a max–min formulation of disentangled representation learning, along with a corresponding objective and theoretical guarantees for its equilibrium; and (iii) through experiments on synthetic benchmarks and real-world datasets, we show that our approach consistently improves upon existing methods disentanglement of shared and condition-specific structure.

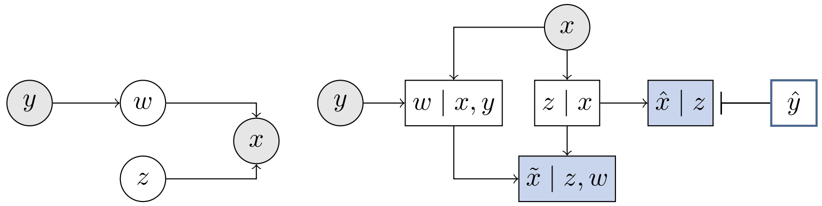
2 DisCoVR: Disentangling Common and Variant Representations

For the task of learning disentangled representations from multi-condition data, we consider a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ consisting of inputs $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ collected from associated condition labels $y_i \in \{1, \ldots, K\}$. For each class k (corresponding to a study or experimental condition), the associated subset $\mathcal{D}_k \coloneqq \{x_i : y_i = k\}$ consists of i.i.d. samples drawn from a class-conditional distribution $p(x \mid y = k)$.

2.1 Model assumptions

We assume that the data is generated by latent variables z and w, such that the joint distribution $p(x, y, z, w)$ factorizes according to the probabilistic graphical model illustrated in Figure 1a, i.e.,

$$p(x, y, z, w) = p(y)\, p(w \mid y)\, p(z)\, p(x \mid z, w). \tag{1}$$



(a) (b)

Figure 1: Graphical overview of our model. (a) Probabilistic graphical model: gray circles denote observed variables, white show latent variables. (b) Encoder–decoder architecture: the inhibition arrow from $\hat{y}$ to $\hat{x}$ corresponds to the adversarial component.

This model encodes two key conditional independence assumptions:

1. *Latent variable conditional independence:* Given the condition y, the latent representations z and w are conditionally independent: $z \perp w \mid y$.
2. *Sufficiency of the shared latent representation:* The input x is conditionally independent of the condition y given w: $x \perp y \mid w$.

However, note that in this formulation, z and w are no longer independent if conditioned also on x, that is, $z \not\perp w \mid x, y$.

2.2 Target posterior structure

In our model, each observation x is generated from two latent variables: z, which is *condition-invariant*, and w, which is *condition-aware* through its dependence on y. Our goal is to learn probabilistic representations where the marginals of z and w preserve this structure, yielding disentangled factors. Thus, we seek to approximate the posterior $p_{z,w|x,y}$.

However, approximating the full posterior with a variational distribution $q_{z,w|x,y}$ is intractable: even for simple variational families such as Gaussians, modeling the dependencies between z and w requires a full covariance structure, which is computationally prohibitive in high dimensions. To mitigate this, we employ a factorized approximation $q_{z|x}\, q_{w|x,y}$.

Our variational approximation is guided by two complementary objectives: (i) $q_{z|x}$ closely approximating the marginal posterior $p_{z|x}$; and (ii) the product $q_{z|x}\, q_{w|x,y}$ closely approximating the true posterior $p_{z,w|x,y}$. Formally, given variational families[1] $\mathcal{Q}_z$ and $\mathcal{Q}_w$ we seek to find $q^*_{z|x} \in \mathcal{Q}_z$ and $q^*_{w|x,y} \in \mathcal{Q}_w$ that minimize the following sum of Kullback-Leibler (KL) divergences:

$$q^*_{z|x},\, q^*_{w|x,y} = \arg \min_{\substack{q_{z|x} \in \mathcal{Q}_z \\ q_{w|x,y} \in \mathcal{Q}_w}} D_{\mathrm{KL}}(q_{z|x} \parallel p_{z|x}) + D_{\mathrm{KL}}(q_{z|x} q_{w|x,y} \parallel p_{z,w|x,y}). \tag{2}$$

[1] Here we consider general families and specify our concrete choices in §2.4.

2.3 Optimization objective

Since direct evaluation of the KL divergences in Equation 2 is intractable, we optimize a surrogate objective consisting of two ELBO terms.

The corresponding ELBO objective to minimize $D_{\mathrm{KL}}(q_{z|x} \parallel p_{z|x})$ is

$$\mathcal{L}_z(q_{z|x}, p; x) \coloneqq \mathbb{E}_{q_{z|x}}\left[\log p(x \mid z)\right] - D_{\mathrm{KL}}(q_{z|x} \parallel p_z), \tag{3}$$

and the ELBO objective for the second KL term, $D_{\mathrm{KL}}(q_{z|x}q_{w|x,y} \parallel p_{z,w|x,y})$, is given by

$$\mathcal{L}_w(q_{w|x,y}, p; x, y) \coloneqq \mathbb{E}_{q_{z|x}}\left[\mathbb{E}_{q_{w|x,y}}[\log\left(p(x|z,w)\right)]\right] - D_{\mathrm{KL}}\left(q_{z|x} \parallel p_z\right) - D_{\mathrm{KL}}\left(q_{w|x,y} \parallel p_{w|y}\right). \tag{4}$$

Note that $\mathcal{L}_w(q_{w|x,y}, p; x, y)$ is the ELBO objective that corresponds to a factorized posterior $q_{z|x}q_{w|x,y}$. In Proposition 2.1 we examine the gap between this objective, and an ELBO term corresponding to a full variational posterior $q_{z,w|x,y}$. This can be interpreted as the cost of enforcing a condition-invariant latent representation, specifically, constraining z to depend only on x. Proposition

Proposition 2.1. *For random variables x, y, z, w following the graphical model in Figure 1a,*

$$\mathrm{ELBO}(q, p; x, y) - \mathcal{L}_w(q_{w|x,y}, p; x, y) = \mathbb{E}_{q_{w|x,y}}\left[KL\left(q_{z|x} \parallel p_{z|w,x,y}\right)\right].$$

where $\mathrm{ELBO}(q, p; x, y) \coloneqq \log p(x \mid y) - D_{\mathrm{KL}}\left(q_{w|x,y} \parallel p_{w|x,y}\right)$.

The proof is provided in Appendix B.1.

Note that a full definition of the objectives in Equations 3 and 4 requires the specification of corresponding prior distributions, namely p_z and $p_{w|y}$. We defer their definitions to §2.4.

Equation 4 provides an evidence lower bound on the conditional log-likelihood $\log p(x \mid y)$. By adding $\log p(y)$, this bound extends to the joint log-marginal likelihood $\log p(x, y)$. Beyond optimizing this objective, we aim to ensure that the marginal distribution over y implicitly induced by the latent representations is consistent with the true $p(y)$.

To this end, we introduce an auxiliary classifier $g(y \mid z)$ as a form of posterior regularization (Ganchev et al., 2010). This classifier captures the residual predictive signal about y in z and is trained by minimizing the expected negative log-likelihood $-\mathbb{E}_{q(z|x)} \log g(y \mid z)$. If z is truly independent of y, then $g(y \mid z)$ will approximate the marginal distribution $p(y)$. By penalizing deviations from this behavior, we enforce the structural constraint $z \perp y$ in the learned representation.

Note that for this term to effectively encourage $q_{z|x}$ to discard condition-specific information, the classifier $g_{y|z} \in \mathcal{G}$ must be trained adversarially, with its own update step. This prevents degenerate solutions in which the loss is minimized without removing information about y from z, for example, by collapsing g to a constant predictor that ignores its input.

Combining the three terms above, we define the objective

$$\mathcal{L}(q_{z|x}, q_{w|x,y}, g_{y|z}; x, y) = \mathcal{L}_z(q_{z|x}, p; x) + \mathcal{L}_w(q_{w|x,y}, p; x, y) - \mathbb{E}_{q_{z|x}} \log g(y \mid z), \tag{5}$$

which can be explicitly expressed as

$$\begin{aligned}\mathcal{L}(q_{z|x}, q_{w|x,y}, g_{y|z}; x, y) \coloneqq& \mathbb{E}_{q_{z|x}}[\log p(x \mid z)] + \mathbb{E}_{q_{z|x}}\left[\mathbb{E}_{q_{w|x,y}}[\log p(x \mid z, w)]\right] \\ &- \mathbb{E}_{q_{z|x}}[\log g(y \mid z)] - 2D_{\mathrm{KL}}(q_{z|x} \parallel p_z) - D_{\mathrm{KL}}\left(q_{w|x,y} \parallel p_{w|y}\right). \end{aligned}\tag{6}$$

Finally, to enable flexible trade-offs between reconstruction expressiveness and disentanglement, we introduce weighting terms $\alpha = (\alpha_1, \alpha_2)$ into the objective following the motivation of β-VAEs (Higgins et al., 2017):

$$\begin{aligned}\mathcal{L}_\alpha(q_{z|x}, q_{w|x,y}, g_{y|z}; x, y) \coloneqq& \mathbb{E}_{q_{z|x}}[\log p(x \mid z)] + \mathbb{E}_{q_{z|x}}\left[\mathbb{E}_{q_{w|x,y}}[\log p(x \mid z, w)]\right] \\ &- \mathbb{E}_{q_{z|x}} \log g(y \mid z) - \alpha_1 D_{\mathrm{KL}}(q_{z|x} \parallel p_z) - \alpha_2 D_{\mathrm{KL}}\left(q_{w|x,y} \parallel p_{w|y}\right). \end{aligned}\tag{7}$$

Accordingly, the mean weighted objective is suitable for max-min optimization of the form:

$$\max_{q_{z|x} \in \mathcal{Q}_z} \max_{q_{w|x,y} \in \mathcal{Q}_w} \min_{g_{y|z} \in \mathcal{G}} \mathbb{E}_{p_{x,y}}\left[\mathcal{L}_\alpha(q_{z|x}, q_{w|x,y}, g_{y|z}; x, y)\right]. \tag{8}$$

2.4 Latent prior models and variational approximations

Prior specification We place a standard Normal prior over the condition-invariant latent variable, $p_z = \mathcal{N}(0, I)$, which reflects a non-informative prior belief over its values.

For the condition-aware latent variable w, we define a class-conditional Gaussian prior $p_{w|y}$. As a simple choice, we let w have the same dimensionality as z and specify

$$p(w \mid y = k) = \mathcal{N}(\mu_k, I), \qquad \mu_k := \mathbb{E}_{p_{x|y=k}}\left[\mathbb{E}_{q_{z|x}}[z]\right]. \tag{9}$$

Here μ_k is the mean of the inferred latent representations z within the k-th class[2].

This specification induces a coupling between the two latent variables through the data distribution. Aligning $p_{w|y}$ with the class-wise expectations of the invariant variable, further encourages $q_{z|x}$ to encode informative representations, since capturing the shared structures will now not only increase $\mathcal{L}_z(q_{z|x}, p; x)$, but also decrease $D_{\mathrm{KL}}\left(q_{w|x,y} \,\|\, p_{w|y}\right)$, and thus increase $\mathcal{L}_w(q_{w|x,y}, p; x, y)$.

Importantly, for a truly condition-agnostic $q_{z|x}$, the conditional expectations μ_k will collapse to a shared mean $\mu := \mathbb{E}_{p_x}\left[\mathbb{E}_{q_{z|x}}[z]\right]$. In this case $p_{w|y}$ becomes a shared prior across classes, centered at a meaningful point in the latent space, rather than an uninformative one.

As the following proposition establishes, this anchoring of the prior $p_{w|y}$ in the variational distribution $q_{z|x}$ preserves the convex–concave structure of the objective, ensuring that the resulting max-min problem has a unique optimal solution.

Proposition 2.2. *Let $\mathcal{Q}_z$ and $\mathcal{Q}_w$ be convex parametric families of variational distributions over z and w, respectively, and let $\mathcal{G}$ denote a convex set of classifiers such that $g(x) \in [0,1]$ for all $g \in \mathcal{G}$. Assume the latent priors are given by $z \sim p(z)$ and $p(w|y) = \mathcal{N}(\mu_y, I)$, where $p(z)$ is a continuous strictly positive distribution, and $\mu_y = \mathbb{E}_{p_{x|y}}\left[\mathbb{E}_{q_{z|x}}[z]\right]$. Then, under standard regularity conditions (see Appendix B.2.1), there exists a unique saddle point:*

$$\left(q^*_{z|x}, q^*_{w|x,y}, g^*_{y|z}\right) = \max_{q_{z|x} \in \mathcal{Q}_z} \max_{q_{w|x,y} \in \mathcal{Q}_w} \min_{g_{y|z} \in \mathcal{G}} \mathcal{L}(q_{z|x}, q_{w|x,y}, g_{y|z}).$$

The proof is provided in Appendix B.2.2.

Specification of variational families We set both variational families $\mathcal{Q}_z$ and $\mathcal{Q}_w$ as d-dimensional Gaussian distributions with diagonal covariance matrices. Accordingly, each variational distribution is parameterized by a mean vector $\mu \in \mathbb{R}^d$ and a vector of variances $\sigma^2 \in \mathbb{R}^d_+$ corresponding to the diagonal of the covariance matrix, yielding $\theta = (\mu, \sigma^2)$.

3 Encoder-decoder model

In order to optimize the objective in Equation 8 with respect to $q_{z|x}$, $q_{w|x,y}$, and $g_{y|z}$ over the dataset $\mathcal{D}$, we introduce an encoder-decoder framework (illustrated in Figure 1b). In this framework, two separate reconstructions of x are generated: one, denoted $\hat{x} \sim p_{x|z}$, where z is sampled from the condition-invariant posterior $q_{z|x}$, and the other, denoted $\tilde{x} \sim p_{x|z,w}$, where in addition, w is sampled from the condition-aware posterior $q_{w|x,y}$. The corresponding algorithm is summarized in Algorithm 1.

Condition agnostic representation An input $x \in \mathcal{X}$ is mapped to the variational parameters $\theta_z = (\mu_z, \sigma^2_z)$ by an encoder neural network $f^z_\phi : \mathcal{X} \to \mathbb{R}^d \times \mathbb{R}^d_+$ parametrized by weights ϕ. A latent encoding $z \sim q_{\theta_z}$ is then sampled and mapped to a reconstruction $\hat{x}$ via a decoder neural network $h^z_\psi : \mathbb{R}^d \to \mathcal{X}$ parametrized by weights ψ.

Adversarial classifier Rather than learning a complex classifier from z to y, we use the reconstruction $\hat{x}$ from z, and predict y from $\hat{x}$ via a simple multinomial logistic regression $g_\beta : \hat{\mathcal{X}} \to [0,1]^K$, with class-specific weights $\beta = \beta_k{}_{k=1}^{K}$, or a shallow MLP. Since $\hat{x}$ is a deterministic function of z, this is equivalent to applying a restricted classifier on z. By the data processing inequality, such

[2] Similarly, for different dimensions of z and w the mean aggregation can be replaced with a neural network that maps the inferred representations z for each class to the parameters of the Gaussian prior.

a classifier can only capture a subset of the information z contains about y; thus, maximizing this lower bound on $I(z; y)$ also maximizes $I(z; y)$ itself. Although this substitution weakens the estimation of the cross-entropy term $-\mathbb{E}qz \mid x \log g(y \mid z)$, from an information-theoretic standpoint, we observed this to be often advantageous in practice: predicting y from $\hat{x}$ reduces the variance introduced by sampling $z \sim q_{\theta_z}$, providing a regularizing effect that prevents q_{θ_z} from overfitting to noisy classifier signals.

Condition aware representation A labeled input pair $(x, y) \in \mathcal{X} \times \{1, \dots, K\}$ is mapped to the parameters $\theta_w = (\mu_w, \sigma_w^2)$ using an encoder neural-network $f_\rho^w : \mathcal{X} \times \{1, \dots, K\} \to \mathbb{R}^d \times \mathbb{R}_+^d$ parametrized by weights ρ. A sample $w \sim q_{\theta_w}$ is then drawn, and the pair (z, w) is mapped to a reconstruction $\tilde{x}$ via a decoder neural-network $h_\eta^{z,w} : \mathbb{R}^d \to \mathcal{X}$, parametrized by weights η. To compute the prior $p_{w|y}$, we estimate the class-specific mean as $\hat{\mu}_k := \frac{1}{n_k} \sum_{i; y_i = k} z_i$ where each $z_i \sim p(z \mid x_i)$ is sampled from the encoder given an input x_i with label $y_i = k$, and n_k is the number of training points with the label $y = k$.

Algorithm 1

```
Input: Data $\mathcal{D} = \{x_{1:n}, y_{1:n}\}$, number of training iterations $J$, initial parameters $\phi^{(0)}, \psi^{(0)}, \rho^{(0)}, \eta^{(0)}, \beta^{(0)}$, learning rates $\gamma_1, \gamma_2$, weighting terms $\alpha = (\alpha_1, \alpha_2)$
for $1 \leq j \leq J$ do
    Compute $\theta_z = f_{\phi^z}^z(x)$ and $\theta_w = f_{\rho^{(j-1)}}^w(x, y)$
    Sample condition invariant and aware latent variables $z \sim q_{\theta_z}$ and $w \sim w_{\theta_w}$
    Compute reconstructions $\hat{x} = h_{\psi^{(j-1)}}^z(z)$ and $\tilde{x} = h_{\eta^{(j-1)}}^{z,w}$
    Compute condition prediction $\hat{y} = g_{\beta^{(j-1)}}(\hat{x})$
    Update classifier parameters:
        $\beta^{(j)} \leftarrow \beta^{(j-1)} - \gamma_1 \nabla_\beta \mathcal{L}_\alpha(q_{z|x}, q_{w|x,y}, g_{y|z})\Big|_{\phi=\phi^{(j-1)}, \psi=\psi^{(j-1)}, \rho=\rho^{(j-1)}, \eta=\eta^{(j-1)}}$
    Update encoder and decoder parameters, $\Omega^{(j)} = \left(\phi^{(j)}, \psi^{(j)}, \rho^{(j)}, \eta^{(j)}\right)$,
        $\Omega^{(j)} \leftarrow \Omega^{(j-1)} - \gamma_2 \nabla_{\phi,\psi,\rho,\eta} \mathcal{L}_\alpha(q_{z|x}, q_{w|x,y}, g_{y|z})\Big|_{\beta=\beta^{(j)}}$
end for
Return: $\phi^{(J)}, \psi^{(J)}, \rho^{(J)}, \eta^{(J)}, \beta^{(J)}$
```

In practice, following standard approaches in VAE-based models, (i) we use a single-sample Monte Carlo estimate to approximate the expectations in Equation 8, and (ii) instead of directly sampling from q_θ, we employ the reparameterization trick to enable differentiable sampling. Specifically, we sample $\epsilon \sim \mathcal{N}(0, I)$ and obtain a sample from q_θ by applying a deterministic transformation of ϵ based on the variational parameters θ.

4 Comparison to previous approaches

Here we review VAE-based methods for disentangled representation learning, which form the primary basis for comparison with our approach. Broader related literature is discussed in Appendix A.

VAEs (Kingma & Welling, 2014) are generative models that learn latent representations by maximizing the evidence lower bound (ELBO) on the data log-likelihood:

$$\mathbb{E}_{q_{z|x}}[\log p(x \mid z)] - D_{\mathrm{KL}}\left(q_{z|x} \parallel p_z\right) \leq \log p(x),$$

where $(x, z) \sim p$, and $z|x \sim q$ is a latent variable inferred from data. VAEs consist of an encoder $q_{z|x}$ that maps inputs to latent distributions, and a decoder $p_{x|z}$ that reconstructs inputs from latent representations. The learning process frames posterior inference as KL-regularized optimization over a variational family $\mathcal{Q}$ aiming to approximate the posterior $p_{z|x}$ under a typically simple prior $p(z)$. Several VAE extensions were proposed to encourage disentanglement. These are discussed below.

Conditional VAEs (Sohn et al., 2015) incorporate supervision to the standard VAE model by conditioning both the encoder and decoder on an observed label y, yielding the following objective:

$$\mathbb{E}_{q_{z|x,y}}[\log p(x \mid z, y)] - D_{\mathrm{KL}}\left(q(z \mid x, y) \| p(z)\right) \leq \log p(x|y).$$

While this allows controlled generation and partial disentanglement between z and y, the latent variable z is still inferred from both x and y and thus can encode condition-specific information.

Conditional Subspace VAEs (CSVAEs) (Klys et al., 2018), explicitly factorize the latent space into a shared component z and a label-specific component w (see Supplementary Figure 1a). Similarly, their hierarchical extension Beker et al. (2024) introduces an intermediate latent variable between x and (z, w). As in our approach, to encourage disentanglement, CSVAEs introduce an adversarial regularization term that penalizes mutual information between z and y, thereby discouraging predictability of y from z. They are learned by optimizing the following lower bound on $\log p(x \mid y)$:

$$\mathbb{E}_{q_{z,w|x,y}}[\log p(x \mid w, y)] - \mathbb{E}_{q_{z|x}}[\textstyle\int g(y \mid z) \log g(y \mid z)\, dy] - D_{\mathrm{KL}}\left(q_{w|x,y} \,\|\, q_{w|y}\right) - D_{\mathrm{KL}}\left(q_{z|x} \,\|\, p_z\right).$$

Domain Invariant VAEs (DIVA) and Characteristic-capturing VAEs (CCVAE) DIVA (Ilse et al., 2020) and CCVAE (Joy et al., 2020), shown in Supplementary Figure 1b, introduce two latent variables, z and w, where w captures label-related features by jointly optimizing a classifier $p(w \mid y)$ jointly alongside the remaining objective. For fully supervised cases, the DIVA model optimizes

$$\mathbb{E}_{q_{z,w|x}}[\log p(x \mid z, w)] + \mathbb{E}_{q_{w|x}}[\log q(y \mid w)] - D_{\mathrm{KL}}\left(q_{z|x} \| p_z\right) - D_{\mathrm{KL}}\left(q_{w|x} \| p_{w|y}\right) \leq \log p(x, y).$$

This objective corresponds to the assumptions $x \perp y \mid z, w$ and $z \perp w$. Similarly, CCVAEs optimize the following lower bound on $\log p(x \mid y)$:

$$\mathbb{E}_{q_{z,w|x,y}}\left[\frac{q(y|w)}{q(y|x)} \log \frac{p(x|z,w)}{q(y|w)}\right] - D_{\mathrm{KL}}\left(q_{z|x} \| p_{z|y}\right) - D_{\mathrm{KL}}\left(q_{w|x} \| p_{w|y}\right) + \log q(y \mid x).$$

Comparison: In prior methods, reconstruction is performed jointly from both representations z and w, via $p(x \mid z, w)$. This design provides no incentive to distribute information meaningfully between z and w: the model can place all relevant information into w, leaving z either uninformative or entangled with w. Our method addresses this limitation through two key components: (i) a *separate reconstruction term from z alone*, which explicitly forces z to capture informative, condition-invariant structure. Theoretically, we show that the additional term $p(\hat{x} \mid z)$ is required to bound the gap between our approximate posterior and the full model ELBO under the factorization assumption; (ii) a *prior over w conditioned on the mean of z*, which, through label-specific aggregates, discourages leakage of class-invariant information into w and reduces redundancy between the two representations.

Another novelty of our approach is a *principled probabilistic objective* that enforces the correct conditional independence. Without an explicit probabilistic model, prior methods sought to enforce $z \perp w \mid x, y$. In contrast, our formulation shows that the proper requirement is the weaker condition $z \perp w \mid y$. As established in Section 2.1, our model satisfies $z \perp w \mid y$, but not $z \perp w \mid x, y$.

To enforce this criterion, we propose a *theoretically grounded variational objective* that uses a conditional ELBO for the dependent representation w, with the *prior over w conditioned on the mean of the independent representation z*.

5 Experiments

Datasets: We evaluate DisCoVR against existing approaches on synthetic data, natural images, and biological data. These datasets were chosen to probe condition-invariant structure and to ensure comparability with prior work: for instance, Swiss rolls and CelebA were used in Klys et al. (2018), and CelebA also in Joy et al. (2020).

Evaluation: When applicable, we evaluate reconstruction quality using negative log-likelihood (NLL), root mean squared error (RMSE), and the absolute deviation from the optimal-Bayes classifier on the reconstructed data, denoted as Δ–Bayes. Disentanglement is quantified via a neural estimator of the mutual information $I(z; w)$ (Belghazi et al., 2018). Full model architectures, hyperparameters, and additional implementation details are provided in Appendix H. Our results show that DisCoVR achieves superior performance across all experiments.

5.1 Simulated data

We begin with controlled synthetic experiments to isolate and visualize disentanglement.

5.1.1 Parametric model

Data generating model: Consider a model where the observed data x is generated as a function of two latent variables z and w, and y are binary labels. Assume that the marginal distributions of the latent variables are given by $z \sim \mathcal{N}(0, 1)$ and $w \sim \mathcal{N}(0, 1)$, and that the data x is generated as the sum of the two latent variables: $x = z + w$. Since z and w are both drawn from $\mathcal{N}(0, 1)$, it follows that $x \sim \mathcal{N}(0, 2)$. Finally, assume that the binary label is determined by the sign of w: $y = 1$ if $w > 0$, and $y = 0$ otherwise.

Optimal disentanglement: Given that z and w are independent and $x = z + w$, we have that $p(z \mid x) = \mathcal{N}\big(z; \frac{x}{2}, \frac{1}{2}\big)$. Hence, given x, the best estimate for z is $\frac{x}{2}$. Note that when ignoring the label y, the conditional distribution $p(w \mid x)$ is $p(w \mid x) = \mathcal{N}\big(w; \frac{x}{2}, \frac{1}{2}\big)$. However, the observation of y (which indicates whether w is positive or negative) truncates this distribution:

$$p(w \mid x, y = 1) = \frac{\mathcal{N}\big(w; \frac{x}{2}, \frac{1}{2}\big)}{\Phi\big(\frac{x}{\sqrt{2}}\big)}, \qquad p(w \mid x, y = 0) = \frac{\mathcal{N}\big(w; \frac{x}{2}, \frac{1}{2}\big)}{1 - \Phi\big(\frac{x}{\sqrt{2}}\big)}. \tag{10}$$

Results: Table 1 shows that DisCoVR (ours) best approximates the analytic posteriors, resulting in the lowest deviation from the optimal Bayes classifier and the best reconstruction.

Table 1: Parametric model results: DisCoVR (ours) outperforms all competitors across all metrics.

	NLL ↓	$D_{\mathrm{KL}}(q_{z\|x} \mid\mid p_{z\|x})$ ↓	$D_{\mathrm{KL}}(q_{w\|x,y} \mid\mid p_{w\|x,y})$ ↓	Δ – Bayes ↓
CSVAE - N.A.	1.810 ± 0.016	6.65 ± 3.46	23.61 ± 0.36	24.83 ± 0.04
CSVAE	1.786 ± 0.022	2.85 ± 1.11	23.98 ± 4.36	24.33 ± 1.28
HCSVAE - N.A.	1.784 ± 0.010	4.01 ± 0.07	25.82 ± 0.38	24.99 ± 0.01
HCSVAE	1.770 ± 0.004	3.99 ± 0.09	26.25 ± 0.59	24.99 ± 0.01
DIVA	1.788 ± 0.008	3.21 ± 1.52	12.88 ± 3.31	3.51 ± 0.32
CCVAE	1.785 ± 0.006	1.77 ± 0.81	12.95 ± 3.35	3.57 ± 0.15
DisCoVR (ours)	$\mathbf{1.769 \pm 0.003}$	$\mathbf{0.17 \pm 0.01}$	$\mathbf{10.10 \pm 0.73}$	$\mathbf{0.1 \pm 0.28}$

5.1.2 Noisy Swiss roll

Dataset: We use a noisy variant of the labeled Swiss Roll dataset (Marsland, 2014; Klys et al., 2018), generating $n = 20,000$ and assigning binary labels based on a lengthwise split, with labels flipped at rate ρ. The common geometry (its projection along the 2D plane) remains intact, the conditional structure along the third axis becomes noisy. Figure 2A illustrates the setup.

Optimal disentanglement: Since the Swiss Roll is sliced at the center and label noise is applied uniformly, marginalizing over labels yields a symmetric spiral centered along the roll—i.e., the marginal posterior $p(z \mid x)$ is label-invariant. In contrast, the conditional component retains a noisy but informative signal, with a uniform noise level of $\rho = 0.3$. As a result, the Bayes optimal classifier trained on any realistic representation is upper-bounded at 70% accuracy. **Results:** Figure 2 presents qualitative and quantitative results, showing that DisCoVR both models the label noise accurately and effectively disentangles shared and condition-specific structure. Notably, DisCoVR captures

Table 2: Noisy Swiss roll results: DisCoVR (ours) yields lowest deviation from optimal-Bayes, maintains low latent leakage, and high reconstruction accuracy.

	$I(z; w)$ ↓	NLL ↓	Δ – Bayes ↓
CSVAE - N.A.	0.047 ± 0.023	3.303 ± 0.003	23.88 ± 12.02
CSVAE	0.031 ± 0.025	$\mathbf{3.302 \pm 0.003}$	17.99 ± 14.58
HCSVAE - N.A.	0.024 ± 0.012	$\mathbf{3.302 \pm 0.002}$	30.00 ± 0.00
HCSVAE	$\mathbf{0.002 \pm 0.001}$	$\mathbf{3.302 \pm 0.002}$	30.00 ± 0.00
DIVA	0.336 ± 0.083	$\mathbf{3.302 \pm 0.003}$	1.88 ± 1.05
CCVAE	0.502 ± 0.089	$\mathbf{3.302 \pm 0.002}$	2.21 ± 0.84
DisCoVR (Ours)	0.005 ± 0.002	$\mathbf{3.302 \pm 0.002}$	$\mathbf{1.14 \pm 0.21}$

the marginal data distribution, successfully recovering the expected spiral pattern, as shown in Figure 2B (right). Additionally, the results in Table 2 show that DisCoVR achieves the lowest deviation from the optimal Bayes classifier and minimal information leakage between latent variables, while preserving reconstruction quality.

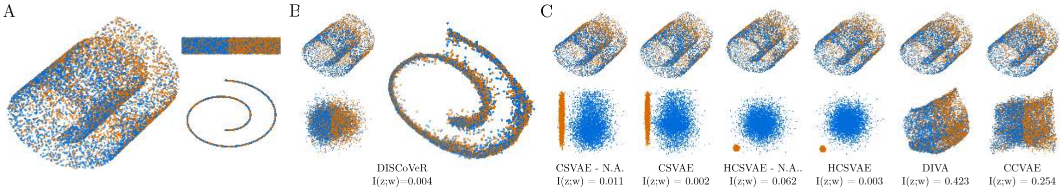


Figure 2: A: Noisy labeled Swiss Roll dataset. B (left): Reconstruction and conditional embedding from DisCoVR. B (right): Marginal reconstruction from the shared embedding, recovering the spiral structure. C: Comparison of reconstruction and conditional embeddings across models.

5.2 Real data

5.2.1 Noisy colored MNIST

Dataset: We use a modified MNIST (Deng, 2012) dataset constructed from $60,000$ duplicated images: in one copy we remove the red channel ($y = 0$) and in the other we remove the green channel ($y = 1$). The digit shape remains intact following the blue channel. Label noise is introduced by flipping labels with probability $\rho \in \{0, 0.1, 0.2, 0.3, 0.4\}$.

Optimal disentanglement: Since the colored images are generated in equal proportions, the marginal reconstruction for retains a single color (see Figure 3).

Results: We evaluate marginal coloring reconstruction by DisCoVR and previous methods. Under no label noise ($\rho = 0$), all methods perform similarly (Supplementary Figure 3). At all non-zero noise levels, DisCoVR consistently outperforms competitors, whose marginal reconstructions are averaged over class-conditioned outputs. Metrics for $\rho = 0.3$ are shown in Supplementary Table 2, with results for other noise levels in Supplementary Figure 4.

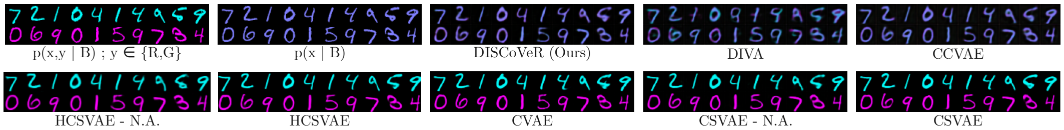


Figure 3: Colored MNIST reconstructions from the label-agnostic representation z for noise level $\rho = 0.3$. DisCoVR is the only model that consistently reconstructs mixed “semi-red/blue” tone (purple) indicating that color information has been removed and matching the true marginal $p(x|B)$.

5.2.2 CelebA-Glasses

Dataset: We use all CelebA (Liu et al., 2015) images labeled with *eyeglasses* attribute ($y = 1$), and twice as many randomly sampled images without ($y = 0$), totaling $n = 35,712$ samples.

Results: Figure 4 shows that DisCoVR accurately reconstructs input images while producing shared embeddings that marginalize over the *eyeglasses* attribute, consistently adding ”pseudo-glasses” to all samples. Competing methods are shown in Supplementary Figure 5, with quantitative results in Supplementary Table 3. While reconstruction quality is comparable across methods, DisCoVR achieves significantly better disentanglement.

Results for an *additional experiment* with the wearing-hat attribute are provided in Appendix G.

5.2.3 Single cell RNA-sequencing (scRNA-seq) data from Lupus dataset

Dataset: We analyze single-cell RNA sequencing from $n = 13,999$ peripheral blood mononuclear cells (PBMCs) collected from 8 lupus patients under two conditions: 7,451 cells control ($y = 0$),

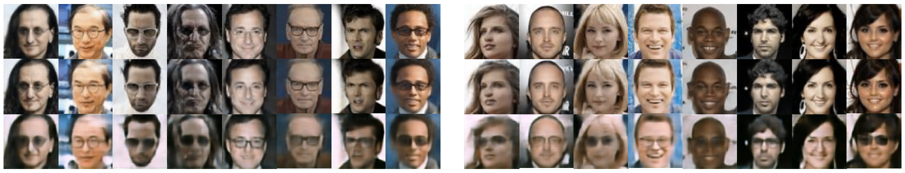


Figure 4: CelebA-Eyeglasses results. Top: Original images with/without eyeglasses. Middle: Full reconstructions by DisCoVR. Bottom: Reconstructions solely from common embeddings z. A shared representation needs to be invariant to y (presence or absence of glasses). Indeed, all reconstructed faces display an intermediate "semi-glasses" appearance, regardless of the original label.

and 6,548 IFN-β stimulation cells. IFN-β stimulation induces notable shifts in gene expression, visible in the UMAP embedding in Figure 5B (left).

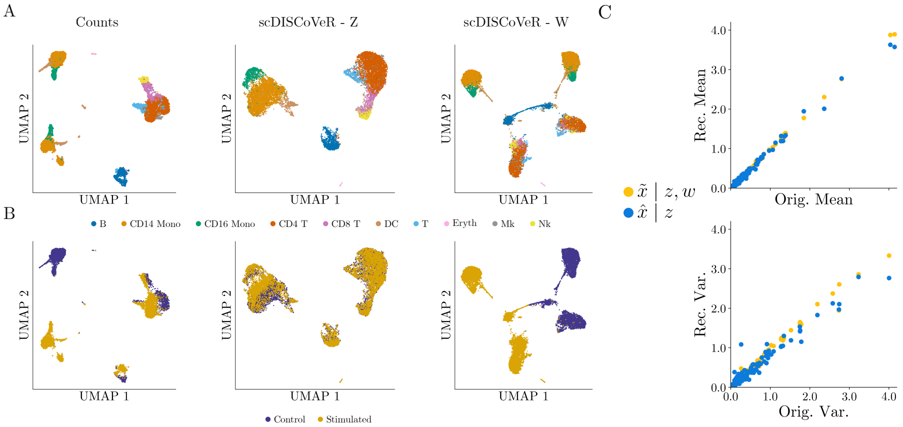


Figure 5: A–B (left): UMAPs of raw gene counts from the IFN-β dataset. A–B (middle): Shared embedding z aligns cells by type while removing stimulation effects. A–B (right): Condition-specific embedding w isolates the stimulation effect. C: Reconstructions from both z and w (yellow) recover empirical gene means and variances, while reconstructions from z alone (blue) miss the stimulation-induced variance, confirming that z discards y while preserving cell-type features.

Results: Supplementary Table 17 shows that DisCoVR effectively achieves the desired behavior with strong empirical performance, where only cell type information is captured in z (Figure 5A, middle) while the effects of IFN-β stimulation are wholly represented in w (Figure 5B, right). Other approaches either (1) achieve mixing in the z space, but compromise on keeping cell types separated or (2) leak information about stimulation into the z space (Supplementary Figure 6).

Facilitating interpretability: By enabling marginalized reconstructions, DisCoVR provides a direct link between shared embeddings and gene expression, offering clearer insight into the effects of IFN-β stimulation, unlike other methods. In Figure 5C, comparing variance across marginal and full reconstructions accurately recovers gene-level differences associated with IFN-β stimulation, including *ISG15*, *FTL*, *CCL8*, *CXCL10*, *CXCL11*, *APOBEC3A*, *IL1RN*, *IFITM3* and *RSAD2*.

6 CONCLUSION

In this work we introduced a variational framework for disentangled representation learning. Our formulation explicitly separates condition-invariant and condition-aware factors. Unlike prior work, DisCoVR incorporates two reconstruction paths, one based solely on the shared latent variable z, and the other on both latent variables z and w. Our model simultaneously learns informative shared representations, and captures structured variation across conditions. Experimental results demonstrate that DisCoVR achieves strong reconstruction, minimal information leakage, and accurate modeling of conditional effects, consistently outperforming existing methods.

REFERENCES

Haleh Akrami, Anand A Joshi, Jian Li, Sergul Aydore, and Richard M Leahy. Brain lesion detection using a robust variational autoencoder and transfer learning. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 786–790. IEEE, 2020.

Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Adversarial invariant feature learning with accuracy constraint for domain generalization. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pp. 315–331. Springer, 2020.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Ozgur Beker, Dreyton Amador, Jose Francisco Pomarino Nima, Simon Van Deursen, Yvon Woappi, and Bianca Dumitrascu. Patches: A representation learning framework for decoding shared and condition-specific transcriptional programs in wound healing. *bioRxiv*, pp. 2024–12, 2024.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/belghazi18a.html.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Pierre Boyeau, Justin Hong, Adam Gayoso, Martin Kim, José L McFaline-Figueroa, Michael I Jordan, Elham Azizi, Can Ergen, and Nir Yosef. Deep generative modeling of sample-level heterogeneity in single-cell genomics. *bioRxiv*, pp. 2022–10, 2022.

Fabio Maria Carlucci, Paolo Russo, Tatiana Tommasi, and Barbara Caputo. Hallucinating agnostic images to generalize across domains. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3227–3234. IEEE, 2019.

Jianting Chen, Ling Ding, Yunxiao Yang, Zaiyuan Di, and Yang Xiang. Domain adversarial active learning for domain generalization classification. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Sheng Cheng, Tejas Gokhale, and Yezhou Yang. Adversarial bayesian augmentation for single-source domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11400–11410, 2023.

Aveen Dayal, Vimal KB, Linga Reddy Cenkeramaddi, C Mohan, Abhinav Kumar, and Vineeth N Balasubramanian. Madg: Margin-based adversarial learning for domain generalization. *Advances in Neural Information Processing Systems*, 36:58938–58952, 2023.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Zhengming Ding and Yun Fu. Deep domain generalization with structured low-rank constraint. *IEEE Transactions on Image Processing*, 27(1):304–313, 2017.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049, 2010.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2551–2559, 2015.

William J Godinez, Eric J Ma, Alexander T Chao, Luying Pei, Peter Skewes-Cox, Stephen M Canham, Jeremy L Jenkins, Joseph M Young, Eric J Martin, and W Armand Guiguemde. Design of potent antimalarials with generative chemistry. *Nature Machine Intelligence*, 4(2):180–186, 2022.

Tejas Gokhale, Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, Chitta Baral, and Yezhou Yang. Improving diversity with adversarially learned transformations for domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 434–443, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pp. 84–92. Springer, 2015.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351. PMLR, 2022.

Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pp. 322–348. PMLR, 2020.

Tom Joy, Sebastian M Schmon, Philip HS Torr, N Siddharth, and Tom Rainforth. Capturing label characteristics in vaes. *arXiv preprint arXiv:2006.10102*, 2020.

Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9619–9628, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

Jack Klys, Jake Snell, and Richard Zemel. Learning latent subspaces in variational autoencoders. *Advances in Neural Information Processing Systems*, 31, 2018.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *International Conference on Machine Learning (ICML) Deep Learning Workshop*, volume 2, 2015.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Erratum: Estimating mutual information [phys. rev. e69, 066138 (2004)]. *Physical Review E*, 83(1), January 2011. ISSN 1550-2376. doi: 10.1103/physreve.83.019903. URL `http://dx.doi.org/10.1103/PhysRevE.83.019903`.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, volume 1, pp. 3, 2008.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1558–1566, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/larsen16.html.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Yunsheng Li and Nuno Vasconcelos. Efficient multi-domain learning by covariance normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5424–5433, 2019.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 2019.

Mario Lovrić, Tomislav Đuričić, Han TN Tran, Hussain Hussain, Emanuel Lacić, Morten A Rasmussen, and Roman Kern. Should we embed in chemistry? a comparison of unsupervised transfer learning with pca, umap, and vae on molecular fingerprints. *Pharmaceuticals*, 14(8):758, 2021.

Massimiliano Mancini, Samuel Rota Bulo, Barbara Caputo, and Elisa Ricci. Best sources forward: Domain generalization through source-specific nets. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 1353–1357. IEEE, 2018.

Stephen Marsland. *Machine Learning: An Algorithmic Perspective, Second Edition*. Chapman & Hall/CRC, 2nd edition, 2014. ISBN 1466583282.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5715–5725, 2017.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Usman Muhammad, Jorma Laaksonen, Djamila Romaissa Beddiar, and Mourad Oussalah. Domain generalization via ensemble stacking for face presentation attack detection. *International Journal of Computer Vision*, 132(12):5759–5782, 2024.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4004–4012, 2016.

Kazuki Omi, Jun Kimata, and Toru Tamaki. Model-agnostic multi-domain learning with domain-specific adapters for action recognition. *IEICE Transactions on Information and Systems*, 105 (12):2119–2126, 2022.

Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. The MIT Press, 2017.

Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Focus on the common good: Group distributional robustness follows. In *International Conference on Learning Representations*, 2021.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in Neural Information Processing Systems*, 30, 2017.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8119–8127, 2018.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

Yuli Slavutsky and Yuval Benjamini. Class distribution shifts in zero-shot learning: Learning robust representations. *Advances in Neural Information Processing Systems*, 37:89213–89248, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in Neural Information Processing Systems*, 29, 2016.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*, 28, 2015.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part III 14*, pp. 443–450. Springer, 2016.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *Advances in Neural Information Processing Systems*, 34:2215–2227, 2021.

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 949–954. IEEE, 2017.

Haohan Wang, Eric P Xing, Zexue He, and Zachary C Lipton. Learning robust representations by projecting superficial statistics out. In *7th International Conference on Learning Representations*, 2019.

Jiaheng Wei, Harikrishna Narasimhan, Ehsan Amid, Wen-Sheng Chu, Yang Liu, and Abhishek Kumar. Distributionally robust post-hoc classifiers under prior shifts. In *International Conference on Learning Representations*, 2023.

Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2840–2848, 2017.

Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1):e9620, 2021.

Tongtong Yuan, Weihong Deng, Jian Tang, Yinan Tang, and Binghui Chen. Signal-to-noise ratio: A robust distance metric for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4815–4824, 2019.

Jiusi Zhang, Xiang Li, Jilun Tian, Yuchen Jiang, Hao Luo, and Shen Yin. A variational local weighted deep sub-domain adaptation network for remaining useful life prediction facing cross-domain condition. *Reliability Engineering & System Safety*, 231:108986, 2023.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021.

Wei Zhu, Le Lu, Jing Xiao, Mei Han, Jiebo Luo, and Adam P Harrison. Localized adversarial domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7108–7118, 2022.

A Additional related work

A.1 Domain generalization

The task of representation disentanglement is closely related to the field of domain generalization (Muandet et al., 2013), which assumes limited or no access to target domain samples and aims to learn representations that can be readily adapted, often via transfer learning, to new, unseen domains.

As noted by Wang et al. (2019), existing methods in domain generalization can be broadly categorized into two main approaches: (i) approaches for reducing the inter-domain differences, often by using adversarial techniques (Ghifary et al., 2015; Wang et al., 2017; Motiian et al., 2017; Li et al., 2018; Carlucci et al., 2019; Li et al., 2018; Wang et al., 2019; Akuzawa et al., 2020; Zhu et al., 2022; Gokhale et al., 2023; Dayal et al., 2023; Cheng et al., 2023; Chen et al., 2024), and (ii) Approaches that construct an ensemble of domain-specific models, and then fuse their representations to form a unified, domain-agnostic representation (Ding & Fu, 2017; Mancini et al., 2018; Zhou et al., 2021; Muhammad et al., 2024).

Additional strategies for domain generalization include contrastive learning approaches (Kim et al., 2021), methods based on distribution alignment via metrics (Muandet et al., 2013; Sun & Saenko, 2016), and techniques utilizing custom network architectures, for instance by incorporating domain-specific adapters between shared layers (Rebuffi et al., 2017; 2018; Li & Vasconcelos, 2019; Omi et al., 2022).

The primary distinction between these methods and ours lies in the explicit probabilistic modeling and disentanglement of domain-invariant and domain-specific factors. Whereas prior approaches typically focus on aligning domains through adversarial training or fusing multiple domain-specific predictors, our method constructs a structured latent space, decomposed into a shared representation z, capturing domain-invariant information, and a conditional component w, which encodes domain-specific variability. This factorization is learned through a tailored variational objective involving an adversarial penalty and two reconstructions —one based on z alone, and another on the full latent pair (z, w), thereby promoting both interpretability and a clean separation of shared and domain-aware features.

A.2 Out of distribution generalization

A.2.1 Enviroment balancing methods

The field of out-of-distribution (OOD) generalization emerged from foundational work on causality and invariance across training environments (Peters et al., 2016; 2017). The central assumption is that each environment exhibits distinct spurious correlations between features and labels; therefore, robust generalization requires models to focus on invariant relationships that hold across environments. To address this distribution shift, many recent approaches adopt a regularized empirical risk minimization framework:

$$\min_{\theta} \sum_{e \in E_{\text{train}}} \ell^e(f_\theta) + \lambda R(f_\theta, E_{\text{train}}), \tag{11}$$

where the regularizer R encourages representations that are stable across environments. Among these, Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) enforces that a single classifier remains optimal across all environments, Variance Risk Extrapolation (VarREx) (Krueger et al., 2021) promotes robustness by minimizing the variance of losses across environments, and CLOvE (Wald et al., 2021) takes a calibration-theoretic perspective, penalizing discrepancies between predicted confidence and correctness across environments.

While these methods focus on enforcing predictive invariance across environments through regularization, our approach instead explicitly enforces conditional independence between the shared latent variable z and an environment-aware variable w.

A.2.2 Distributionally robust methods

An alternative line of work for handling distribution shifts is Distributionally Robust Optimization (DRO) (Ben-Tal et al., 2013; Duchi et al., 2021; Duchi & Namkoong, 2021; Wei et al., 2023), which avoids assuming a fixed data-generating distribution. Instead, DRO methods optimize performance under the worst-case scenario over a family of plausible distributions. A prominent variant, known as group DRO (Sagawa et al., 2019; Piratla et al., 2021), introduces group-level structure that may correlate with spurious features, potentially leading to biased predictions. In settings where group labels are not directly observed, several strategies have been proposed, including reweighting high-loss examples (Liu et al., 2021) and balancing class-group combinations through data sub-sampling (Idrissi et al., 2022).

However, these approaches assume that the label space remains fixed between training and test time, limiting their applicability in adaptation to new domains, environments or conditions.

A.3 Zero-shot learning

Zero-shot learning systems (Fei-Fei et al., 2006; Larochelle et al., 2008) aim to classify instances from novel, previously unseen classes at test time. In contrast to the out-of-distribution (OOD) generalization setting, these approaches typically do not assume the presence or structure of a distribution shift. Instead, a common strategy is to learn data representations that capture class-agnostic similarity, enabling the model to determine whether two instances belong to the same class without requiring knowledge of the class identity itself. Such methods include contrastive-learning (Hadsell et al., 2006), siamese neural networks (Koch et al., 2015), triplet networks (Hoffer & Ailon, 2015), and other more recent variations (Oh Song et al., 2016; Sohn, 2016; Wu et al., 2017; Yuan et al., 2019). Recent work has begun to address the impact of class distribution shifts in zero-shot settings. For instance, Slavutsky & Benjamini (2024) integrate environment-based regularization—motivated by OOD generalization—with zero-shot learning by simulating distribution shifts through hierarchical sampling, enabling the model to learn representations that are robust to shifts in class distributions.

While this line of work shares our motivation of improving robustness under unseen conditions, it primarily addresses the problem of class-level generalization through similarity-based learning, rather than explicitly modeling and disentangling the latent factors—such as domain or environment—that drive distributional variation across tasks.

B Proofs

B.1 Proof of Proposition 2.1

Proof.

$$\mathrm{ELBO}(q, p; x, y) - \mathcal{L}_w(q_{w|x,y}, p; x, y) \tag{12}$$

$$= \left[\log p(x \mid y) - D_{\mathrm{KL}}(q_{w|x,y} \,\|\, p_{w|x,y})\right] - \left[\log p(x \mid y) - D_{\mathrm{KL}}\left(q_{z|x} q_{w|x,y} \,\|\, p_{z,w|x,y}\right)\right] \tag{13}$$

$$= D_{\mathrm{KL}}\left(q_{z|x} q_{w|x,y} \,\|\, p_{z,w|x,y}\right) - D_{\mathrm{KL}}\left(q_{w|x,y} \,\|\, p_{w|x,y}\right) \tag{14}$$

$$= \mathbb{E}_{q_{w|x,y}}\left[\mathbb{E}_{q_{z|x}}\left[\log q(z \mid x) + \log q(w|x,y) - \log p(z, w \mid x, y)\right]\right] \tag{15}$$

$$- \mathbb{E}_{q_{w|x,y}}\left[\log q(w|x,y) - \log p(w|x,y)\right] \tag{16}$$

$$= \mathbb{E}_{q_{w|x,y}}\left[\mathbb{E}_{q_{z|x}}\left[\log q(z \mid x) - \log p(z, w \mid x, y) + \log p(w|x,y)\right]\right] \tag{17}$$

$$= \mathbb{E}_{q_{w|x,y}}\left[\mathbb{E}_{q_{z|x}}\left[\log q(z \mid x) - \log p(z \mid w \mid x, y)\right]\right] \tag{18}$$

$$= \mathbb{E}_{q_{w|x,y}}\left[\mathrm{KL}\left(q_{z|x} \,\|\, p_{z|w,x,y}\right)\right]. \tag{19}$$

□

B.2 Game equilibrium

B.2.1 Regularity conditions

To ensure that expectations and KL-terms in the game objective $\mathcal{L}(q_{z|x}, q_{w|x,y}, g_{y|z})$ render the functionals strictly concave in $q_{z|x}$, strictly concave in $q_{w|x,y}$, and strictly convex in g, the following regularity conditions are required:

1. The likelihoods $p(x|z), p(x|z,w), p(y|x)$ are strictly positive, continuous densities.

2. The variational families Q_z and Q_w, and the set of achievable classifiers $\mathcal{G}$ are non-empty, convex and compact.
3. $\log p(x|z,w)$ and $\log g(y|x)$ are integrable.

B.2.2 Proof of Proposition 2.2

Proof. Since $\mathcal{L}_z(q_{z|x}, p; x)$ is the standard ELBO objective, we have that

$$\mathcal{L}_z(q_{z|x}, p; x) = \log p(x) - D_{\mathrm{KL}}\left(q_{z|x} \,\|\, p_{z|x}\right). \tag{20}$$

Similarly, we have that

$$\mathcal{L}_w(q_{w|x,y}, p; x, y) = \log p(x \mid y) - D_{\mathrm{KL}}\left(q_{z|x} q_{w|x,y} \,\|\, p_{z,w|x,y}\right). \tag{21}$$

Thus,

$$\begin{aligned}\mathcal{L}(q_{z|x}, q_{w|x,y}, g_{y|z}) = \mathbb{E}_{p_{x,y}} \big[&\log p(x) - D_{\mathrm{KL}}\left(q_{z|x} \,\|\, p_{z|x}\right) &(22)\\ &+ \log p(x \mid y) - D_{\mathrm{KL}}\left(q_{z|x} q_{w|x,y} \,\|\, p_{z,w|x,y}\right) &(23)\\ &- \mathbb{E}_{q_{z|x}} \log g(y \mid z)\big]. &(24)\end{aligned}$$

For fixed $q_{z|x}$, the adversarial classifier minimizes:

$$-\mathbb{E}_{p_{x,y}} \mathbb{E}_{q_{z|x}} \log g(y \mid z), \tag{25}$$

which is the population cross-entropy and is strictly convex in $g(y|z)$, and thus has a unique solution.

It remains to show that the terms in the objective function that depend on $q_{z|x}$ and $q_{w|x,y}$, are strictly concave in each argument when the others are held fixed.

Focusing on the terms dependent on $q_{w|x,y}$ first, define

$$\begin{aligned}\ell_w &:= - D_{\mathrm{KL}}\left(q_{z|x} q_{w|x,y} \,\|\, p_{z,w|x,y}\right) &(26)\\ &= - \iint q\left(z \mid x\right) q(w \mid x,y) \left[\log q\left(z \mid x\right) + \log q(w \mid x,y) - \log p(z,w \mid x,y)\right] dz\, dw \\ &= - \int q\left(z \mid x\right) \log q\left(z \mid x\right)\, dz + \int q(w \mid x,y) \log q(w \mid x,y)\, dw &(27)\\ &\quad - \iint q\left(z \mid x\right) q(w \mid x,y) \log p(z,w \mid x,y)\, dz\, dw &(28)\\ &= H(q_{z|x}) + H(q_{w|x,y}) + \mathbb{E}_{q_{z|x}} \mathbb{E}_{q_{w|x,y}} \log p(z,w \mid x,y). &(29)\end{aligned}$$

Note that

$$\mathbb{E}_{q_{z|x}} \mathbb{E}_{q_{w|x,y}} \log p(z,w \mid x,y) \tag{30}$$

is linear in $q_{w|x,y}$, and since $H(q_{w|x,y})$ is strictly concave in $q_{w|x,y}$, we have that $\mathbb{E}_{p_{x,y}}[\ell_w]$ is strictly concave in $q_{w|x,y}$.

Similarly, define

$$\ell_z := -D_{\mathrm{KL}}\left(q_{z|x} \,\|\, p_{z|x}\right) - D_{\mathrm{KL}}\left(q_{z|x} q_{w|x,y} \,\|\, p_{z,w|x,y}\right). \tag{31}$$

By convexity of KL divergence in its first argument, $-D_{\mathrm{KL}}\left(q_{z|x} \,\|\, p_{z|x}\right)$ is strictly concave in $q_{z|x}$.

Focusing on the second KL term, from Equation 29 we have that

$$-D_{\mathrm{KL}}\left(q_{z|x} q_{w|x,y} \,\|\, p_{z,w|x,y}\right) = H(q_{z|x}) + H(q_{w|x,y}) + \mathbb{E}_{q_{z|x}} \mathbb{E}_{q_{w|x,y}} \log p(z,w \mid x,y), \tag{32}$$

where $H(q_{z|x})$ is strictly concave in $q_{z|x}$.

Recall that we assumed that $p(w|y)$ depends on $q(z|x)$. Under our model

$$p(x,y,z,w) = p(y) p(w \mid y) p(z) p(x \mid z,w), \tag{33}$$

yielding

$$p\left(z, w, \mid x, y\right) = p(w \mid y) \frac{p(z) p(x \mid z,w)}{p(x \mid y)}. \tag{34}$$

Hence,

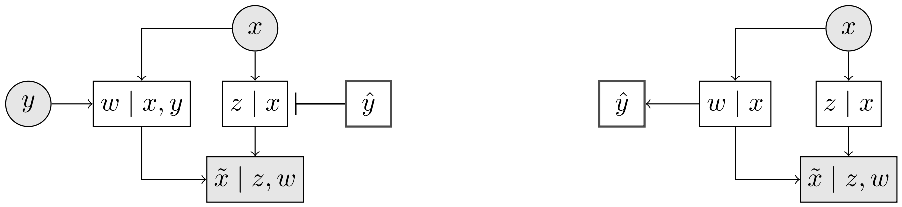
$$\mathbb{E}_{q_{z|x}}\mathbb{E}_{q_{w|x,y}} \log p(z, w \mid x, y) = \mathbb{E}_{q_{z|x}}\mathbb{E}_{q_{w|x,y}} \left[\log p(w \mid y) + \log \frac{p(z)p(x \mid z, w)}{p(x \mid y)}\right],$$

where $p(w \mid y) = \mathcal{N}(w; \mu_y, I)$ with $\mu_y = \mathbb{E}_{p_{x|y}}\left[\mathbb{E}_{q_{z|x}}[z]\right]$. Therefore,

$$\mathbb{E}_{q_{z|x}}\mathbb{E}_{q_{w|x,y}}[\log p(w \mid y)] = -\frac{1}{2}\left[d\log(2\pi) + \mathbb{E}_{q_{z|x}}\mathbb{E}_{q_{w|x,y}}\|w - \mu_y\|^2\right] \tag{35}$$

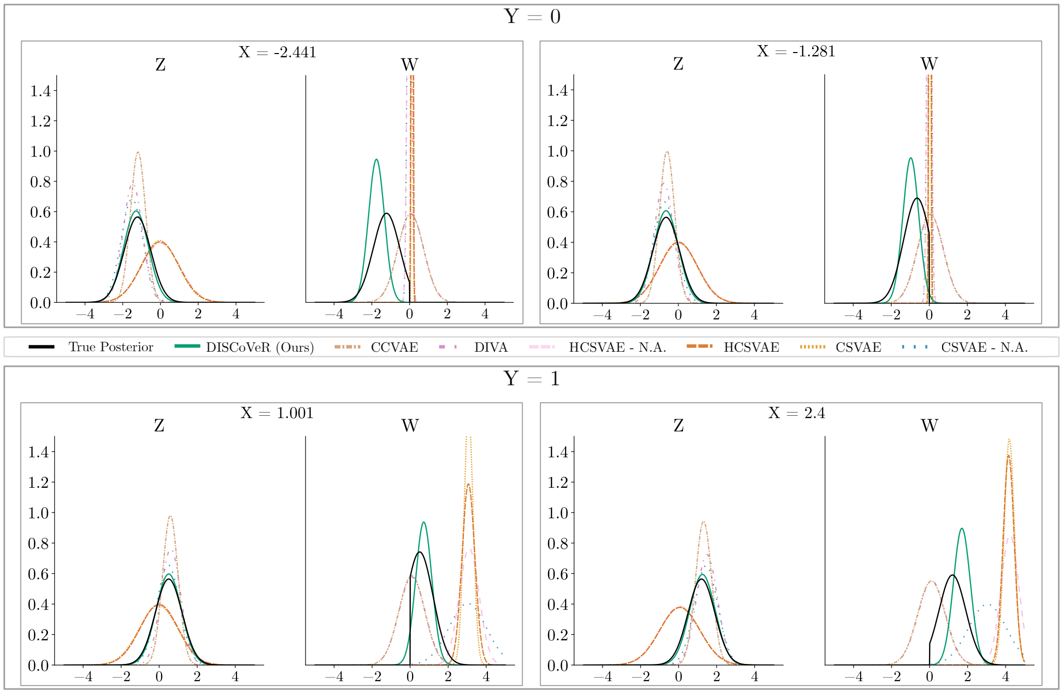
where $-\|w - \mu_y\|^2$ is a quadratic form in μ_y, which is linear in $q_{z|x}$, and thus $\mathbb{E}_{q_{z|x}}\mathbb{E}_{q_{w|x,y}}[\log p(w \mid y)]$ is strictly concave in $q_{z|x}$. Hence, $-D_{\text{KL}}\left(q_{z|x}q_{w|x,y} \,\|\, p_{z,w|x,y}\right)$ is strictly concave in $q_{z|x}$, and thus so is $\mathbb{E}_{p_{x,y}}[\ell_z]$. □

C Supplementary Figures

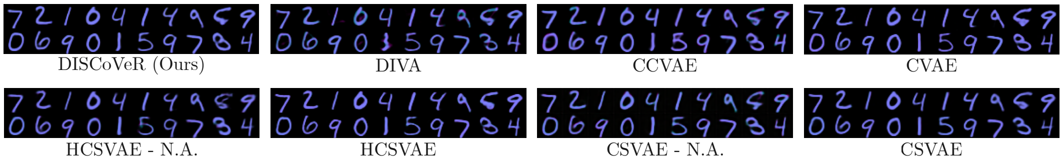


(a) (b)

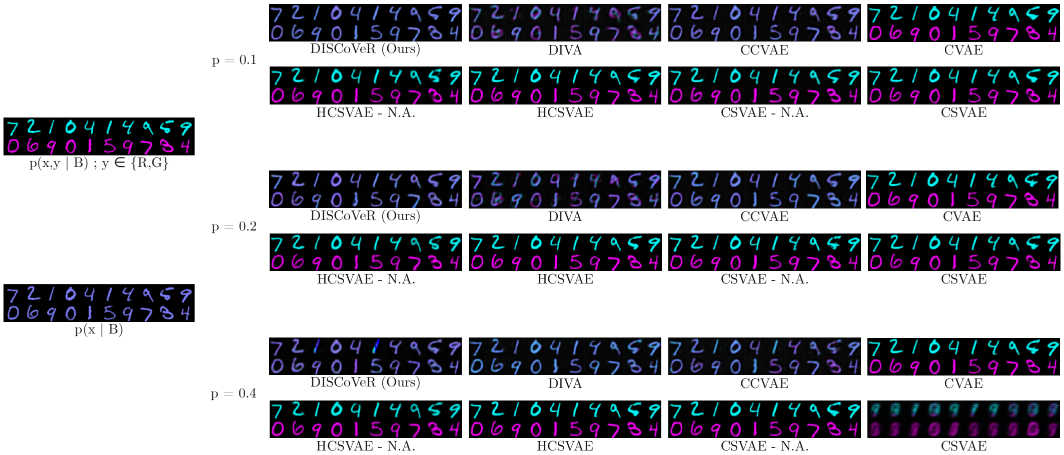
Supplementary Figure 1: Encoder-decoder structures for previous approaches. (a) CSVAE. (b) DIVA - CCVAE.



Supplementary Figure 2: Comparison of approximate variational posteriors against the true posterior for latent variables z, w for different values of x with $y = 0$ (top) and $y = 1$ (bottom).

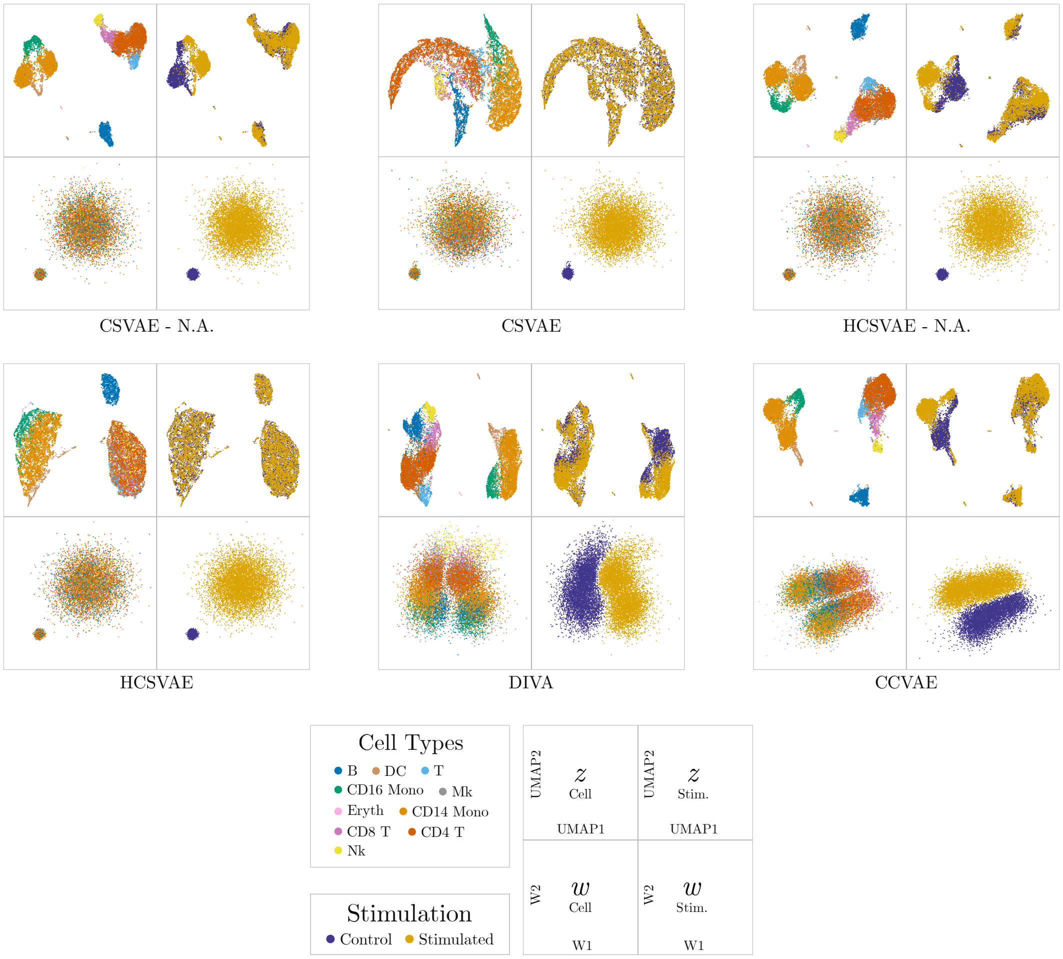


Supplementary Figure 3: Colored MNIST results for no noise.



Supplementary Figure 4: Colored MNIST visual results across the remaining noise levels.

Supplementary Figure 5: Reconstruction performance for other models on the CelebA-Glasses dataset. Top: Original samples from the data. Bottom: Reconstructions by the given model.



Supplementary Figure 6: Embeddings obtained by other models on the Kang dataset. For each block, top (resp. bottom) rows are z (resp. w) embeddings, while left (resp. right) columns are colored by cell type (resp. stimulation).

D Supplementary tables for experimental results

Supplementary Table 1: RMSE for the Colored MNIST dataset without any label noise.

	Marginal RMSE ($p = 0$) ↓
CSVAE - No Adv.	$\mathbf{0.064 \pm 0.002}$
CSVAE	0.079 ± 0.008
HCSVAE - No Adv.	0.094 ± 0.004
HCSVAE	0.079 ± 0.030
DIVA	0.065 ± 0.005
CCVAE	0.065 ± 0.006
DisCoVR (Ours)	$\mathbf{0.064 \pm 0.000}$

Supplementary Table 2: RMSE calculated between the estimated and true marginal across different levels of label noise on the Colored MNIST dataset. p defines label flip probability. Bold denotes best performance.

	Marginal RMSE ↓			
	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$
CSVAE - No Adv.	0.141 ± 0.002	0.141 ± 0.003	0.142 ± 0.002	0.143 ± 0.002
CSVAE	0.135 ± 0.022	0.152 ± 0.018	0.181 ± 0.007	0.173 ± 0.008
HCSVAE - No Adv.	0.150 ± 0.001	0.150 ± 0.000	0.151 ± 0.000	0.151 ± 0.001
HCSVAE	0.139 ± 0.003	0.141 ± 0.001	0.141 ± 0.001	0.141 ± 0.001
DIVA	0.115 ± 0.011	0.102 ± 0.013	0.106 ± 0.010	0.113 ± 0.014
CCVAE	0.092 ± 0.002	0.103 ± 0.014	0.099 ± 0.011	0.092 ± 0.005
DisCoVR (Ours)	$\mathbf{0.073 \pm 0.001}$	$\mathbf{0.083 \pm 0.004}$	$\mathbf{0.087 \pm 0.002}$	$\mathbf{0.087 \pm 0.001}$

Supplementary Table 3: Model performances on the CelebA-Glasses dataset. Bold denotes best performance.

	$I(z; w)$ ↓	NLL (↓)
CSVAE - No Adv.	0.048 ± 0.014	137.522 ± 0.155
CSVAE	0.079 ± 0.029	145.989 ± 0.336
HCSVAE - No Adv.	0.055 ± 0.012	131.813 ± 0.21
HCSVAE	0.055 ± 0.014	137.319 ± 0.265
DIVA	0.188 ± 0.028	143.528 ± 0.02
CCVAE	0.083 ± 0.022	$\mathbf{131.764 \pm 0.006}$
DisCoVR (Ours)	$\mathbf{0.030 \pm 0.011}$	135.677 ± 0.007
DisCoVR - Common (Ours)	—	374.114 ± 0.05

E Additional disentanglement metrics

We provide an extended disentanglement assessment using multiple metrics. Because mutual information is difficult to estimate reliably, we report two estimators—MINE and kNN. Although their absolute values differ, the relative rankings of the methods remain consistent as can be seen in the ranking tables. In addition to these mutual-information estimates, we also report the following metrics, which quantifying the level of label information captured by w compared to z :

Mutual Information Gap (MIG)

$$\text{MIG}(w; z) = \frac{I(y; w) - I(y; z)}{H(y)}$$

Mutual Information Completeness (MIC)

$$\text{MIC}(w; z) = \frac{I(y; w)}{I(y; w) + I(y; z)}$$

E.1 Parametric model

CSVAE and its variants impose a fully separable prior, thereby forcing separability even when the true latent structure is not separable (see Table 1). In contrast, DisCoVR learns informative conditional embeddings that closely track the true posterior without requiring ground-truth knowledge of a truncated or fully separable prior, and it outperforms both DIVA and CCVAE.

Replacing the prior in DisCoVR with a fully separable predefined prior on w yields consistent embeddings with the ground-truth structure while retaining the benefits of separability.

Supplementary Table 4: Additional disentanglement metrics calculated with kNN mutual information estimation for the parametric model dataset with $k = 20$. Bold indicates closest to true posterior within group.

Assumption	Model	$I(y; z)$	$I(y; w)$	$I(w; z)$	MIG$(w; z)$	MIC$(w; z)$	$I(w; z \mid y)$
Fully Separable	CSVAE - N.A.	$\mathbf{0.069 \pm 0.034}$	0.634 ± 0.002	$\mathbf{0.098 \pm 0.034}$	0.063 ± 0.004	$\mathbf{0.904 \pm 0.048}$	0.000 ± 0.002
	CSVAE	0.024 ± 0.048	0.620 ± 0.044	0.047 ± 0.050	$\mathbf{0.067 \pm 0.010}$	0.963 ± 0.073	0.013 ± 0.009
	HCSVAE - N.A.	0.000 ± 0.000	$\mathbf{0.643 \pm 0.000}$	0.000 ± 0.001	0.072 ± 0.000	1.000 ± 0.000	0.000 ± 0.000
	HCSVAE	0.000 ± 0.000	$\mathbf{0.643 \pm 0.001}$	0.001 ± 0.001	0.072 ± 0.000	1.000 ± 0.000	0.000 ± 0.001
	DisCoVR (CSVAE prior)	0.000 ± 0.000	$\mathbf{0.643 \pm 0.000}$	0.051 ± 0.007	0.072 ± 0.000	1.000 ± 0.000	$\mathbf{0.031 \pm 0.005}$
Flexible	DIVA	0.021 ± 0.042	0.091 ± 0.046	0.000 ± 0.000	0.008 ± 0.010	0.800 ± 0.400	0.000 ± 0.000
	CCVAE	$\mathbf{0.022 \pm 0.043}$	0.090 ± 0.045	0.000 ± 0.000	0.008 ± 0.010	0.800 ± 0.400	0.000 ± 0.000
	DisCoVR (our prior)	0.010 ± 0.006	$\mathbf{0.151 \pm 0.007}$	$\mathbf{0.108 \pm 0.029}$	$\mathbf{0.016 \pm 0.001}$	$\mathbf{0.938 \pm 0.035}$	$\mathbf{0.072 \pm 0.020}$
Fully Separable	Posterior (no truncation)	0.057 ± 0.001	0.057 ± 0.000	0.144 ± 0.003	0.000 ± 0.000	0.499 ± 0.006	0.090 ± 0.002
	True Posterior	0.058 ± 0.003	0.643 ± 0.000	0.144 ± 0.005	0.066 ± 0.000	0.917 ± 0.004	0.055 ± 0.003

Supplementary Table 5: Additional disentanglement metrics calculated with MINE mutual information estimation for the parametric model dataset. Bold indicates closest to true posterior within group.

Assumption	Model	$I(y; z)$	$I(y; w)$	$I(w; z)$	MIG$(w; z)$	MIC$(w; z)$	$I(w; z \mid y)$
Fully Separable	CSVAE - N.A.	$\mathbf{0.096 \pm 0.037}$	0.528 ± 0.026	$\mathbf{0.096 \pm 0.034}$	$\mathbf{0.048 \pm 0.006}$	$\mathbf{0.848 \pm 0.057}$	0.001 ± 0.001
	CSVAE	0.033 ± 0.055	$\mathbf{0.526 \pm 0.045}$	0.031 ± 0.045	0.055 ± 0.010	0.945 ± 0.091	0.009 ± 0.005
	HCSVAE - N.A.	0.000 ± 0.000	0.543 ± 0.018	0.000 ± 0.000	0.061 ± 0.002	1.000 ± 0.000	0.000 ± 0.000
	HCSVAE	0.000 ± 0.000	0.543 ± 0.022	0.000 ± 0.000	0.061 ± 0.002	1.000 ± 0.000	0.001 ± 0.000
	DisCoVR (CSVAE prior)	0.020 ± 0.004	0.543 ± 0.018	0.033 ± 0.005	0.058 ± 0.002	0.964 ± 0.008	$\mathbf{0.030 \pm 0.004}$
Flexible	DIVA	0.027 ± 0.053	0.113 ± 0.056	0.001 ± 0.001	0.010 ± 0.012	0.798 ± 0.398	0.001 ± 0.000
	CCVAE	0.026 ± 0.052	0.115 ± 0.058	0.001 ± 0.001	0.010 ± 0.012	0.799 ± 0.399	0.001 ± 0.000
	DisCoVR (ours)	$\mathbf{0.037 \pm 0.006}$	$\mathbf{0.176 \pm 0.008}$	$\mathbf{0.109 \pm 0.025}$	$\mathbf{0.016 \pm 0.001}$	$\mathbf{0.825 \pm 0.026}$	$\mathbf{0.073 \pm 0.018}$
Fully Separable	Posterior (no truncation)	0.084 ± 0.003	0.083 ± 0.003	0.137 ± 0.004	0.000 ± 0.000	0.497 ± 0.009	0.088 ± 0.003
	True Posterior	0.085 ± 0.005	0.493 ± 0.011	0.139 ± 0.005	0.046 ± 0.002	0.853 ± 0.009	0.057 ± 0.004

Supplementary Table 6: Rank (1 = closest to True Posterior) of each method with respect to the true posterior for metrics calculated with kNN mutual information estimation with $k = 20$. Colors indicate rank within each block: red = worse (farther), green = better (closer).

Assumption	Model	$I(y; z)$	$I(y; w)$	$I(w; z)$	MIG$(w; z)$	MIC$(w; z)$	$I(w; z \mid y)$
Fully Separable	CSVAE - N.A.	1	4	1	2	1	3
	CSVAE	2	5	3	1	2	2
	HCSVAE - N.A.	3	1	5	3	3	3
	HCSVAE	3	1	4	3	3	3
	DisCoVR (CSVAE prior)	3	1	2	3	3	1
Flexible	DIVA	2	2	2	2	2	2
	CCVAE	1	3	2	2	2	2
	DisCoVR (our prior)	3	1	1	1	1	1

Supplementary Table 7: Rank (1 = closest to True Posterior) of each method with respect to the True Posterior for metrics calculated with MINE mutual information estimation. Colors indicate rank within each block: red = worse (farther), green = better (closer).

Assumption	Model	$I(y; z)$	$I(y; w)$	$I(w; z)$	MIG$(w; z)$	MIC$(w; z)$	$I(w; z \mid y)$
Fully Separable	CSVAE - N.A.	1	2	1	1	1	3
	CSVAE	2	1	3	2	2	2
	HCSVAE - N.A.	4	3	4	4	4	5
	HCSVAE	4	3	4	4	4	3
	DisCoVR (CSVAE prior)	3	3	2	3	3	1
Flexible	DIVA	2	3	2	2	3	2
	CCVAE	3	2	2	2	2	2
	DisCoVR (ours)	1	1	1	1	1	1

E.2 Noisy Swiss Roll

When the observed labels are noisy, DisCoVR outperforms other methods, obtaining embeddings close to the ground truth.

Supplementary Table 8: Additional disentanglement metrics calculated with kNN mutual information estimation for the Noisy Swiss Roll ($p = 0.3$) dataset with $k = 20$. Bold indicates closest to ground truth within group.

Assumption	Model	$I(y;z)$	$I(y;w)$	$I(w;z)$	MIG$(w;z)$	MIC$(w;z)$	$I(w;z \mid y)$
Fully Separable	CSVAE - N.A.	0.041 ± 0.007	0.525 ± 0.221	0.362 ± 0.180	0.057 ± 0.026	0.888 ± 0.098	0.266 ± 0.152
	CSVAE	0.018 ± 0.026	**0.429 ± 0.254**	0.240 ± 0.181	**0.048 ± 0.032**	0.912 ± 0.129	0.186 ± 0.146
	HCSVAE - N.A.	0.029 ± 0.007	0.642 ± 0.000	0.065 ± 0.013	0.072 ± 0.001	0.957 ± 0.010	0.009 ± 0.019
	HCSVAE	**0.001 ± 0.002**	0.641 ± 0.001	**0.005 ± 0.004**	0.075 ± 0.000	**0.999 ± 0.003**	**0.000 ± 0.000**
Flexible	DIVA	0.034 ± 0.013	0.036 ± 0.011	2.633 ± 0.360	0.000 ± 0.003	0.515 ± 0.159	2.185 ± 0.332
	CCVAE	0.040 ± 0.015	0.030 ± 0.007	2.952 ± 0.124	−0.001 ± 0.003	0.447 ± 0.153	2.462 ± 0.118
	DisCoVR (ours)	**0.000 ± 0.000**	**0.049 ± 0.002**	**0.029 ± 0.011**	**0.006 ± 0.000**	**1.000 ± 0.000**	**0.014 ± 0.008**
Noisy	Ground Truth	0.000 ± 0.000	0.055 ± 0.002	0.000 ± 0.000	0.007 ± 0.000	1.000 ± 0.000	0.000 ± 0.000

Supplementary Table 9: Additional disentanglement metrics calculated with MINE mutual information estimation for the Noisy Swiss Roll ($p = 0.3$) dataset. Bold indicates closest to ground truth within group.

Assumption	Model	$I(y;z)$	$I(y;w)$	$I(w;z)$	MIG$(w;z)$	MIC$(w;z)$	$I(w;z \mid y)$
Fully Separable	CSVAE - N.A.	0.046 ± 0.022	0.422 ± 0.186	0.050 ± 0.020	0.044 ± 0.023	0.834 ± 0.184	0.029 ± 0.017
	CSVAE	0.023 ± 0.027	**0.373 ± 0.232**	0.027 ± 0.020	**0.041 ± 0.028**	0.877 ± 0.198	0.024 ± 0.017
	HCSVAE - N.A.	0.023 ± 0.014	0.585 ± 0.011	0.026 ± 0.012	0.066 ± 0.002	0.963 ± 0.021	0.006 ± 0.002
	HCSVAE	**0.002 ± 0.000**	0.570 ± 0.011	**0.002 ± 0.001**	0.067 ± 0.001	**0.997 ± 0.001**	**0.003 ± 0.001**
Flexible	DIVA	0.041 ± 0.024	0.043 ± 0.026	0.313 ± 0.084	**0.000 ± 0.006**	0.507 ± 0.296	0.345 ± 0.065
	CCVAE	0.056 ± 0.020	**0.036 ± 0.020**	0.507 ± 0.114	−0.002 ± 0.004	0.390 ± 0.226	0.494 ± 0.099
	DisCoVR (ours)	**0.001 ± 0.000**	0.069 ± 0.002	**0.004 ± 0.002**	0.008 ± 0.000	**0.983 ± 0.004**	**0.006 ± 0.002**
Noisy	Ground Truth	0.000 ± 0.000	0.024 ± 0.018	0.000 ± 0.001	0.003 ± 0.002	0.985 ± 0.048	0.002 ± 0.001

Supplementary Table 10: Rank (1 = closest to Ground Truth) of each method with respect to the Ground Truth for metrics calculated with kNN mutual information estimation with $k = 20$ on the Noisy Swiss Roll ($p = 0.3$) dataset. Colors indicate rank within each block: red = worse (farther), green = better (closer).

Assumption	Method	$I(y;z)$	$I(y;w)$	$I(w;z)$	MIG$(w;z)$	MIC$(w;z)$	$I(w;z \mid y)$
Fully Separable	CSVAE - N.A.	4	2	4	2	4	4
	CSVAE	2	1	3	1	3	3
	HCSVAE - N.A.	3	4	2	3	2	2
	HCSVAE	1	3	1	4	1	1
Flexible	DIVA	2	2	2	2	2	2
	CCVAE	3	3	3	3	3	3
	DisCoVR (ours)	1	1	1	1	1	1

Supplementary Table 11: Rank (1 = closest to Ground Truth) of each method with respect to the Ground Truth for metrics calculated with MINE mutual information estimation on the Noisy Swiss Roll ($p = 0.3$) dataset. Colors indicate rank within each block: red = worse (farther), green = better (closer).

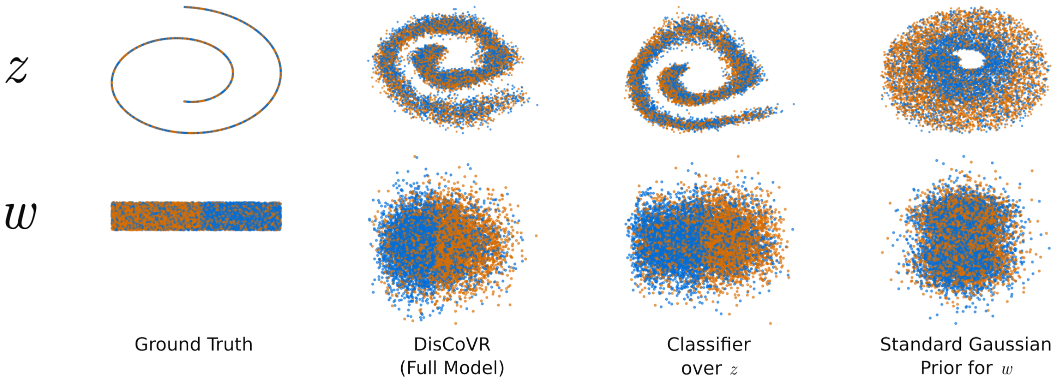
Assumption	Model	$I(y;z)$	$I(y;w)$	$I(w;z)$	$\mathrm{MIG}(w;z)$	$\mathrm{MIC}(w;z)$	$I(w;z \mid y)$
Fully Separable	CSVAE - N.A.	4	2	4	2	4	4
	CSVAE	2	1	3	1	3	3
	HCSVAE - N.A.	2	4	2	3	2	2
	HCSVAE	1	3	1	4	1	1
Flexible	DIVA	2	2	2	1	2	2
	CCVAE	3	1	3	2	3	3
	DisCoVR (ours)	1	3	1	2	1	1

F Ablations on model components

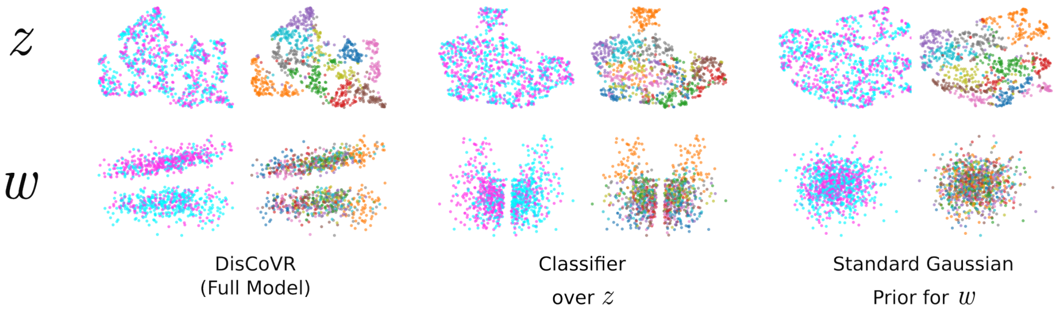
We evaluate the contribution of each model component by examining two variations: (1) applying the classifier directly to z, and (2) replacing the conditional prior on w with a standard Gaussian.

When training the classifier directly on z we were able achieve results qualitatively similar to those obtained using the reconstruction $\hat{x}$, but doing so requires substantially more parameter tuning.

An unconditional standard Gaussian prior for w, causes w to collapse into a representation redundant with z, removing meaningful separation.



Supplementary Figure 7: Ablation study on the Noisy Swiss Roll ($p = 0.3$) dataset.



Supplementary Figure 8: Ablation study on the Noisy Colored MNIST ($p = 0.3$) dataset. For each setting: left column denotes coloring by noisy labels, right column denotes coloring by digit (shape, not included in the label).

G Additional experiment on CelebA-Hats

We performed an additional experiment on the CelebA dataset, with the attribute $Wearing\_hat$ denoting the y label. Supplementary Table 12 outlines the results of this experiment. DisCoVR is the only method that exhibits high disentanglement for z, w without compromising reconstruction quality.

Supplementary Table 12: Model performances of a single experiment on CelebA-Hats. Bold denotes best performance.

	$I(z; w) \downarrow$	NLL (↓)
CSVAE - N.A.	0.360	653.537
CSVAE	0.213	**351.082**
HCSVAE - N.A.	0.135	2608.442
HCSVAE	0.192	673.674
DIVA	0.553	356.090
CCVAE	0.856	347.940
DisCoVR (Ours)	**0.059**	353.271
DisCoVR (Ours) - Common	-	437.144

H Implementation details

H.1 Considerations and reproducibility

We run all experiments on a single H100 GPU. Reported means and standard deviations for tables are conducted over 10 repetitions of the experiment with different random seeds. All models are trained using the AdamW (Loshchilov & Hutter, 2019) optimizer until validation loss stops decreasing for 50 epochs. Wherever provided, we use mutual information neural estimation (MINE, Belghazi et al. (2018)) and k-Nearest Neighbor (kNN) mutual information estimation Kraskov et al. (2011) to obtain mutual information estimates. For Naive Bayes classifiers, we use the implementation provided by *scikit-learn* (Pedregosa et al., 2011). To use ideal hyperparameters for each method, we consult the original implementation whenever possible, and conduct a simple grid-search to produce originally described model behavior. Implementations of all methods compared in this study, including DisCoVR, as well as code to reproduce our results, is attached to this submission and will be made public upon acceptance. Models compared in the study admit a weighting term for each term in the loss function, of which most are shared across different approaches. We use the following shorthands for each of the terms:

$$\begin{aligned}
&\text{Rec.} \rightarrow \mathbb{E}_{q_{z|x}}\left[\mathbb{E}_{q_{w|x,y}}[\log p(x \mid z, w)]\right] \\
&D_{\mathrm{KL}}(Z) \rightarrow D_{\mathrm{KL}}(q_{z|x} \,\|\, p_z) \\
&D_{\mathrm{KL}}(W) \rightarrow D_{\mathrm{KL}}\left(q_{w|x,y} \,\|\, p_{w|y}\right) \\
&\text{Adv.} \rightarrow -\mathbb{E}_{q_{z|x}}[\log g(y \mid z)] \\
&\text{Class.} \rightarrow \mathbb{E}_{q_{w|x,y}}[\log q(y \mid w)] \\
&\text{Rec. - }(Z) \rightarrow \mathbb{E}_{q_{z|x}}[\log p(x \mid z)]
\end{aligned}$$

Below, we provide additional details for the hyperparameters used in each experiment, and any other external resources used to obtain the corresponding sections' results. In addition, we include details regarding runtime and memory footprint of running experiments with the models included in our study.

Supplementary Table 13: Time spent per epoch during training for each dataset.

	P.M.	N.S.R	CMNIST	CelebA	scRNA-seq
CSVAE - N.A.	10.91s	8.02s	14s	54.43s	4.78s
CSVAE	12.71s	8.98s	14.41s	74.68s	4.99s
HCSVAE - N.A.	15.6s	10.92s	17s	44.3s	6.34s
HCSVAE	16.51s	11.8s	16.8s	68.53s	5.3s
DIVA	11.1s	8s	18.59s	48.96s	3.55s
CCVAE	12.1s	8.44s	17.9s	49.65s	5.25s
DisCoVR(Ours)	12.18s	8.73s	21.8s	109.86s	6.09s

Supplementary Table 14: Model inference time for a single batch for each dataset.

	P.M.	N.S.R	CMNIST	CelebA	scRNA-seq
CSVAE - N.A.	72ms	28ms	57ms	329ms	100ms
CSVAE	40ms	29ms	59ms	187ms	96ms
HCSVAE - N.A.	36ms	30ms	55ms	214ms	95ms
HCSVAE	37ms	39ms	52ms	191ms	119ms
DIVA	25ms	26ms	60ms	154ms	42ms
CCVAE	27ms	28ms	42ms	163ms	93ms
DisCoVR(Ours)	32ms	27ms	63ms	205ms	123ms

Supplementary Table 15: Memory footprint of running an experiment for each dataset.

	P.M.	N.S.R	CMNIST	CelebA	scRNA-seq
CSVAE - N.A.	53 MiB	255 MiB	1988 MiB	4868 MiB	298 MiB
CSVAE	253 MiB	255 MiB	2378 MiB	4812 MiB	300 MiB
HCSVAE - N.A.	254 MiB	255 MiB	2558 MiB	4588 MiB	292 MiB
HCSVAE	253 MiB	256 MiB	2998 MiB	4466 MiB	294 MiB
DIVA	253 MiB	255 MiB	2634 MiB	4996 MiB	300 MiB
CCVAE	253 MiB	255 MiB	3066 MiB	4998 MiB	300MiB
DisCoVR (Ours)	254 MiB	257 MiB	3612 MiB	7078 MiB	308MiB

H.1.1 Parametric model

for the parametric model, we consider $z, w \in \mathbb{R}$ and use multi-layer perceptrons (MLPs) with $n_{hidden} = 2, d_{hidden} = 8$ to parameterize approximate posteriors, the generative model and classifiers. For all models, we use learning rate $\gamma = 0.001$. A more detailed table of model-specific loss weights is provided in Supplementary Table 16.

Supplementary Table 16: Loss weights for the parametric model experiment.

	Rec.	$D_{\text{KL}}(Z)$	$D_{\text{KL}}(W)$	Adv.	Class.	Rec. - (Z)
CSVAE - No Adv.	1	1	1	—	—	—
CSVAE	2.5	1	0.5	20	—	—
HCSVAE - No Adv.	1	1	0.5	—	—	—
HCSVAE	2.5	1	0.5	20	—	—
DIVA	1	1	1	—	1	—
CCVAE	1	1	1	—	1	—
DisCoVR (Ours)	0.75	0.9	0.2	0.8	—	0.25

Supplementary Table 17: K-Means NMI for embeddings across stimulation (y) and cell type (common structure).

	w - Stimulation (↑)	z - Cell Type (↑)	z - Stimulation (↓)
CSVAE - No Adv.	0.949 ± 0.003	0.702 ± 0.015	0.187 ± 0.0
CSVAE	0.939 ± 0.002	0.406 ± 0.001	$\mathbf{0.002 \pm 0.0}$
HCSVAE - No Adv.	0.933 ± 0.006	0.628 ± 0.016	0.091 ± 0.002
HCSVAE	0.931 ± 0.005	0.433 ± 0.001	0.003 ± 0.0
DIVA	0.801 ± 0.0	0.628 ± 0.011	0.056 ± 0.0
CCVAE	0.604 ± 0.0	0.683 ± 0.016	0.103 ± 0.0
DisCoVR (Ours)	$\mathbf{0.946 \pm 0.002}$	$\mathbf{0.688 \pm 0.0}$	$\mathbf{0.002 \pm 0.0}$

H.1.2 NOISY SWISS ROLL

For this experiment, we consider $z, w \in \mathbb{R}^2$ and use MLPs with $n_{hidden} = 2, d_{hidden} = 128$ to parameterize approximate posteriors, the generative model and classifiers. For all models, we use learning rate $\gamma = 0.001$. A more detailed table of model-specific hyperparameters is provided in Supplementary Table 18.

Supplementary Table 18: Loss weights for the noisy Swiss roll experiment.

	Rec.	$D_{\text{KL}}(Z)$	$D_{\text{KL}}(W)$	Adv.	Class.	Rec. - (Z)
CSVAE - No Adv.	20	0.2	1	—	—	—
CSVAE	20	0.2	1	50	—	—
HCSVAE - No Adv.	20	0.2	1	—	—	—
HCSVAE	20	0.5	1	50	—	—
DIVA	20	0.2	0.2	—	1	—
CCVAE	20	0.2	0.2	—	1	—
DisCoVR (Ours)	0.9	0.2	0.2	8	—	0.1

H.1.3 NOISY COLORED MNIST

For this experiment, we consider $z \in \mathbb{R}^{20}$, $w \in \mathbb{R}^2$ and use convolutional neural networks (CNNs) to parameterize approximate posteriors and the generative model. For this example, DisCoVR can support z, w with different sizes, by parameterizing $p(w \mid y)$ through neural networks. For all models, we use learning rate $\gamma = 0.0001$. We detail the architectures and model-specific hyperparameters in Supplementary Tables 19 - 22. All other neural networks are formulated as MLPs with $n_{hidden} = 2, d_{hidden} = 4096$.

Supplementary Table 19: Image encoder architecture for noisy colored MNIST. Parameters for Conv2d are input / output channels. Parameters for MaxPool2D are kernel size and stride. Parameter for the linear layer is the output size. For variances, outputs are passed through an additional Softplus layer to ensure non-negativity.

Block	Details
1	Conv2d(3,32) + BatchNorm2D + ReLU
2	Conv2d(32,32) + BatchNorm2D + ReLU + MaxPool2D(2,2)
3	Conv2d(32,64) + BatchNorm2D + ReLU + MaxPool2D(2,2)
4	Conv2d(64,128) + BatchNorm2D + ReLU + MaxPool2D(2,2)
5	Linear(4096) + BatchNorm1D + ReLu
6	Linear(4096) + BatchNorm1D + ReLu
7	Linear(d_{latent})

Supplementary Table 20: Image decoder architecture for noisy colored MNIST. Parameters for Conv2d are input / output channels. Parameters for MaxPool2D are kernel size and stride. Parameter for the linear layer is the output size.

Block	Details
1	Linear(4096) + BatchNorm1D + ReLu
2	Linear(4096) + BatchNorm1D + ReLu
3	Linear(1152) + Unflatten
4	Upsample(2) + Conv2d(128, 64) + BatchNorm2D + ReLU
5	Upsample(2) + Conv2d(64, 32) + BatchNorm2D + ReLU
6	Upsample(2) + Conv2d(32, 32) + BatchNorm2D + ReLU
7	Conv2d(32, 3) + Sigmoid

Supplementary Table 21: Latent classifier architecture for noisy colored MNIST. Outputs parameterize logits of class probabilities.

Block	Details
1	Linear(4096) + BatchNorm1D + ReLu
2	Linear(4096) + BatchNorm1D + ReLu
3	Linear(2)

Supplementary Table 22: Loss weights for the noisy colored MNIST experiment.

	Rec.	$D_{\mathrm{KL}}(Z)$	$D_{\mathrm{KL}}(W)$	Adv.	Class.	Rec. - (Z)
CSVAE - No Adv.	1	0.0001	1	—	—	—
CSVAE	1	0.0001	1	1	—	—
HCSVAE - No Adv.	1000	0.0001	1	—	—	—
HCSVAE	10000	0.0001	1	1	—	—
DIVA	1	0.0001	0.0001	—	1	—
CCVAE	1	0.0001	0.0001	—	1	—
DisCoVR (Ours)	0.5	0.0001	0.0001	0.1	—	0.5

H.1.4 CELEBA-GLASSES

Motivated by the previous application of Klys et al. (2018), our choices follow those outlined in Larsen et al. (2016). We provide a detailed table of model-specific hyperparameters in Supplementary Table 23:

Supplementary Table 23: Loss weights for the CelebA-Glasses experiment.

	Rec.	$D_{\mathrm{KL}}(Z)$	$D_{\mathrm{KL}}(W)$	Adv.	Class.	Rec. - (Z)
CSVAE - No Adv.	1	0.0001	1	—	—	—
CSVAE	1000	0.0001	1	1	—	—
HCSVAE - No Adv.	1000	0.0001	1	—	—	—
HCSVAE	10000	0.0001	1	1	—	—
DIVA	100000	0.0001	0.0001	—	1	—
CCVAE	100000	0.0001	0.0001	—	1	—
DisCoVR (Ours)	1000000	0.0001	0.0001	2000	—	100000

H.1.5 SCRNA-SEQ

Following on the previous applications by Lopez et al. (2018), we use $z \in \mathbb{R}^{10}$, $w \in \mathbb{R}^2$. For DisCoVR we match the dimensions and use $w \in \mathbb{R}^{10}$ to avoid parameterizing the prior $p(w \mid z, y)$ with an additional neural network. We use MLPs with $n_{hidden} = 1, d_{hidden} = 128$ to parameterize approximate posteriors, the generative model and classifiers. We calculate K-Means NMI through *scikit-learn* (Pedregosa et al., 2011) by calling the `normalized_mutual_info_score` function with the original labels and the clusterings obtained by running KMeans on (1) the entire latent embedding and (2) single dimensions of the embedding and report the highest score. A more detailed table of model-specific hyperparameters is provided in Supplementary Table 24:

Supplementary Table 24: Loss weights for the scRNA-seq experiment.

	Rec.	$D_{\mathrm{KL}}(Z)$	$D_{\mathrm{KL}}(W)$	Adv.	Class.	Rec. - (Z)
CSVAE - No Adv.	1	0.0001	1	—	—	—
CSVAE	1	0.0001	1	100	—	—
HCSVAE - No Adv.	1	0.0001	1	—	—	—
HCSVAE	1	0.0001	1	100	—	—
DIVA	1	0.0001	0.0001	—	1	—
CCVAE	1	0.0001	0.0001	—	1	—
DisCoVR (Ours)	0.9	0.0001	0.0001	100	—	0.1

H.2 SUMMARY OF THE SCVI GENERATIVE MODEL FOR 5.2.3

Given batch key b and G genes, the generative model of scVI for a single cell $x_i \in \mathbb{N}^G$ is formulated as:

$$\begin{aligned} z_i &\sim \mathcal{N}(0, 1) \\ \rho_i &= f_\theta(z_i, b_i) \\ \pi_{ig} &= h^g_\phi(z_i, b_i) \\ x_{ig} &\sim \text{ZINB}(l_i \rho_i, \theta_g, \pi_{ig}) \end{aligned}$$

Here, g indexes genes, $l_i = \sum_g x_{ig}$ denotes the total number of counts for a single cell, z_i denotes the latent representation of the cell, and ρ_i denotes the normalized expression of the cell. f_θ is formulated as a neural network with a final softmax layer. h_ϕ is a neural network used to parameterize zero-inflation probabilities for the generative zero-inflated negative binomial (ZINB) distribution.

As such, for a single batch, the formulation of scVI is equivalent to the VAE with a ZINB likelihood. While all other models can be extended easily, DisCoVR requires reconstructions as a proxy for the adversarial loss. For this formulation, we directly treat the normalized expressions ρ_i as the adversarial reconstructions $\hat{x}$.