PGM: Project Report

Characterizing Labels with CCVAE: Latent Representations and Interventions

Ambre Etienne
ambre.etienne@ensae.fr

Mahdi Ranjbar
mahdi.ranjbar@ens-paris-saclay.fr

1 Introduction

The quest to comprehend and manipulate the characteristic factors of perceptual observations has long been central to machine intelligence. Variational autoencoders (VAEs) offer a versatile framework for meaningful representation learning, often leveraging labels. However, conventional approaches directly associating latent variables with labels can limit manipulation tasks. In response, we propose the Characteristic Capturing VAE (CCVAE) [4], explicitly capturing label characteristics in the latent space, offering enhanced manipulation capabilities while maintaining predictive accuracy.

Our model leverages the interplay between labels and inputs, capturing more information than labels alone convey. Conceptually, this work shares similarities with interventions on VAEs for text data [5], yet distinguishes itself by leveraging labels for explicit characteristic information capture.

2 Characteristic Capturing Variational Autoencoders

2.1 VAEs and SSVAEs

The aim of our project is to understand the rich characteristic information associated with labels. We consider VAEs because they blend deep autoencoder unsupervised learning with generative latent-variable models, representing data as distributions. Employing variational inference, VAEs use neural networks for encoding and decoding to learn representations and construct an approximate posterior. The objective involves maximizing the marginal likelihood through the evidence lower bound (ELBO) :

$$\log p_\theta(x) = \log \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(z,x)}{q_\phi(z|x)}\right] \geq \mathbb{E}_{q_\phi(z|x)}\left[\log(\frac{p_\theta(z,x)}{q_\phi(z|x)})\right] \equiv \mathcal{L}(x;\phi,\theta)$$

Semi-supervised Variational Autoencoders (SSVAEs) integrate labeled data into the VAE framework. They decompose the log-marginal likelihood into the sum of labeled and unlabeled data log-likelihoods, each bounded by their respective ELBOs. SSVAEs commonly split the latent space into z_y (representing labeled information) and $z_{\setminus y}$ (for unlabeled data). The encoder imputes z_y values for unlabeled instances, making SSVAEs versatile for applications beyond semi-supervised classification, such as learning meaningful representations

and manipulating data based on labeled information. But in this project our focus is on structuring the latent space to disentangle label-associated characteristics.

2.2 CCVAEs

Assuming that labels directly correspond to a part of the latent space (z_y) can cause issues in capturing rich label-related characteristics and manipulating data. For example, difficulty manipulating a characteristic without modifying the labels. The assumption can limit interventions and cause a mismatch between the VAE prior and data distribution. So direct correspondence between labels and latent variables may not be necessary, and the assumption can be counterproductive, limiting the range of interventions.

The article from Tom Joy [4] proposes Characteristic-Capturing Variational Autoencoders (CCVAEs) to address issues in treating labels as direct components of the latent space in VAEs. It suggests conditioning latent variables on labels to capture characteristics effectively. CCVAEs split the latent space into characteristic (z_c) and non-characteristic ($z_{\setminus c}$) components.

However, the characteristics of different labels become entangled within z_c. To disentangle it, we partition the latent space, such that the classification of particular labels y_i only has access to particular latents z_c^i and thus $\log q_\varphi(y|z_c) = \sum_i \log q_{\varphi_i}(y^i|z_c^i)$. This forces the characteristic information needed to classify y^i to be stored only in the corresponding z_c^i. Additionally, we can introduce a factorized set of generative models $p(z_c|y) = \prod_i p(z_c^i|y^i)$, enabling easy generation and manipulation of z_c^i individually. The final graphical model is illustrated in Figure 6.

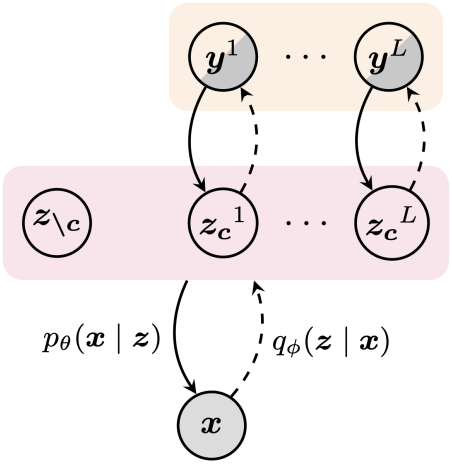


Figure 1: CCVAE graphical model

To do so, the article introduces a model objective, L_{CCVAE}, that formulates a lower bound on the full model log-likelihood, accounting for supervised and unsupervised subsets.

The supervised objective is defined as:

$$\log p_{\theta,\psi}(x,y) \geq \mathbb{E}_{q_{\phi,\phi}(z|x,y)} \log q_{\phi,\phi}(z|x,y) \equiv L_{CCVAE}(x,y)$$

Where $p_\psi(z|y) = p(z)p_\psi(z_c|y)$. Leveraging conditional independence and Bayes' rule, this leads to the following equation:

$$\begin{aligned} L_{CCVAE}(x,y) = \mathbb{E}_{q_\phi(z|x)} q_{\phi,\phi}(y|x) \log q_\phi(y|z_c) q_\phi(z|x) \\ + \log q_{\phi,\phi}(y|x) + \log p(y) \end{aligned}$$

A classifier term, $\log q_{\phi,\phi}(y|x)$, naturally emerges, crucially demonstrating that not placing labels directly in the latent space avoids the detachment of mappings observed in previous models. The classifier strength can be adjusted, but the approach proves insensitive, avoiding the need for manual tuning. For unlabeled data, variational inference is performed, and the unsupervised objective, $L_{CCVAE}(x))$, is derived as a standard ELBO. Combining the 2 equations yields the following lower bound:

$$\log p(D) \geq \sum_{(x,y)\in S} L_{CCVAE}(x,y) + \sum_{x\in U} L_{CCVAE}(x)$$

3 Interventions

To demonstrate the qualitative aspects of representation disentanglement, we often assess it through independent exploration of latent dimensions, commonly known as latent traversals. Our study highlights the efficacy of CCVAE in interventions, showcasing its capacity to selectively isolate label-specific characteristics and provide precise control over interventions.

To achieve this, we employ CCVAE to conduct latent traversals and compare its performance against a modified version of DIVA, as proposed by Ilse et al. (2019) [3]. Our aim is to underscore the sophistication of CCVAE in handling such tasks.

Our evaluation is conducted in a multi-label context using the CelebA dataset, in which we extract a detailed description of 18 distinct labels. The encoder and decoder architectures are adapted from Higgins et al. (2016), with specific modifications to accommodate label-predictive distribution and conditional prior.

In the context of conventional interventions, where one or more labels undergo changes, we can effortlessly resample the associated latent variable z_c^i, thereby generating new characteristics aligned with the altered labels. Moreover, CCVAE offers flexibility in alternative interventions, such as identifying the closest z_c^i to the original that induces a label change—reminiscent of adversarial attacks. Alternatively, manipulation of z_c^i is possible without altering the class itself, enabling exploration of characteristics consistent with the specified labels.

4 Experiments

4.1 Interventions

4.1.1 CelebA

For our experiments, we define the generative and inference networks for CCVAEs as follows.

The approximate posterior is represented as

$$q_\phi(z|x) = \mathcal{N}(z_c, z_{\setminus c}|\mu_\phi(x), \text{diag}(\sigma^2_\phi(x)))$$

with $\mu_\phi(x)$ and $\text{diag}(\sigma^2_\phi(x))$ following the architecture from Higgins et al. (2016) [2].

The generative model $p_\theta(x|z)$ is represented by a Laplace distribution, again parameterized using the architecture from Higgins et al. (2016).

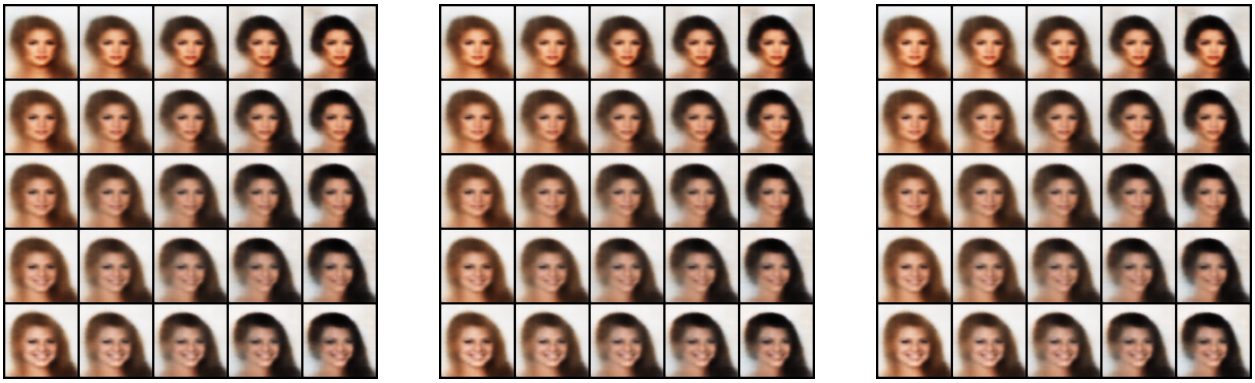
The label predictive distribution $q_\phi(y|z_c)$ is represented as $Ber(y|\pi_\phi(z_c))$ with $\pi_\phi(z_c)$ being a diagonal transformation forcing the factorization $q_\phi(y|z_c) = \prod_i q_{\psi_i}(y_i|z_{c_i})$.

The conditional prior is given as $p_\psi(z_c|y) = \mathcal{N}(z_c|\mu_\psi(y), \text{diag}(\sigma^2_\psi(y)))$, with the appropriate factorization, where the parameters are represented by an MLP.

Finally, the prior placed on the portion of the latent space reserved for unlabelled latent variables is $p(z_{\setminus c}) = \mathcal{N}(z_{\setminus c}|0, I)$.

For the latent space $z_c \in \mathbb{R}^{m_c}$ and $z_{\setminus c} \in \mathbb{R}^{m_{\setminus c}}$, where $m = m_c + m_{\setminus c}$ with $m_c = 18$ and $m_{\setminus c} = 27$. The architectures of the encoder, the decoder, the classifier and the conditional prior are given in the appendix.

The model underwent optimization using the Adam optimizer with a learning rate set to 2×10^{-4}. Training sessions were conducted for 10, 50, and 100 epochs, and the outcomes are presented in Figure 4. Notably, as the number of epochs increases, the latent traversal exhibits enhanced performance, allowing for a more distinct discernment of variations in characteristics.



(a) 10 epochs (b) 50 epochs (c) 100 epochs

Figure 2: CCVAE latent walk Smiling and Black Hair

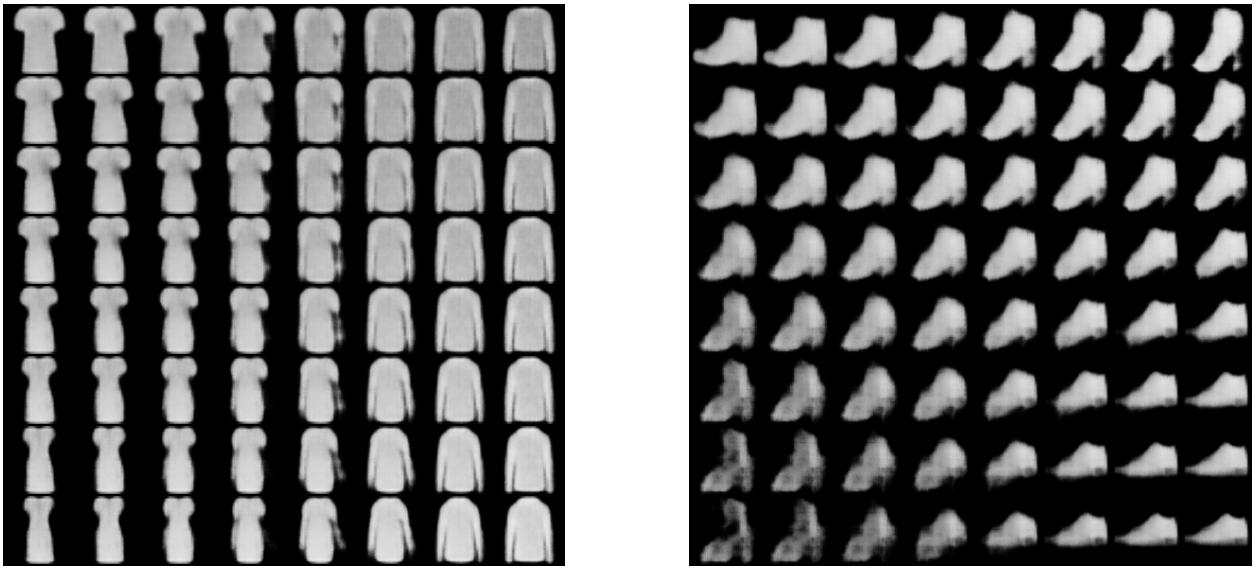
4.1.2 FashionMNIST

CCVAE demonstrates its versatility not only in addressing multi-label challenges but also in effectively handling multi-class problems. In this context, we present results obtained from

the multi-class setting using FashionMNIST dataset, even though this aspect is somewhat tangential to our primary focus. Despite its tangential nature, we include these results for the sake of comprehensive exploration.

Unlike the multi-label scenario, where encapsulating representations for each label is straightforward, the multi-class setting introduces complexities. In the context of Fashion-MNIST, the challenge lies in determining how one might adjust the representation of an item recognized as, for instance, a sneaker, while simultaneously maintaining its representation as a handbag. In this unique context, there is essentially only one label, albeit with multiple values. To address this, we exercise flexibility in structuring the latent space for CCVAE.

Explicit choices are made regarding the latent space's configuration—whether $z_c \in \mathbb{R}$ or $z_c \in \mathbb{R}^N$, or, conversely, whether all the representation is stored in z_c (i.e., $z_{\setminus c} = \emptyset$). Furthermore, the label predictive distribution $q_\phi(y|z_c)$ is represented as $Cat(y|\pi_\phi(z_c))$. The factorization $q_\phi(y|z_c) = \prod_i q_{\psi_i}(y_i|z_{c_i})$ is not enforced in this context. Instead, we employ a transformation which is parameterized by a function $F : \mathbb{R}^N \to \mathbb{R}^M$, where M is the number of possible classes. Latent traversals are also performed in the multi-class setting, utilizing linear interpolation on a polytope defined by the network's outputs $\mu_\psi(y)$ for four distinct classes. The reconstructions are visually presented in 3.



(a) T-Shirt - Dress - Shirt - Coat

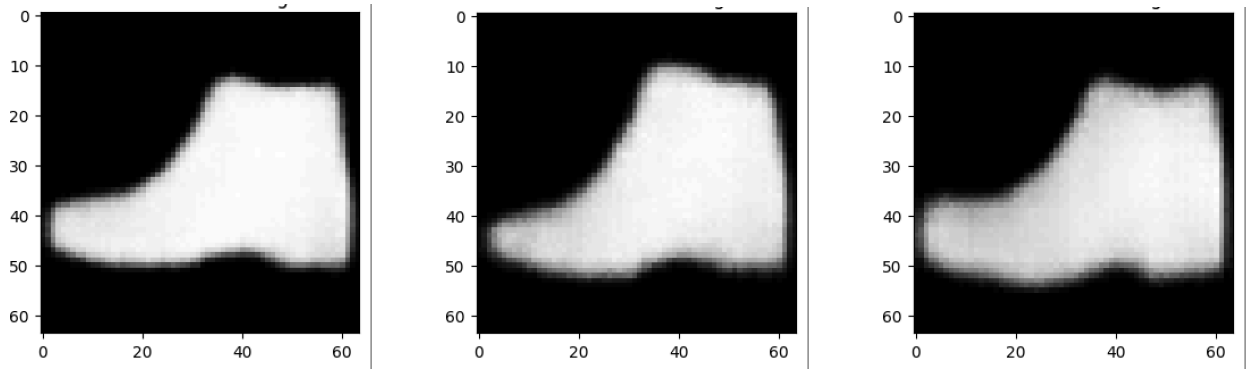
(b) Sandal - Ankle Boot - Sneaker

Figure 3: CCVAE latent traversals for FashionMNIST

We train CCVAE using a learning rate of 2×10^{-4} for 100 epochs. In 3a, we observe CCVAE's ability to transition from a t-shirt to a dress by elongating the length. In 3b, we demonstrate how CCVAE facilitates the transformation from flat boots to high-heeled boots by creating and increasing the heel.

4.2 Reconstruction

We evaluated the role of the parameter fraction of supervision in the reconstruction of image task and the results can be seen in the following figures **??**:



(a) filly unsupervised (b) supervision fraction =0.5 (c) fully supervised

Figure 4: Reconstruction varying the supervision fraction parameter

5 Extended work

5.1 Introduction

Being interested in Natural Language Processing, we decided to extend our work to this domain and consider to study a deep generative model for unsupervised text style transfer [1]. This model uses a probabilistic approach, hypothesizing a latent sequence to transform sequences between domains without supervision. The model employs a recurrent language model as a prior and an encoder-decoder for transduction. While computing marginal data likelihood is challenging, amortized variational inference provides a practical surrogate. The text connects the variational objective to other unsupervised style transfer and machine translation techniques, unifying non-generative objectives like backtranslation and adversarial loss. Although in the paper several task in text style transfer such as formality transfer, word decipherment and so on are discussed, we only focus our attention to sentiment transfer which means paraphrasing a sentence with a different sentiment while preserving the original content. Evaluation consist of considering three aspects: attribute control, content preservation, and fluency. A successful system needs to perform well with respect to all three aspects.

5.2 Architecture

There is an encoder-decoder architecture based on the standard attentional Seq2Seq model. Distributions $q(\bar{y}|x)$ and $q(\bar{x}|y)$ represent inference networks that approximate the model's true posterior. Critically, parameters are shared between the generative model and inference networks to tie the learning problems for both domains.

5.3 Learning Objective

Ideally, we aim to improve learning by optimizing the log data likelihood. However, the model's neural parameterization complicates direct computation of the data likelihood using dynamic programming. To address this challenge, using amortized variational inference the

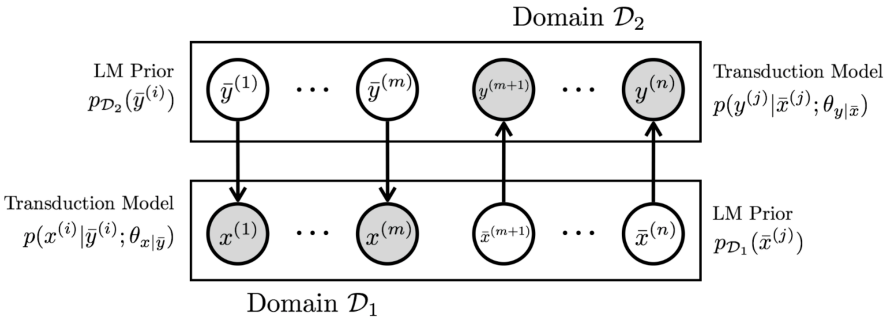


Figure 5: LM graphical model: shaded circles denote the observed variables and unshaded circles denote the latents. The generator is parameterized as an encoder-decoder architecture and the prior on the latent variable is a pretrained language model.

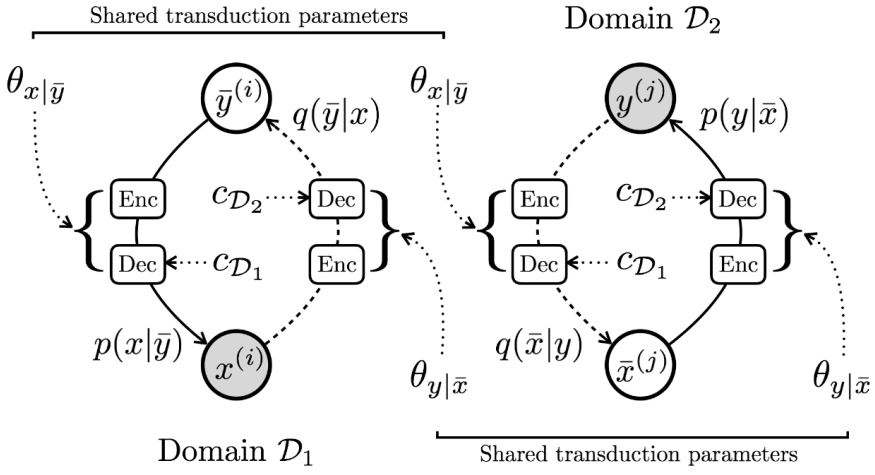


Figure 6: Model Architecture

learning objective is the evidence lower bound (ELBO) on log marginal likelihood which is shown in Figure 7.
The reconstruction and KL terms still involve intractable expectations due to the marginalization over the latent sequence, thus we need to approximate their gradients. However, when working with the vast possibilities of the latent text variables, traditional methods like REINFORCE result in too much variability. So, to get around the problem, it turns out that there is a simpler approach—greedy decoding without keeping track of gradients—to approximate the reconstruction term. It's important to note that even though the inference networks still get gradients from the prior through the KL term, their parameters are shared with the decoders, which do receive gradients from the reconstruction.

5.4 Data

We worked with the Yelp reviews dataset, consisting of 250,000 negative sentences and 380,000 positive sentences. Additionally, we have a smaller test set containing 1,000 human-annotated parallel sentences. The positive sentiment domain is labeled as D1 and the negative sentiment domain is labeled as D2. To evaluate the output, Self-BLEU and BLEU are used.The Self-BLEU and BLEU represent score against the original sentence and the reference, respectively.

5.5 Result

The results of the model in the style transfer task is shown in the tables 1 and 2. It can be noticed that the model perfectly trandfer the style both from positive to negative and inverse.

Original	**Style Transfer Model**
The service was incredibly slow, and the staff was unhelpful.	The service was unhurried, and the staff was unreservedly helpful.
This product is a complete waste of money, don't buy it.	This product is worth every penny, definitely buy it.
The customer support is terrible, never responds on time.	The customer support is exceptional, always responsive.
The book was poorly written and lacked any depth.	The book was well-written and full of depth.

Table 1: Negative to Positive Sentiment Transfer Examples

5.6 Conclusion

The probabilistic framework interprets text style transfer as amortized variational inference in a generative model and the resulting objective is different from previous state of the art result.

Original	Style Transfer Model
The hotel offers luxurious amenities and impeccable service.	The hotel lacks basic amenities and provides substandard service.
The team's synergy leads to remarkable outcomes.	The team's collaboration is non-existent; they struggle with basic tasks.
The novel delves into complex human relationships with rich characters.	The novel oversimplifies relationships with uninteresting characters.

Table 2: Positive to Negative Sentiment Transfer Examples

6 Conclusion

In conclusion, our exploration of CCVAE's capabilities in representation disentanglement across both multi-label and multi-class settings has unveiled its remarkable flexibility and effectiveness. By delving into diverse datasets, including CelebA and FashionMNIST, we have demonstrated CCVAE's proficiency in isolating and manipulating specific label characteristics. The model's adaptability to various latent space configurations and its successful handling of unconventional interventions underscore its robustness. This study not only contributes to the understanding of disentangled representations but also emphasizes CCVAE's potential applicability in a wide range of complex scenarios, showcasing its versatility in capturing intricate relationships within diverse datasets.

References

[1] Junxian He et al. "A probabilistic formulation of unsupervised text style transfer". In: 2020.

[2] Irina Higgins et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=Sy2fzU9gl.

[3] Maximilian Ilse et al. *DIVA: Domain Invariant Variational Autoencoders*. 2019. arXiv: 1905.10427 [stat.ML].

[4] Tom Joy et al. *Capturing Label Characteristics in VAEs*. 2022. arXiv: 2006.10102 [cs.LG].

[5] Jonas Mueller, David Gifford, and Tommi Jaakkola. "Sequence to Better Sequence: Continuous Revision of Combinatorial Structures". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 2536–2544. URL: https://proceedings.mlr.press/v70/mueller17a.html.

$$
\begin{aligned}
&\log p(X, Y; \theta_{x|\bar{y}}, \theta_{y|\bar{x}}) \\
&\geq \mathcal{L}_{\text{ELBO}}(X, Y; \theta_{x|\bar{y}}, \theta_{y|\bar{x}}, \phi_{\bar{x}|y}, \phi_{\bar{y}|x}) \\
&= \sum\nolimits_i \left[\mathbb{E}_{q(\bar{y}|x^{(i)}; \phi_{\bar{y}|x})}[\log p(x^{(i)}|\bar{y}; \theta_{x|\bar{y}})] - D_{\text{KL}}\left(q(\bar{y}|x^{(i)}; \phi_{\bar{y}|x}) || p_{\mathcal{D}_2}(\bar{y})\right) \right] \\
&+ \sum\nolimits_j \underbrace{\left[\mathbb{E}_{q(\bar{x}|y^{(j)}; \phi_{\bar{x}|y})}[\log p(y^{(j)}|\bar{x}; \theta_{y|\bar{x}})] \right.}_{\text{Reconstruction likelihood}} \underbrace{\left. - D_{\text{KL}}\left(q(\bar{x}|y^{(j)}; \phi_{\bar{x}|y}) || p_{\mathcal{D}_1}(\bar{x})\right) \right]}_{\text{KL regularizer}}
\end{aligned}
$$

Figure 7: Learning Objective