Out-of-Distribution Detection with An Adaptive Likelihood Ratio on Informative Hierarchical VAE

Anonymous Author(s)
Affiliation
Address
email

Abstract

Unsupervised out-of-distribution (OOD) detection is essential for the reliability of machine learning. In the literature, existing work has shown that higher-level semantics captured by hierarchical VAEs can be used to detect OOD instances. However, we empirically show that, the inheirt "*posterior collapse*" of hierarchical VAEs would seriously limit their capacity for OOD detection. Based on a thorough analysis, we propose an informative hierarchical VAE to alleviate this issue through enhancing the connections between the data sample and its multi-layer stochastic latent representations during training. Furthermore, we propose a novel score function for unsupervised OOD detection, referred to as Adaptive Likelihood Ratio, which can selectively aggregate the semantic information on multiple hidden layers of hierarchical VAEs, leading to a strong separability between in-distribution and OOD samples. Experimental results demonstrate that our method can significantly outperform existing state-of-the-art unsupervised OOD detection approaches.

1 Introduction

Despite achieving great success in real-world applications recently, existing machine learning (ML) systems are still designed to be tested on the in-distribution dataset, whose statistics are similar to those of the training set [1]. However, when applied to deal with the dataset consisting of out-of-distribution (OOD) samples, whose statistics are extremely different from those of the training set, these ML systems would produce a series of incorrect judgments [2, 3]. Considering the fact that OOD data is very common in real-world applications, pre-execution OOD detection is increasingly attractive to make sure the reliability and safety of ML systems. Although several supervised methods [4, 5, 6, 7, 8, 9, 10, 11, 12] have achieved great success in OOD detection, the unsupervised ones are more practical since the category labels of in-distribution samples are often missing in real-world applications, which brings more challenges for OOD detection and is also the focus of this work.

Without labels, likelihood-based models could be a promising way for unsupervised OOD detection, such as flow-based models [13], auto-regressive models [14, 15], and variational autoencoders (VAEs) [16, 17]. Unfortunately, some recent studies have shown that, in some cases, these generative models tend to achieve higher likelihoods on certain types of OOD samples [1, 18, 19, 20, 21], which makes the OOD detection methods based on thresholding likelihood scores problematic. To address this issue, based on the prior knowledge collected from OOD samples, some improvements have been made for unsupervised OOD detection, *i.e.*, Ren et al. [1] take additional datasets as the OOD validation sets for choosing the best hyperparameters; Hendrycks et al. [22] introduce an auxiliary dataset to teach the network to learn more expressive representations for OOD detection; However, the behaviour of borrowing the prior knowledge of OOD data is usually unreasonable in practice, because we will never know the statistic information of OOD samples to be dealt with.

To conduct purely unsupervised OOD detection without the labels or prior assumptions, deep ensemble method named WAIC [18] is developed by making full use of the difference between the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

density estimations of multiple independent models trained on the in-distribution data. Recently, through capturing the semantic information with multi-layer latent variables, Maaløe et al. [23] and Havtorn et al. [21] develop various score functions based on Likelihood-Ratio, which help hierarchical VAEs achieve competitive performance in unsupervised OOD detection. However, in our implementation, we find that the phenomenon of "*posterior collapse*" in hierarchical VAEs still limits their performance on OOD detection, where the main reason could be that "*posterior collapse*" will make the high-level latent variables meaningless and cannot provide faithful summaries for the data.

With this insight in hand, in this paper, we start from rethinking the cause of "*posterior collapse*" in hierarchical VAEs, and then theoretically explain why "*posterior collapse*" will limit the OOD detection performance of these Likelihood-Ratio based methods. Further, we develop an informative hierarchical VAE to alleviate "*posterior collapse*" and a novel Adaptive Likelihood Ratio score function for unsupervised OOD detection. The major contributions of this work include:

- With a thorough analysis of "*posterior collapse*" in hierarchical VAEs, we enhance the connections between the data sample and its multiple latent representations in the expected log-likelihood term of evidence lower bound (ELBO) for training, and develop a novel informative hierarchical VAE to extract more expressive hierarchical latent representations.
- We theoretically explain why alleviating "*posterior collapse*" in hierarchical VAEs can help the performance of Likelihood Ratio on OOD detection, and then develop a novel score function for fully unsupervised OOD detection, termed Adaptive Likelihood Ratio, which owns fewer hyperparameters to be tuned and can make full use of the semantic divergences between in-distribution and OOD samples across all hidden layers of hierarchical VAEs.
- Combing the informative hierarchical VAE with the Adaptive Likelihood Ratio, we demonstrate that our method can achieve state-of-the-art OOD detection performance across a wide range of benchmarks in an unsupervised manner.

2 Background and Related Works

2.1 Hierarchical Variational Autoencoder

Preliminary: Extending the basic VAE [16], a hierarchical VAE [24, 25] is defined by the observation $\boldsymbol{x}$ that depends on a hierarchy of stochastic latent variables $\boldsymbol{z} = \boldsymbol{z}_1, ..., \boldsymbol{z}_L$, where the generative model is defined with a top-down structure, formulated as $p_\theta(\boldsymbol{x}, \boldsymbol{z}) = p_\theta(\boldsymbol{x}|\boldsymbol{z}_1) \prod_{l=1}^{L-1} p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}) p_\theta(\boldsymbol{z}_L)$; the inference model is designed to approximate the posterior over these latent variables, commonly factorized with a top-down structure as $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \prod_{l=1}^{L-1} q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}) q_\phi(\boldsymbol{z}_L|\boldsymbol{x})$ or a bottom-up structure as $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \prod_{l=1}^{L-1} q_\phi(\boldsymbol{z}_{l+1}|\boldsymbol{z}_l) q_\phi(\boldsymbol{z}_1|\boldsymbol{x})$. The demonstration of these structures could be seen in Fig. 1. The parameters of the generative and infernce models, denoted as θ and ϕ respectively, can be jointly optimized by maximizing the evidence lower bound (ELBO) expressed as

$$\mathcal{L} = \mathbb{E}_{p(\boldsymbol{x})} \left[\mathcal{L}_x(\, \boldsymbol{x}; \theta, \phi) \right], \tag{1}$$

where $\mathcal{L}_x$ is denoted as

$$\mathcal{L}_x = \log p(\boldsymbol{x}) - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) || p_\theta(\boldsymbol{z}|\boldsymbol{x})) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}) \right] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) || p_\theta(\boldsymbol{z})), \tag{2}$$

where $D_{\mathrm{KL}}(\cdot||\cdot)$ denotes the KL divergence and maximizing ELBO is equivalent to minimize the divergence between the varaitional distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and true posterior $p_\theta(\boldsymbol{z}|\boldsymbol{x})$.

Related Works: While the hierarchy of latent stochastic variables can improve the generation capability of standard VAEs, in practice, the posterior of higher-level stochastic latent variables have a tendency to *collapse* into the prior, called "*posterior collapse*". To address this issue, Sønderby et al. [24] propose a Ladder VAE (LVAE) to change the bottom up inference process into a top-down one; Vahdat and Kautz [25] develop a variant of the LVAE, which carefully designs a sophisticated network architecture to achieve better generation quality; Maaløe et al. [23] combine the bottom-up inference with the top-down inference by proposing a bidirectional inference scheme. Despite obtaining performance improvements with more flexible inference networks, these hierarchical VAEs still lack a theoretical guide to alleviate the phenomenon of "*posterior collapse*".

2.2 Out-of-distribution Detection with Variational Autoencdoer

Problem Formulation: Suppose that there are a set of N training samples $\{\boldsymbol{x}_i\}_{i=1}^N$ drawn from the data distribution $\boldsymbol{x}_i \sim p(\boldsymbol{x})$, after training a VAE to model the generation of these data samples,

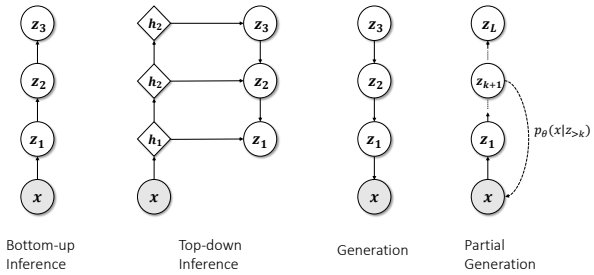


Figure 1: Illustration of the usual structures of both inference network and generative model in hierarchical VAEs.

the generative model $p_\theta(\boldsymbol{x})$ is supposed to detect whether a testing sample $\boldsymbol{x}$ is an outlier, where the outlier should have a low density estimation under the true data distribution $p(\boldsymbol{x})$. However, on the contrary, former likelihood-based methods found the generative model always assign higher $p_\theta(\boldsymbol{x})$ for OOD data than in-distribution data [1, 18, 19, 20]. Luckily, although the likelihood based methods with VAE are rarely investigated, they have revealed the potential to better address this problem without the help of labels or assumptions about the prior.

Related Works: A pioneering VAE-based OOD detection method is Likelihood Regret (LRe) [26] obtained by iteratively finetuning the decoder parameters of VAE, which is time-consuming but achieves competitive performance in an unsupervised manner. Maaløe et al. [23] reveal the potential of hierarchical VAE for OOD detection and Havtorn et al. [21] further propose a score function to improve the model performance, which needs to specify the backbone model. The backbone models in likelihood-based methods [1, 18, 27] can be directly replaced with VAEs, but these methods usually underperform flow-based models like Glow [13] or autoregressive models like PixelCNN [15].

3 From Informative Hierarchical VAE to Adaptive Likelihood Ratio

3.1 Rethinking of "*posterior collapse*" in Hierarchical VAEs

Firstly, let's understand the cause of "*posterior collapse*" in hierarchical VAEs theoretically. Taking an L-layer hierarchical VAE with a top-down inference network as an example, the set of latent variables can be separated as the lower-level variables $\boldsymbol{z}_{\leq k} = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_k\}$ and the higher-level ones $\boldsymbol{z}_{>k} = \{\boldsymbol{z}_{k+1}, ..., \boldsymbol{z}_L\}$, where $k \in \{0, ..., L-1\}$, then the ELBO in Eq. (1) can be reformulated as

$$\mathcal{L} = \mathbb{E}_{p(\boldsymbol{x})}\left[\mathbb{E}_{q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_1)\right] - \sum\nolimits_{l=1}^{L} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right], \tag{3}$$

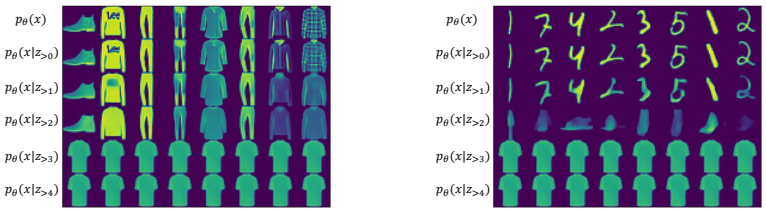
where $q_\phi(\boldsymbol{z}_L|\boldsymbol{z}_{L+1}) := q_\phi(\boldsymbol{z}_L|\boldsymbol{x})$, $p_\theta(\boldsymbol{z}_L|\boldsymbol{z}_{L+1}) := p_\theta(\boldsymbol{z}_L)$, and the main contribution to the expected log-likelihood term is coming from the lower-level latent variables $\boldsymbol{z}_{\leq k}$ before the kth hidden layer [23]. Once the generation capacity of the generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{\leq k})$ is powerful enough to reconstruct the observation $\boldsymbol{x}$ well, the variational posteriors of higher-level latent variables $\boldsymbol{z}_{>k}$ will be optimized to be close to their priors, i.e., $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x}) \approx p_\theta(\boldsymbol{z}_{>k})$, leading the representations learned by VAE at higher layers to be meaningless and cannot provide faithful summaries for $\boldsymbol{x}$, which is well-known as the phenomenon of "*posterior collapse*" or "*latent variable collapse*" [24, 28].

To find the potential solutions to alleviating "*posterior collapse*", in the following, we reinterpret this phenomenon from the perspective of information theory [29] by extending the findings in [28] to a hierarchical VAE scenario. For ease of undertanding, we define the mutual information between the data $\boldsymbol{x}$ and the higher-level latent variables $\boldsymbol{z}_{>k}$ as

$$\mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}_{>k}) = -\mathcal{H}_q(\boldsymbol{z}_{>k}|\boldsymbol{x}) + \mathcal{H}_q(\boldsymbol{z}_{>k}) = \mathbb{E}_{p(\boldsymbol{x})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x}) - \mathbb{E}_{q_\phi(\boldsymbol{z}_{>k})} \log q_\phi(\boldsymbol{z}_{>k}),$$

which is induced by the variational posterior $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})$. Then KL term in Eq. (3) can be rewritten as

$$\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{L} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + \mathbb{E}_{p(\boldsymbol{x})}\left[D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})||p_\theta(\boldsymbol{z}_{>k}))\right] \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + \mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}_{>k}) + D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_{>k})||p_\theta(\boldsymbol{z}_{>k})),
\end{aligned} \tag{4}$$



(a) FashionMNIST (in) (b) MNIST (out)

Figure 2: Illustration of "*posterior collapse*" in a 5-layer hierarchical VAE trained on FashionMNIST (in) by visualizing its reconstructions conditioned on various $\boldsymbol{z}_{>k}$ for both in-distribution and OOD samples. The first row is the input $\boldsymbol{x}$, and the other rows are generated from the partial generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k})$ by taking $\boldsymbol{z}_{>k}$ drawn from $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})$ as input, where $k \in \{0, ..., 4\}$.

where $q_\phi(\boldsymbol{z}_{>k}) = \mathbb{E}_{p(\boldsymbol{x})}\left[q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})\right]$ and the detailed derivation can be found in Appendix A. By substituting Eq. (4) into Eq. (3), due to the non-negativity of mutual information and KL divergence, we can find that maximizing the ELBO is opposite to maximizing the mutual information $\mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}_{>k})$. When $\mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}_{>k})$ is minimized to zero, the variational posterior $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})$ will be independent of the data $\boldsymbol{x}$, which leads to the phenomenon of "*posterior collapse*".

3.2 Why "*posterior collapse*" limits Likelihood-Ratio for OOD Detection

OOD detection has become one of the most important applications of VAEs, which can be applied to filter OOD samples by setting a threshold on the score of log-likelihood term. However, some recent studies have shown that, in some cases, VAEs tend to achieve higher likelihoods on certain types of OOD samples [30], which makes the OOD detection rules based on likelihood threshold problematic. Recently, Havtorn et al. [21] reinterpreted this problematic behavior by providing evidence that the low-level features learned by VAEs generalize well across datasets and dominate the estimated likelihoods. Inspired by the alternative log-likelihood lower bound [23] that partly replaces the inference network with the generative model to highlight high-level features, formulated as

$$\mathcal{L}_{\boldsymbol{x}}^{>k} = \log p(\boldsymbol{x}) - D_{\mathrm{KL}}(p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})), \tag{5}$$

Havtorn et al. [21] considered to subtract $\mathcal{L}_{\boldsymbol{x}}^{>k}$ from $\mathcal{L}_{\boldsymbol{x}}$ to cancel out the data distribution $\log p(\boldsymbol{x})$, resulting in a likelihood-ratio score for unsupervised OOD detection as

$$\mathcal{LLR}^{>k} = D_{\mathrm{KL}}(p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})) - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})), \tag{6}$$

which discards the likelihood term to prevent the low-level features from dominating and measures divergence in the latent space to ensure that data should be in-distribution across all feature levels.

To intuitively understand the nature of success in the likelihood-ratio score and illustrate why alleviating *"posterior collapse"* in hierarchical VAEs can improve its performance on OOD detection, we provide an insightful analysis on $\mathcal{LLR}^{>k}$ in Eq. (6) by reformulating it as follows:

$$\begin{aligned}\mathcal{LLR}^{>k} &= \mathbb{E}_{q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}\left[D_{\mathrm{KL}}(p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})||p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k}, \boldsymbol{x})) - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})||p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k}, \boldsymbol{x})\right] \\ &\approx \mathbb{E}_{q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}\left[D_{\mathrm{KL}}(p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})||q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k}))\right],\end{aligned} \tag{7}$$

where the detailed derivations can be found in Appendix B. Eq. (7) shows that, when the inference network $q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$ can approximate the true posterior $p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k}, \boldsymbol{x})$ very well, thorough equivalent replacement, $\mathcal{LLR}^{>k}$ will approach the expected KL divergence between the prior $p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$ and variational posterior $q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$. More specifically, conditioned on the expectation of $\boldsymbol{z}_{>k}$ drawn from its variational posterior $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})$, $\mathcal{LLR}^{>k}$ is developed to calculate the summation of k KL divergence terms, measuring the distance between the lower-level variables $\boldsymbol{z}_{\leq k}$ drawn from the generative model $p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$ and those from the variational inference network $q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$.

The premise of applying $\mathcal{LLR}^{>k}$ to OOD detection is that, after training a hierarahical VAE on in-distribution samples, for each OOD sample, the latent variables $\boldsymbol{z}_{\leq k}$ generated from the generative model $p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$ will be clearly distinct from those drawn from the variational inference network $q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$. For ease of understanding the principle, after training a 5-layer hierarchical VAE on FashionMNIST, for each hidden layer, we exhibit the reconstructions of both in-distribution (FashionMNIST) and OOD (MNIST) data samples with the partial generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k})$

conditioned on the latent variables $\boldsymbol{z}_{>k}$ drawn from $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})$ as shown in Fig. 2. From the results, we can find that, when setting $k = 2$, the reconstructions of MNIST samples tend to reflect the high-level semantic structures learned from FashionMNIST, indicating that the generation mechanism of $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>2})$ seems to prevent accurate reconstruction of out-of-distribution data, which implies that a score function based on the distance between $p_\theta(\boldsymbol{z}_{\leq 2}|\boldsymbol{z}_{>2})$ and $q_\phi(\boldsymbol{z}_{\leq 2}|\boldsymbol{z}_{>2})$, like $\mathcal{LLR}^{>2}$ in Eq. (7), could be a promising metric for OOD detection.

However, when the phenomenon of "*posterior collapse*" occurs, the variational posterior $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})$ will be independent of the data $\boldsymbol{x}$, resulting in $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x}) \approx p_\theta(\boldsymbol{z}_{>k})$, and the reconstructions of in-distribution and OOD samples, which are generated from the partial generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k})$, will be almost the same, such as the visualization examples shown in Fig. 2 by setting $k = 3$ or $k = 4$. In that case, for each in-distribution sample, the latent variables $\boldsymbol{z}_{\leq k}$ generated $p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$ will be clearly distinct from those drawn from $q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$, and similar conclusions can also be achieved by OOD samples, which will reduce the variance of $\mathcal{LLR}^{>k}$ scores between in-distribution and OOD samples and further bring troubles for OOD detection with $\mathcal{LLR}^{>k}$.

3.3 Informative Hierarchical VAE to Alleviate "*posterior collapse*"

Recall to the conflict between the ELBO objective and $\mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}_{>k})$ as discussed in Sec. 3.1, which causes "*posterior collapse*" in hierarchical VAEs, there could be two main approaches to alleviate this phenomenon, including: 1) downweight the KL term, like applying a warm-up scheme on it [31], which is the most common heuristic in practice but still cannot essentially address this issue; 2) enhance the connections between the observation and its multi-layer stochastic latent representations in the expected log-likelihood term, like modifying the generative process described by $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ [23].

In this paper, focused on exploring the potential of alleviating "*posterior collapse*" with the second approach, we try to introduce skip-connection-liked structures into expected log-likelihood term to enhance the connections between the data $\boldsymbol{x}$ and the latent variables $\boldsymbol{z} = \boldsymbol{z}_1, ..., \boldsymbol{z}_L$. However, constrained by the layer-by-layer generation process of hierarchical VAE, there remains a great challenge to introduce physical skip connections into the generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, because arbitrarily adding or concatenating the stochastic hidden layers at different semantic levels will hurt the hierarchy of these multi-layer latent representations. We emphasize that the skip connections between the single stochastic layer and multiple deterministic layers [28] cannot be extended for hierarchical VAEs with multiple stochastic hidden layers, but our developed method below can be applied to any existing hierarchical VAE, which is one of the main contributions of this paper.

Generally speaking, moving beyond downweighting $\mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}_{>k})$ included in the KL term or modifying the structure of generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, our main idea is to upweight the mutual information between the data $\boldsymbol{x}$ and the higher-level variables $\boldsymbol{z}_{>k}$, which is denoted as

$$\mathcal{I}_p(\boldsymbol{x}, \boldsymbol{z}_{>k}) = -\mathcal{H}_p(\boldsymbol{x}|\boldsymbol{z}_{>k}) + \mathcal{H}_p(\boldsymbol{x}) = \mathbb{E}_{p(\boldsymbol{x})p_\theta(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k}) - \mathbb{E}_{p(\boldsymbol{x})} \log p(\boldsymbol{x}), \tag{8}$$

in the objective function of hierarchical VAEs. In Eq. (8), the first item can be approximated by $\mathcal{H}_{p,q}(\boldsymbol{x}|\boldsymbol{z}_{>k}) = \mathbb{E}_{p(\boldsymbol{x})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k})$ and the second item $\mathcal{H}_p(\boldsymbol{x})$ is a positive constant, leading to $\mathcal{I}_p(\boldsymbol{x}, \boldsymbol{z}_{>k}) \geq \mathcal{H}_{p,q}(\boldsymbol{x}|\boldsymbol{z}_{>k})$. Thus, targeted at directly maximizing multiple $\mathcal{H}_{p,q}(\boldsymbol{x}|\boldsymbol{z}_{>k})$ in the objective function, we develop an informative loss for training hierarchical VAEs, denoted as

$$\mathcal{L}^{in} = \mathbb{E}_{p(\boldsymbol{x})} \left[\frac{1}{L} \sum\nolimits_{k=0}^{L-1} \mathbb{E}_{q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k}) \right] - \sum\nolimits_{l=1}^{L} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})) \right] \tag{9}$$

where $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k}) = \mathbb{E}_{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})} \left[p_\theta(\boldsymbol{x}|\boldsymbol{z}_{\leq k}) \right]$ describes a partial generative model to reconstruct the observation $\boldsymbol{x}$ taking $\boldsymbol{z}_{>k}$ drawn from the variational inference network $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})$ as input; the weight $1/L$ before each expected term is introduced to keep numerical stability. The developed $\mathcal{L}^{in}$ not only inherits the terms of ELBO in Eq. (1), which helps preserve the original model properties of VAE, but also introduces virtual skip-connection-liked structures with partial generative models to enhance the connections between $\boldsymbol{x}$ and $\boldsymbol{z}_{>k}$, contributing to alleviating "*posterior collapse*".

To avoid directly calculating $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k})$ in practice, inspired by [21, 23], the informative loss in Eq. (9) can be optimized by maximizing its lower bound $\hat{\mathcal{L}}^{in}$, expressed as

$$\begin{aligned} \mathcal{L}^{in} \geq \hat{\mathcal{L}}^{in} =& \mathbb{E}_{p(\boldsymbol{x})} \left[\frac{1}{L} \sum\nolimits_{k=0}^{L-1} \mathbb{E}_{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_{\leq k}) \right] \right] \\ & - \mathbb{E}_{p(\boldsymbol{x})} \left[\sum\nolimits_{l=1}^{L} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})) \right], \end{aligned} \tag{10}$$

where the expected log-likelihood term is the summation of L components and each component denoted as $\mathcal{LL}^{>k} = \mathbb{E}_{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_{\leq k})]$ can be obtained by replacing the inference network $q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$ in the original log-likelihood term as described in Eq. (3) with the generative model $p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$. Intuitively, for each expected log-likelihood term $\mathcal{LL}^{>k}$, after sampling the top variables $\boldsymbol{z}_{>k}$ from the variational posterior $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})$, these variables $\boldsymbol{z}_{>k}$ will be forced to reconstruct the observation $\boldsymbol{x}$ with the partial generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k})$, which builds the straightforward connections between $\boldsymbol{x}$ and $\boldsymbol{z}_{>k}$ to alleviate "*posterior collapse*".

We refer to the hierarchical VAE trained with the lower bound of informative loss in Eq. (10) as informative hierarchical VAE. We note that this method is applicable for hierarchical VAEs with either top-down or bottom-up inference network to explicitly utilize the generative hierarchy of the multi-layered stochastic variables during training, and can be flexibly extended in future works.

3.4 Adaptive Likelihood Ratio for OOD detection

Besides "*posterior collapse*", an inappropriate choice of k in the likelihood-ratio score function, denoted as $\mathcal{LLR}^{>k}$ in Eq. (6), will also bring negative impact on the performance of applying hierarchical VAEs for OOD detection. Recall to the visualization examples in Fig. 2, when setting $k = 0$ or $k = 1$, the reconstruction quality of either in-distribution or OOD samples is surprisingly high, leading to the $\mathcal{LLR}^{>k}$ scores of both in-distribution and OOD samples are relatively small and further making it difficult to distinguish whether the data sample is OOD or not. Moving beyond cherry picking the hyperparameter k on testing OOD samples, which is unreasonable for unsupervised OOD detection, we develop a novel adaptive likelihood-ratio score function $\mathcal{LLR}^{ada}$, described as

$$\mathcal{LLR}^{ada} = \sum\nolimits_{k=0}^{L-1} \frac{\mathcal{R}(\boldsymbol{x}, \boldsymbol{z}_{>k-1})}{\mathcal{R}(\boldsymbol{x}, \boldsymbol{z}_{>k})}(\mathcal{LLR}^{>k} - \mathcal{LLR}^{>k-1}), \tag{11}$$

where $\mathcal{R}(\boldsymbol{x}, \boldsymbol{z}_{>k})$ is designed to measure the relevance between the data sample $\boldsymbol{x}$ and its latent variables $\boldsymbol{z}_{>k}$ sampled from the variational posterior $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})$, specifically defining $\mathcal{LLR}^{>-1} := 0$ and $\mathcal{R}(\boldsymbol{x}, \boldsymbol{z}_{>-1}) := \mathcal{R}(\boldsymbol{x}, \boldsymbol{z}_{>0})$. More specifically, there are many choices for the definition of $\mathcal{R}(\boldsymbol{x}, \boldsymbol{z}_{>k})$, but in the following experiments, we only use the log-likelihood score for brevity, by specifically defining $\mathcal{R}(\boldsymbol{x}, \boldsymbol{z}_{>k}) := 1/\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k})$.

The intuition of designing $\mathcal{LLR}^{ada}$ is to move beyond the choose of k but adaptively enhance the importance of some discriminative terms, like $\mathcal{LLR}^{>2}$, in the overall score function for OOD detection. With $\mathcal{R}(\boldsymbol{x}, \boldsymbol{z}_{>k})$ to measure the relevance between $\boldsymbol{x}$ and $\boldsymbol{z}_{>k}$, we find that the adaptive weight $\frac{\mathcal{R}(\boldsymbol{x}, \boldsymbol{z}_{>k-1})}{\mathcal{R}(\boldsymbol{x}, \boldsymbol{z}_{>k})}$ will be relatively large when the data information drop rapidly at the current hidden layer, like $k = 2$ in Fig. 2, which can be naturally applied as the importance weights to enlarge the gap between the metric scores of in-distribution and OOD samples. Compared to the previous score functions for OOD detection [21, 26], the developed $\mathcal{LLR}^{ada}$ in Eq. (11) owns less hyperparameters to be tuned, making its performance more stable on various benchmarks. More discussions about $\mathcal{LLR}^{ada}$ can be found in Appendix C.

4 Experiments

4.1 Experimental setup

Datasets: Following [1, 21, 26], we compare our method with previous works on two dataset pairs, including: FashionMNIST [32] (in) / MNIST [33] (out) and CIFAR10 [34] (in) / SVHN [35] (out), where the suffix "in" and "out" denote the in-distribution dataset and OOD dataset, respectively. To better evaluate the generalization ability of these methods, we introduce additional OOD datasets: for FashionMNIST/MNIST pair, we add KMNIST [36], notMNIST [37], Omniglot [38] and SmallNORB [39] datasets; for CIFAR10/SVHN pair, we add CelebA [40], Places365 [41], Flower102 [42] and LFWPeople [43] datasets. More details about datasets can be found in Appendix D.

Evaluation and Metrics: We follow the evaluation procedure in Havtorn et al. [21], where all methods are trained on the training split of the in-distribution dataset, and their OOD detection performance is evaluated on both the testing split of the in-distribution dataset and OOD dataset. Following previous works' evaluation approaches [4, 5, 22], we adopt two popular threshold-independent evaluation metrics, including Area Under the Receiver Operator Characteristic (AUROC↑) and Area

Table 1: The comparisons of the 5-layer informative hierarchical VAE with $\mathcal{LLR}^{ada}$ and other OOD detection methods. The state-of-the-art results achieved by the methods of the category "Not ensembles" of "Unsupervised" have been bold.

FashinMNIST(in)/MNIST(out)						CIFAR10(in)/SVHN(out)					
Labels		**Prior**		**Unsupervised**		**Labels**		**Prior**		**Unsupervised**	
Method	AUROC↑	Mehod	AUROC↑	Method	AUROC↑	Method	AUROC↑	Mehod	AUROC↑	Method	AUROC↑
CP [4]	73.4	LR(PC) [1]	99.4	*-Ensembles*		MD [8]	99.7	LR(PC) [1]	93.0	*-Ensembles*	
CP(Ent) [4]	74.6	LR(BC) [1]	45.5	WAIC(5VAE) [18]	76.6	LMD [44]	27.9	LR(VAE) [1]	26.5	WAIC(5Glow) [18]	99.0
ODIN [7]	75.2	CP(OOD) [1]	87.7	WAIC(5PC) [18]	22.1	EN [10]	98.9	OE [22]	98.4	WAIC(5PC) [18]	62.8
VIB [5]	94.1	CP(Cal) [1]	90.4	*-Not Ensembles*		iDE [12]	95.7	IC(Glow) [27]	95.0	*-Not Ensembles*	
MD(CNN) [8]	94.2	IC(Glow) [27]	99.8	LRe [26]	**98.8**			IC(PC++) [27]	92.9	LRe [26]	87.5
MD(DN) [8]	98.6	IC(PC++) [27]	96.7	HVK [21]	98.4			IC(HVAE) [27]	83.3	HVK [21]	89.1
DE [4]	85.7			$\mathcal{LLR}^{ada}$(Ours)	98.0					$\mathcal{LLR}^{ada}$(Ours)	**94.2**

Table 2: The comparisons of the 3-layer informative hierarchical VAEs with various score functions and other unsupervised OOD detection methods.

FashinMNIST(in)/MNIST(out)				CIFAR10(in)/SVHN(out)			
Method	AUROC↑	AUPRC↑	FPR80↓	Method	AUROC↑	AUPRC↑	FPR80↓
WAIC(5PC) [18]	22.1	40.1	91.1	WAIC(5PC) [18]	62.8	61.6	65.7
HVK [21]	**98.4**	**98.4**	1.3	HVK [21]	89.1	87.5	17.2
-Ours:				***-Ours:***			
$\mathcal{L}$	55.3	51.8	67.9	$\mathcal{L}$	49.9	51.0	79.4
$\mathcal{LLR}^{>1}$	97.5	97.0	2.8	$\mathcal{LLR}^{>1}$	68.4	71.3	61.8
$\mathcal{LLR}^{>2}$	97.4	97.7	**1.2**	$\mathcal{LLR}^{>2}$	**93.0**	**92.5**	**10.8**
$\mathcal{LLR}^{ada}$	97.0	97.6	**0.9**	$\mathcal{LLR}^{ada}$	**92.6**	**91.8**	**11.1**

Under the Precision Recall Curve (AUPRC↑), and another metric False Positive Rate at 80% true positive rate (FPR80↓), where the arrow indicates the direction of improvement.

Baselines: The comparisons in our experiments mainly include two aspects: **i**) the comparisons with previous OOD detection methods to see whether our method can achieve competitive performance; **ii**) the comparisons with several hierarchical VAEs to see whether the new training objective of our method can lead to better performance. For the comparisons in **i**, the baselines can be divided into three categories: "**Labels**": methods using in-distribution data labels [4, 5, 7, 8, 44, 45]; "**Prior**": methods using the prior knowledge collected from OOD data [1, 22, 27]; and "**Unsupervised**": methods without any OOD-specific assumptions [18, 21, 26]. For the comparisons in **ii**, we compare our method with a normal bottom-up inference hierarchical VAE (HVAE) [16], which is also the backbone of our method, and its two major variants: a top-down inference hierarchical VAE named Ladder VAE (LVAE) [24] and a bidirectional inference hierarchical VAE (BIVA) [23]. More details of these baselines and the categories they belong to can be found in Appendix E.

Implementation Details: For the comparisons on FashionMNIST(in)/MNIST(out), we set the network structure of hierarchical VAEs as [16, 8, 4] and [32, 24, 16, 8, 4] from shallow to deep, respectively. For CIFAR10(in)/SVHN(out), we set the network structure as [128, 64, 32] and [128, 64, 32, 28, 24], respectively. For optimization, we adopt the same Adam optimizer [46] with a learning rate of 3e-4. We train all models in comparison by setting the batch size as 128 and the max epoch as 1000. All experiments are performed on a PC with an NVIDIA RTX 3090 GPU and the our code is implemented with PyTorch [47]. More implementation details can be found in Appendix F.

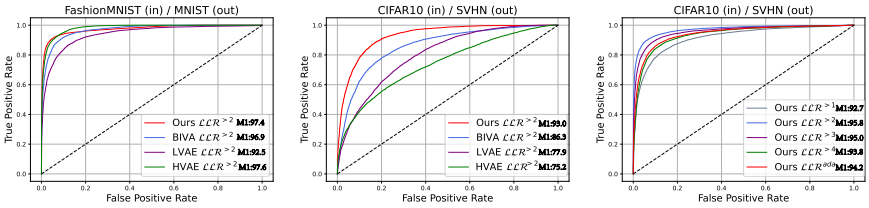
4.2 Quantitative Comparisons

Overall Comparisons: Following the experimental settings in Sec. 4.1, we exhibit the experimental results in in Tab. 1. From the results, we can find that our method with the $\mathcal{LLR}^{ada}$ score function is comparable with those non-ensemble completely unsupervised methods in FashionMNIST/MNIST, and significantly outperform them in CIFAR10/SVHN. We emphasize that, without utilizing the labels of in-distribution samples [4, 5, 7, 8, 44, 45] or the prior knowledge collected from OOD samples [1, 22, 27], our method can still achieve competitive performance with these methods.

Effectiveness of $\mathcal{LLR}^{ada}$: Focused on the comparison between $\mathcal{LLR}^{>k}$ and $\mathcal{LLR}^{ada}$ exhibited in Tab. 2 and Fig. 3(c), when the performance of $\mathcal{LLR}^{>k}$ is sensitive to the selection of k, we can find that the performance of $\mathcal{LLR}^{ada}$ can approach the best performance achieved by $\mathcal{LLR}^{>k}$ with the optimal k, as shown in the right part of Tab. 2 (CIFAR/SVHN). Furthermore, when the performance of $\mathcal{LLR}^{>k}$ is stable, the developed $\mathcal{LLR}^{ada}$ can still achieve comparable OOD detection performance, as shown in the left part of Tab. 2 (FashionMNIST/MNIST) and Fig. 3(c). The experimental results above demonstrate the adaptability of our developed $\mathcal{LLR}^{ada}$.

Table 3: The comparisons of the OOD detection performance of various 3-layer hierarchical VAEs with the same $\mathcal{LLR}^{ada}$ score function. “M1” refers to the metric AUROC↑, “M2” refers to the metric AUPRC↑, and “M3” refers to the metric "FPR80↓".

	Trained on FashionMNIST.												**Trained on CIFAR10.**											
OOD	KMNIST			Omniglot			notMNIST			SmallNORB			CelebA			Places365			Flower102			LFWPeople		
Model	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
HVAE [21]	86.4	89.7	29.0	99.8	99.9	**0.00**	85.3	88.1	29.2	**100**	**100**	**0.00**	39.8	44.7	90.0	40.1	46.6	94.0	45.2	51.7	92.0	42.5	48.2	92.5
LVAE [24]	85.9	87.7	24.8	93.1	96.0	0.5	94.0	93.6	6.0	97.3	97.7	0.8	53.1	54.2	80.5	56.2	53.7	74.4	56.5	52.3	70.9	63.0	65.8	61.6
BIVA [23]	86.5	87.0	27.0	**100**	**100**	**0.00**	96.4	97.0	2.4	98.7	98.6	1.9	70.5	67.8	53.2	60.1	**63.0**	74.6	61.9	69.2	84.4	75.2	74.0	44.6
Ours	**95.0**	**95.1**	**7.1**	**100**	**100**	**0.00**	**99.7**	**99.8**	**0.00**	**100**	**100**	0.1	**72.1**	**70.5**	**49.0**	**63.3**	62.1	**62.6**	**63.4**	**70.1**	**71.2**	**83.0**	**83.4**	**29.0**

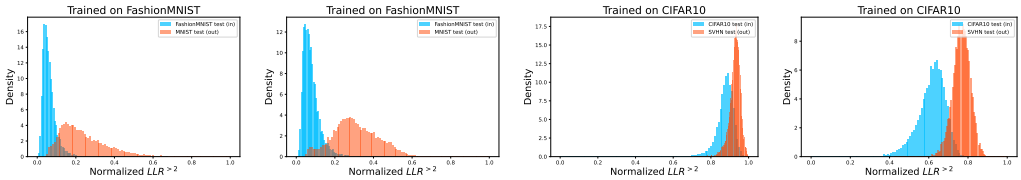


(a) (b) (c)

Figure 3: Plots of the ROC curves on the OOD detection performance of various hierarchical VAEs. (a)∼(b): ROC curves on FashionMNIST (in) / MNIST (out) and CIFAR10 (in) / SVHN (out) with the same score function $\mathcal{LLR}^{>2}$. (c): ROC curves on CIFAR10 (in) / SVHN (out) with various score functions ($\mathcal{LLR}^{>k}$ and $\mathcal{LLR}^{ada}$). "M1" denotes for the AUROC value.

Effectiveness of Informative Hierarchical VAE: Taking the same $\mathcal{LLR}^{ada}$ as the score function, we compare the performance of informative hierarchical VAE with other hierarchical VAEs in Tab. 3. From the results, we can find that our model outperforms others, indicating the effectiveness of alleviating “*posterior collapse*” with informative hierarchical VAE for OOD detection.

To evaluate the degree of “*posterior collapse*”, we use the ROC curves in Fig. 3 and the empirical densities in Fig. 4 to compare the performance of $\mathcal{LLR}^{>2}$ scores based on the top-level latent variables $\boldsymbol{z}_2$ of these VAEs. The ROC results in Fig. 3 demonstrate the superiority of the informative hierarchical VAE, confirming that our model can provide more expressive higher-level latent representations for OOD detection. The empirical densities in Fig. 4 show that $\boldsymbol{z}_2$ learned by our model on the CIFAR10 (in)/SVHN (out) pairs owns better separability than those learned by BIVA.

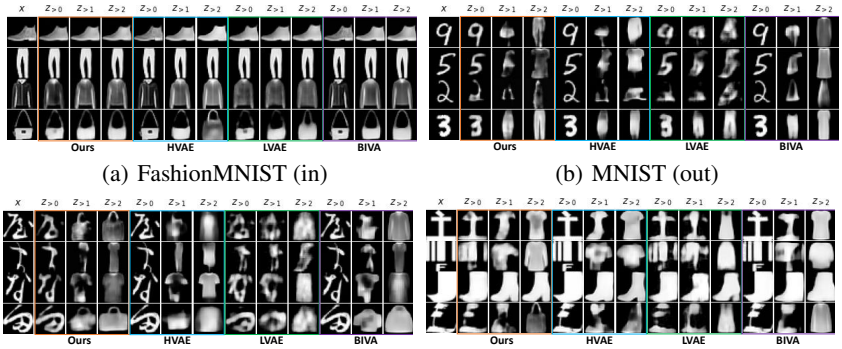


(a) BIVA (b) Ours (c) BIVA (d) Ours

Figure 4: Empirical densities under the score function $\mathcal{LLR}^{>2}$ on FashionMNIST (in)/MNIST (out) and CIFAR10 (in)/SVHN (out) dataset pairs.

4.3 Qualitative Analysis

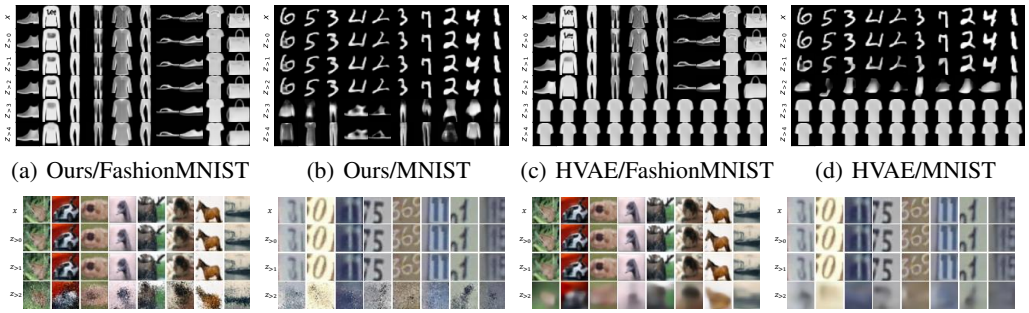
Meaningful Semantic Space Learned by Informative Hierarchical VAE: Following the same procedure as Fig. 2, we visualize more reconstructed samples on various benchmarks in Fig. 5. From the visualized results, for both in-distribution and OOD samples, we can find that the quality of the reconstructions generated from $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>0})$ is surprisedly high, indicating that these reconstructions are almost dominated by the low-level features, which potentially explains the previous problematic phenomenon that these methods based on single-layer likelihood will assign higher likelihood scores for OOD samples and fail on OOD detection. Focused on the reconstructions generated by $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>2})$, we can find that the developed informative hierarchical VAE can provide more realistic and clear reconstructed samples for both in-distribution and OOD inputs, indicating that our model can learn a more meaningful high-level latent semantic space than other models. Furthermore, based on providing higher-quality reconstruction for OOD samples, the gap between the metric scores of



(a) FashionMNIST (in) (b) MNIST (out)

(c) KMNIST (out) (d) notMNIST (out)

Figure 5: Comparisons of reconstructions with 3-layer hierarchical VAEs trained on FashionMNIST, where the leftmost column in each subfigure is the input $\boldsymbol{x}$ and the column noted with $\boldsymbol{z}_{>k}$ means the generation from the partial generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k})$ with $k \in \{0, 1, 2\}$.



(a) Ours/FashionMNIST (b) Ours/MNIST (c) HVAE/FashionMNIST (d) HVAE/MNIST

(e) Ours/CIFAR10 (in) (f) Ours/SVHN (out) (g) HVAE/CIFAR10 (in) (h) HVAE/SVHN (out)

Figure 6: Comparisons of the degree of "*posterior collapse*" in 5-layer hierarchical VAEs. For each subfigure, the 1st row is the input image $\boldsymbol{x}$ and the ith row is generated from $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k})$.

in-distribution and OOD samples in informative hierarchical VAE tend to be larger than other models, leading to a better OOD performance as shown in Tab. 3.

Alleviating "*Posterior Collapse*" with Informative Hierarchical VAE: As discussed in Sec. 3.1, "*posterior collapse*" will cause the higher-level latent variables to become uninformative. Considering the developed informative hierarchical VAE shares the same network structure with the basic HVAE, in this part, we focus on evaluating whether our model can alleviate "*posterior collapse*" in higher layers when the network depth becomes deeper. As shown in Fig. 6, for both in-distribution data and OOD data samples, the overall quality of the reconstructions generated by our model is significantly higher than those generated by the basic HVAE. Specifically, for FashionMNIST (in)/MNIST (out), the reconstructions generated by $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>3})$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>4})$ of HVAE are almost the same, indicating that its posterior described by $q_\phi(\boldsymbol{z}_{>3}|\boldsymbol{x})$ or $q_\phi(\boldsymbol{z}_{>4}|\boldsymbol{x})$ tend to collapse to a prior distribution about T-shirts. On the contrary, the reconstructions generated by the our model are still realistic, where the reconstructions of MINST (out) samples generated by $q_\phi(\boldsymbol{z}_{>3}|\boldsymbol{x})$ or $q_\phi(\boldsymbol{z}_{>4}|\boldsymbol{x})$ preserve the semantic structural information learned form FashionMNIST (in), explaining the underlying reason why our model can achieve better OOD detection performance. Similar conclusions can be achieved by the experimental results on CIFAR10 (in)/SVHN (out) as shown in Fig. 6.

5 Conclusion

In this paper, after presenting a thorough analysis of "*posterior collapse*", we develop a novel informative hierarchical VAE to extract more expressive hierarchical latent representations by alleviating "*posterior collapse*". Then we theoretically explain why "*posterior collapse*" will limit the performance of existing hierarchical VAEs, and develop a novel Adaptive Likelihood Ratio score function for unsupervised OOD detection. Experiments demonstrate the effectiveness of our method, whose main thought can be borrowed other hierarchical VAEs to improving their performance on downstream tasks relied on the hierarchy of latent representations.

References

[1] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[3] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[4] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[5] Alexander A Alemi, Ian Fischer, and Joshua V Dillon. Uncertainty in the variational information bottleneck. *arXiv preprint arXiv:1807.00906*, 2018.

[6] Zhuo Huang, Chao Xue, Bo Han, Jian Yang, and Chen Gong. Universal semi-supervised learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[8] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[9] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized ODIN: detecting out-of-distribution image without learning from out-of-distribution data. In *Conference on Computer Vision and Pattern Recognition*, pages 10948–10957, 2020.

[10] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.

[11] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, pages 144–157, 2021.

[12] Ramneet Kaur, Susmit Jha, Anirban Roy, Sangdon Park, Edgar Dobriban, Oleg Sokolsky, and Insup Lee. idecode: In-distribution equivariance for conformal out-of-distribution detection. *arXiv preprint arXiv:2201.02331*, 2022.

[13] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[14] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

[15] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[17] 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. 2016.

[18] Hyunsun Choi, Eric Jang, and Alexander A Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

[19] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

[20] Alireza Shafaei, Mark Schmidt, and James J Little. Does your model know the digit 6 is not a cat? a less biased evaluation of" outlier" detectors. 2018.

[21] Jakob D Drachmann Havtorn, Jes Frellsen, Soren Hauberg, and Lars Maaløe. Hierarchical vaes know what they don't know. In *International Conference on Machine Learning*, pages 4117–4128. PMLR, 2021.

[22] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[23] Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. Biva: A very deep hierarchy of latent variables for generative modeling. *Advances in neural information processing systems*, 32, 2019.

[24] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016.

[25] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020.

[26] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *Advances in neural information processing systems*, 33: 20685–20696, 2020.

[27] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*, 2019.

[28] Adji B Dieng, Yoon Kim, Alexander M Rush, and David M Blei. Avoiding latent variable collapse with generative skip models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2397–2405. PMLR, 2019.

[29] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[30] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

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

[32] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[33] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[34] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[35] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[36] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

[37] Yaroslav Bulatov. notMNIST dataset. `http://yaroslavvb.blogspot.com/2011/09/notmnist-dataset.html`.

[38] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[39] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–104. IEEE, 2004.

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

[41] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[42] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[43] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[44] Saikiran Bulusu, Bhavya Kailkhura, Bo Li, Pramod K Varshney, and Dawn Song. Anomalous example detection in deep learning: A survey. *IEEE Access*, 8:132330–132347, 2020.

[45] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[46] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[48] Yuhta Takida, Wei-Hsiang Liao, Toshimitsu Uesaka, Shusuke Takahashi, and Yuki Mitsufuji. Preventing posterior collapse induced by oversmoothing in gaussian vae. *arXiv preprint arXiv:2102.08663*, 2021.

[49] Dihong Jiang, Sun Sun, and Yaoliang Yu. Revisiting flow generative models for out-of-distribution detection. In *International Conference on Learning Representations*, 2021.

[50] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

[51] Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Abstract
   (b) Did you describe the limitations of your work? [Yes] See Appendix H
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A and Appendix B
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplemental material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.1 and Appendix F
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Appendix G
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.1 and Appendix F
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See supplemental material
   (b) Did you mention the license of the assets? [Yes] See supplemental material
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See supplemental material
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

Due to the page limitation, we put more technical details here for ease of understanding. We will appreciate it so much if the audience can have a careful reading of the following Appendix.

A More discussions about *"Posterior Collapse"*

Here, we present a detailed reinterpretation for intuitively understanding the conflict of the optimization between the ELBO $\mathcal{L}$ and the mutual information $\mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}_{>k})$. Then, we provide additional experiments to meausre the "posterior collapse" in higher layer of the latent variables $\boldsymbol{z}_{>k}$.

A.1 Detailed Derivations for Section 3.1

First, we present a detailed formulation of the aggregated posterior $q_\phi(\boldsymbol{z}_{>k})$ as

$$q_\phi(\boldsymbol{z}_{>k}) = \mathbb{E}_{p(\boldsymbol{x})} q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x}). \tag{12}$$

Under the setting of top-down inference structure, with the aggregated posterior $q_\phi(\boldsymbol{z}_{>k})$, where $q_\phi(\boldsymbol{z}_L|\boldsymbol{z}_{L+1}) := q_\phi(\boldsymbol{z}_L|\boldsymbol{x})$, and $p_\theta(\boldsymbol{z}_L|\boldsymbol{z}_{L+1}) := p_\theta(\boldsymbol{z}_L)$, the KL term in Eq. (3) can be rewritten as follows:

$$\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{L} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + \mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=k+1}^{L} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + \mathbb{E}_{p(\boldsymbol{x})}[\sum_{l=k+1}^{L} \mathbb{E}_{q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})} \log \frac{q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})}{p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})}] \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + \mathbb{E}_{p(\boldsymbol{x})}[\sum_{l=k+1}^{L} \mathbb{E}_{q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log \frac{q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})}{p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})}] \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + \mathbb{E}_{p(\boldsymbol{x})}[\mathbb{E}_{q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log \prod_{l=k+1}^{L} \frac{q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})}{p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})}] \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + \mathbb{E}_{p(\boldsymbol{x})}\left[D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})||p_\theta(\boldsymbol{z}_{>k}))\right] \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + \mathbb{E}_{p(\boldsymbol{x})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log \frac{q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}_{>k})} \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + \mathbb{E}_{p(\boldsymbol{x})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log q_\phi(\boldsymbol{z}_{>k}|x) \\
&- \mathbb{E}_{p(\boldsymbol{x})q(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log p_\theta(\boldsymbol{z}_{>k}) \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + \mathbb{E}_{p(\boldsymbol{x})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log q_\phi(\boldsymbol{z}_{>k}|x) \\
&- \mathbb{E}_{q_\phi(\boldsymbol{z}_{>k})} \log q_\theta(\boldsymbol{z}_{>k}) + \mathbb{E}_{q_\phi(\boldsymbol{z}_{>k})} \log q_\theta(\boldsymbol{z}_{>k}) - \mathbb{E}_{p(\boldsymbol{x})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log p_\theta(\boldsymbol{z}_{>k}) \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + E_{p(\boldsymbol{x})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log q_\phi(\boldsymbol{z}_{>k}|x) \\
&- \mathbb{E}_{q_\phi(\boldsymbol{z}_{>k})} \log q_\theta(\boldsymbol{z}_{>k}) + \mathbb{E}_{q_\phi(\boldsymbol{z}_{>k})} \log q_\theta(\boldsymbol{z}_{>k}) - \mathbb{E}_{q_\phi(\boldsymbol{z}_{>k})} \log p_\theta(\boldsymbol{z}_{>k}) \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + \mathbb{E}_{p(\boldsymbol{x})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x}) \\
&- \mathbb{E}_{q_\phi(\boldsymbol{z}_{>K})} \log q_\phi(\boldsymbol{z}_{>k}) + D_{KL}(q_\phi(\boldsymbol{z}_{>k})||p_\theta(\boldsymbol{z}_{>k})) \\
=&\mathbb{E}_{p(\boldsymbol{x})}\left[\sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] + \mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}_{>k}) + D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_{>k})||p_\theta(\boldsymbol{z}_{>k})).
\end{aligned} \tag{13}$$

Thus, the ELBO in Eq. (3) can be rewritten as

$$\begin{aligned}\mathcal{L} =& \mathbb{E}_{p(\boldsymbol{x})}\left[\mathbb{E}_{q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_1)\right] - \sum\nolimits_{l=1}^{k} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{l+1})||p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{l+1}))\right] \\ & - \mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}_{>k}) - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_{>k})||p_\theta(\boldsymbol{z}_{>k}).\end{aligned} \tag{14}$$

Therefore, maximizing the ELBO will be opposite to maximizing the $\mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}_{>k})$, which leads the higher-layers posterior latent variables $\boldsymbol{z}_{>k}$ to be independent of the input data $\boldsymbol{x}$ and collapse to the uninformative $p_\theta(\boldsymbol{z}_{>k})$.

A.2 Quantitative experiments on "Posterior Collapse"

Since the "posterior collapse" of the in-distribution data would lead to a larger Likelihood Ratio $\mathcal{LLR}^{>k}$ in high layers, which is harmful for OOD detection, we add an additional experiment here to testify the "posterior collapse" with the metric *average bits per dim* as shown in Table 4. The $\mathcal{L}_{\boldsymbol{x}}^{>L-1}$ is the ELBO for the partial generative model, which could be used to evaluate the reconstruction quality of $\boldsymbol{z}_L$, and more details can be found in Eq. (15).

Table 4: The average bits per dim and the OOD detection performance of four hierarchical VAEs. The average bits per dim is calculated in the testing split of the in-distribution dataset and the OOD detection performance is tested with $\mathcal{LLR}^{>L-1}$, where $L = 5$ in the FashionMNIST/MNIST pair and $L = 3$ in the CIFAR10/SVHN pair. Note that, $\mathcal{LLR}^{>L-1} = \mathcal{L}_{\boldsymbol{x}} - \mathcal{L}_{\boldsymbol{x}}^{>L-1}$.

FashionMNIST(in)/MNIST(out)					
	Avg. bits per dim		OOD Detection		
Metthod	$\mathcal{L}_{\boldsymbol{x}}$	$\mathcal{L}_{\boldsymbol{x}}^{>4}$	AUROC↑	AUPRC↑	FPR80↓
HVAE(5)	2.67	11.0	33.7	38.7	70.8
LVAE(5)	**2.61**	5.91	64.3	61.5	59.5
BIVA(5)	2.70	11.1	35.3	39.2	69.7
Ours(5)	3.45	**3.54**	**98.2**	**98.3**	**1.5**
CIFAR10(in)/SVHN(out)					
	Avg. bits per dim		OOD Detection		
Method	$\mathcal{L}_{\boldsymbol{x}}$	$\mathcal{L}_{\boldsymbol{x}}^{>2}$	AUROC↑	AUPRC↑	FPR80↓
HVAE(3)	3.82	40.01	74.1	76.4	54.7
LVAE(3)	3.85	14.32	80.1	78.8	36.1
BIVA(3)	**3.49**	20.42	86.1	85.2	22.6
Ours(3)	6.29	**6.40**	**93.0**	**92.5**	**10.8**

As the results shown in Table 4, although the baselines for comparison (HVAE, LVAE, and BIVA) can obtain better reconstruction performance on $\mathcal{L}_{\boldsymbol{x}}$, they still suffer from a large shrink in $\mathcal{L}_{\boldsymbol{x}}^{>L-1}$, which is mainly caused by the "posterior collapse". On the contrary, the developed informative HVAE can acquire stable performance from $\mathcal{L}_{\boldsymbol{x}}$ to $\mathcal{L}_{\boldsymbol{x}}^{>L-1}$, resulting in a smaller Likelihood Ratio $\mathcal{LLR}^{>L-1}$, which illustrates why our method can achieve much better performance on unsupervised OOD detection.

B Derivations for analyzing the $\mathcal{LLR}^{>k}$ in Section 3.2

For ease of understanding of the Eq. (6), we give a detailed derivation below, which is mostly based on the Havtorn et al. [21].

First, we define a looser ELBO for each observation $\boldsymbol{x}$ as below of the partial generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k}) = \mathbb{E}_{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})}\left[p_\theta(\boldsymbol{x}|\boldsymbol{z}_{\leq k})\right]$, which reconstructs the observation $\boldsymbol{x}$ by taking $\boldsymbol{z}_{>k}$ drawn from the variational inference network $q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})$,

$$\mathcal{L}_{\boldsymbol{x}}^{>k} = \log p(\boldsymbol{x}) - D_{KL}(p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})), \tag{15}$$

recall to the common ELBO for each observation $\boldsymbol{x}$ as

$$\mathcal{L}_{\boldsymbol{x}} = \log p(\boldsymbol{x}) - D_{KL}(q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})), \tag{16}$$

then, the $\mathcal{LLR}^{>k}$ is defined as

$$\mathcal{LLR}^{>k} = \mathcal{L}_{\boldsymbol{x}} - \mathcal{L}_{\boldsymbol{x}}^{>k}. \tag{17}$$

Further, the detailed derivation for Eq. (7) is as follows:

$$\begin{aligned}
&\mathcal{LLR}^{>k} \\
=&\mathcal{L}_{\boldsymbol{x}} - \mathcal{L}_{\boldsymbol{x}}^{>k} \\
=&D_{\mathrm{KL}}(p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})) - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})) \\
=&\mathbb{E}_{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log \frac{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} \\
&- \mathbb{E}_{q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})} \log \frac{q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} \\
=&\mathbb{E}_{q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}[\mathbb{E}_{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})} \log \frac{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} \\
&- \mathbb{E}_{q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})} \log \frac{q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})}] \\
=&\mathbb{E}_{q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}[\mathbb{E}_{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})} \log \frac{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k},\boldsymbol{x})p_\theta(\boldsymbol{z}_{>k}|\boldsymbol{x})} \\
&- \mathbb{E}_{q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})} \log \frac{q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k},\boldsymbol{x})p_\theta(\boldsymbol{z}_{>k}|\boldsymbol{x})}] \\
=&\mathbb{E}_{q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}[\mathbb{E}_{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})} \log \frac{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})}{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k},\boldsymbol{x})} + \frac{q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}_{>k}|\boldsymbol{x})} \\
&- \mathbb{E}_{q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})} \log \frac{q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})}{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k},\boldsymbol{x})} - \frac{q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}_{>k}|\boldsymbol{x})}] \\
=&\mathbb{E}_{q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}[\mathbb{E}_{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})} \log \frac{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})}{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k},\boldsymbol{x})} - \mathbb{E}_{q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})} \log \frac{q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})}{p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k},\boldsymbol{x})}] \\
=&\mathbb{E}_{q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}\left[D_{\mathrm{KL}}(p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})||p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k},\boldsymbol{x})) - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})||p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k},\boldsymbol{x}))\right] \\
\approx&\mathbb{E}_{q_\phi(\boldsymbol{z}_{>k}|\boldsymbol{x})}\left[D_{\mathrm{KL}}(p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})||q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k}))\right],
\end{aligned} \tag{18}$$

when the approximated posterior $q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$ is closer to the true posterior $p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k},\boldsymbol{x})$ after training, *i.e.*, $D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})||p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k},\boldsymbol{x})) \approx 0$ and $D_{\mathrm{KL}}(p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})||p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k},\boldsymbol{x})) \approx D_{\mathrm{KL}}(p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})||q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k}))$.

C More discussions about $\mathcal{LLR}^{ada}$

Recall to the visualization exhibited in Fig. 2, when setting $k = 0$ or $k = 1$, we can find that the quality of the reconstructions generated from $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>0})$ is surprisedly high, indicating that the KL-divergence between $p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$ and $q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$ is small for both in-distribution and OOD samples, which makes it problematic for OOD detection with single-layer likelihood $\mathcal{LLR}^{>0}$ or $\mathcal{LLR}^{>1}$; when setting $k = 2$, the KL divergence between $p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$ and $q_\phi(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$ will be relative small for in-distribution samples, but large for OOD samples, which is the main reason for the success of $\mathcal{LLR}^{>2}$; however, when setting $k = 3$ or $k = 4$, the latent variables $\boldsymbol{z}_{\leq k}$ generated $p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$ will be clearly distinct from those drawn from $q_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k},\boldsymbol{x})$ for both in-distribution and OOD samples, resulting in that the performance of OOD detection with $\mathcal{LLR}^{>3}$ or $\mathcal{LLR}^{>4}$ will be worse than $LLR^{>2}$. The reason why OOD detection based on $\mathcal{LLR}^{>3}$ or $\mathcal{LLR}^{>4}$ can outperform OOD detection based on $\mathcal{LLR}^{>0}$ or $\mathcal{LLR}^{>1}$ is that the generative model $p_\theta(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})$ can still learn the generation mechanism of in-distribution samples at higher hidden layers, even when "*posterior collapse*" occurs.

Based on these findings, the intuition of designing $\mathcal{LLR}^{ada}$ is to move beyond the choose of k but enhance the importance of the score over some discriminative layers, like $\mathcal{LLR}^{>2}$, in the overall metric for OOD detection. To achieve these goals, we use $\mathcal{R}(\boldsymbol{x}, \boldsymbol{z}_{>k})$ to measure the relevance between $\boldsymbol{x}$ and $\boldsymbol{z}_{>k}$, and the adaptive weight $\frac{\mathcal{R}(\boldsymbol{x},\boldsymbol{z}_{>k-1})}{\mathcal{R}(\boldsymbol{x},\boldsymbol{z}_{>k})}$ in $\mathcal{LLR}^{ada}$ will be relatively large when the data information drop rapidly at the current hidden layer, like $k = 2$. We admit that "*posterior collapse*" will still hurt the performance of $\mathcal{LLR}^{ada}$, leading to worse performance than $\mathcal{LLR}^{>k}$ with the optimal k, but it can avoid the unreasonable hyper-parameter adjustment based on OOD samples. Moreover, with the informative hierarchical VAE to alleviate "*posterior collapse*", the performance of $\mathcal{LLR}^{ada}$ will be even better than $\mathcal{LLR}^{>2}$ on some datasets.

D More Details of the Datasets

We use additional datasets to evaluate the OOD detection performance.

For FashionMNIST/MNIST pair, we add KMNIST [36], notMNIST [37], Omniglot [38] and SmallNORB [39] datasets. **KMNIST** is a dataset, adapted from Kuzushiji Dataset, as a drop-in replacement for MNIST dataset, which contains 70,000 28×28 grayscale images. **notMNIST** is a dataset made by 547,838 28×28 grayscale images of extracted glyphs from some publicly available fonts with letters A-J taken from different fonts. **Omniglot** contains 32,460 28×28 grayscale images of 1623 different handwritten characters from 50 different alphabets. **SmallNORB** contains 97,200 28×28 grayscale images of 50 toys belonging to 5 generic categories: four-legged animals, human figures, airplanes, trucks, and cars.

For CIFAR10/SVHN pair, we add CelebA [40], Places365 [41], Flower102 [42] and LFWPeople [43] datasets. **CelebA** is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations. **Places365** contains 1.8 million train images from 365 scene categories, 50 images per category in the validation set, and 900 images per category in the testing set. **Flowers102** is an image classification dataset consisting of 102 flower categories, where flowers were chosen to be flowers commonly occurring in the United Kingdom and each class consists of between 40 and 258 images. **LFWPeople** contains more than 13,000 images of faces collected from the web. All these datasets' images would be randomly cropped into the dimension of 32×32×3 before sending into the models.

E Details of the Baselines

Due to the space limitation, we use the abbreviation for each baseline in Table 1. Here, we give a detailed description for each baseline of the three categories:

- **"Labels"** (Methods using in-distribution data labels y): maximum softmax classification probability (CP) method [4] and its variants, denoted as "CP", "CP(OOD)" with OOD as noise class, "CP(Cal)" with calibration on OOD and "CP(Ent)" with entropy of softmax classification probability $p(y|x)$, and Mahalanobis distance (MD) method [8], latent Mahalanobis distance (LMD) method [44], ODIN method [7], VIB method [5] and deep ensembles (DE) method [45] with 20 classifiers;
- **"Prior"** (Methods using prior knowledge assumption of OOD): Likelihood Ratio (LR) method [1] with different backbones, denoted as "LR(PC)" with backbone PixcelCNN, "LR(VAE)" with VAE and "LR(BC)" with binary classifier), Outlier exposure (OE) method [22] and Input complexity (IC) method [27] with different backbones, denoted as "IC(PC)" with backbone PixcelCNN, "IC(Glow)" with backbone Glow and "IC(HVAE)" with backbone HVAE;
- **"Unsupervised"** (Methods with no OOD-specific assumptions): Ensemble methods: WAIC method [18] with different backbones, denoted as "WAIC(5Glow)" with 5 Glow models, "WAIC(5VAE)" with 5 VAE models and "WAIC(5PC)" with 5 PixcelCNN models; Not ensembles methods: Likelihood regret method [26] and its variant "Likelihood regret(z)", Log-Likelihood Ratio ($\mathcal{LLR}$) method [21], which achieved the best performance with "$\mathcal{LLR}^{>1}$(HVAE)" (hyperparameter $k = 1$ and backbone method 3-layer HVAE trained on binarized data) for FashionMNIST(in)/MNIST(out) pair and "$\mathcal{LLR}^{>2}$(BIVA)" (hyperparameter $k = 2$ and backbone method 10-layer BIVA) for CIFAR10(in)/SVHN(out) pair. For this

$\mathcal{LLR}$ method, we denote their best combinations' result in Tab. 1 as "HVK" (Hierarchical VAEs Know what they don't know).

F Details of the Implementation

To make sure the new training objective in Eq. (10) can really lead to an informative hierarchical VAE, we do not use the warm-up trick or the free bits trick. However, we apply the warm-up trick (200 epochs for the Warmup anneal period) and free bits trick (2 nats per $\boldsymbol{z}_i$ and 400 epochs for the free bits period) to the other three hierarchical VAEs (HVAE, LVAE, and BIVA) to empirically alleviate the posterior collapse, which is proven in Havtorn et al. [21] and we follow their procedure to train these three hierarchical VAEs.

Following Havtorn et al. [21], the VAE-based methods' results are computed with 1000 importance samples. However, our method's results only get slight improvement after sampling, considering the computation burden brought by it, we report the results of our method without importance sampling. But we also use the importance sampling for other baseline hierarchical VAEs (HVAE, LVAE, and BIVA).

G Error Bar

We randomly run 5 seeds for our method in experiments and report the error bar as below.

Table 5: Error bar for our method (3 layer) on the performance of OOD detection under the metric AUROC↑, AUPRC↑, and FPR80↓.

Trained on FashionMNIST. Use LLR^{ada}.															
OOD	MNIST			KMNIST			Omniglot			notMNIST			SmallNORB		
Metric	AUROC↑	AUPRC↑	FPR80↓	AUROC↑	AUPRC↑	FPR80↓	AUROC↑	AUPRC↑	FPR80↓	AUROC↑	AUPRC↑	FPR80↓	AUROC↑	AUPRC↑	FPR80↓
Ours	97.0±0.5	97.6±0.6	0.9±0.05	95.0±1.1	95.1±0.9	7.1±0.8	100±0.0	100±0.0	0.00±0.0	99.7±0.1	99.8±0.1	0.00±0.01	100±0.0	100±0.0	0.1±0.0
Trained on CIFAR10. Use $\mathcal{LLR}^{ada}$.															
	SVHN			CelebA			Places365			Flower102			LFWPeople		
Metric	AUROC↑	AUPRC↑	FPR80↓	AUROC↑	AUPRC↑	FPR80↓	AUROC↑	AUPRC↑	FPR80↓	AUROC↑	AUPRC↑	FPR80↓	AUROC↑	AUPRC↑	FPR80↓
Ours	92.6±0.4	91.8±0.5	11.1±0.2	72.1±0.8	70.5±0.7	49.0±0.5	63.3±0.5	62.1±0.4	62.6±1.0	63.4±0.3	70.1±0.4	71.2±1.2	83.0±1.3	83.40.9	29.0±0.6

H Limitation

The developed informative hierarchical VAE can alleviate "posterior collapse" to a certain degree, but still cannot completely avoid the appearing of this phenomenon in VAEs. Then, the developed $\mathcal{LLR}^{ada}$ is not the optimal choice of score function of unsupervised OOD detection, which needs to be investigated in the future work.

I Broader Impact

The developed method in this paper can be straightforwardly applied to real-word applications based on hierarchical VAEs, and alleviate their "posterior collapse" to achieve better model performance. The adaptive score function can be used for purely unsupervised OOD detection, which can boost the reliability and safety of recent machine learning (ML) systems.

J Comparison with other methods to alleviate "posterior collapse"

We provide more comparisons with the methods designed for alleviating the "posterior collapse", including "Warm-up", "OverSmooth", "BIVA", and Our method "Informative". "Warm-up" [24] gradually increases the weight of the KL-divergence term in the learning objective from 0 to 1 and we set the warm-up epochs as 200 with a total training epoch as 1000. "OverSmooth" [48] sets the σ_x as a one-dimensional parameter and updated according to the training objective, where the reconstruction likelihood function is $p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mu_{\boldsymbol{x}}(\boldsymbol{z}), \sigma_{\boldsymbol{x}}^2\boldsymbol{I})$. "BIVA" [23] introduces a

bidirectional inference and generative network architecture, but it changes the original structure of vanilla hierarchical VAE and may hurt the hierarchy of the latent variables. To provide an intuitive comparison of these methods's effect on alleviating "posterior collapse", we introduce the OOD detection performance on a vanilla hierachical VAE, termd as "Vanilla", as an additional baseline. Then, we testify these methods' performance under 3 different score methods for OOD detection, where $\mathcal{L}_x$ represents the evidence lower bound (ELBO) of the VAE, $\mathcal{LLR}^{>L-1}$ represents the gap between partial ELBO of the highest-level latent variables and the the ELBO of lowest-level latent variables, and $\mathcal{LLR}^{ada}$ is an adaptive score that evalutes the whole hierarchy of the latent variables.

As shown in Table 6 and Table 7, the developed informative hierarchical VAE, termed as "Informative", outperforms other methods significantly especially under the score $\mathcal{LLR}^{>L-1}$ and $\mathcal{LLR}^{ada}$.

HVAE: FashionMNIST (in) / MNIST (out))			
Score + Methods	AUROC% ↑	AUPRC% ↑	FPR80% ↓
$\mathcal{L}_x$ + Vanilla [16]	15.3	33.2	96.0
$\mathcal{L}_x$ + Warm-up [24]	26.3	36.2	86.8
$\mathcal{L}_x$ + OverSmooth [48]	45.4	31.2	90.4
$\mathcal{L}_x$ + BIVA [23]	26.1	35.9	91.3
$\mathcal{L}_x$ + Informative(ours)	**49.9**	**51.0**	**79.4**
$\mathcal{LLR}^{>L-1}$ + Vanilla	33.3	38.7	71.3
$\mathcal{LLR}^{>L-1}$ + Warm-up	47.9	44.0	66.4
$\mathcal{LLR}^{>L-1}$ + OverSmooth	81.1	66.5	23.4
$\mathcal{LLR}^{>L-1}$ + BIVA	35.3	39.2	69.7
$\mathcal{LLR}^{>L-1}$ + Informative(ours)	**98.2**	**98.3**	**1.5**
$\mathcal{LLR}^{ada}$ + Vanilla	59.8	50.6	52.9
$\mathcal{LLR}^{ada}$ + Warm-up	68.1	59.0	49.5
$\mathcal{LLR}^{ada}$ + OverSmooth	80.9	66.3	23.5
$\mathcal{LLR}^{ada}$ + BIVA	35.1	38.8	69.0
$\mathcal{LLR}^{ada}$ + Informative(ours)	**98.0**	**97.6**	**1.6**

Table 6: Comparison of the OOD detection performance with 3 score methods for different methods designed for alleviating posterior collapse. All these methods are based on a 5-layer ($L = 5$) HVAE trained on FashionMNIST (in-distribution) for detecting MNIST as OOD data.

HVAE: CIFAR10 (in) / SVHN (out)			
Score + Methods	AUROC% ↑	AUPRC% ↑	FPR80% ↓
$\mathcal{L}_x$ + Vanilla	49.5	51.2	82.6
$\mathcal{L}_x$ + Warm-up	49.5	**51.3**	**70.3**
$\mathcal{L}_x$ + OverSmooth	14.4	32.9	98.3
$\mathcal{L}_x$ + BIVA	13.3	32.6	99.6
$\mathcal{L}_x$ + Informative(ours)	**49.9**	51.0	79.4
$\mathcal{LLR}^{>L-1}$ + Vanilla	63.2	66.7	70.3
$\mathcal{LLR}^{>L-1}$ + Warm-up	75.2	79.1	49.4
$\mathcal{LLR}^{>L-1}$ + OverSmooth	83.4	79.8	25.6
$\mathcal{LLR}^{>L-1}$ + BIVA	86.3	86.5	22.2
$\mathcal{LLR}^{>L-1}$ + Informative(ours)	**93.0**	**92.5**	**10.8**
$\mathcal{LLR}^{ada}$ + Vanilla	63.1	65.2	69.5
$\mathcal{LLR}^{ada}$ + Warm-up	80.1	81.6	37.7
$\mathcal{LLR}^{ada}$ + OverSmooth	83.3	80.7	25.8
$\mathcal{LLR}^{ada}$ + BIVA	86.3	86.6	21.9
$\mathcal{LLR}^{ada}$ + Informative(ours)	**92.6**	**91.8**	**11.1**

Table 7: Comparison of the OOD detection performance with 3 score methods for different methods designed for alleviating posterior collapse. All these methods are based on a 3-layer ($L = 3$) HVAE trained on CIFAR10 (in-distribution) for detecting SVHN as OOD data.

K Comparison with non-VAE methods

Likelihood-based methods are promising to detect the OOD data in an unsupervised manner, since they could give an estimation of the data $\boldsymbol{x}$'s likelihood $p(\boldsymbol{x})$ under the learned distribution p of in-distribution data. Among likelihood-based methods, Flow-based models [13], auto-regressive models [14], and variational auto-encoder (VAE) models are popular for OOD detection tasks. HVK [21] achieves the state-of-the-art of OOD detection under the unsupervised setting. Though our method could outperform HVK, it is still interesting to have an comparison with other two type of models (Flow-based and auto-regressive models) to see whether our methods could also outperform them.

We add 3 non-VAE methods as our baselines for comparison: "Glow", "Flow+Group", and "PixcelCNN++". "Glow" [19] try to do OOD detection with Glow model and find the counterfactual behaviour of assigning high likelihood to OOD data specially in the model family of Flow models. "FLow+Group" [49] is an SOTA flow-based OOD detection method but is not initially designed for our setting, but for group OOD detection, where their model needs to justify whether a batch of sample $\{x_1, x_2, ..., x_n\}, (n > 1)$ is an OOD batch. Note that, the batch size n is set as 5, 10, and 20 in their paper. Luckily, "FLow+Group" also modify their method via data augmentaion to the point OOD detection situation, i.e., $n = 1$, which is the same as our setting, and we directly report their OOD detection results in their Appendix F. "PixcelCNN++" [1] propose to use an auto-regressive model (PixcelCNN++) for OOD detection with the help of additional OOD datasets like NotMNIST dataset, and we report the results of this model under our setting (no additional datasets to help training).

As shown in Table 8 and Table 9, Our method could also outperform other non-VAE methods.

FashionMNIST (in) / MNIST (out)			
Models	AUROC % ↑	AUPRC % ↑	FPR80 % ↓
- **non-VAE methods**			
Glow [19]	6.29	31.5	99.2
Flow+Group [49]	90.0	-	-
PixcelCNN++ [1]	8.90	32.0	99.0
Ours	98.0	97.6	1.60

Table 8: Comparison of 3 non-VAE methods for the unsupervised OOD detection task of detecting MNIST as OOD data with models trained on in-distribution dataset FashionMNIST.

CIFAR10 (ID) / SVHN (OOD)			
Models	AUROC % ↑	AUPRC % ↑	FPR80 % ↓
- **non-VAE methods**			
Glow	7.62	31.7	98.6
Flow+Group	85.0	-	-
PixcelCNN++	9.50	32.0	100.
Ours	92.6	91.8	11.1

Table 9: Comparison of 3 non-VAE methods for the unsupervised OOD detection task of detecting SVHN as OOD data with models trained on in-distribution dataset CIFAR10.

L Comparisons of Score Functions

OOD detection methods need to assign a score for each data sample and detect the OOD data out according to this score. However, since the score method $\mathcal{LLR}^{>k}$ proposed by HVK [21] needs to select the optimal hyperparameter k to achieve the best OOD detection performance, which is unreasonable under the unsupervised setting, we design a novel score function named $\mathcal{LLR}^{ada}$ that does not need to select the k. Thanks for the Reviewer 9EQd's awesome suggestion, there could be another way for automatically selecting the k, where the k is selected based on the largest R-ratio in Eq. (11). To make a further investigation, we provide a comparison between this score function, termed "$\mathcal{LLR}^{opt_k}$", and our $\mathcal{LLR}^{ada}$.

As shown in Table 10 and Table 11, $\mathcal{LLR}^{opt_k}$ could achieve a promising OOD detection performance but still underperform our developed $\mathcal{LLR}^{ada}$.

A more intuitive and numerical analysis about these methods has been provided in Appendix M.

FashionMNIST (in) / MNIST (out)			
Score	AUROC % ↑	AUPRC % ↑	FPR80 % ↓
$\mathcal{L}_x$	55.3	51.8	67.9
$\mathcal{LLR}^{>1}$	97.5	97.0	2.8
$\mathcal{LLR}^{>2}$	97.4	97.7	1.2
$\mathcal{LLR}^{opt_k}$	94.3	94.0	6.13
$\mathcal{LLR}^{ada}$	97.0	97.6	0.9

Table 10: Comparison of different score methods for OOD detection based on a 3-layer HVAE trained with informative loss on in-distribution dataset FashionMNIST.

CIFAR (in) / SVHN (out)			
Score	AUROC % ↑	AUPRC % ↑	FPR80 % ↓
$\mathcal{L}_x$	49.9	51.0	79.4
$\mathcal{LLR}^{>1}$	68.4	71.3	61.8
$\mathcal{LLR}^{>2}$	93.0	92.5	10.8
$\mathcal{LLR}^{opt_k}$	88.4	90.7	11.5
$\mathcal{LLR}^{ada}$	92.6	91.8	11.1

Table 11: Comparison of different score methods for OOD detection based on a 3-layer HVAE trained with informative loss on in-distribution dataset CIFAR10.

M Measure $\mathcal{LLR}^{ada}$ on vanilla HVAE without Informative loss

It would be interesting to see whether the $\mathcal{LLR}^{ada}$ will be effective on a vanilla hierarchical VAE trained without the informative loss. Take a 5-layer HVAE trained on FashionMNIST for example, we give a comparison with different score functions for OOD detection as shown in Table 12.

To better undertand the underlying mechanism of these score methods, we compute the mean negative log-likelihood $-\log p_\theta(x|z_{>k})$ for reconstruction and log-likelihood ratio ($\mathcal{LLR}^{>k}$) for each layer of a 5-layer vanilla HVAE in Table 13.

Since the values $\mathcal{LLR}^{>k}$ of in-distribution data is closer or larger than OOD data, it is not surprising that the $\mathcal{LLR}^{>k}$ cannot achieve promising OOD detection performance.

However, the performance could be significantly improved with the score $\mathcal{LLR}^{opt_k}$ and $\mathcal{LLR}^{ada}$. Specifically, for score $\mathcal{LLR}^{opt_k}$, it would highly possible to assign $\mathcal{LLR}^{>1}$ (1.65×10^3) for in-distribution data and assign $\mathcal{LLR}^{>2}$ (2.90×10^3) for OOD data, which makes it easier to detect OOD data. For score $\mathcal{LLR}^{ada}$, its average score for in-distribution data is $1.65 + \frac{4.87}{3.50} * (2.97 - 1.65) + \frac{7.81}{4.87}(5.90 - 2.97) + \frac{7.81}{7.81} * (5.90 - 5.90) = 8.173(10^3)$, but for OOD data, the average score is $0.74 + \frac{4.32}{2.05}(2.9 - 0.74) + \frac{6.89}{4.32}(5.46 - 2.9) + \frac{6.89}{6.89}(5.46 - 5.46) = 11.869(10^3)$. Since the average score $\mathcal{LLR}^{ada}$ of OOD data is much larger than in-distribution data, our developed $\mathcal{LLR}^{ada}$ could be a more promising score function to achieve better OOD detection performance.

N Numerical analysis to illustrate "posterior collapse" in Fig. 2

It is interesting to investigate the numerical changes when the "posterior collapse" occurs, such as the data samples visualized in Fig. 2. As shown in Table 14, from shallow to deep (Layer 1 to Layer 4), we can find that the KL-divergence of HVAE gradually reduces to 0, which indicates that the posteriors of 4-th and 5-th hidden layers collapse to their priors, resulting in that the high-level latent variables $\boldsymbol{z}_4$ and $\boldsymbol{z}_5$ sampled from the posterior $q_\phi(\boldsymbol{z}_k|\boldsymbol{z}_{>k})$ have no information of input data $\boldsymbol{x}$.

(Vanilla) HVAE: FashionMNIST (in) / MNIST (out)			
Method	AUROC % ↑	AUPRC % ↑	FPR80 % ↓
$\mathcal{L}_x$	15.3	33.3	96.0
$\mathcal{L}_x^{>1}$	14.7	33.2	94.7
$\mathcal{L}_x^{>2}$	37.0	39.6	79.7
$\mathcal{L}_x^{>3}$	23.2	35.6	80.6
$\mathcal{L}_x^{>4}$	23.1	35.6	80.6
$\mathcal{LLR}^{>1}$	18.2	34.0	91.8
$\mathcal{LLR}^{>2}$	45.5	43.3	72.4
$\mathcal{LLR}^{>3}$	33.2	38.6	71.3
$\mathcal{LLR}^{>4}$	33.3	38.6	71.3
$\mathcal{LLR}^{opt_k}$	49.2	45.8	67.5
$\mathcal{LLR}^{ada}$	**59.8**	**50.6**	**52.9**

Table 12: Comparison of the effect of different score methods on Vanilla VAE.

5-Layer Vanilla HVAE				
	FashionMNIST (in)		MNIST (out)	
Layer #	$-\log p(x\|z_{>i})$	$\mathcal{LLR}^{>i}$	$-\log p(x\|z_{>i})$	$\mathcal{LLR}^{>i}$
0	1.59×10^3	N/A	1.15×10^3	N/A
1	3.50×10^3	1.65×10^3	2.05×10^3	7.40×10^2
2	4.87×10^3	2.97×10^3	4.32×10^3	2.90×10^3
3	7.81×10^3	5.90×10^3	6.89×10^3	5.46×10^3
4	7.81×10^3	5.90×10^3	6.89×10^3	5.46×10^3

Table 13: The mean negative log-likelihood for reconstruction and log-likelihood ratio (LLR) for each layer on a 5-layer Vanilla HVAE.

Further, we use t-SNE method to visualize the learned latent data representations in Fig. 7. Note that, different colors in Fig. 7 indicates different classes of data samples. As shown in Fig. 7, we can find that the latent space of HVAE gradually collapses to a non-informative prior distribution, while the learned latent space of our method is still informative at higher layers.

To see more reconstructed data samples of the partial generative models $p(x|z_{>3})$ and $p(x|z_{>4})$,

To intuitively demonstrate that the posterior of our method does not collapse to a single point, we visualize the data samples generated from $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k})$ by taking the latent variables $\boldsymbol{z}_k$ sampled from the posterior $q_\phi(\boldsymbol{z}_k|\boldsymbol{z}_{>k}, \boldsymbol{x})$ as input, where $\boldsymbol{x}$ is a fixed data point. As shown in Fig. 8, the diversity of the generated samples demonstrate that the posterior $q_\phi(\boldsymbol{z}_k|\boldsymbol{z}_{>k}, \boldsymbol{x})$ collapses to its prior distribution $p_\theta(\boldsymbol{z}_k|\boldsymbol{z}_{>k})$ rather than a single point.

KL-divergence in different layers		
Layer index #	HVAE	Ours
1	2.59×10^2	2.28×10^1
2	4.99×10^1	1.42×10^1
3	1.02×10^1	2.74×10^2
4	5.75×10^{-4}	4.09×10^1
5	5.00×10^{-4}	1.97×10^1

Table 14: The KL-divergence for each layer's latent variables.

O Comparisons of Reconstruction and Generation quality

The reconstruction and generation capability are two important model properties of VAE, and here we compare our method with the vanilla HVAE on both these two aspects.

Firstly, we quantitatively evaluate the reconstruction capability of the partial generative models $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k})$ conditioned on latent variables $\boldsymbol{z}_{>k}$ at different hidden layers, and report the average bit

Avg. bits per dim for reconstruction log-likelihood $\log p_\theta(x\|z_{>k})$						
	HVAE			Ours		
Dataset	$z_{>0}$	$z_{>1}$	$z_{>2}$	$z_{>0}$	$z_{>1}$	$z_{>2}$
FashionMNIST	2.953	7.656	9.608	3.025	4.019	4.233
CIFAR	2.181	9.207	18.22	2.193	2.508	5.778

Table 15: Comparison of the reconstruction quality under the metric "Average bits per dim" of the reconstruction log-likelihood for a 3-layer HVAE and our method.

per dim results in Table 15. From the results, we can see that our method can achieve a comparable reconstruction quality with other baselines on $p_\theta(x|z_{>0})$ and significantly outperform them on the reconstruction conditioned on higher-layer latent variables.

For the generation capability, we qualitatively visualize the data samples generated from artial generative models $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{>k})$ conditioned on $\boldsymbol{z}_{>k}$ drawn from the prior distribution $p_\theta(\boldsymbol{z}_L) = \mathcal{N}(0, I)$. From the results shown in in Fig. 9, we can find that the quality and diversity of data samples generated by our model significantly outperform those generated by HVAE, indicating the benefits of alleviating "posterior collapse."

Thus, the aforementioned experimental results demonstrate that our method can perverse the versatility of VAE.

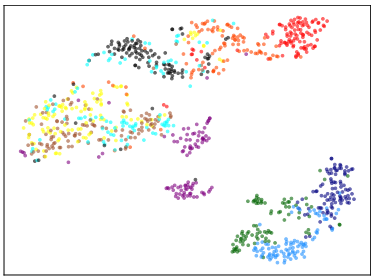
P Comparison on more natural images

To investigate the effectiveness of our method on more natural images, we provide additional comparisons on the other datasets, including LFWPeople [43] (people's faces in the wild), Flower102 [42] (102 types of flowers), Food101 [50] (101 types of food), Places365 [41] (365 scene categories), and Tiny-ImageNet [51] (containing 200 categories of images). With the same score function $\mathcal{LLR}^{ada}$, we train the VAE-based model on each of these datasets, where each image is resized as $32\times32\times3$ before training, and then use it for OOD detection on SVHN dataset. The unified network structure of these 3-layer models are [64, 32, 16] from shallow to deep.

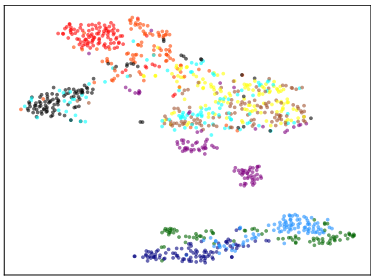
As shown in Table 16, LVAE, BIVA, and our method can generally outperform the vanilla HVAE, while our method could still significantly achieve the best performance on these dataset pairs, which indicates the generality of our method on unsupervised OOD detection.

Trained on ID dataset and Detecting SVHN as OOD data				
ID Dataset	Methods	AUROC % ↑	AUPRC % ↑	FPR80 % ↓
Tiny-ImageNet	HVAE	75.2	75.5	50.7
	LVAE	78.8	75.2	34.9
	BIVA	80.7	76.8	32.5
	Ours	**91.6**	**92.6**	**11.0**
LFWPeople	HVAE	78.4	79.0	46.1
	LVAE	79.0	82.6	39.0
	BIVA	75.6	81.3	57.0
	Ours	**88.5**	**91.8**	**14.5**
Flower102	HVAE	73.2	77.9	60.3
	LVAE	73.3	74.5	48.6
	BIVA	88.2	87.5	21.6
	Ours	**91.6**	**91.7**	**11.4**
Places365	HVAE	54.2	55.7	82.0
	LVAE	58.2	60.5	80.2
	BIVA	72.9	74.3	51.6
	Ours	**87.3**	**89.6**	**19.1**
Food101	HVAE	74.6	74.0	47.7
	LVAE	80.3	84.5	35.6
	BIVA	75.7	79.7	46.2
	Ours	**92.1**	**93.2**	**9.54**

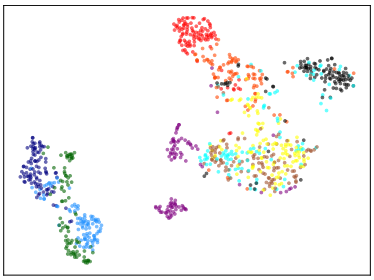
Table 16: Comparison on more dataset pairs. All these methods are trained on in-distribution (ID) dataset and then evaluated on the OOD detection performance with detecting SVHN as OOD dataset.



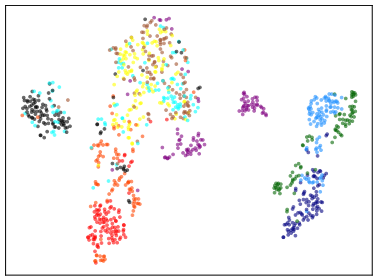
(a) latent space of z_1 of HVAE



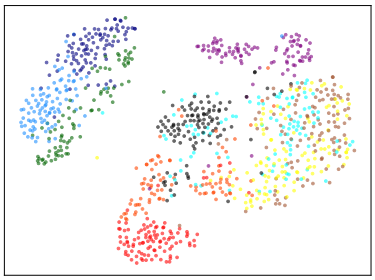
(b) latent space of z_1 of Ours



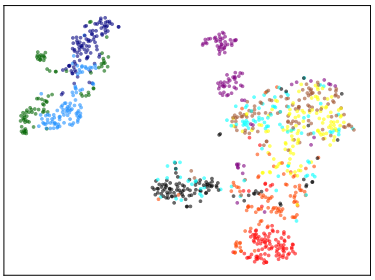
(c) latent space of z_2 of HVAE



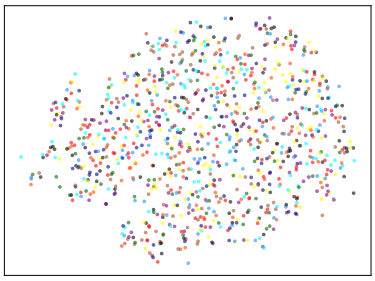
(d) latent space of z_2 of Ours



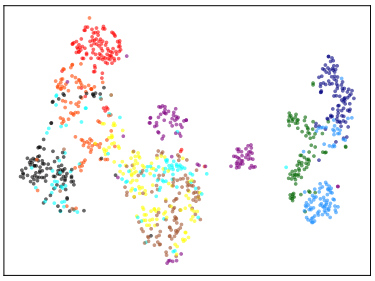
(e) latent space of z_3 of HVAE



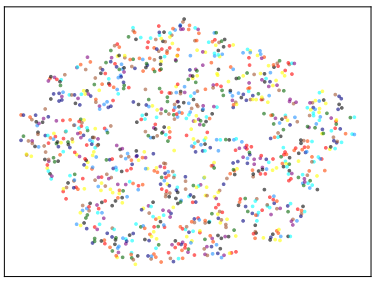
(f) latent space of z_3 of Ours



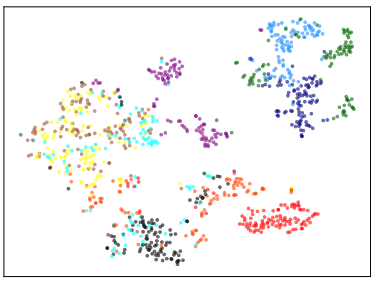
(g) latent space of z_4 of HVAE



(h) latent space of z_4 of Ours

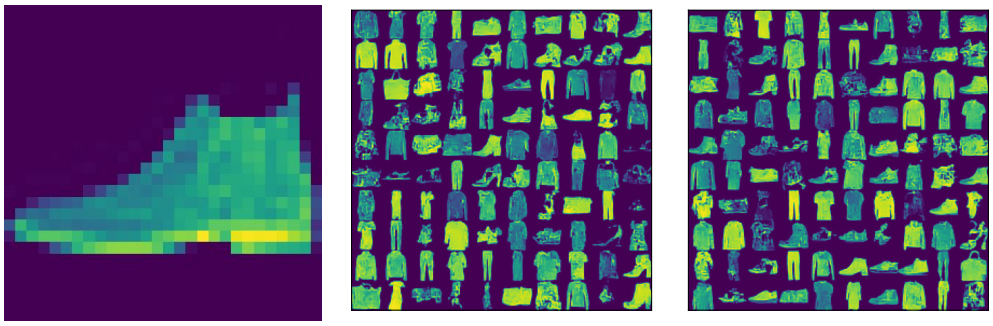


(i) latent space of z_5 of HVAE



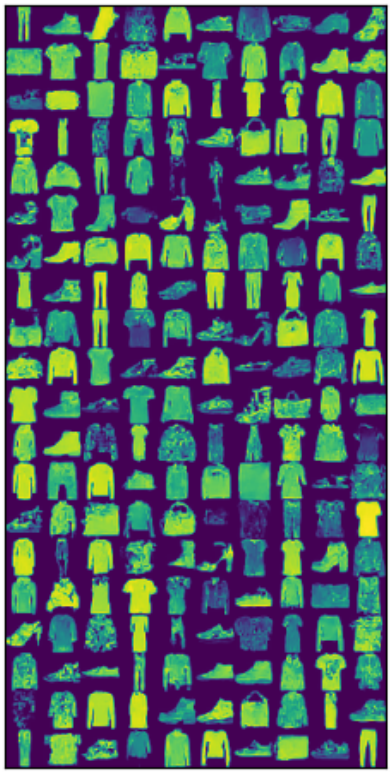
(j) latent space of z_5 of Ours

Figure 7: The learned each layer's latent space of z_i of HVAE and Our method. Different colors indicates that the z is inferred from different classes of input x.



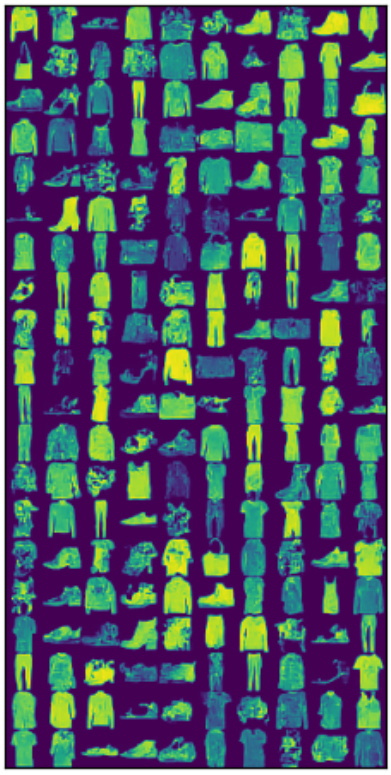
(a) Input data x (b) Partial Generation samples from $p(x|z_{>3})$ (c) Partial Generation samples from $p(x|z_{>4})$

Figure 8: Partial generation samples from $p(x|z_{>3})$ and $p(x|z_{>4})$ of HVAE by taking the latent variables $\boldsymbol{z}_k$ sampled from the posterior $q_\phi(\boldsymbol{z}_k|\boldsymbol{z}_{>k}, \boldsymbol{x})$ as input, where $\boldsymbol{x}$ is a fixed data point.

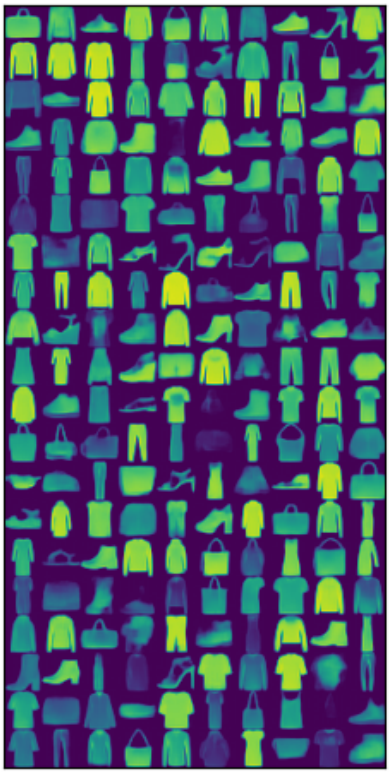


(a) Generated from a 3-layer HVAE

(b) Generated from a 3-layer our model



(c) Generated from a 5-layer HVAE



(d) Generated from a 5-layer our model

Figure 9: Generated samples from Prior distribution.