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. [27].

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 [43], notMNIST [44], Omniglot [45] and SmallNORB [46] 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 [47], Places365 [48], Flower102 [49] and LFWPeople [50] 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 [5] 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 [9], latent Mahalanobis distance (LMD) method [38], ODIN method [8], VIB method [6] and deep ensembles (DE) method [51] 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 [28] and Input complexity (IC) method [33] 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 [24] 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 [32] and its variant "Likelihood regret(z)", Log-Likelihood Ratio ($\mathcal{LLR}$) method [27], 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. [27] and we follow their procedure to train these three hierarchical VAEs.

Following Havtorn et al. [27], 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.

Considering the computational footprint change, we take the vanilla VAE equipped with Likelihood Ratio as the baseline for analysis. For the space complexity, our method doesn't introduce any additional model parameters or memory cost. For the time complexity, compared to the baseline, our method requires additional $L-1$ times computation cost to calculate those expected log-likelihood terms in the loss function during training, specifically $\frac{1}{L}\sum_{k=0}^{L-1}\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{z}_{\leq k}|\boldsymbol{z}_{>k})q_{\boldsymbol{\phi}}(\boldsymbol{z}_{>}k|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_{\leq k})]$, where L denotes the number of layers and will be a relative small number in practice.

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" [30] 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" [54] 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" [29] 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 [20]	15.3	33.2	96.0
$\mathcal{L}_x$ + Warm-up [30]	26.3	36.2	86.8
$\mathcal{L}_x$ + OverSmooth [54]	45.4	31.2	90.4
$\mathcal{L}_x$ + BIVA [29]	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.

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 [17], auto-regressive models [18], and variational auto-encoder (VAE) models are popular for OOD detection tasks. HVK [27] 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" [25] 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" [55] is an SOTA flow-based OOD detection method but is not initially designed for

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.

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 [25]	6.29	31.5	99.2
Flow+Group [55]	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 [27] 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.

(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.

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 $\boldsymbol{x}$.

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.

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.

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.

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.

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 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.

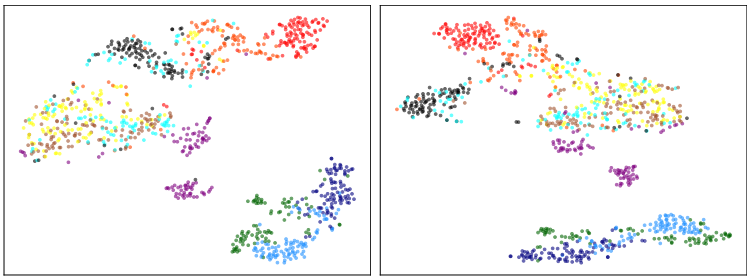
P Comparison on more natural images

To investigate the effectiveness of our method on more natural images, we provide additional comparisons on the other datasets, including LFWPeople [50] (people's faces in the wild), Flower102 [49] (102 types of flowers), Food101 [56] (101 types of food), Places365 [48] (365 scene categories), and Tiny-ImageNet [57] (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 32×32×3 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.

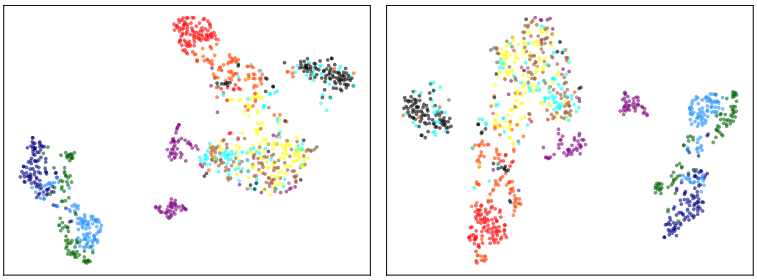
As shown in Table 16, LVAE, BIVA, and our method can generally outperform the vanilla HVAE, while our method could still significantly achieve the best performance on these dataset pairs, which indicates the generality of our method on unsupervised OOD detection.

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**

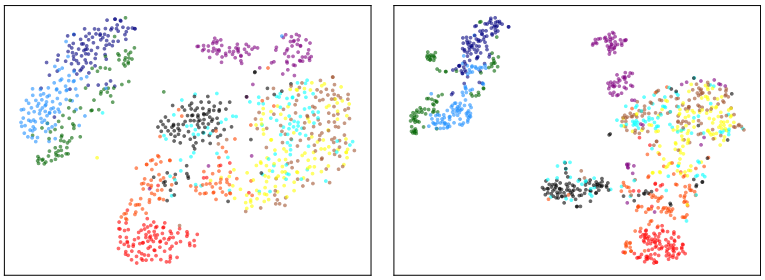
Table 16: Comparison on more dataset pairs. All these methods are trained on in-distribution (ID) dataset and then evaluated on the OOD detection performance with detecting SVHN as OOD dataset.



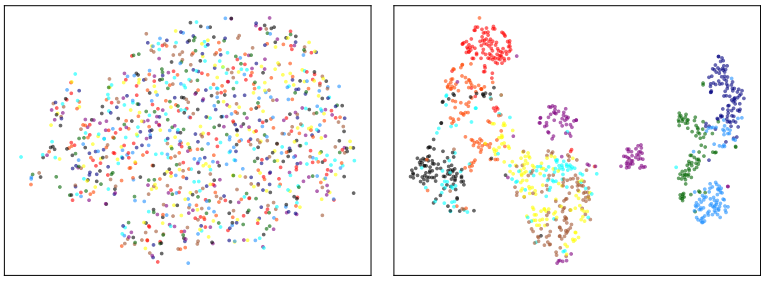
(a) latent space of z_1 of HVAE (b) latent space of z_1 of Ours



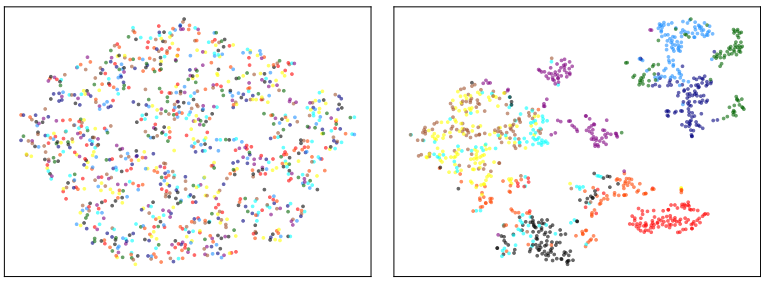
(c) latent space of z_2 of HVAE (d) latent space of z_2 of Ours



(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

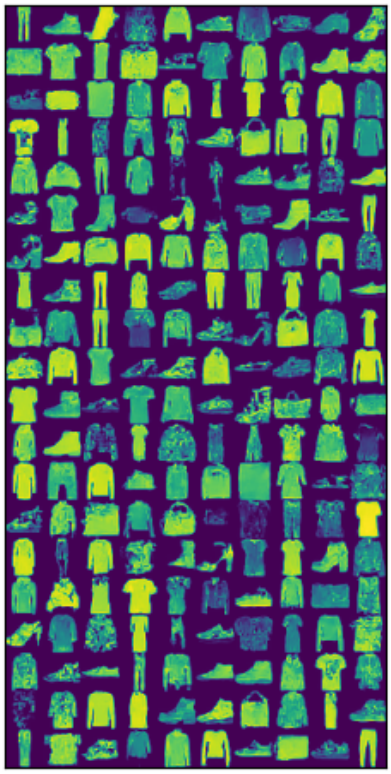


(i) latent space of z_5 of HVAE (j) latent space of z_5 of Ours

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.

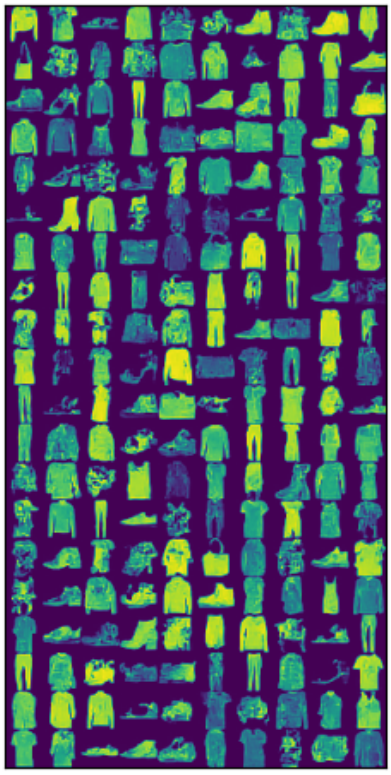
(a) Input data x (b) Partial Generation samples from $p(x|z_{>3})$ (c) Partial Generation samples from $p(x|z_{>4})$

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.

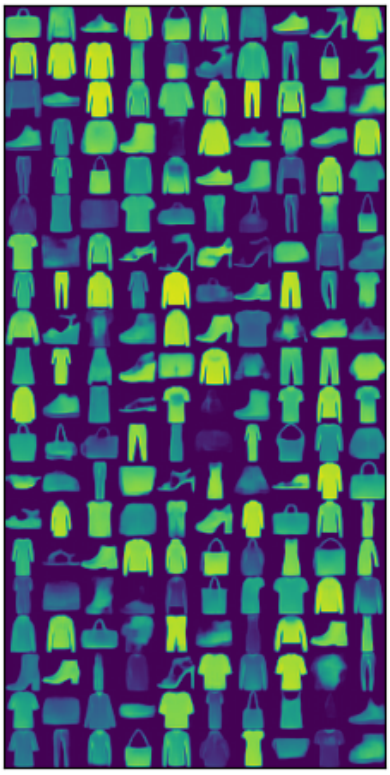


(a) Generated from a 3-layer HVAE

(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.