Additional Experiments for Rebuttal

Table 1: Comparisons between our methods and other VAE-based OOD detection methods on the "harder tasks" (CelebA(ID) / CIFARs(OOD)). Bold numbers are superior results.

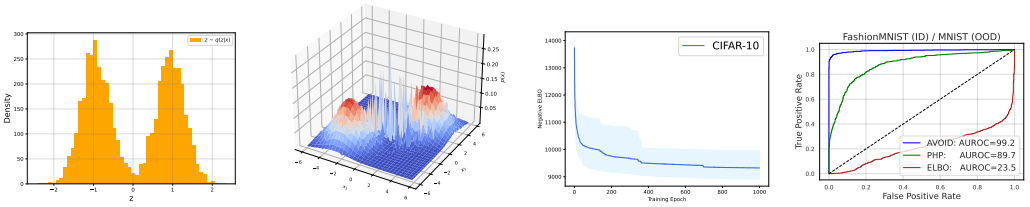
CelebA(ID) / CIFAR-10(OOD)				**CelebA(ID) / CIFAR-100(OOD)**			
Method	AUROC↑	AUPRC↑	FPR80↓	Method	AUROC↑	AUPRC↑	FPR80↓
ELBO [25]	27.8	37.5	96.3	ELBO [25]	33.1	41.9	96.7
HVK [17]	40.1	43.8	88.1	HVK [17]	45.2	49.0	91.2
$\mathcal{LLR}^{ada}$ [18]	58.0	62.5	77.3	$\mathcal{LLR}^{ada}$ [18]	52.5	58.8	85.6
-Ours				***-Ours***			
PHP	69.5	63.7	50.2	PHP	68.9	64.2	50.6
DEC	73.3	67.7	45.5	DEC	73.7	67.0	46.4
AVOID	**75.6**	**70.3**	**43.4**	AVOID	**75.5**	**69.8**	**42.1**

Table 2: Comparisons on more OOD datasets between our method and other VAE-based OOD detection methods with VAEs trained on CelebA(ID). Bold numbers are superior results.

AUROC↑ with models trained on **CelebA (ID)**									
OOD datasets	SVHN	STL10	Places365	LFWPeople	SUN	GTSRB	DTD	Const	Random
ELBO [25]	27.2	56.9	50.2	52.2	27.1	67.9	54.5	1.24	**100**
HVK [17]	36.8	59.7	59.1	**59.9**	54.3	49.8	61.5	92.9	74.4
$\mathcal{LLR}^{ada}$ [18]	91.2	61.5	55.7	58.6	58.8	42.3	68.1	90.2	73.4
-Ours									
PHP	56.9	59.9	63.5	52.5	67.2	72.0	63.2	53.4	100
DEC	99.7	60.1	60.9	55.7	66.1	67.8	68.5	97.0	100
AVOID	**95.8**	**67.6**	**68.4**	55.9	**73.7**	**75.6**	**76.3**	**97.1**	**100**

Table 3: Ablation study examining the effects of dataset size (data amount) and model capacity (number of convolutional neural network (CNN) layers) on the OOD detection performance of ELBO. Results indicate that increasing the amount of data and the number of CNN layers does not yield significant improvements.

FashionMNIST(ID) / MNIST(OOD)						CIFAR-10(ID) / SVHN(OOD)					
	Num. of Layers						Num. of Layers				
Data Amount	3	6	9	12	15	Data Amount	3	6	9	12	15
10000	9.45	14.0	13.2	14.2	14.6	10000	14.4	12.8	16.9	20.5	20.3
30000	16.3	14.5	15.3	14.5	15.8	30000	24.6	25.3	25.9	24.4	23.9
60000	23.5	25.1	23.0	20.3	19.8	50000	24.9	22.6	23.5	28.1	24.0



(a) $q_{\text{id}}(\boldsymbol{z})$ (b) Estimated $p_\theta(\boldsymbol{x})$ (c) Training curve (d) ROC curve

Figure 1: **(a)** and **(b)**: visualization of $q_{\text{id}}(\boldsymbol{z})$ and estimated $p_\theta(\boldsymbol{x})$ by ELBO on a synthesized 2D multi-modal dataset. The data amount here is 10 times larger than in Figure 3 of the main paper, increasing from 10,000 to 100,000 samples. The VAE used is a non-linear deep one based on a 10-layer MLP, in contrast to the 3-layer MLP used in Figure 3 of the main paper; Results indicate that the $q_{id}(\boldsymbol{z})$ is still not equal to $p(\boldsymbol{z}) = \mathcal{N}(0, \mathbf{I})$ and the *overestimation* issue still exists. **(c)**: Training curve of the negative ELBO on the CIFAR-10 dataset, obtained from five random runs with different seeds. **(d)**: ROC curve and its corresponding values for the AVOID, PHP, and ELBO methods.