We thank all reviewers for their useful comments. First we address the common concern regarding the **importance and impact of our work** in the following three arguments:

1. Although classifier based OOD detectors often outperform OOD detectors based on generative models, we should note that many machine learning systems deployed in practice are trained on unlabeled data, and generative models are widely used in tasks like language modeling and image editing. Therefore, OOD detection with unsupervised generative models is an important task.

2. Among likelihood based generative models, VAEs have the promising property of extracting representations from data, and they are particularly popular in applications (for example, VAE is a component of most image-to-image translation frameworks). Recent work obtains impressive OOD detection on flow/auto-regressive models, but our motivation is that, if VAE is being used in a task, it would be beneficial to enable the VAE itself to detect OOD inputs, rather than training a separate generative model merely for OOD detection. However, as we show, recent SOTA OOD scores for generative models cannot be applied to VAEs. Our work makes an important contribution by introducing an easy yet effective OOD score that works well on VAEs.

3. It is most natural to apply LR to VAEs due to the bottleneck structure serving as a regularization for optimizable model configuration, however, later we also explore the possibility of applying LR to flow/pixelCNN. In Table 1, we report the AUCROC on some tasks obtained from LR on Glow and PixelCNN. We only optimize the last coupling block and the last masked conv layer for Glow and PixelCNN respectively to avoid overfitting on a single input. Our results are competitive to those of previous SOTA methods on generative models. We therefore highlight LR as a general approach for OOD detection of all likelihood based generative models, with unique effectiveness on VAEs.

	Glow	PixelCNN
MNIST	0.995	0.987
KMNIST	0.990	0.973
NotMNIST	0.991	0.995
Noise	0.989	0.964
Constant	1	1

(a) Models trained on Fashion MNIST

	Glow	PixelCNN
SVHN	0.894	0.889
CelebA	0.735	0.774
LSUN	0.643	0.686
Noise	1	0.975
Constant	0.989	1

(b) Models trained on CIFAR-10

Table 1: AUCROC obtained from Likelihood Regret on Glow and PixelCNN.

We now address detailed concerns of each reviewer.

Reviewer 1: We apologize for the confusion caused by the notation. Indeed, the decoder should be denoted with θ. Our VAE is based on the DCGAN structure, which is fully convolutional and the last FC layer is replaced by conv layer that reduces the spatial dimension to 1×1 (which returns the vectors for posterior mean/variance). We tried different network structures, including those with FC layers, and have found little difference in terms of OOD detection. The precise definition of OOD samples is interesting and debatable, as we only have an in-distribution *dataset* rather than a true distribution. Strictly speaking, both blue circle and cars are OOD, but the former is much harder to detect, especially for generative models.

Reviewer 2: Please refer to items 1-3 for concerns regarding why we focus on VAE's OOD detection. In short, our motivation is to let VAEs detect OOD when it is used, and our extended experiments on flow/pixelCNN indicate that LR is general across generative models. As for why optimizing the encoder works better, for in-dist samples, optimizing the encoder is more constrained than directly optimizing z, which prevents the latent variables from moving too much. We will provided a detailed explanation in the revision. We will carefully fix writting issues as pointed out.

Reviewer 3: Please refer to items 1-3 for concerns regarding why we focus on VAE's OOD detection. We found that β-VAEs behave similarly to plain VAEs in terms of OOD detection. We will include more details on applying our method to other VAE variants. We will also include the MNIST results in the appendix as suggested. In particular, we obtain AUCROC 0.963 on your mentioned MNIST vs. EMNIST (VAE's likelihood gives 0.781). CIFAR-10 vs. CIFAR-100 is indeed a very hard task for generative models. We tested LR on this task and obtained AUCROC 0.582 (compared to 0.489 using likelihood), while likelihood ratio gives 0.564 and IC gives 0.535. This result is also comparable to IC on pixelCNN. We thank the reviewer for pointing out writing issues and typos in the code. These will be fixed, suggested additions will be included, as well as the correct reference for the definition of OOD.

Reviewer 4: Please refer to items 1-3 for concerns regarding why we focus on VAE's OOD detection and applying LR to other generative models. We only tested on image data because most recent OOD papers, both classifier based and generative model based, focus on image data. The Likelihood Ratio paper by Ren et al. introduces a gene sequence OOD dataset, and we tested our Likelihood Regret on a VAE trained on this dataset. Our LR obtains 0.745 AUCROC (the VAE's likelihood baseline gives 0.538). This result is competitive to Ren et al. (which uses a auto-regressive model whose likelihood gives 0.626 and their proposed likelihood ratio improves to 0.732).