We are very grateful to all reviewers for their detailed, insightful, and constructive comments and questions. We find many of the comments truly helpful to improve the quality of the paper, and some of them actually enlightened us, correcting some of our initial claims that turn out to be wrong.

Due to lack of time, we apologize for not responding to all questions, and not conducting all extra empirical evaluations suggested by the reviewers. But we believe that they are very important, and we will pursue them in our ongoing study.

Our responses (blue) to reviewers' comments/questions (***black/bold/italic***) are as follows.

Table 1: Test data log-likelihood scores for the **Binary MNIST**. Our results are in the column titled "CNN". The column "FC" is excerpted from [VLAE19] (Table 2).

	CNN	FC
VAE	-84.49	-85.38
SA$^{(1)}$	-83.64	-85.20
SA$^{(2)}$	-83.79	-85.10
SA$^{(4)}$	-83.85	-85.43
SA$^{(8)}$	-84.02	-85.24
IAF$^{(1)}$	-83.37	-84.26
IAF$^{(2)}$	-83.15	-84.16
IAF$^{(4)}$	-83.08	-84.03
IAF$^{(8)}$	-83.12	-83.80
HF$^{(1)}$	-83.82	-85.27
HF$^{(2)}$	-83.70	-85.31
HF$^{(4)}$	-83.87	-85.22
HF$^{(8)}$	-83.76	-85.41
ME$^{(2)}$	-83.77	-
ME$^{(3)}$	-83.81	-
ME$^{(4)}$	-83.83	-
ME$^{(5)}$	-83.75	-
VLAE$^{(2)}$	-	-83.72
VLAE$^{(3)}$	-	-83.84
VLAE$^{(4)}$	-	-83.73
VLAE$^{(5)}$	-	-83.60
RME$^{(2)}$	-83.14	-
RME$^{(3)}$	-83.14	-
RME$^{(4)}$	-83.09	-
RME$^{(5)}$	-83.15	-

1. Claims on semi-amortized variational inference (SAVI) methods:

We agree that we made somewhat exaggerated claims on the drawbacks of the SAVI methods. Specifically, some SAVI methods do not suffer from either step size adjustment or Hessian evaluation. We will refine our claims, and also refer to these SAVI methods.

2. About our claim that the normalizing flow models (eg, IAF) tend to overfit:

It turns out that it was our faulty claim. As the reviewers (esp., R1) suggested, we checked the training data likelihood scores of the IAF model, and found that oftentimes the training performance was even worse than the vanilla VAE. This implies that the failure of the normalizing flow models may not be because of overfitting, but difficulty of optimizing the complex encoder model. This is also argued/stated in the related work [VLAE19] (the second paragraph in Sec. 5.1).

3. The increased number of parameters in the proposed model:

In the paper, we only emphasized the merits of our model without mentioning the drawbacks. Clearly, our model has more parameters to estimate than the SAVI methods. The increased training time (due to the loop over the mixture components for each iteration) is yet another weak point. We will summarize and state the drawbacks of our model in the revised version.

4. Performance of the proposed model with a non-Gaussian (eg, Bernoulli) decoder, or binarized input images:

Our empirical evaluations were predominantly conducted with the convolutional architectures on real-valued image data. Several reviewers wondered how the proposed model would perform on different data/network setups. For the performance of our model with non-convolutional (fully connected) network architectures, please refer to Table 5 and 6 in our supplementary material at the time of paper submission.

For the binarized input images, we have conducted extra experiments on the **Binary MNIST** dataset. Please see Table 1 (on the right) for the results. We have set the latent dimension $\dim(\mathbf{z}) = 50$, and used the same CNN architectures as in our paper, except that the decoder output is changed from Gaussian to Bernoulli. We also include the reported results from [VLAE19] for comparison, which employed the same latent dimension 50 and fully connected encoder/decoder networks with similar model complexity as our CNNs'. Due to lack of time, we only report mean scores averaged over three runs. As shown, IAF and our RME performs equally the best, although the performance differences among the competing approaches are not very pronounced compared to real-valued image cases.

5. The KL bound parameter C:

We haven't tested extensively the impact of C. We once tried a few different values $\{100, 500, 1000\}$, and checked that the performance did not vary significantly. Although we guess that having too large or too small value of C would deteriorate the performance, we should do more rigorous empirical study in our ongoing work.

6. About RME(1). Due to the coordinate descent learning, it is not equivalent to the VAE:

Although RME(1), the single component mixture model, is optimized with the same loss function, its coordinate descent learning may yield a local optimum different from that of the VAE. We haven't compared the two models specifically, but we believe it is worth testing it empirically.

7. The VampPrior model as a baseline:

We agree that VampPrior can be another reasonable baseline in the sense that the model can potentially avoid component collapsing. Although VampPrior adopts a mixture-type *prior* while ours adopts a mixture *encoder*, understanding the relationship between the two, either empirically or theoretically, must be an intriguing research problem.

References

[VLAE19] "Variational Laplace Autoencoders", Park et al., ICML 2019