We thank the reviewers for their thoughtful feedback, and address the main questions raised in turn.

R1: 1) We clarify that although sparsemax can be applied during training, it can also be applied post hoc, just like vanilla softmax. Since we present a post hoc sparsification method, we apply sparsemax post hoc in our experiments for fair comparison. Both our method and sparsemax compute implicit thresholds for the distribution. However, both methods (1) do not need to be tuned for each network or dataset, and (2) automatically adapt to individual input features, unlike a static threshold. We consider a static threshold baseline in our MNIST experiment in Fig. 2 for intuition. We select a static threshold of 0.14 using cross-validation with the NLL metric. However, this static threshold caused even more false negatives than sparsemax. By visual inspection, it is impossible to choose a single static threshold that would outperform our method in balancing sparsity and multimodality in Fig. 2 (either additional false negatives or false positives would result). While we could tune a static threshold for each individual input query for the MNIST task (even/odd), this would be intractable for a continuous trajectory input query as in the behavior prediction task.

2) We will add discussion of the work by Laha et al. to our paper. They augment the sparsemax technique to control sparsity, presenting sparsegen-lin (sparsemax with regularization) and sparsegen-hg (sparsemax with a scaled input). Both of these methods require hyperparameter tuning which may be computationally prohibitive on larger networks. We applied these methods to the MNIST experiment in Fig. 2, and found that they either pruned the same latent classes as sparsemax or, with certain parameter choices, produced even sparser distributions with more false negatives. Our method better balances multimodality and sparsity while adapting to different feature inputs *without parameter tuning*; indeed, such an "auto-tuning" property is a key advantage of our method.

In Eq. 8, $\hat{\beta}_{jk}$ are the linear layer weights learned by the neural network a priori, akin to the definition of the bias parameters (line 106). We will update our paper to include this definition, a brief summary of sparsemax, and a correction to the typo in Appendix B, Eq. 21.

R2: i) In addition to application-based papers and evidential theory in deep learning, our literature review covers (1) theoretical works on training VAEs with discrete latent spaces and (2) existing alternatives to softmax (see Sec. 4). There are indirectly related works in low-dimensional encodings for VAEs that focus on regularization [1] and enforcing structure in the latent space [2] during training, but they do not sparsify the latent space post hoc. We will include more VAE references in our updated literature review, and welcome further suggestions.

[1] Mathieu et al. Disentangling Disentanglement in Variational Autoencoders. In ICML, 2019.
[2] Kosiorek et al. Sequential attend, infer, repeat: Generative modelling of moving objects. In NeurIPS, 2018.

ii) Currently, we introduce the CVAE in lines 33–38 and provide details of the CVAE architectures we consider in lines 160–179, Fig. 1, and in Appendices C, H, I. We will update our paper to further describe CVAEs (e.g., the ELBO loss).

iii) We address R2's question: "could the approach amplify errors or mistakes in the discrete latent space?" as follows. By construction, both our method and sparsemax amplify the likelihood of the latent classes selected by each technique. However, our method strikes a better balance than sparsemax between sparsity and multimodality of the latent distribution. As demonstrated in our experiments, specifically in the MNIST experiment referred to by R2, our method improves the test time sampling performance of the CVAE by pruning the majority of the erroneous latent classes, while keeping the correct ones (see Appendix D, Fig. 6). When considering 25 samples for the "even" input, softmax produces eight incorrect samples, while our method produces only three (the incorrect 9 image). Conversely, sparsemax severely collapses the uncertainty in the latent distribution, removing correct latent classes from the distribution.

Regarding R2's question about the accuracy in Table 1, this performance is due to the quality of the VQVAE images generated using the PixelCNN prior network. With additional computational resources and the full ImageNet dataset (as in the VQVAE paper), this performance could be further optimized. However, we were still able to learn sufficiently good quality encodings using the VQVAE architecture and generate reasonable images (see Appendix H) to demonstrate the relative performance of softmax, sparsemax, and our proposed method on the latent distribution of the VQVAE.

We will include that z_k are one-hot encodings in line 68. In line 95, j is the index into the feature vector (defined in lines 93–95). Regarding R2's question on why modeling real-world tasks requires highly multimodal distributions, consider the example in lines 31–32. If a prediction model receives a trajectory of a person walking straight, the model should predict multiple plausible paths (e.g., continuing straight or turning left or right), constituting multimodal output.

R3: In both our method and sparsemax, it is not possible to manually tune the level of sparsity achieved in the latent space. Rather, by design, both methods compute an implicit threshold for each input query. Our method automatically balances the objectives of sparsity and multimodality by keeping only the latent classes that receive direct evidence from the network's features and weights, as described in Secs. 2.3–2.4. Although sparsemax reduces the latent space by a larger percentage than our technique, this negatively impacts its performance; our experiments demonstrate that sparsemax collapses the multimodality in the latent space distribution, resulting in undesirable pruning of correct latent classes. We convey these ideas in lines 199–204, but we will make them clearer in an updated version of the paper.