We would like to thank the reviewers for their thoughtful comments and questions.

The probabilistic framework we developed is intended to provide a set of tools for understanding the relationship between neural activity and behavior. These tools can compress, segment, and generate behavioral videos, as well as decode those videos from neural activity. We address reviewer concerns for these four tasks separately.

All three reviewers noted that compressing the videos with a convolutional autoencoder (CAE) did not seem to qualitatively outperform compression with a simple linear model (though the CAE did perform better quantitatively). We agree with this observation, and note that the use of the CAE is not critical to downstream analyses. However, by using the CAE we achieve the same MSE in pixel space as the linear model with fewer latents, thereby reducing the number of parameters in the subsequent segmentation and decoding models.

Reviewers 1 and 2 raised the concern that the discrete behavioral syllables inferred by the autoregressive hidden Markov model (ARHMM) were not interpretable, as has been demonstrated in previous work. First of all, we want to emphasize that interpretability of the behavioral syllables is only qualitative and is not our main concern; rather, we use the ARHMM as a prior model of behavior which is then incorporated into the Bayesian decoder. However, we agree that clear, nameable syllables are ultimately useful beyond decoding. We do in fact find some degree of interpretability in the behavioral syllables in the WFCI dataset, where a clear trial structure exists. In Fig. 3C we show the inferred behavioral syllables across many repeats of the trial. There is, for example, a syllable that almost always directly follows the lever grab (maroon), and another syllable that directly follows the spout movements (light blue). Reviewer 2 raised the concern that the syllable sequence is not the same across trials, but this could reflect trial-to-trial variability in animal behavior, which is typically not taken into account in standard analyses linking neural activity and behavior. Though we do not further pursue this variability in the manuscript (e.g. how behavioral variability is related to correct versus error trials) we think this is an extremely interesting application for these methods.

Reviewer 1 noted that the levers in the WFCI dataset (the round structures in the lower part of the video) are being reconstructed, which is indeed interesting. In this task, the mouse grabs the levers once they are moved closer. Our reconstructions indicate that information about the lever movement is present in the neural activity from which we are decoding, perhaps in a visual area.

Reviewers 2 and 3 suggested several comparison models to decode behavioral video from neural data. Reviewer 3 suggested fitting other models directly from neural activity to behavioral video, such as an RNN. If accepted, we will add this comparison to the manuscript (both quantitatively, through MSE in the pixel space, and qualitatively with reconstructed videos). Reviewer 2 also suggested comparing the decoding of the WFCI dataset to the trial-averaged video, given the stereotyped behavior introduced by the trial structure. We think some variant of this approach is a good comparison for understanding if variability in neural activity is actually able to decode variability in the behavior. Reviewer 3 asked how training this model end-to-end compares to the piece-wise training presented in our manuscript. We think this is an interesting question, and have already begun preliminary efforts in this direction, though cannot currently comment on the differences.

Finally, Reviewer 1 suggested the application of this approach to freely moving animals. We have in fact been doing this, and plan to share the results of that analysis in the future.