Author Response for 'Shaping Belief States with Generative Environment Models for RL'

We are grateful to all constructive and actionable feedback provided by the reviewers. We will account for all comments and suggestions in the revised version. We are especially thankful for the detailed feedback provided by **R2** and **R3**. We believe to have addressed the key concerns raised by the reviewers below.

General Comments: R1 expressed concerns regarding the novelty of this work, observing that most of the experiments in the paper involve published components. We also understand **R1**'s concerns with our main hypothesis as it has not been explored in the literature yet. We are working to improve our explanations in section 2.2 based on all feedback that we received.

We emphasize that careful empirical experimentation in ML can also bring valuable insights to the community. While the many ingredients of our paper exist in isolation, very little is known about how they interact with each other. As we explore complex RL agents, these different components must work together. Thus an empirical study of these interactions was timely.

As **R1** rightly observed, one of the most important contributions was to formulate and empirically evaluate a hypothesis regarding the formation of useful belief states in environment models. This involves the interaction between overshoot, probabilistic models and memory. Studying these factors require an intersectional empirical study such as this paper. Our hypothesis can be split in two parts: **H1** Overshoot is useful to learn models with long-term dependencies and **H2** Probabilistic models benefit more from overshoot than Deterministic models.

By the chain rule of probabilities we can always write the joint pdf of all observations as a product of next-step conditional pdfs, so we know that is not a problem in principle. But this is an incomplete view of the problem as it disregards the properties of the data. It is known theoretically and empirically that the properties of the data are as important as choosing the right model (e.g. imbalanced datasets require calibration, non-iid data require causal corrections). A more relevant question is: What is necessary to learn such models efficiently? We could formulate these as more formal statements about the statistical properties of the data collected by an agent walking in a 3D environment but this is beyond the scope of this paper. We expand these hypothesis here for clarity. **H1**: This is a problem of data imbalance and causal learning. Two successive frames observed by an agent walking in a 3D environment are highly correlated with each other, this implies that a model can predict the next frame with high accuracy without knowing much about the environment (e.g. by learning to displace the previous frame). However, two distant frames have a much less relation to each other (imagine two frames captured on opposing sides of a wall). In this case the probabilistic model cannot predict one frame by merely displacing the other, it is necessary to use a more global representation of the environment for such predictions. **H2**: With overshooting, another problem emerges: Due to the partial observability of the environment, the entropy of future frames conditioned on the past grows with the overshoot length. It is clear that deterministic models cannot perform multi-modal predictions. For sufficiently long overshoots, any deterministic prediction will inevitably converge to the average of all possible frames that could be seen. Since this prediction is independent of the belief-state, deterministic models should not benefit from long overshoots. Our experiments provide strong evidence for **H1** and **H2** on numerous complex environments.

Overshoot just adds more labels: Our experiments reject this hypothesis. If this was the case, increasing the overshoot length would not affect the asymptotic error in top-down view reconstruction, only the convergence time. As we can see in Figures 3 and 4 this is not the case.

Relevance to RL: We wholeheartedly agree that doing planning with our models would be a great follow up work. But we also believe that our experiments show benefits to the performance of a strong baseline model-free RL agent in complex environments due to the model. We appreciate the suggestion and will expand the discussion about model-based RL.

Why a new environment? We appreciate the observation and will expand a lot more the details in the revision. One of the main reasons for a new environment was that we believe that the benefits of expressive belief-state models will become more evident in combinatorial and compositional environments, where the agents are expected to perform a variety of different tasks. Our experiments indicate a substantial data-efficiency gain in these environments, Figure 7, and also add more evidence to the hypothesis put forward in the paper.

Comparison to [20]: Indeed, reference [20] is the closest to our paper. For this reason, we dedicated an entire paragraph in section 3 discussing our innovations relative to [20]. Some of the key components of our paper, namely the interaction between overshoot and generative models are not addressed in [20].

It is not surprising to decode the top-view from the LSTM state: Learning to extract the top-view of an environment uniquely from actions and first-person-views is by far not a trivial problem. As discussed in section 2.3, all known solutions involve a substantial injection of prior knowledge in the models.