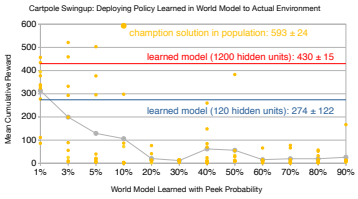
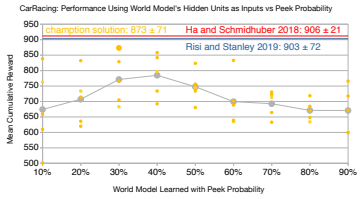
We thank the reviewers for their feedback and comments on the manuscript! Indeed, this work was largely an exploration of the idea: "Can a world model be learned without using a forward-predictive loss?" Owing to the intense research effort spent on learning forward predictive models in the presence of many a flavor of forward-predictive auxiliary losses, we focused our efforts on simple tasks that best illustrated the trade-offs of our exploration. We are assured to see that we have achieved this as all 3 reviewers thought our method was interesting, and that our presentation was clear. While adding more complex test environments would certainly strengthen our work, we assert that the experiments presented here provide a compelling sketch for a different way to train world models.

Because of the relative maturity of the model-based literature, we were initially hesitant to make direct comparisons with model-based baselines because doing so immediately invites a comparison that we did not intend to draw. However, we fully agree that including such a baseline is useful, informative, and, if nothing else, establishes a long term goal for this program. We have now included model-based baselines for the swing-up cartpole and car-racing tasks (see below), and we have launched more detailed baseline experiments for car-racing (suggested by R3) and gridworld.





Cartpole Swingup	Reward
Explicitly learned model from data (120 hidden units)	274 ± 122
Explicitly learned model from data (1200 hidden units)	430 ± 15
Champion solution in our population (120 hidden units)	593 ± 24

CarRacing-v0	Reward
Ha and Schmidhuber (NeurIPS 2018)	906 ± 21
Risi and Stanley (GECCO 2019)	903 ± 72
Champion solution in our population	873 ± 71

The full baselines and discussion will be readily available in the revision. To our surprise, we find it interesting that our approach can produce models that outperform an explicitly learned model with the same architecture size (120 units) for cartpole transfer task, but this advantage goes away if we scale up the learned model size by 10x. For car-racing, only the best solution out of a population of trials gets near the performance of population-based baselines from the literature, but honestly we did not expect our new implicit approach of learning world models to be this close in performance.

With respect to R1, R2 and R3's comments on prior work (especially the Predictron), we of course will add these overlooked references, and provide a discussion of similarities and differences of these works. In particular for the Predictron, both methods are after the same goals: a latent model emerging from optimization. The predictron does this via learning representations, dynamics, and value estimates end-to-end over imagined rollouts of varying lengths. The primary innovation of observational dropout is to recouple the "imagined" rollouts back to the real observation space—i.e., any sequence of transitions within our learned world models always end back on a "real" frame from the environment, with "real" environment steps in the intervening frames. Additionally, where the predictron derives consistency by demanding that its value estimates obey a Bellman equation, we only demand that the learned dynamics model facilitate the learning of a policy. While similar in spirit, we hope the difference here is clear—it's interesting, for example, to consider how a predictron could be combined with observational dropout, where at each step, instead of necessarily seeing another ground-truth frame from the environment, the predictron sees its own 1-step prediction.

Regarding R1 and R2's concerns on generalization (whether to "harder" or "unseen" tasks): we agree that understanding generalization is crucial, and regret that we did not have more space to discuss this point in the manuscript. In sum: for high observational dropout, too much information is lost, and the world model does not have the capability to accurately reconstruct the system due to noise. At low dropout, the world model might as well not be present, because the policy can ignore the occasionally noisy frame. The extent to which the world model can then transfer to "unseen tasks" is dependent on how well the world model fully captures the dynamics of the system. In some sense, world models learned in this way are "bad" at generalizing, because they are trained to only learn the "relevant" dynamics of the task being studied (as in the grid world task, where it only learned transitions in some but not all directions). In another sense, they're "good" at generalizing, because the "relevant" dynamics of the current task might also be the relevant dynamics of "unseen" tasks (i.e., as in balance cartpole, initialized at an angle not seen during training). It is difficult to make a precise statement here, because the quality of the learned world model truly depends on the dynamics being studied.

R3 had an interesting comment regarding the relative power of the policy versus the world model—indeed, a sufficiently powerful policy could extract meaningful information from the world model, even if the world model was not actually forward predictive at all. In this limit, the joint policy + world model system starts looking like a strange sort of recurrent network. We agree with this assessment, and will add a short discussion highlighting this limit. An alternative interpretation is that our procedure is a way of encouraging a recurrent model's internal state to be dynamically relevant/interpretable—first by mediating a policy's observation space via that lossy dynamics module, and then by constraining the policy itself to be simple, so that it is not, in R3's words, "sufficiently powerful". We are absolutely interested in pursuing this interpretation through an information bottleneck lens, but felt it was out of the scope of this work. Finally, we absolutely will add more details on the number of training steps used for the gridworld tasks (an equal number was used for both architectures, to answer R3), as well as details on the observation space of cartpole. We hope this response has addressed the reviewer's concerns, and would appreciate a reevaluation of your scores!