Paper: Generalization of Reinforcement Learners with Working and Episodic Memory

We thank the reviewers for their thoughtful and constructive feedback on our manuscript. We are excited about this work and glad for their help in improving it.

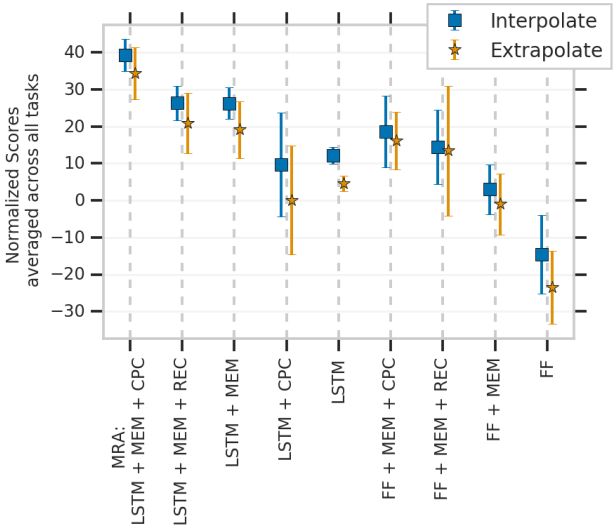
We can confirm that the full task suite will be released at the time of publication (Reviewers 4 and 5) and will include videos for each task. This should help both contextualize each task's difficulty and illustrate what it involves.

Reviewer 3 noted the Section 2 task descriptions could be better presented. We have reformatted it so that "the order of the figures matches the order that the tasks are presented in the main text" as suggested. We moved Figure 1, an overview of the 3D task types, to the beginning of the section to reduce confusion. Due to space constraints we are not able to show figures for all 13 tasks in this section, but these per-task figures will be in the Appendix and will be referenced when describing each task. We also changed our description of IMPALA to match Reviewer 5's suggestion.

Regarding the task suite, Reviewer 4 raised a thoughtful consideration on whether "most of the findings translate when some confounding elements of the tasks are removed e.g. by making the environment 2D, or by removing the need to navigate around". Some 3D tasks in the suite already have '2D-like' semi-counterparts that do not require navigation, which we hope may be able to shed light on whether our findings would hold up if translated from 3D to 2D versions. Namely, for the *Spot the Difference* tasks (Basic and Passive), removing the 3D partial-observability and navigation features would produce something close to the PsychLab *Change Detection* task (included in our task suite) in terms of which aspects of memory and of the agent's other abilities are being evaluated. We consider the PsychLab tasks '2D-like' because everything is fully observable and the agent has a first-person point of view from a fixed point, without any need to navigate. Based on our heatmap results, we found that *Spot the Difference: Passive*, which is the simplest *Spot the Difference* level, was overall harder than *Change Detection* for our ablation models. That said, a full analysis between these tasks and newly created fully 2D analogs runs the risk of overwhelming a single paper. We thus leave a full enquiry into this issue for future work.

Reviewer 4 noted they "would have liked to see included in the testbeds versions of environments (or some evaluation metrics) for which the proposed method fails". Two examples where our agent and baselines perform poorly (as can be seen in Figure 5) are *Spot the Difference: Motion* and *What Then Where*. We will highlight this more explicitly in our main text.

On our generalization results, we acknowledge that the heatmap numbers are not as easy to read as we would like them to be (Reviewer 5). Due to space challenges we also had to relegate our generalization-related plots per task to the Appendix. We thank Reviewer 3 for their excellent suggestion to make a figure that combines information across tasks, which we hope can tie in with our findings about generalization more clearly. We have created the attached plot showing each model's normalized performance on training and holdout, averaged across all tasks, and will include it in the main text.



Reviewer 5 asked for clarification on 'the differences in performance between CPC and REC'. Why CPC (contrastive predictive coding) was more helpful for the harder tasks in our suite and REC (reconstruction loss) on the simpler ones certainly bears further investigation. From the literature, Guo et al (2018) find that in partially observable environments such as first-person-view navigation tasks, where each observation provides only a partial and possibly noisy view of the environment, it is vital for the agent to learn a representation that encodes its uncertainty about the underlying state of the environment, and further find that CPC was more useful in getting the learned representation to encode the agent's position and orientation on visually complex 3D tasks, whereas one-step frame prediction was more useful on visually simple tasks.

Review 4 asked for some clarification on 'the hypothesis presented in lines 268-270'. The hypothesis was that CPC captures subtler differences than REC. From the loss functions of these two methods, it should be clear that while REC requires features capable of reconstructing the full scene on a per-pixel basis, CPC is satisfied with a representation that is distinguishable from the alternatives. This is not always a good thing, as this means CPC (and mutual information maximizers in general) can have a problem with representing high amount of information (see Ozair et al, 2019). But since only a few bits are needed for the episodic memories in our tasks (e.g. textures and shadows are irrelevant), it's probable that CPC's representational strategy is superior on the more challenging tasks in our suite.