We thank all reviewers for their time and useful feedback.

General point about Figure 2: We apologize for a typo in the figure. Y_t should be U_t as defined in Section 2 (the value target). There was a late notation change that wasn't reflected in the figure during edits. Thanks to reviewer R1-3 for comments about the figure, we'll improve clarity as well.

R1:

- **About off-policyness:** This is a great point. In our experiments, the model targets the expected ϕ under the behavior policy (which is different but close to π), so this model target could be different from the expected ϕ under π. However, even without any correction, this is a valid model of the future which we learn to interpret for better value predictions (Note: the hindsight and normal value weights are not shared in this case, $\theta_1 \neq \eta_1$). Different correction schemes are possible, but we wanted to keep the approach simple; we'll add a note in the paper. One related note: R2D2 computes n-step returns without any correction for off-policyness, but IMPALA corrects the value targets with VTrace.
- **"(Weber 2017),"** We apologize, that's an omission and we meant to cite it since it inspired some of the design choices (using the model as feature).

R2:

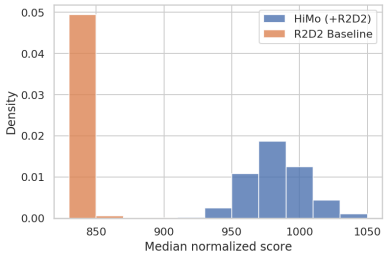
- **about baselines:** The reason we selected model-free baselines was that 1) our method in a way sits between classical model-free and model-based methods and 2) model-free methods have been dominating empirically in many domains. So we wanted to see whether our approach could contribute beyond the best performing method. Given the recent results of MuZero in Atari, that would also be a good candidate but 1) we cannot directly compare fairly against the published results since any difference in the setup/network architecture would not provide a well-controlled experiment and 2) MuZero is considerably more expensive to run than R2D2 compute-wise to reproduce in a comparable setting. Nevertheless, we have started a broader experimental investigation by providing a comparison to a model-based approach in Fig 1 and preliminary results of HCA in Portal choice (where it did not perform well). And we plan to study more comparisons in the future.

R3:

- **"(189) seems unclear / unmotivated."** This is motivated by the Imaginative Agents line of work (Weber et al. 2017) of using the model output as an input feature to an agent. We'll expand to make this clearer.
- **"Moving the related work section up"** That's a good suggestion, we'll move it.

R4:

- **"Unless you can get a really big effect, ..."** Many major results in our field were achieved by combining 'incremental' performance gains together. Our approach, which all reviewers agreed is novel, is evaluated in two smaller scale domains where the performance is clearly superior to the baseline (Fig 1R, Fig 4a) and where it's feasible to provide a detailed empirical analysis. The goal of the Atari experiment was to demonstrate that the same setup could work at scale in combination with SOTA RL agents. Our results show that the proposed approach combines easily with these agents and it significantly improves their performance in many games. Note that R2D2 already gets the maximum score in a number of games, so there is a ceiling to possible improvements. Nonetheless, the aggregate median metric shows an overall improvement (832.5% to 965%) and the detailed scores per game show that the performance improved significantly in a number of games.
- **"demonstrate why the features learned by the algorithm are interesting"** An analysis of the features learned by HiMo in the Portal Choice domain is already provided in Figure 4-c. It shows that the features capture the room color identity in practice, which is exactly the right information to model in hindsight.
- **"quantification of the uncertainty"** All individual domains were evaluated using multiple seeds, and the corresponding plots indicate uncertainty intervals. We have additionally updated our main Atari result, which aggregate results across games, to also quantify the uncertainty. This was done by applying bootstrap sampling to the evaluation episodes across the 3 seeds (repeated 5000 times), resulting in 95% CI intervals of [941.05%-1028.67%] for HiMo's median normalized score, and [832.5%-838.38%] for the R2D2 baseline's median score. A 2-sided K-S test allows us to reject the hypothesis that these were drawn from the same underlying distribution ($p \leq 1e^{-10}$)



Empirical median distributions in Atari obtained from a bootstrap procedure.

- **"algorithm does work faster, but the baseline still converged to the optimal performance."** R4 acknowledges that our algorithm learns faster, but this is exactly our claim. Vanilla model-free RL methods will converge to the optimal performance eventually (modulo some assumptions) but they may be slow getting there. Our approach attempts to leverage the same trajectory data in a more effective way to improve the learning process. This is what we demonstrate in the experiments.
- **"Value Prediction Network"** Please note this paper is already cited, cf. [13].