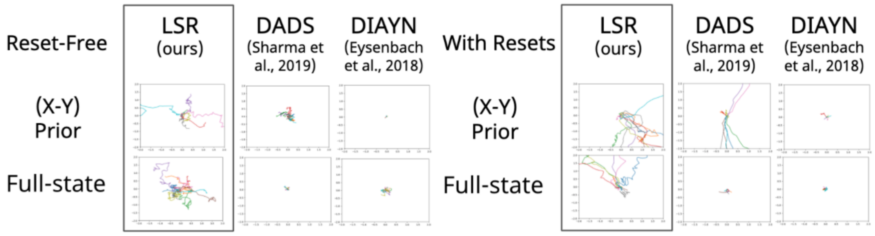
We thank the reviewers for the thoughtful feedback and suggestions. We address specific questions and include suggested evaluations. Any clarifications that were raised and not specifically discussed here due to space will be addressed in the final version, and we will include all requested clarifications and discussion of prior work.

R1 – "some concern that the downstream HRL experiments are somewhat unfair to the baseline methods (DADS and DIAYN)" We emphasize that the experiment is set up in a fair way, in the sense that all methods get the same state representation. That said, we agree that DADS and DIAYN are capable of performing better with additional supervision using the (x,y) prior (as is our method!). To address this concern, we present results below to help characterize the different skill learning capabilities across different settings. We will include quantitative comparisons for all of these in the final paper.

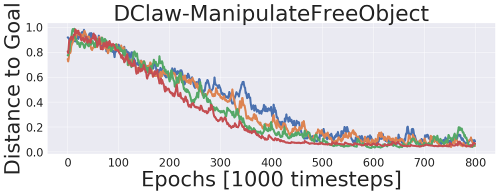


All methods learn meaningful skills, but DADS skills with the (x, y) prior travel much further. LSR learns skills that travel far both with and without the xy prior, and LSR learns substantially better skills in both settings in the reset-free setting. We will also investigate other design choices suggested by R1 in the final version of the paper.

R1,R3 – "how to choose the task for the forward policy" / "is it reasonable to assume that an RL agent deployed in the real-world will have the ability to consistently verify that it has successfully completed it's task" We agree that this is an important real world challenge that our approach does not attempt to address. We will highlight this in the final version as an important challenge for building a full real-world learning system. Multiple prior works do already study these questions, including methods for learning rewards and verifying the success of a task, learning tasks from demonstrations, etc. With regards to the interesting suggestion of learning a zero or few-shot classifier, it would be interesting future work to investigate how prior [Xie et al., CoRL 2018] could be adopted or made compatible to the assumptions in our work.

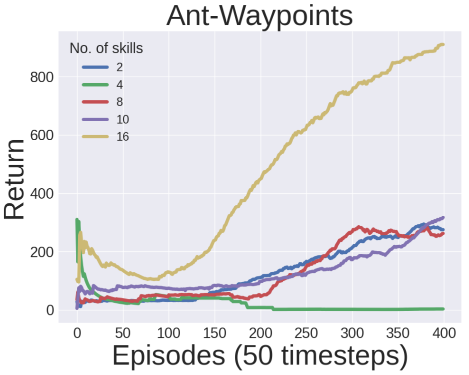
R2,R3: "Not all RL environments can be "reset" by reaching a state" / "Perhaps add an assumption that the underlying MDP is irreducible" These comments raise good points which we will address by adding a "Limitations and Future Work" section to the final version of the paper. In short, the reset-free learning setting that we consider implicitly assumes that the underlying MDP is resettable from different states. This is equivalent to limiting our approach to MDPs which are irreducible. Prior work on reset-free learning has also made this assumption [Chatzilygeroudis et al. RAS 2016, Eysenbach et al. ICLR 2019, Zhu et al. ICLR 2020]. Interesting future directions to address this issue would be design systems that explore conservatively (to avoid scenarios such as ones described by R2) or make use of limited human supervision in non-reversible MDPs when the agent encounters a non-communicating state.

R4: 'the stability of the proposed objective" We found that our approach is reasonably stable across different seeds in our method, which we plot individually in the plot on the right. We will add a detailed hyperparameter stability analysis in the final version, as we did not have time to complete one during the rebuttal phase.



R3: "did you control for hyperparameter tuning" We followed the base hyperparameter exactly as provided in prior work [Zhu et al. ICLR 2020], which the authors tuned specifically for their approach. We did not re-tune these base parameters beyond the initial learning rate and tuned the scale term λ introduced in our approach using a gridsearch.

R4 – "The importance of this hyperparameter is not mentioned" In the paper, we chose this hyperparameter to be roughly on the order of prior work that considered similar domains. We agree however, that this is an important hyperparameter, as it can have a complex effect on downstream RL performance. We provide on the right a plot showing the effect of varying the number of skills on the `Ant-Waypoints` tasks. We find that increasing the number of skills improves performance (16, yellow curve), the value used in our experiments (10, purple curve) is far from the best, and many other values attain similar performance.


Ant-Waypoints
No. of skills
Return
Episodes (50 timesteps)