We thank the reviewers and are happy they found the paper is well-written (R1, R2, R3, R4), the method is theoretically sound (R1, R2), the experiments are appropriate and comprehensive (R2, R3, R4), the results are convincing (R1, R3, R4), and the ablation studies are “tremendously useful” and helpful for making design choices (R1, R2, R3, R4). We address individual questions below.

R1: POMDP claims. We’ll update the paper to stress that our method is not equipped to solve the POMDP, e.g. reducing uncertainty. It was not our intent to claim that it was. Although we use the POMDP formalism to derive our algorithm, we agree with R1 that our experiments do not address the capability of our method to solve POMDPs.

R1: SOTA claims. We’ll remove the SOTA claims in light of the recent works CURL, RAD, and DrQ [1]. We strongly agree that this is distracting and less important than the content of the paper, and we had not intended to present that as the main contribution. Regarding the results: to our knowledge, DrQ (which came out on 28 April 2020, and would be reasonably considered concurrent with NeurIPS submissions) is the best among these methods, and SLAC is comparable to DrQ in the 4 DM control tasks, as shown in Figure 1. Note that the DrQ and CURL papers report outdated and lower-performing results for our method, compared to the results from this submission. That said, we will remove claims about state of the art results, and add results for all methods.

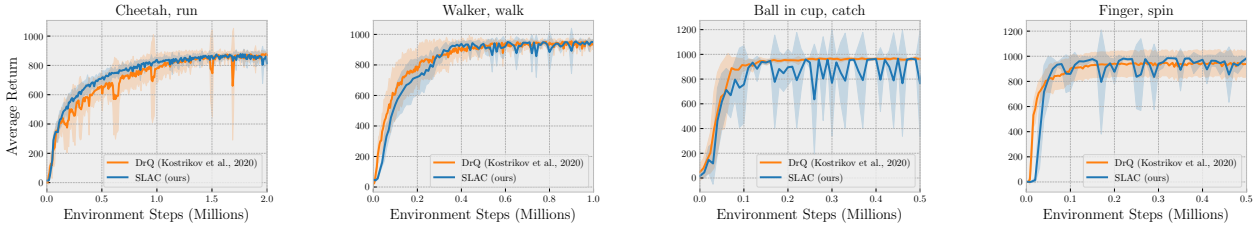

Figure 1: SLAC (ours) achieves comparable performance as DrQ (Kostrikov et al., 2020 [1]) in the 4 DM control tasks. Note that the plots include the initial exploration steps (10k for SLAC and 1k for DrQ), in which data is collected from a uniformly random policy.

R2: Divergence and instability. We hypothesize that it’s caused in part by the constantly changing latent space of the model and the model overfitting to the most recent data (the replay buffer contains the 100 most recent episodes). We believe that our method could benefit from early stopping or a schedule for model updates (e.g. having a "target" model, similar to target networks in RL). This merits further investigation, which we leave for future work.

R2: Task-dependent representations. The latent representation does receive some task-dependent information (in the form of reward prediction), although we acknowledge it could further incorporate more feedback from the value functions and policy. We chose to keep them separate but we recognize that it’s worth investigating futher.

R3: "The two main components of the algorithm can be drawn from existing methods in a straightforward way." We acknowledge that the individual components are based on previous work, and we will discuss this more clearly in the paper. Although the resulting algorithm is simple, we believe that formally establishing a link between them is important to help us understand the implications of certain design choices. E.g., one interesting finding is that, as indicated by Equation (21) in the Appendix, the optimal policy of the SLAC objective is optimal with respect to the expectation over the belief of the Q value of the learned MDP. This is equivalent to the Q-MDP heuristic, which "amounts to assuming that any uncertainty in the agent’s current belief state will be gone after the next action" (Littman et al., 1995 [2]). We’ll update the paper to emphasize these observations.

R3: "Overly optimistic / risk seeking." We believe we don’t have that same issue since the agent isn’t allowed to control the dynamics of the *future* (i.e., the policy objective in Equation (10) updates only the policy parameters). Furthermore, the history-dependent policies mitigate a similar issue, except that it’s caused by something different: When executing a latent-conditioned policy, the latent distribution is available but the specific latent state is unknown. Any attempt to choose any latent (e.g. the mode or a random sample) results in an overly confident agent since the policy assumes that the given latent is correct. By conditioning the policy in the observations, we instead obtain the Q-MDP heuristic (see paragraph above). We acknowledge that our empirical evaluation doesn’t exemplify this. This issue doesn’t arise if the belief is narrow. We’ll update the paper to clarify this.

R3: Correctness and clarity. We’ll update the paper with clarifications on the points brought up by R3. Most importantly, the true state of the environment is never available.

R1: Reference in the related work: We’ll add the missing reference to the paper.

[1] Kostrikov et al. Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels. arXiv:2004.13649, 2020.
[2] Littman et al. Learning policies for partially observable environments: Scaling up. Machine Learning Proceedings, 1995.