We thank all reviewers for their efforts and thoughtful comments, which are helpful for improving our paper.

Response to Reviewer #1: Q1. Ablation on entropy loss and clarification on model loss. Policy entropy maximization and model error minimization are deduced by our objective of optimizing mutual information between real and imaginary trajectories. Since the model loss term has the same form as Dreamer, we directly use the same model learning component as Dreamer that adopts multi-step prediction and removes latent overshooting used in PlaNet. Our main difference from Dreamer is the policy entropy maximization and thus our comparison experiments with Dreamer can be seen as the ablation on the entropy loss, as shown in Section 5.2. We will add these clarifications to our paper and further refine the presentation of entropy loss and model loss. In addition, we also conduct ablation in Section 4.3 and 5.3 (BIRD vs. soft-BIRD) to show that simply encouraging policy entropy by incorporating soft learning into model-based RL does not work.

Response to Reviewer #2: Q1. Cheetah and Quadruped are not presented. As shown in Figure 1, BIRD also outperforms Dreamer on high-dimensional (Cheetah and Quadruped) or sparse-reward tasks (Cartpole Swingup Sparse).

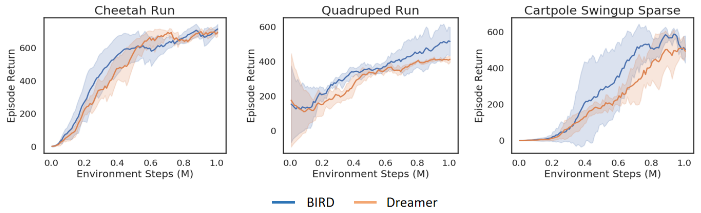

Figure 1: Results on more tasks.

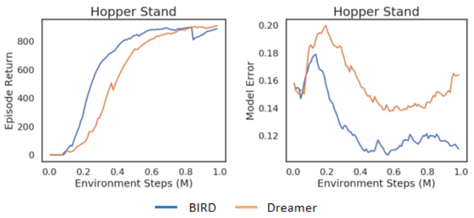

Figure 2: Comparison of model error.

Q2. Some papers regularizing the discrepancy between imagined and real trajectories are not discussed or compared. These related papers use MPC (described in Line 29-41 of our paper) for sampling-based planning and do not show effectiveness on RL tasks with image inputs. Compared to the MPC-based approaches that generate many rollouts to select the highest performing action sequence, our paper builds upon analytic value gradients that can directly propagate gradients through a differentiable world model and is more computationally efficient on domains that require learning from pixels. Our paper focuses on visual control tasks, and thus we only compare with state-of-the-art algorithms of these tasks (i.e., PlaNet and Dreamer). We will properly cite all these papers as the reviewer suggested and add these discussions in the next version.

Q3. A rigorous comparison (with Dreamer) of prediction error. Image reconstruction error will be dominated by image background and cannot reflect the prediction error on latent state. Thus we calculate the model error as the discrepancy between latent states that predicted by model and encoded from posterior image observations. As shown in Figure 2, BIRD that significantly outperforms Dreamer has a much lower model error.

Response to Reviewer #3: Q1. It is unclear how important SVG and D4PG are to the proposed algorithm. Because our proposed framework adopts an end-to-end model-based RL paradigm, as a differentiable model-based policy optimization method, SVG is an indispensable module of our framework, which accounts for the reason we cannot remove SVG to conduct ablation study on it. In addition, our framework does not include D4PG module so we also cannot conduct ablation on D4PG. We only use it as a baseline to demonstrate that our proposed algorithm can significantly outperform a popular model-free RL algorithm.

Q2. There are also no experiments which take states as input. This paper focuses on visual control tasks and aims at improving the state-of-the-art RL algorithm (Dreamer) for these tasks, and thus we follow experiment settings used by Dreamer.

Q3. What happens if we do not include the mutual information term? Mutual information term consists of model error and policy entropy. If we only remove policy entropy, our algorithm becomes Dreamer. The comparison experiments in Section 5.2 show that our model outperforms Dreamer. If we further remove model error, the performance of learned policy will be much worse.

Response to Reviewer #5: Q1. I'd expect a bit more significant improvement over recent approaches. Learning policy from raw visual observation has always been a challenging problem for RL algorithms. We significantly improve the state-of-the-art visual control approach (i.e., Dreamer) on Hopper Stand and Hopper Hop and achieve slightly better or comparable performance on other 10 tasks by stochastic value gradients in conjunction with mutual information maximization, which provide a promising avenue for model-based policy learning from pixels.

Q2. More seeds. We run experiments with more seeds and the results are similar. We will add these results in the next version.