

# Planning to Explore via Self-Supervised World Models

Ramanan Sekar<sup>1\*</sup> Oleh Rybkin<sup>1\*</sup> Kostas Daniilidis<sup>1</sup> Pieter Abbeel<sup>2</sup> Danijar Hafner<sup>3,4</sup> Deepak Pathak<sup>5,6</sup>

## Abstract

Reinforcement learning allows solving complex tasks, however, the learning tends to be task-specific and the sample efficiency remains a challenge. We present Plan2Explore, a self-supervised reinforcement learning agent that tackles both these challenges through a new approach to self-supervised exploration and fast adaptation to new tasks, which need not be known during exploration. During exploration, unlike prior methods which retrospectively compute the novelty of observations after the agent has already reached them, our agent acts efficiently by leveraging planning to seek out expected future novelty. After exploration, the agent quickly adapts to multiple downstream tasks in a zero or a few-shot manner. We evaluate on challenging control tasks from high-dimensional image inputs. Without any training supervision or task-specific interaction, Plan2Explore outperforms prior self-supervised exploration methods, and in fact, almost matches the performances oracle which has access to rewards. Videos and code: <https://ramanans1.github.io/plan2explore/>

## 1. Introduction

The dominant approach in sensorimotor control is to train the agent on one or more pre-specified tasks either via rewards in reinforcement learning, or via demonstrations in imitation learning. However, learning each task from scratch is often inefficient, requiring a large amount of task-specific environment interaction for solving each task. How can an agent quickly generalize to unseen tasks it has never experienced before in a zero or few-shot manner?

<sup>\*</sup>Equal contribution <sup>1</sup>University of Pennsylvania <sup>2</sup>UC Berkeley <sup>3</sup>Google Research, Brain Team <sup>4</sup>University of Toronto <sup>5</sup>Carnegie Mellon University <sup>6</sup>Facebook AI Research. Correspondence to: Oleh Rybkin <oleh@seas.upenn.edu>.

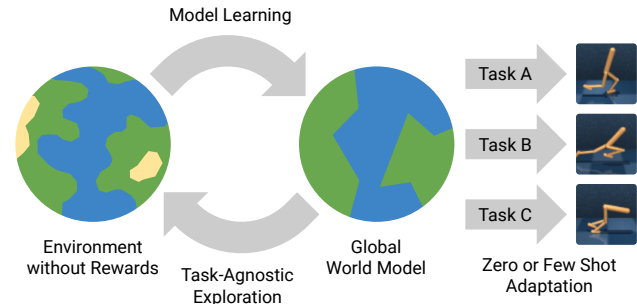


Figure 1. The agent first leverages planning to explore in a self-supervised manner, without task-specific rewards, to efficiently learn a global world model. After the exploration phase, it receives reward functions at test time to adapt to multiple downstream tasks, such as standing, walking, running, and flipping using either zero or few tasks-specific interactions.

**Task-agnostic RL** Because data collection is often expensive, it would be ideal to not keep collecting data for each new task. Hence, we explore the environment once without reward to collect a diverse dataset for later solving any downstream task, as shown in Figure 1. After the task-agnostic exploration phase, the agent is provided with downstream reward functions and needs to solve the tasks with limited or no further environment interaction. Such a self-supervised approach would allow solving various tasks without having to repeat the expensive data collection for each new task.

**Intrinsic motivation** To explore complex environments in the absence of rewards, the agent needs to follow a form of intrinsic motivation that is computed from inputs that could be high-dimensional images. For example, an agent could seek inputs that it cannot yet predict accurately (Schmidhuber, 1991b; Oudeyer et al., 2007; Pathak et al., 2017), maximally influence its inputs (Klyubin et al., 2005; Eysenbach et al., 2018), or visit rare states (Poupart et al., 2006; Lehman & Stanley, 2011; Bellemare et al., 2016; Burda et al., 2018). However, most prior methods learn a model-free exploration policy to act in the environment which needs large amounts of samples for finetuning or adaptation when presented with rewards for downstream tasks.

**Retrospective novelty** Model-free exploration methods not only require large amounts of experience to adapt to downstream tasks, they can also be inefficient during exploration. These agents usually first act in the environment,

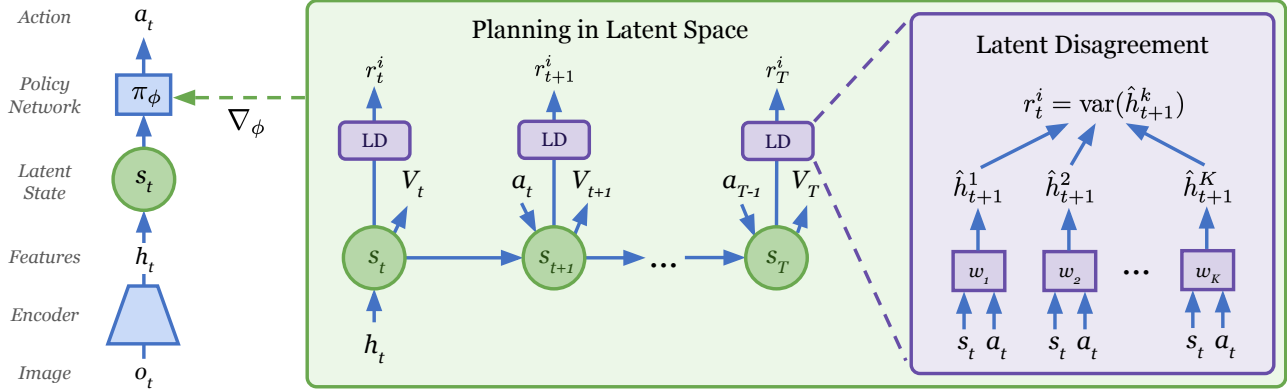


Figure 2. Overview of Plan2Explore. Each observation  $o_t$  at time  $t$  is first encoded into features  $h_t$  which are then used to infer a recurrent latent state  $s_t$ . At each training step, the agent leverages planning to explore by imagining the consequences of the actions of policy  $\pi_\phi$  using the current world model. The planning objective is to maximize expected novelty  $r_t^i$  over all future time steps, computed as the disagreement in the predicted next image embedding  $h_{t+1}$  from an ensemble of learned transition dynamics  $w_k$ . This planning objective is backpropagated all the way through the imagined rollout states to improve the exploration policy  $\pi_\phi$ . The learned model is used for planning to explore in latent space, and the data collected during exploration is in turn used to improve the model. This world model is then later used to plan for novel tasks at test time by replacing novelty reward with task reward.

collect trajectories and then calculate an intrinsic reward as the agent’s current estimate of novelty. This approach misses out on efficiency by operating retrospectively, that is, the novelty of inputs is computed after the agent has already reached them. For instance, in curiosity (Pathak et al., 2017), novelty is measured by computing error between the prediction of the next state and the ground truth only after the agent has visited the next state. Hence, it seeks out previously novel inputs that have already been visited and would not be novel anymore. Instead, one should directly seek out future inputs that are expected to be novel.

**Planning to explore** We address both of these challenges – quick adaptation and expected future novelty – within a common framework while learning directly from high-dimensional image inputs. Instead of maximizing intrinsic rewards in retrospect, we learn a world model to plan ahead and seek out the expected novelty of future situations. This lets us learn the exploration policy purely from imagined model states, without causing additional environment interaction (Sun et al., 2011; Shyam et al., 2019). The exploration policy is optimized purely from trajectories imagined under the model to maximize the intrinsic rewards computed by the model itself. After the exploration phase, the learned world model is used to train downstream task policies in imagination via offline reinforcement learning, without any further environment interaction.

**Challenges** The key challenges for planning to explore are to train an accurate world model from high-dimensional inputs and to define an effective exploration objective. We focus on world models that predict ahead in a compact latent space and have recently been shown to solve challenging control tasks from images (Hafner et al., 2019; Zhang et al., 2019). Predicting future compact representations facilitates

accurate long-term predictions and lets us efficiently predict thousands of future sequences in parallel for policy learning.

An ideal exploration objective should seek out inputs that the agent can learn the most from (epistemic uncertainty) while being robust to stochastic parts of the environment that cannot be learned accurately (aleatoric uncertainty). This is formalized in the expected information gain (Lindley, 1956), which we approximate as the disagreement in predictions of an ensemble of one-step models. These one-step models are trained alongside the world model and mimic its transition function. The disagreement is positive for novel states, but given enough samples, it eventually reduces to zero even for stochastic environments because all one-step predictions converge to the mean of the next input (Pathak et al., 2019).

**Contributions** We introduce Plan2Explore, a self-supervised reinforcement learning agent that leverages planning to efficiently explore visual environments without rewards. Across 20 challenging control tasks without access to proprioceptive states or rewards, Plan2Explore achieves state-of-the-art zero-shot and adaptation performance. Moreover, we empirically study the questions:

- How does planning to explore via latent disagreement compare to a supervised oracle and other model-free and model-based intrinsic reward objectives?
- How much task-specific experience is enough to fine-tune a self-supervised model to reach the performance of a task-specific agent?
- To what degree does a self-supervised model generalize to unseen tasks compared to a task-specific model trained on a different task in the same environment?
- What is the advantage of maximizing expected future novelty in comparison to retrospective novelty?

## 2. Control with Latent Dynamics Models

World models summarize past experience into a representation of the environment that enables predicting imagined future sequences (Sutton, 1991; Watter et al., 2015; Ha & Schmidhuber, 2018). When sensory inputs are high-dimensional observations, predicting compact latent states  $s_t$  lets us predict many future sequences in parallel due to memory efficiency.<sup>1</sup> Specifically, we use the latent dynamics model of PlaNet (Hafner et al., 2019), that consists of the following key components that are illustrated in Figure 2,

$$\begin{aligned}
 \text{Image encoder:} & \quad h_t = e_\theta(o_t) \\
 \text{Posterior dynamics:} & \quad q_\theta(s_t \mid s_{t-1}, a_{t-1}, h_t) \\
 \text{Prior dynamics:} & \quad p_\theta(s_t \mid s_{t-1}, a_{t-1}) \\
 \text{Reward predictor:} & \quad p_\theta(r_t \mid s_t) \\
 \text{Image decoder:} & \quad p_\theta(o_t \mid s_t).
 \end{aligned} \tag{1}$$

The image encoder is implemented as a CNN, and the posterior and prior dynamics share an RSSM (Hafner et al., 2019). The temporal prior predicts forward without access to the corresponding image. The reward predictor and image decoder provide a rich learning signal to the dynamics. The distributions are parameterized as diagonal Gaussians. All model components are trained jointly similar to a variational autoencoder (VAE) (Kingma & Welling, 2013; Rezende et al., 2014) by maximizing the evidence lower bound (ELBO).

Given this learned world model, we need to derive behaviors from it. Instead of online planning, we use Dreamer (Hafner et al., 2020) to efficiently learn a parametric policy inside the world model that considers long-term rewards. Specifically, we learn two neural networks that operate on latent states of the model. The state-value estimates the sum of future rewards and the actor tries to maximize these predicted values,

$$\text{Actor: } \pi(a_t \mid s_t) \quad \text{Value: } V(s_t). \tag{2}$$

The learned world model is used to predict the sequences of future latent states under the current actor starting from the latent states obtained by encoding images from the replay buffer. The value function is computed at each latent state and the actor policy is trained to maximize the predicted values by propagating their gradients through the neural network dynamics model as shown in Figure 2.

## 3. Planning to Explore

We consider a learning setup with two phases, as illustrated in Figure 1. During self-supervised exploration, the agent gathers information about the environment and summarizes

<sup>1</sup>The latent model states  $s_t$  are not to be confused with the unknown true environment states.

---

### Algorithm 1 Planning to Explore via Latent Disagreement

---

```

1: initialize: Dataset D from a few random episodes.
2:   World model M.
3:   Latent disagreement ensemble E.
4:   Exploration actor-critic  $\pi_{LD}$ .
5: while exploring do
6:   Train M on D.
7:   Train E on D.
8:   Train  $\pi_{LD}$  on LD reward in imagination of M.
9:   Execute  $\pi_{LD}$  in the environment to expand D.
10: end while
11: return Task-agnostic D and M.
    
```

---



---

### Algorithm 2 Zero and Few-Shot Task Adaptation

---

```

1: input: World model M.
2:   Dataset D without rewards.
3:   Reward function R.
4: initialize: Latent-space reward predictor  $\hat{R}$ .
5:   Task actor-critic  $\pi_R$ .
6: while adapting do
7:   Distill R into  $\hat{R}$  for sequences in D.
8:   Train  $\pi_R$  on  $\hat{R}$  in imagination of M.
9:   Execute  $\pi_R$  for the task and report performance.
10:  Optionally, add task-specific episode to D and repeat.
11: end while
12: return Task actor-critic  $\pi_R$ .
    
```

---

this past experience in the form of a parametric world model. After exploration, the agent is given a downstream task in the form of a reward function that it should adapt to with no or limited additional environment interaction.

During exploration, the agent begins by learning a global world model using data collected so far, and then this model is in turn used to direct agent’s exploration to collect more data, as described in Algorithm 1. This is achieved by training an exploration policy inside of the world model to seek out novel states. Novelty is estimated by ensemble disagreement in latent predictions made by 1-step transition models trained alongside the global recurrent world model. More details to follow in Section 3.1.

During adaptation, we can efficiently optimize a task policy by imagination inside of the world model, as shown in Algorithm 2. Since our self-supervised model is trained without being biased toward a specific task, a single trained model can be used to solve multiple downstream tasks.

### 3.1. Latent Disagreement

To efficiently learn a world model of an unknown environment, a successful strategy should explore the environment such as to collect new experience that improves the model

the most. For this, we quantify the model’s uncertainty about its predictions for different latent states. An exploration policy then seeks out states with high uncertainty. The model is then trained on the newly acquired trajectories and reduces its uncertainty in these and the process is repeated.

Quantifying uncertainty is a long-standing open challenge in deep learning (MacKay, 1992; Gal, 2016). In this paper, we use ensemble disagreement as an empirically successful method for quantifying uncertainty (Lakshminarayanan et al., 2017; Osband et al., 2018). As shown in Figure 2, we train a bootstrap ensemble (Breiman, 1996) to predict, from each model state, the next encoder features. The variance of the ensemble serves as an estimate of uncertainty.

Intuitively, because the ensemble models have different initialization and observe data in a different order, their predictions differ for unseen inputs. Once the data is added to the training set, however, the models will converge towards more similar predictions, and the disagreement decreases. Eventually, once the whole environment is explored, the models should converge to identical predictions.

Formally, we define a bootstrap ensemble of one-step predictive models with parameters  $\{w_k \mid k \in [1; K]\}$ . Each of these models takes a model state  $s_t$  and action  $a_t$  as input and predicts the next image embedding  $h_{t+1}$ . The models are trained with the mean squared error, which is equivalent to Gaussian log-likelihood,

$$\begin{aligned} \text{Ensemble predictors: } q(h_{t+1} \mid w_k, s_t, a_t) \\ q(h_{t+1} \mid w_k, s_t, a_t) \triangleq \mathcal{N}(\mu(w_k, s_t, a_t), 1). \end{aligned} \quad (3)$$

We quantify model uncertainty as the variance over predicted means of the different ensemble members and use this disagreement as the intrinsic reward  $ir_t \triangleq D(s_t, a_t)$  to train the exploration policy,

$$\begin{aligned} D(s_t, a_t) &\triangleq \text{Var}(\{\mu(w_k, s_t, a_t) \mid k \in [1; K]\}) \\ &= \frac{1}{K-1} \sum_k (\mu(w_k, s_t, a_t) - \mu')^2, \\ \mu' &\triangleq \frac{1}{K} \sum_k \mu(w_k, s_t, a_t). \end{aligned} \quad (4)$$

The intrinsic reward is non-stationary because the world model and the ensemble predictors change throughout exploration. Indeed, once certain states are visited by the agent and the model gets trained on them, these states will become less interesting for the agent and the intrinsic reward for visiting them will decrease.

We learn the exploration policy using Dreamer (Section 2). Since the intrinsic reward is computed in the compact representation space of the latent dynamics model, we can optimize the learned actor and value from imagined latent trajectories without generating images. This lets us efficiently

optimize the intrinsic reward without additional environment interaction. Furthermore, the ensemble of lightweight 1-step models adds little computational overhead as they are trained together efficiently in parallel across all time steps.

### 3.2. Expected Information Gain

Latent disagreement has an information-theoretic interpretation. This subsection derives our method from the amount of information gained by interacting with the environment, which has its roots in optimal Bayesian experiment design (Lindley, 1956; MacKay, 1992).

Because the true dynamics are unknown, the agent treats the optimal dynamics parameters as a random variable  $w$ . To explore the environment as efficiently as possible, the agent should seek out future states that are informative of our belief over the parameters.

Mutual information formalizes the amount of bits that a future trajectory provides about the optimal model parameters on average. We aim to find a policy that shapes the distribution over future states to maximize the mutual information between the image embeddings  $h_{1:T}$  and parameters  $w$ ,

$$I(h_{t+1}; w \mid s_t, a_t) \quad (5)$$

We operate on latent image embeddings to save computation. To select the most promising data during exploration, the agent maximizes the expected information gain,

$$a_t^* \triangleq \arg \max_{a_t} I(h_{t+1}; w \mid s_t, a_t). \quad (6)$$

This expected information gain can be rewritten as conditional entropy of trajectories subtracted from marginal entropy of trajectories, which correspond to, respectively, the aleatoric and the total uncertainty of the model,

$$\begin{aligned} I(h_{t+1}; w \mid s_t, a_t) \\ = H(h_{t+1} \mid s_t, a_t) - H(h_{t+1} \mid w, s_t, a_t). \end{aligned} \quad (7)$$

We see that the information gain corresponds to the episodic uncertainty, i.e. the reducible uncertainty of the model that is left after subtracting the expected amount of data noise from the total uncertainty.

Trained via squared error, our ensemble members are conditional Gaussians with means produced by neural networks and fixed variances. The ensemble can be seen as a mixture distribution of parameter point masses,

$$p(w) \triangleq \frac{1}{K} \sum_k \delta(w - w_k) \quad (8)$$

$$p(h_{t+1} \mid w_k, s_t, a_t) \triangleq \mathcal{N}(h_{t+1} \mid \mu(w_k, s_t, a_t), \sigma^2).$$

Because the variance is fixed, the conditional entropy does not depend on the state or action in our case ( $D$  is the



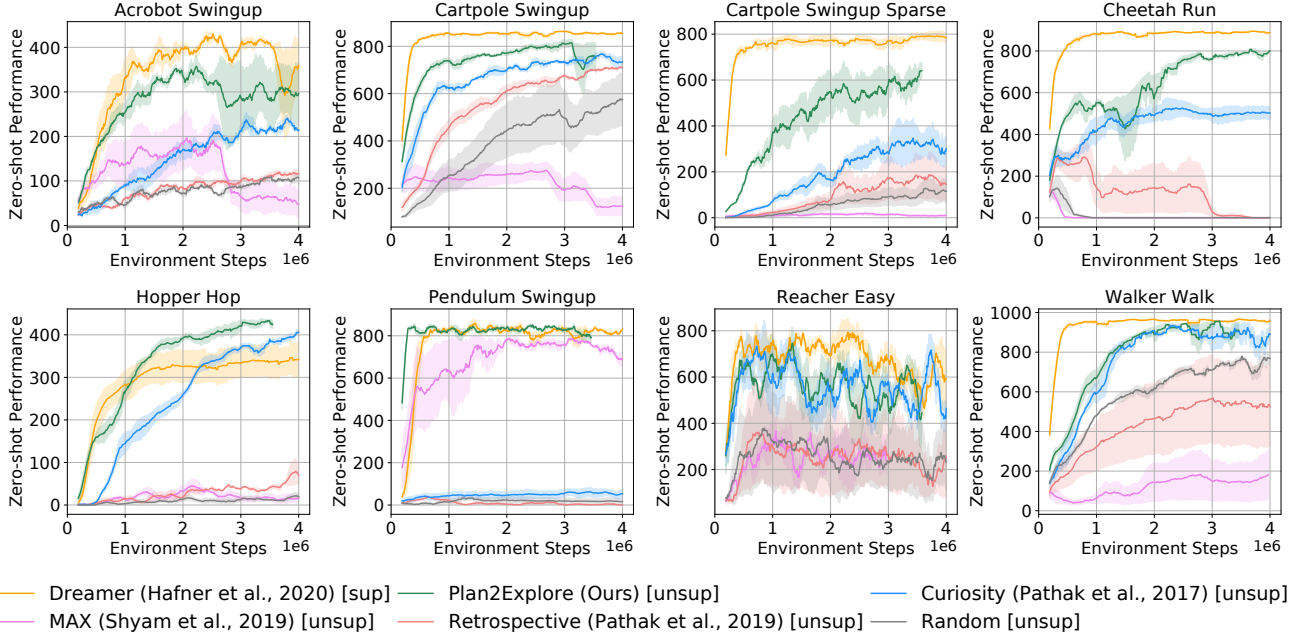


Figure 3. Zero-shot RL performance from raw pixels. After training the agent without rewards, we provide it with a task by specifying the reward function at test time. Throughout the exploration, we take snapshots of the agent to train a task policy on the final task and plot its zero-shot performance. We see that Plan2Explore achieves state-of-the-art zero-shot task performance on a range of tasks, and even demonstrates competitive performance to Dreamer (Hafner et al., 2020), a state-of-the-art supervised RL agent. This indicates that Plan2Explore is able to explore and learn a global model of the environment that is useful for adapting to new tasks, demonstrating the potential of self-supervised RL. Results on all 20 tasks are in the appendix Figure 6 and videos on the project website.

dimensionality of the predicted embedding),

$$\begin{aligned} H(h_{t+1} | w, s_t, a_t) &= \frac{1}{K} \sum_k H(h_{t+1} | w_k, s_t, a_t) \\ &= \frac{D}{K} \sum_k \ln \sigma_k(s_t, a_t) + \text{const.} \end{aligned} \quad (9)$$

We note that this fixed variance approach is applicable even in environments with heteroscedastic variance, where it will measure the information gain about the mean prediction.

Maximizing information gain then means to simply maximize the marginal entropy of the ensemble prediction. For this, we make the following observation: the marginal entropy is maximized when the ensemble means are far apart (disagreement) so the modes overlap the least, maximally spreading out probability mass. As the marginal entropy has no closed-form expression suitable for optimization, we instead use the empirical variance over ensemble means to measure how far apart they are,

$$\begin{aligned} D(s_t, a_t) &\triangleq \frac{1}{K-1} \sum_k (\mu(w_k, s_t, a_t) - \mu')^2, \\ \mu' &\triangleq \frac{1}{K} \sum_k \mu(w_k, s_t, a_t). \end{aligned} \quad (10)$$

To summarize, our exploration objective defined in Section 3.1, which maximizes the variance of ensemble means,

approximates the information gain and thus should find trajectories that will efficiently reduce the model uncertainty.

## 4. Experimental Setup

**Environment Details** We use the DM Control Suite (Tassa et al., 2018), a standard benchmark for continuous control. All experiments use visual observations only, of size  $64 \times 64 \times 3$  pixels. The episode length is 1000 steps and we apply an action repeat of  $R = 2$  for all the tasks. We run every experiment with three different random seeds with standard deviation shown in the shaded region. Further details are in the appendix.

**Implementation** We use (Hafner et al., 2020) with the original hyperparameters unless specified otherwise to optimize both exploration and task policies of Plan2Explore. We found that additional capacity provided by increasing the hidden size of the GRU in the latent dynamics model to 400 and the deterministic and stochastic components of the latent space to 60 helped performance. For a fair comparison, we maintain this model size for Dreamer and other baselines. For latent disagreement, we use an ensemble of 5 one-step prediction models implemented as 2 hidden-layer MLP. Full details are in the appendix.

**Baselines** We compare our agent to a state-of-the-art task-oriented agent that receives rewards throughout training,

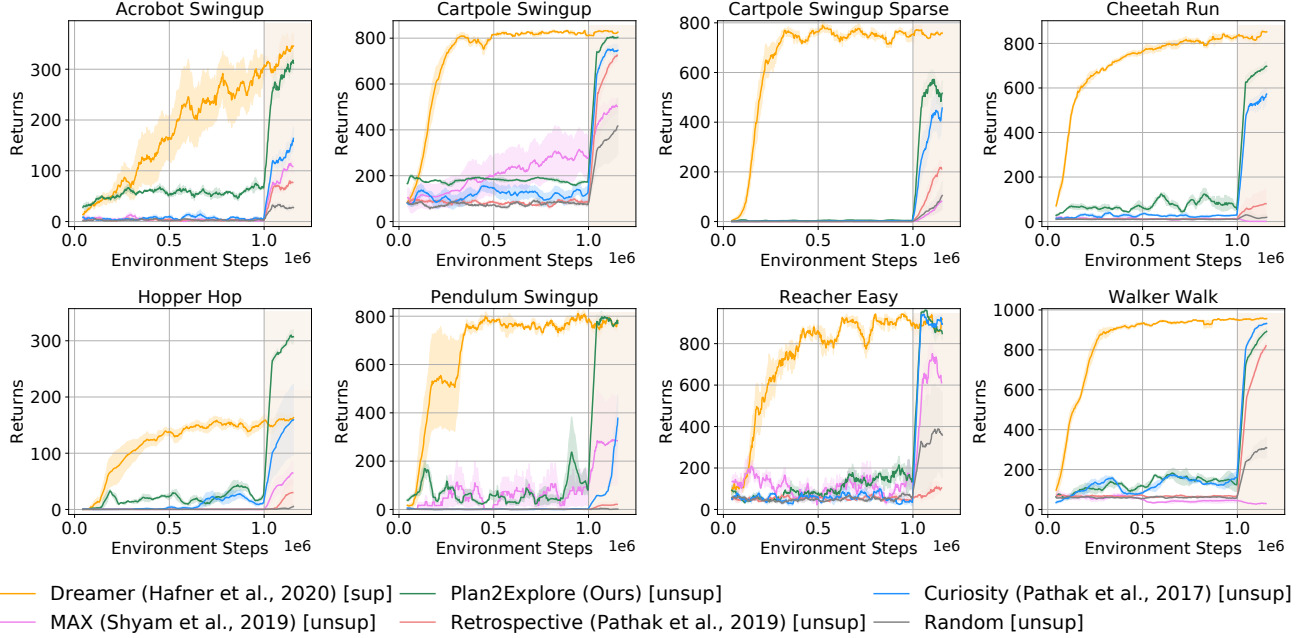


Figure 4. Performance on few-shot adaptation from raw pixels without state-space input. After the exploration phase of 1M steps (white background), during which the agent does not observe the reward and thus does not solve the task, we let the agent collect a small amount of data from the environment (shaded background). We see that Plan2Explore is able to explore the environment efficiently in only 1000 episodes, and then adapt its behavior immediately after observing the reward. Plan2Explore adapts rapidly, producing effective behavior competitive to state-of-the-art supervised reinforcement learning in just a few collected episodes.

Dreamer (Hafner et al., 2020). We also compare to state-of-the-art unsupervised agents: Curiosity (Pathak et al., 2017) and Model-based Active Exploration (Shyam et al., 2019, MAX). Because Curiosity is inefficient during fine-tuning and would not be able to solve a task in a zero-shot way, we adapt it into the model-based setting. We further adapt MAX to work with image observations as (Shyam et al., 2019) only addresses learning from low-dimensional states. We use (Hafner et al., 2020) as the base agent for all methods to provide a fair comparison. We additionally compare to a random data collection policy that uniformly samples from the action space of the environment. All methods share the same model hyperparameters to provide a fair comparison.

## 5. Results and Analysis

Our experiments focus on evaluating whether our proposed Plan2Explore agent efficiently explores and builds a model of the world that allows quick adaptation to solve tasks in zero or few-shot manner. The rest of the subsections are organized in terms of the key scientific questions we would like to investigate as discussed in the introduction.

### 5.1. Does the model transfer to solve tasks zero-shot?

To test whether Plan2Explore has learned a global model of the environment that can be used to solve new tasks, we

evaluate the zero-shot performance of our agent. Our agent learns a model without using any task-specific information. After that, a separate downstream agent is trained in imagination, which optimizes the task reward using only the self-supervised world model and no new interaction with the world. To specify the task, we provide the agent with the reward function that is used to label its replay buffer with rewards and train a reward predictor. This process is described in the Algorithm 2, with step 10 omitted.

In Figure 3, we compare the zero-shot performance of our downstream agent with different amounts of exploration data. This is done by training the downstream agent in imagination at each training checkpoint. The same architecture and hyper-parameters are used for all the methods for a fair comparison. We see that Plan2Explore overall performs better than prior state-of-the-art exploration strategies from high dimensional pixel input, sometimes being the only successful unsupervised method. Moreover, the zero-shot performance of Plan2Explore is competitive to Dreamer, even outperforming it in the hopper hop task.

Plan2Explore was able to successfully learn a good model of the environment and efficiently derive task-oriented behaviors from this model. We emphasize that Plan2Explore explores without task rewards, and Dreamer is the oracle as it is given task rewards during exploration. Yet, Plan2Explore almost matches the performance of this oracle.

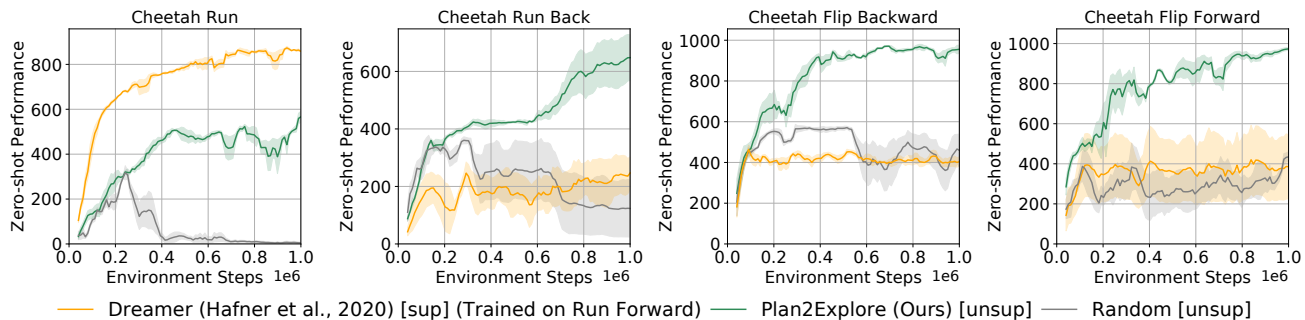


Figure 5. Do task-specific models generalize? We test Plan2Explore on zero-shot performance on four different tasks in the cheetah environment from raw pixels without state-space input. Throughout the exploration, we take snapshots of policy to plot its zero-shot performance. In addition to random exploration, we compare to an oracle agent, Dreamer, that uses the data collected when trained on the run forward task with rewards. Although Dreamer trained on ‘run forward’ is able to solve the task it is trained on, it struggles on the other tasks, indicating that it has not learned a global world model.

## 5.2. How much task-specific interaction is needed for finetuning to reach the supervised oracle?

While zero-shot learning might suffice for some tasks, in general, we will want to adapt our model of the world to task-specific information. In this section, we test whether few-shot adaptation of the model to a particular task is competitive to training a fully supervised task-specific model. To adapt our model, we only add 100–150 supervised episodes which falls under ‘few-shot’ adaptation. Furthermore, in this setup, to evaluate the data efficiency of Plan2Explore we set the number of exploratory episodes to only 1000.

In the exploration phase of Figure 4, i.e., left of the vertical line, our agent does not aim to solve the task, as it is still unknown, however, we expect that during some period of exploration it will *coincidentally* achieve higher rewards as it explores the parts of the state space relevant for the task. The performance of unsupervised methods is coincidental until 1000 episodes and then it switches to task-oriented behavior for the remaining 150 episodes, while for supervised, it is task-oriented throughout. That’s why we see a big jump for unsupervised methods where the shaded region begins.

In the few-shot learning setting, Plan2Explore eventually performs competitively to Dreamer on all tasks, significantly outperforming it on the hopper task. Plan2Explore is also able to adapt quickly or similar to other unsupervised agents on all tasks. These results show that a self-supervised agent, when presented with a task specification, should be able to rapidly adapt its model to the task information, matching or outperforming the fully supervised agent trained only for that task. Moreover, Plan2Explore is able to learn this general model with a small number of samples, matching Dreamer, which is fully task-specific, in data efficiency. This shows the potential of an unsupervised pre-training in reinforcement learning. Please refer to the appendix for detailed quantitative results.

## 5.3. Do self-supervised models generalize better than supervised task-specific models?

If the quality of our learned model is good, it should be transferable to multiple tasks. In this section, we test the quality of the learned model on generalization to multiple tasks in the same environment. We devise a set of three new tasks for the Cheetah environment, specifically, running backward, flipping forward, and flipping backward. We evaluate the zero-shot performance of Plan2Explore and additionally compare it to a Dreamer agent that is only allowed to collect data on the running forward task and then tested on zero-shot performance on the three other tasks.

Figure 5 shows that while Dreamer performs well on the task it is trained on, running forward, it fails to solve all other tasks, performing comparably to random exploration. It even fails to generalize to the running backward task. In contrast, Plan2Explore performs well across all tasks, outperforming Dreamer on the other three tasks. This indicates that the model learned by Plan2Explore is indeed global, while the model learned by Dreamer, which is task-oriented, fails to generalize to different tasks.

## 5.4. What is the advantage of maximizing expected novelty in comparison to retrospective novelty?

Our Plan2Explore agent is able to measure expected novelty by imagining future states that have not been visited yet. A model-free agent, in contrast, is only trained on the states from the replay buffer, and only gets to see the novelty in retrospect, after the state has been visited. Here, we evaluate the advantages of computing expected versus retrospective novelty by comparing Plan2Explore to a one-step planning agent. The one-step planning agent is not able to plan to visit states that are more than one step away from the replay buffer and is somewhat similar to a Q-learning agent with a particular parametrization of the Q-function. We refer to this approach as Retrospective Disagreement. Figures 3 and 4

show the performance of this approach. Our agent achieves superior performance, which is consistent with our intuition about the importance of computing expected novelty.

## 6. Related Work

**Exploration** Exploration is crucial for efficient reinforcement learning (Kakade & Langford, 2002). In tabular settings, it is efficiently addressed with exploration bonuses based on state visitation counts (Strehl & Littman, 2008; Jaksch et al., 2010) or fully Bayesian approaches (Duff & Barto, 2002; Poupart et al., 2006), however, these approaches are hard to generalize to high-dimensional inputs, such as images. Recently methods are able to scale early ideas to high dimensional data via pseudo-count measures of state visitation (Bellemare et al., 2016; Ostrovski et al., 2018). Osband et al. (2016) derived an efficient approximation to the Thompson sampling procedure via ensembles of Q-functions. Osband et al. (2018); Lowrey et al. (2018) use ensembles of Q-functions to track the posterior of the value functions. In contrast to these task-oriented methods, our approach uses neither reward nor state at training time.

**Self-Supervised RL** One way to learn skills without extrinsic rewards is to use intrinsic motivation as sole objective (Oudeyer & Kaplan, 2009). Practical examples of such approaches focus on maximizing prediction error as curiosity bonus (Pathak et al., 2017; Burda et al., 2019; Haber et al., 2018). These approaches can also be understood as maximizing the agent’s surprise (Schmidhuber, 1991a; Achiam & Sastry, 2017). Similar to our work, other recent approaches use the notion of model disagreement to encourage visiting states with the highest potential to improve the model (Burda et al., 2018; Pathak et al., 2019), motivated by the active learning literature (Seung et al., 1992; McCallum & Nigam, 1998). An alternative is to explore by generating goals from prior experience (Andrychowicz et al., 2017; Nair et al., 2018). However, most of these approaches are model-free and expensive to fine-tune to a new task, requiring millions of environment steps for fine-tuning.

**Model-based Control** Early work on model-based reinforcement learning used Gaussian processes and time-varying linear dynamical systems and has shown significant improvements in data efficiency over model-free agents (Kaelbling et al., 1996; Deisenroth & Rasmussen, 2011; Levine & Koltun, 2013) when low-dimensional state information is available. Recent work on latent dynamics models has shown that model-based agents can achieve performance competitive with model-free agents while attaining much higher data efficiency, and even scale to high-dimensional observations (Chua et al., 2018; Buesing et al., 2018; Ebert et al., 2018; Ha & Schmidhuber, 2018; Hafner et al., 2019; Nagabandi et al., 2019). We base our agent on a state-of-the-art model-based agent, Dreamer (Hafner et al.,

2020), and use it to perform self-supervised exploration in order to solve tasks in a few-shot manner.

Certain prior work has considered model-based exploration (Amos et al., 2018; Sharma et al., 2019), but was not shown to scale to complex visual observations, only using proprioceptive information. Other work (Ebert et al., 2018; Pathak et al., 2018) has demonstrated the possibility of self-supervised model-based learning with visual observations. However, these approaches do not integrate exploration and model learning together, instead of performing them in stages (Pathak et al., 2018) or just using random exploration (Ebert et al., 2018), which makes them difficult to scale to long-horizon problems.

The idea of actively exploring to collect the most informative data goes back to the formulation of the information gain (Lindley, 1956). MacKay (1992) described how a learning system might optimize Bayesian objectives for active data selection based on the information gain. Sun et al. (2011) derived a model-based reinforcement learning agent that can optimize the infinite-horizon information gain and experimented with it in tabular settings. The closest works to ours are Shyam et al. (2019); Henaff (2019), which use a measurement of disagreement or information gain through ensembles of neural networks in order to incentivize exploration. However, these approaches are restricted to setups where low-dimensional states are available, whereas we design a latent state approach that scales to high-dimensional observations. Moreover, we provide a theoretical connection between information gain and model disagreement. Concurrent with us, Ball et al. (2020) discuss the connection between information gain and model disagreement for task-specific exploration from low-dimensional state space.

## 7. Discussion

We presented Plan2Explore, a self-supervised reinforcement learning method that learns a world model of its environment through unsupervised exploration and uses this model to solve tasks in a zero-shot or few-shot manner. We derived connections of our method to the expected information gain, a principled objective for exploration. Building on recent work on learning dynamics models and behaviors from images, we constructed a model-based zero-shot reinforcement learning agent that was able to achieve state-of-the-art zero-shot task performance on the DeepMind Control Suite. Moreover, the agent’s zero-shot performance was competitive to Dreamer, a state-of-the-art supervised reinforcement learning agent on some tasks, with the few-shot performance eventually matching or outperforming the supervised agent. By presenting a method that can learn effective behavior for many different tasks in a scalable and data-efficient manner, we hope this work constitutes a step toward building scalable real-world reinforcement learning systems.



**Acknowledgements** We thank Rowan McAllister, Aviral Kumar, Vijay Balasubramanian, and the members of GRASP for fruitful discussions. This work was supported in part by NSF-IIS-1703319, ONR N00014-17-1-2093, ARL DCIST CRA W911NF-17-2-0181, Curious Minded Machines grant from Honda Research and DARPA Machine Common Sense grant.

## References

- Achiam, J. and Sastry, S. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv:1703.01732*, 2017. 8
- Amos, B., Dinh, L., Cabi, S., Rothrl, T., Muldal, A., Erez, T., Tassa, Y., de Freitas, N., and Denil, M. Learning awareness models. In *ICLR*, 2018. 8
- Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., and Zaremba, W. Hindsight experience replay. In *NIPS*, 2017. 8
- Ball, P., Parker-Holder, J., Pacchiano, A., Choromanski, K., and Roberts, S. Ready policy one: World building through active learning. *arXiv preprint arXiv:2002.02693*, 2020. 8
- Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Sutton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. In *NIPS*, 2016. 1, 8
- Breiman, L. Bagging predictors. *Machine learning*, 24(2): 123–140, 1996. 4
- Buesing, L., Weber, T., Racaniere, S., Eslami, S., Rezende, D., Reichert, D. P., Viola, F., Besse, F., Gregor, K., Hassabis, D., et al. Learning and querying fast generative models for reinforcement learning. *arXiv preprint arXiv:1802.03006*, 2018. 8
- Burda, Y., Edwards, H., Storkey, A., and Klimov, O. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018. 1, 8
- Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A. Large-scale study of curiosity-driven learning. *ICLR*, 2019. 8
- Chua, K., Calandra, R., McAllister, R., and Levine, S. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *arXiv preprint arXiv:1805.12114*, 2018. 8
- Deisenroth, M. and Rasmussen, C. E. Pilco: A model-based and data-efficient approach to policy search. In *ICML*, 2011. 8
- Duff, M. O. and Barto, A. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, University of Massachusetts at Amherst, 2002. 8
- Ebert, F., Finn, C., Dasari, S., Xie, A., Lee, A., and Levine, S. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv:1812.00568*, 2018. 8
- Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. Diversity is all you need: Learning skills without a reward function. *arXiv:1802.06070*, 2018. 1
- Gal, Y. Uncertainty in deep learning. *University of Cambridge*, 1:3, 2016. 4
- Ha, D. and Schmidhuber, J. World models. *arXiv preprint arXiv:1803.10122*, 2018. 3, 8
- Haber, N., Mrowca, D., Wang, S., Fei-Fei, L. F., and Yamins, D. L. Learning to play with intrinsically-motivated, self-aware agents. In *NeurIPS*, 2018. 8
- Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., and Davidson, J. Learning latent dynamics for planning from pixels. *ICML*, 2019. 2, 3, 8, 11
- Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination. *ICLR*, 2020. 3, 5, 6, 8, 13
- Henaff, M. Explicit explore-exploit algorithms in continuous state spaces. In *NeurIPS*, 2019. 8
- Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *JMLR*, 2010. 8
- Kaelbling, L. P., Littman, M. L., and Moore, A. W. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 1996. 8
- Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pp. 267–274, 2002. 8
- Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3
- Klyubin, A. S., Polani, D., and Nehaniv, C. L. Empowerment: A universal agent-centric measure of control. In *Evolutionary Computation*, 2005. 1
- Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, 2017. 4
- Lehman, J. and Stanley, K. O. Evolving a diversity of virtual creatures through novelty search and local competition. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, 2011. 1

- Levine, S. and Koltun, V. Guided policy search. In *International Conference on Machine Learning*, pp. 1–9, 2013. 8
- Lindley, D. V. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pp. 986–1005, 1956. 2, 4, 8
- Lowrey, K., Rajeswaran, A., Kakade, S., Todorov, E., and Mordatch, I. Plan online, learn offline: Efficient learning and exploration via model-based control. *arXiv preprint arXiv:1811.01848*, 2018. 8
- MacKay, D. J. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992. 4, 8
- McCallumzy, A. K. and Nigamy, K. Employing em and pool-based active learning for text classification. In *ICML*, 1998. 8
- Nagabandi, A., Konoglie, K., Levine, S., and Kumar, V. Deep dynamics models for learning dexterous manipulation. *arXiv preprint arXiv:1909.11652*, 2019. 8
- Nair, A. V., Pong, V., Dalal, M., Bahl, S., Lin, S., and Levine, S. Visual reinforcement learning with imagined goals. In *NeurIPS*, 2018. 8
- Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. Deep exploration via bootstrapped dqn. In *NIPS*, 2016. 8
- Osband, I., Aslanides, J., and Cassirer, A. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 8617–8629, 2018. 4, 8
- Ostrovski, G., Bellemare, M. G., Oord, A. v. d., and Munos, R. Count-based exploration with neural density models. *ICML*, 2018. 8
- Oudeyer, P.-Y. and Kaplan, F. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 2009. 8
- Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. Intrinsic motivation systems for autonomous mental development. *Evolutionary Computation*, 2007. 1
- Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017. 1, 2, 6, 8, 11
- Pathak, D., Mahmoudieh, P., Luo, G., Agrawal, P., Chen, D., Shentu, Y., Shelhamer, E., Malik, J., Efros, A. A., and Darrell, T. Zero-shot visual imitation. In *ICLR*, 2018. 8
- Pathak, D., Gandhi, D., and Gupta, A. Self-supervised exploration via disagreement. *ICML*, 2019. 2, 8
- Poupart, P., Vlassis, N., Hoey, J., and Regan, K. An analytic solution to discrete bayesian reinforcement learning. In *ICML*, 2006. 1, 8
- Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014. 3
- Schmidhuber, J. Curious model-building control systems. In *Neural Networks, 1991. 1991 IEEE International Joint Conference on*, pp. 1458–1463. IEEE, 1991a. 8
- Schmidhuber, J. A possibility for implementing curiosity and boredom in model-building neural controllers. In *From animals to animats: Proceedings of the first international conference on simulation of adaptive behavior*, 1991b. 1
- Seung, H., Oppen, M., and Sompolinsky, H. Query by committee. *COLT*, 1992. 8
- Sharma, A., Gu, S., Levine, S., Kumar, V., and Hausman, K. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019. 8
- Shyam, P., Jaśkowski, W., and Gomez, F. Model-Based Active Exploration. In *ICML*, 2019. 2, 6, 8
- Strehl, A. and Littman, M. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 2008. 8
- Sun, Y., Gomez, F., and Schmidhuber, J. Planning to be surprised: Optimal bayesian exploration in dynamic environments. In *AGI*, 2011. 2, 8
- Sutton, R. S. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4): 160–163, 1991. 3
- Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., de Las Casas, D., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., Lillicrap, T., and Riedmiller, M. DeepMind control suite. Technical report, DeepMind, January 2018. URL <https://arxiv.org/abs/1801.00690>. 5, 11
- Watter, M., Springenberg, J., Boedecker, J., and Riedmiller, M. Embed to control: A locally linear latent dynamics model for control from raw images. In *NIPS*, 2015. 3
- Zhang, M., Vikram, S., Smith, L., Abbeel, P., Johnson, M., and Levine, S. Solar: deep structured representations for model-based reinforcement learning. In *ICML*, 2019. 2

## A. Appendix

**Results DM Control Suite** In Figure 6, we show the performance of our agent on all 20 DM Control Suite tasks from pixels. In addition, we show videos corresponding to all the plots on the project website: <https://ramanans1.github.io/plan2explore/>

**Convention for plots** We run every experiment with three different random seeds. The shaded area of the graphs shows the standard deviation in performance. All plot curves are smoothed with a moving mean that takes into account a window of the past 20 data points. Only Figure 5 was smoothed with a window of past 5 data points so as to provide cleaner looking plots that indicate the general trend. Low variance in all the curves consistently across all figures suggests that our approach is very reproducible.

**Rewards of new tasks** To test the generalization performance of our agent, we define three new tasks in the Cheetah environment:

- **Cheetah Run Backward** Analogous to the forward running task, the reward  $r$  is linearly proportional to the backward velocity  $v_b$  up to a maximum of 10m/s, which means  $r(v_b) = \max(0, \min(v_b/10, 1))$ , where  $v_b = -v$  and  $v$  is the forward velocity of the Cheetah.
- **Cheetah Flip Backward** The reward  $r$  is linearly proportional to the backward angular velocity  $\omega_b$  up to a maximum of 5rad/s, which means  $r(\omega_b) = \max(0, \min(\omega_b/5, 1))$ , where  $\omega_b = -\omega$  and  $\omega$  is the angular velocity about the positive Z-axis, as defined in DeepMind Control Suite.
- **Cheetah Flip Forward** The reward  $r$  is linearly proportional to the forward angular velocity  $\omega$  up to a maximum of 5rad/s, which means  $r(\omega) = \max(0, \min(\omega/5, 1))$ .

**Environment** We use the DeepMind Control Suite (Tassa et al., 2018) tasks, a standard benchmark of tasks for continuous control agents. All experiments are performed with only visual observations. We use RGB visual observations with  $64 \times 64$  resolution. We have selected a diverse set of 8 tasks that feature sparse rewards, high dimensional action spaces, and environments with unstable equilibria and environments that require a long planning horizon. We use episode length of 1000 steps and a fixed action repeat of  $R = 2$  for all the tasks.

**Agent implementation** For implementing latent disagreement, we use an ensemble of 5 one-step prediction models with a 2 hidden-layer MLP, which takes in the RNN-state of RSSM and the action as inputs, and predicts the encoder

features, which have a dimension of 1024. We scale the disagreement of the predictions by 10,000 for the final intrinsic reward, this was found to increase performance in some environments. We do not normalize the rewards, both extrinsic and intrinsic. This setup for the one-step model was chosen over 3 other variants, in which we tried predicting the deterministic, stochastic, and the combined features of RSSM respectively. The performance benefits of this ensemble over the variants potentially come from the large parametrization that comes with predicting the large encoder features.

**Baselines** We note that while Curiosity (Pathak et al., 2017) uses  $L_2$  loss to train the model, the RSSM loss is different (see (Hafner et al., 2019)); we use the full RSSM loss as the intrinsic reward for the Curiosity comparison, as we found it produces the best performance. Note that this reward can only be computed when ground truth data is available and needs a separate reward predictor to optimize it in a model-based fashion.

Table 1. Zero-shot performance at 3.5 million environment steps (corresponding to 1.75 agent steps times 2 for action repeat). We report the average performance of the last 20 episodes before the 3.5 million steps point. The performance is computed by executing the mode of the actor without action noise. Among the agents that receive no task rewards, the highest performance of each task is highlighted. The corresponding training curves are visualized in Figure 6.

Zero-shot performance	Plan2Explore	Curiosity	Random	MAX	Retrospective	Dreamer
Task-agnostic experience	3.5M	3.5M	3.5M	3.5M	3.5M	—
Task-specific experience	—	—	—	—	—	3.5M
Acrobot Swingup	<b>280.23</b>	219.55	107.38	64.30	110.84	408.27
Cartpole Balance	950.97	917.10	<b>963.40</b>	—	—	970.28
Cartpole Balance Sparse	<b>860.38</b>	695.83	764.48	—	—	926.9
Cartpole Swingup	<b>759.65</b>	747.488	516.04	144.05	700.59	855.55
Cartpole Swingup Sparse	<b>602.71</b>	324.5	94.89	9.23	180.85	789.79
Cheetah Run	<b>784.45</b>	495.55	0.78	0.76	9.11	888.84
Cup Catch	<b>962.81</b>	<b>963.13</b>	660.35	—	—	963.4
Finger Spin	655.4	661.96	<b>676.5</b>	—	—	333.73
Finger Turn Easy	401.64	266.96	<b>495.21</b>	—	—	551.31
Finger Turn Hard	270.83	289.65	<b>464.01</b>	—	—	435.56
Hopper Hop	<b>432.58</b>	389.64	12.11	17.39	41.32	336.57
Hopper Stand	841.53	<b>889.87</b>	180.86	—	—	923.74
Pendulum Swingup	<b>792.71</b>	56.80	16.96	748.53	1.383	829.21
Quadruped Run	<b>223.96</b>	164.02	139.53	—	—	373.25
Quadruped Walk	182.87	<b>368.45</b>	129.73	—	—	921.25
Reacher Easy	<b>530.56</b>	416.31	229.23	242.13	230.68	544.15
Reacher Hard	66.76	<b>123.5</b>	4.10	—	—	438.34
Walker Run	429.30	<b>446.45</b>	318.61	—	—	783.95
Walker Stand	331.20	<b>459.29</b>	301.65	—	—	655.80
Walker Walk	<b>911.04</b>	889.17	766.41	148.02	538.84	965.51
Task Average	<b>563.58</b>	489.26	342.11	—	—	694.77

Table 2. Adaptation performance after 1M task-agnostic environment steps, followed by 150K task-specific environment steps (agent steps are half as much due to the action repeat of 2). We report the average performance of the last 20 episodes before the 1.15M steps point. The performance is computed by executing the mode of the actor without action noise. Among the self-supervised agents, the highest performance of each task is highlighted. The corresponding training curves are visualized in Figure 4.

Adaptation performance	Plan2Explore	Curiosity	Random	MAX	Retrospective	Dreamer
Task-agnostic experience	1M	1M	1M	1M	1M	—
Task-specific experience	150K	150K	150K	150K	150K	1.15M
Acrobot Swingup	<b>312.03</b>	163.71	27.54	108.39	76.92	345.51
Cartpole Swingup	<b>803.53</b>	747.10	416.82	501.93	725.81	826.07
Cartpole Swingup Sparse	<b>516.56</b>	456.8	104.88	82.06	211.81	758.45
Cheetah Run	<b>697.80</b>	572.67	18.91	0.76	79.90	852.03
Hopper Hop	<b>307.16</b>	159.45	5.21	64.95	29.97	163.32
Pendulum Swingup	<b>771.51</b>	377.51	1.45	284.53	21.23	781.36
Reacher Easy	848.65	<b>894.29</b>	358.56	611.65	104.03	918.86
Walker Walk	892.63	<b>932.03</b>	308.51	29.39	820.54	956.53
Task Average	<b>643.73</b>	537.95	155.23	210.46	258.78	700.27



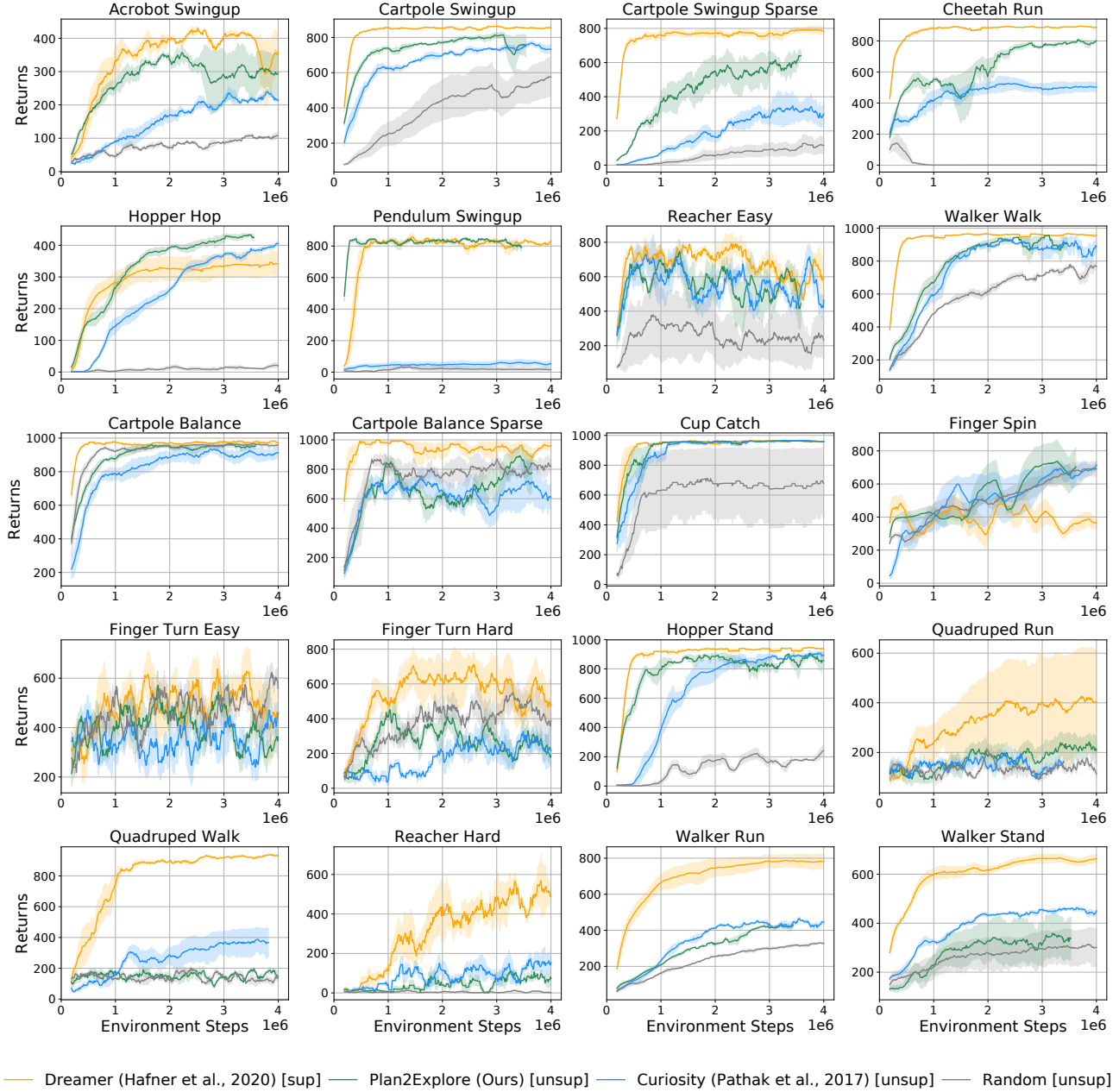


Figure 6. We evaluate the zero-shot performance of the self-supervised agents as well as supervised performance of Dreamer on all tasks from the DM control suite. All agents operate from raw pixels. The experimental protocol is the same as in Figure 3 of the main paper. To produce this plot, we take snapshots of the agent throughout exploration to train a task policy on the downstream task and plot its zero-shot performance. We use the same hyperparameters for all environments. We see that Plan2Explore achieves state-of-the-art zero-shot task performance on a range of tasks. Moreover, even though Plan2Explore is a self-supervised agent, it demonstrates competitive performance to Dreamer (Hafner et al., 2020), a state-of-the-art supervised reinforcement learning agent. This shows that self-supervised exploration is competitive to task-specific approaches in these continuous control tasks.