Dream to Control
Learning Behaviors by Latent Imagination

Danijar Hafner[12], Timothy Lillicrap[3], Jimmy Ba[2], Mohammad Norouzi[1]
[1] Google Brain   [2] University of Toronto   [3] DeepMind

1 We introduce Dreamer

1. We present Dreamer, an RL agent that solves visual control tasks purely by latent imagination inside of a world model.

2. To efficiently learn behaviors, we backpropagate multi-step value gradients through imagined latent transitions.

3. Dreamer learns off-policy and is applicable to both continuous and discrete actions and early episode ends.

4. Dreamer exceeds top model-free agents in final performance, data-efficiency, and wall-clock time.

2 Agent Overview

(a) Learn dynamics from experience   (b) Learn behavior in imagination   (c) Act in the environment

Algorithm 1: Dreamer

Initialize dataset with a few random episodes.
while *not converged* **do**
 for *number of gradient steps* **do**
  Draw sub sequences from dataset at random.
  Update world model using representation learning.
  Use action model to imagine trajectories from all latent states.
  Update value model for Bellman consistency on imagined rewards.
  Update action model to maximize imagined values.
 Collect episode to grow the dataset using the action model.

3 Learning Latent Dynamics

World models are an explicit way to summarize an agent's experience and represent its knowledge about the world.

When inputs are large, predicting forward in a compact latent space lets us imagine thousands of trajectories in parallel.

Dreamer is compatible with any representation objective, e.g. reconstruction, contrastive estimation, or reward prediction.

- encode images
- compute states
- predict rewards
- reconstruction

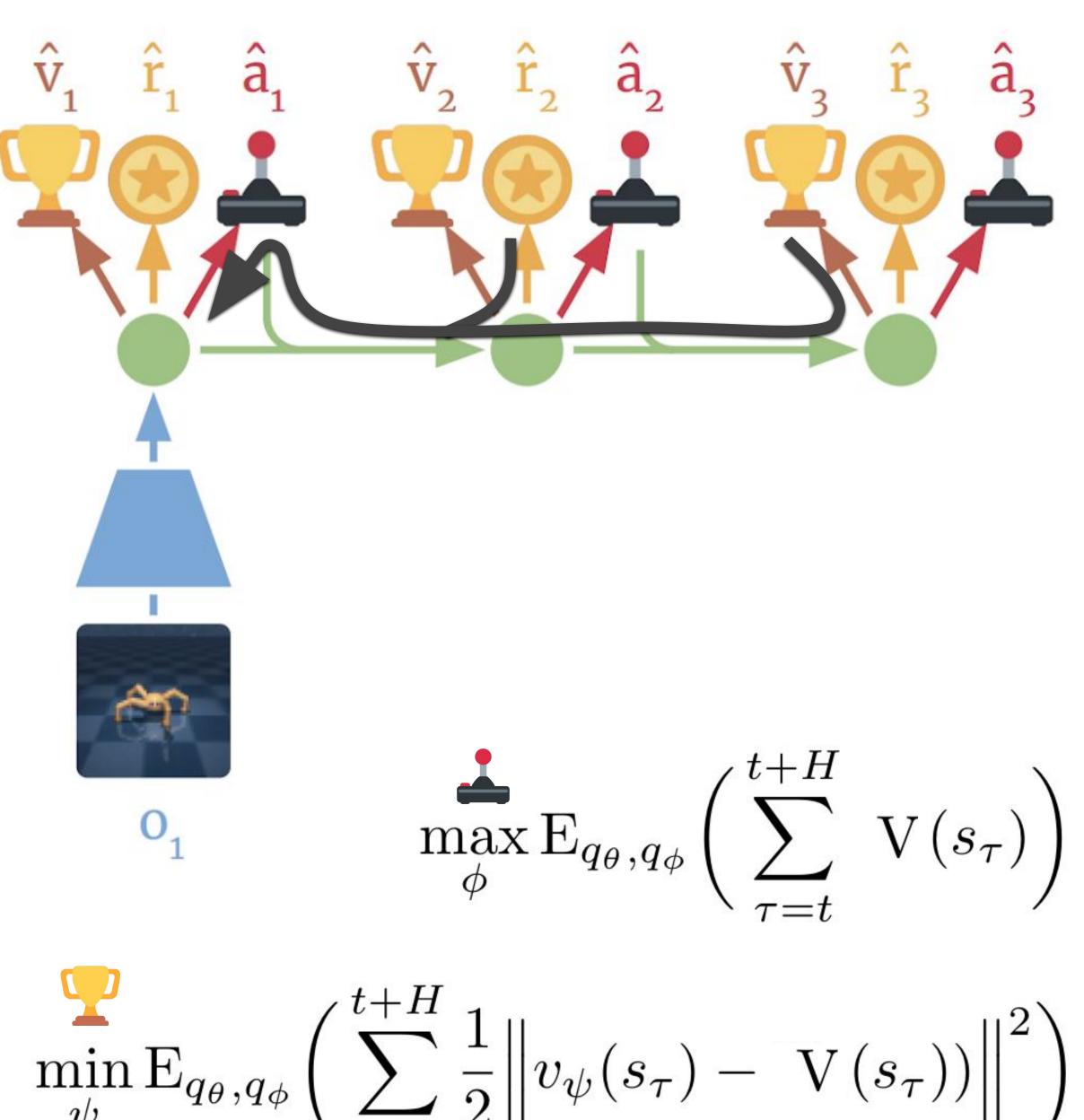## 4 Learning Behaviors in Imagination

Dreamer learns to predict actions and values from the compact model states.

Start imagined trajectory from each latent state of the current data batch.

Value model optimizes Bellman consistency for current action model.

Action model maximizes values by backpropagating through latent transitions.
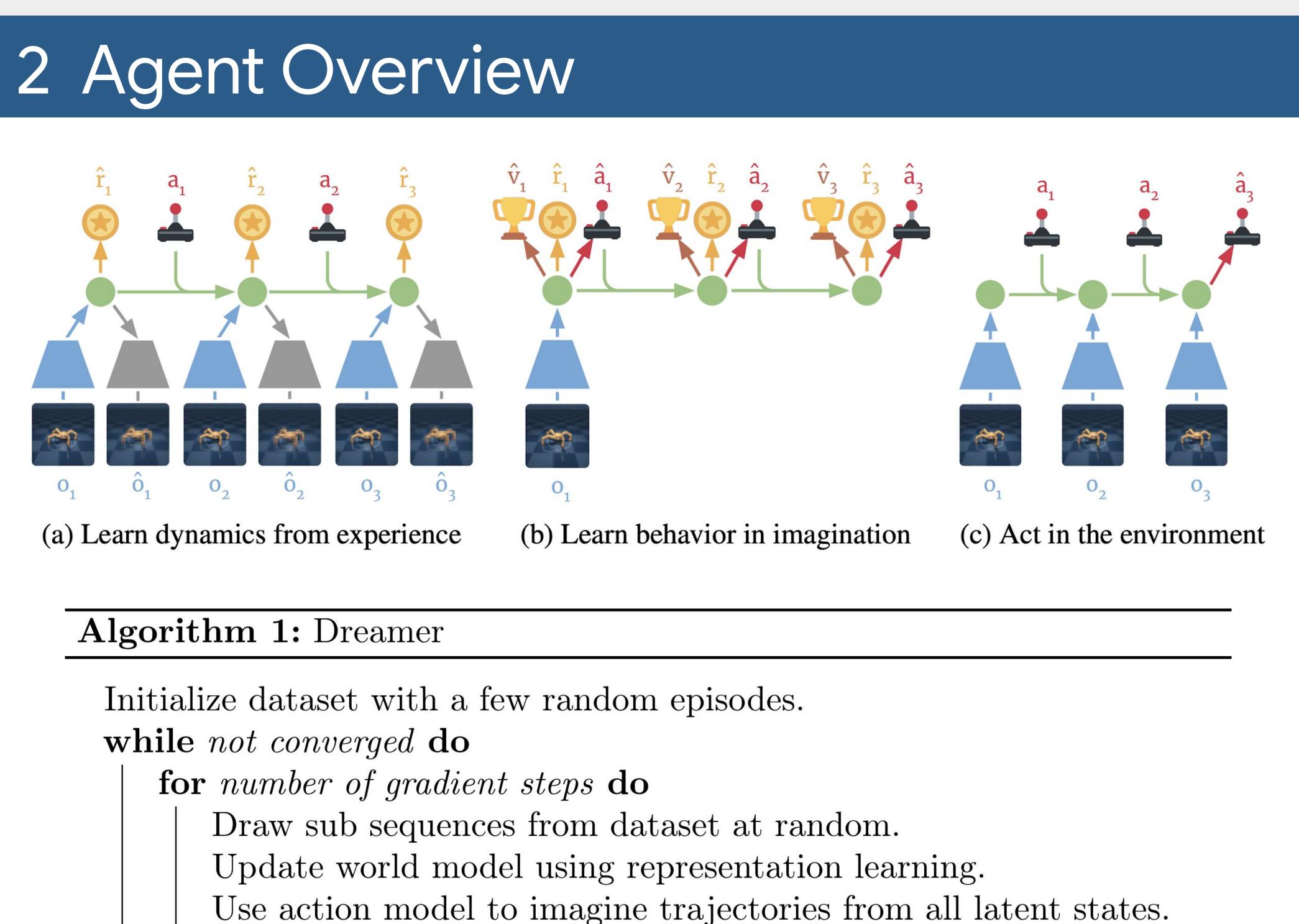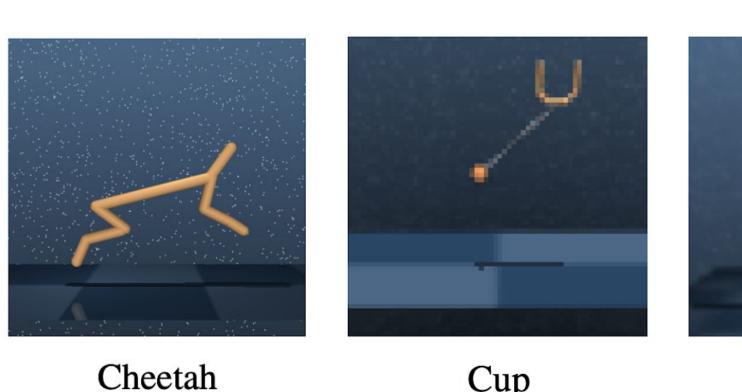
Reparam. gradients for latent states and actions, ST for discrete actions.

$$\max_{\phi} E_{q_\theta, q_\phi} \left(\sum_{\tau=t}^{t+H} V(s_\tau) \right)$$

$$\min_{\psi} E_{q_\theta, q_\phi} \left(\sum_{\tau=t}^{t+H} \frac{1}{2} \left\| v_\psi(s_\tau) - V(s_\tau) \right\|^2 \right)$$

$$V(s_\tau) \overset{e.g.}{\doteq} r(s_\tau) + \gamma r(s_{\tau+1}) + \gamma^2 v_\psi(s_{\tau+2})$$

5 Comparison to Existing Methods

Cheetah   Cup   Acrobot   Hopper   Walker   Quadruped
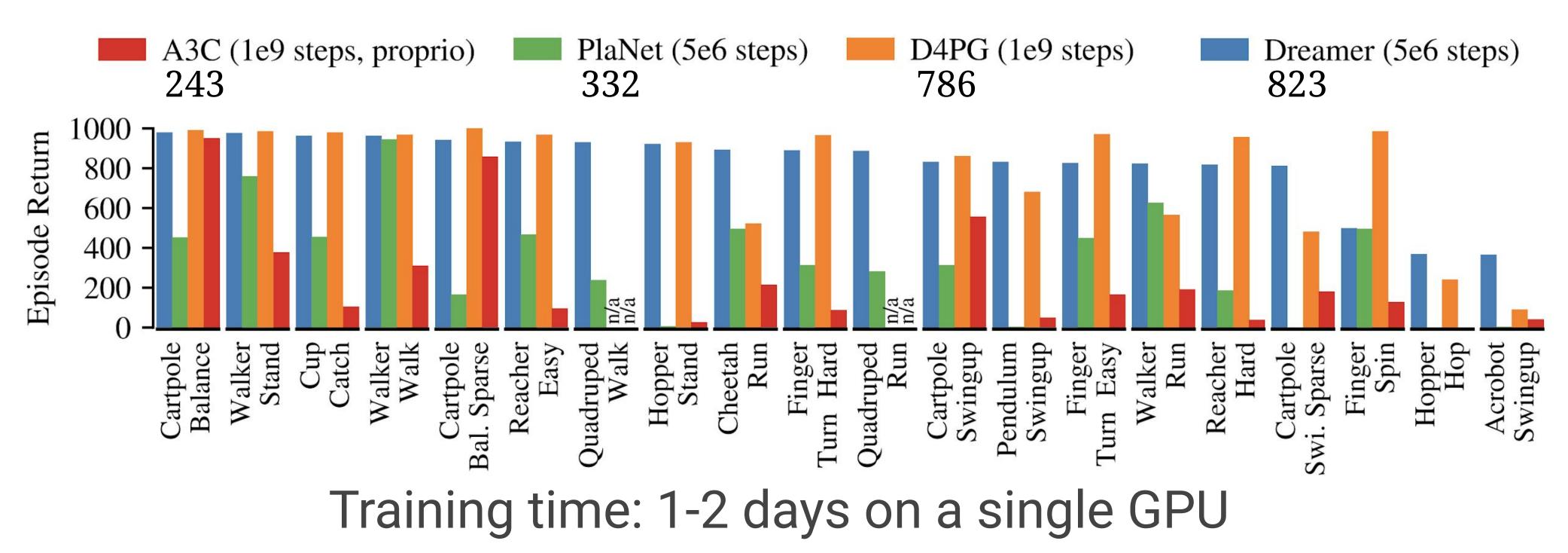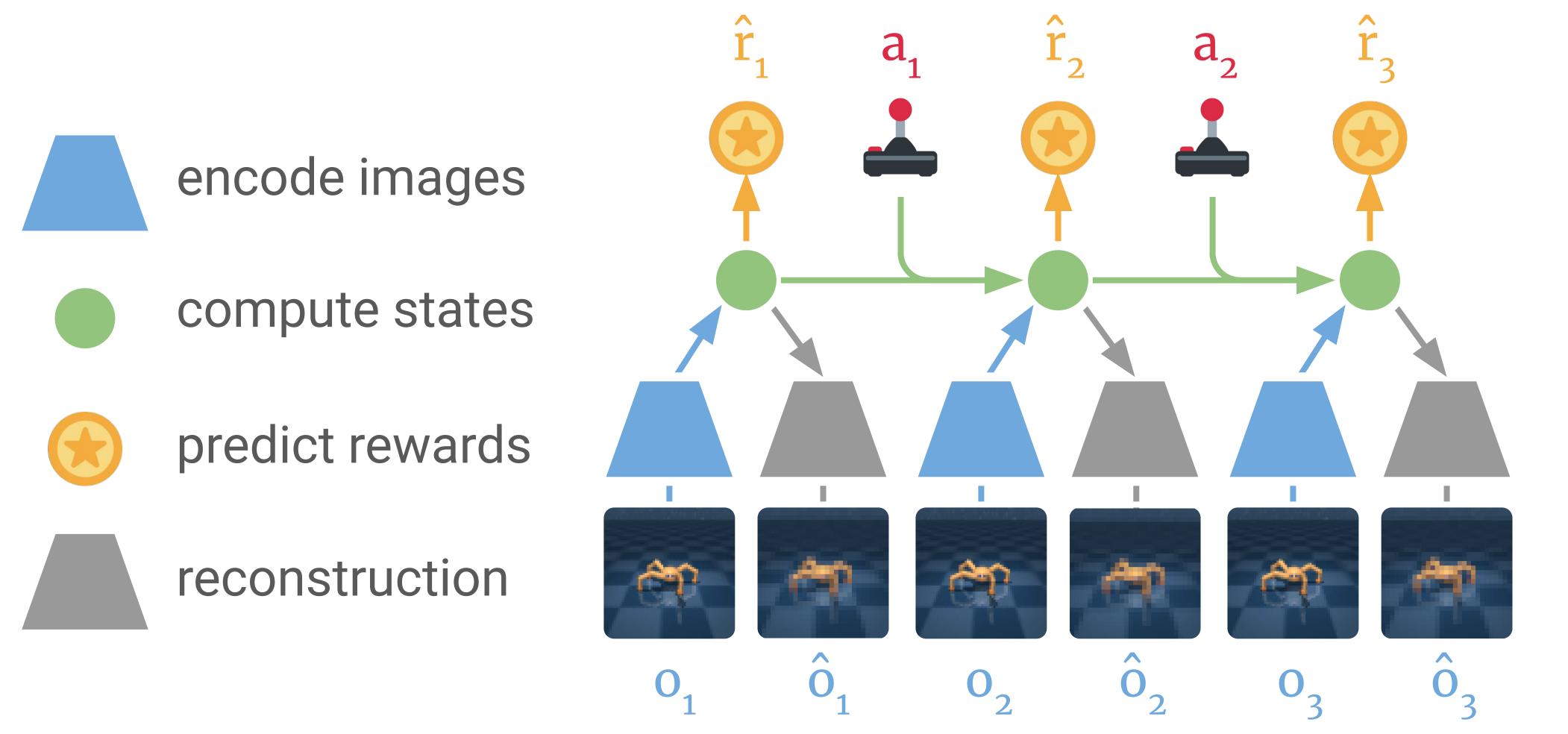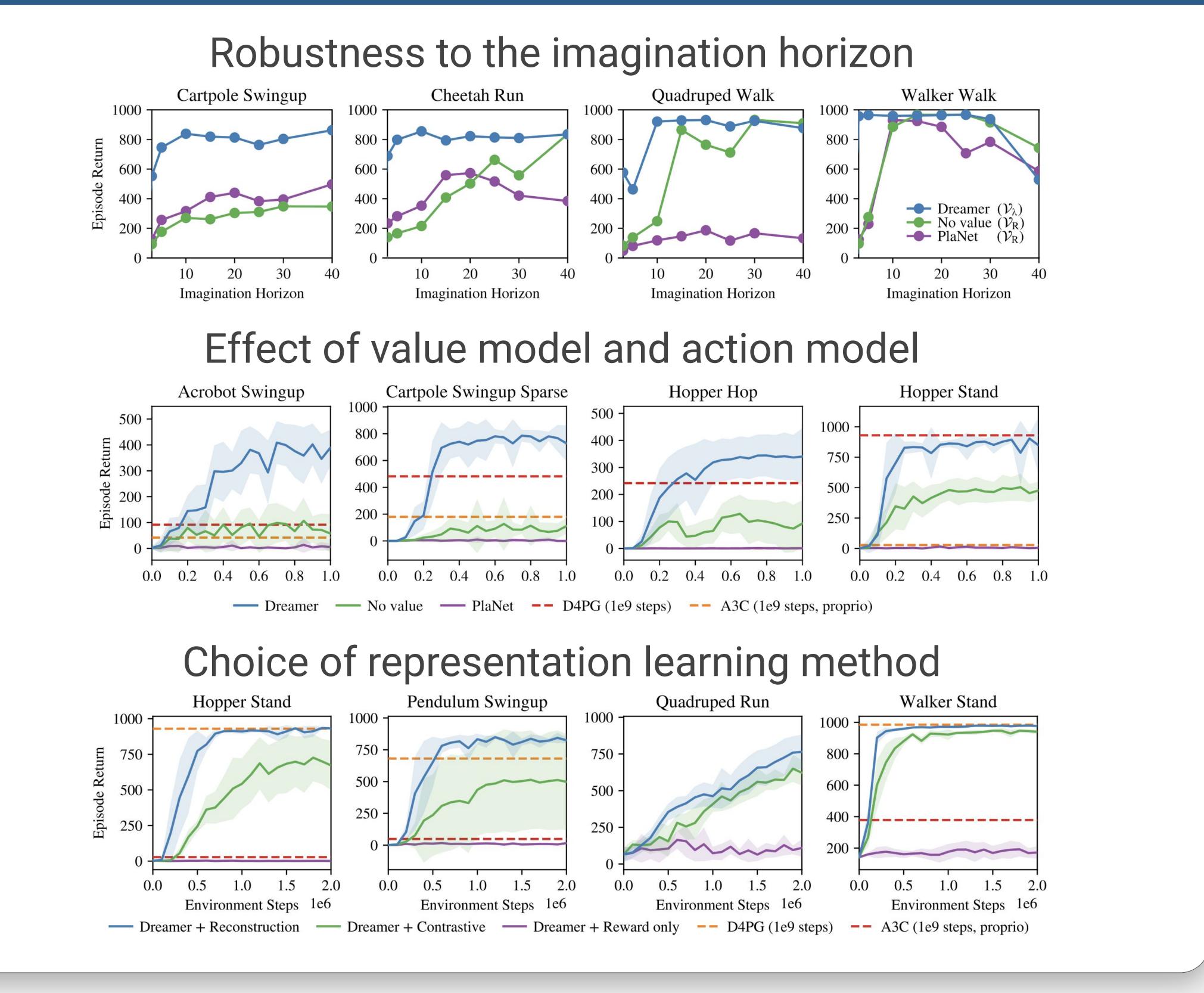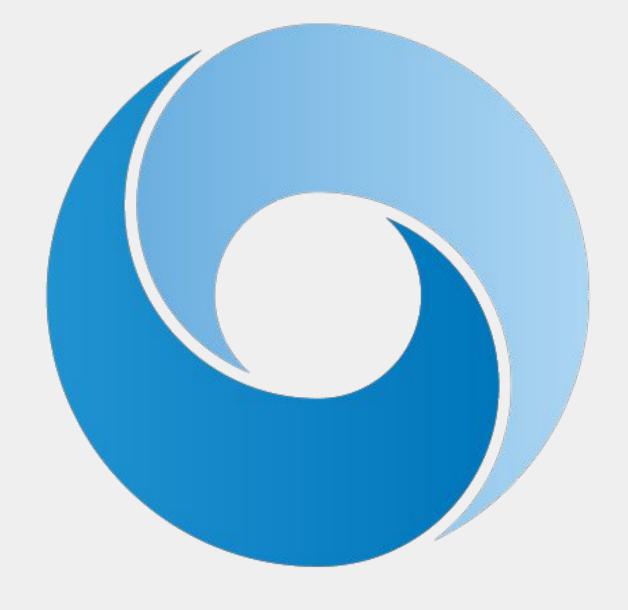
20 <u>visual control tasks</u> posing challenges incl. contact dynamics, sparse rewards, many degrees of freedom, and 3D environments.

Dreamer outperforms the previous best D4PG agent in final performance, sample-efficiency, and wall-clock time. We use the same hyper parameters across all tasks.

A3C (1e9 steps, proprio) 243   PlaNet (5e6 steps) 332   D4PG (1e9 steps) 786   Dreamer (5e6 steps) 823

Training time: 1-2 days on a single GPU

6 Ablation Studies

Robustness to the imagination horizon

Effect of value model and action model

Choice of representation learning method