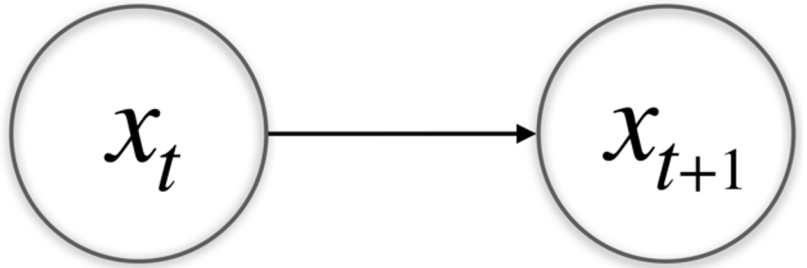
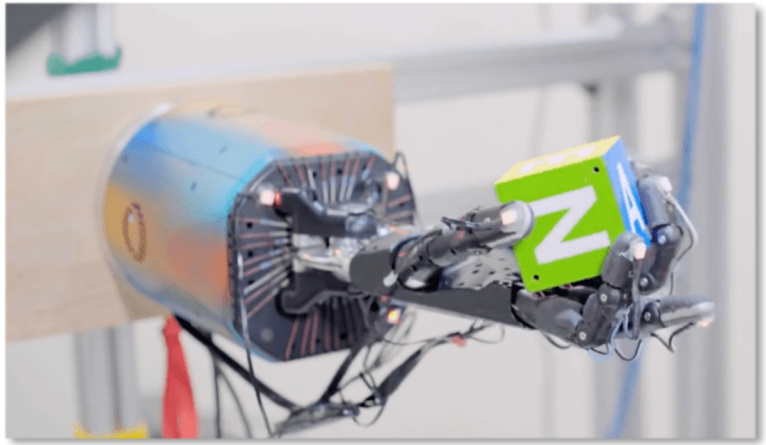
Visual Model-Based Reinforcement Learning

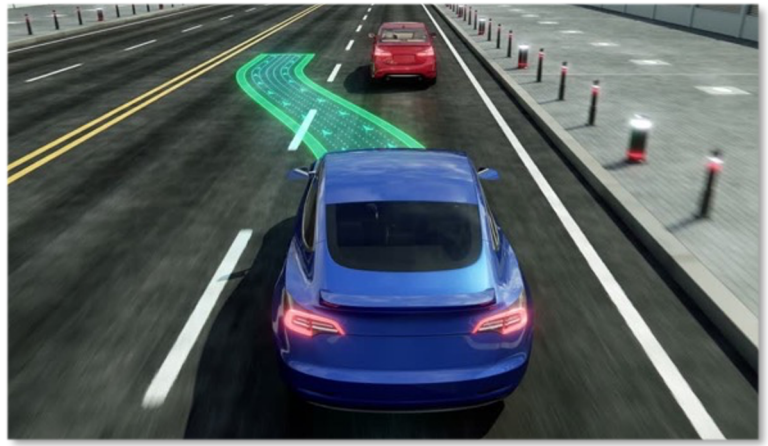


Oleh Rybkin
oleh@seas.upenn.edu

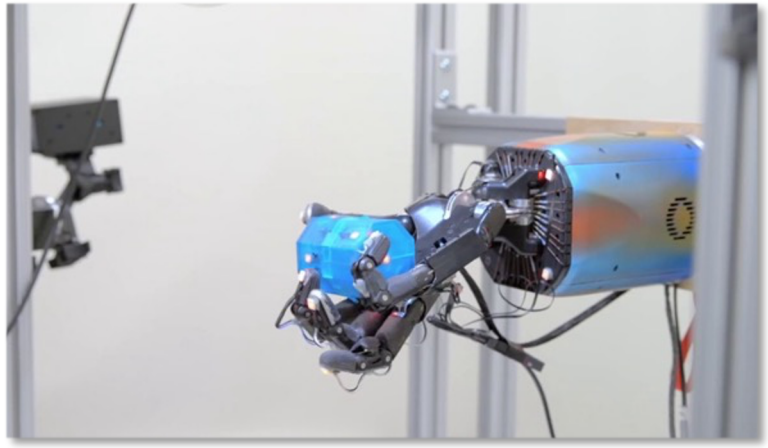
Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

How do we build intelligent machines?



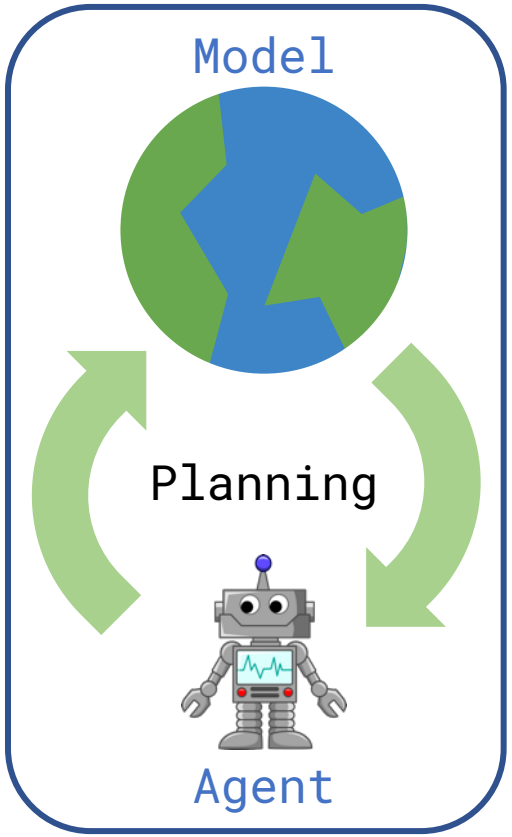


vs.



Environment

$p_\theta(o_{t+1}|o_{0:t}, a_{0:t})$



$$\arg\max_{a_{0:T}} \mathbb{E}_{\prod_{t=0}^{T} p_\theta(o_{t+1}|o_{0:t},a_{0:t})} \left[\sum_t r(o_t) \right]$$

Data Collection

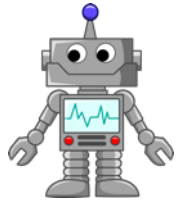
Environment

Model Learning

$$\arg\max_{\theta} \prod_{t=0}^{T} p_\theta(o_{t+1}|o_{0:t}, a_{0:t})$$

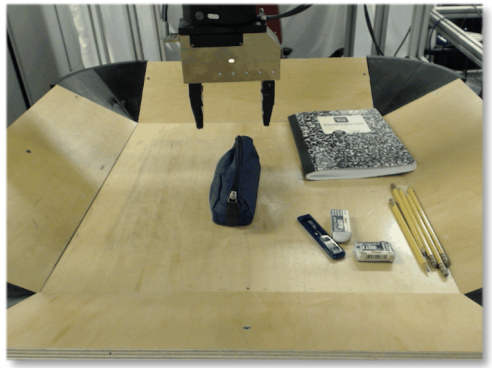
Model

Planning

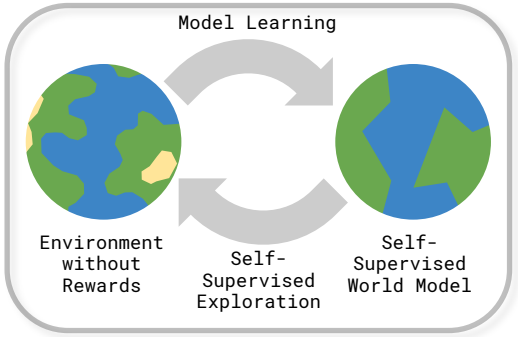


Agent

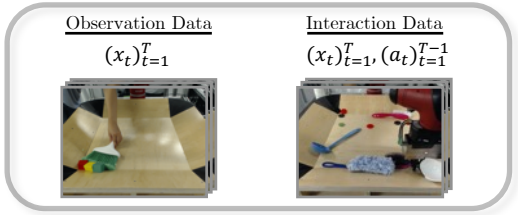
Prediction



Control

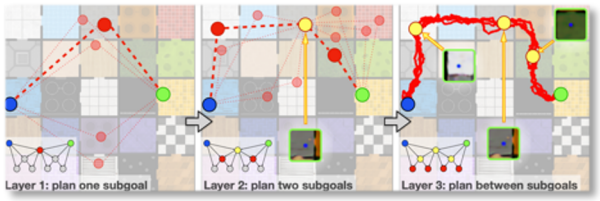


Self-supervised data collection

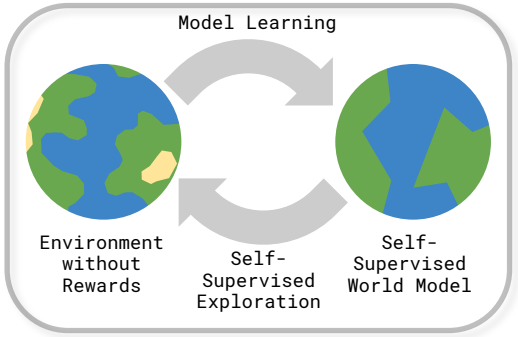


Leveraging human-collected data

Probabilistic latent variable models



Long-horizon hierarchical planning

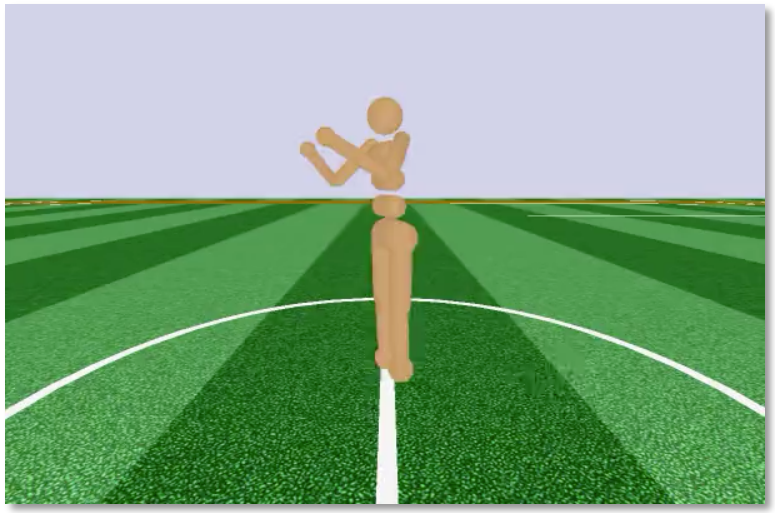


Self-supervised data collection

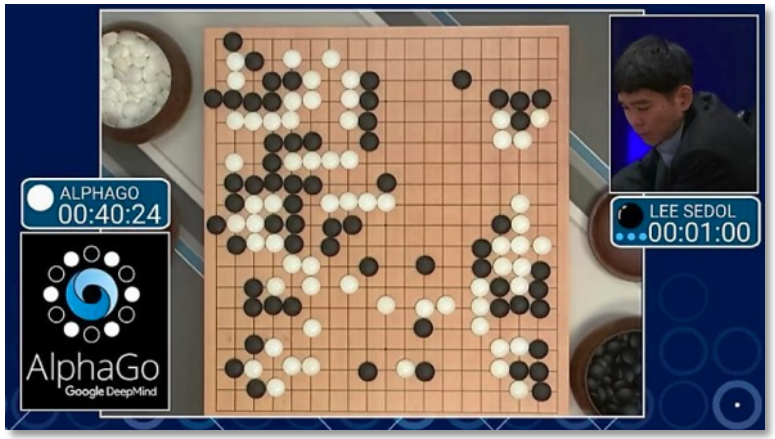
Can we learn general and reusable knowledge?

Reinforcement Learning: task-specific, difficult to generalize

[Mnih *et al.*, Nature 2015]



[Schulman *et al.*, 2015, 2017]



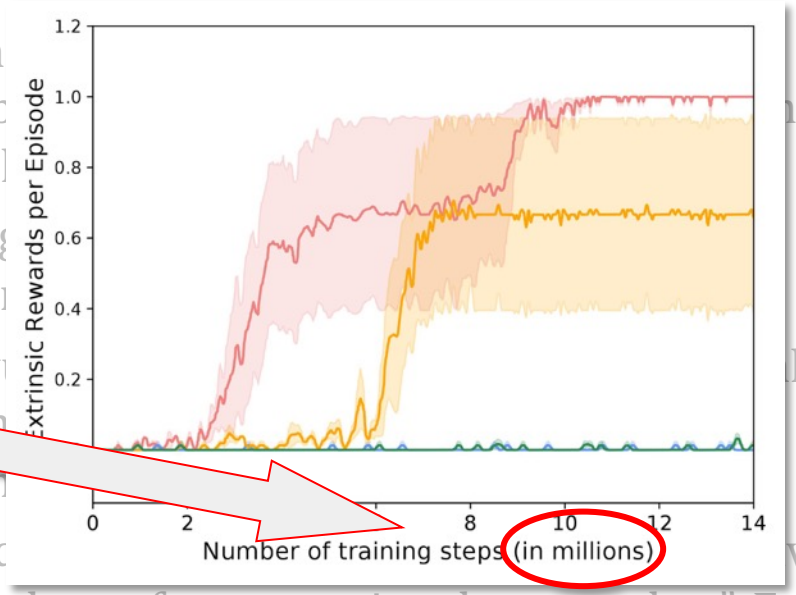
[Silver *et al.*, Nature 2016]

[Kalashnikov *et al.*, CoRL 2018]

Self-Supervised Reinforcement Learning

- Schmidhuber, “A possibility for implementing curiosity and boredom in model building neural controllers”. SAB, 1991.
- Singh *et al.,* “Intrinsically Motivated Reinforcement Learning”. NeurIPS, 2004.
- Klyubin et al., “Empowerment: A Universal Agent-Centric Measure of Control”. Evolutionary Computation, 2005.
- Oudeyer & Kaplan, “What is intrinsic motivation? A typology of computational approaches”. Frontiers in Neurorobotics, 2009.
- Sun *et al.,* “Planning to Be Surprised: Optimal Bayesian Exploration in Dynamic Environments”. AGI, 2011
- Mohamed & Rezende, “Variational information maximisation for intrinsically motivated reinforcement learning”. NeurIPS, 2015.
- Gregor *et al.,* “Variational intrinsic control”. ICLR Workshop, 2017.
- Pathak *et al.,* “Curiosity-driven Exploration by Self-supervised Exploration”. ICML 2017
- Pathak *et al.*, “Self-Supervised Exploration via Disagreement”. ICML, 2019.
- Hansen *et al.,* “Fast Task Inference with Variational Intrinsic Successor Features”. ICLR 2020.

Self-Supervised Reinforcement Learning



- Sch… imp… buil… …model
- Sing… Rei…
- Kly… …l Agent-Con… Con…
- Ouc… …vation? A typology of computational approaches". Frontiers in Neurorobotics, 2009

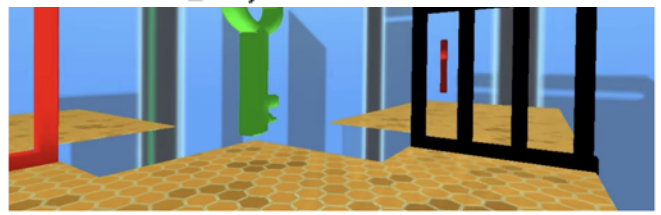
- Gregor *et al.*, "Variational intrinsic contr… …kshop, 2017.
- Pathak *et al.*, "Curiosit… …rvised Exploration". IC…
- Path… …via Disagreement". IC…
- Han… …inference with Variational Intrinsic Succe… …res". ICLR 2020.

Inefficient Adaptation [millions of samples]

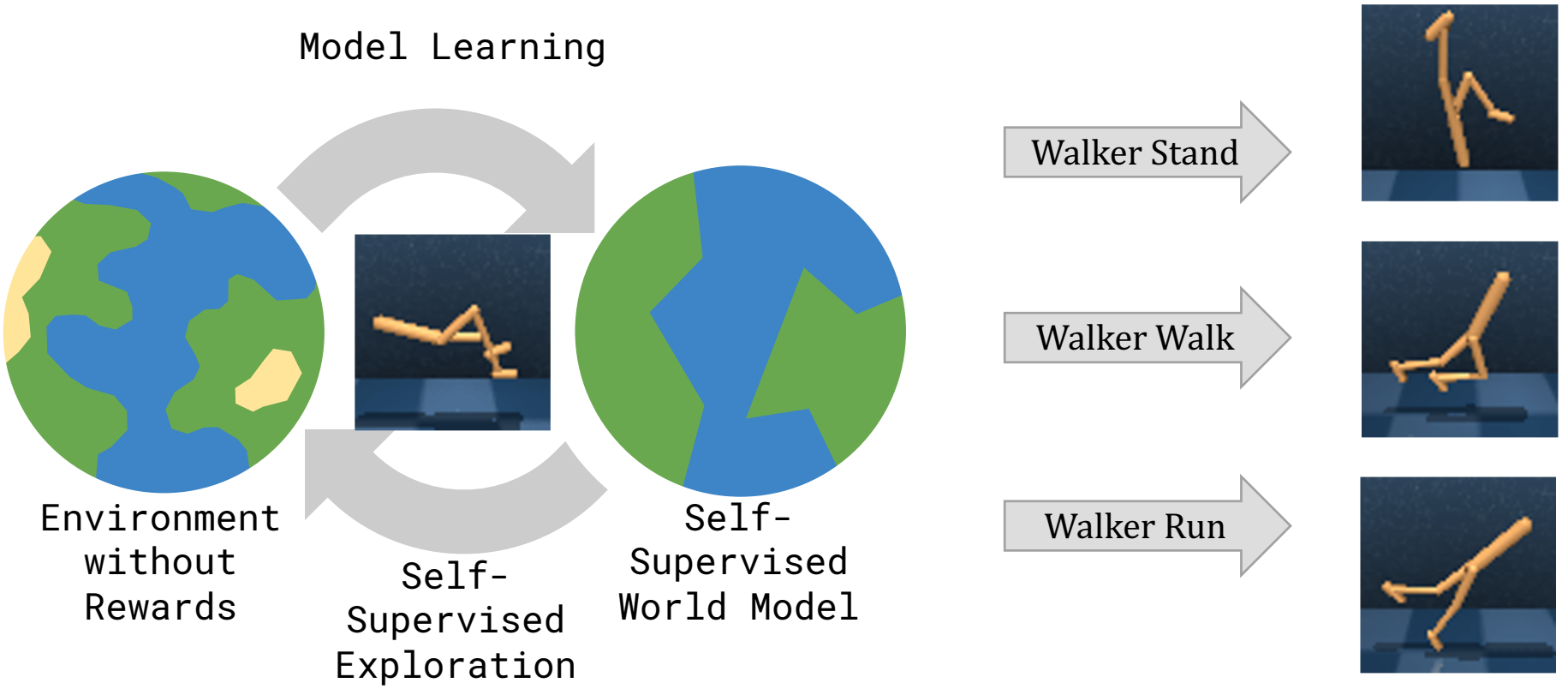
To evaluate VISR, we impose a two-phase setup on the full suite of 57 Atari games (Bellemare et al., 2013). Agents are allowed a long unsupervised training phase ($250M$ steps) without access to rewards, followed by a short test phase with rewards ($100k$ steps). The full VISR

(a) learn to explore in Level-1

(b) explore faster in Level-2

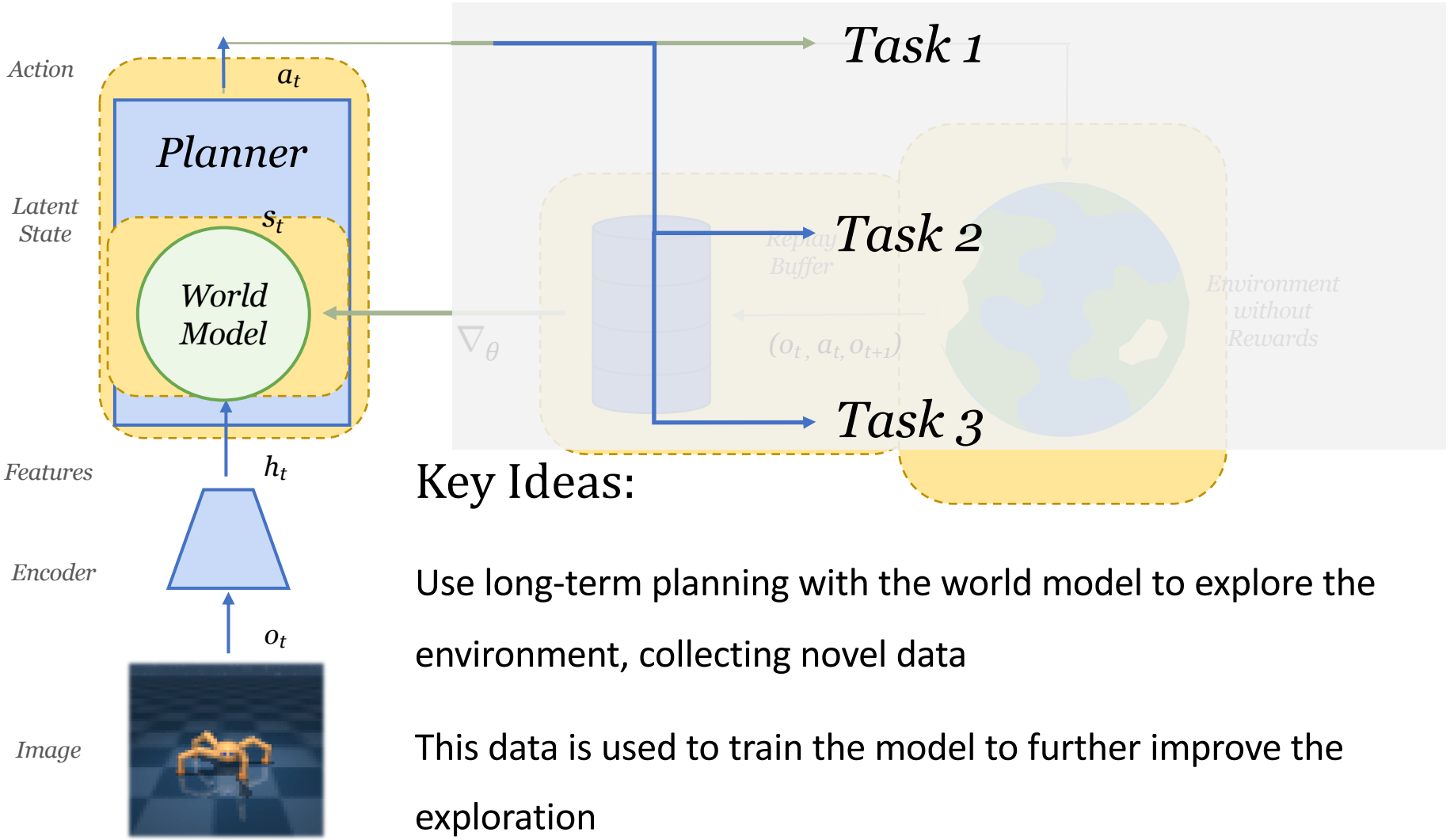


Model-building exploration



Sekar, Rybkin*, Daniilidis,Abbeel,Hafner, Pathak. Planning to Explore via Self-Supervised World Models. ICML, 2020*

Planning to Explore



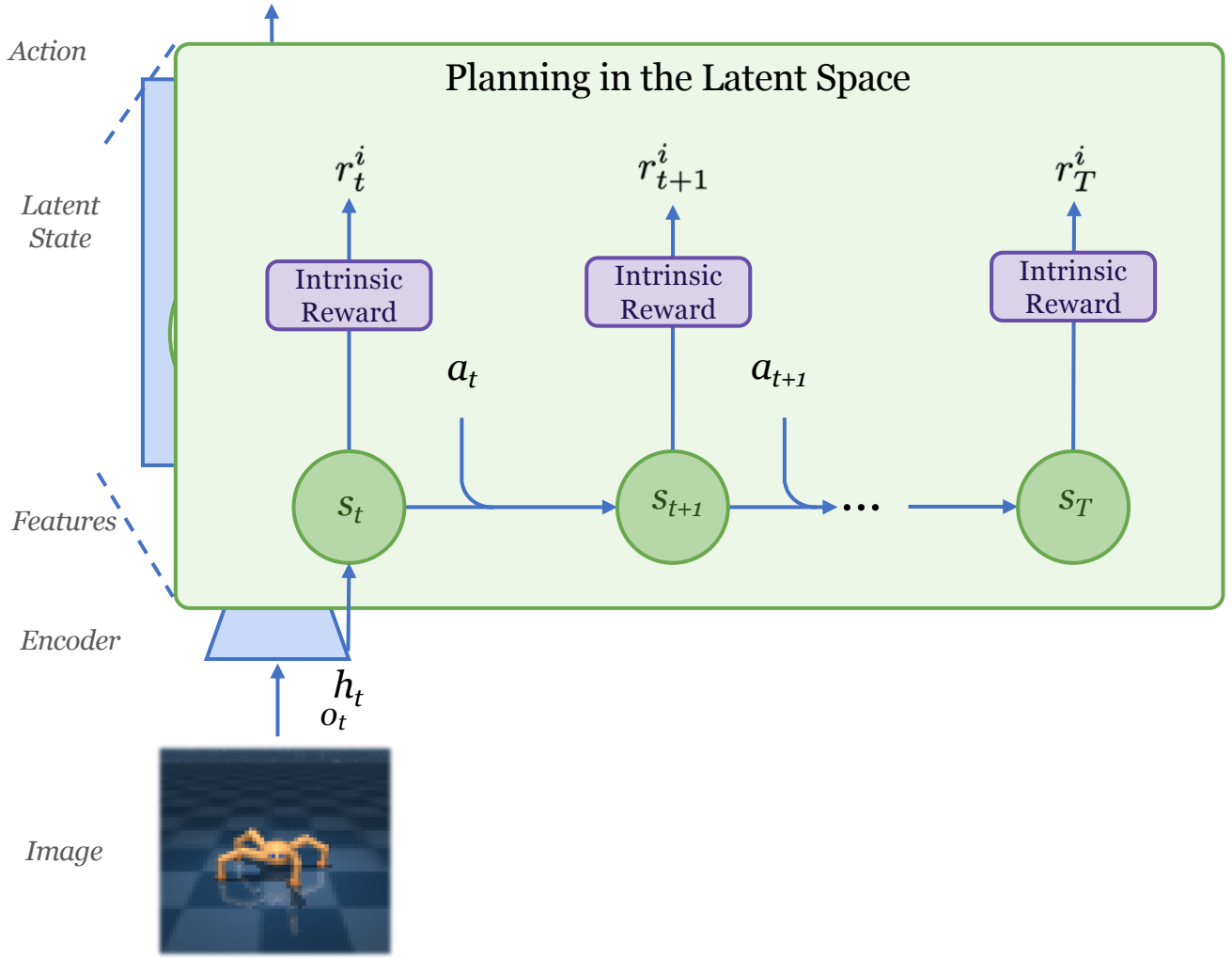
Key Ideas:

Use long-term planning with the world model to explore the environment, collecting novel data

This data is used to train the model to further improve the exploration

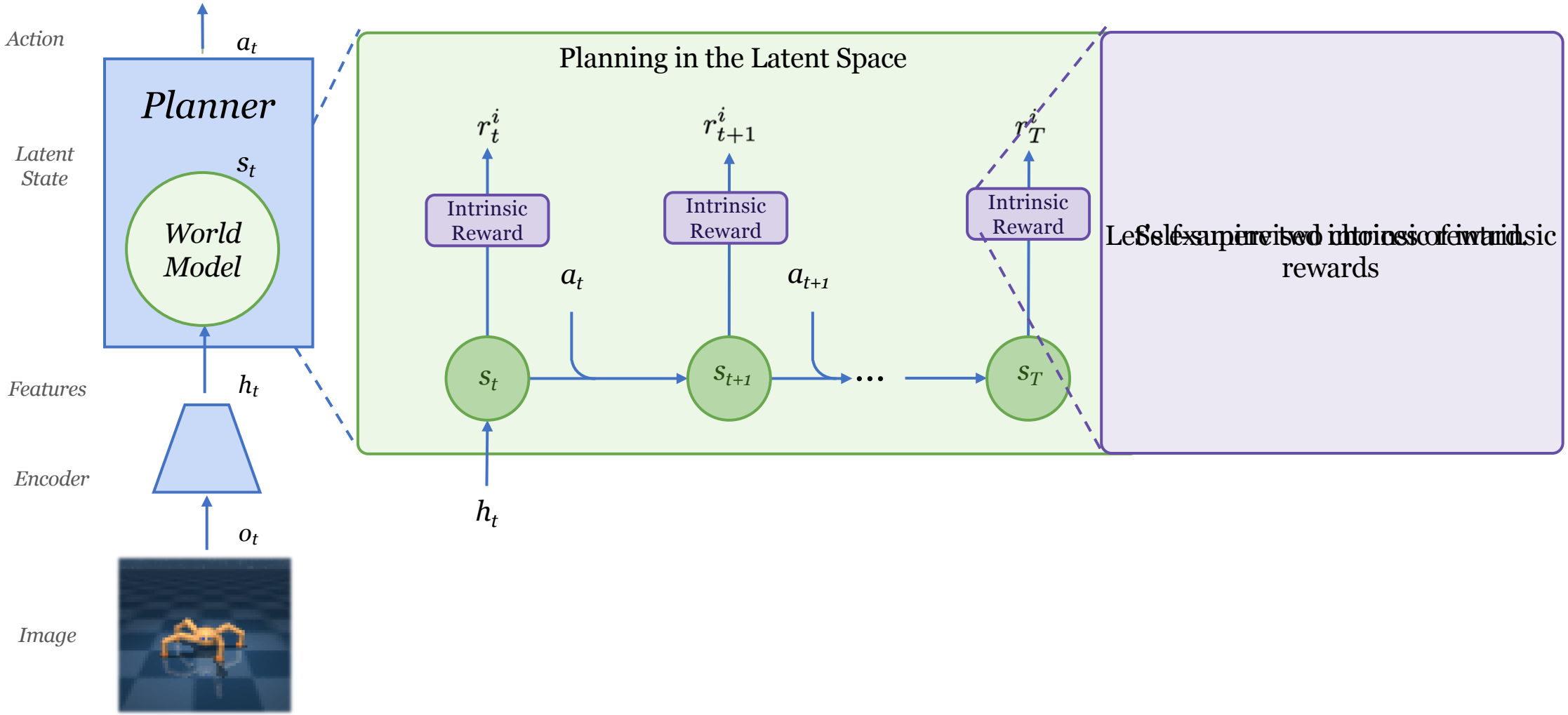
Same model is later used to plan for new tasks at test time

Planning to Explore

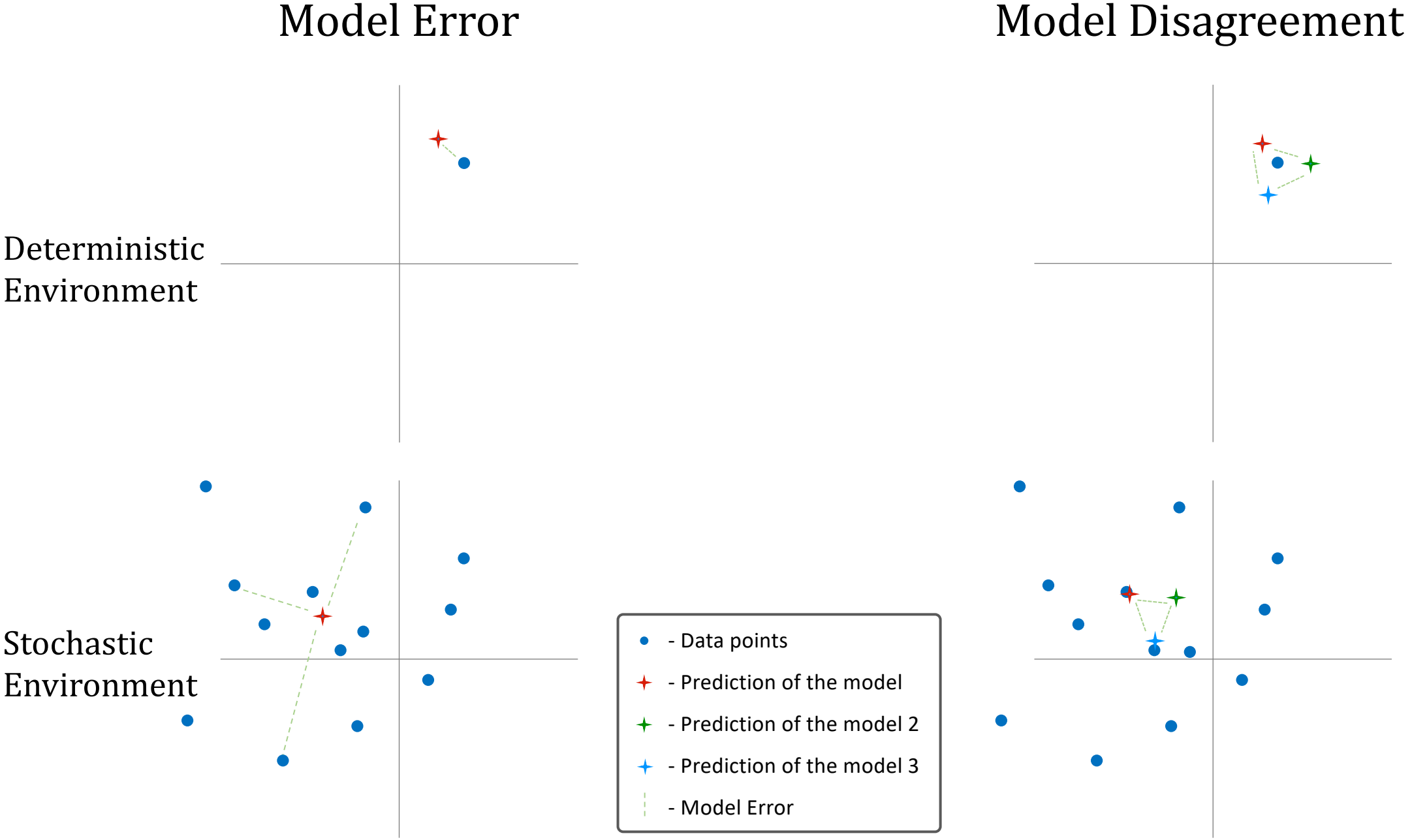


Sekar, Rybkin*, Daniilidis,Abbeel,Hafner, Pathak. Planning to Explore via Self-Supervised World Models. ICML, 2020*

Planning to Explore

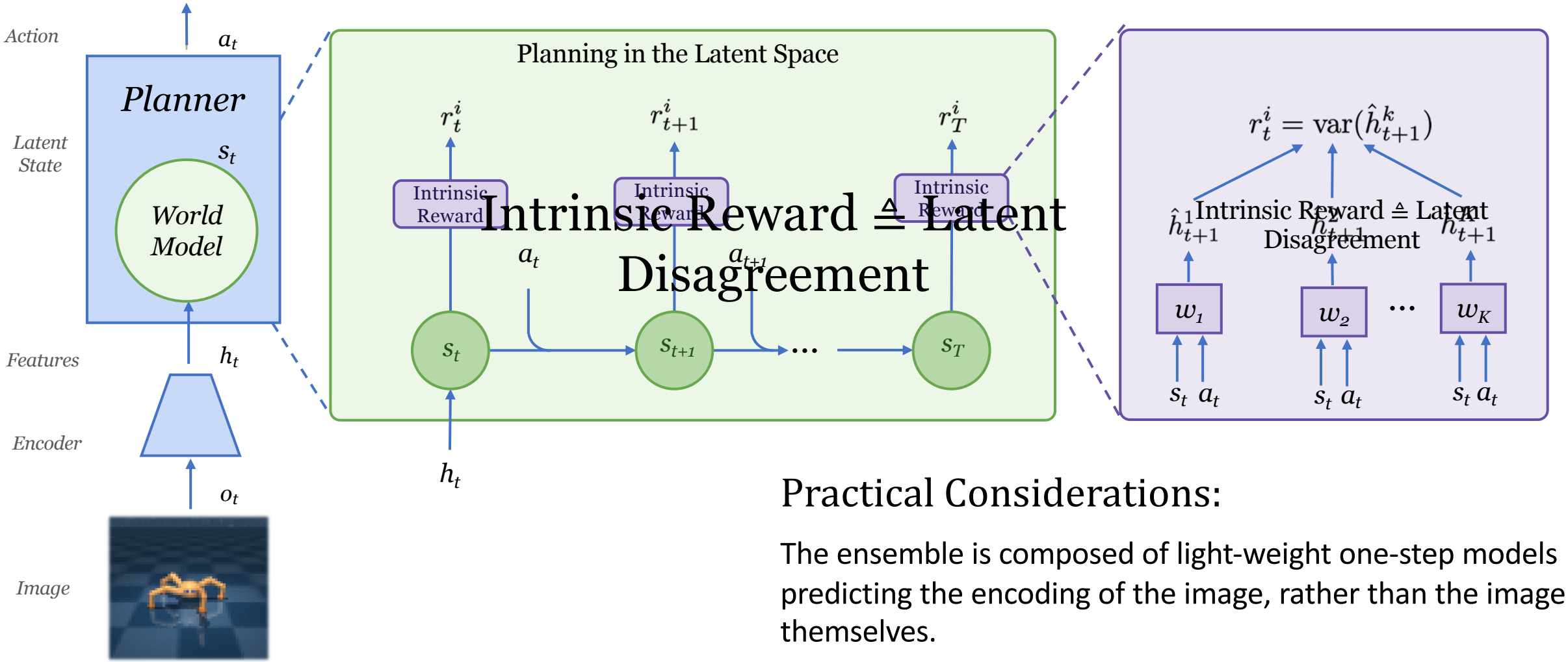


Sekar*, Rybkin*, Daniilidis,Abbeel,Hafner, Pathak. Planning to Explore via Self-Supervised World Models. ICML, 2020



Sekar, Rybkin*, Daniilidis,Abbeel,Hafner, Pathak. Planning to Explore via Self-Supervised World Models. ICML, 2020*

Planning to Explore

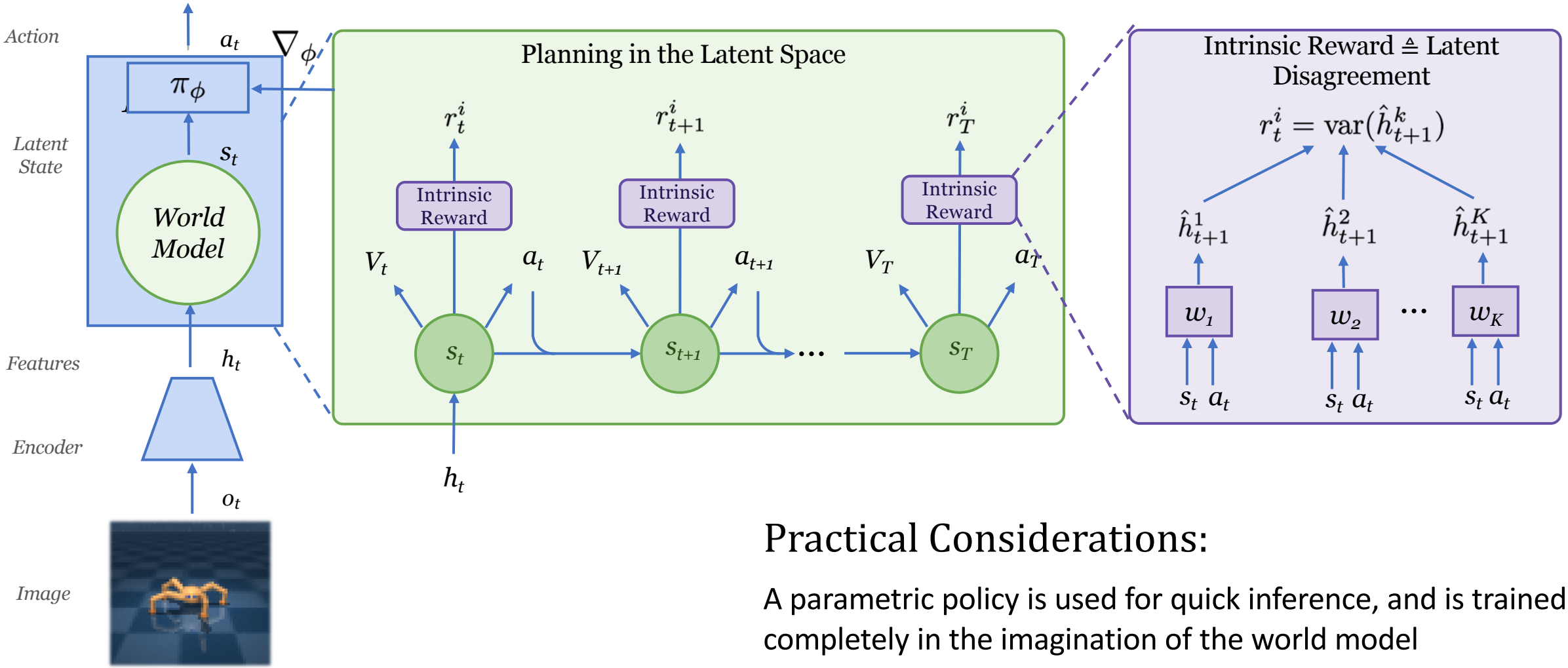


Practical Considerations:

The ensemble is composed of light-weight one-step models predicting the encoding of the image, rather than the image themselves.

Sekar, Rybkin*, Daniilidis, Abbeel, Hafner, Pathak. Planning to Explore via Self-Supervised World Models. ICML, 2020*

Planning to Explore

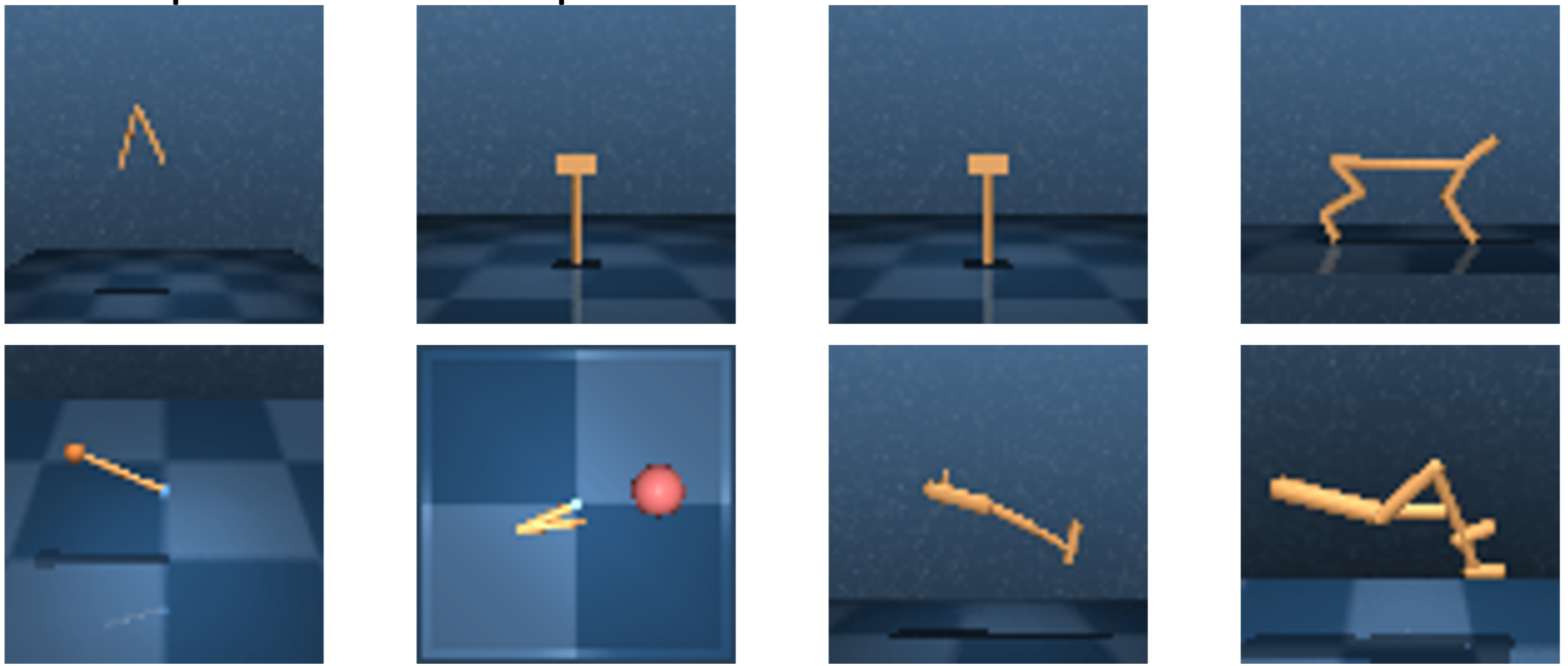


Practical Considerations:

A parametric policy is used for quick inference, and is trained completely in the imagination of the world model

Sekar, Rybkin*, Daniilidis,Abbeel,Hafner, Pathak. Planning to Explore via Self-Supervised World Models. ICML, 2020*

Self-Supervised Exploration Results

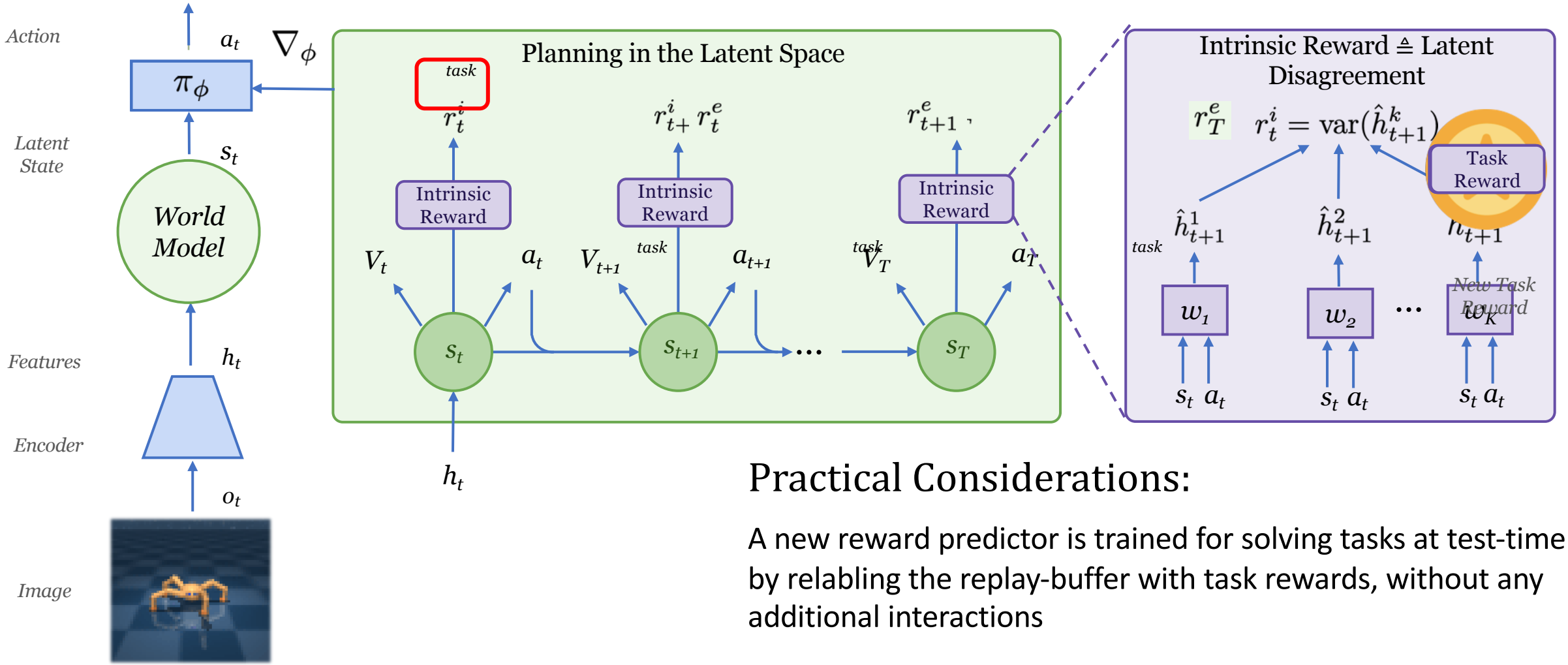


Sekar, Rybkin*, Daniilidis,Abbeel,Hafner, Pathak. Planning to Explore via Self-Supervised World Models. ICML, 2020*

How do we go from exploration to solving tasks?

Sekar, Rybkin*, Daniilidis,Abbeel,Hafner, Pathak. Planning to Explore via Self-Supervised World Models. ICML, 2020*

Exploration ➡ Tasks



Practical Considerations:

A new reward predictor is trained for solving tasks at test-time by relabling the replay-buffer with task rewards, without any additional interactions

Sekar, Rybkin*, Daniilidis,Abbeel,Hafner, Pathak. Planning to Explore via Self-Supervised World Models. ICML, 2020*

Experiments Outline

1. Solving a new-task in zero-shot

2. What if we add 20 supervised episodes?

3. Multi-task performance

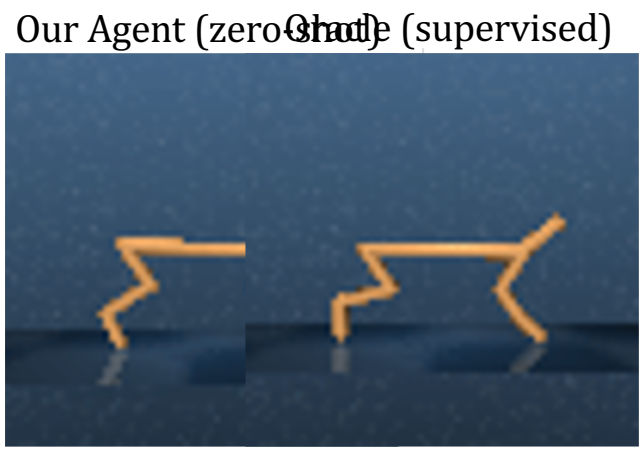
Experiments Outline

1. Solving a new-task in zero-shot

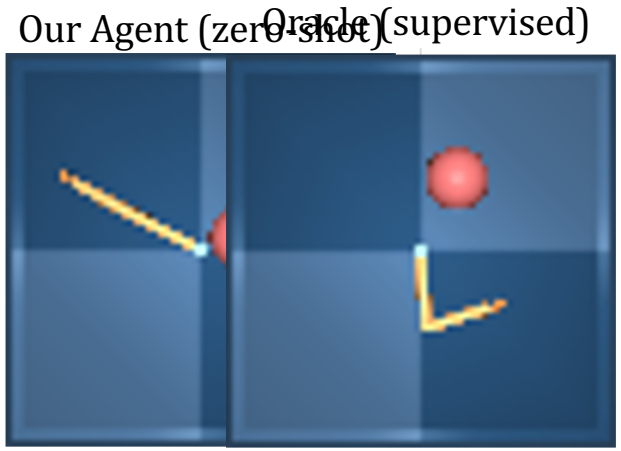
2. What if we add 20 supervised episodes?

3. Multi-task performance

Zero-Shot Reinforcement Learning

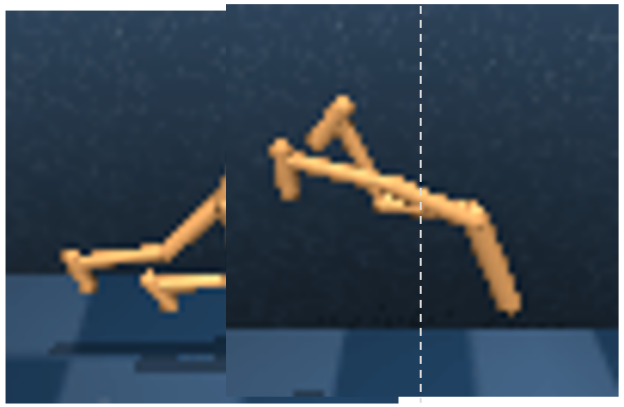


Cheetah Run



Reacher Easy

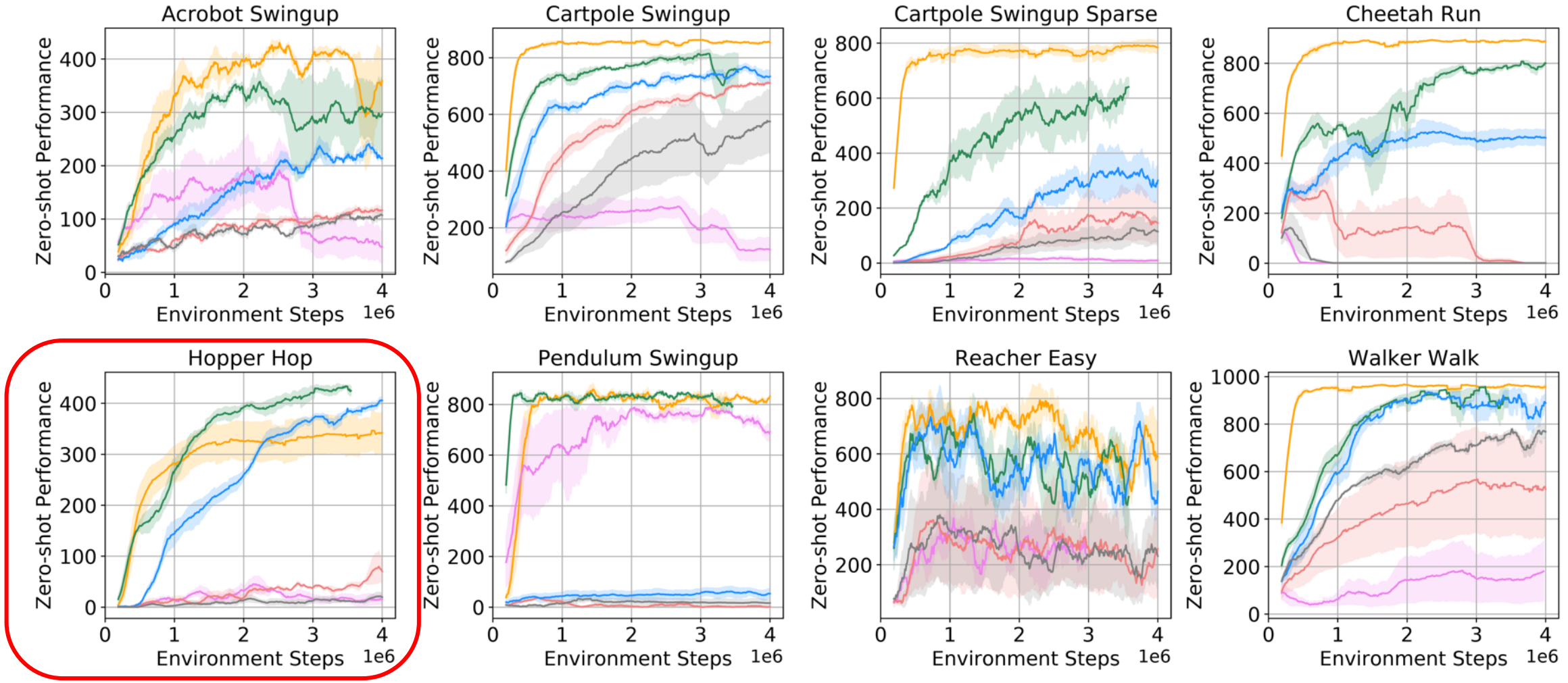
Hopper Hop



Walker Walk

Sekar, Rybkin*, Daniilidis,Abbeel,Hafner, Pathak. Planning to Explore via Self-Supervised World Models. ICML, 2020*

Zero-Shot Reinforcement Learning



Sekar, Rybkin*, Daniilidis,Abbeel,Hafner, Pathak. Planning to Explore via Self-Supervised World Models. ICML, 2020*

Experiments Outline

1. Solving a new-task in zero-shot

2. What if we add 20 supervised episodes?

3. Multi-task performance

Few-Shot Adaptation

Our Agent (few-shot)

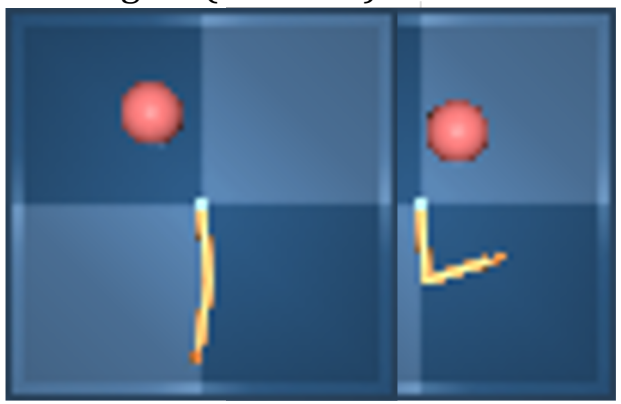
Oracle (supervised)

Cheetah Run

Hopper Hop

Our Agent (few-shot)

Oracle (supervised)



Reacher Easy

Walker Walk

Sekar, Rybkin*, Daniilidis,Abbeel,Hafner, Pathak. Planning to Explore via Self-Supervised World Models. ICML, 2020*

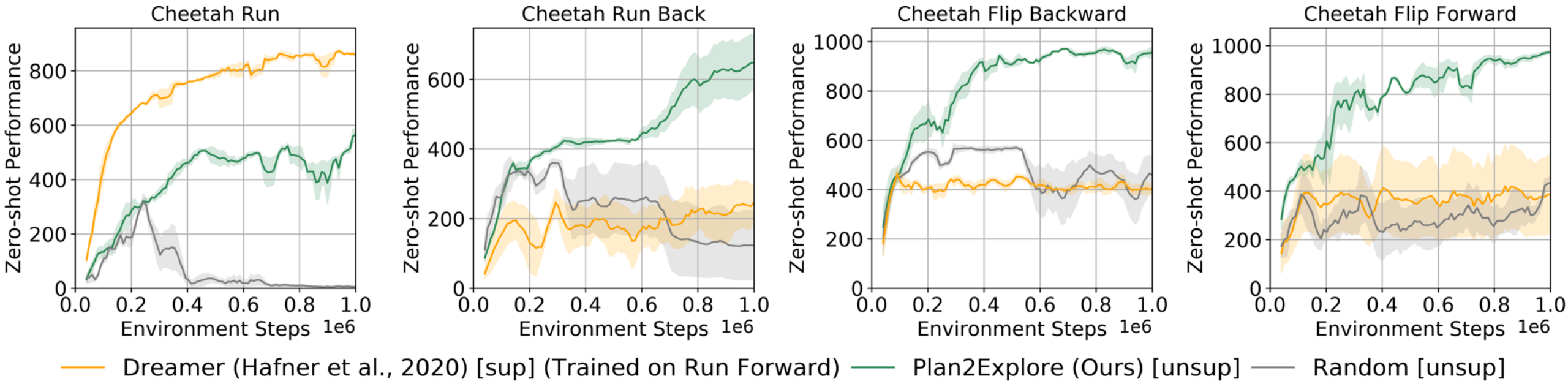
Experiments Outline

1. Solving a new-task in zero-shot

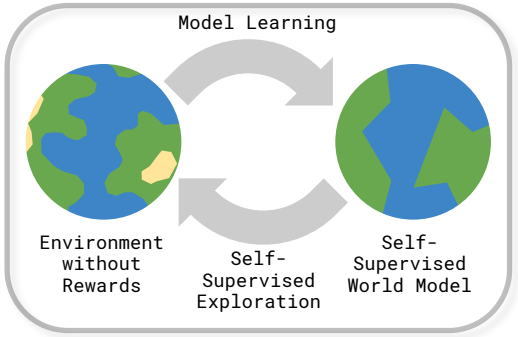
2. What if we add 20 supervised episodes?

3. Multi-task performance

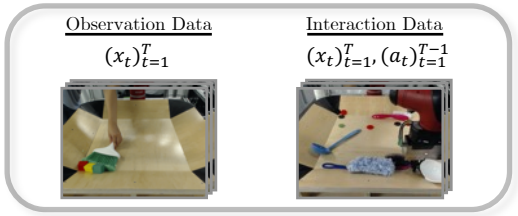
Can one model be used for multiple tasks?



Sekar, Rybkin*, Daniilidis,Abbeel,Hafner, Pathak. Planning to Explore via Self-Supervised World Models. ICML, 2020*



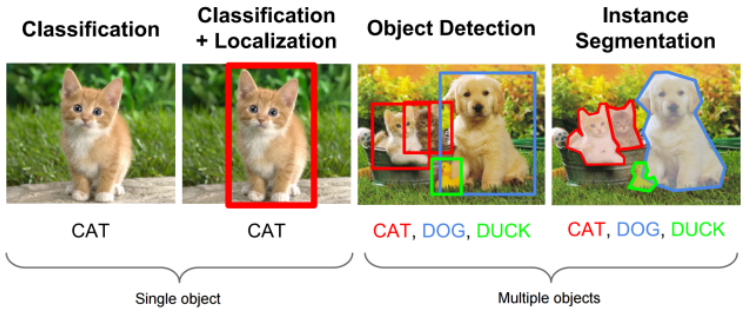
Self-supervised data collection

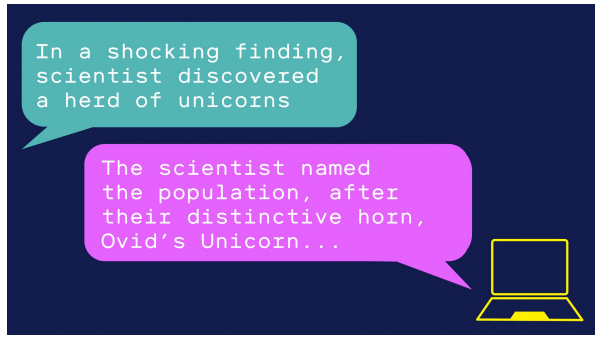


Leveraging human-collected data

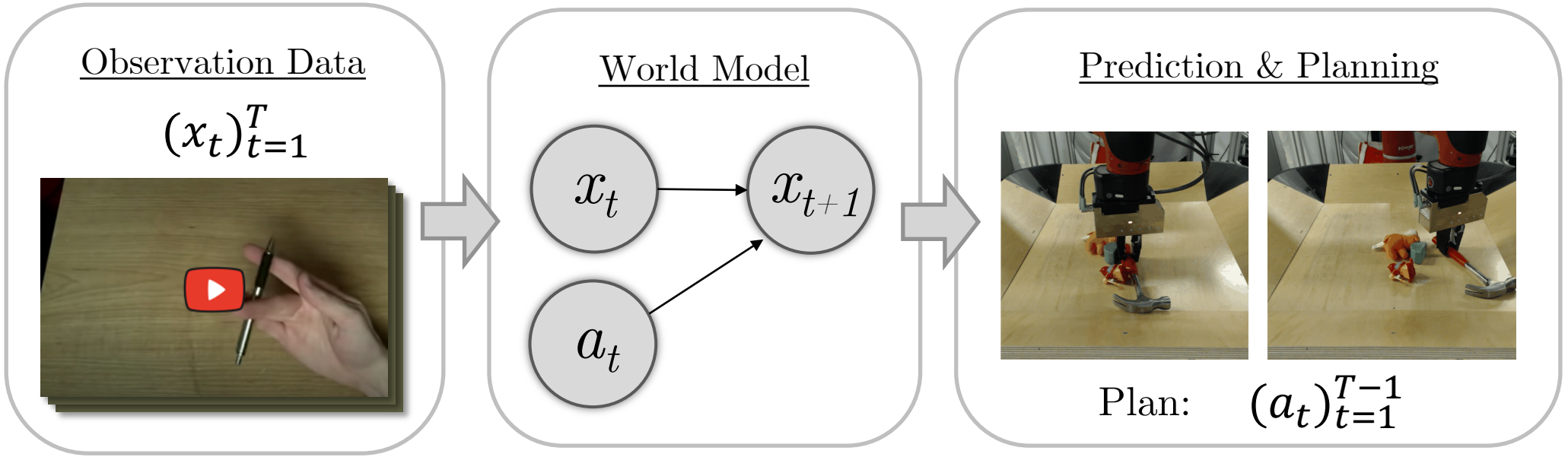

IMAGENET

WIKIPEDIA


Classification
Classification + Localization
Object Detection
Instance Segmentation
CAT
CAT
CAT, DOG, DUCK
CAT, DOG, DUCK
Single object
Multiple objects


In a shocking finding, scientist discovered a herd of unicorns
The scientist named the population, after their distinctive horn, Ovid's Unicorn...

Learning from Observation



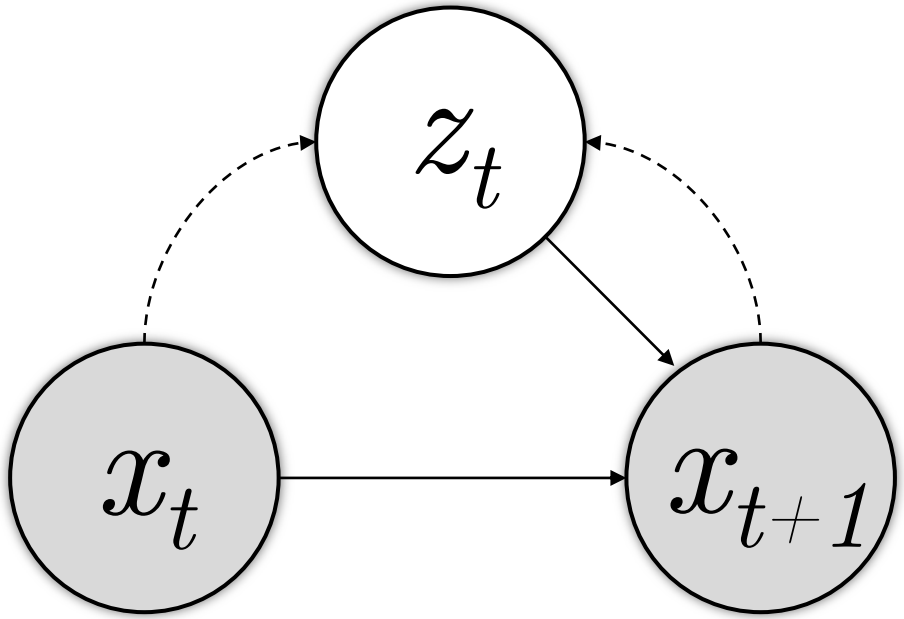
Rybkin, Pertsch*, Derpanis, Daniilidis, Jaegle. Learning what you can do before doing anything. ICLR, 2019*
Schmeckpeper, Xie, Rybkin, Tian, Daniilidis, Levine, Finn. Learning Predictive Models from Observation and Interaction. ECCV, 2020.

Learning Action Representations

$p(z_t)$

$p(x_{t+1}|x_{0:t}, z_{0:t})$

$q(z_t|x_{0:t+1})$



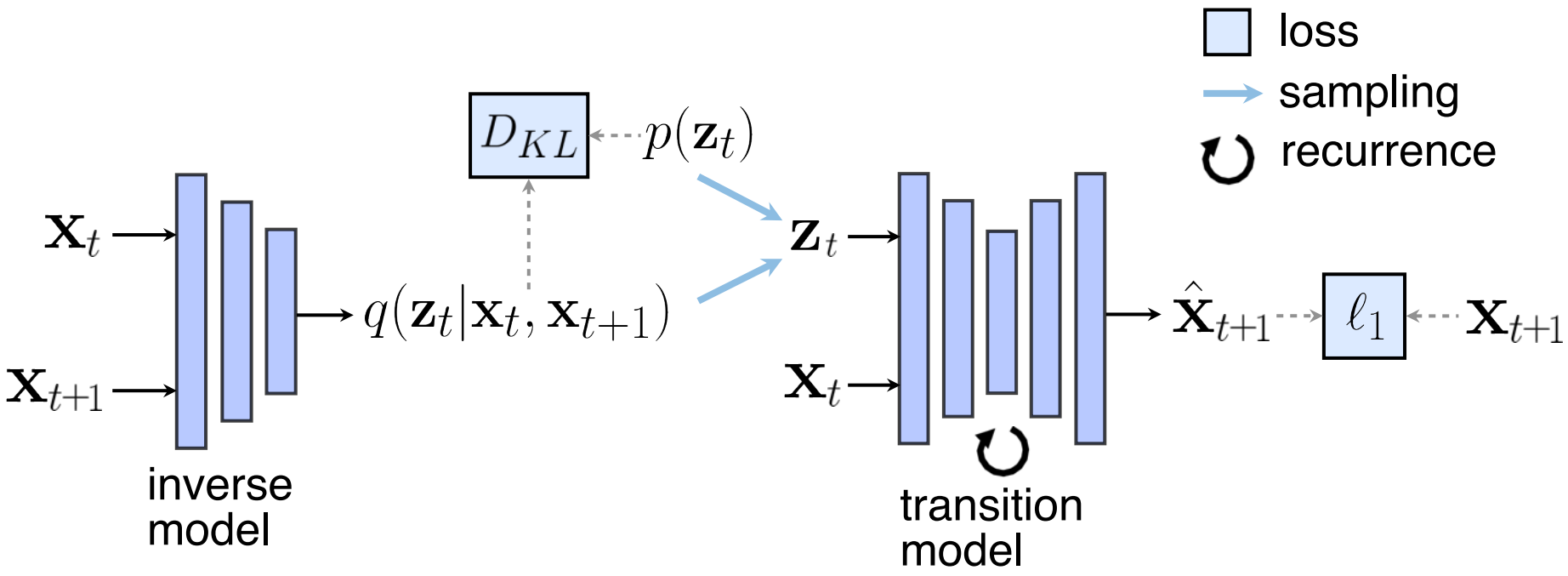
Rybkin, Pertsch*, Derpanis, Daniilidis, Jaegle. Learning what you can do before doing anything. ICLR, 2019*

Learning Action Representations

$p(z_t)$

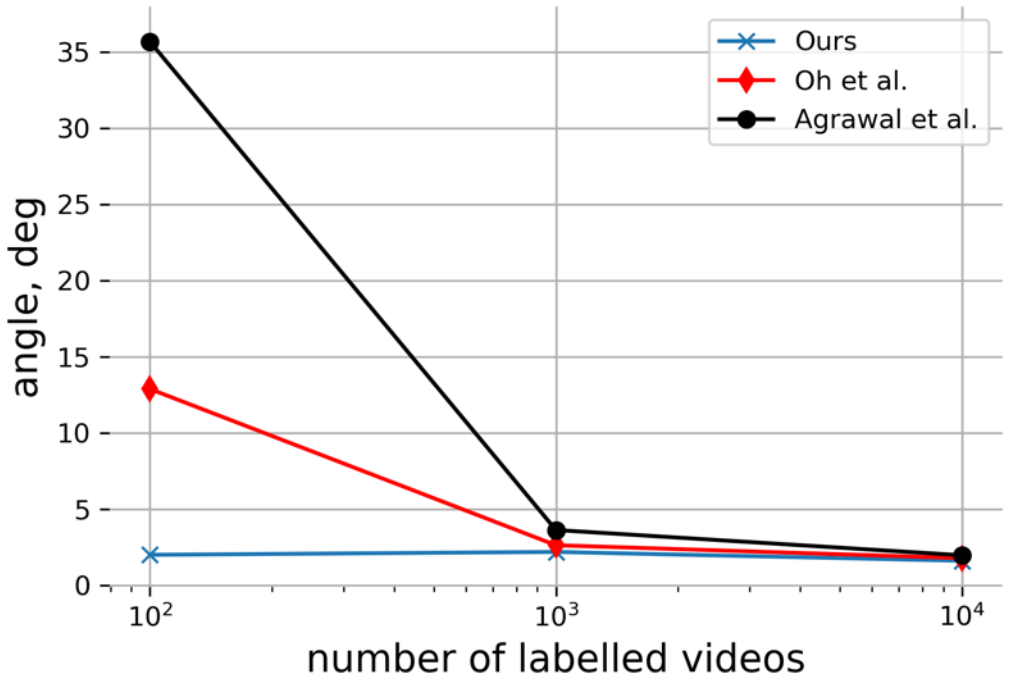
$p(x_{t+1}|x_{0:t}, z_{0:t})$

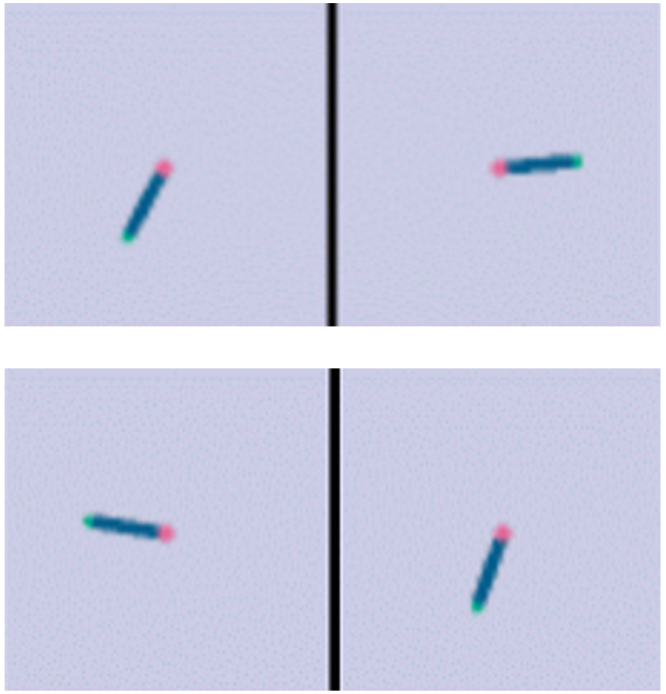
$q(z_t|x_{0:t+1})$



Rybkin, Pertsch*, Derpanis, Daniilidis, Jaegle. Learning what you can do before doing anything. ICLR, 2019*

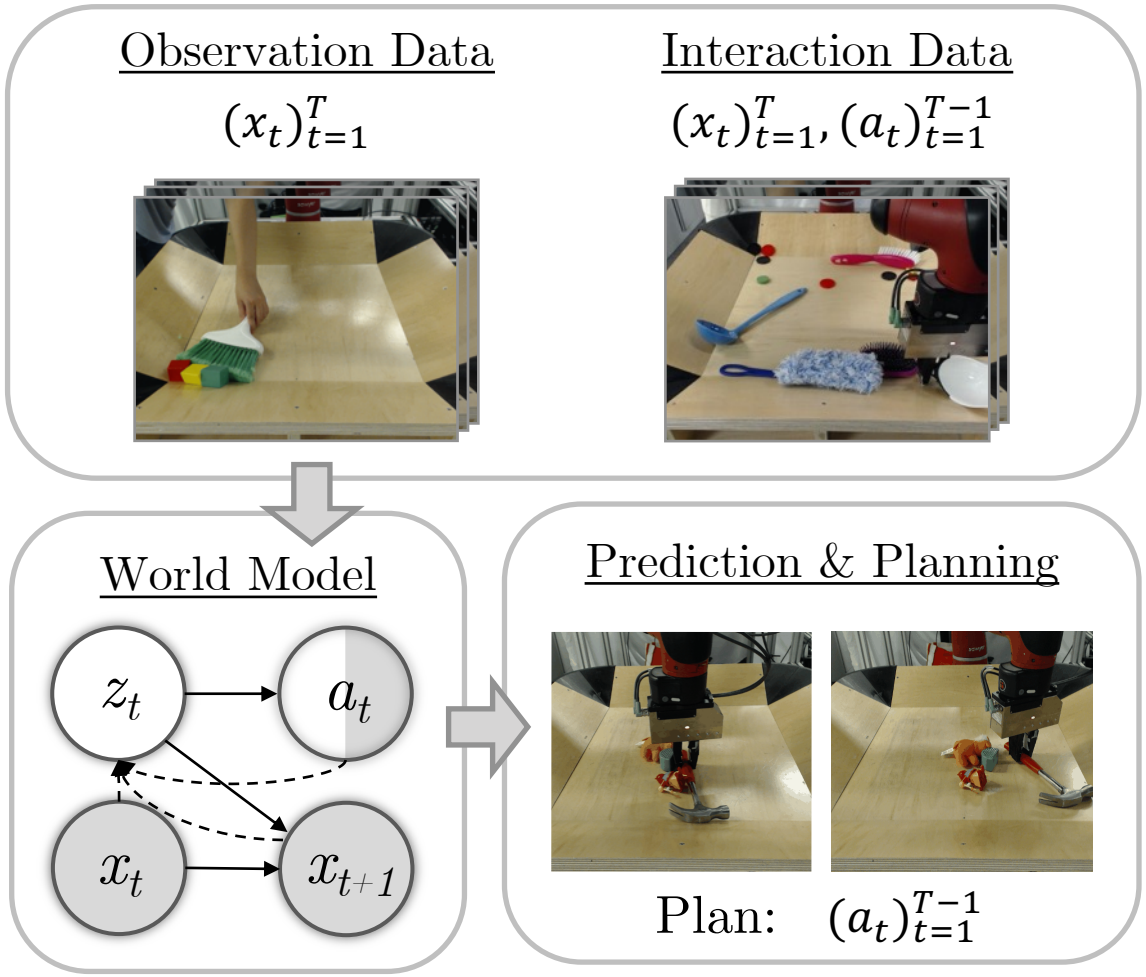
Learning Action Representations





Rybkin, Pertsch*, Derpanis, Daniilidis, Jaegle. Learning what you can do before doing anything. ICLR, 2019*

Learning From Observation and Interaction



Schmeckpeper, Han, Daniilidis, Rybkin. Visual Planning with Semi-Supervised Action Representations. ICML workshops, 2019
Schmeckpeper, Xie, Rybkin, Tian, Daniilidis, Levine, Finn. Learning Predictive Models from Observation and Interaction. ECCV, 2020.

Learning From Observation and Interaction

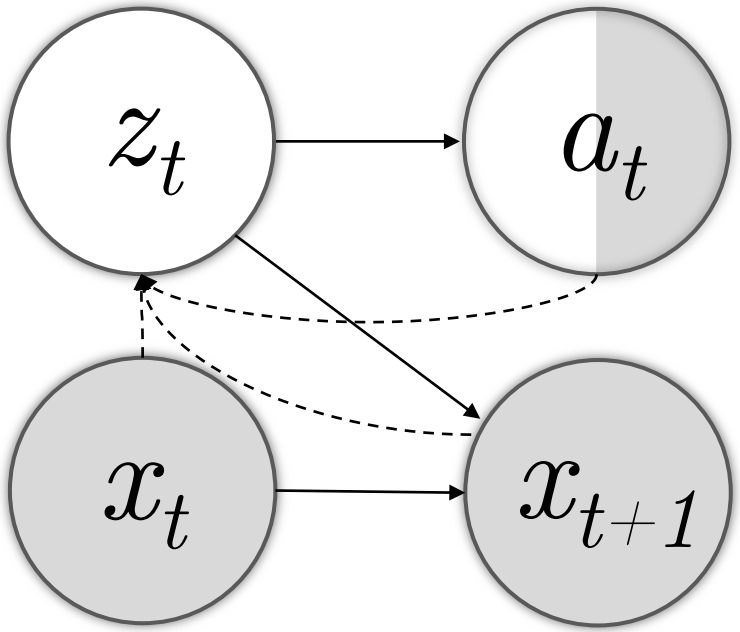
$p(z_t)$

$p(x_{t+1}|x_{0:t}, z_{0:t})$

$p(a_t|z_t)$

$q(z_t|x_{0:t+1})$

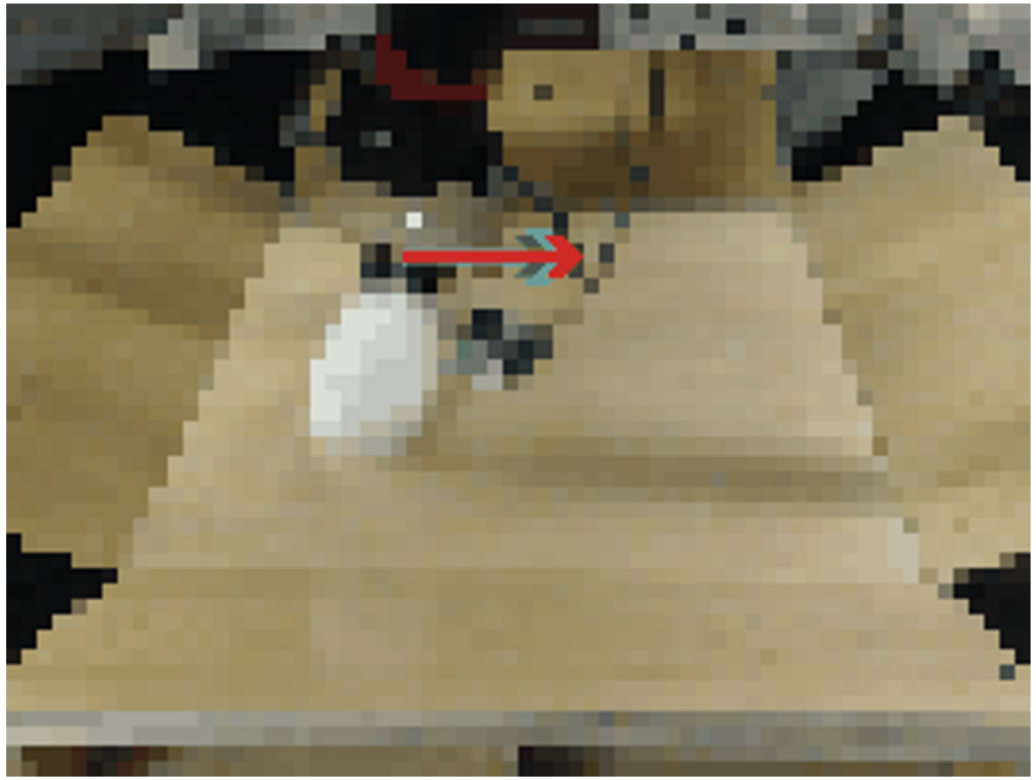
$q(z_t|a_t)$



Schmeckpeper, Han, Daniilidis, Rybkin. Visual Planning with Semi-Supervised Action Representations. ICML workshops, 2019

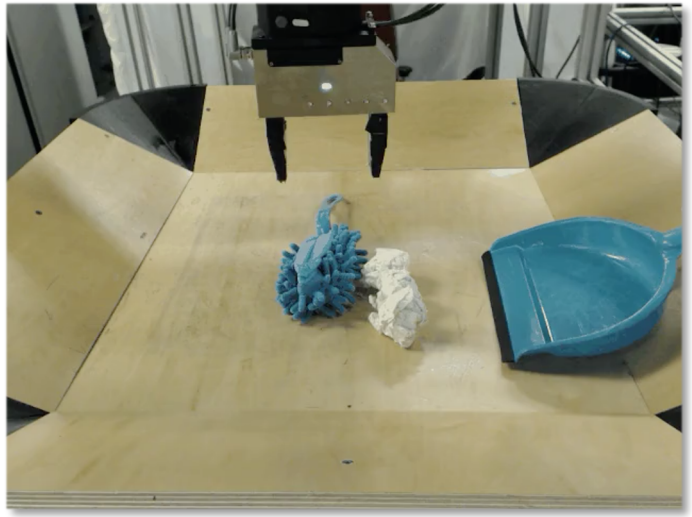
Schmeckpeper, Xie, Rybkin, Tian, Daniilidis, Levine, Finn. Learning Predictive Models from Observation and Interaction. ECCV, 2020.

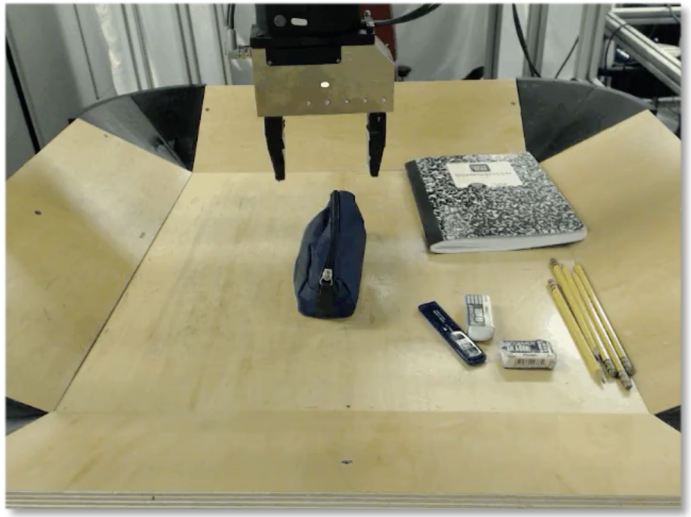
Action Prediction

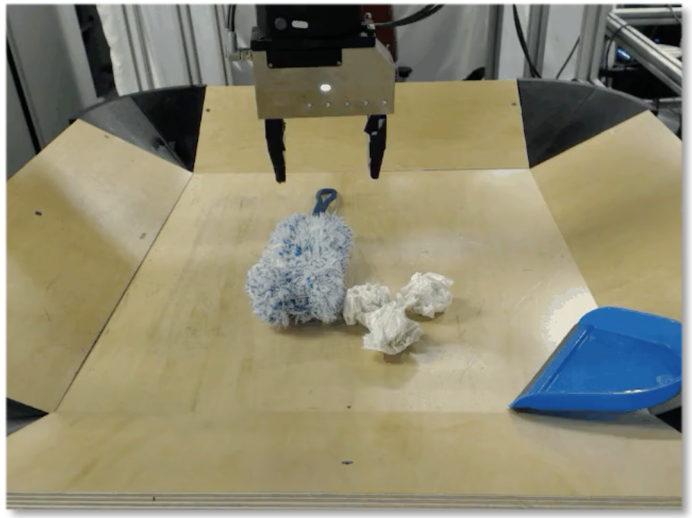


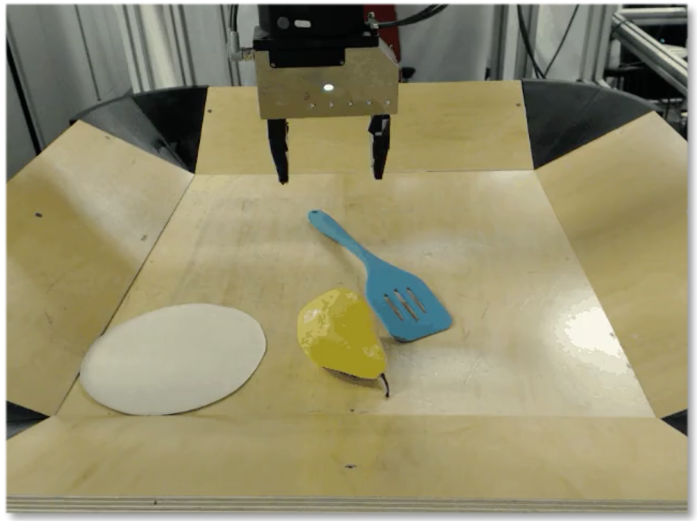
Schmeckpeper, Han, Daniilidis, Rybkin. Visual Planning with Semi-Supervised Action Representations. ICML workshops, 2019
Schmeckpeper, Xie, Rybkin, Tian, Daniilidis, Levine, Finn. Learning Predictive Models from Observation and Interaction. ECCV, 2020.

Robotic Planning

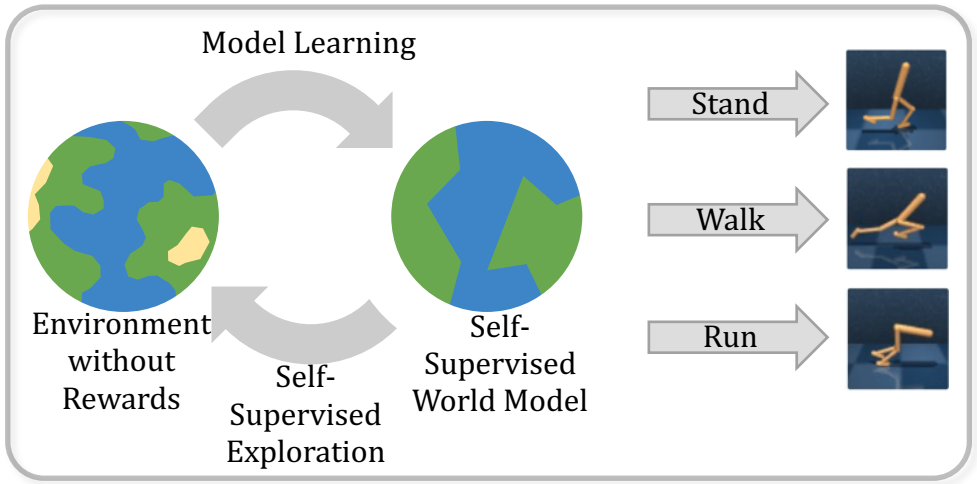


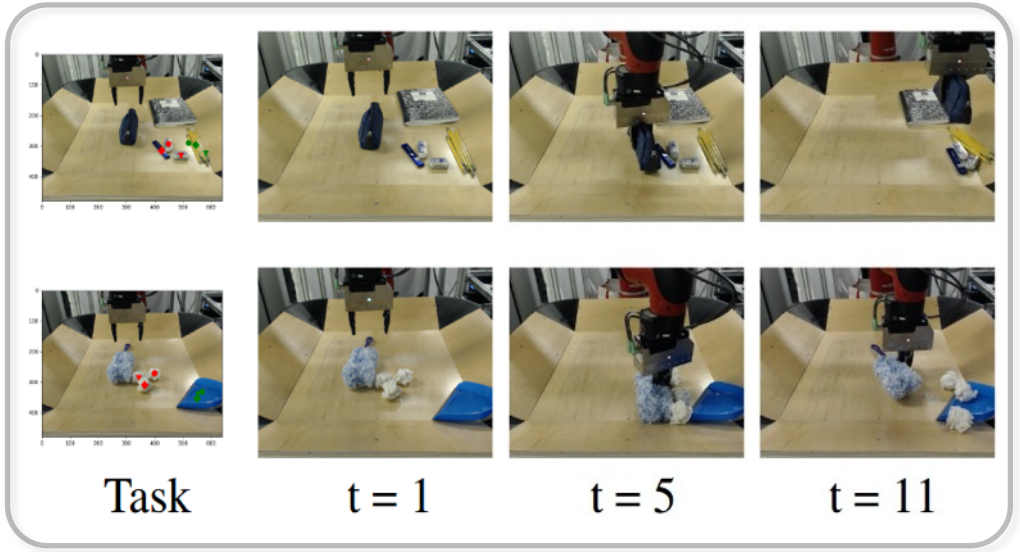




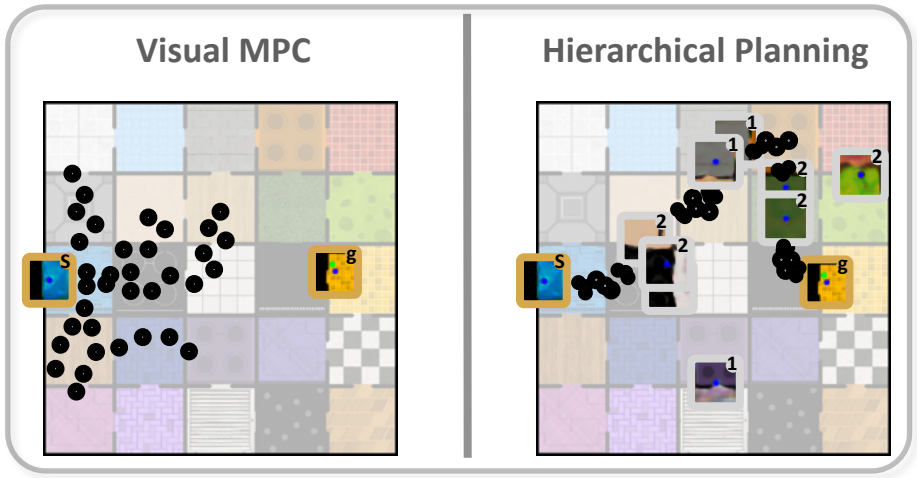


Schmeckpeper, Han, Daniilidis, Rybkin. Visual Planning with Semi-Supervised Action Representations. ICML workshops, 2019
Schmeckpeper, Xie, Rybkin, Tian, Daniilidis, Levine, Finn. Learning Predictive Models from Observation and Interaction. ECCV, 2020.


Model Learning
Environment without Rewards
Self-Supervised Exploration
Self-Supervised World Model
Stand
Walk
Run


Task
t = 1
t = 5
t = 11


Optimal σ-VAE
β-VAE, β=10-1
β-VAE, β=10-2
β-VAE, β=10-3
β-VAE, β=10-4
β-VAE, β=10-5
β-VAE, β=10-6


Visual MPC
Hierarchical Planning
s
g

Thanks to my amazing collaborators!


Karl Pertsch


Ramanan Sekar


Danijar Hafner


Pieter Abbeel

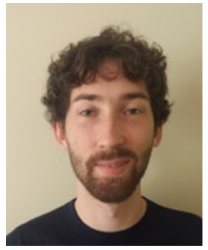

Deepak Pathak


Frederik Ebert


Dinesh Jayaraman


Chelsea Finn


Sergey Levine


Karl Schmeckpeper


Stephen Tian


Annie Xie


Joseph Lim

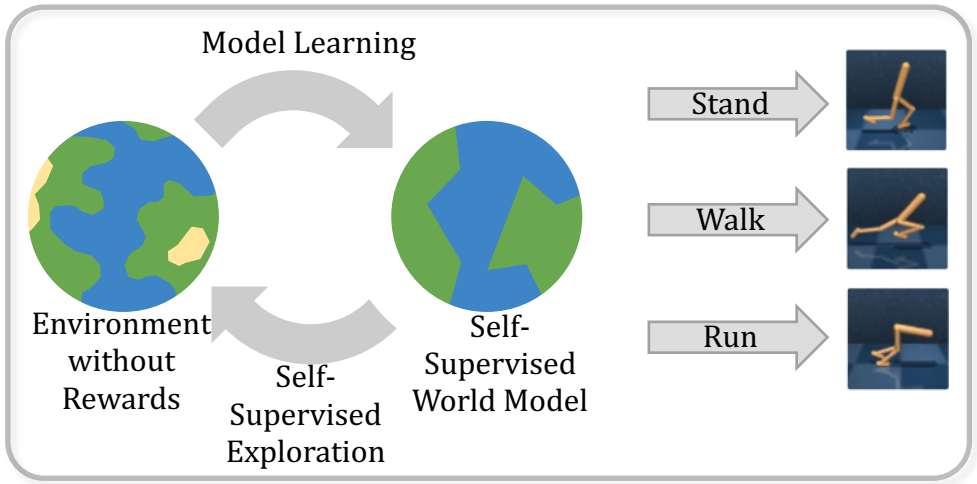

Jingyun Yang

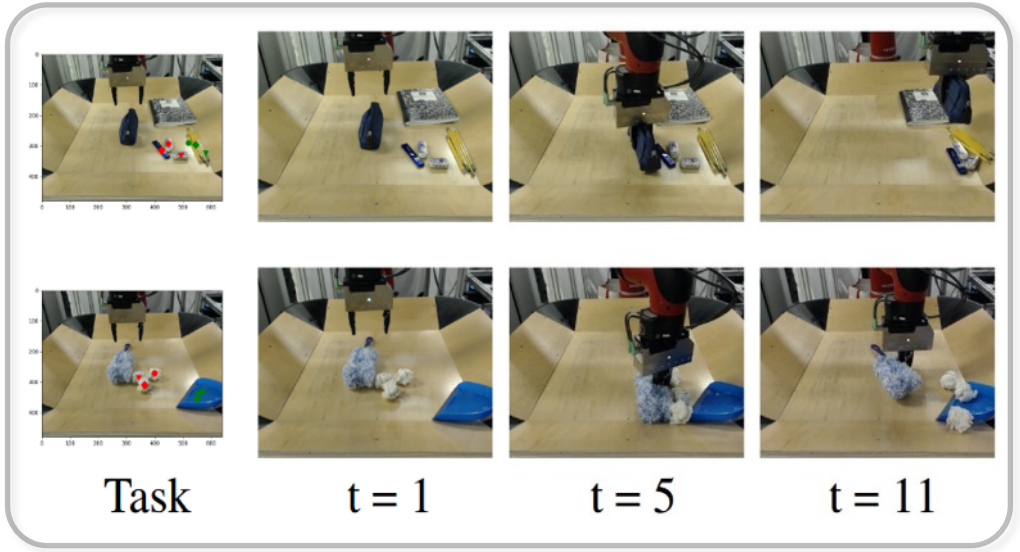

Shenghao Zhou

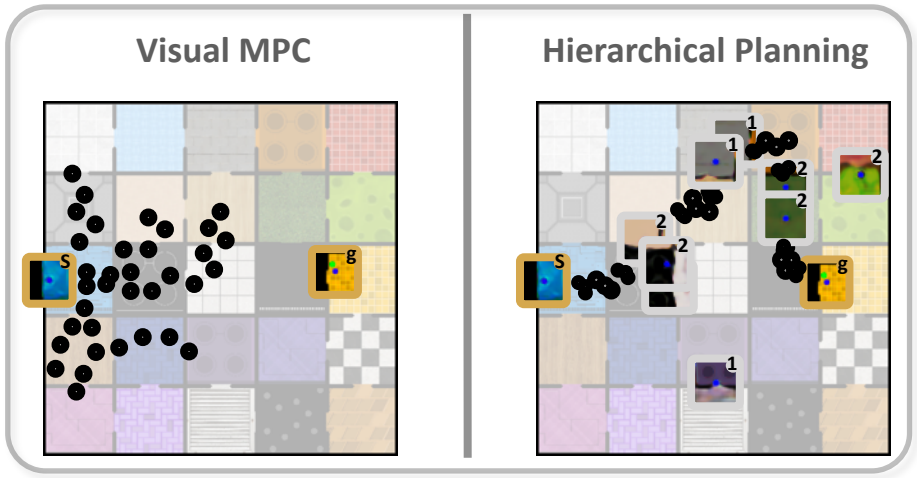

Kosta Derpanis


Andrew Jaegle


Kostas Daniilidis







Questions?

Oleh Rybkin
oleh@seas.upenn.edu