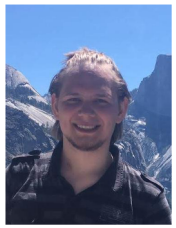
PLANNING TO EXPLORE VIA SELF-SUPERVISED WORLD MODELS

ICML 2020

Ramanan Sekar*

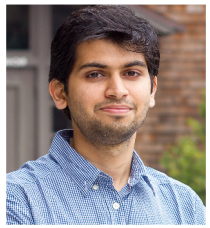


Oleh Rybkin*

Kostas Daniilidis

Pieter Abbeel

Danijar Hafner

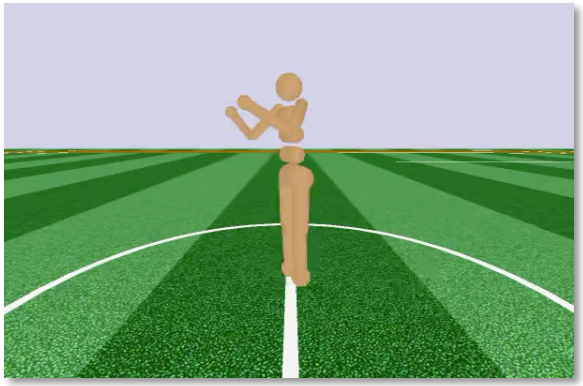


Deepak Pathak

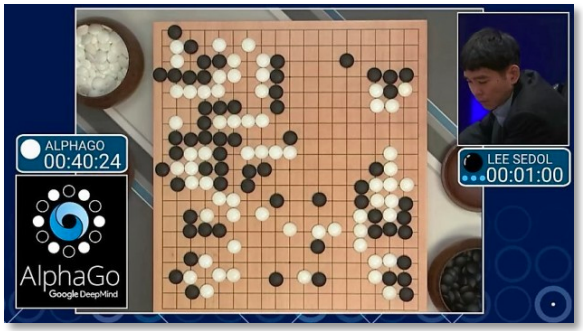
* equal contribution

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 RL

- Schmidhuber, “A possibility for implementing curiosity and boredom in model building neural controllers”. SAB, 1991.
- Singh *et al., “*Intrinsically Motivated Reinforcement Learning”. NeurIPS, 2004.
- Oudeyer, P.-Y. and Kaplan, F. What is intrinsic motivation? a typology of computational approaches. Frontiers in neurorobotics, 2009.
- Schmidhuber, “Formal theory of creativity, fun, and intrinsic motivation (1990-2010)”. TAMD, 2010.
- Sun *et al., “*Planning to Be Surprised: Optimal Bayesian Exploration in Dynamic Environments”. AGI, 2011
- Mohamed *et.al.,* “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.
- Savinov *et al.,* “Episodic curiosity through reachability”. ICLR 2019.
- Eysenbach *et al.,* “Diversity is all you need: Learn skills without a reward function”. ICLR 2019.
- Shyam *et al.,* “Model-based Active Exploration”. ICML, 2019
- Sharma *et al.*, “Dynamics-Aware Unsupervised Discovery of Skills”. ICLR, 2020.
- Finn, Levine., “deep visual foresight for planning robot motion”. ICRA, 2017.
- Pathak *et al.,* “Zero-Shot Visual Imitation”. ICLR, 2018.

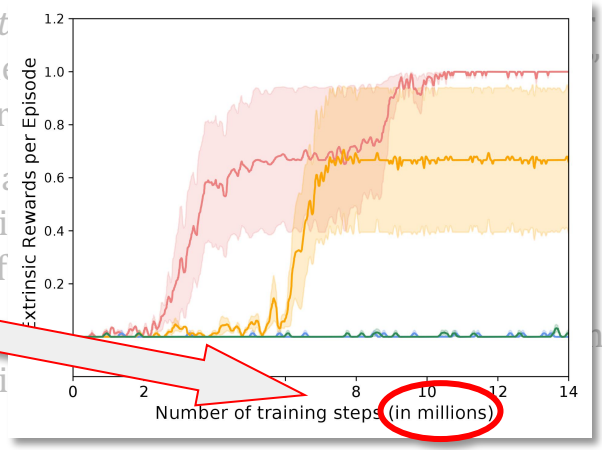
Self-Supervised RL

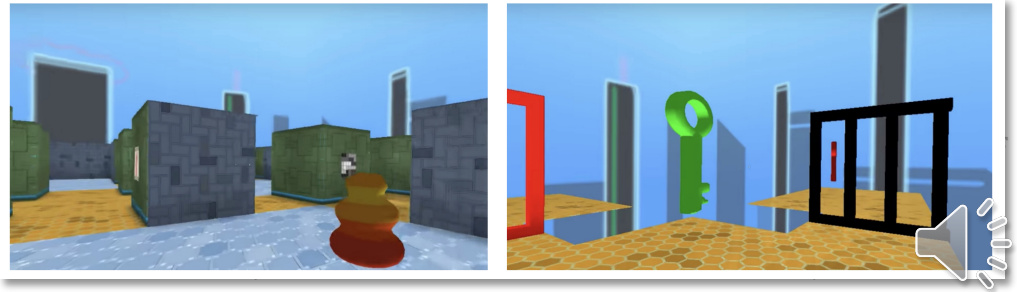
(a) learn to explore in Level-1 (b) explore faster in Level-2

- Gregor *et.al.,* "Variational intrinsic control". ICLR Workshop, 2017.
- Pathak *et.al.,* "Curiosity-driven E… supervised Exploration".
- Ostrovski *et…* …ural dens…
- Pat… …xploration via Disa… …919.
- Savin… …, "Episodic curiosity through reachability". ICLR 2019.

Inefficient Adaptation [millions of samples]

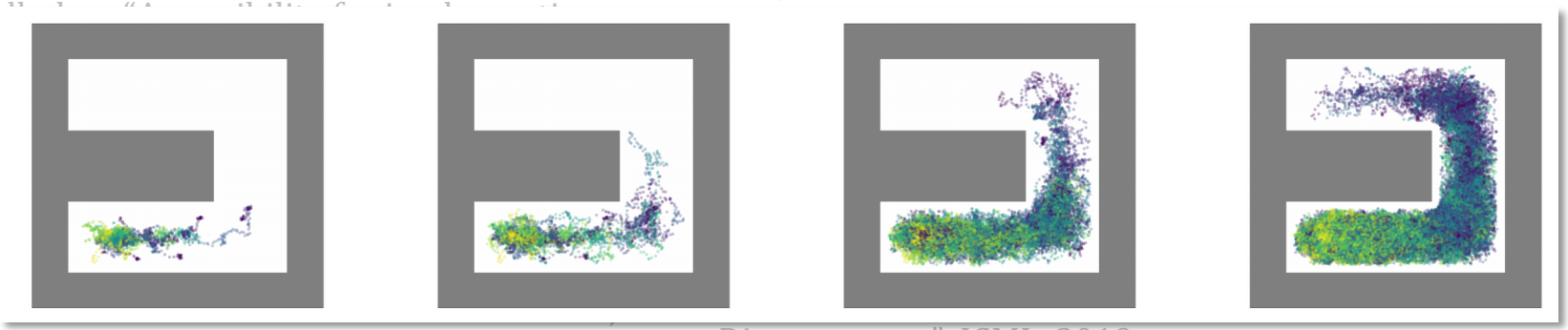
- Eysenbach *et al.,* "Diversity is all you need: Learn skills





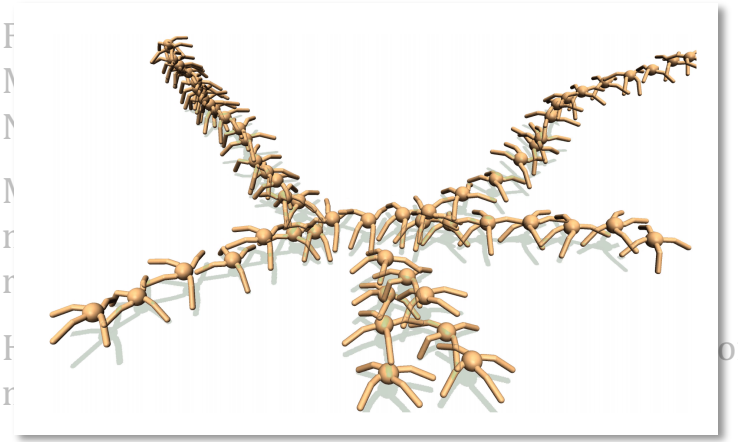
Self-Supervised RL

- Gregor *et.al.,* "Variational intrinsic control". ICLR



Disagreement". ICML, 2019.

- Savinov *et al.,* "Episodic curiosity through reachability". ICLR 2019.

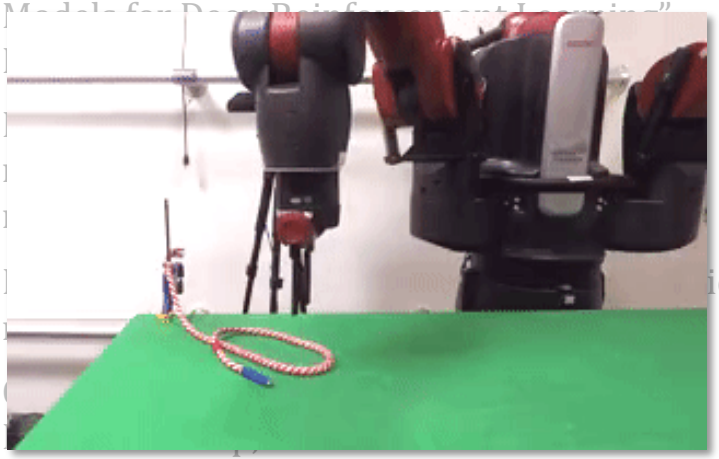


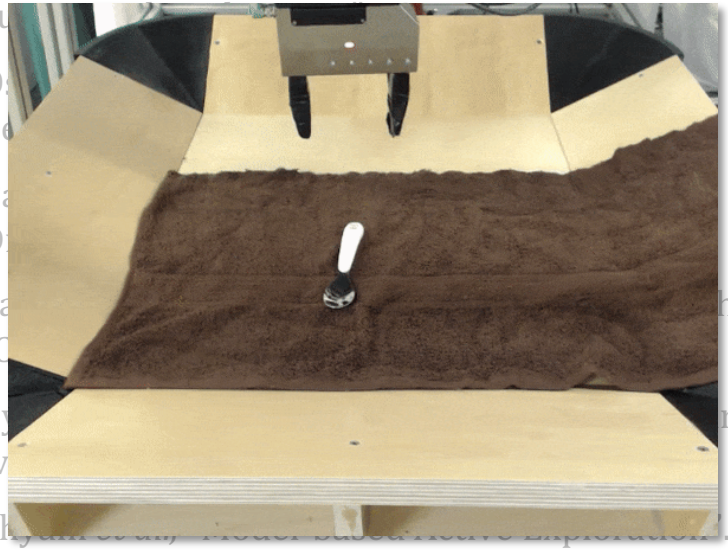
- Eysenbach *et al.,* "Diversity is all you need: Learning skills without a reward f
- Shya L, 2019
- Shar upervised Discovery of Skills , 2020.

Does not scale to images

- Pathak *et al.,* "Zero-Shot Visual Imitation". ICLR, 2018.

Self-Supervised RL

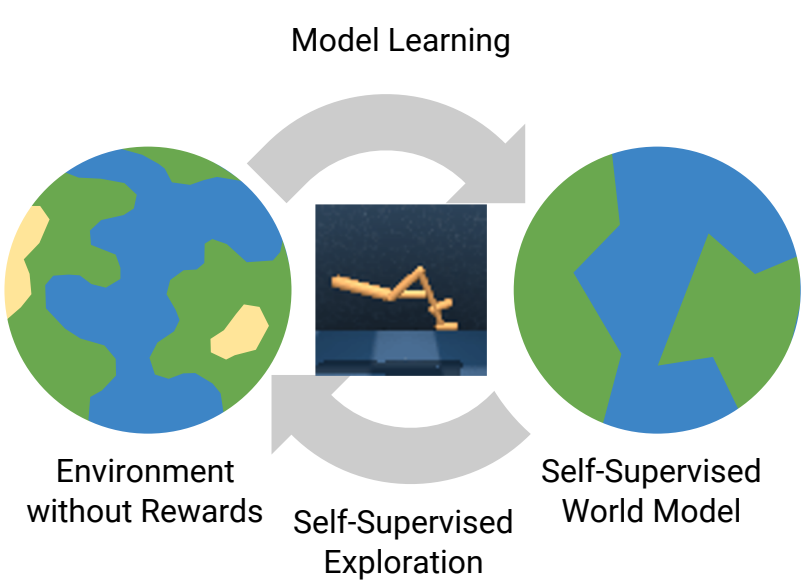


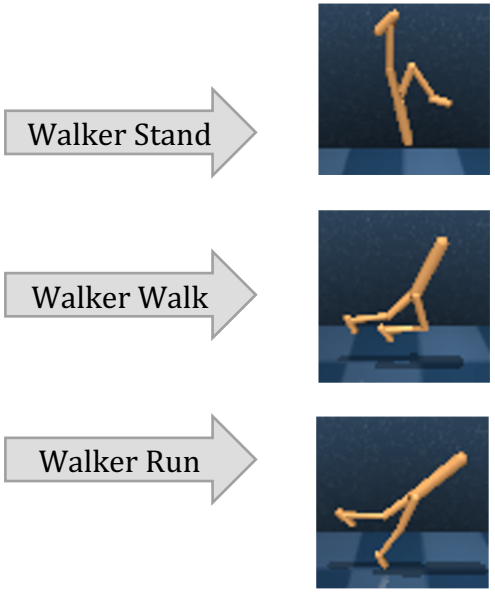


- Pathak *et.al.,* “Curiosity-driven Exploration by Self-
- neural
- hability”.
- rn skills
- ICML, 2019
- Sharma *et al.*, “Dynamics-Aware Unsupervised Discovery of Skills”. ICLR, 2020.

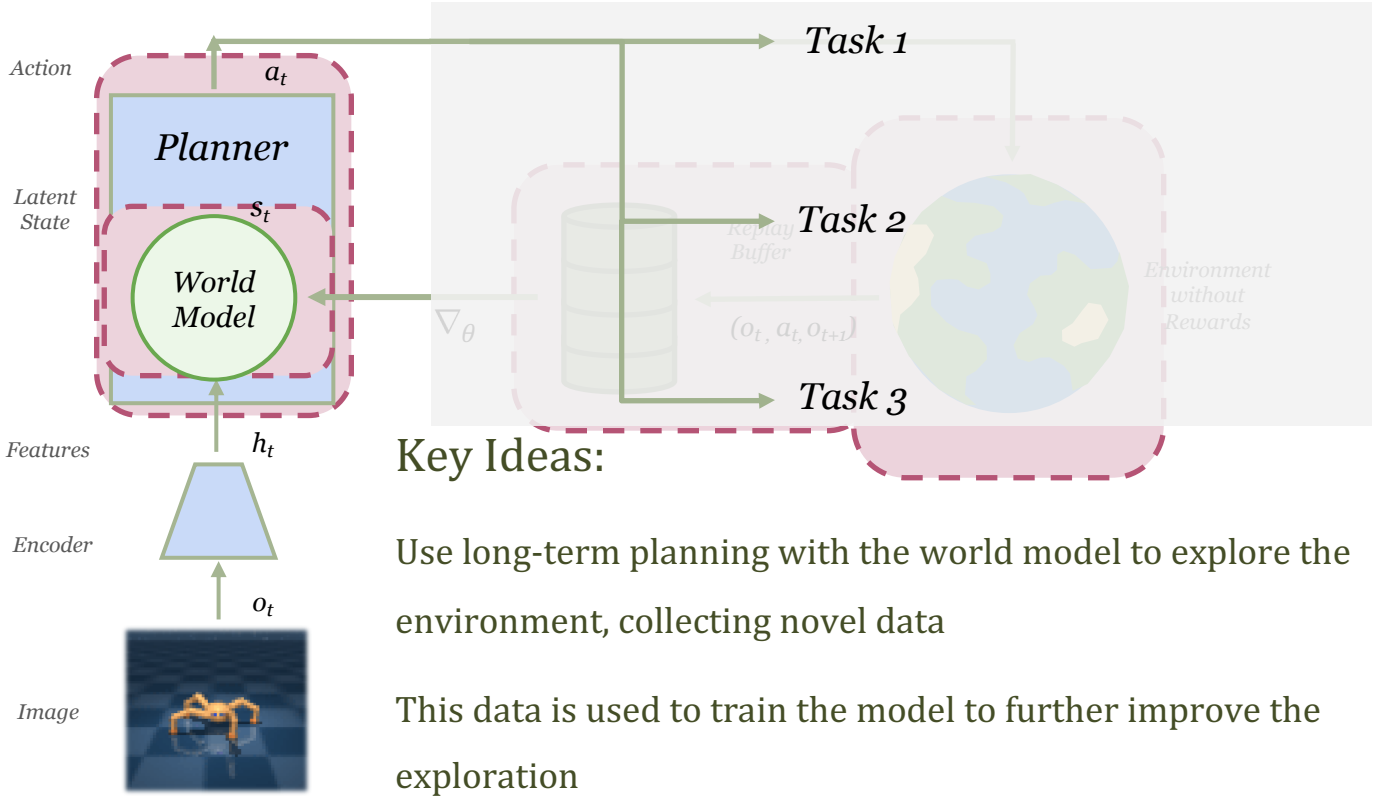
- Finn, Levine., “deep visual foresight for planning robot motion”. ICRA, 2017.
- Pathak *et al.,* “Zero-Shot Visual Imitation”. ICLR, 2018.

Planning to Explore





Planning to Explore



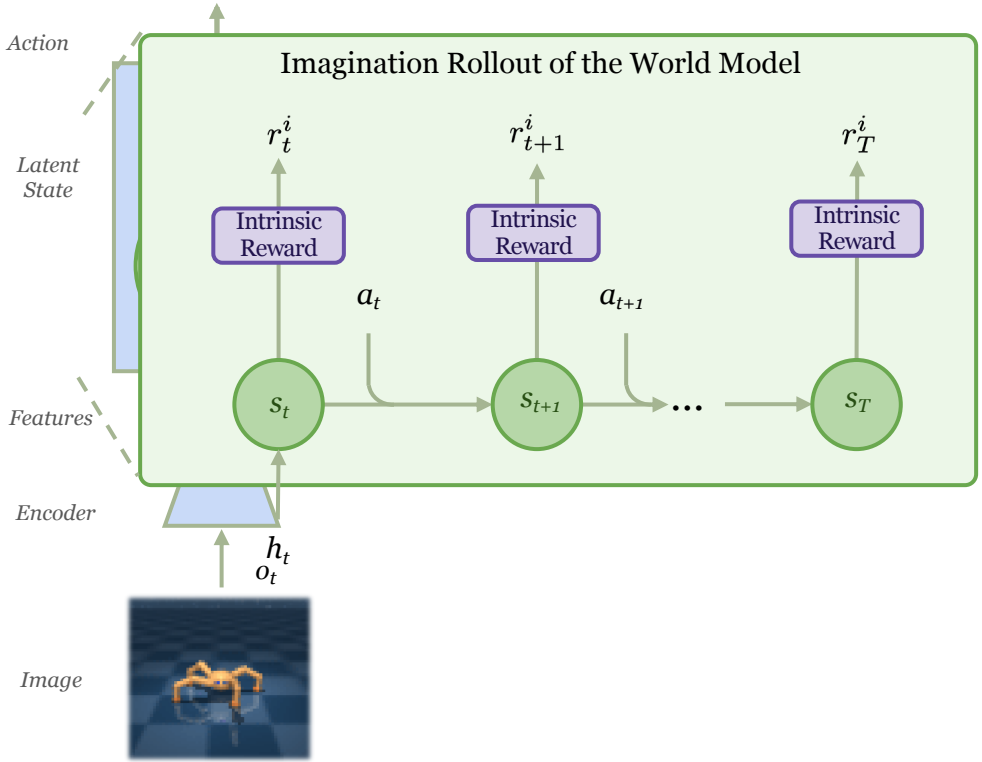
Key Ideas:

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

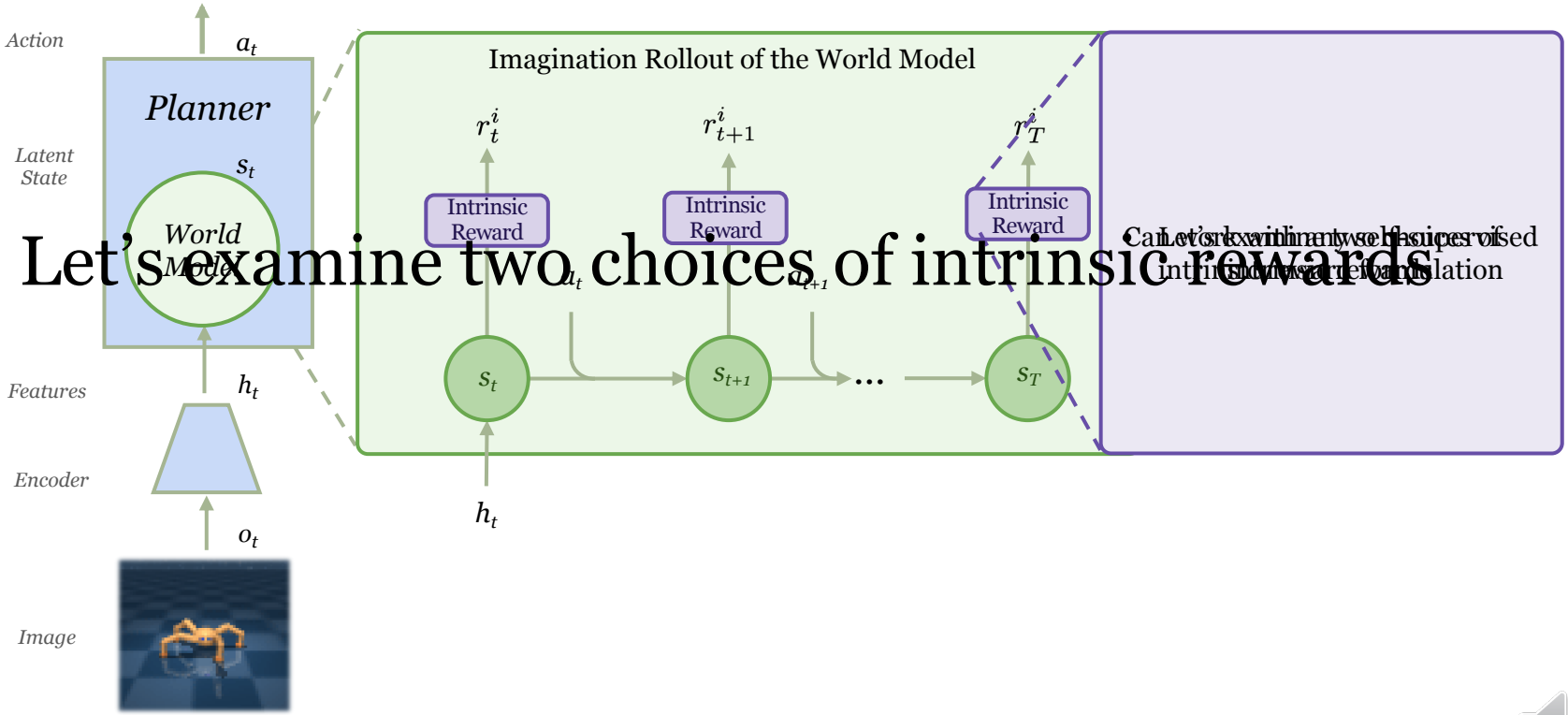
This data is used to train the model to further improve the exploration

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

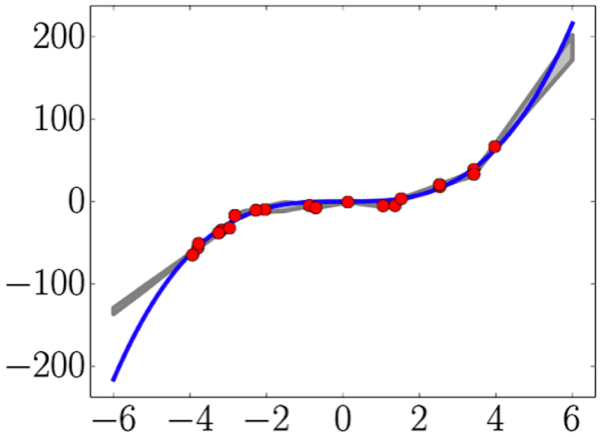
Planning to Explore



Planning to Explore

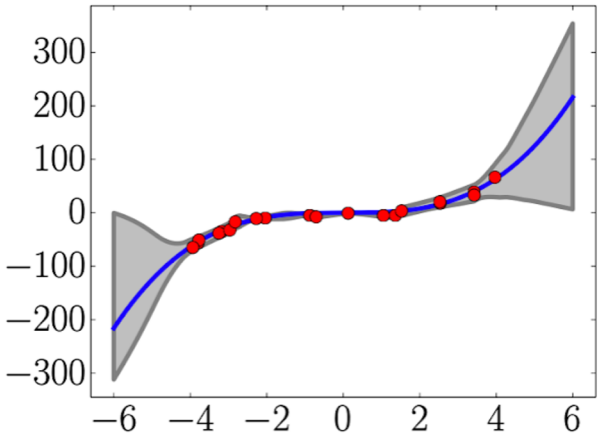


Model Error



- Schmidhuber'01, Pathak'17
- High outside training data

Model Disagreement



- Lakshminarayanan'17, Pathak'19, Shyam'19
- High only outside training data

- True function
- Data points
- Prediction of the model
- Model Variance

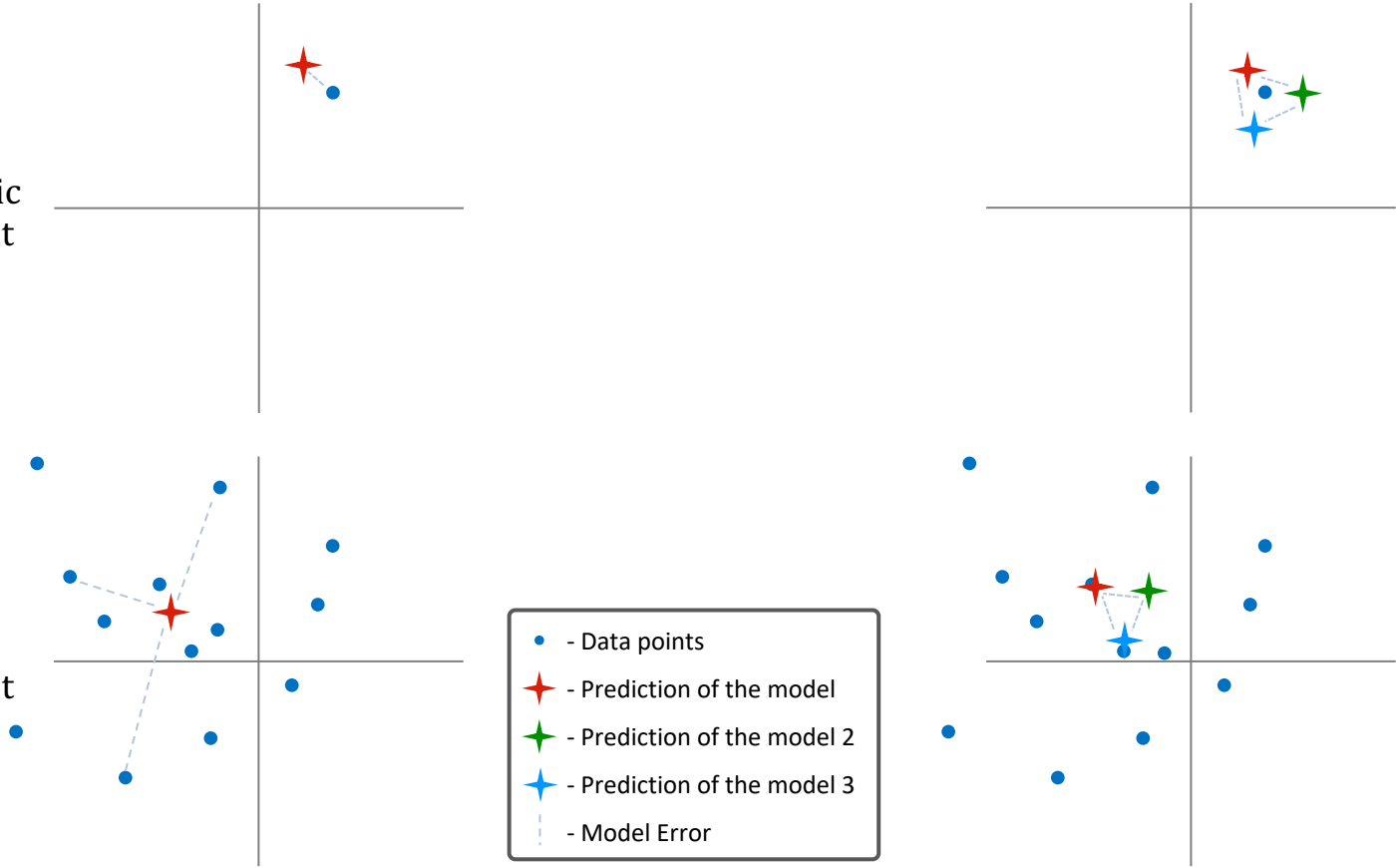
[Figure from Lakshminarayanan'17]

Model Error

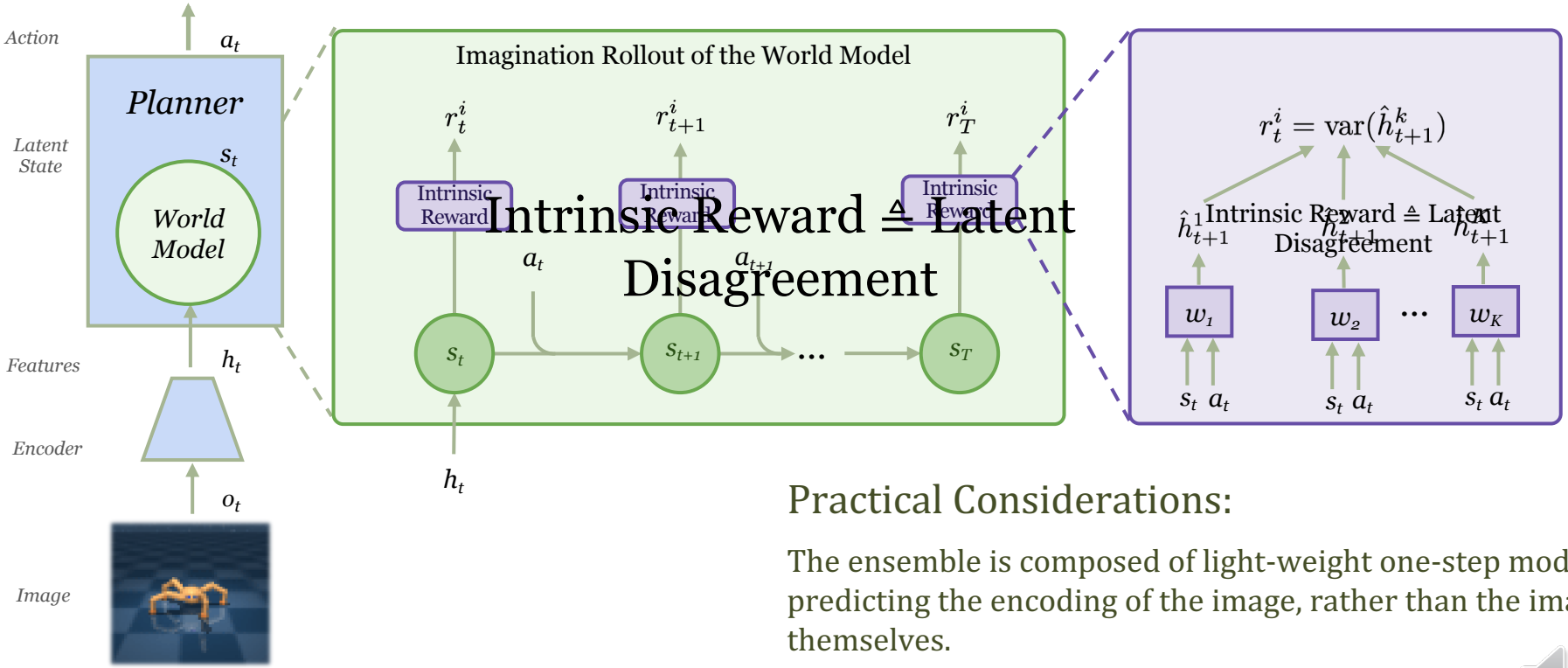
Model Disagreement

Deterministic Environment

Stochastic Environment



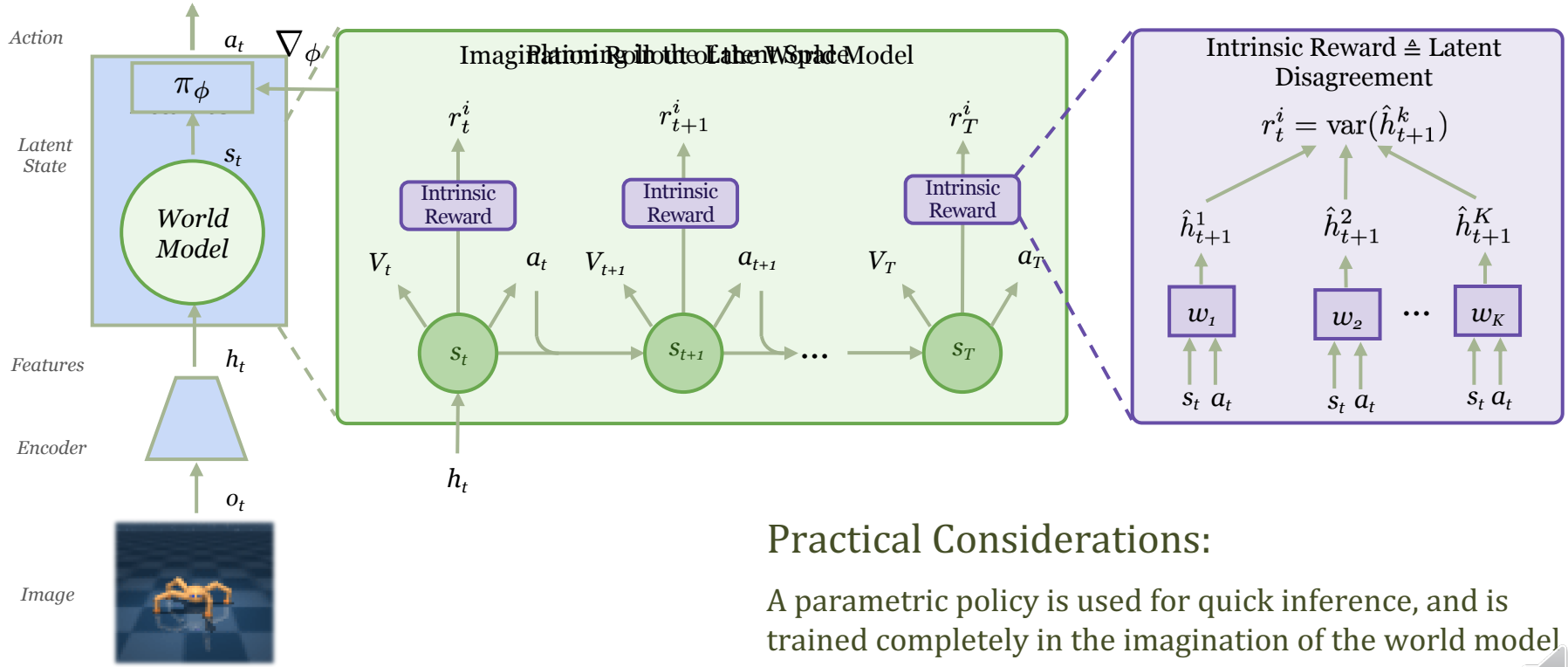
Planning to Explore



Practical Considerations:

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

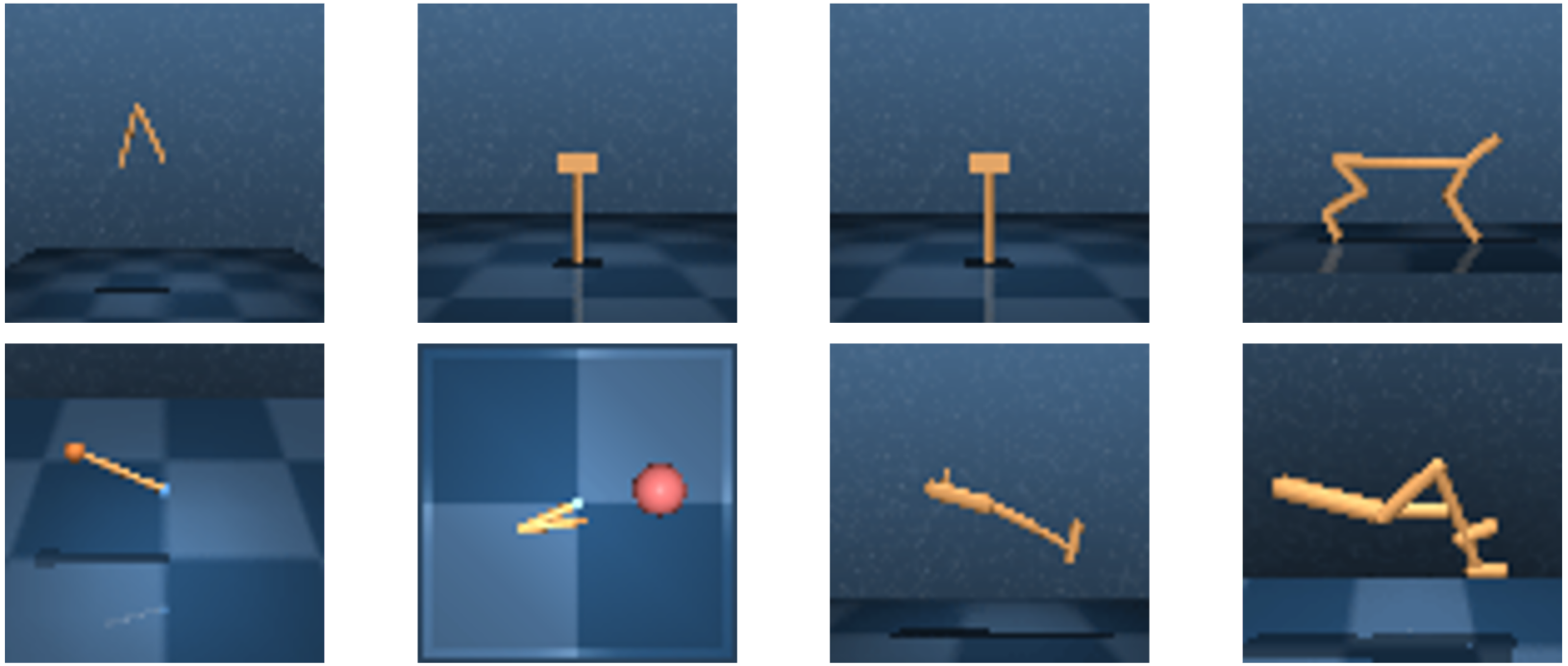
Planning to Explore



Practical Considerations:

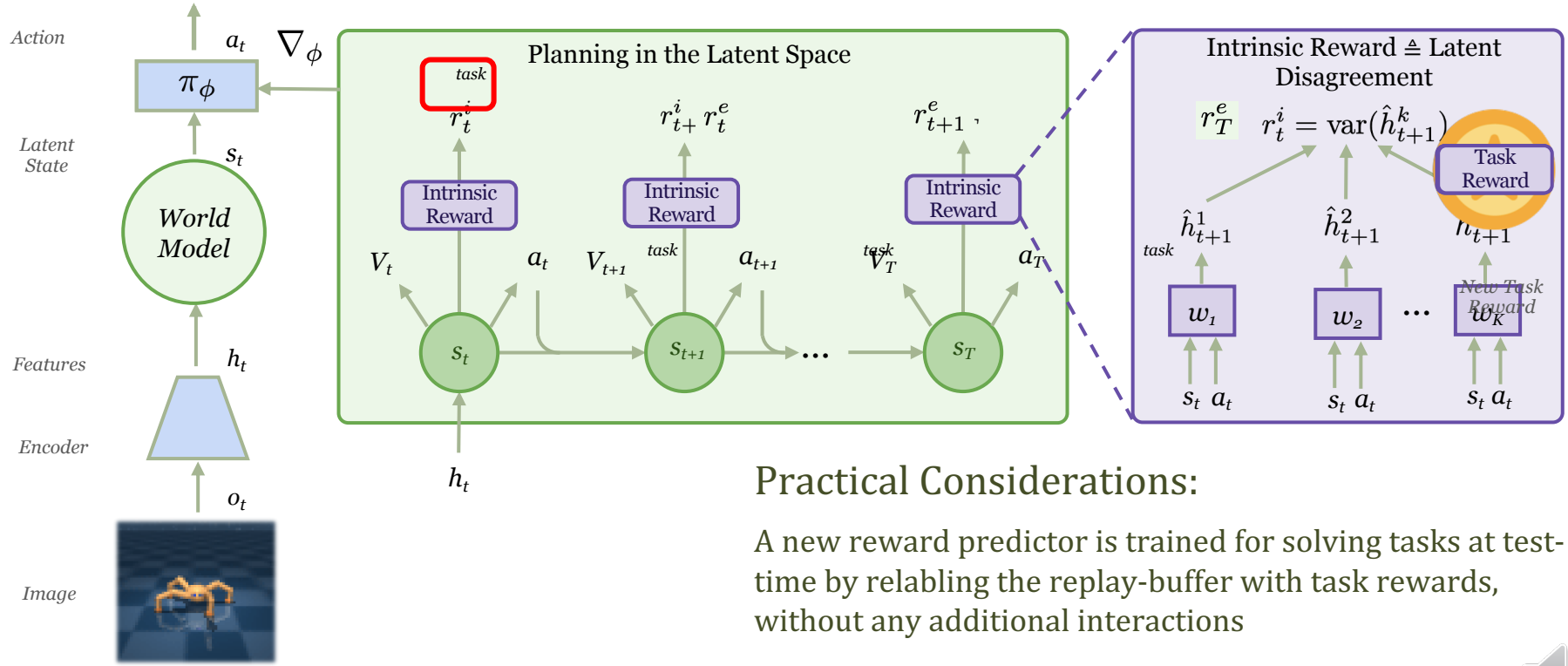
A parametric policy is used for quick inference, and is trained completely in the imagination of the world model

Self-Supervised Exploration Results



How do we go from exploration to solving tasks?

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

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

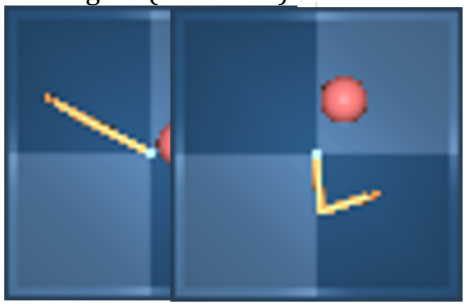
Zero-Shot Reinforcement Learning

Our Agent (zero-shot) Oracle (supervised)

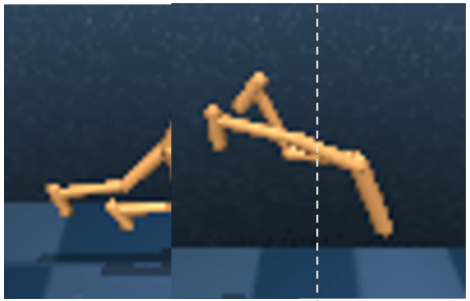
Cheetah Run

Hopper Hop

Our Agent (zero-shot) Oracle (supervised)

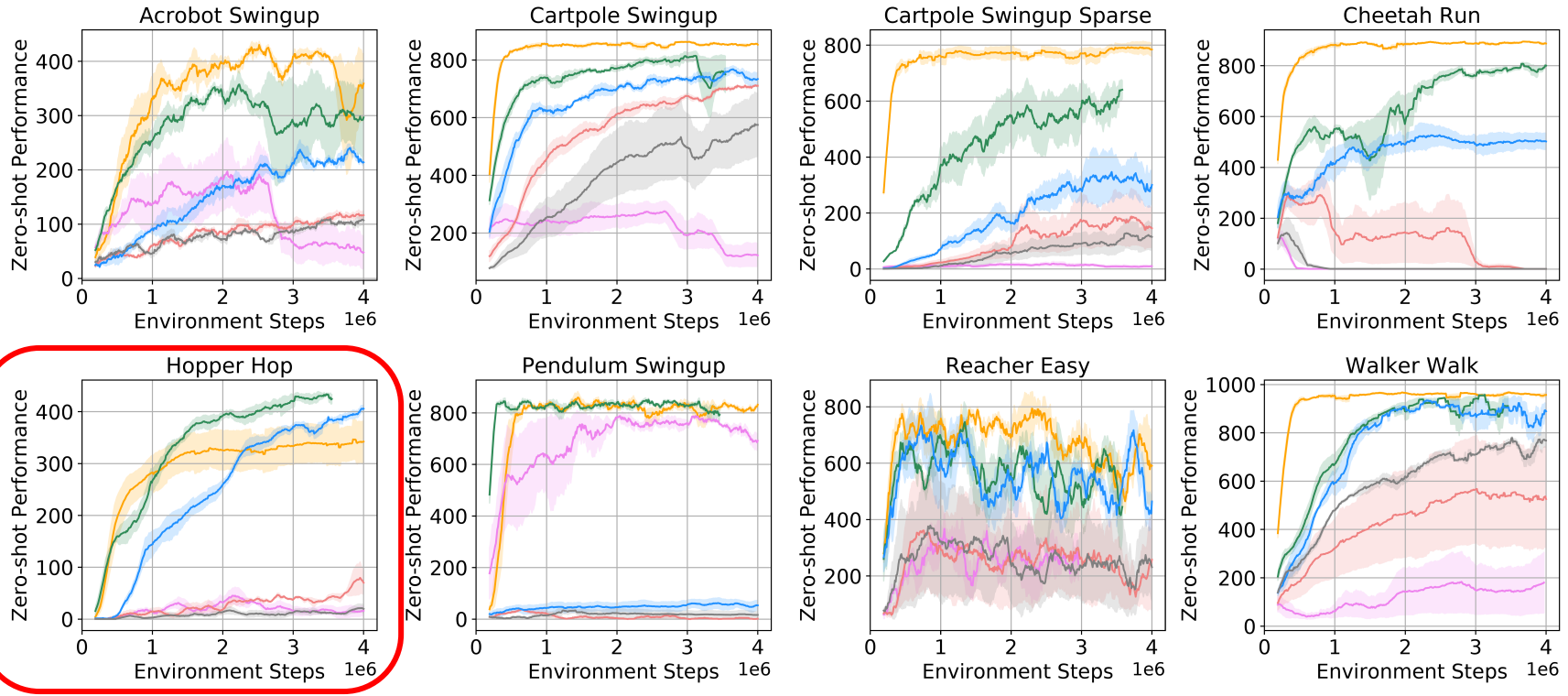


Reacher Easy



Walker Walk

Zero-Shot Reinforcement Learning



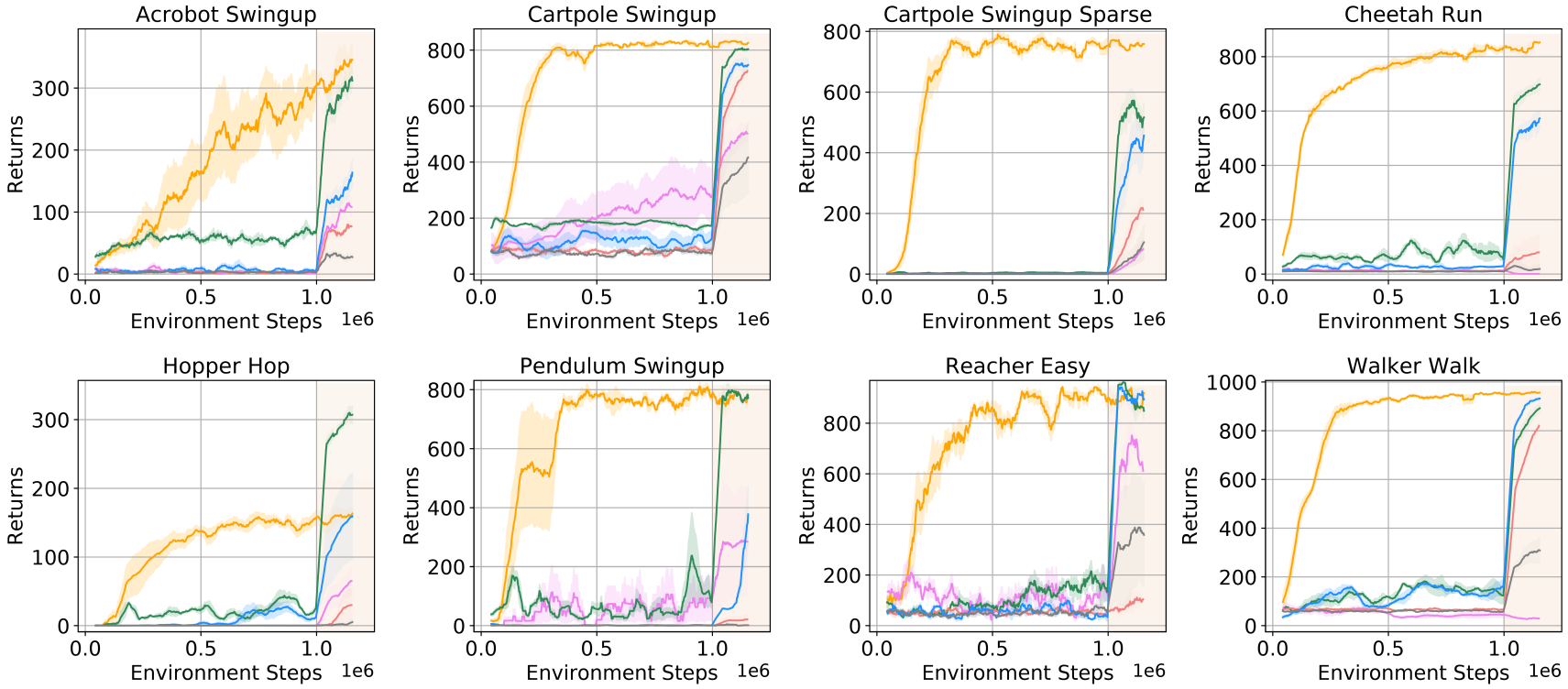
Experiments Outline

1. Solving a new-task in zero-shot

2. What if we add 20 supervised episodes?

3. Multi-task performance

Few-Shot Adaptation

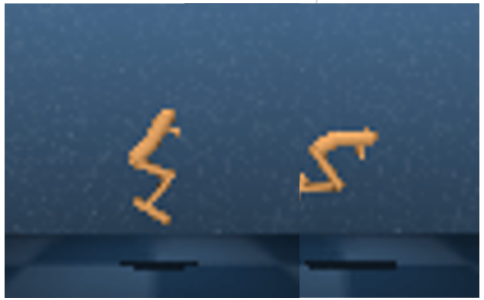


Few-Shot Adaptation

Our Agent (few-shot)

Oracle (supervised)

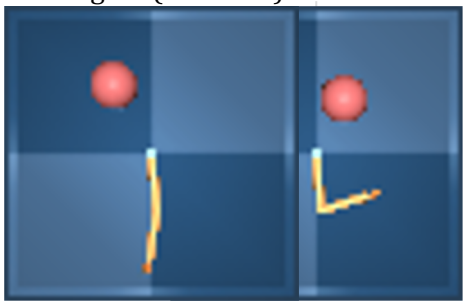
Cheetah Run



Hopper Hop

Our Agent (few-shot)

Oracle (supervised)



Reacher Easy

Walker Walk

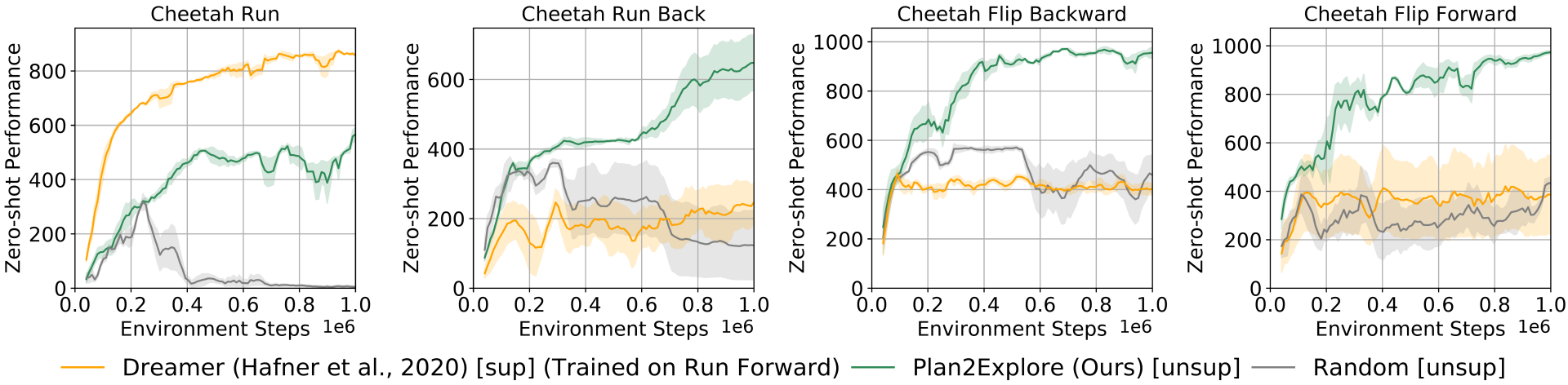
Experiments Outline

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?



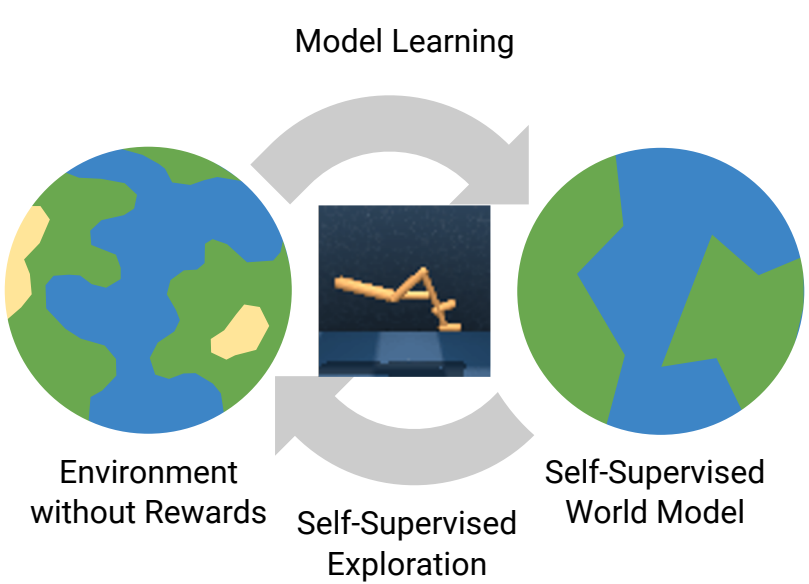
Key Takeaways

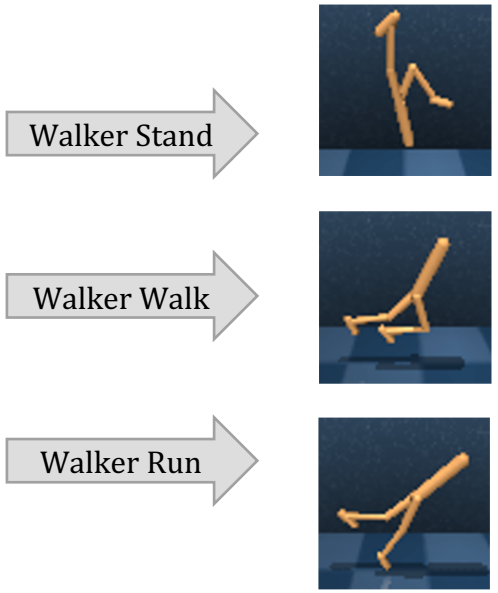
Self-supervised performance comparable to Supervised Oracle

Few supervised samples provide large boost in performance

Perform several tasks by training dynamics only once

Planning to Explore





Code and videos at:

https://ramanans1.github.io/plan2explore/

Thank you!