Understanding and Improving Model-Based Deep Reinforcement Learning

Jessica B. Hamrick
jhamrick@deepmind.com

DeepMind

Imperial College London
ICARL Seminar Series
15 February 2023

Reasoning with a world model

*"If the organism carries a **'small-scale model' of external reality** and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilise the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it."*

–Kenneth Craik, *The Nature of Explanation* (1943)

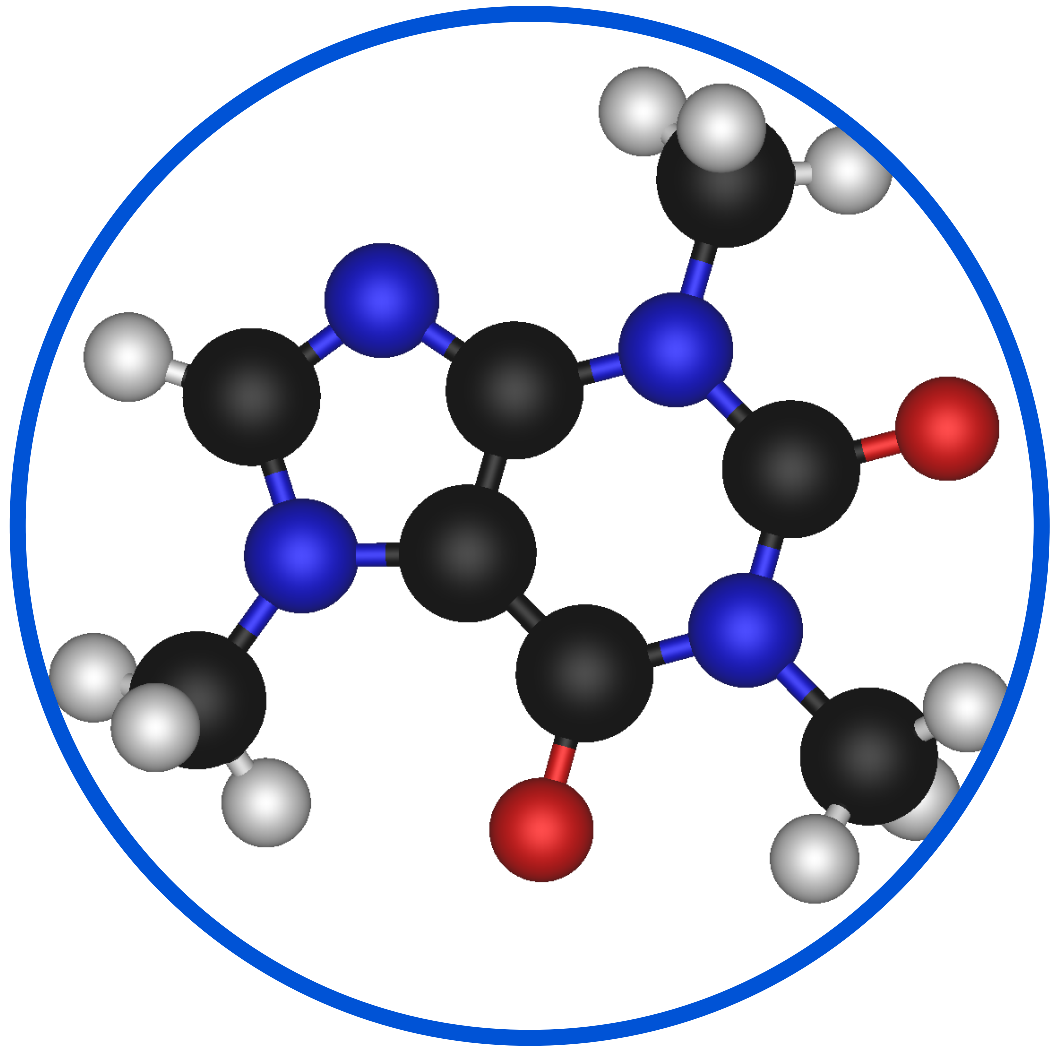
Silver et al. (2016)

Silver et al. (2016)

OpenAI et al. (2019)

Silver et al. (2016)

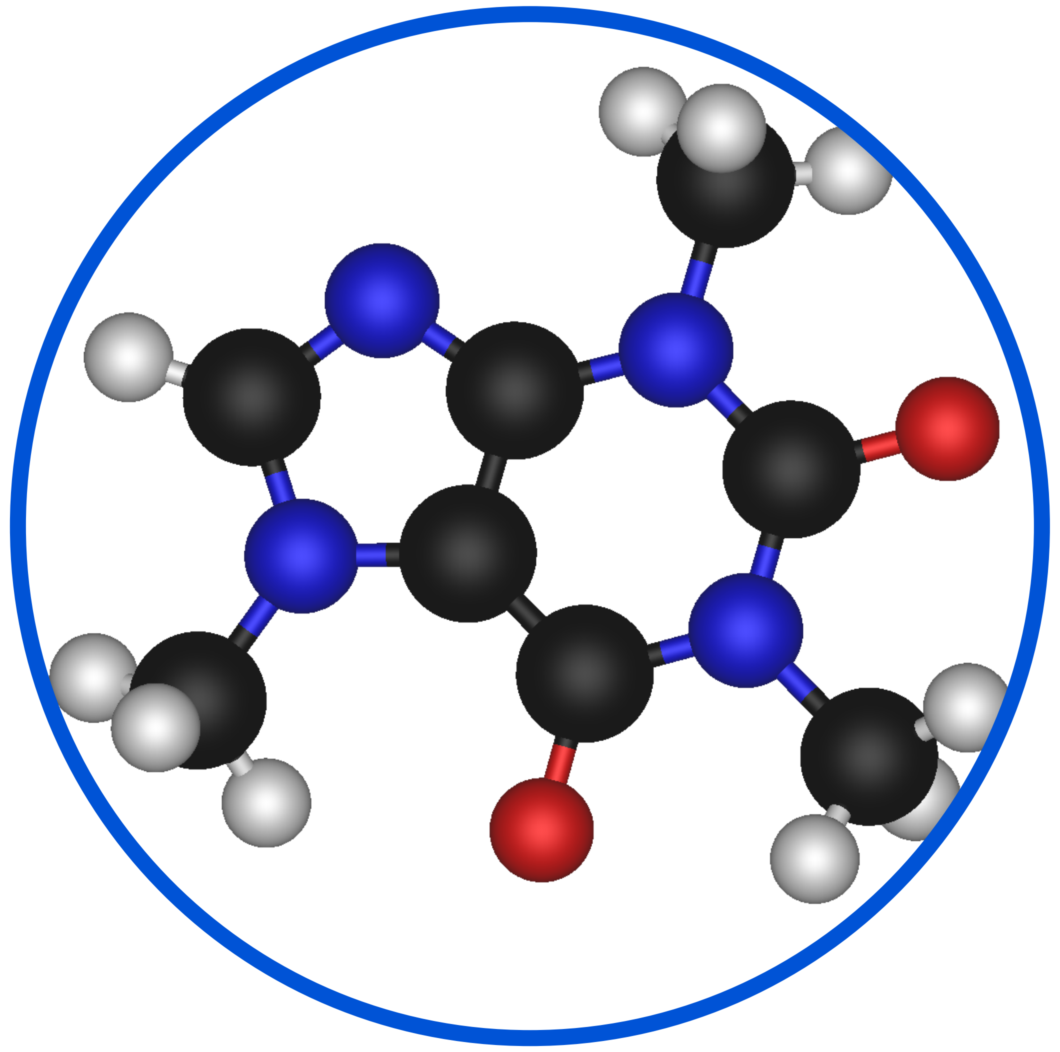
OpenAI et al. (2019)



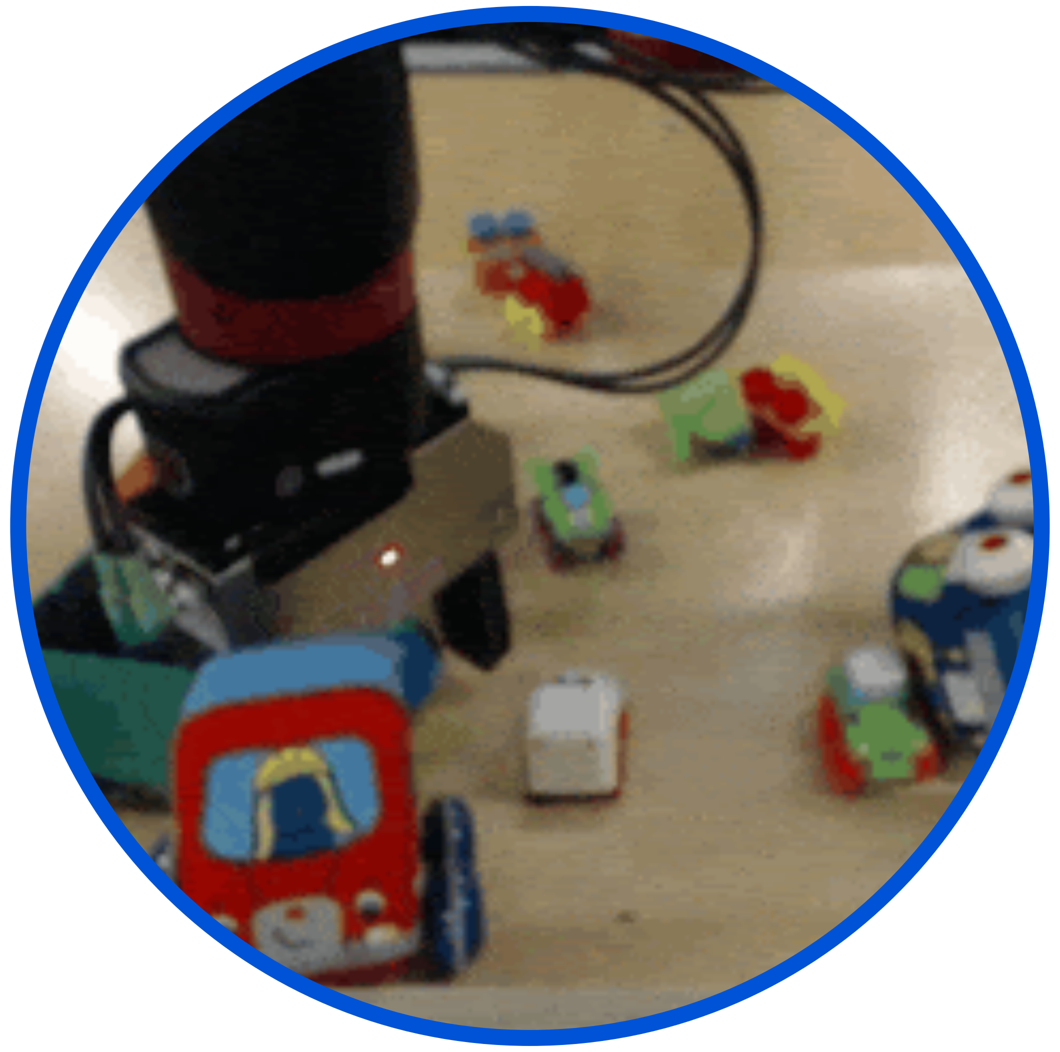
Segler et al. (2018)

Silver et al. (2016)

OpenAI et al. (2019)



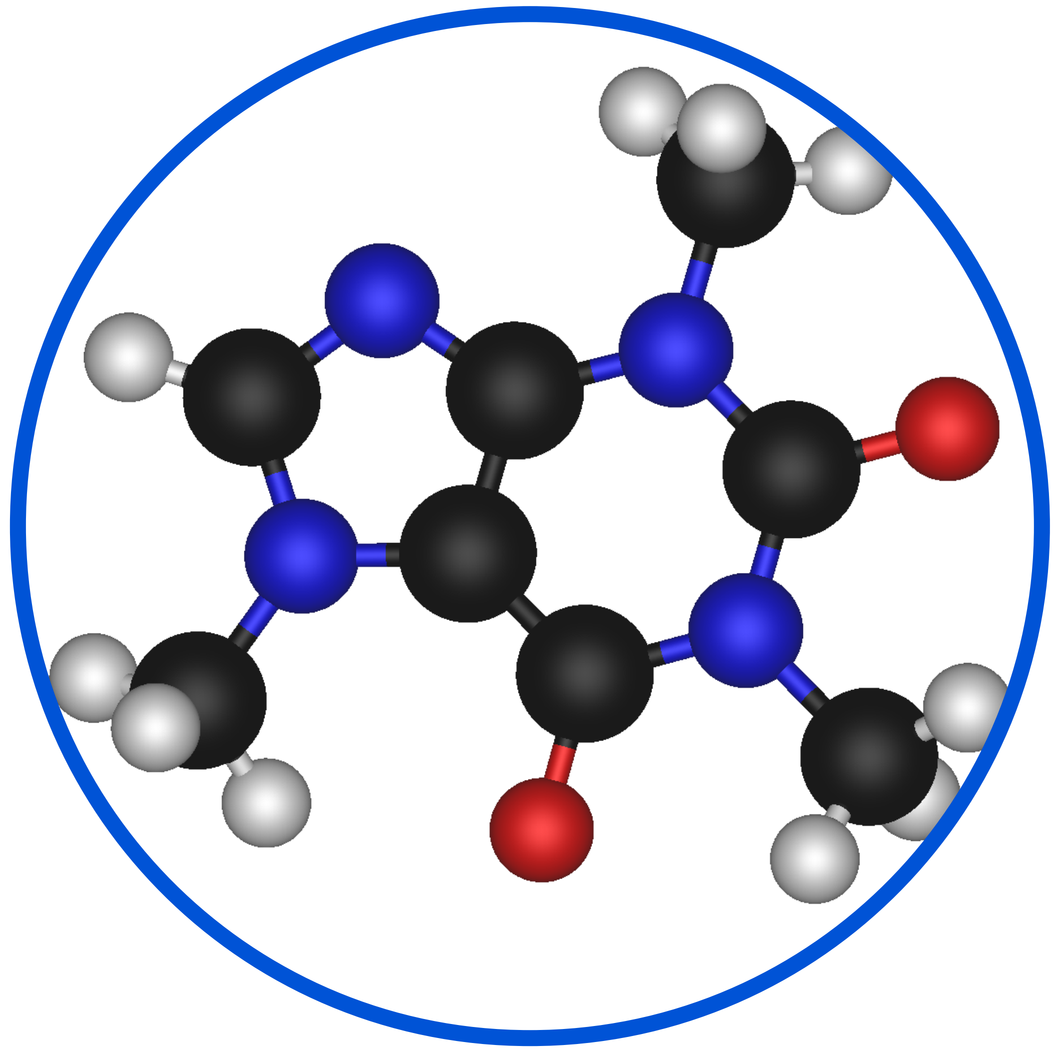
Segler et al. (2018)

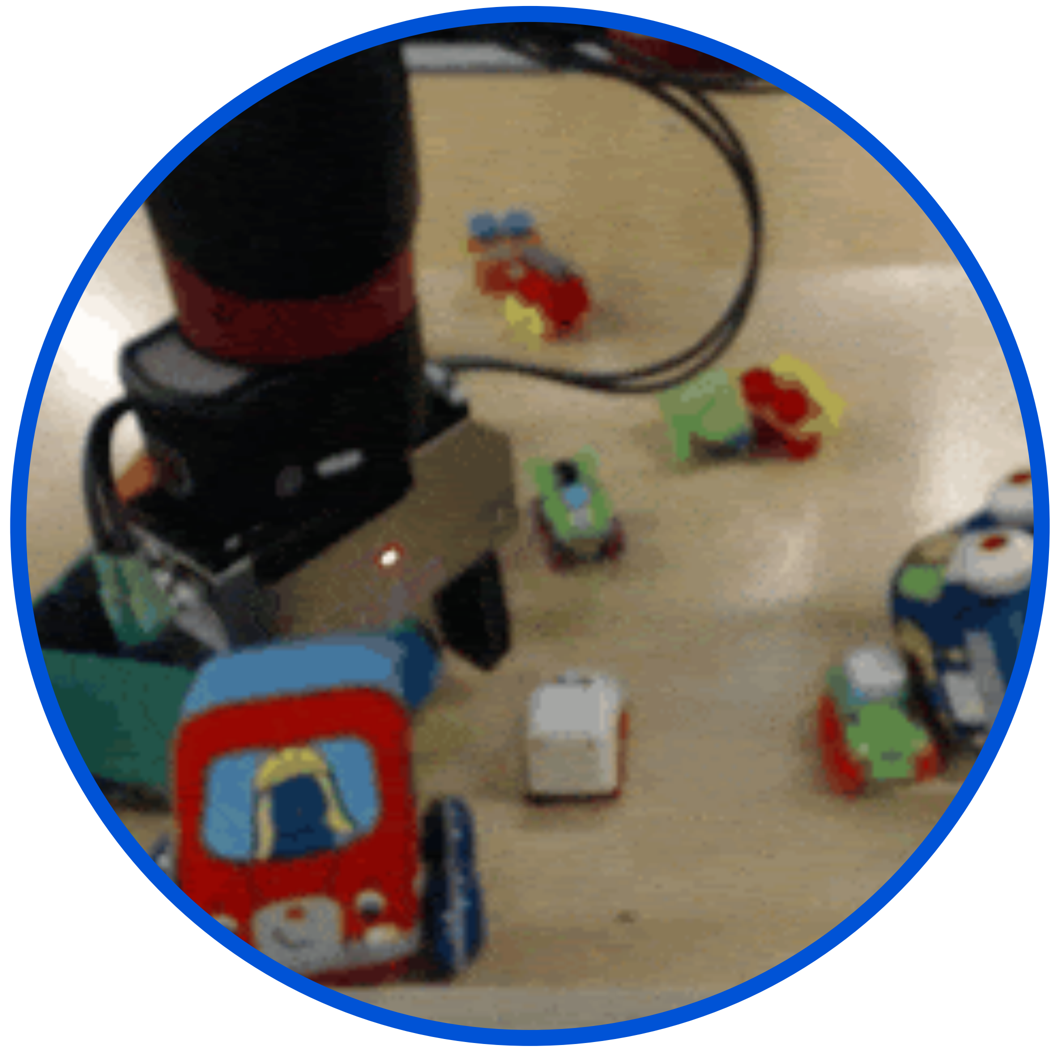


Finn et al. (2018)


Silver et al. (2016)

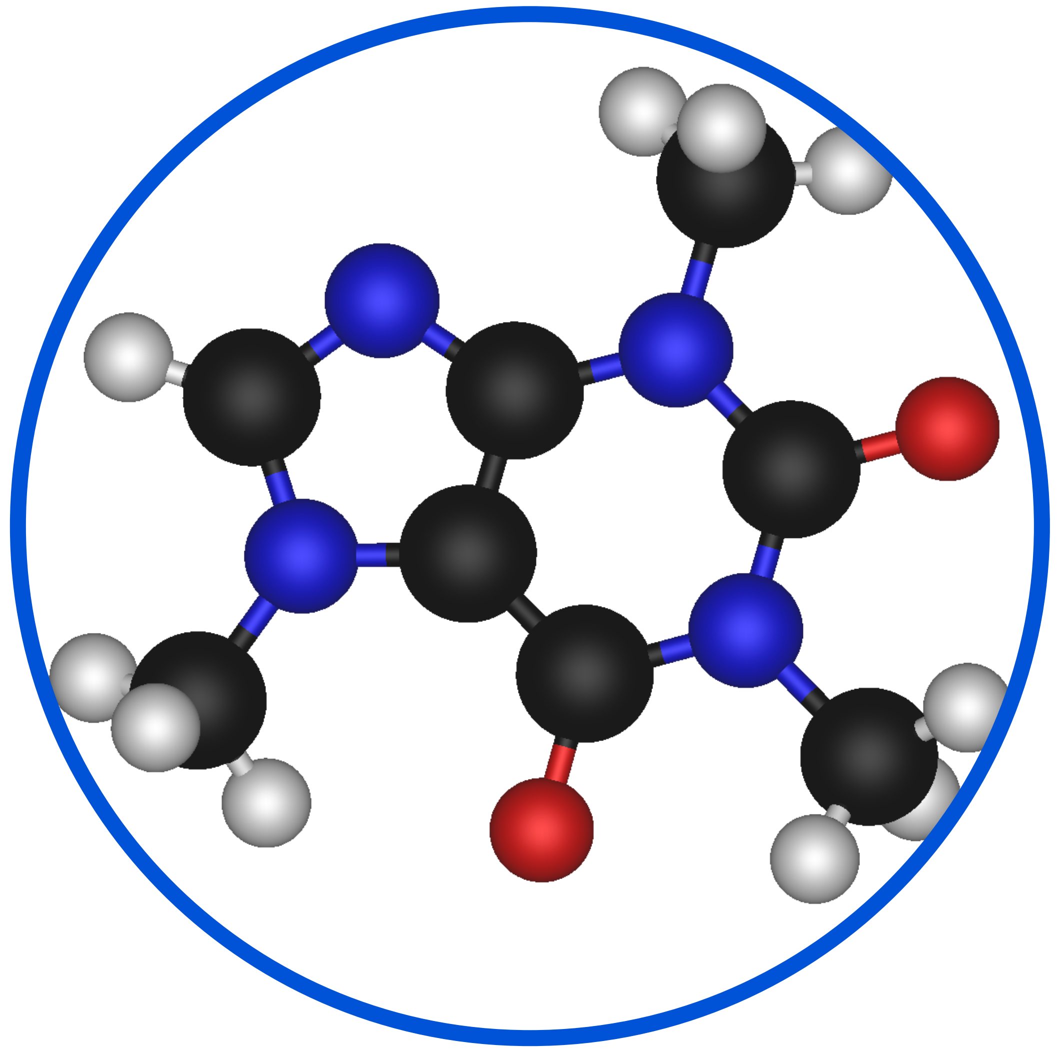

OpenAI et al. (2019)


Segler et al. (2018)

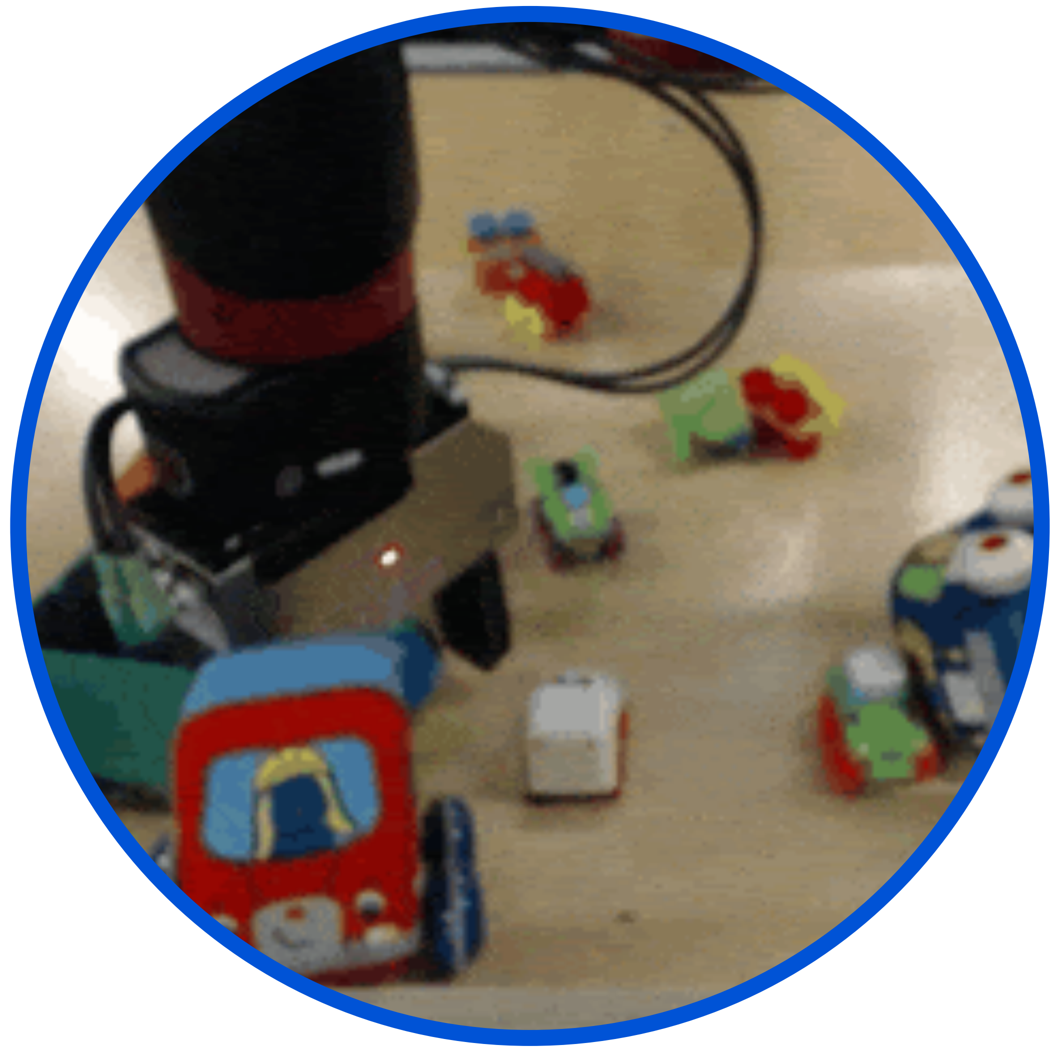

Finn et al. (2018)

Silver et al. (2016)

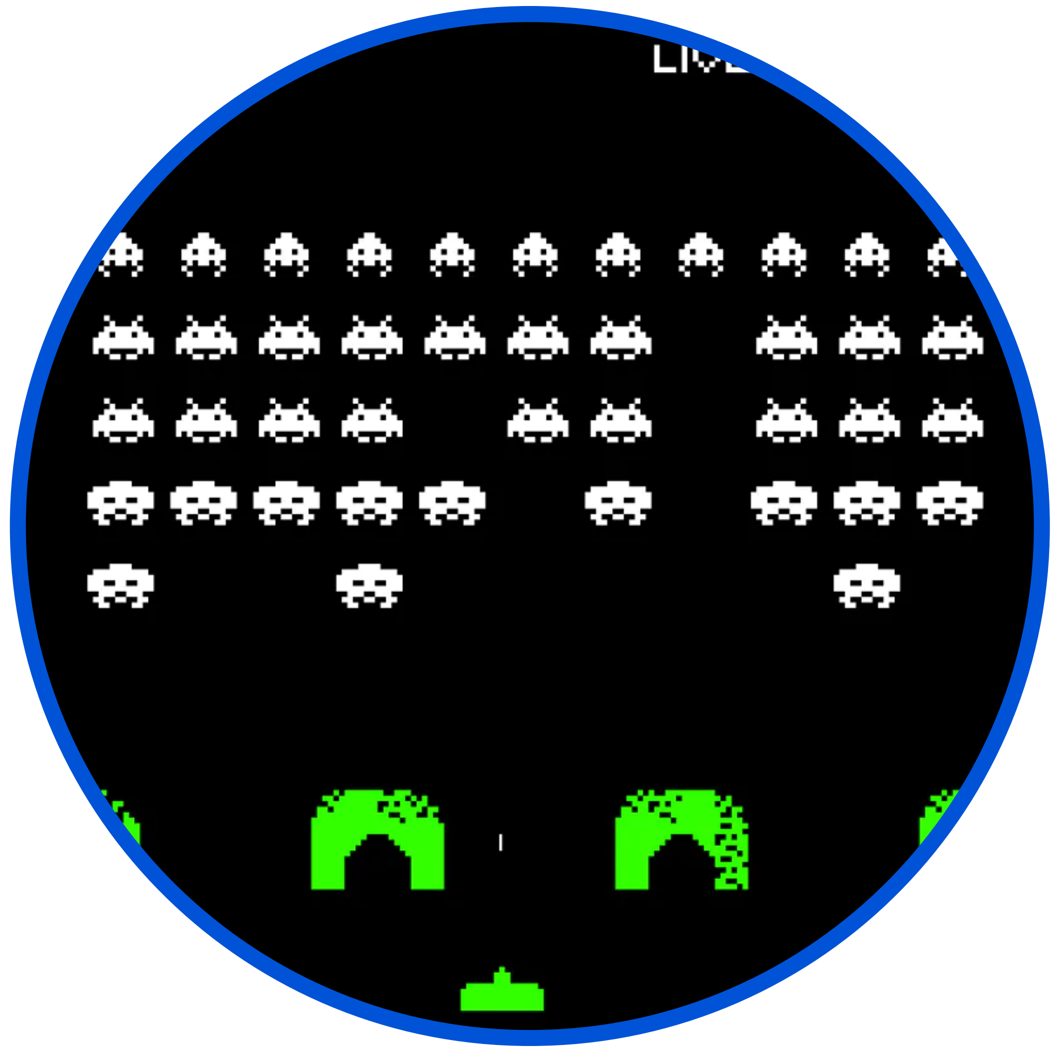
OpenAI et al. (2019)



Segler et al. (2018)



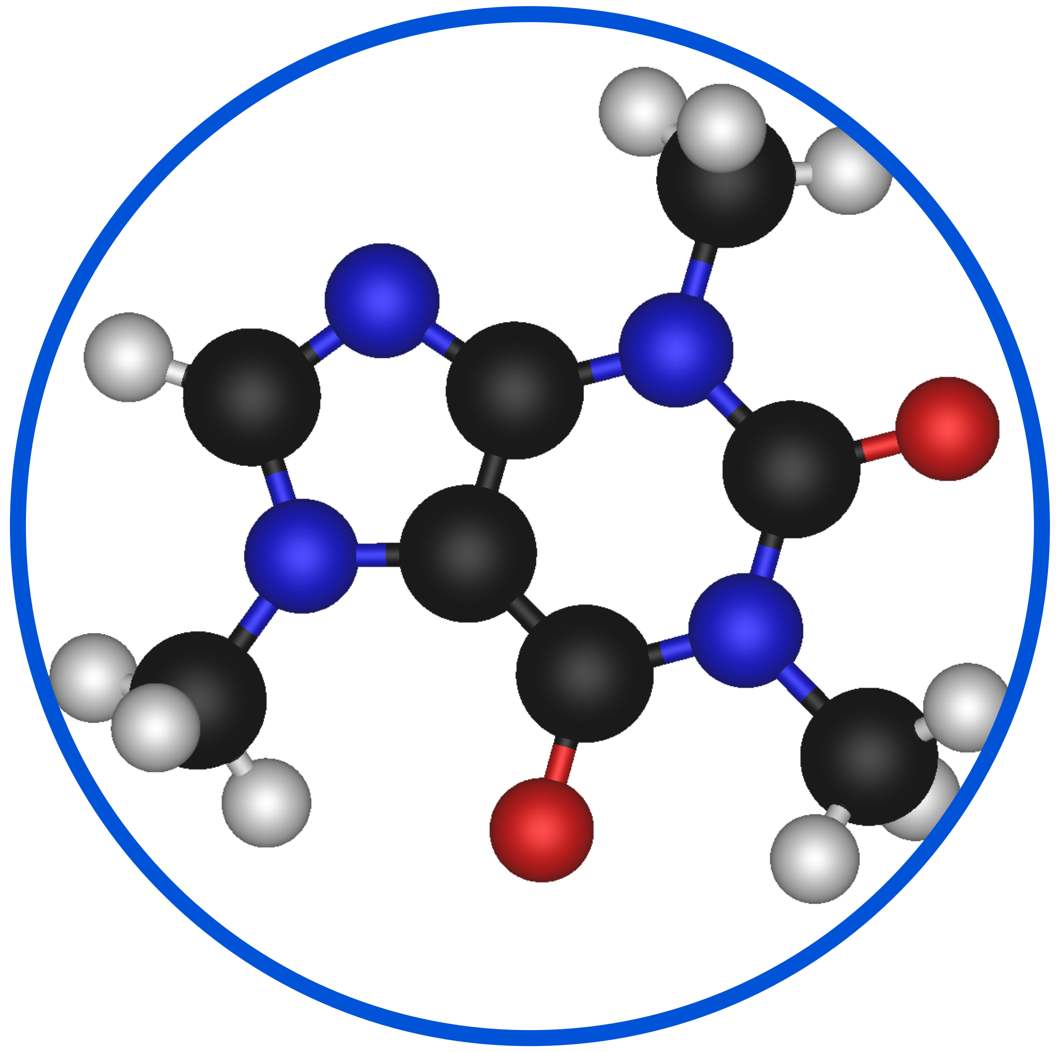
Finn et al. (2018)



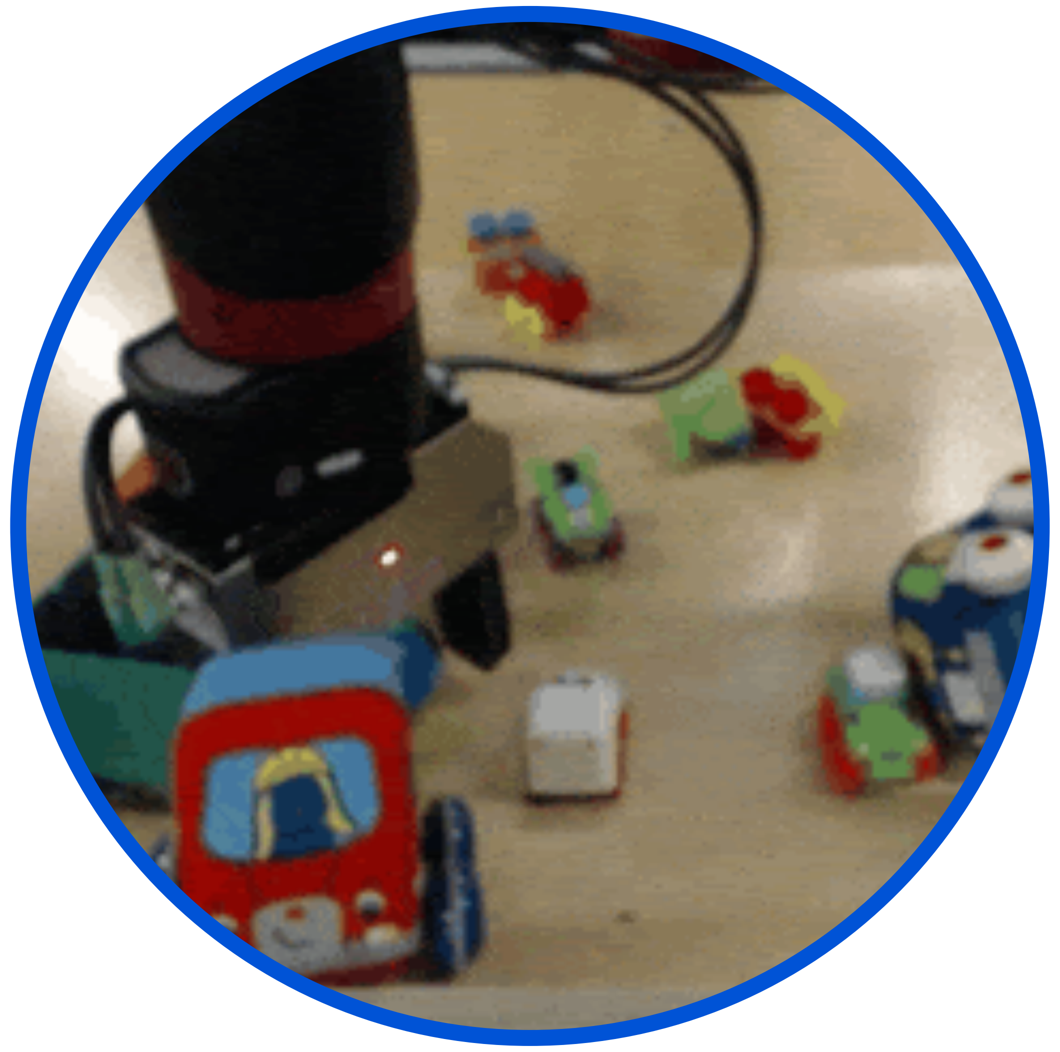
Schrittwieser et al. (2020)

Silver et al. (2016)

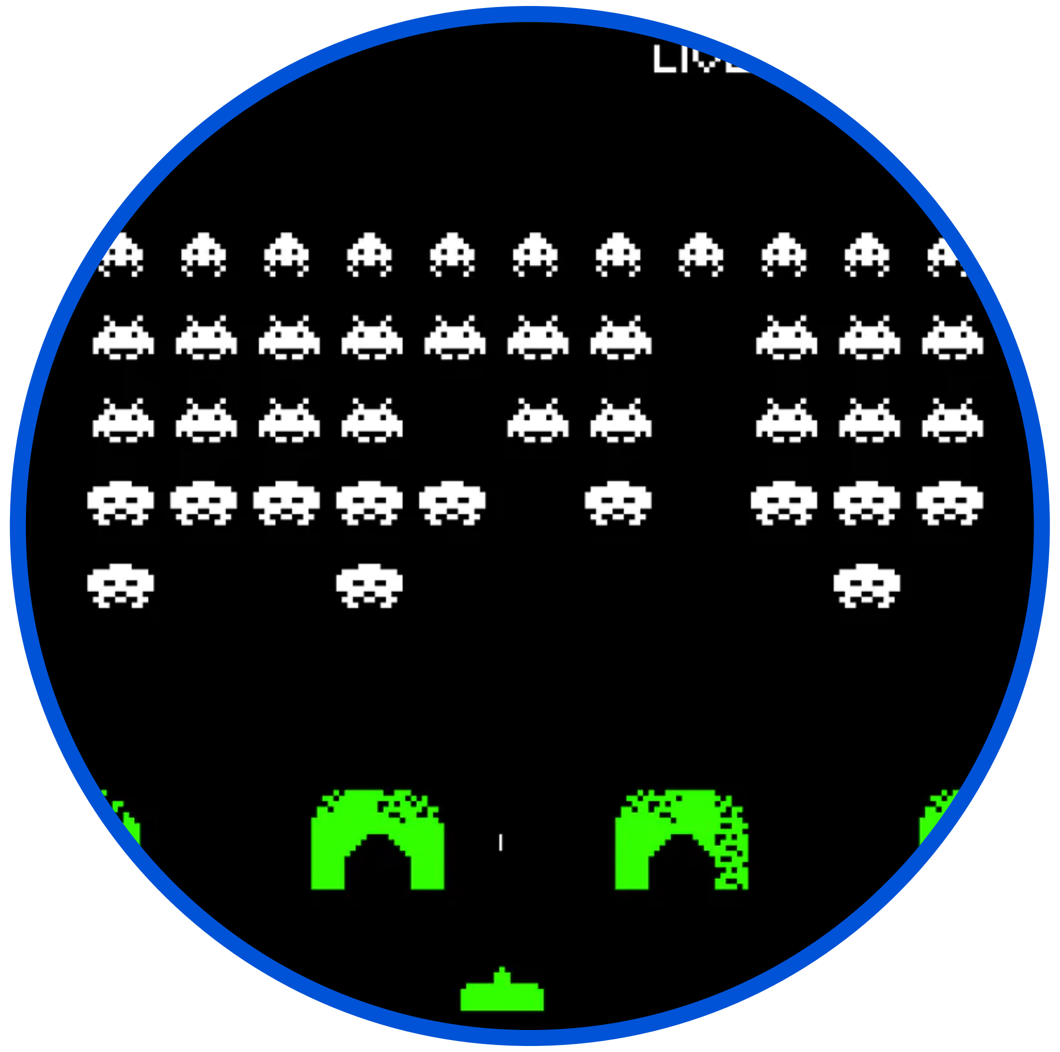
OpenAI et al. (2019)



Segler et al. (2018)



Finn et al. (2018)

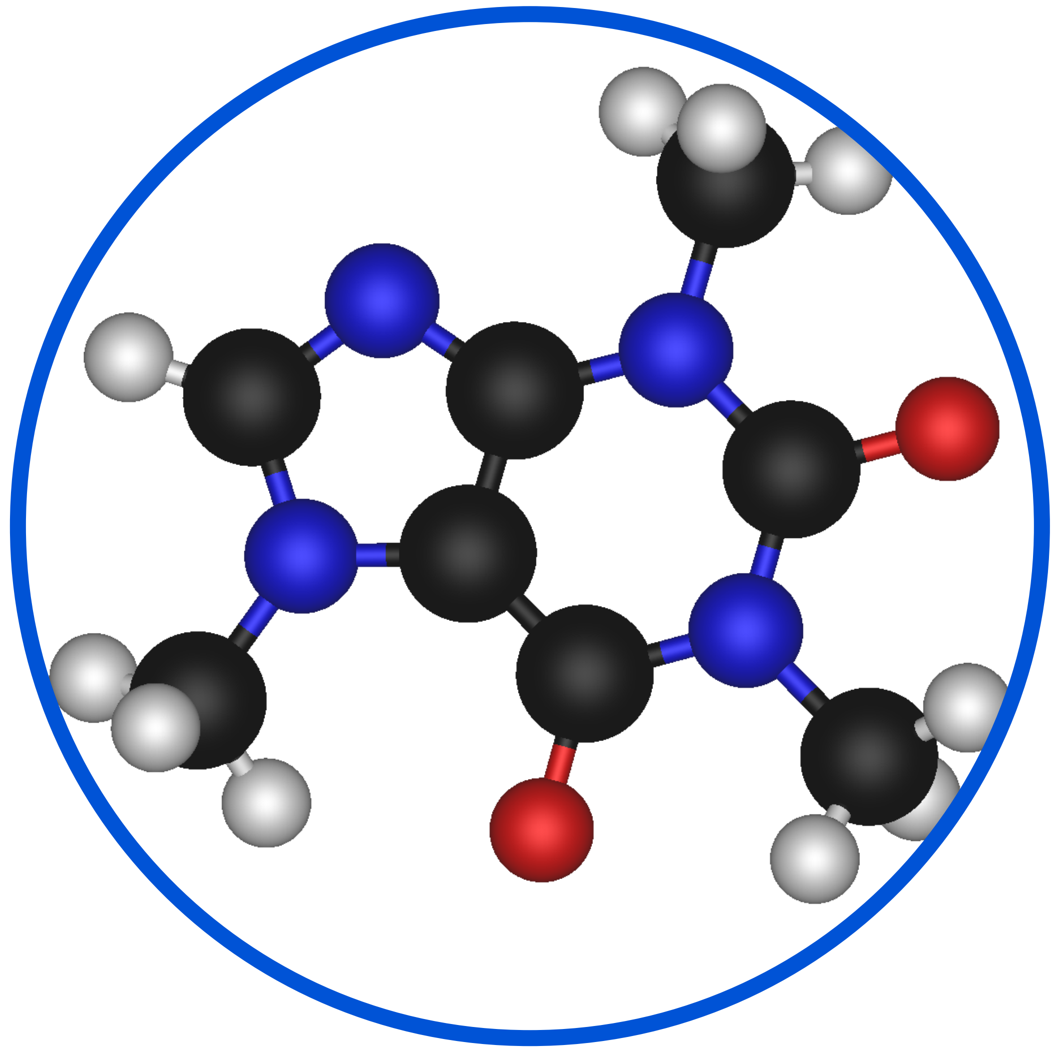


Schrittwieser et al. (2020)

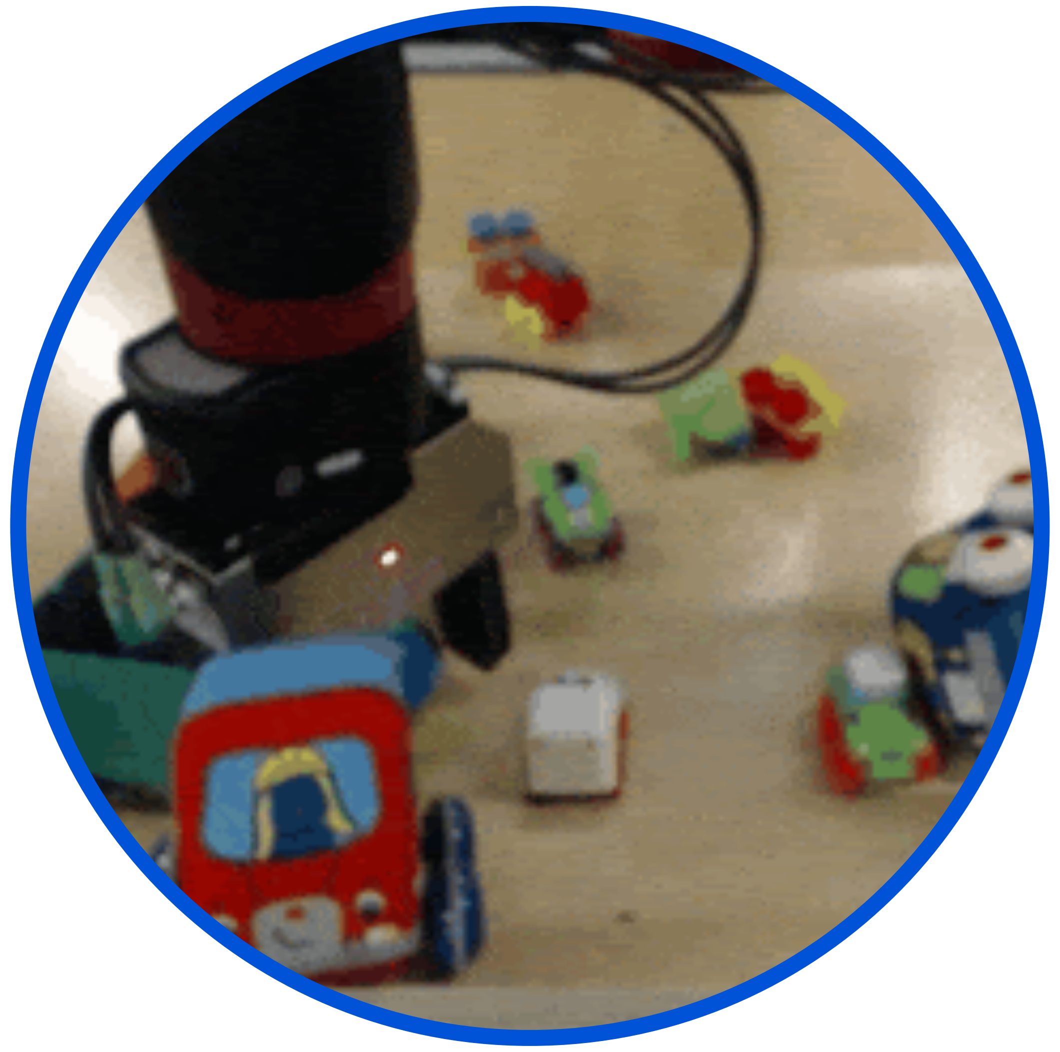
Luo et al. (2019)

Silver et al. (2016)

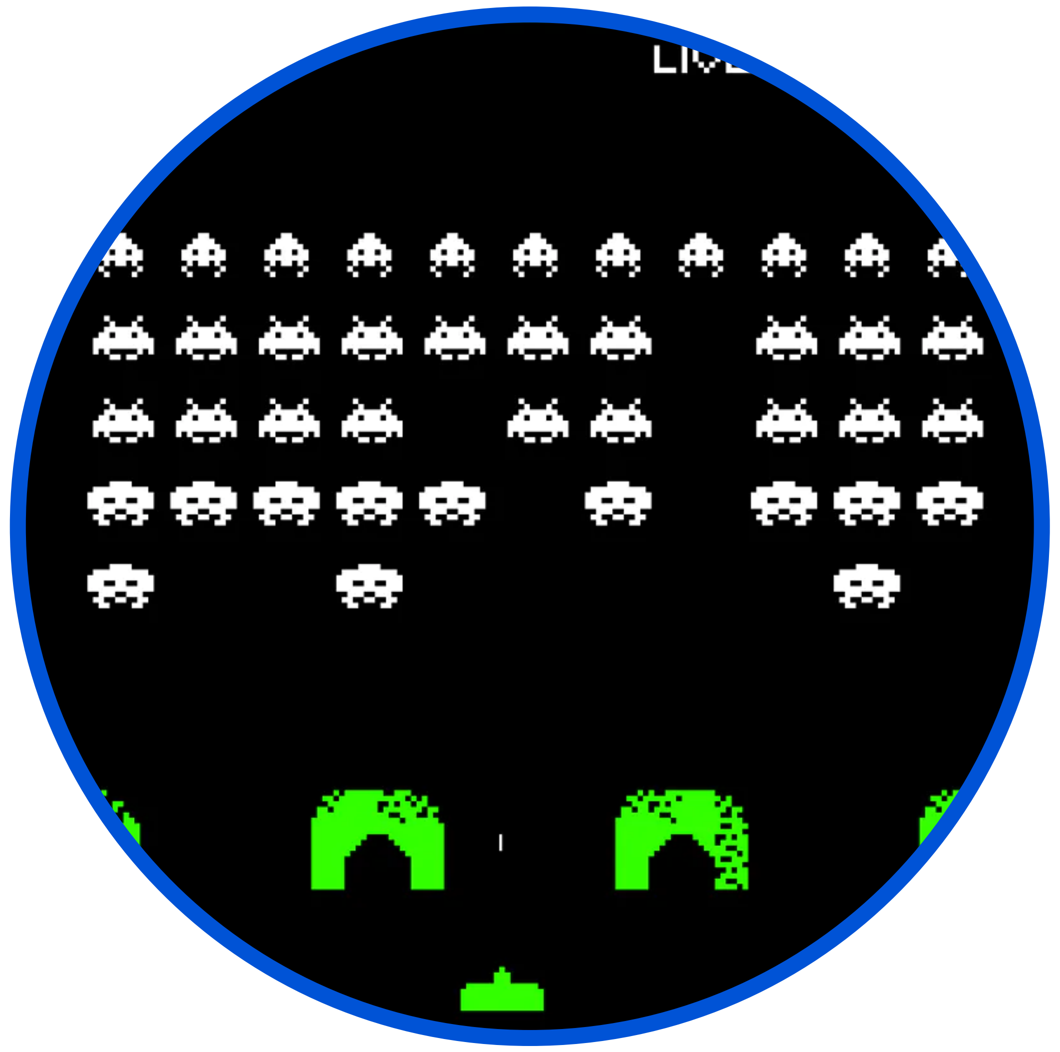
OpenAI et al. (2019)



Segler et al. (2018)

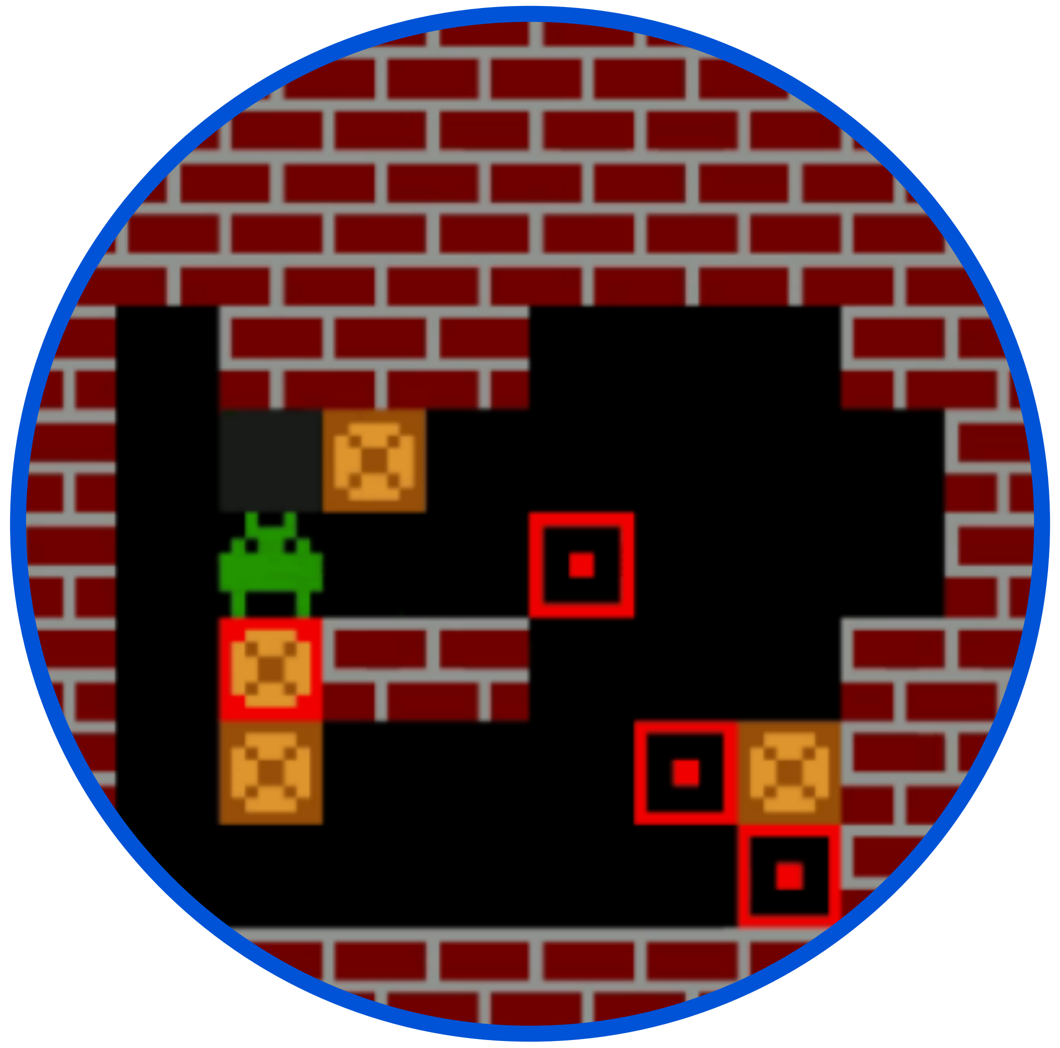


Finn et al. (2018)



Schrittwieser et al. (2020)

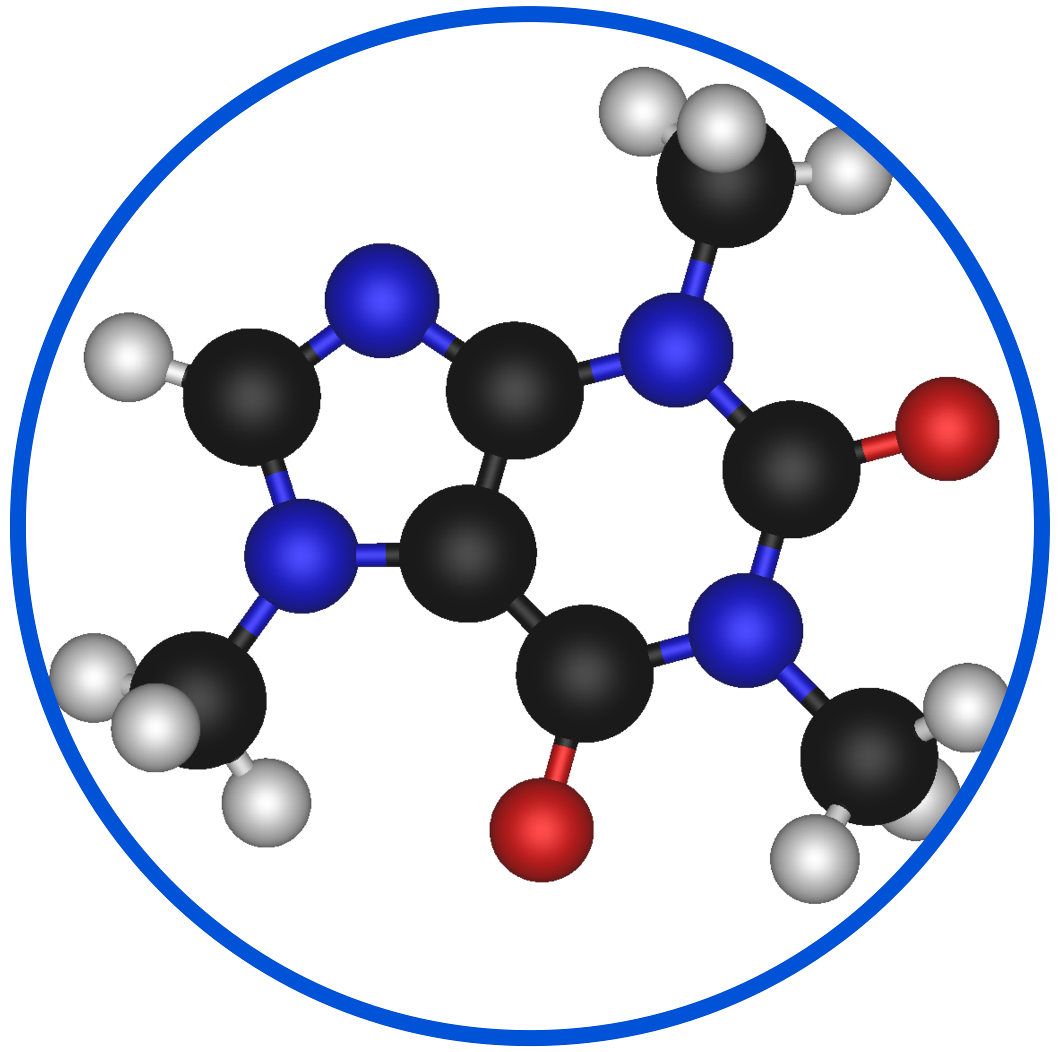
Luo et al. (2019)



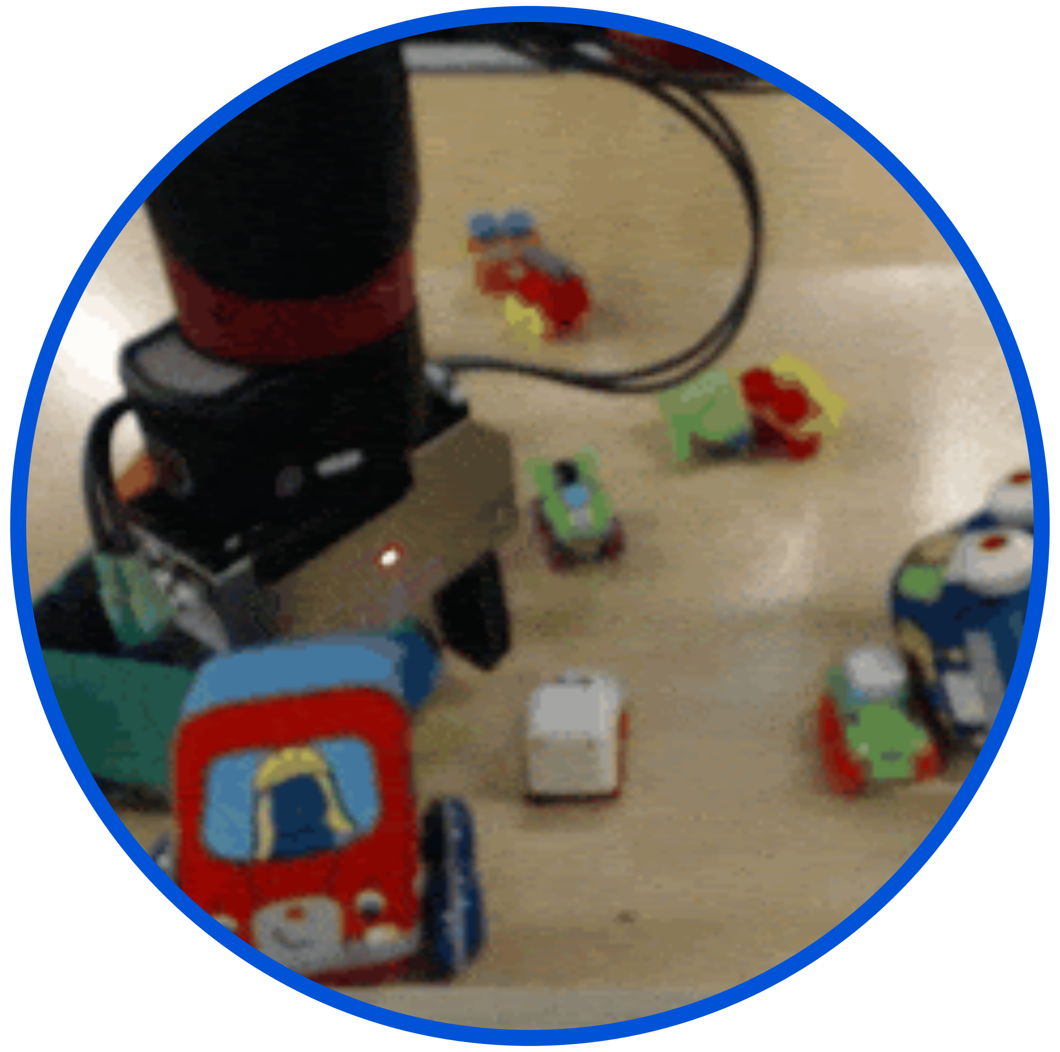
Weber et al. (2017)

Silver et al. (2016)

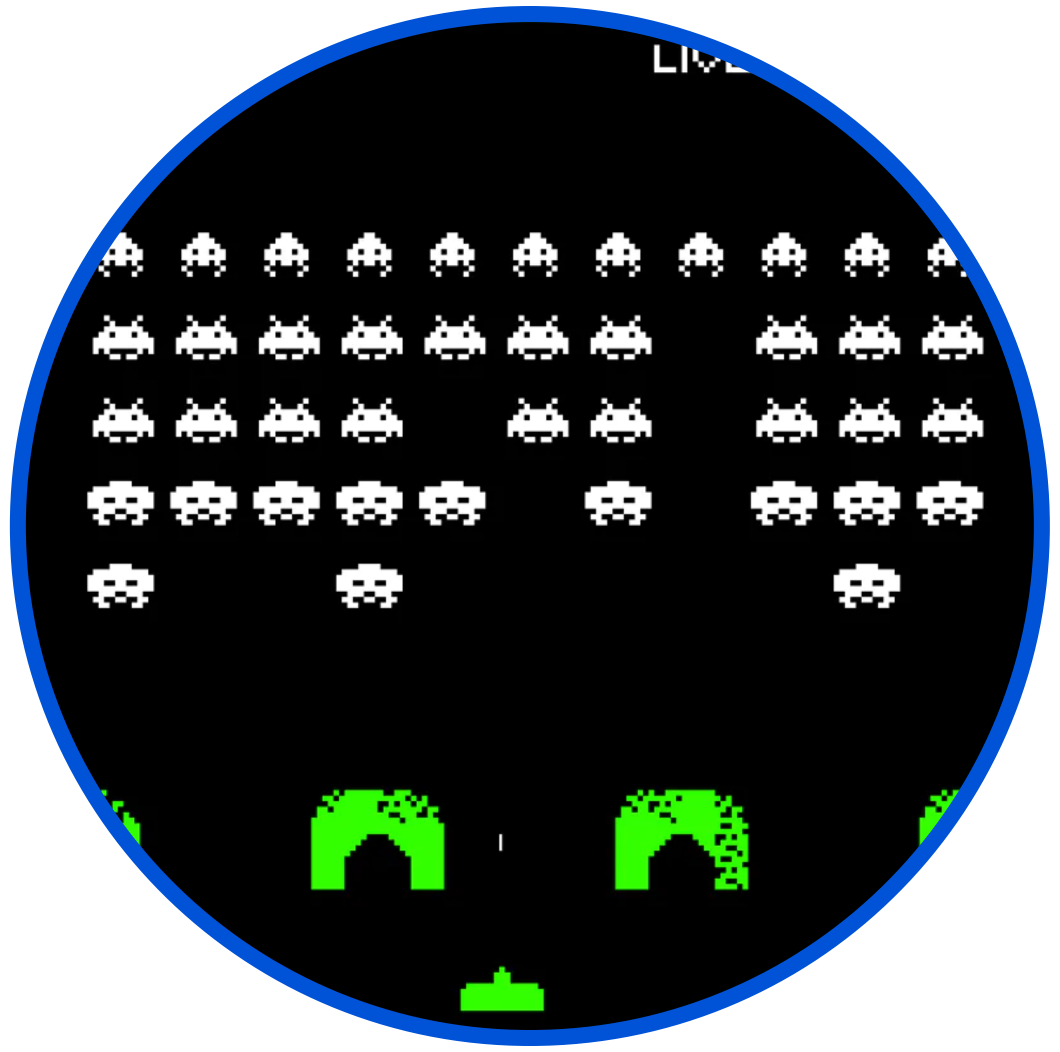
OpenAI et al. (2019)



Segler et al. (2018)

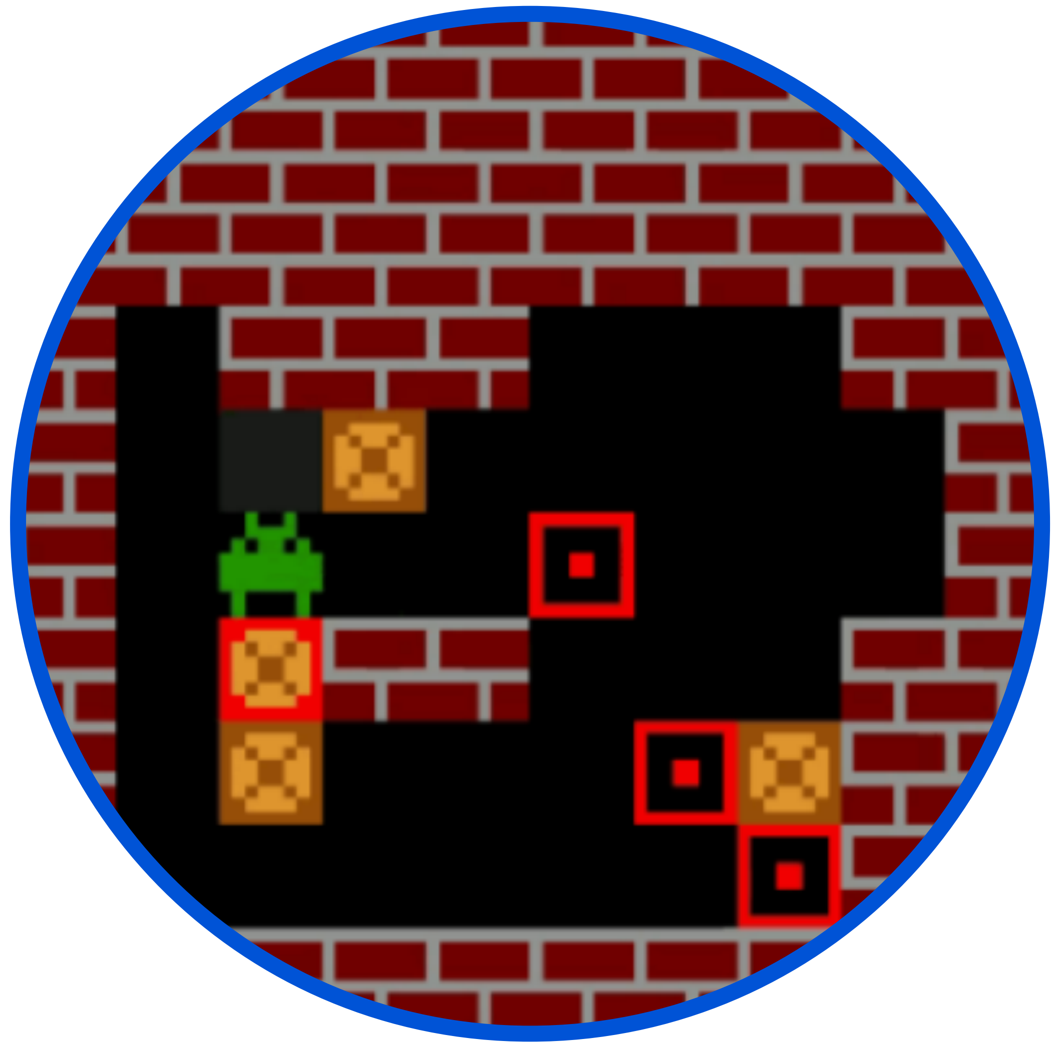


Finn et al. (2018)

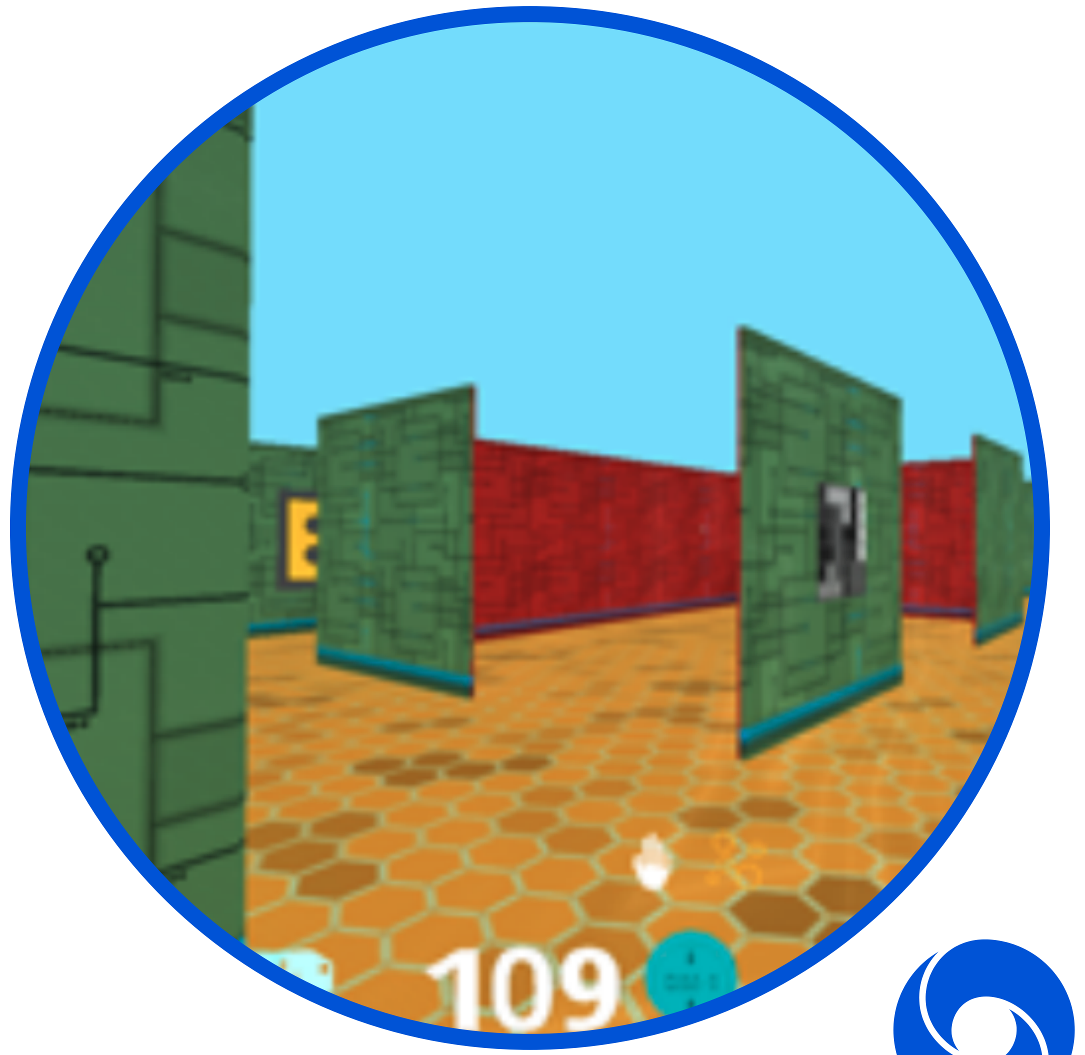


Schrittwieser et al. (2020)

Luo et al. (2019)



Weber et al. (2017)



Hafner et al. (2019)

The promise of model-based RL

"Model-free algorithms are in turn far from the state of the art in domains that require ***precise and sophisticated lookahead***, such as chess and Go"
-Schrittwieser et al. (2019)

"By employing search, we can find strong move sequences potentially ***far away*** from the apprentice policy, accelerating learning in complex scenarios"
-Anthony et al. (2017)

"....predictive models can enable a real robot to manipulate ***previously unseen*** objects and solve new tasks"
-Ebert et al. (2018)

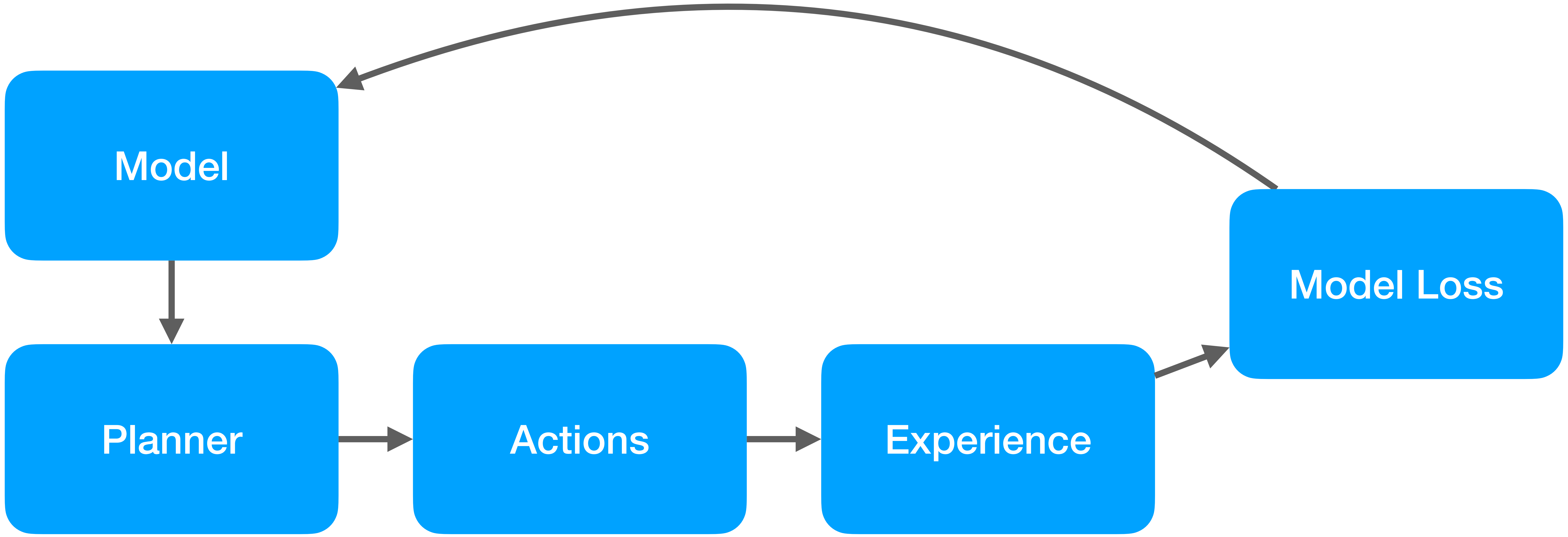
"Model-based planning is an essential ingredient of human intelligence, enabling ***flexible adaptation*** to new tasks and goals"
-Lake et al. (2016)

"...a flexible and general strategy such as mental simulation allows us to reason about a wide range of scenarios, even ***novel*** ones..."
-Hamrick (2017)

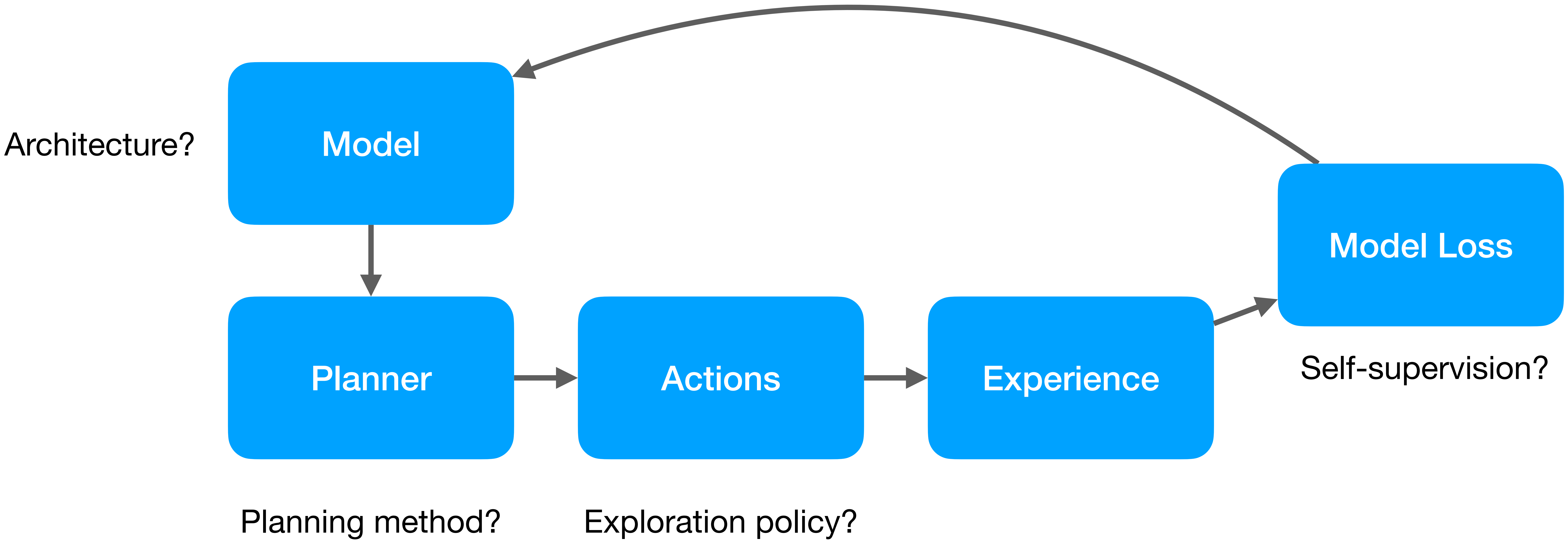
"...[models] enable better ***generalization*** across states, remain valid across tasks in the same environment, and exploit additional unsupervised learning signals..."
-Weber et al. (2017)

The best MBRL systems are *complicated*

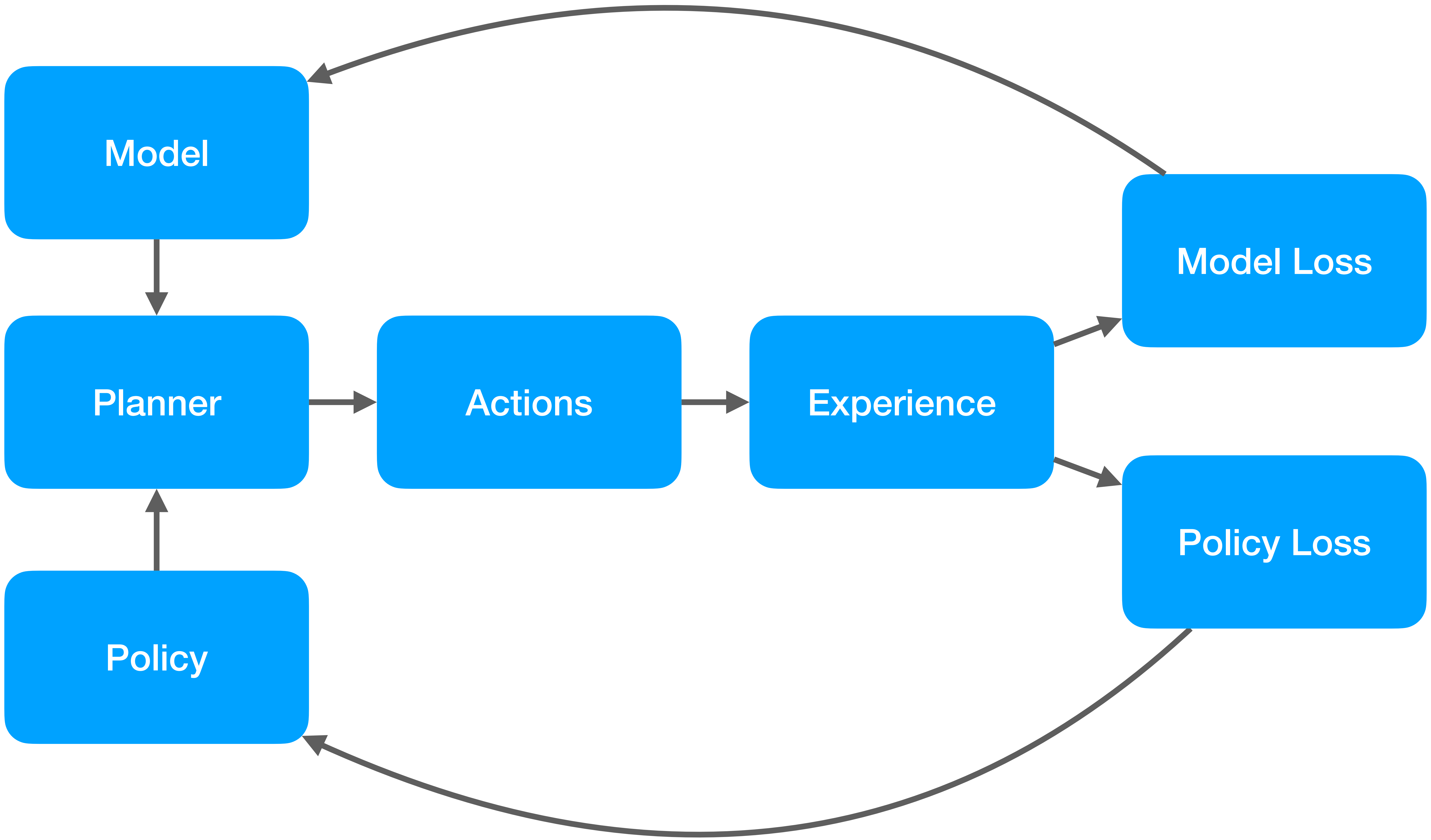
Pure planning



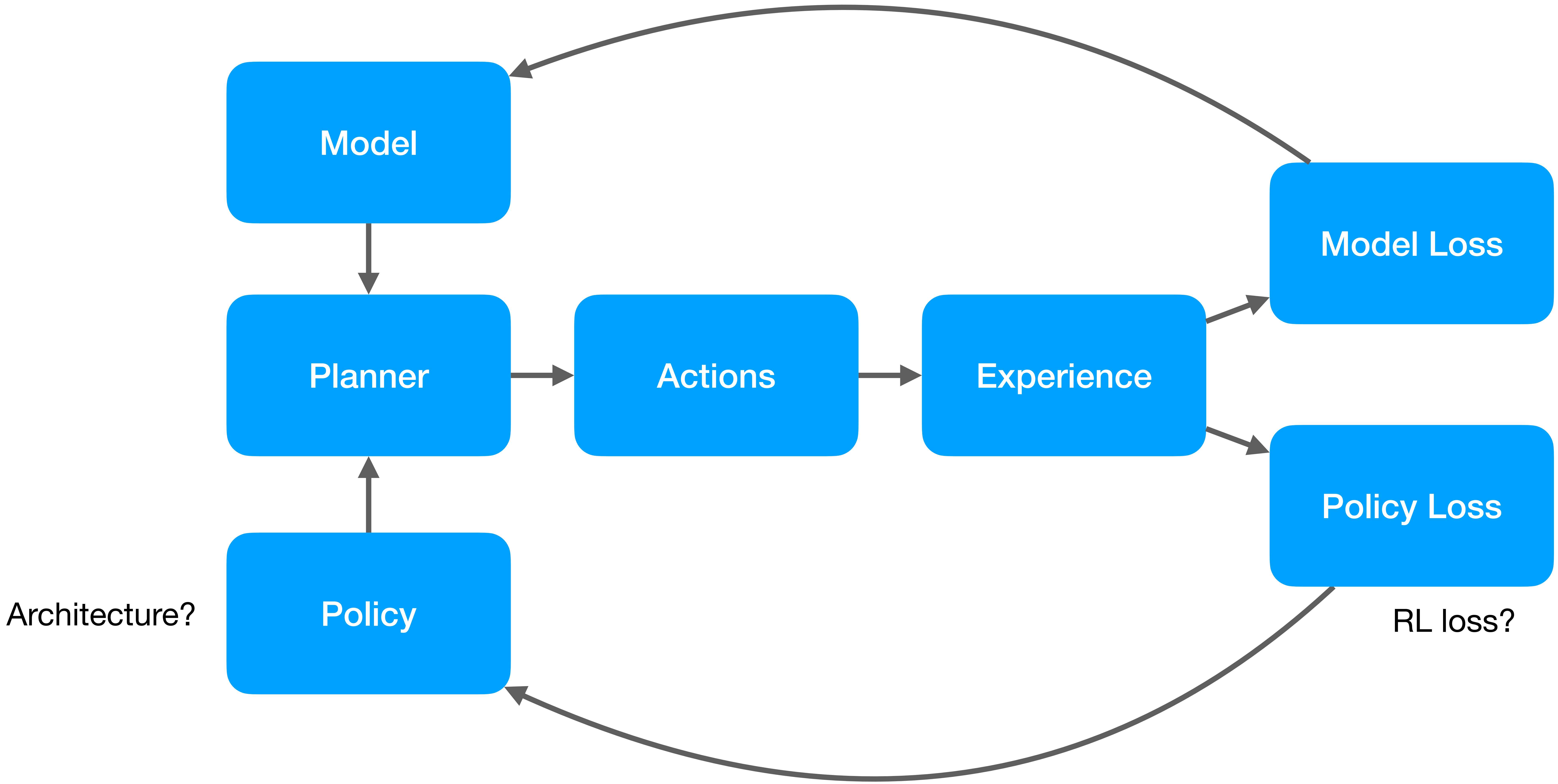
Pure planning



Guided planning

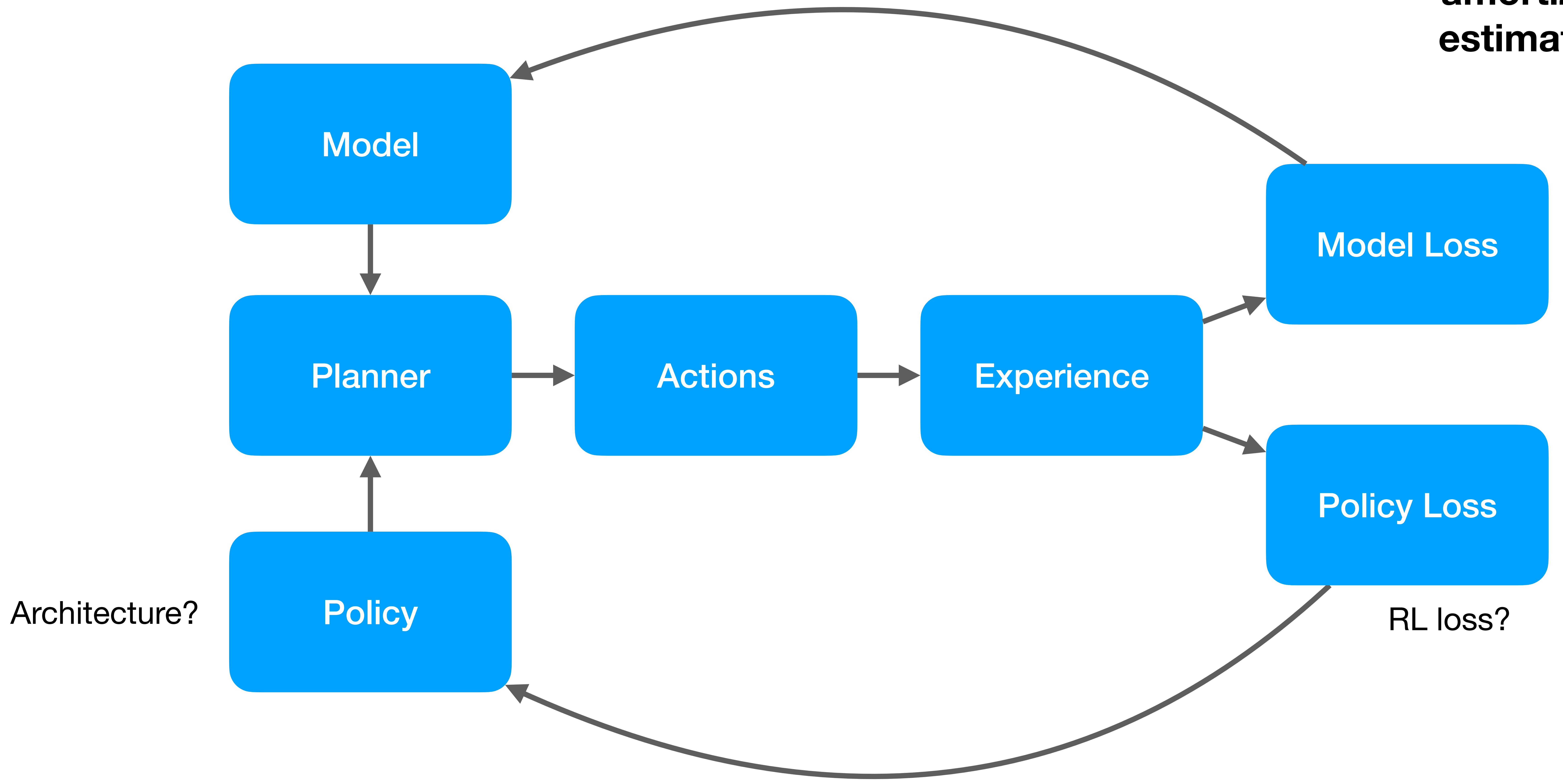


Guided planning



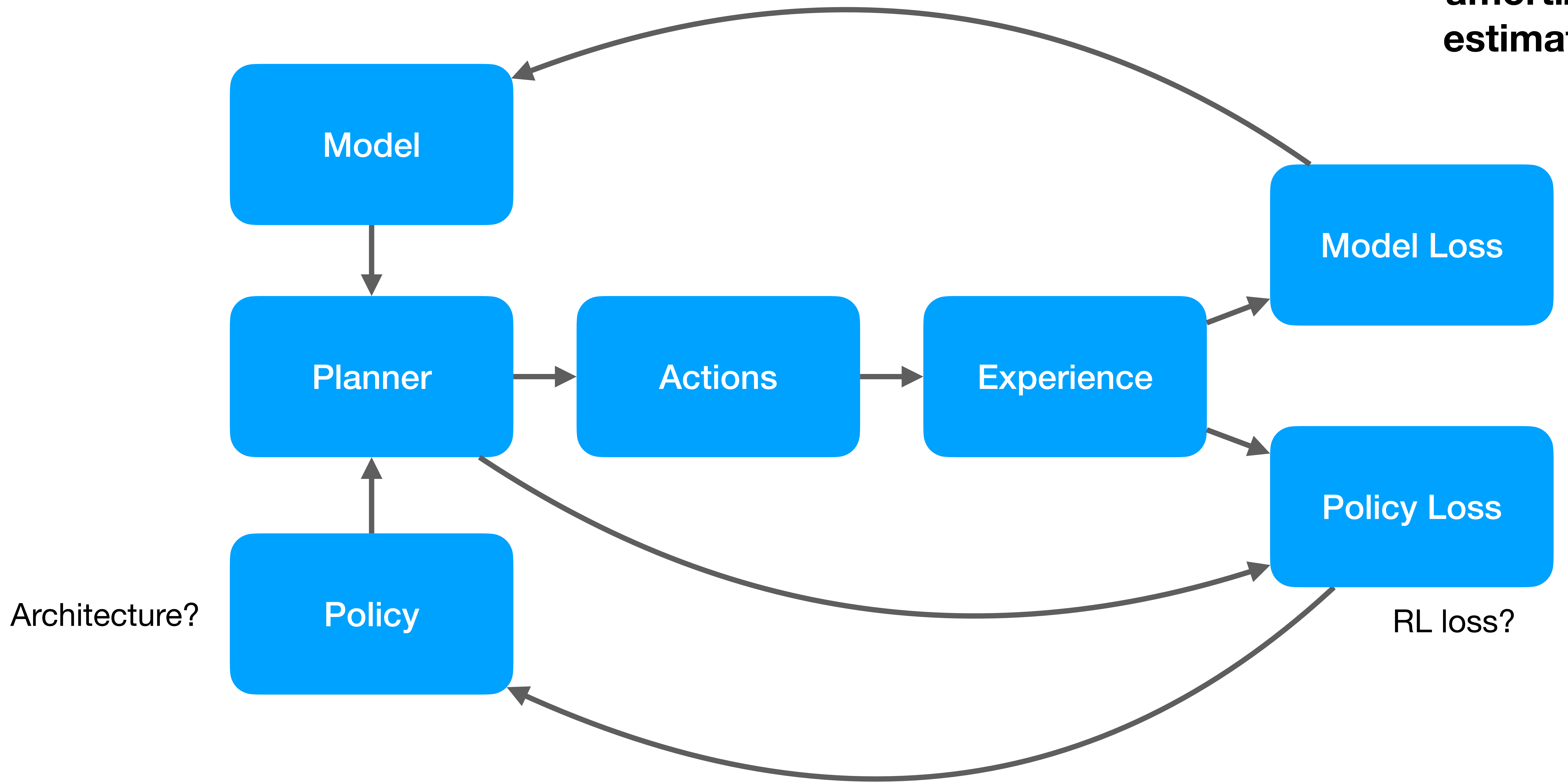
Guided planning

Hamrick et al. (2020). **Combining Q-learning and search with amortized value estimates.** ICLR.

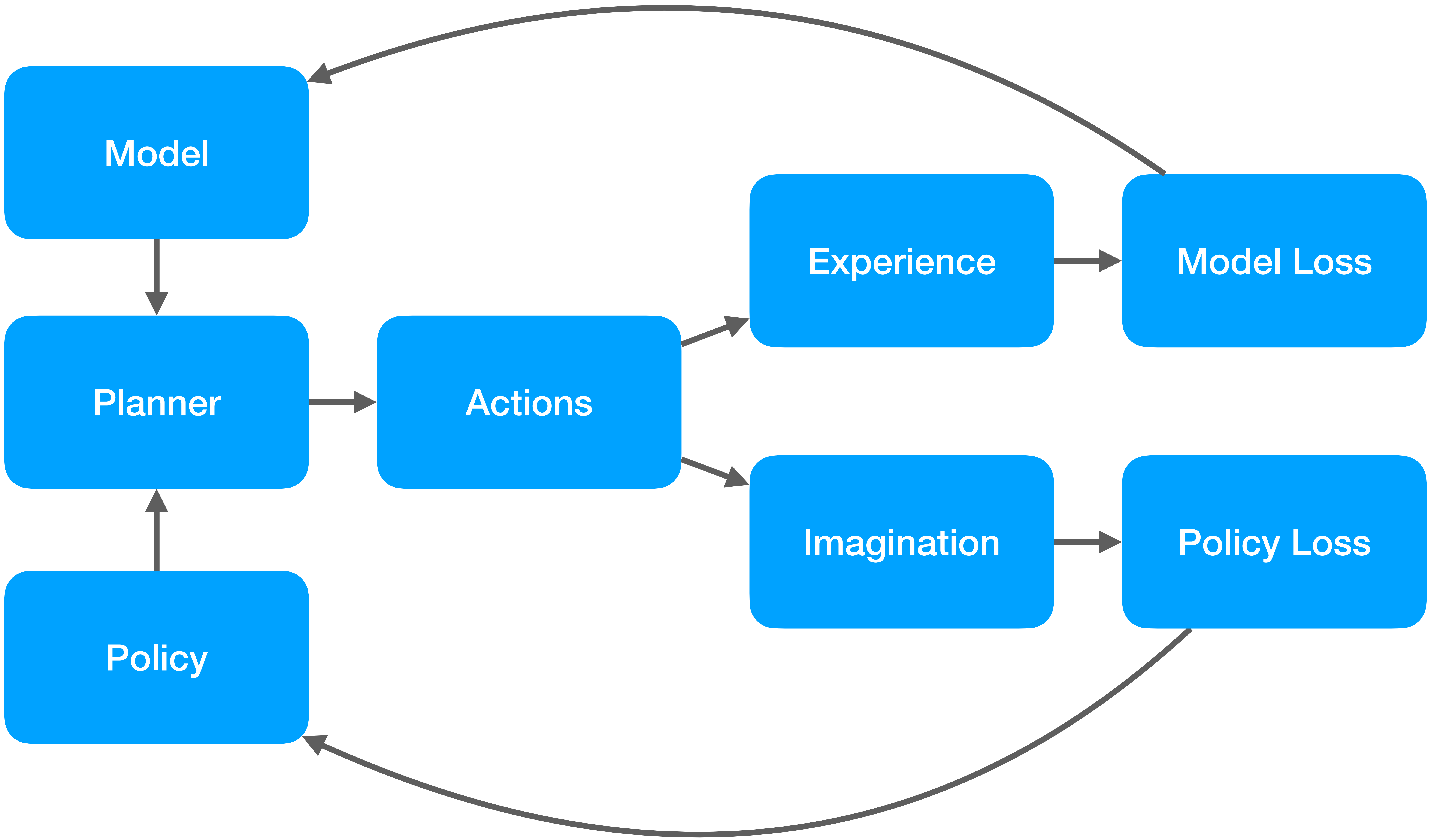


Expert iteration

Hamrick et al. (2020). **Combining Q-learning and search with amortized value estimates**. ICLR.

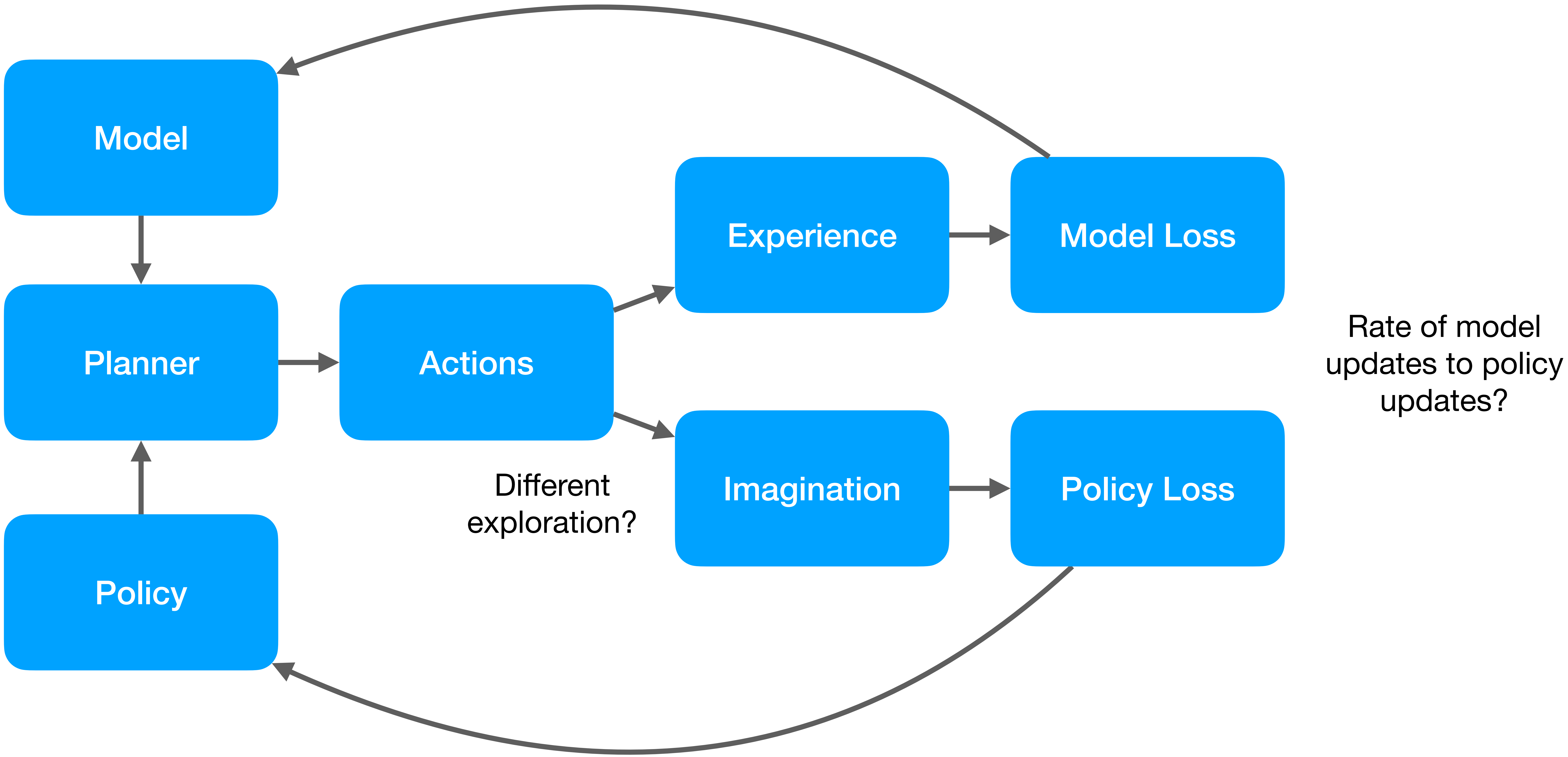


Dyna



Jessica Hamrick - jhamrick@deepmind.com

Dyna



Model
Planner
Policy
Actions
Experience
Imagination
Model Loss
Policy Loss
Different exploration?
Rate of model updates to policy updates?

Outline

- **Understanding MBRL**
  Hamrick et al. (2021). On the role of planning in model based reinforcement learning. ICLR.

- **Understanding and improving generalization**
  Anand, Walker et al. (2022). Procedural generalization by planning with self-supervised world models. ICLR.

- **Understanding and improving transfer**
  Walker, Vértes, Li et al. (2023). Investigating the role of model-based learning in exploration and transfer. Under review.

- **The future of MBRL**

Outline

- **Understanding MBRL**
  Hamrick et al. (2021). On the role of planning in model based reinforcement learning. ICLR.

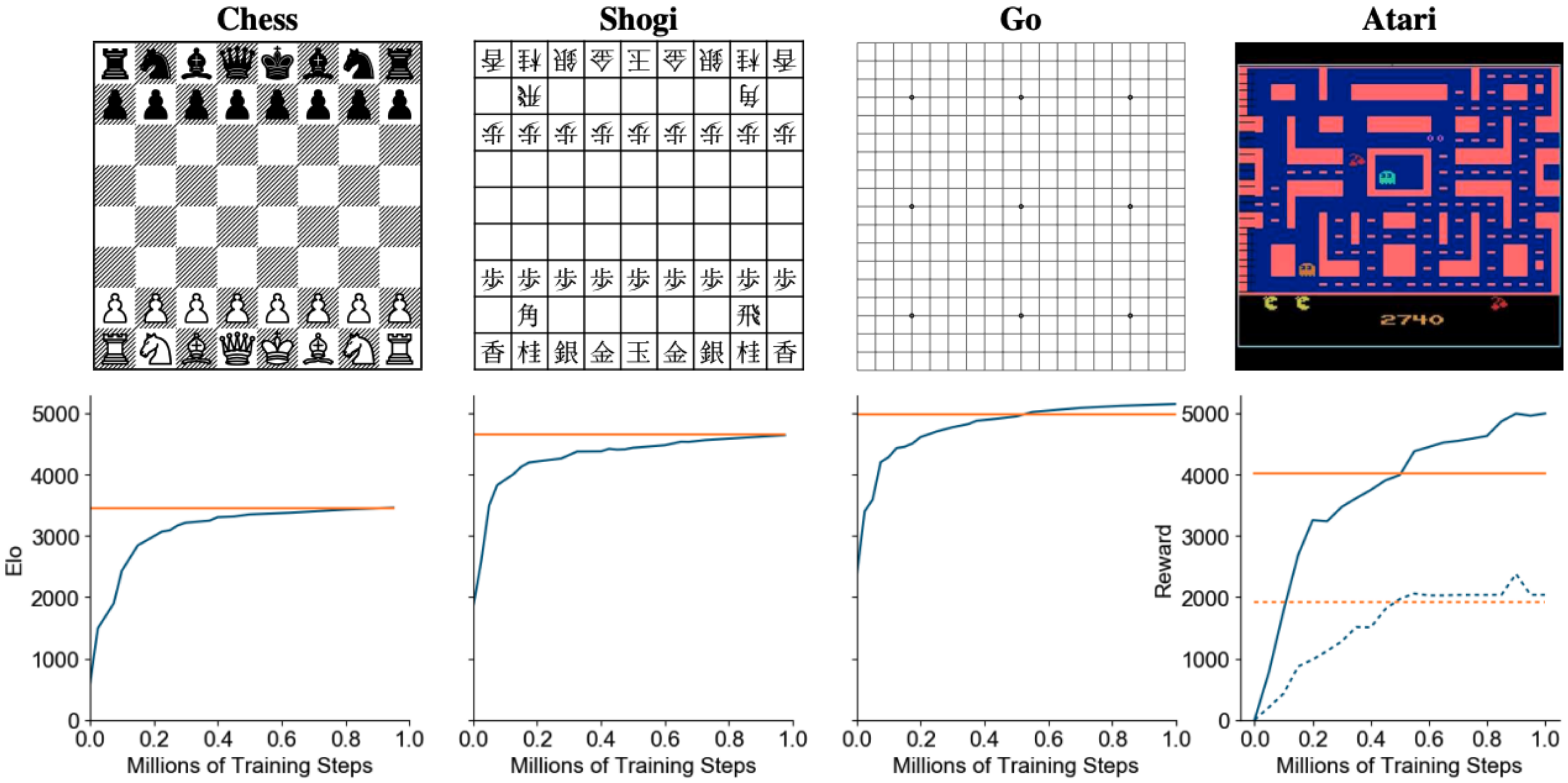
- **Understanding and improving generalization**
  Anand, Walker et al. (2022). Procedural generalization by planning with self-supervised world models. ICLR.

- **Understanding and improving transfer**
  Walker, Vértes, Li et al. (2023). Investigating the role of model-based learning in exploration and transfer. Under review.

- **The future of MBRL**

MuZero

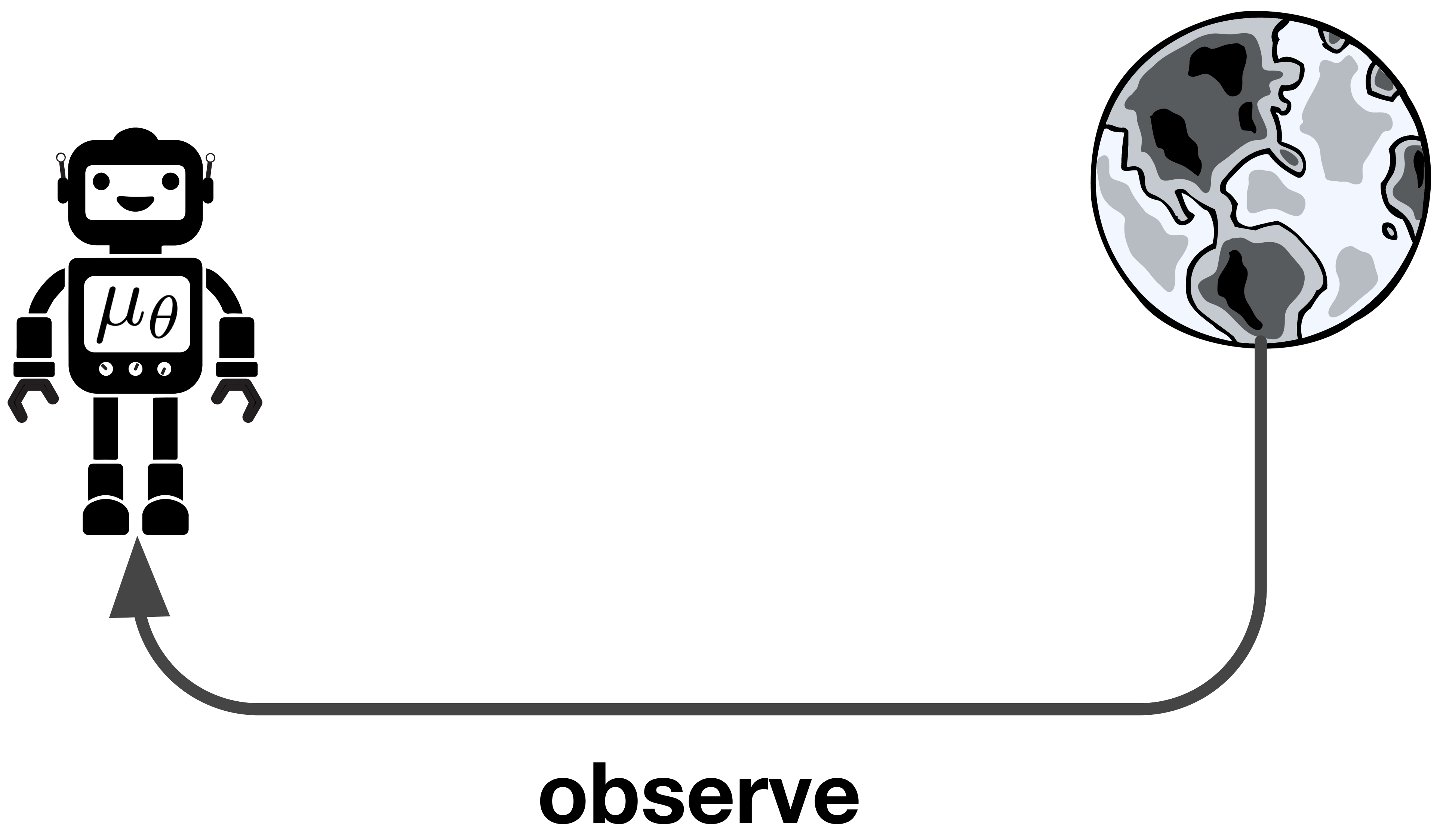
Schrittwieser et al. (2019)



Jessica Hamrick - jhamrick@deepmind.com

MuZero

Schrittwieser et al. (2019)

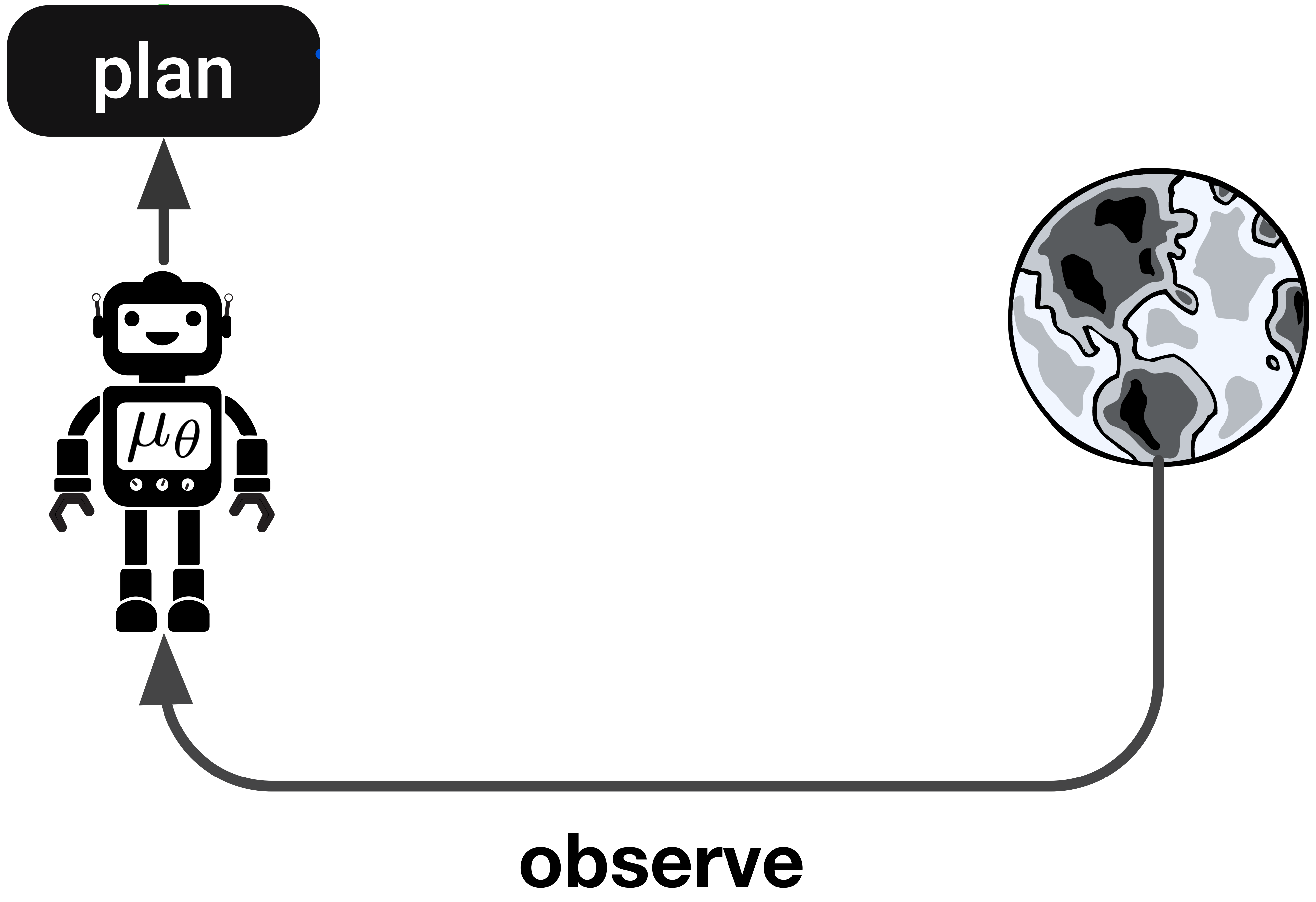


MuZero

Schrittwieser et al. (2019)

*Guide MCTS using learned **policy and value functions***

policy: where to search?
model: what will happen?
value: is what will happen good?



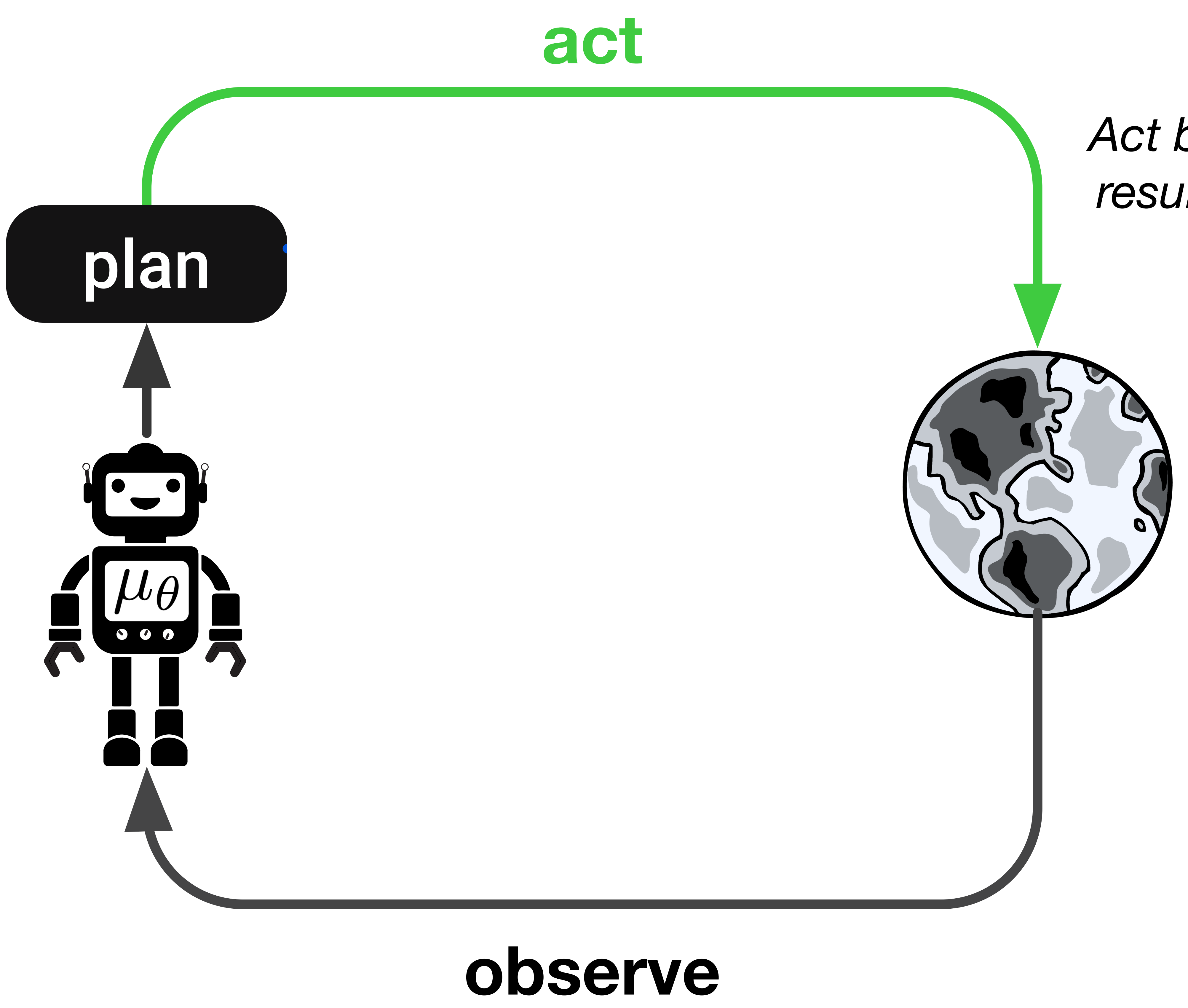
(MCTS = Monte Carlo Tree Search)

MuZero

Schrittwieser et al. (2019)

Guide MCTS using learned ***policy and value functions***

policy: where to search?
model: what will happen?
value: is what will happen good?



Act based on the results of search

(MCTS = Monte Carlo Tree Search)

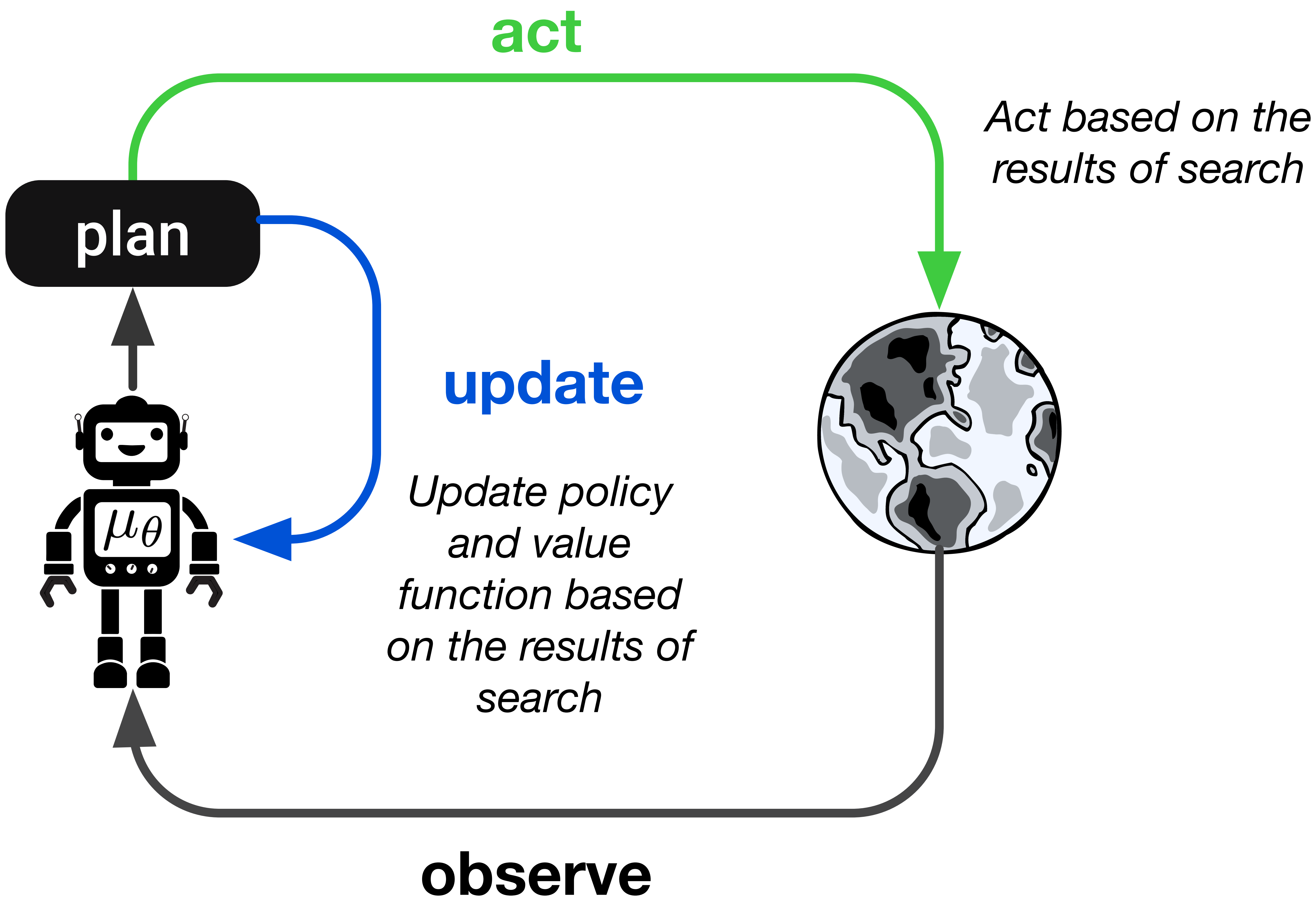
Jessica Hamrick - jhamrick@deepmind.com

MuZero

Schrittwieser et al. (2019)

*Guide MCTS using learned **policy and value functions***

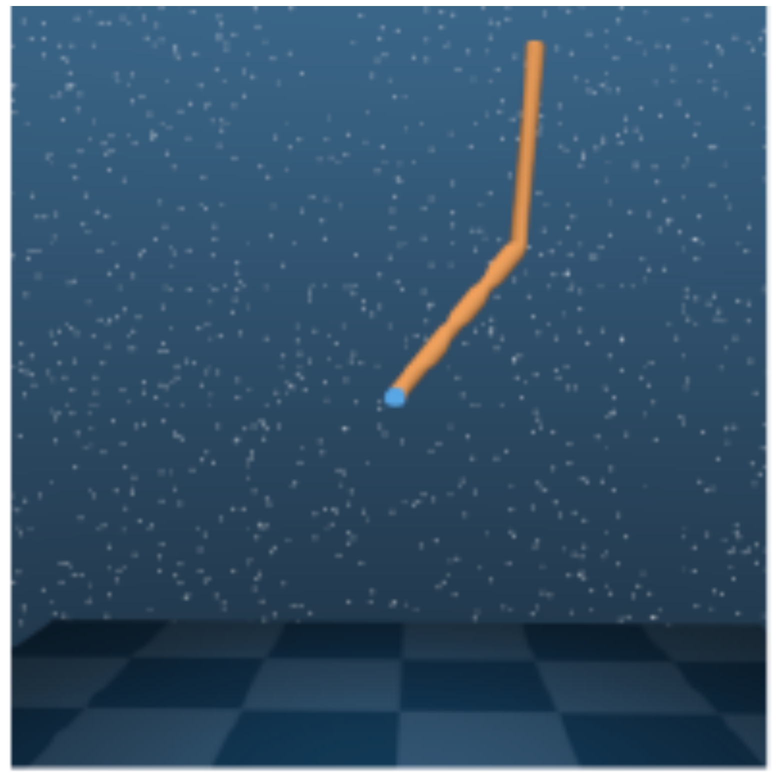
policy: where to search?
model: what will happen?
value: is what will happen good?



(MCTS = Monte Carlo Tree Search)

Jessica Hamrick - jhamrick@deepmind.com

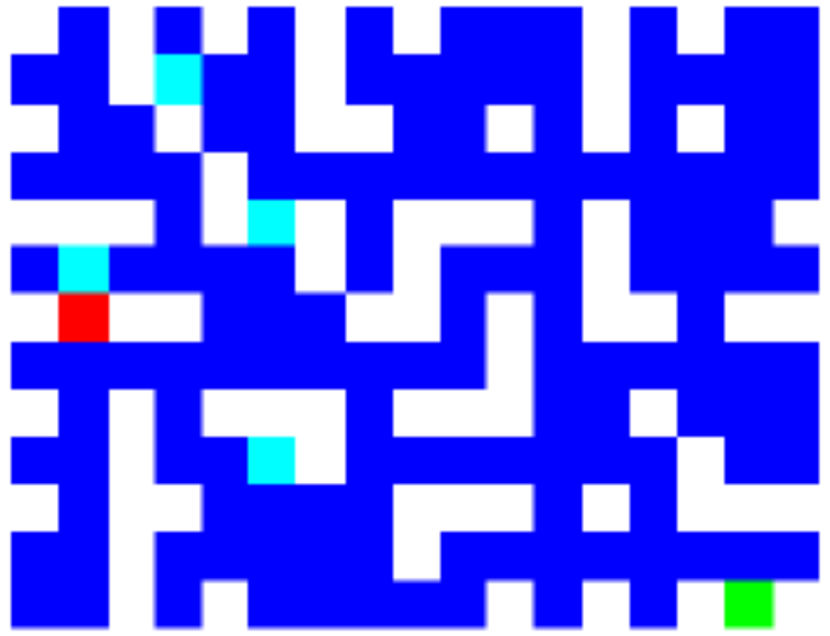
Environments



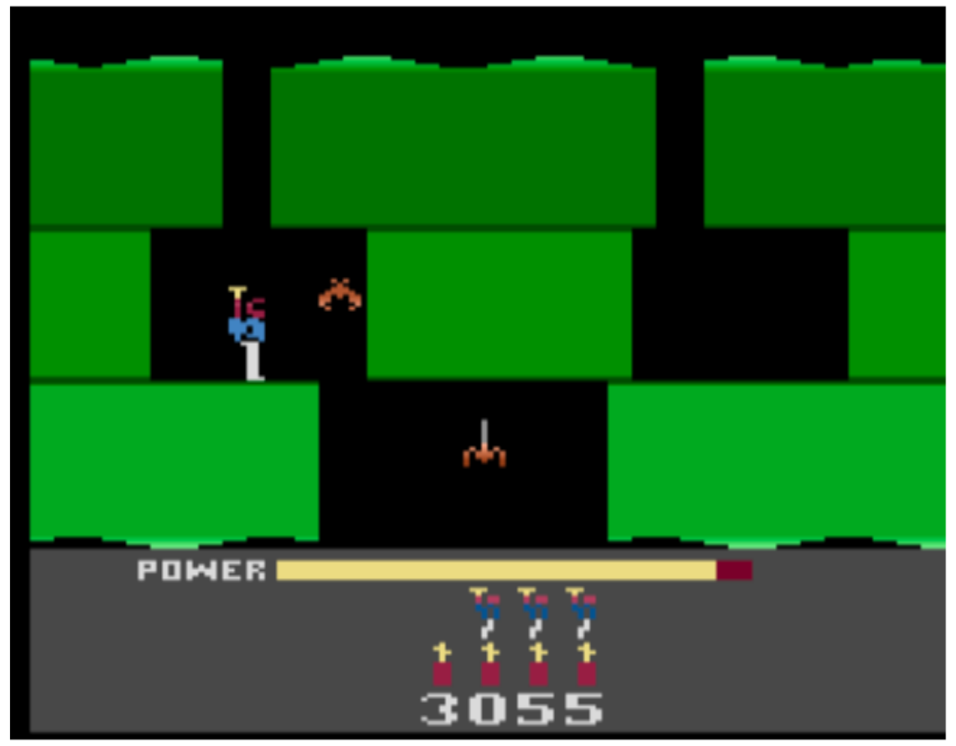
Acrobot
(Swingup Sparse)

Cheetah
(Run)

Humanoid
(Stand)

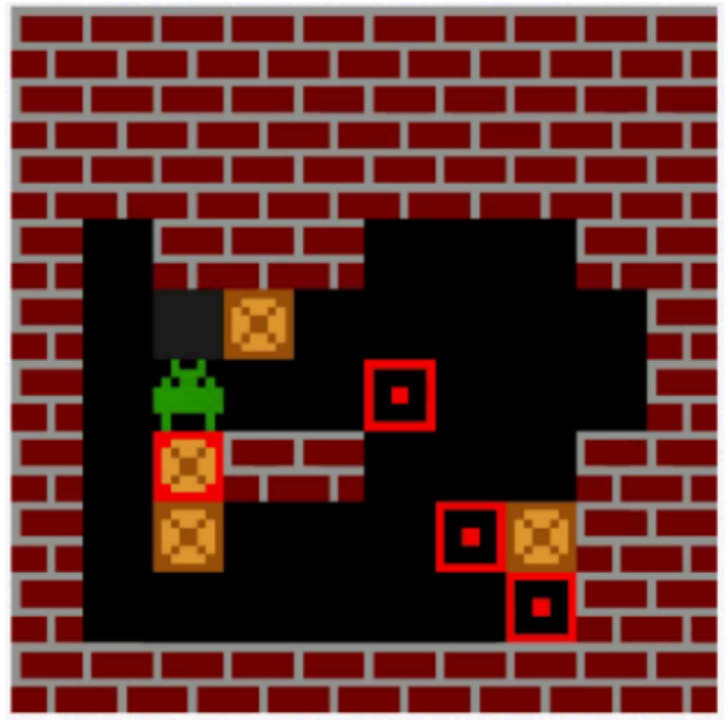


Minipacman
(Procedural)

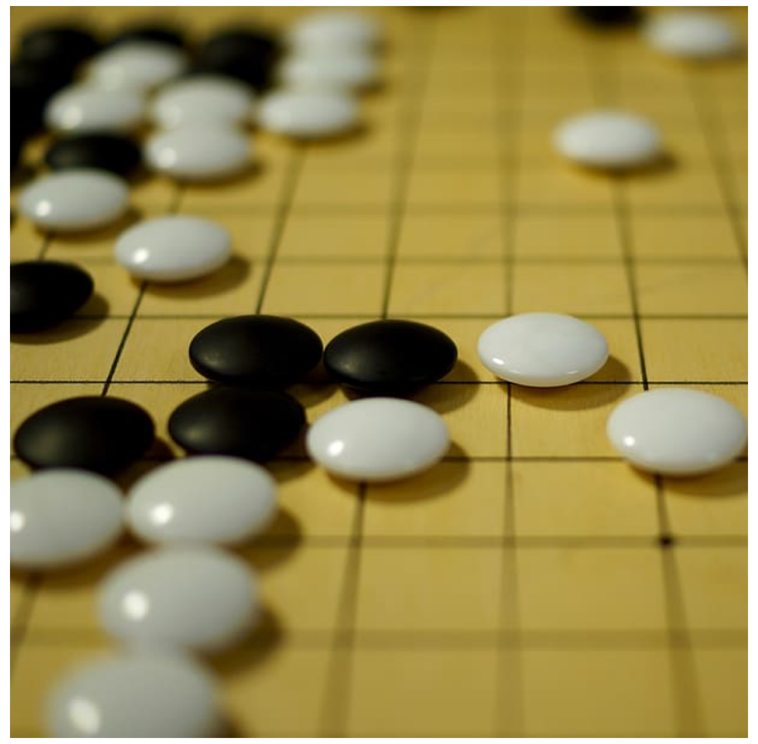


Hero

Ms. Pacman



Sokoban



9x9 Go

Jessica Hamrick - jhamrick@deepmind.com

Q1: How does planning benefit model-based RL agents?

Q1: How does planning benefit model-based RL agents?

Q2: Within planning, what algorithmic choices drive performance?

Q1: How does planning benefit model-based RL agents?

Q2: Within planning, what algorithmic choices drive performance?

Q3: To what extent does planning improve zero-shot generalization?

Q1: How does planning benefit model-based RL agents?

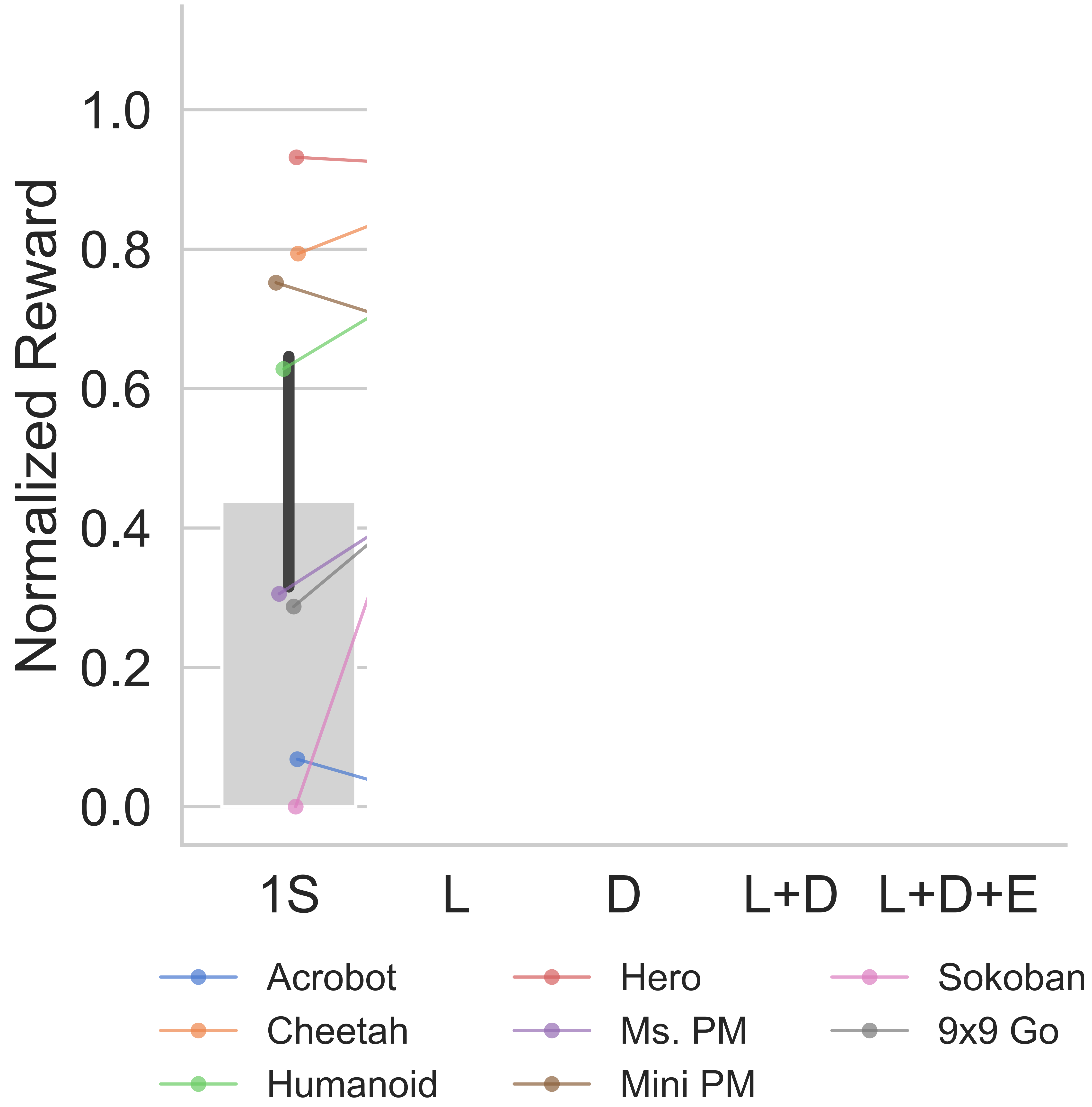
Q2: Within planning, what algorithmic choices drive performance?

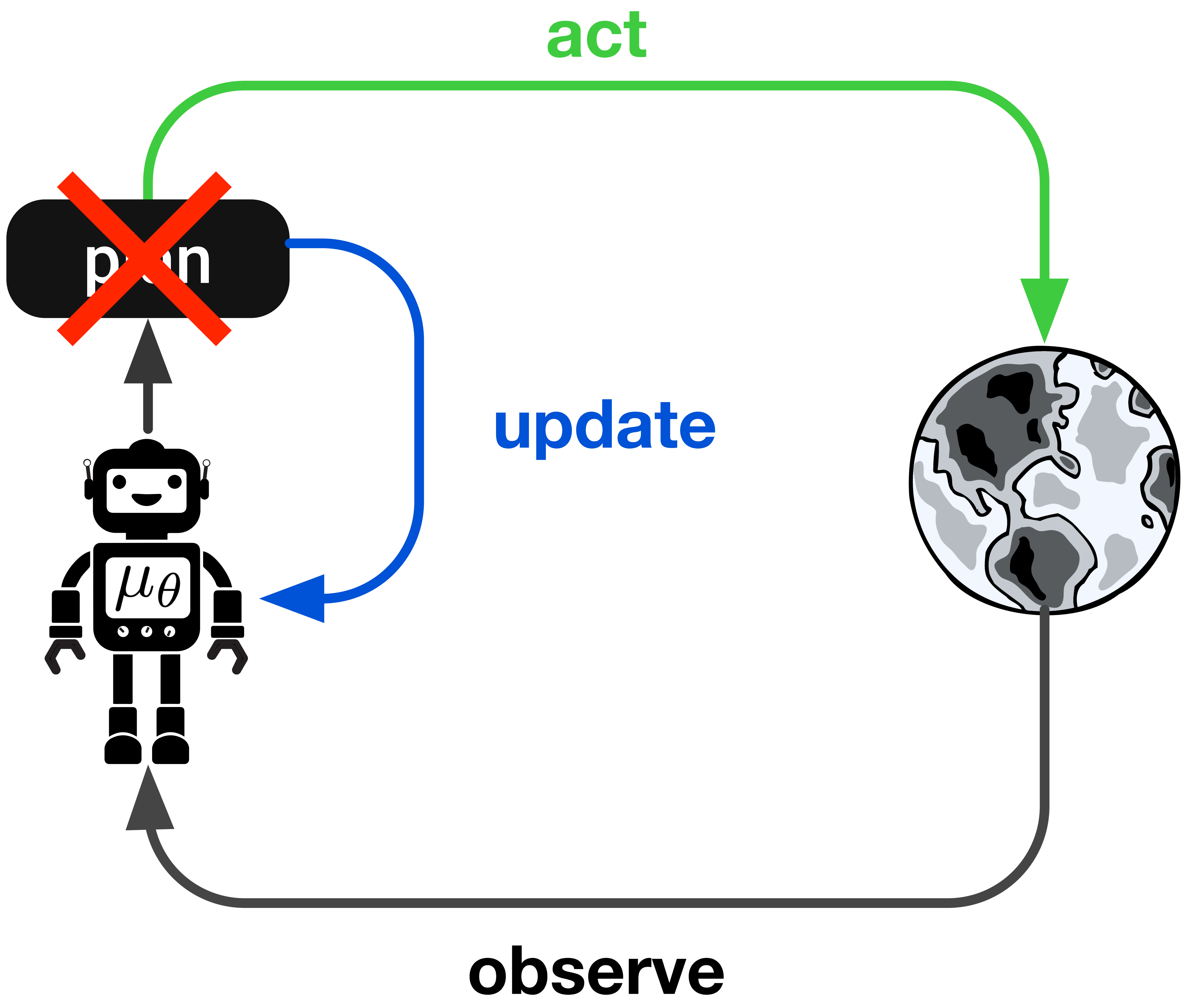
Q3: To what extent does planning improve zero-shot generalization?

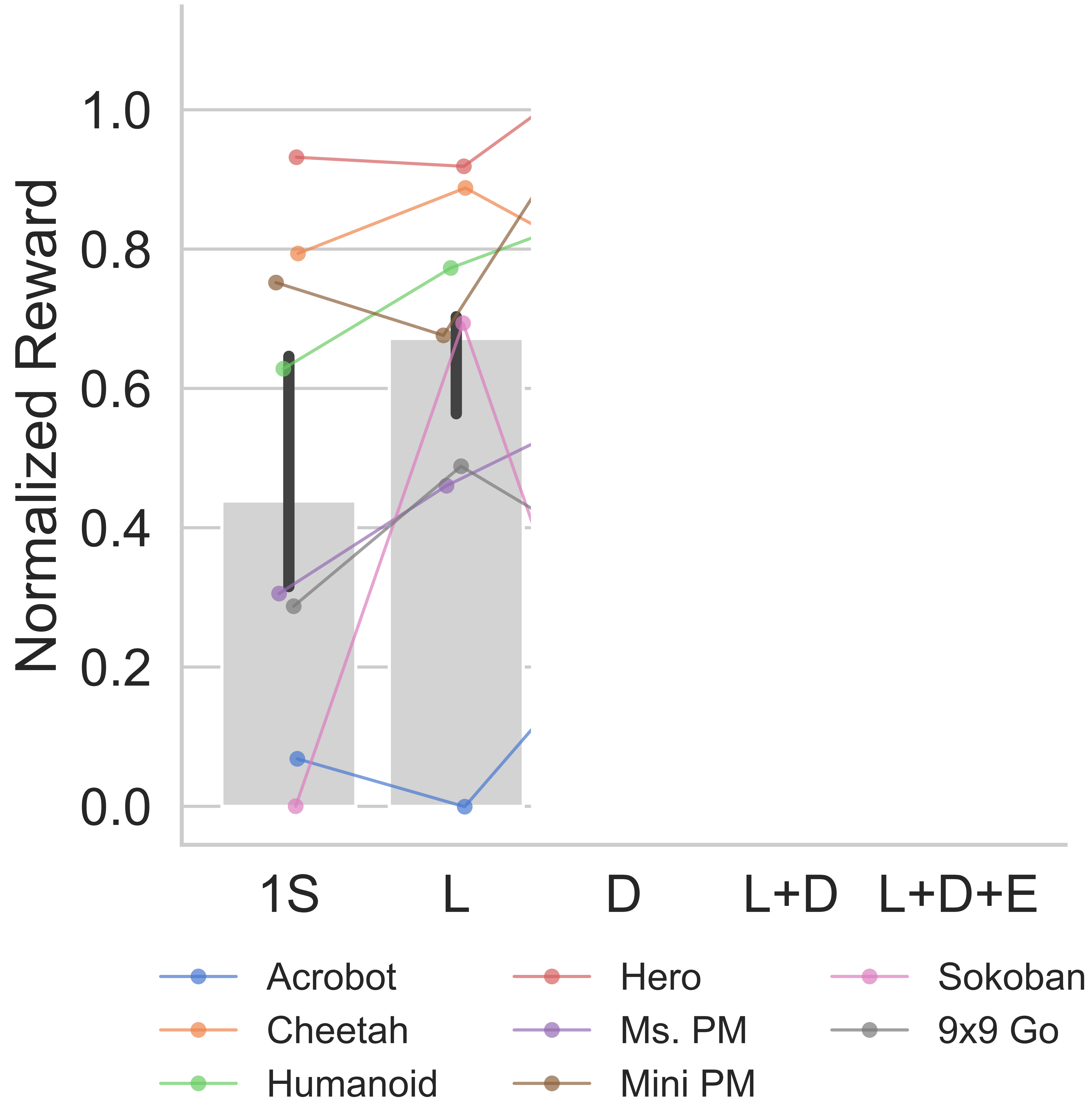
Using search in different ways

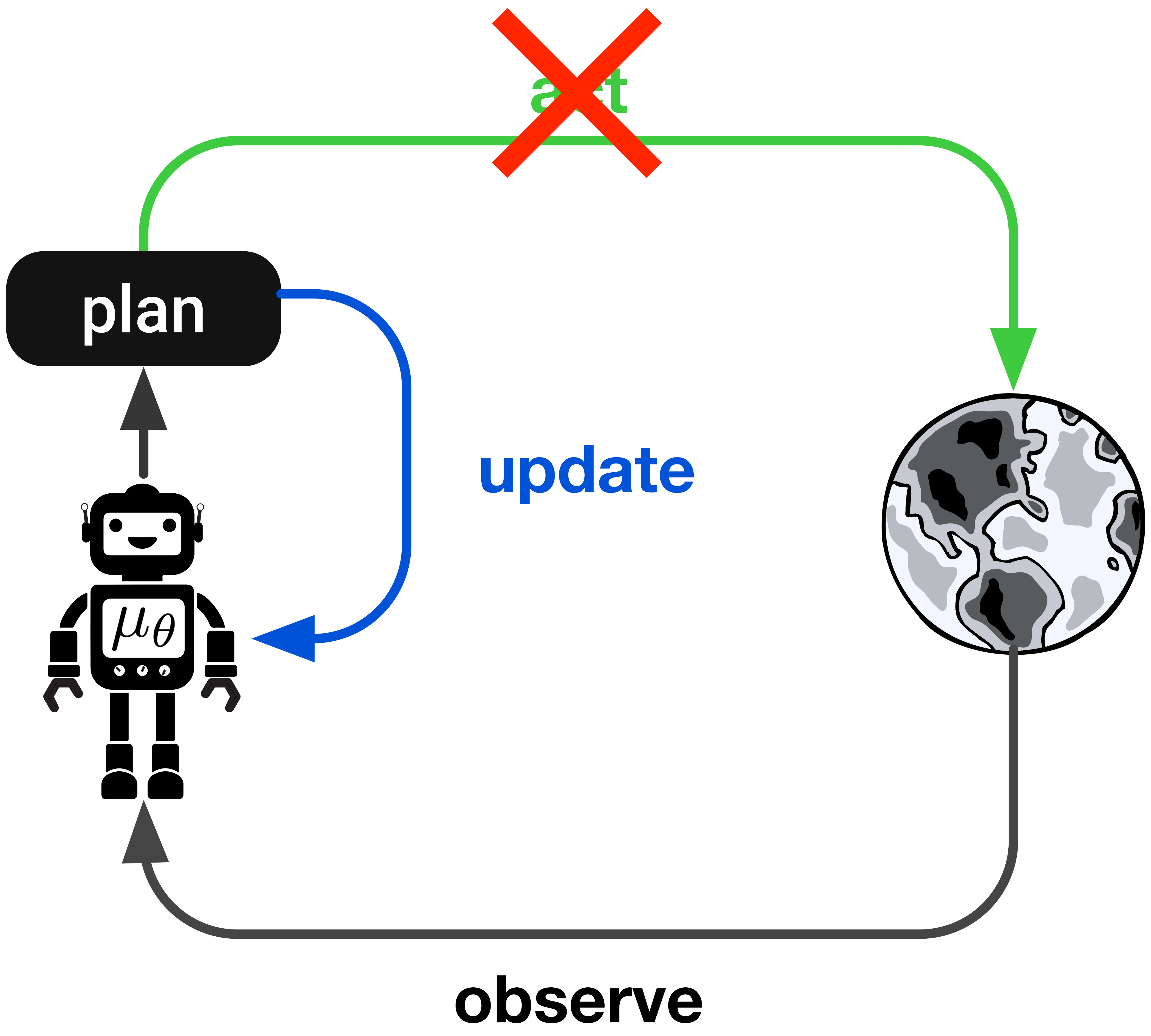
	Train Update	Train Act	Test Act
One-Step	1-step search	prior	prior
Learn	Full search	prior	prior
Data	1-step search	Full search	prior
Learn+Data	Full search	Full search	prior
Learn+Data+Eval (vanilla MuZero)	Full search	Full search	Full search

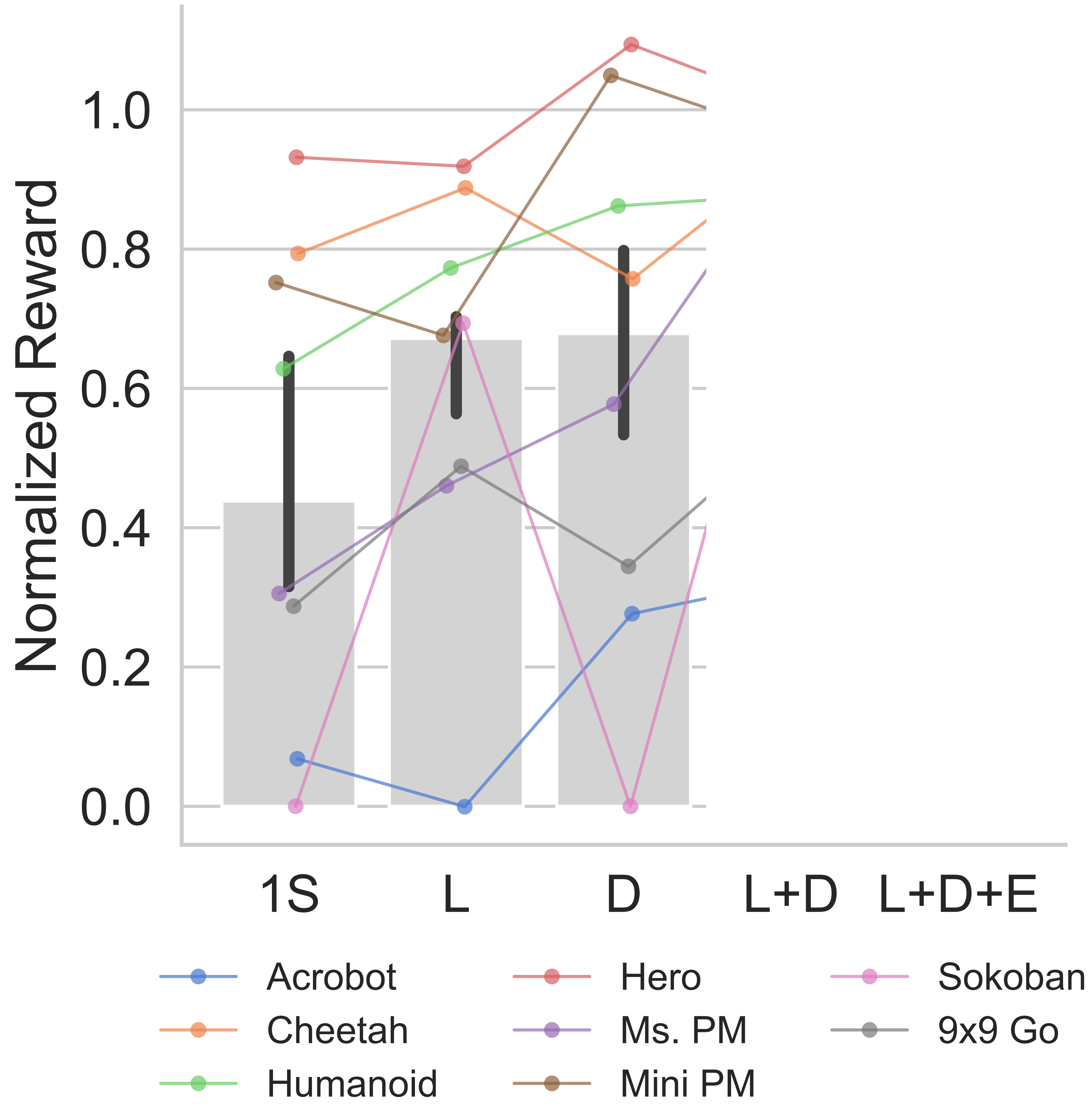
Hamrick et al. (2021). On the role of planning in model-based deep reinforcement learning. ICLR.

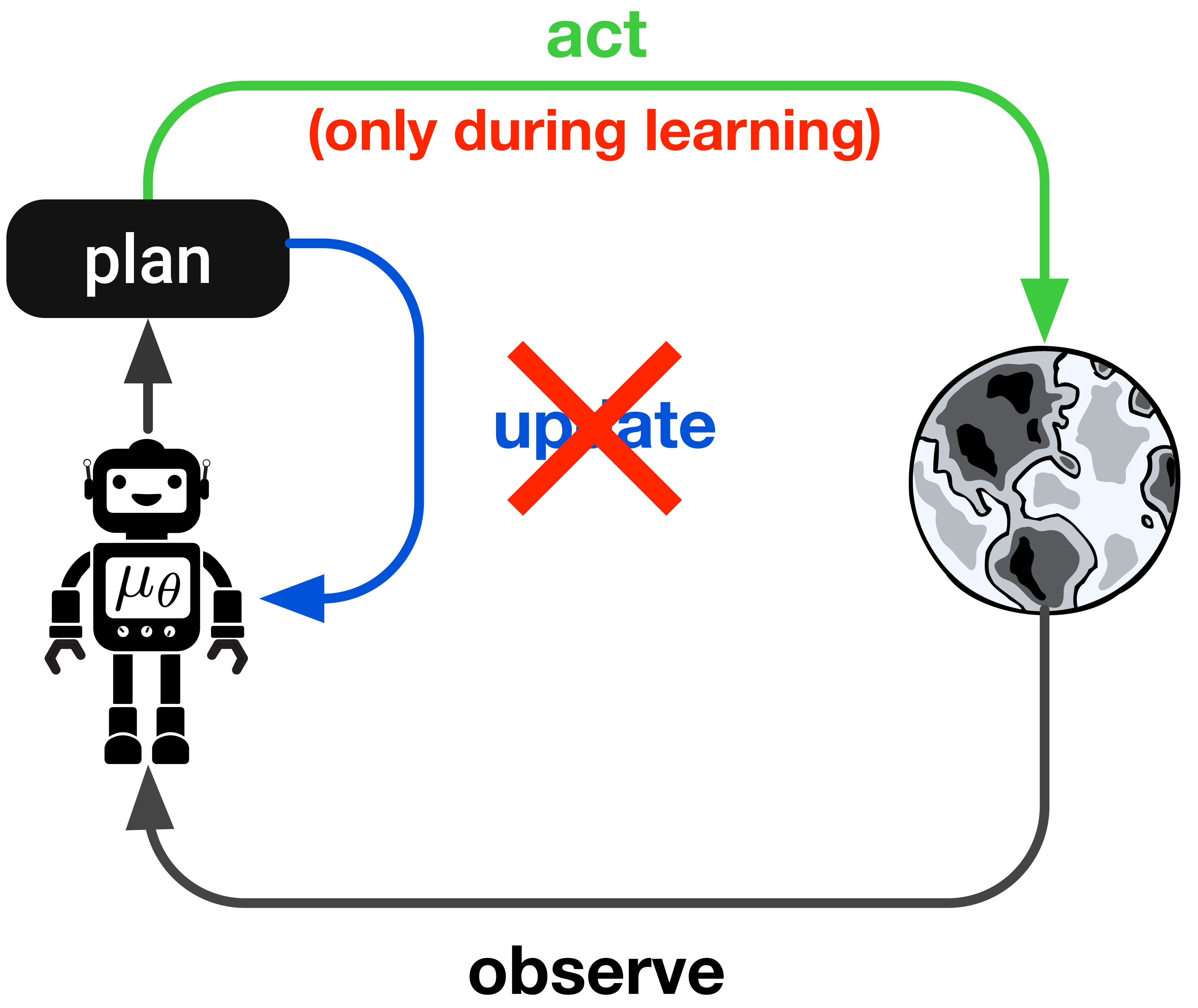




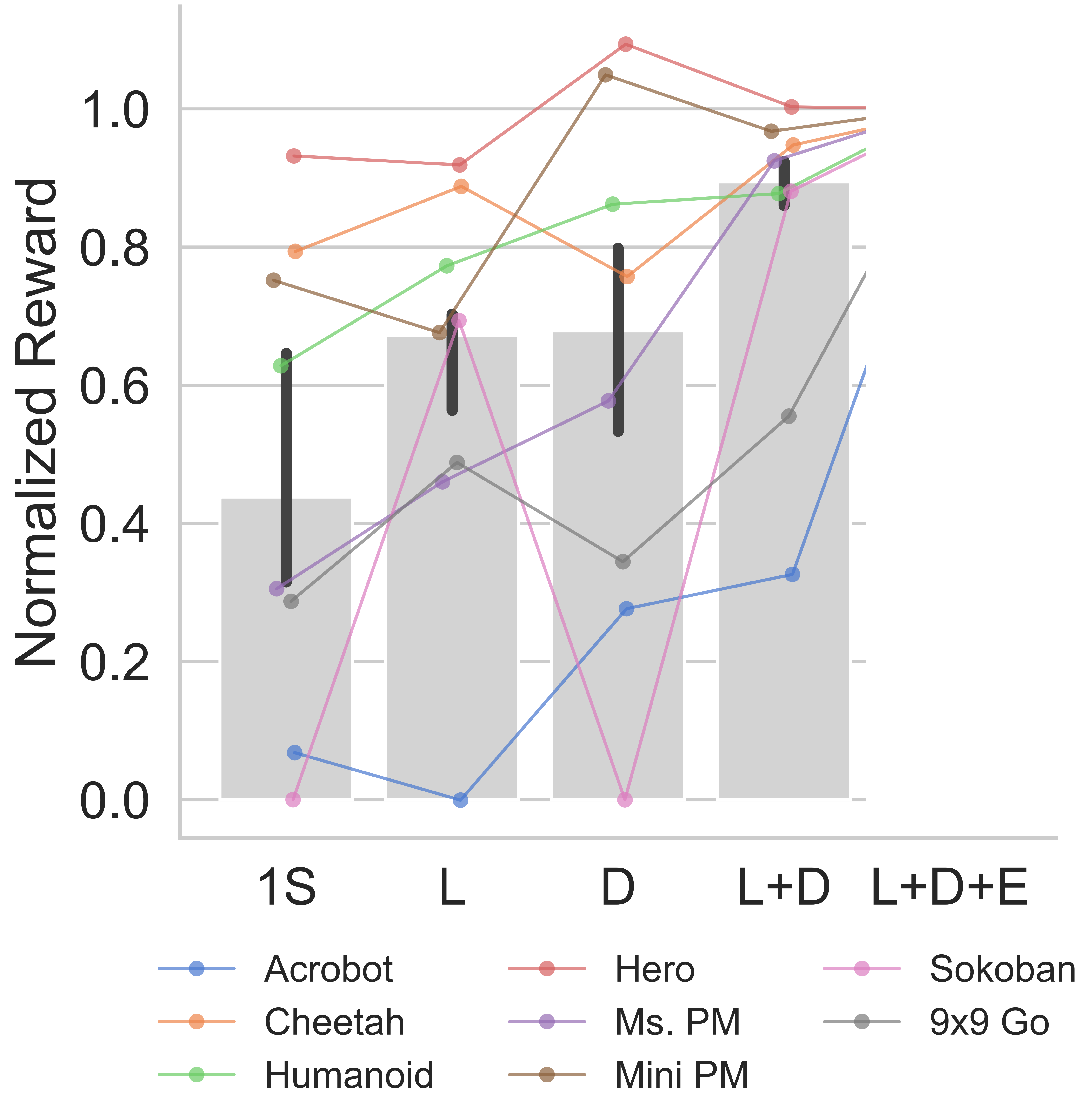

Normalized Reward
1.0
0.8
0.6
0.4
0.2
0.0
1S
L
D
L+D
L+D+E
Acrobot
Cheetah
Humanoid
Hero
Ms. PM
Mini PM
Sokoban
9x9 Go

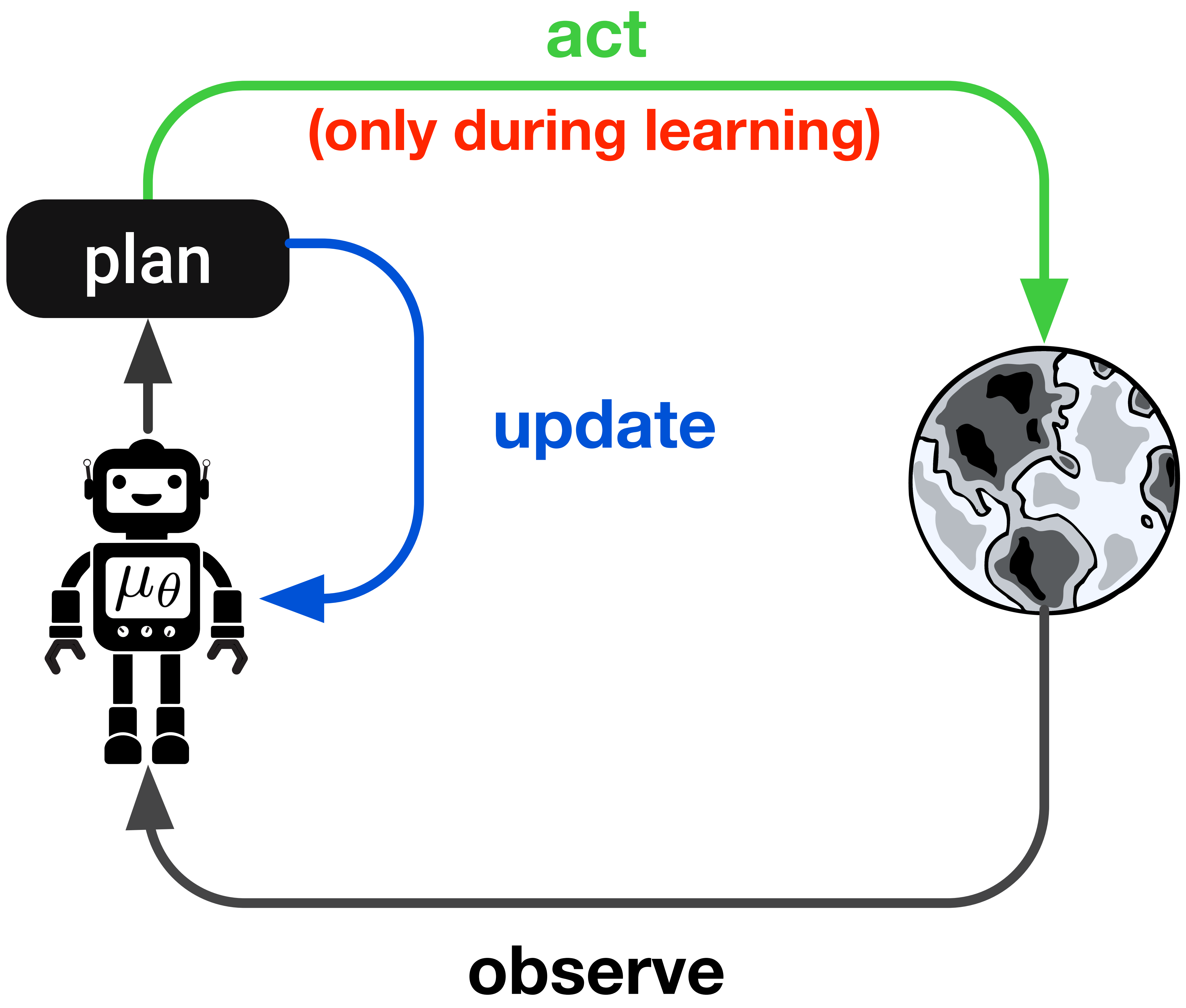

act
plan
update
μ_θ
observe

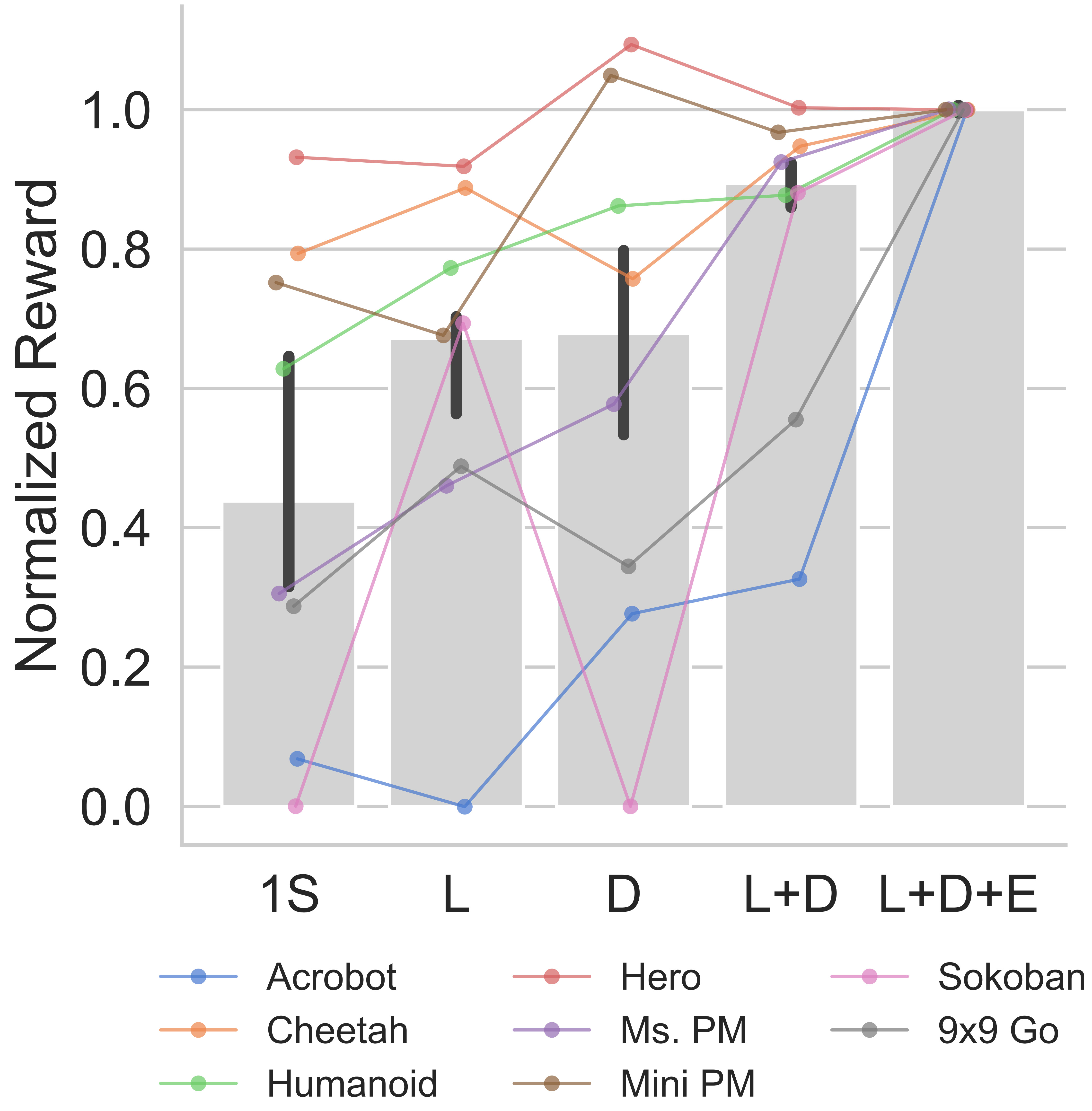

Normalized Reward
1.0
0.8
0.6
0.4
0.2
0.0
1S
L
D
L+D
L+D+E
Acrobot
Cheetah
Humanoid
Hero
Ms. PM
Mini PM
Sokoban
9x9 Go

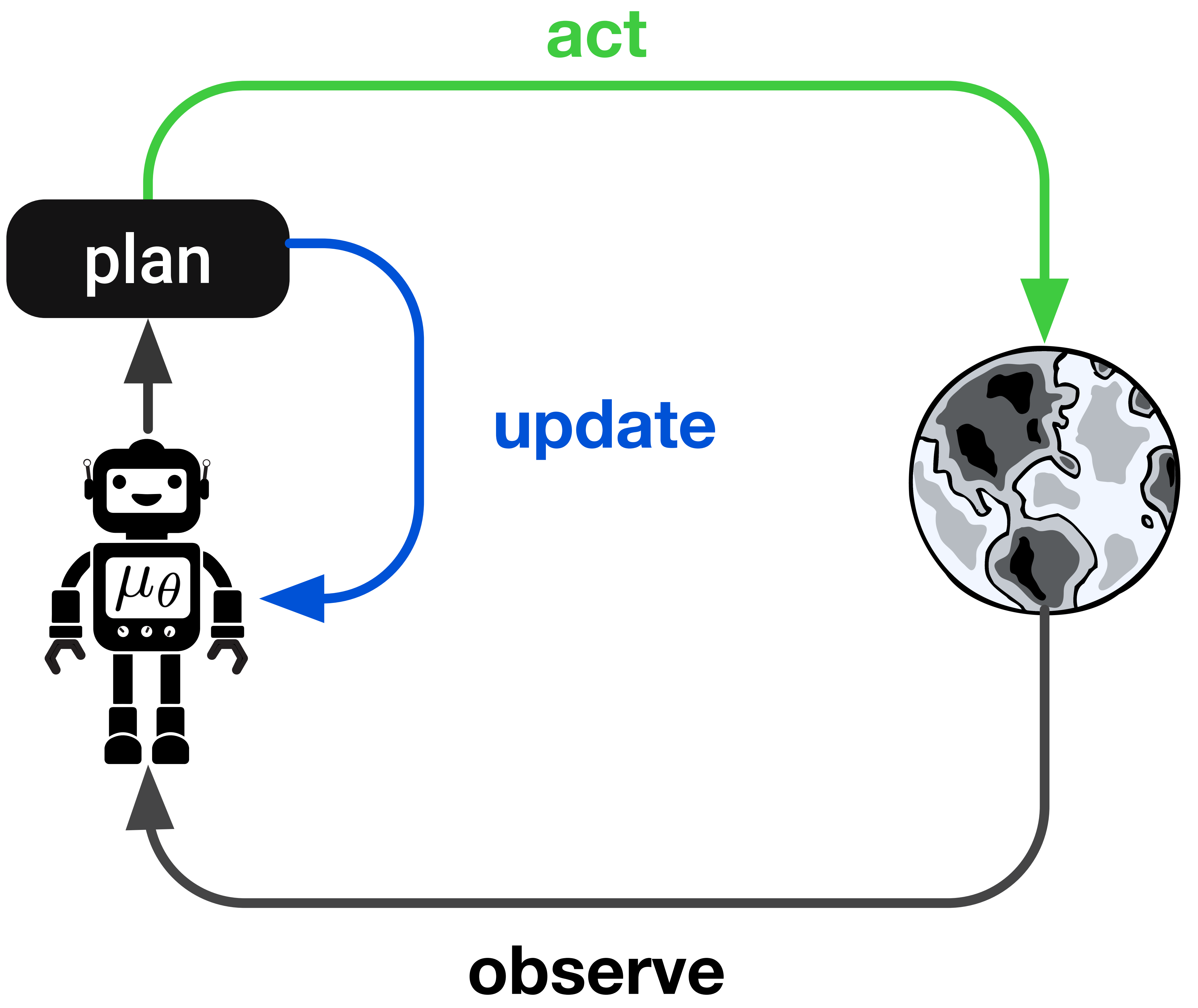

act
(only during learning)
plan
update
μ_θ
observe

Hamrick et al. (2021). On the role of planning in model-based deep reinforcement learning. ICLR.






Normalized Reward
0.0
0.2
0.4
0.6
0.8
1.0
1S
L
D
L+D
L+D+E
Acrobot
Cheetah
Humanoid
Hero
Ms. PM
Mini PM
Sokoban
9x9 Go


act
plan
update
μ_θ
observe

Q1: How does planning benefit model-based RL agents?

A: Primarily by constructing targets for learning & acting to obtain a useful data distribution.

Q2: Within planning, what algorithmic choices drive performance?

Q3: To what extent does planning improve zero-shot generalization?

Q1: How does planning benefit model-based RL agents?

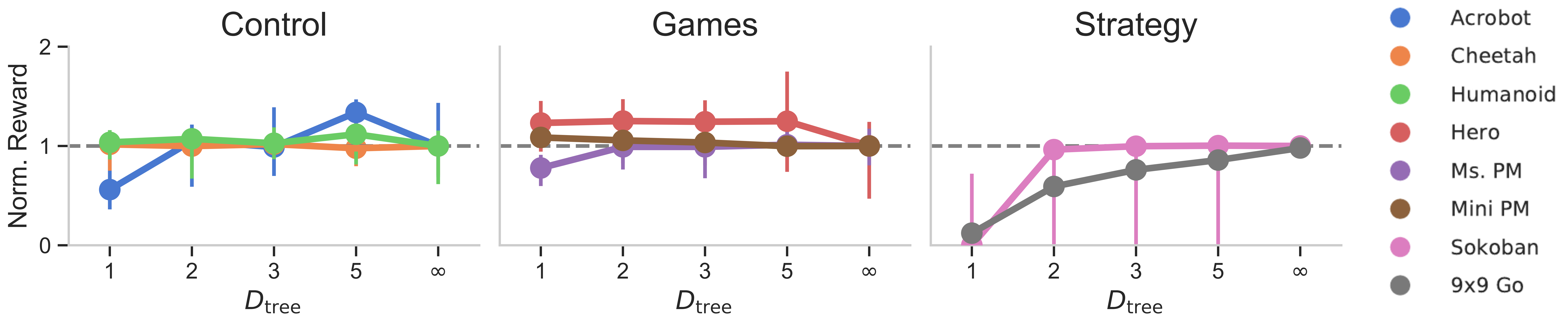
A: Primarily by constructing targets for learning & acting to obtain a useful data distribution.

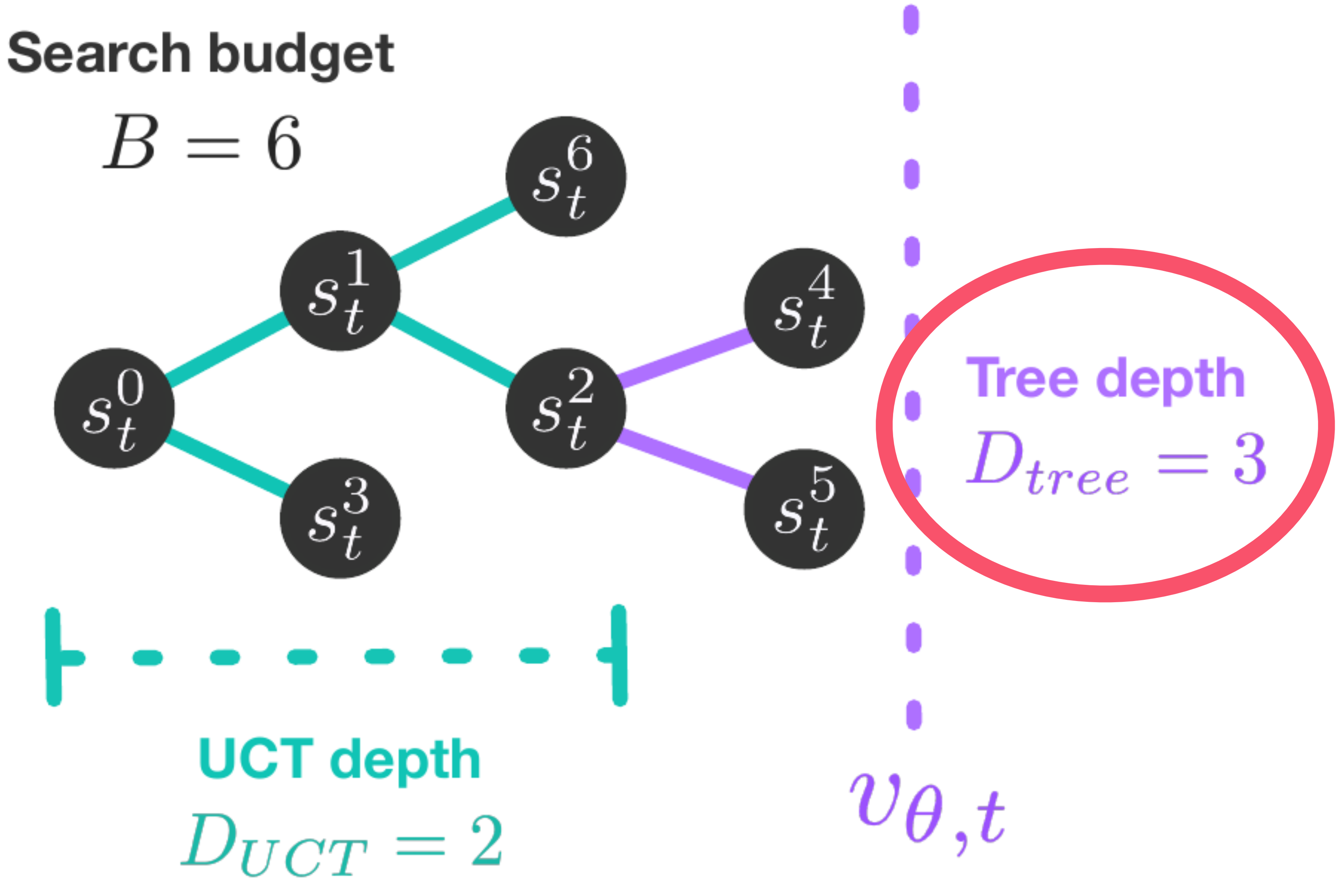
Q2: Within planning, what algorithmic choices drive performance?

Q3: To what extent does planning improve zero-shot generalization?

Effect of tree depth

$D_{UCT} = \infty$; B = 10 (Minipacman), 25 (Sokoban), 150 (Go), or 50 (otherwise)

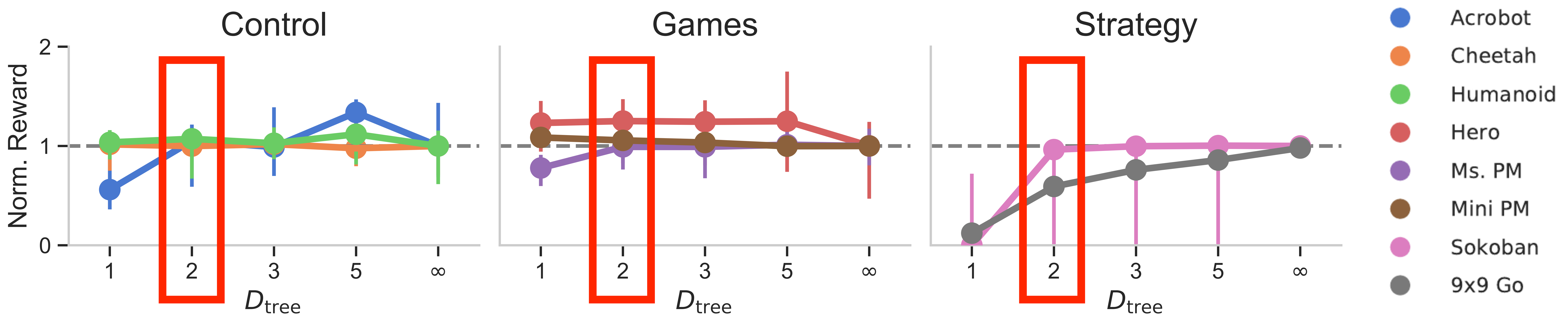


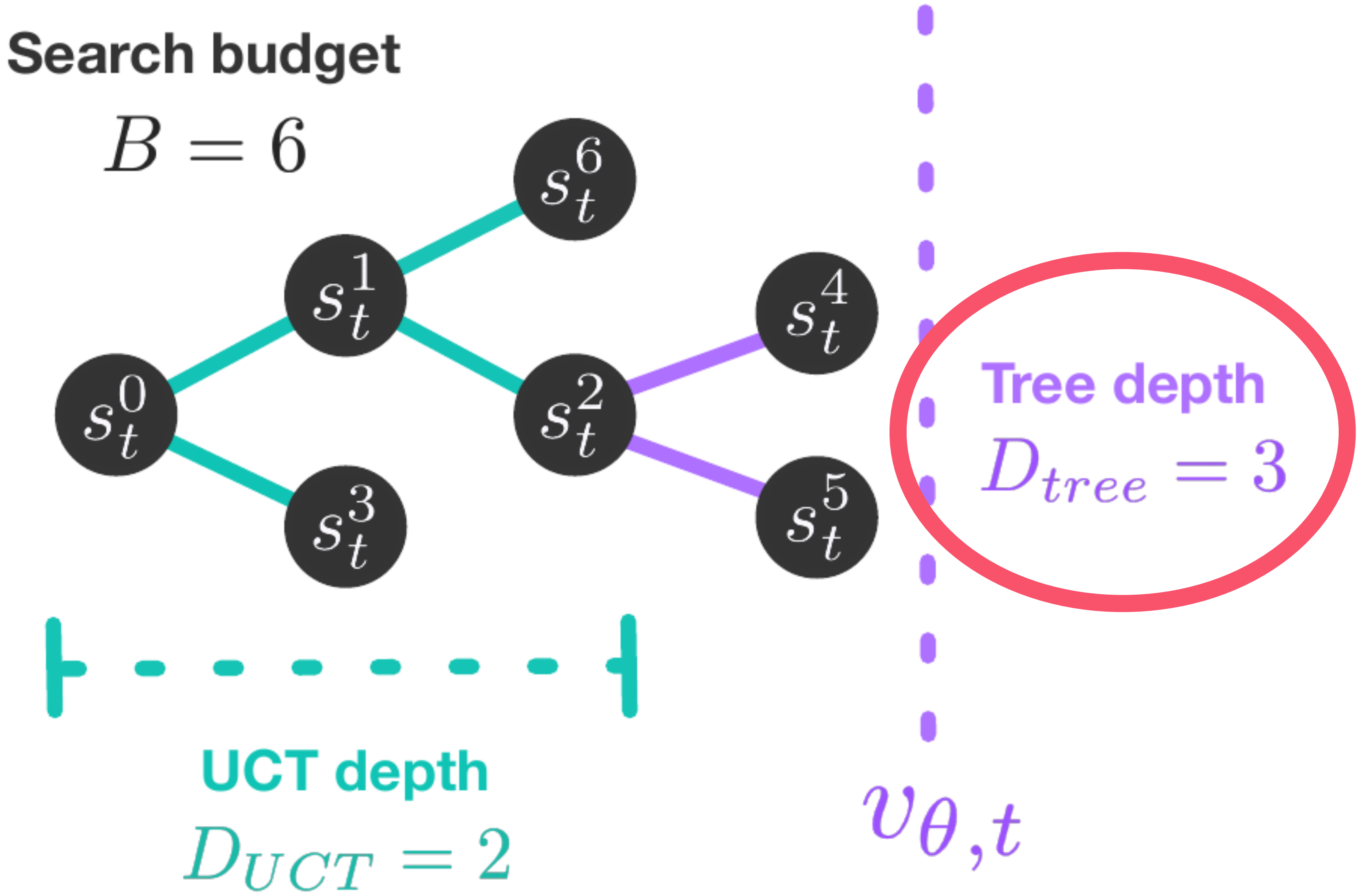


A little bit of lookahead is useful, but it does not need to be very deep to get good performance.

Effect of tree depth

$D_{UCT} = \infty$; B = 10 (Minipacman), 25 (Sokoban), 150 (Go), or 50 (otherwise)

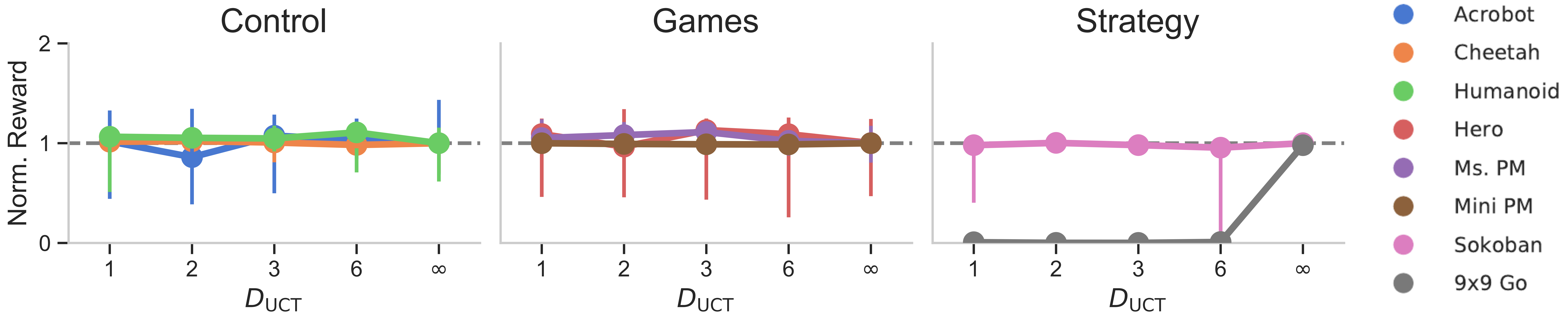


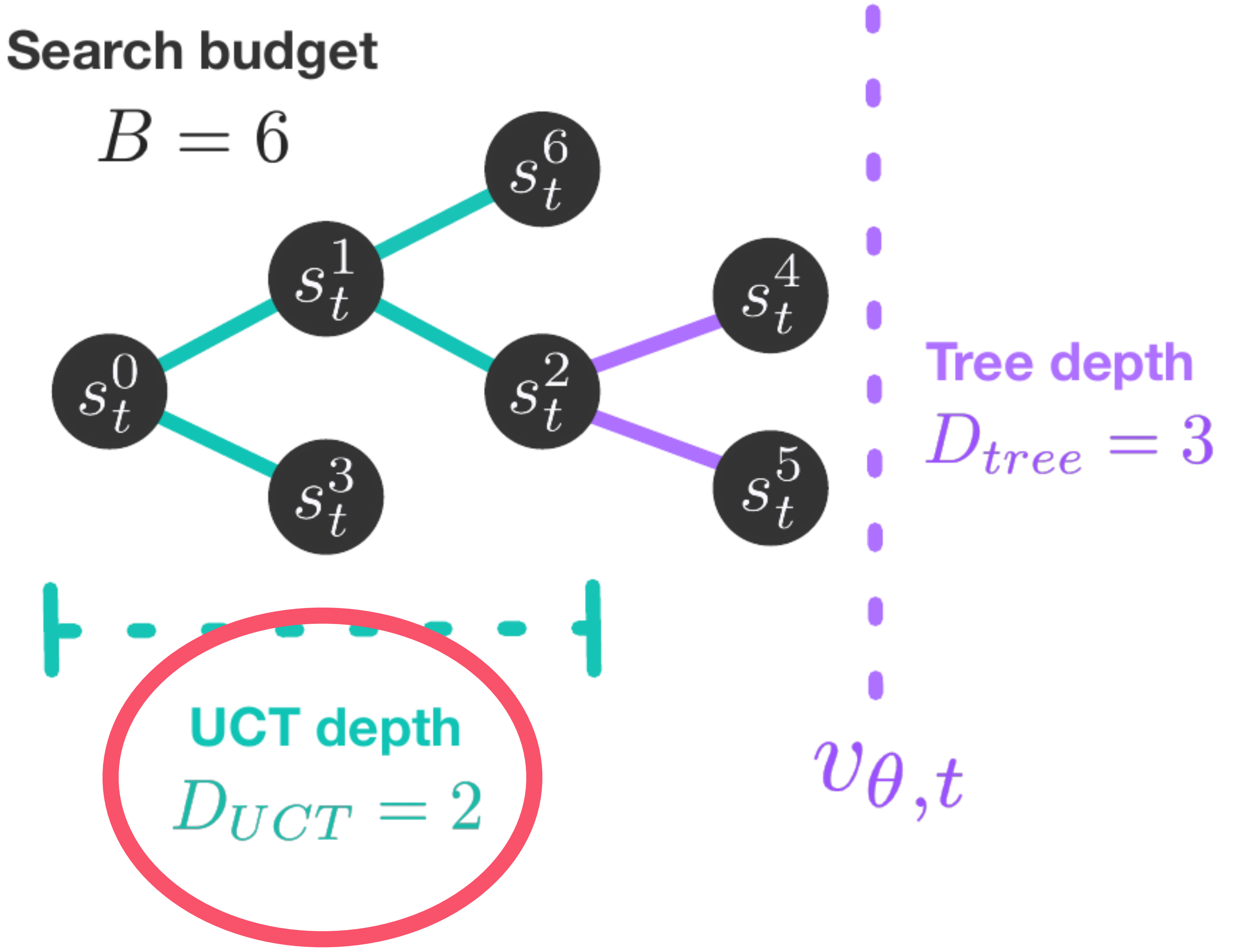


A little bit of lookahead is useful, but it does not need to be very deep to get good performance.

Effect of UCT depth

$D_{tree} = \infty$; B = 10 (Minipacman), 25 (Sokoban), 150 (Go), or 50 (otherwise)

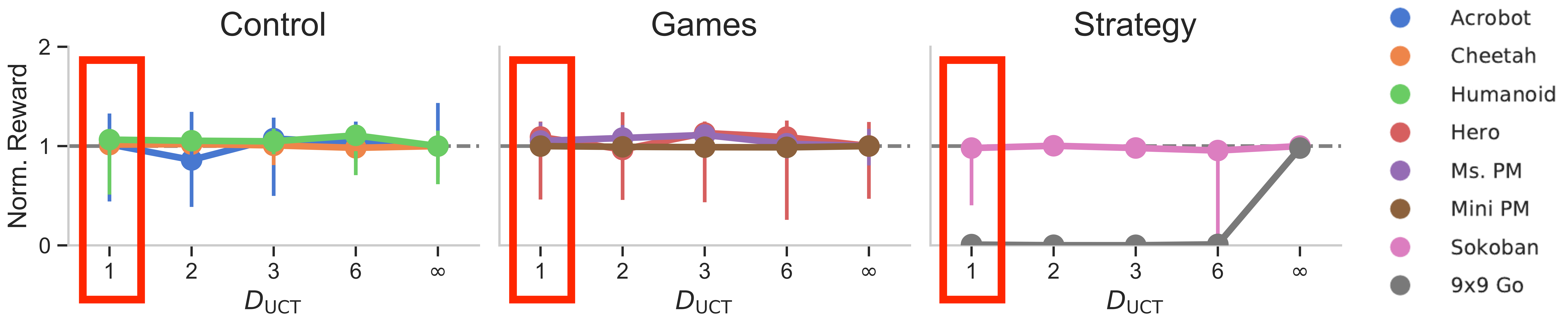


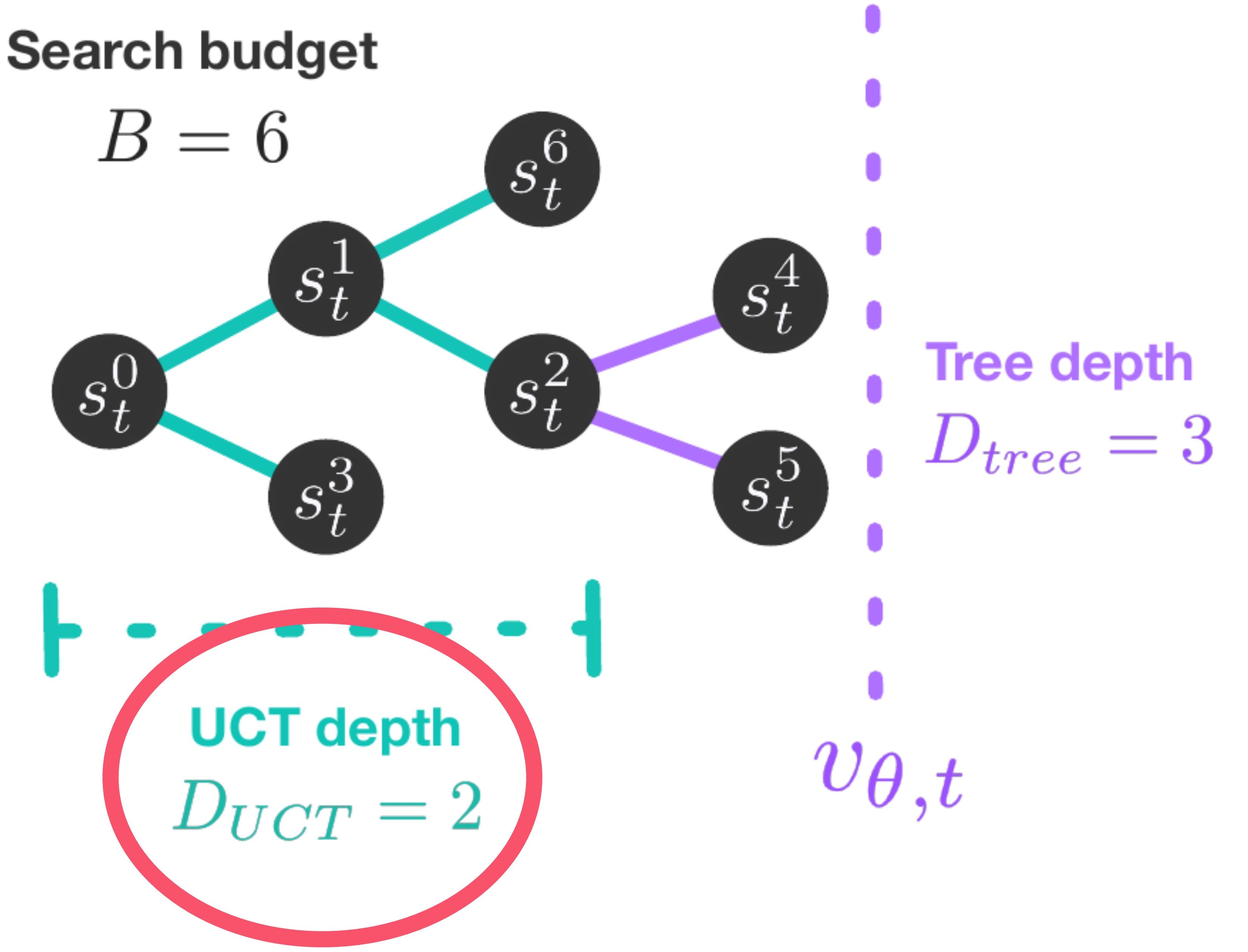


Complex planning ("precise and sophisticated lookahead") does not seem to be needed in common MBRL environments.

Effect of UCT depth

$D_{tree} = \infty$; B = 10 (Minipacman), 25 (Sokoban), 150 (Go), or 50 (otherwise)

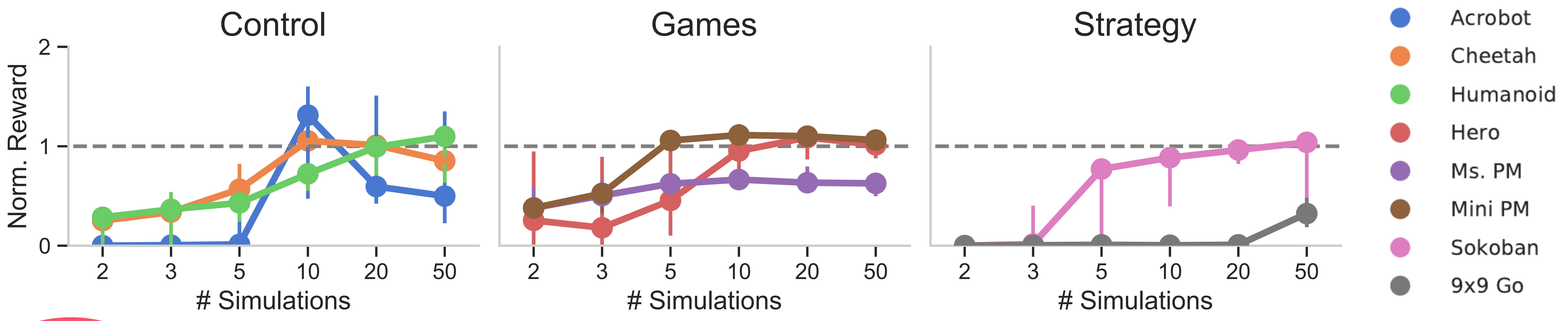


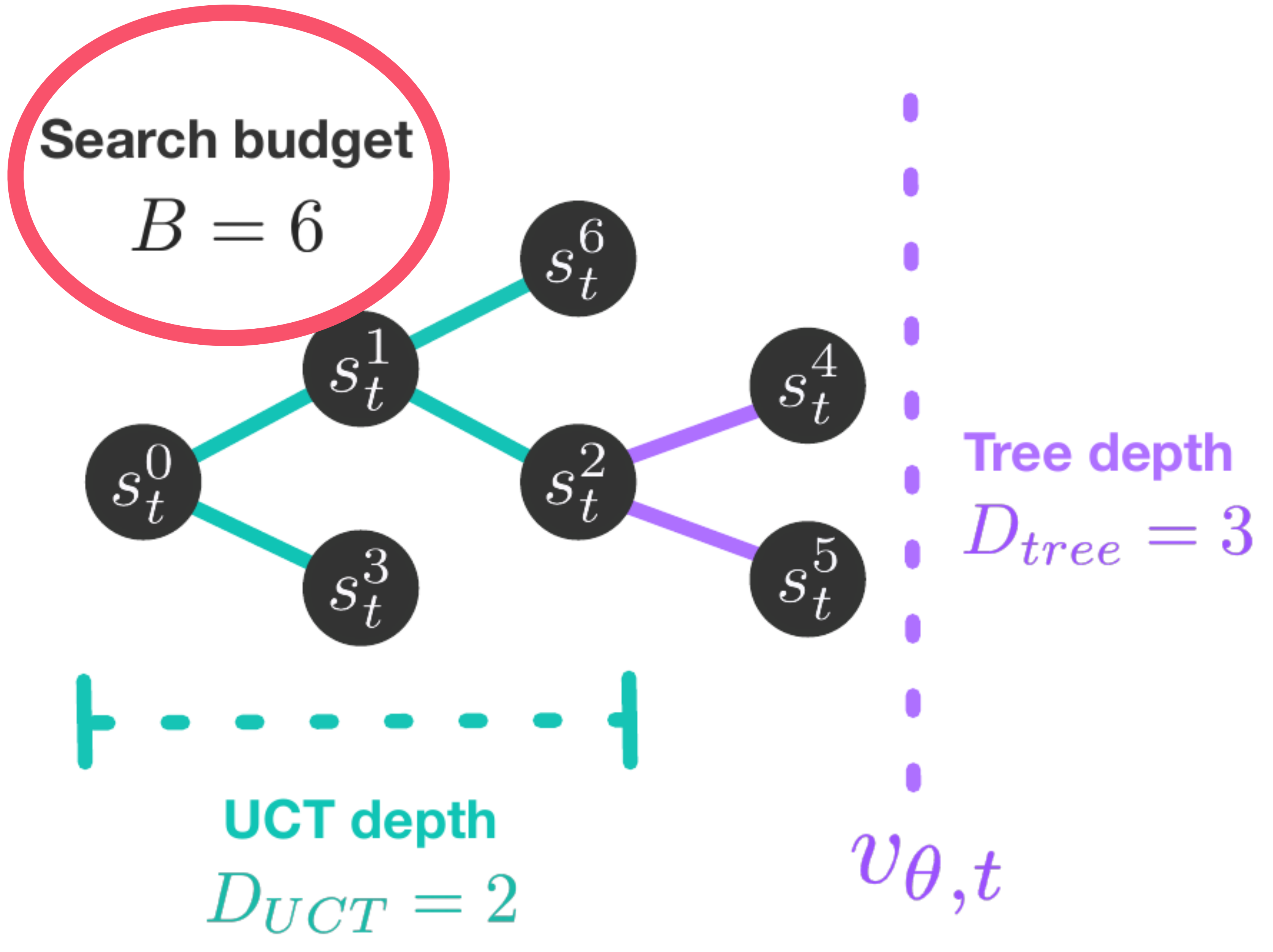


Complex planning ("precise and sophisticated lookahead") does not seem to be needed in common MBRL environments.

Effect of search budget

$D_{UCT} = 1$ (except Go, where $D_{UCT} = \infty$); $D_{tree} = \infty$

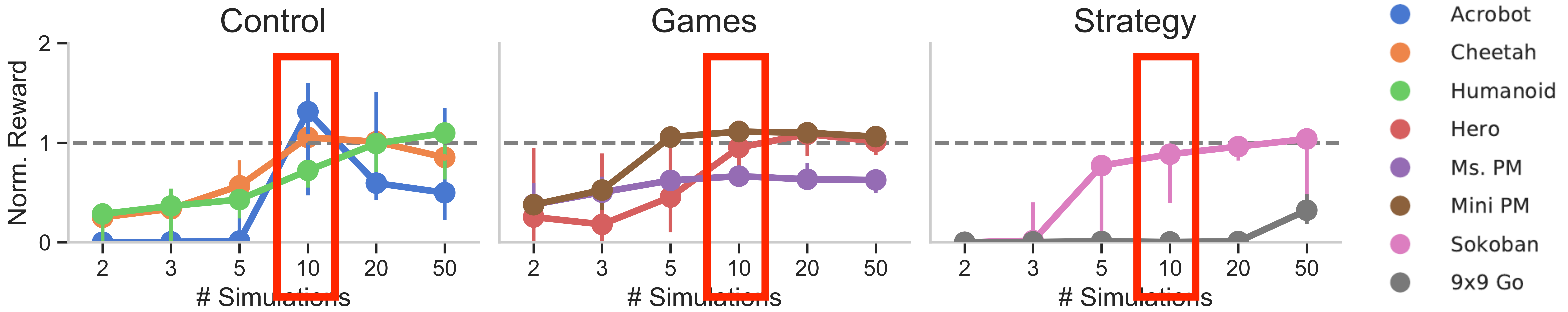


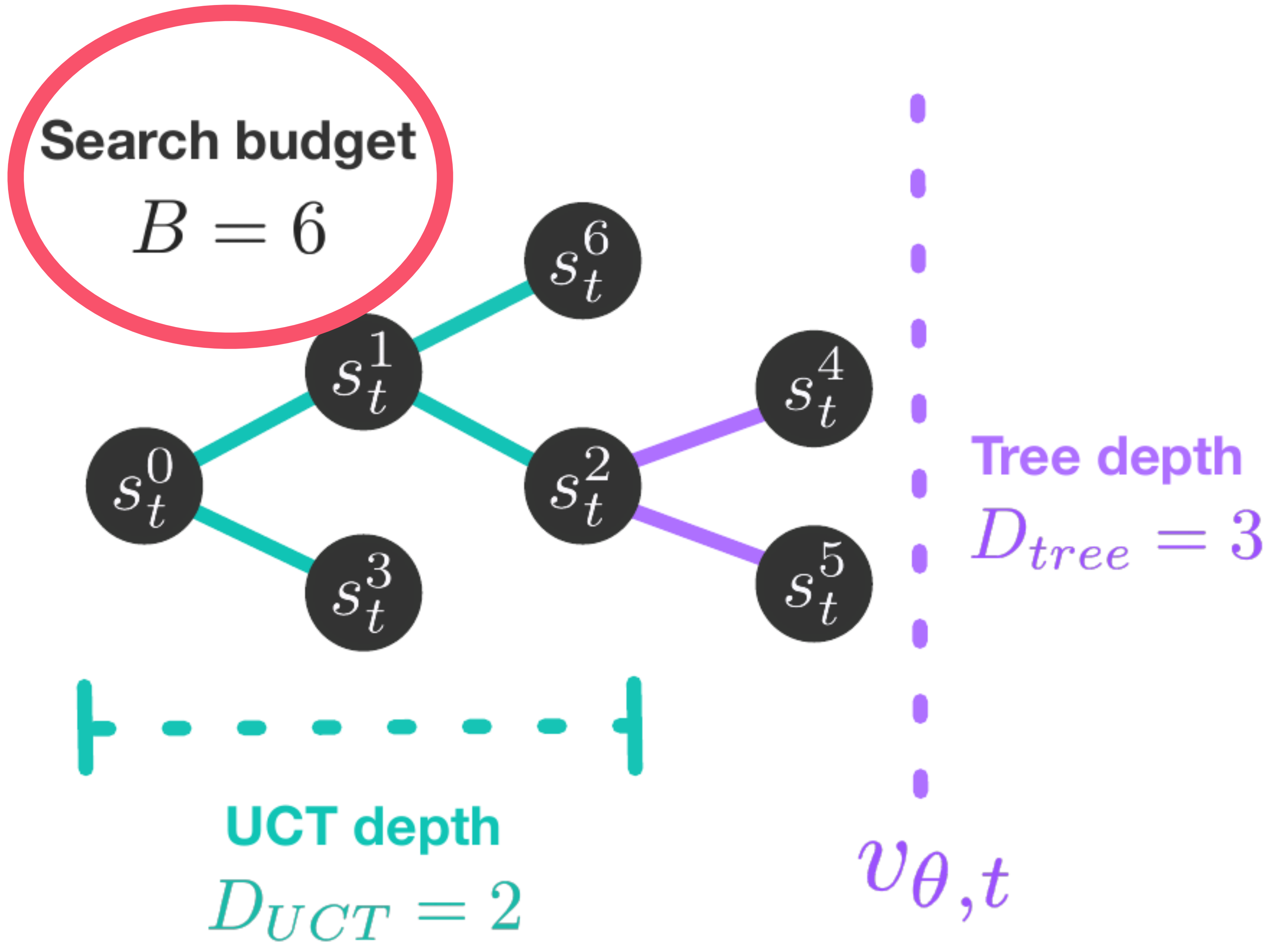


Moderate amounts of search (neither too much nor too little) results in best performance.

Effect of search budget

D_{UCT} = 1 (except Go, where $D_{UCT} = \infty$); $D_{tree} = \infty$





Moderate amounts of search (neither too much nor too little) results in best performance.

Q1: How does planning benefit model-based RL agents?

A: Primarily by constructing targets for learning & acting to obtain a useful data distribution.

Q2: Within planning, what algorithmic choices drive performance?

A: Number of simulations during training. Planning depth and complexity matter less.

Q3: To what extent does planning improve zero-shot generalization?

Q1: How does planning benefit model-based RL agents?

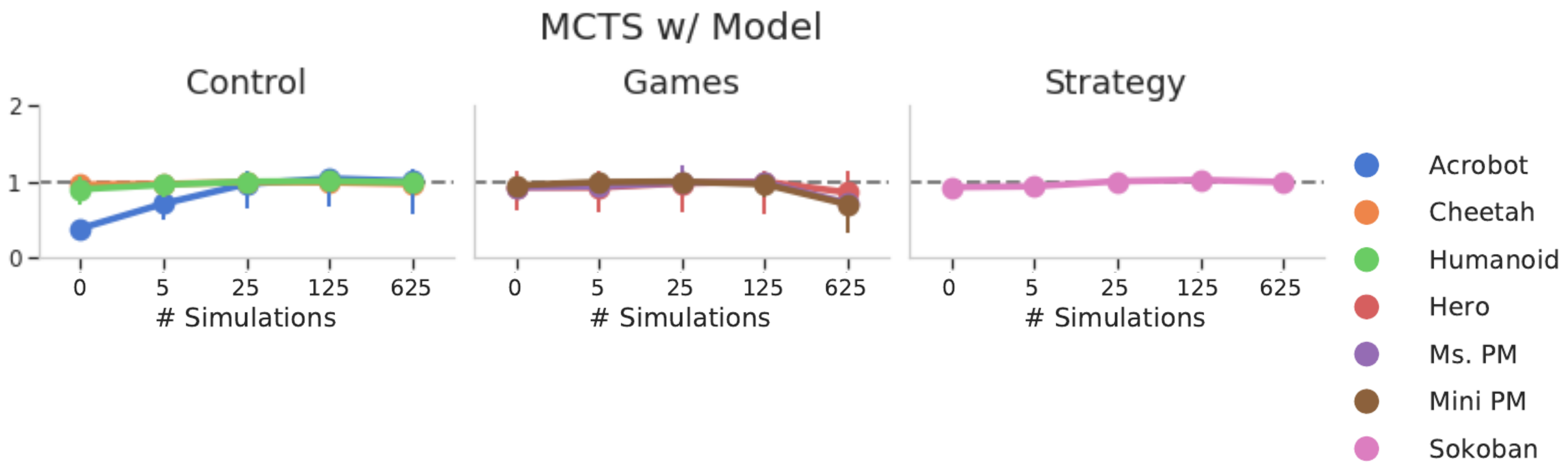
A: Primarily by constructing targets for learning & acting to obtain a useful data distribution.

Q2: Within planning, what algorithmic choices drive performance?

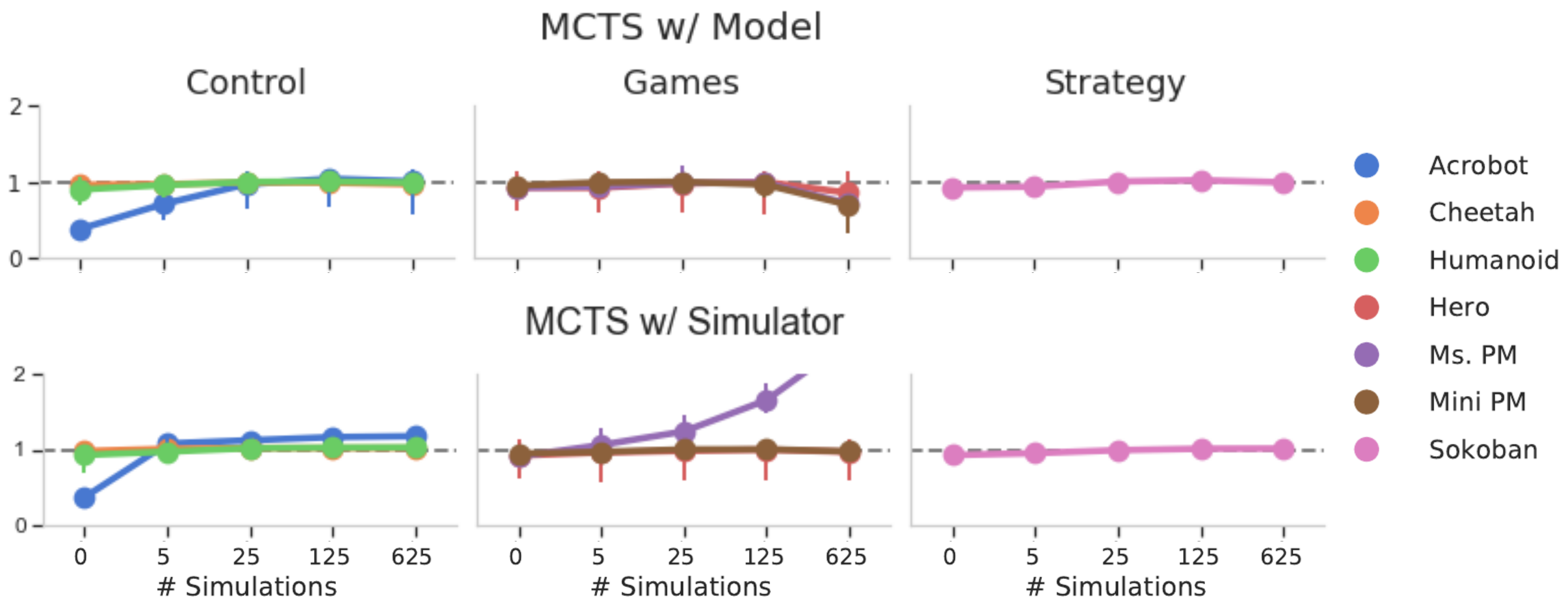
A: Number of simulations during training. Planning depth and complexity matter less.

Q3: To what extent does planning improve zero-shot generalization?

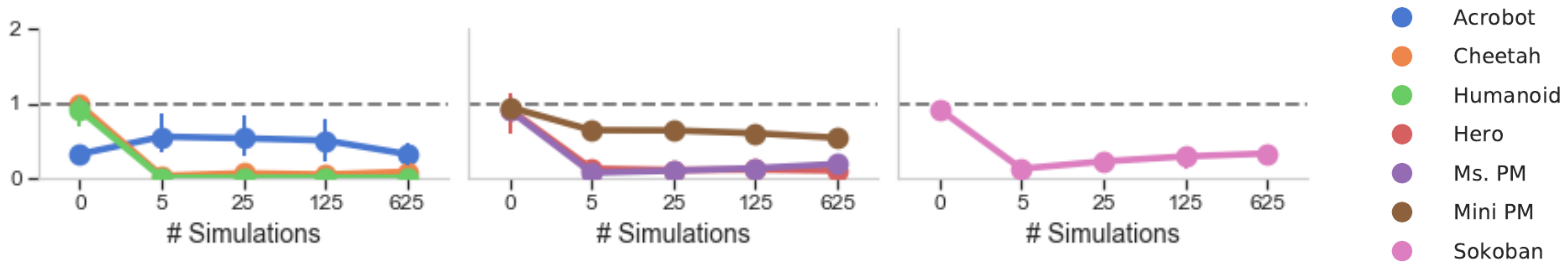
Model generalization to new search budgets



Model generalization to new search budgets

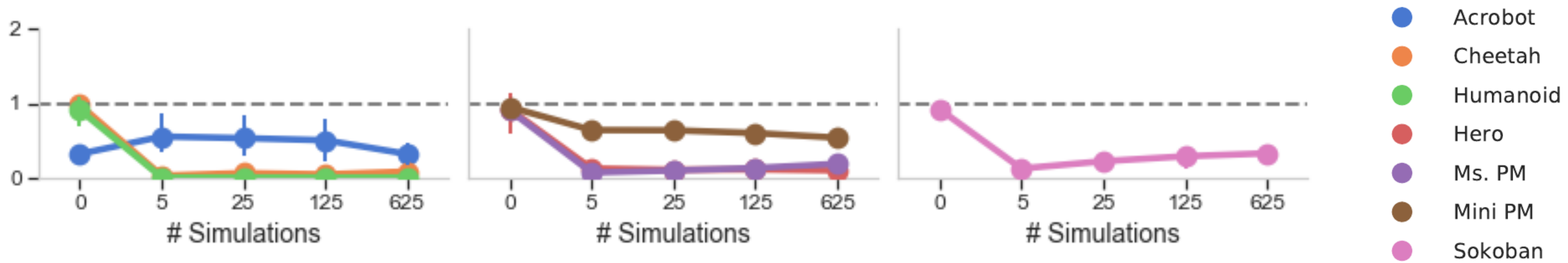


Value generalization to new planners (BFS)



Jessica Hamrick - jhamrick@deepmind.com

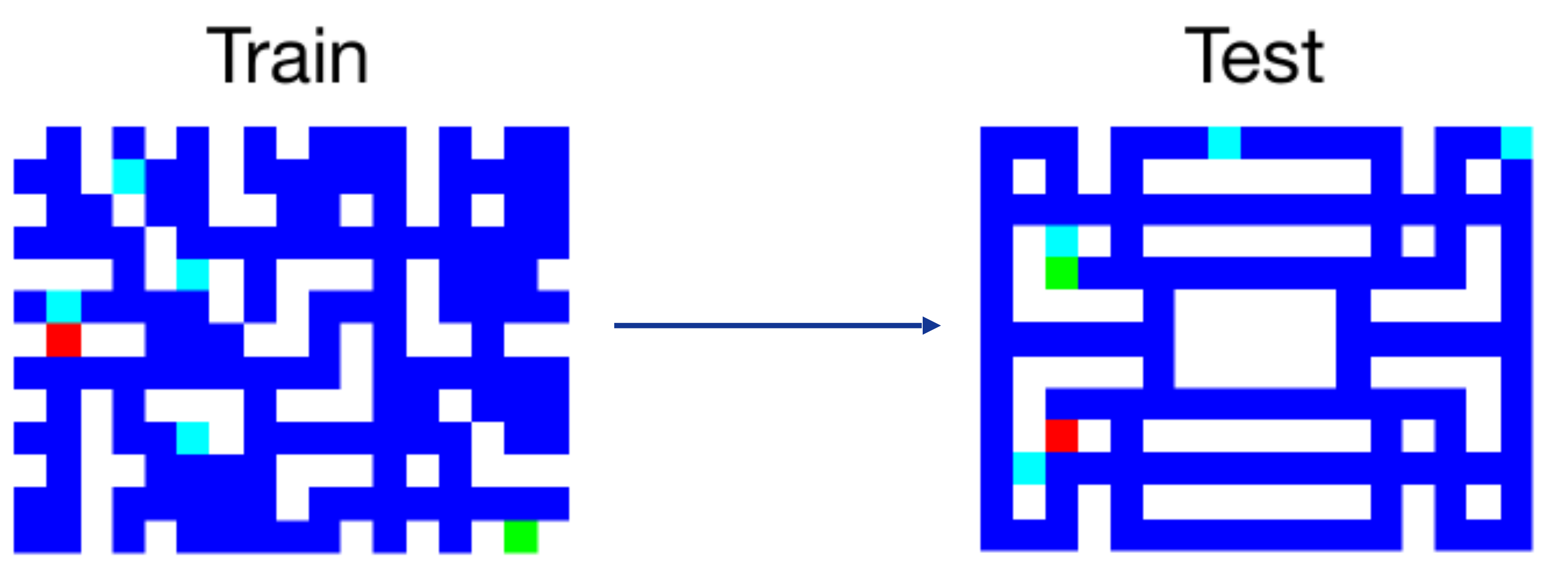
Value generalization to new planners (BFS)

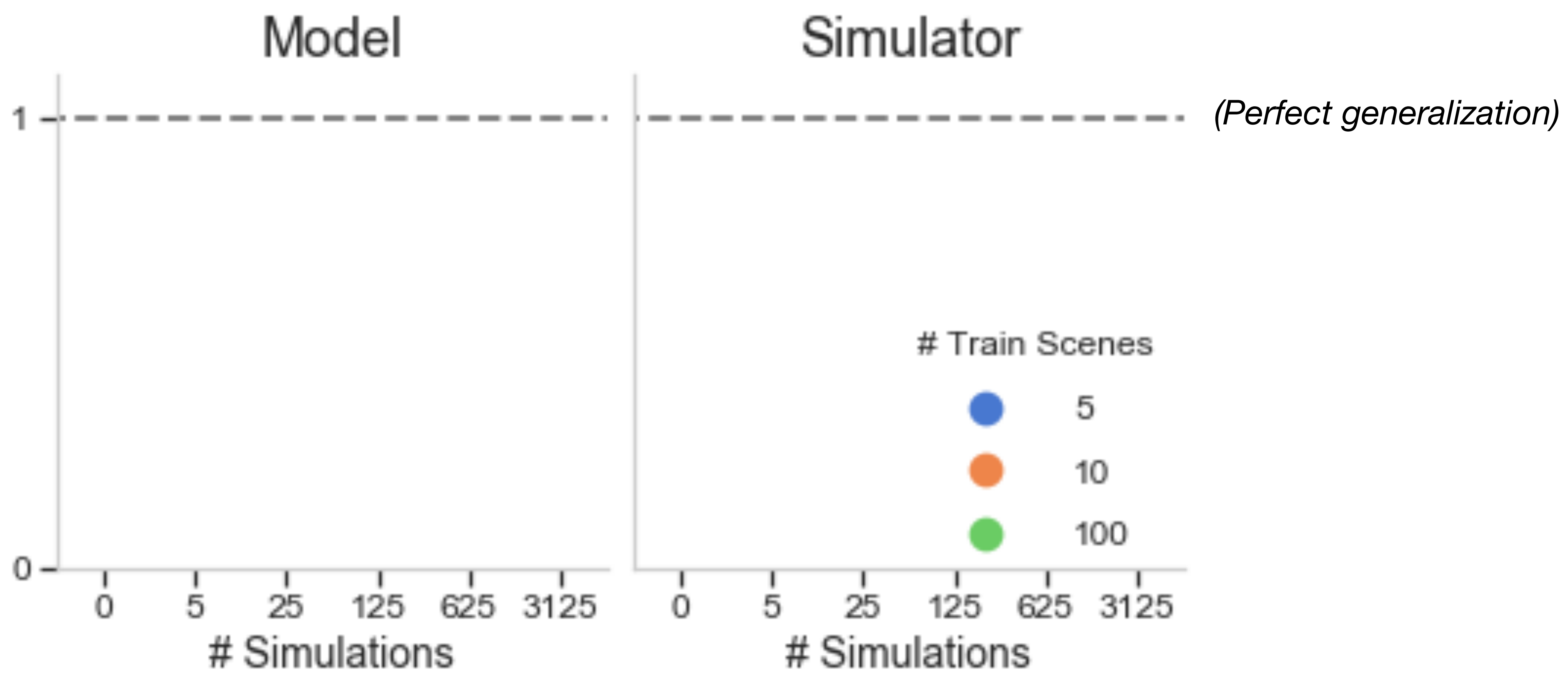


Errors in the model of the world (i.e. transition function) are not the only types of error to be concerned about.

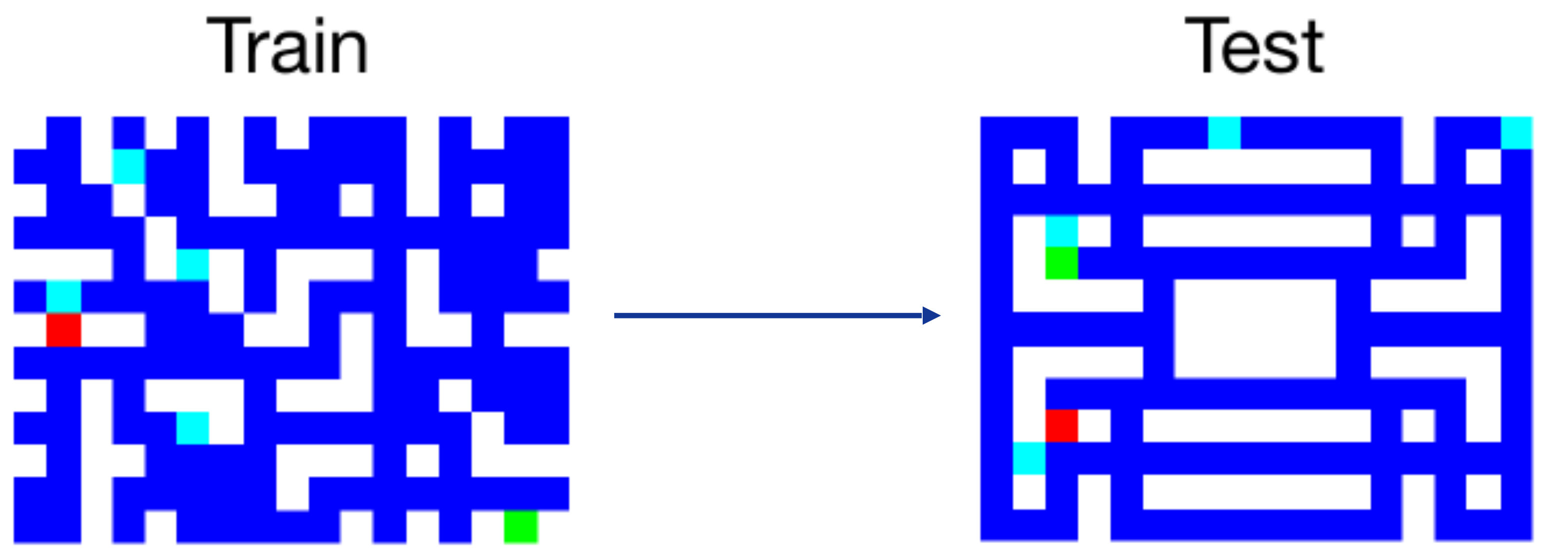
Jessica Hamrick - jhamrick@deepmind.com

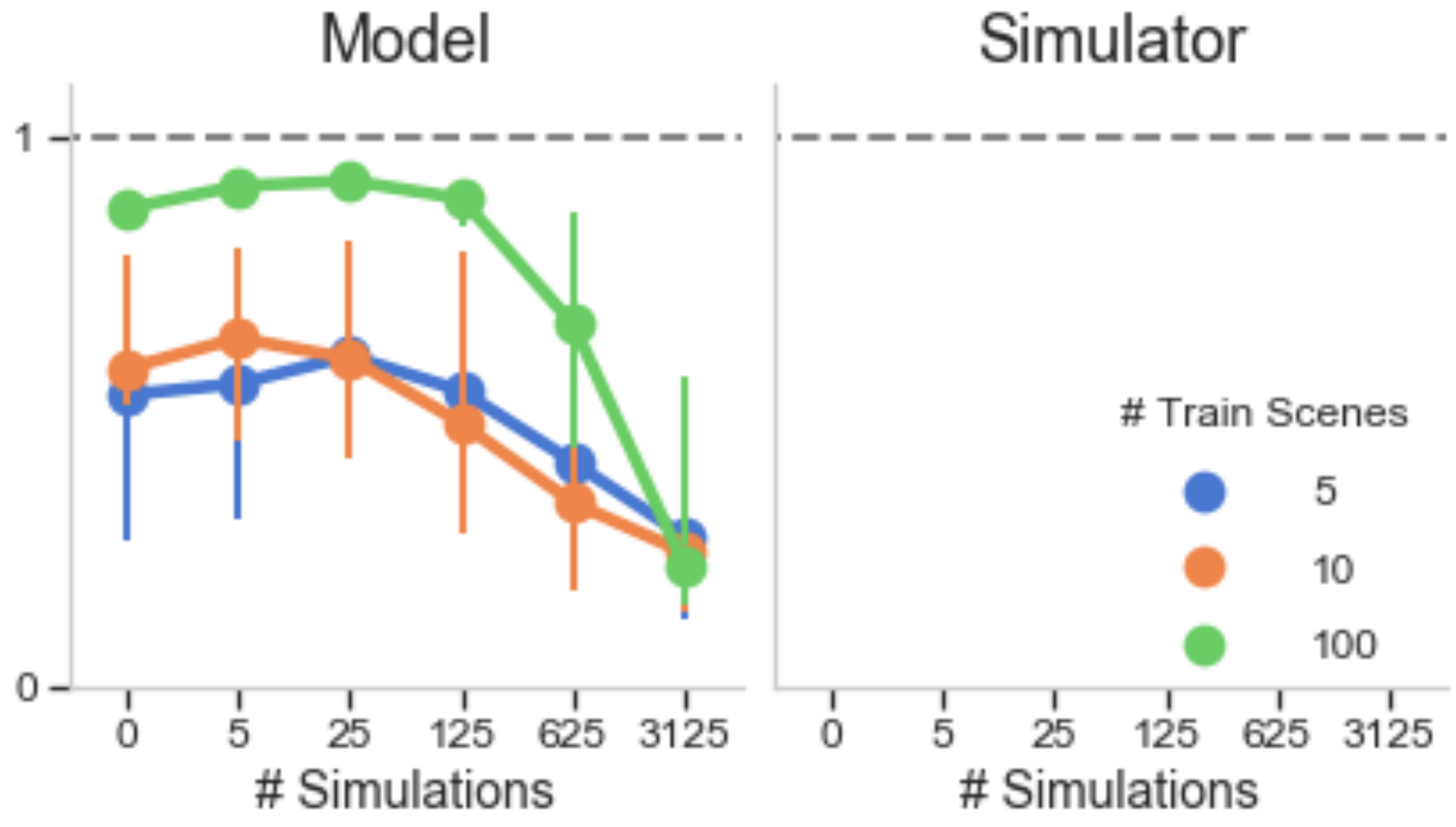
Generalizing to new mazes





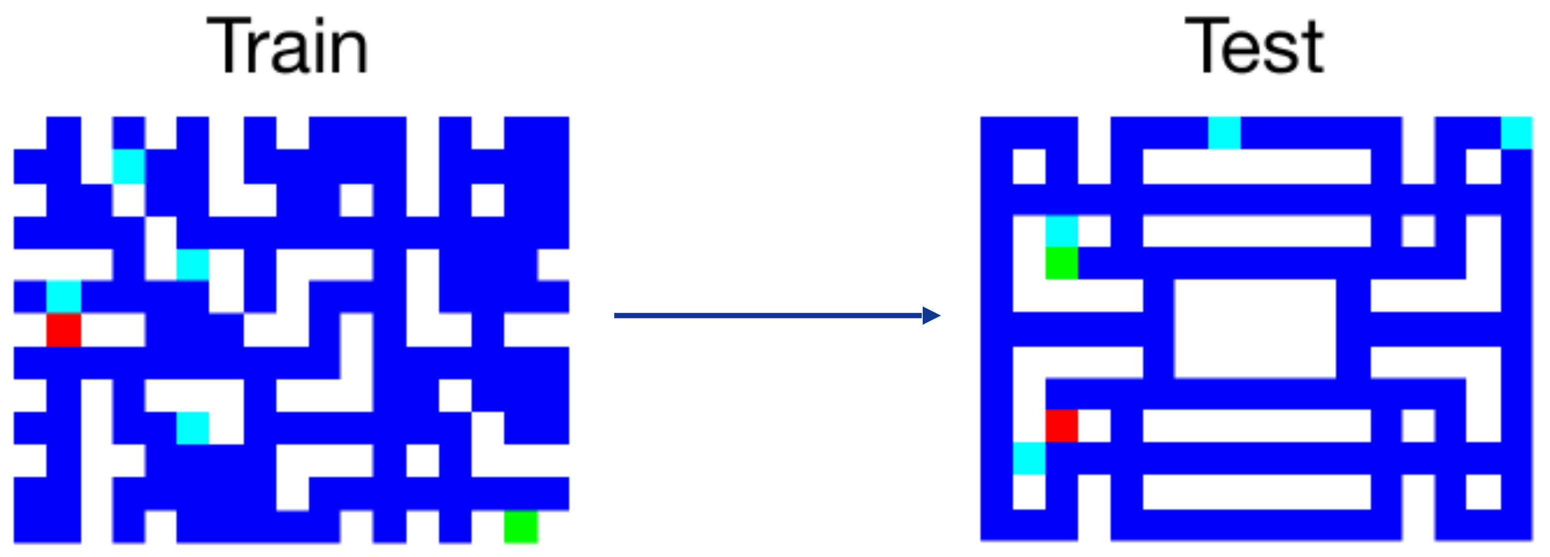
Generalizing to new mazes

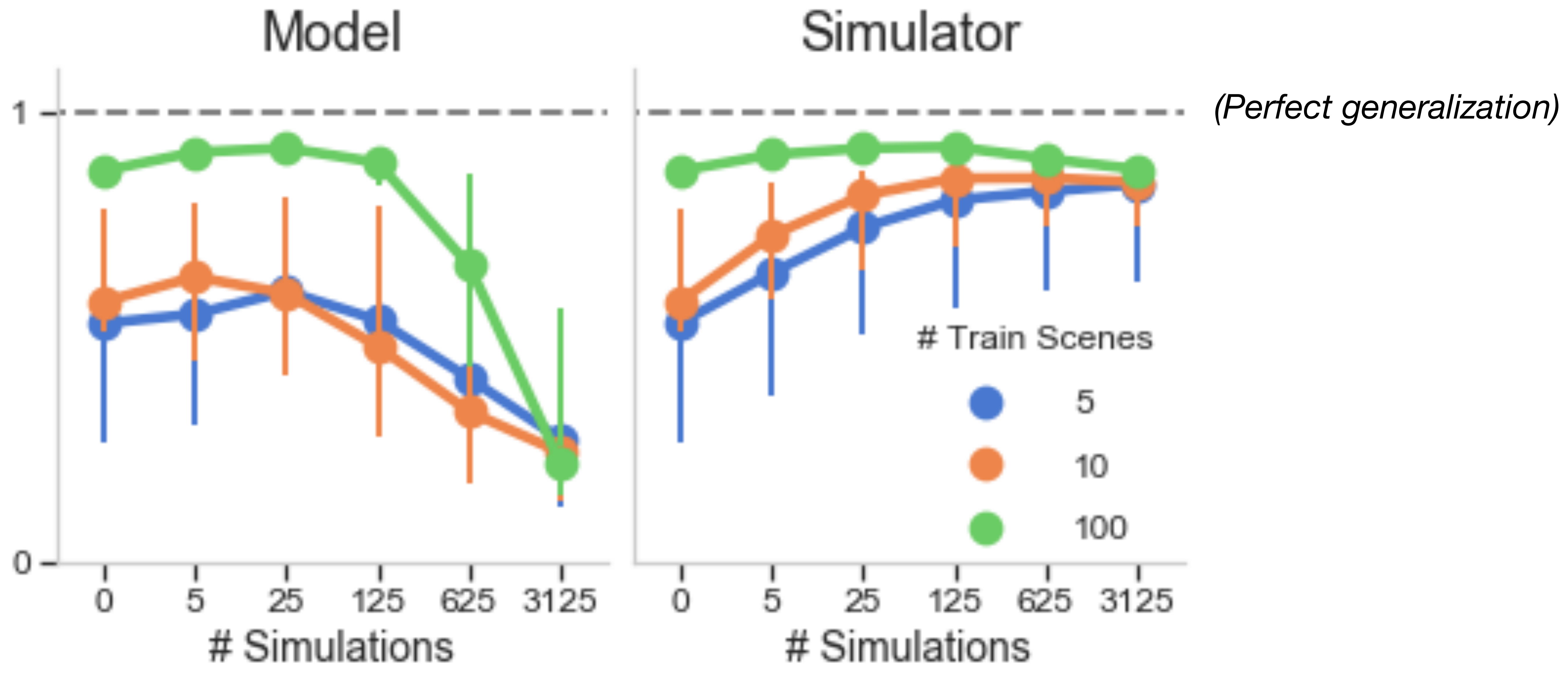




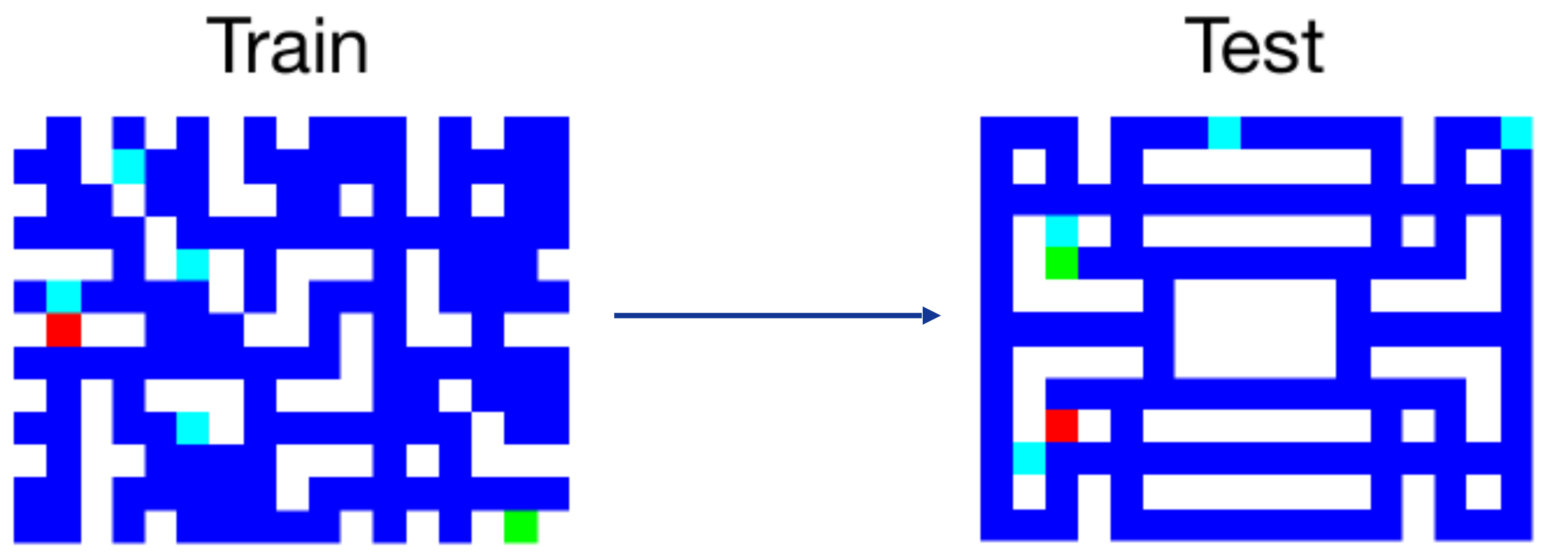
(Perfect generalization)

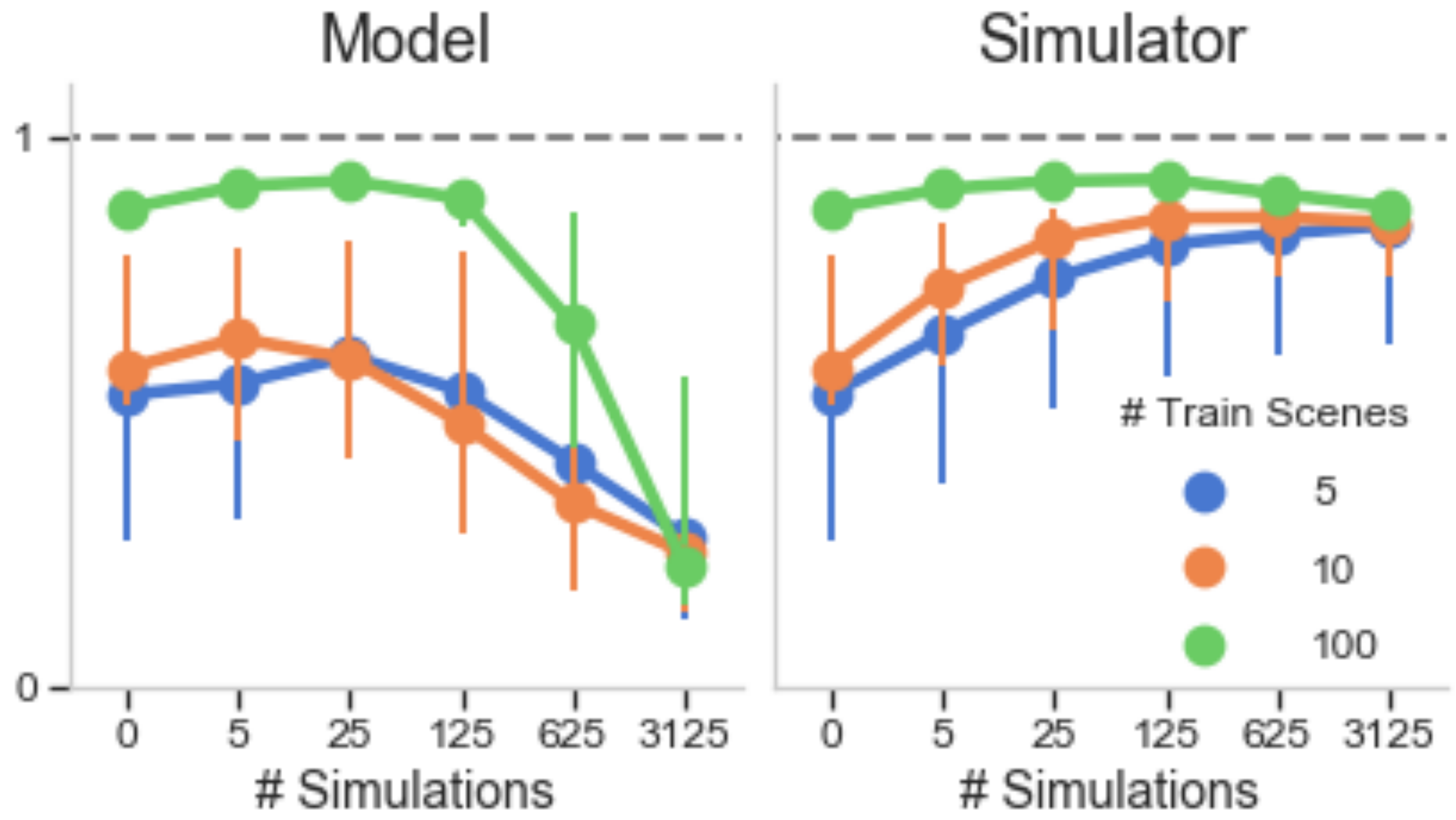
Generalizing to new mazes





Generalizing to new mazes





(Perfect generalization)

Planning—even with a perfect model—does not guarantee good generalization performance.

Jessica Hamrick - jhamrick@deepmind.com

Q1: How does planning benefit model-based RL agents?

A: Primarily by constructing targets for learning & acting to obtain a useful data distribution.

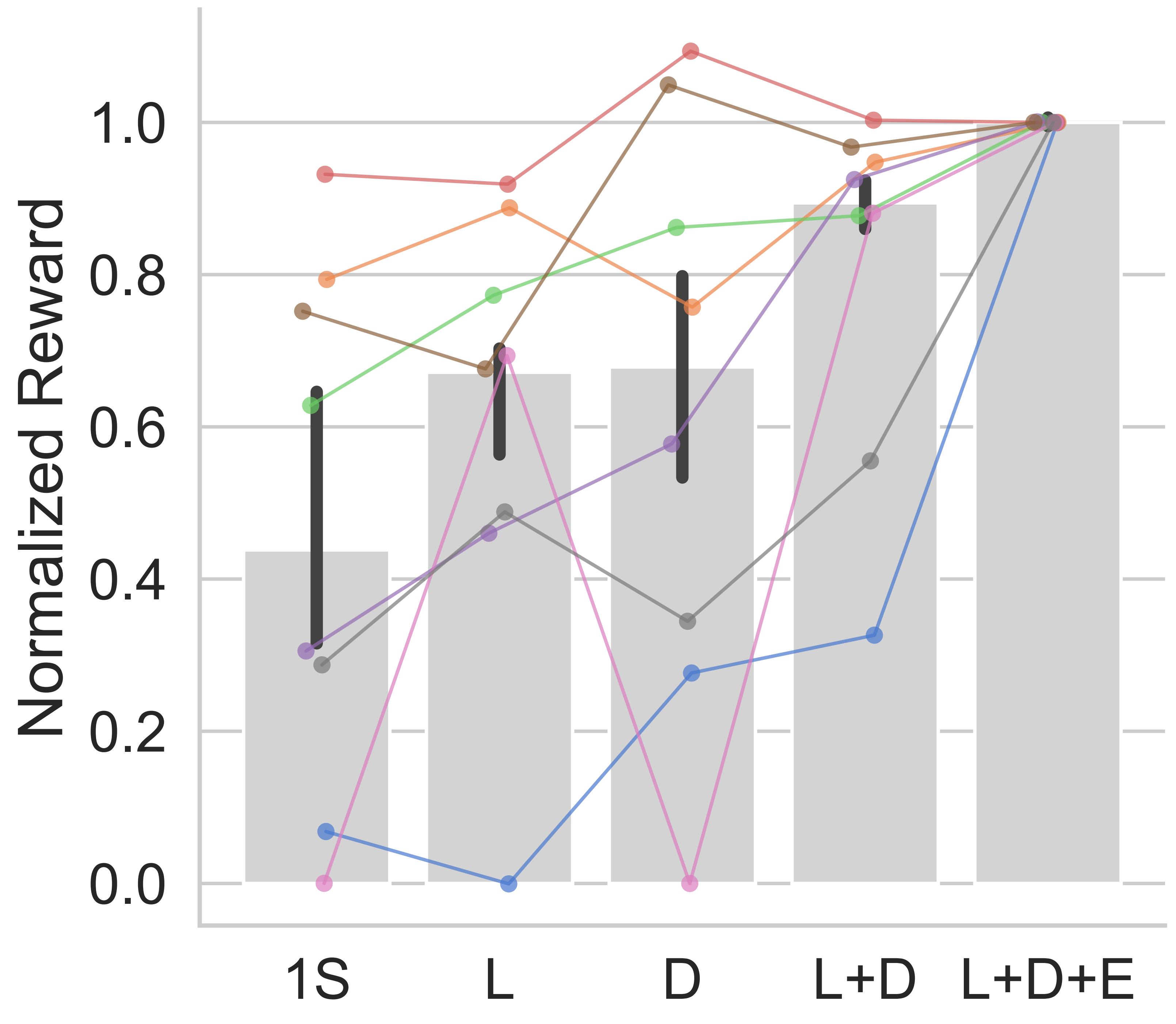
Q2: Within planning, what algorithmic choices drive performance?

A: Number of simulations during training. Planning depth and complexity matter less.

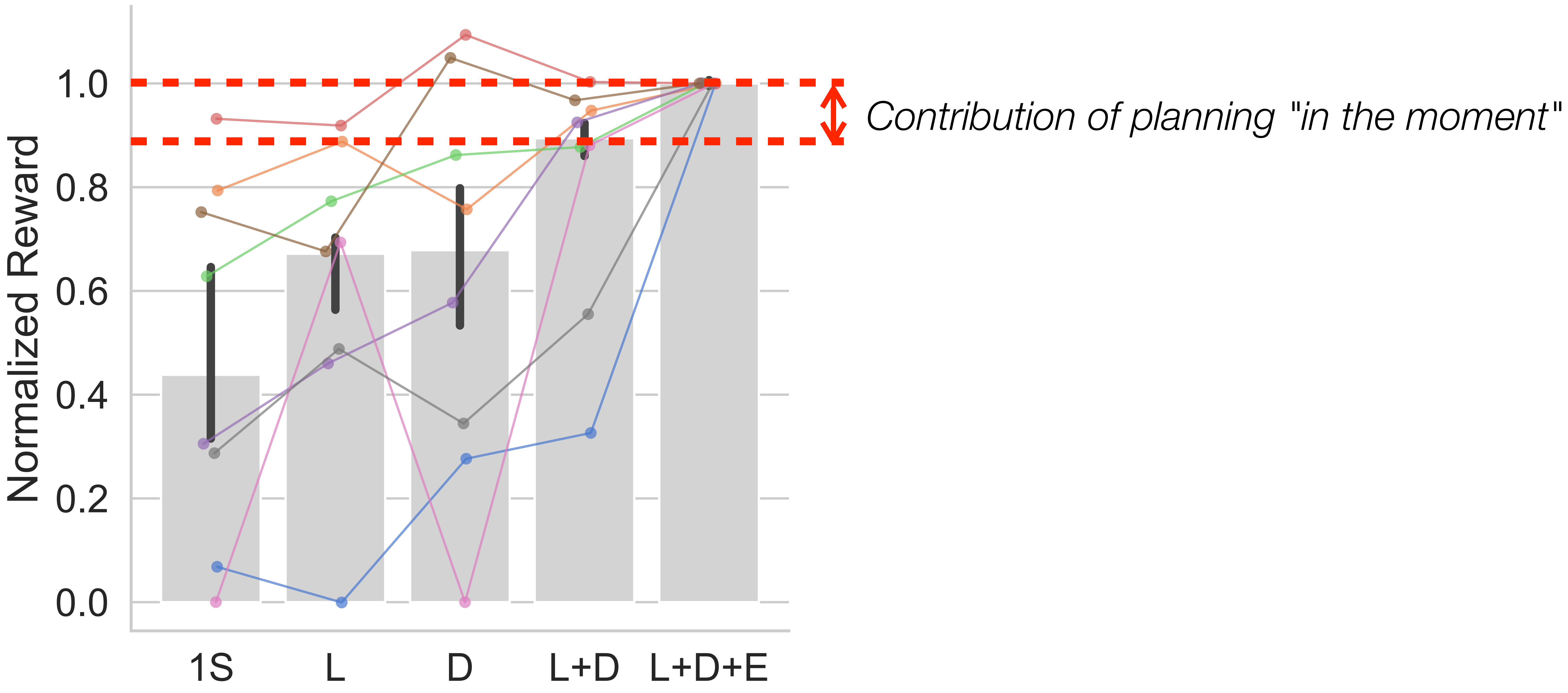
Q3: To what extent does planning improve zero-shot generalization?

A: Not as much as you might think, even with a perfect model!

Interim Takeaway #1: Planning seems to be most useful during learning and less so at test time (in most environments).

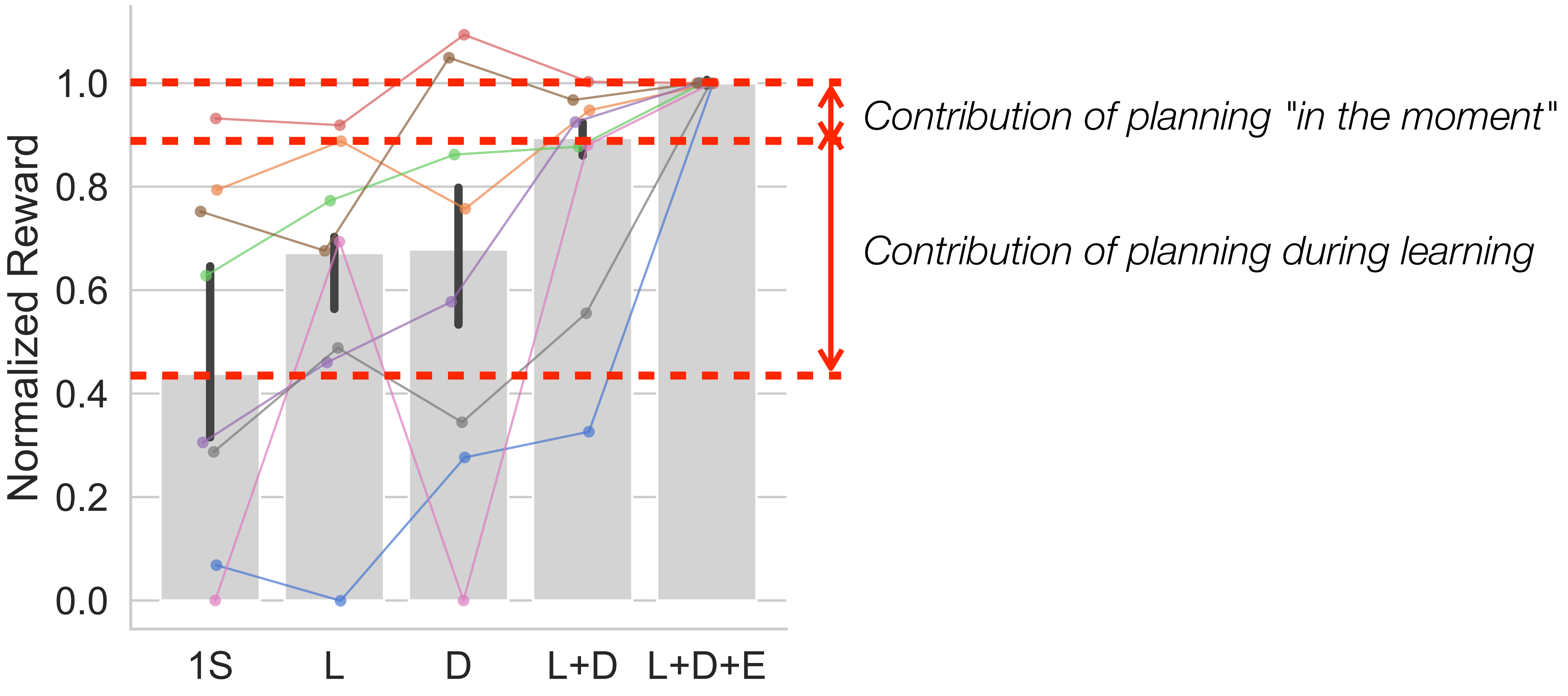


Interim Takeaway #1: Planning seems to be most useful during learning and less so at test time (in most environments).



Jessica Hamrick - jhamrick@deepmind.com

Interim Takeaway #1: Planning seems to be most useful during learning and less so at test time (in most environments).



Interim Takeaway #1: Planning seems to be most useful during learning and less so at test time (in most environments).

Interim takeaway #2: Effective planning requires having good representations for multiple components (policy/value/model).

Outline

- **Understanding MBRL**
  Hamrick et al. (2021). On the role of planning in model based reinforcement learning. ICLR.

- **Understanding and improving generalization**
  Anand, Walker et al. (2022). Procedural generalization by planning with self-supervised world models. ICLR.

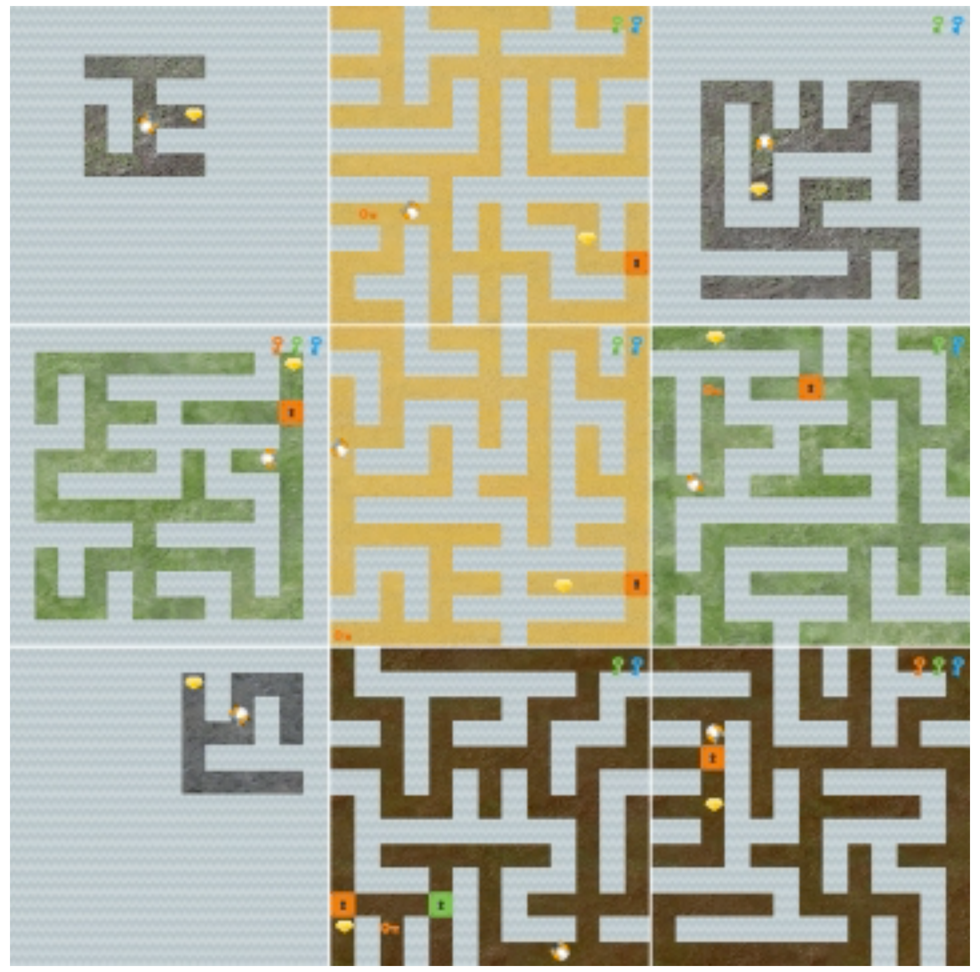
- **Understanding and improving transfer**
  Walker, Vértes, Li et al. (2023). Investigating the role of model-based learning in exploration and transfer. Under review.

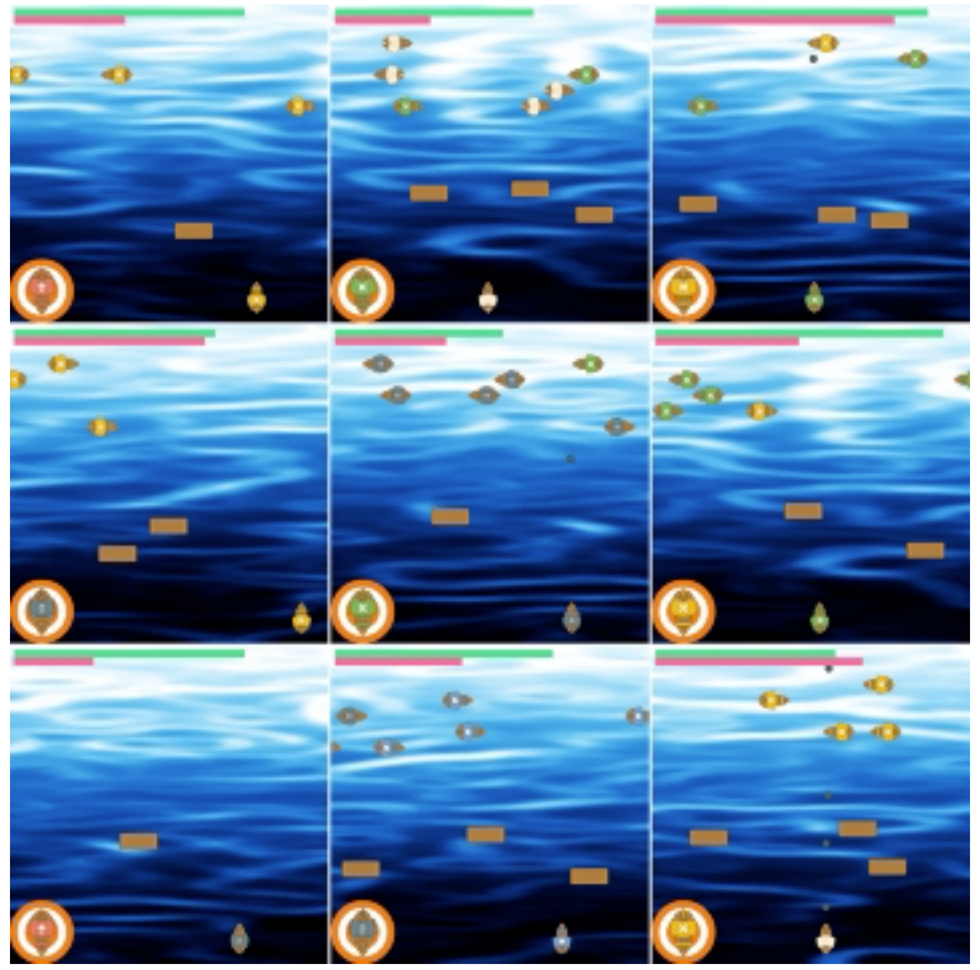
- **The future of MBRL**

Procedural generalization

Train on a **procedurally-generated** distribution of environments
Zero-shot generalization to unseen environments
(e.g. Procgen, Cobbe et al., 2020)

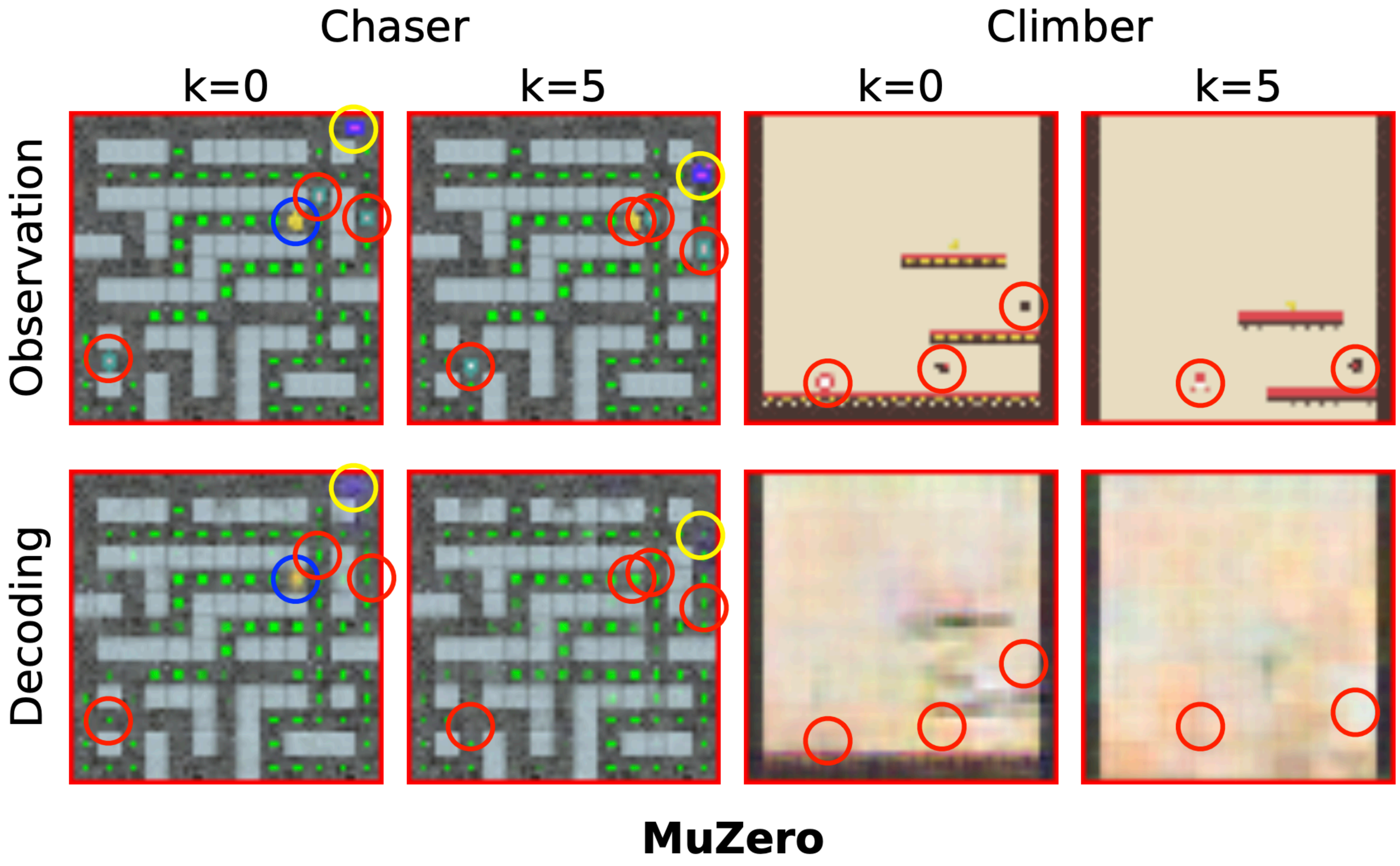
Procedural generalization



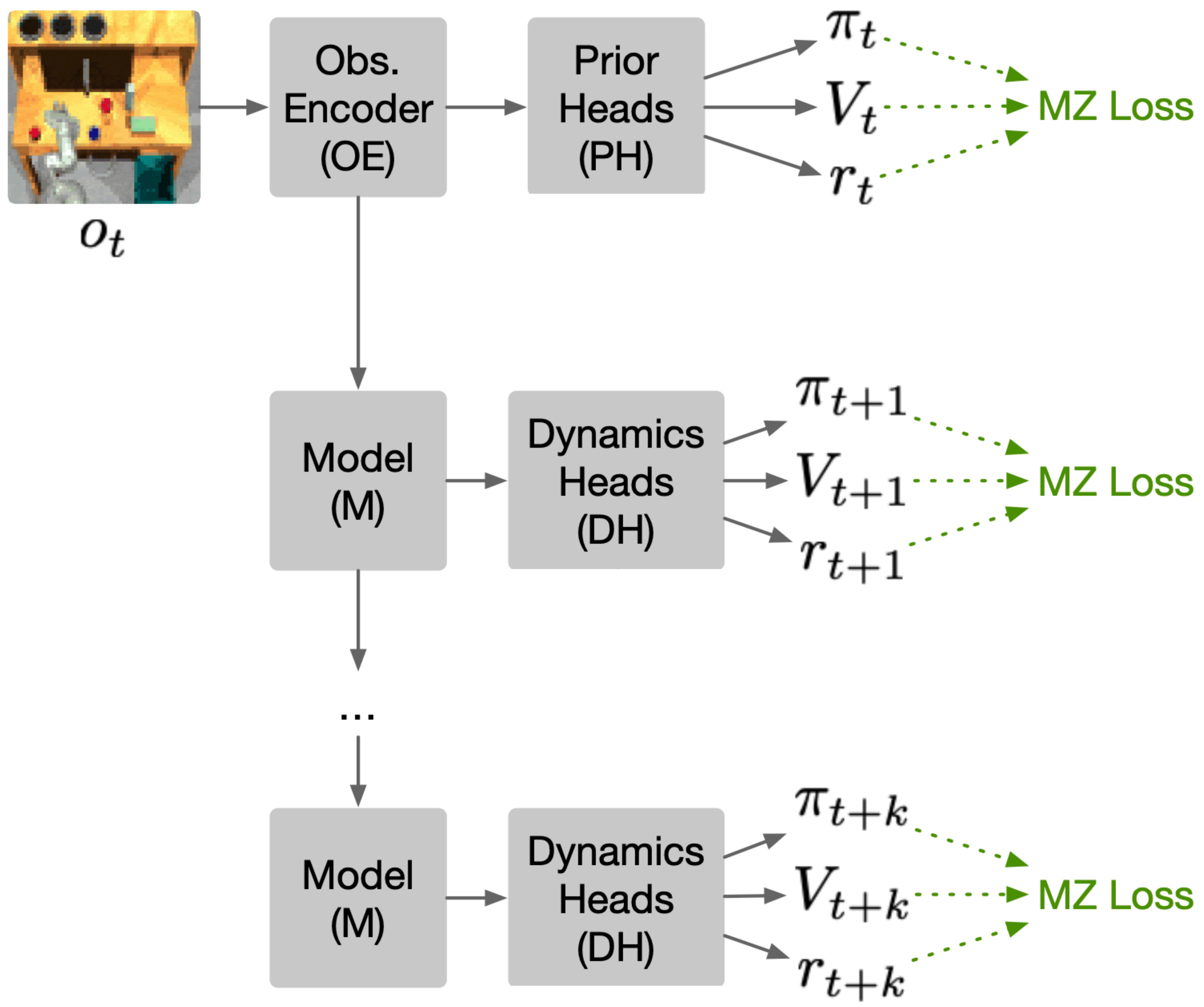


Train on a **procedurally-generated** distribution of environments
Zero-shot generalization to unseen environments
(e.g. Procgen, Cobbe et al., 2020)

Failure of representation

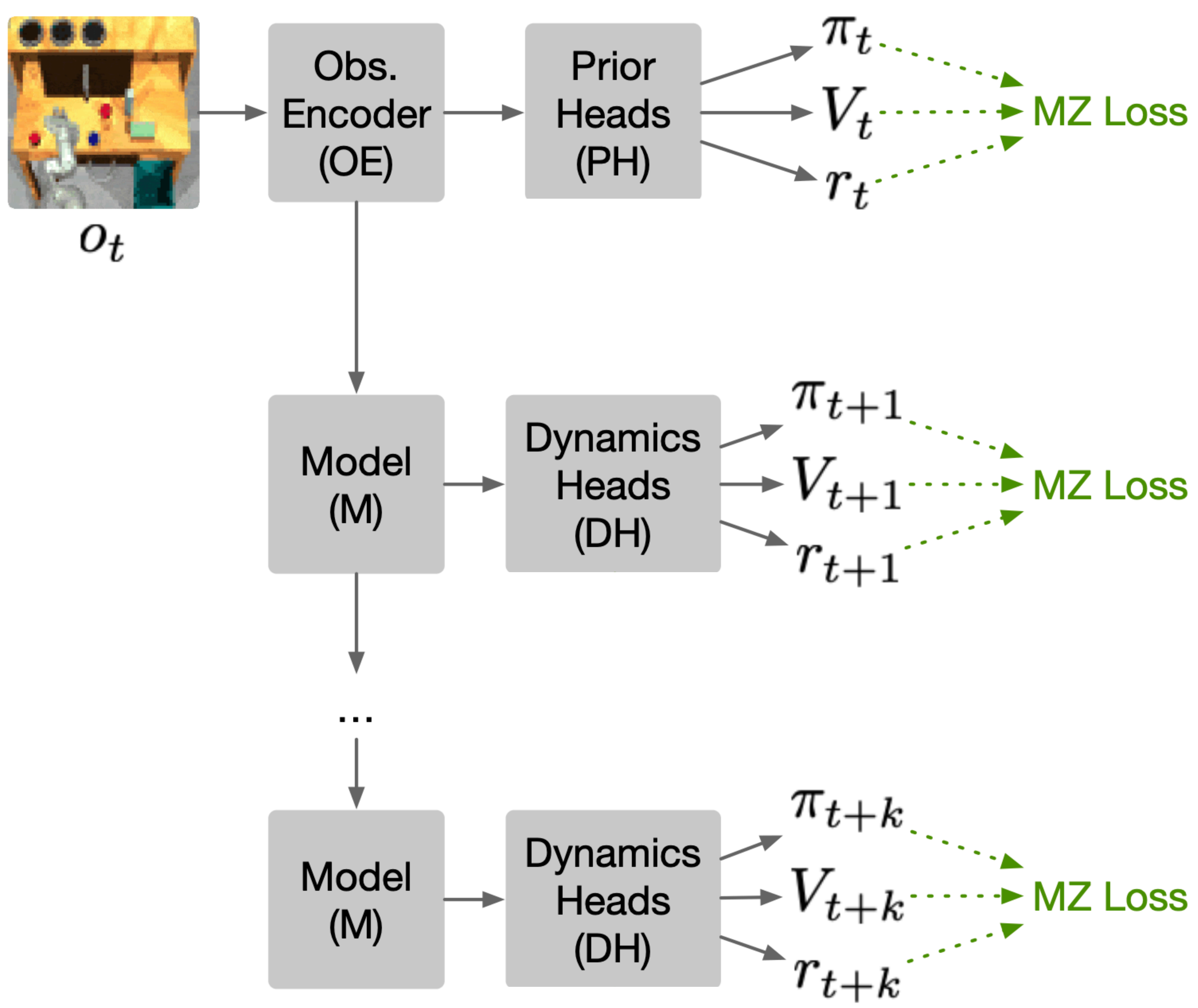


Improving MuZero with self-supervision



MZ loss: for *k=0...K*

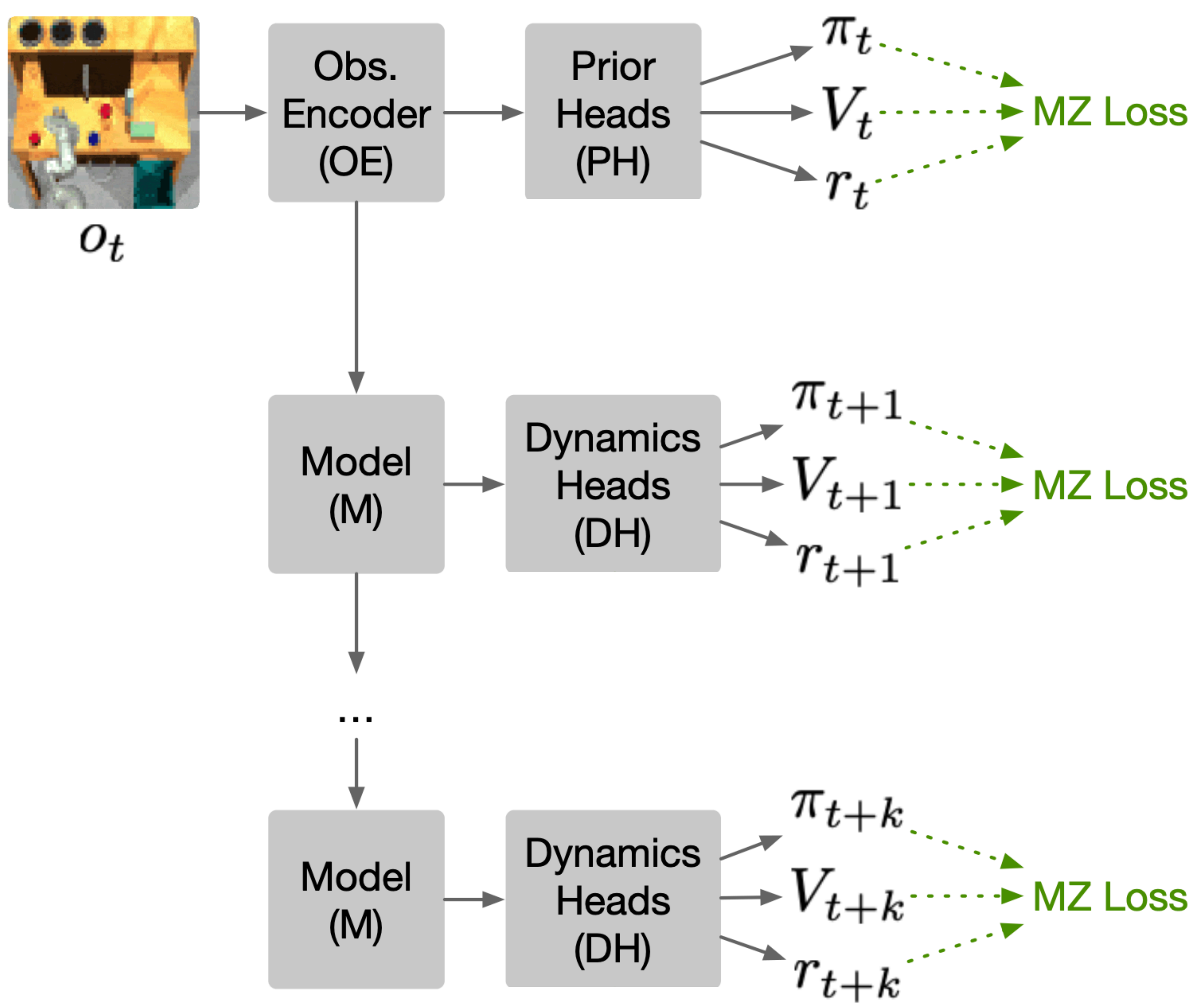
Improving MuZero with self-supervision



MZ loss: for *k=0...K*

- *Policy*: imitate the search policy at time t+k

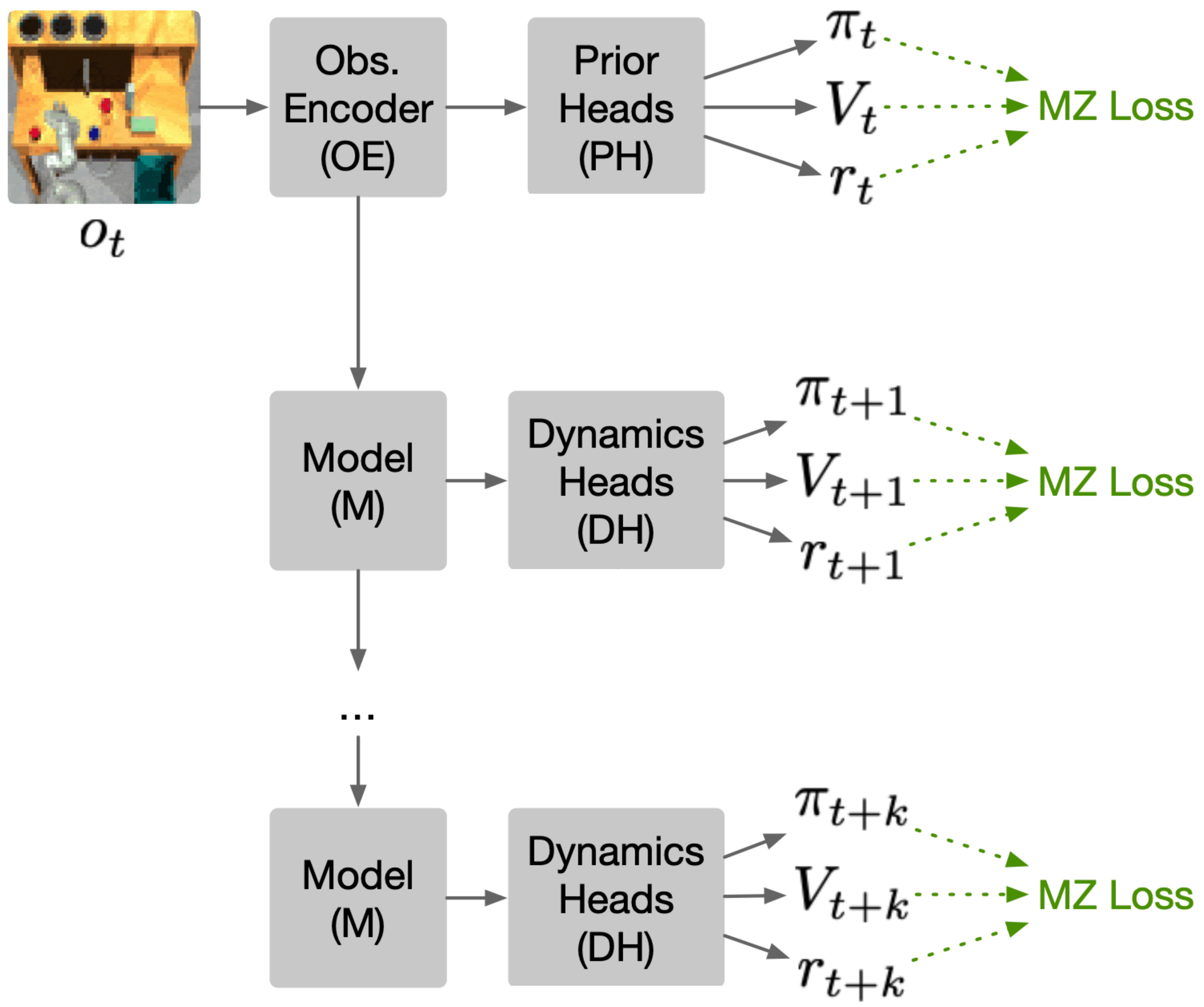
Improving MuZero with self-supervision



MZ loss: for *k=0...K*

- *Policy*: imitate the search policy at time t+k
- *Value*: predict n-step bootstrapped return, with bootstrapped values estimated via MCTS at time t+k+n

Improving MuZero with self-supervision

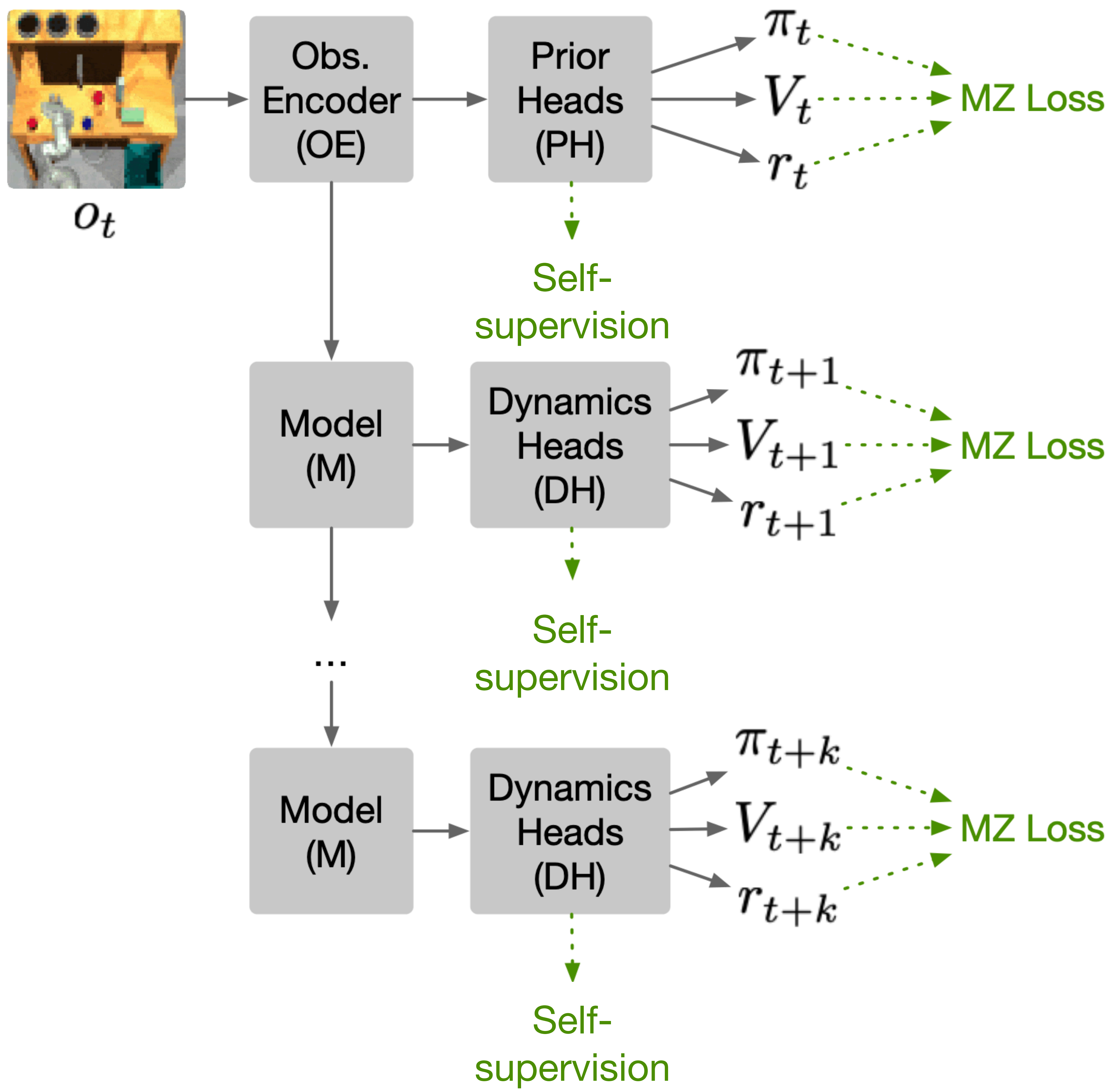


MZ loss: for *k=0...K*

- *Policy*: imitate the search policy at time t+k
- *Value*: predict n-step bootstrapped return, with bootstrapped values estimated via MCTS at time t+k+n
- *Reward*: observed environment reward at time t+k

Jessica Hamrick - jhamrick@deepmind.com

Improving MuZero with self-supervision



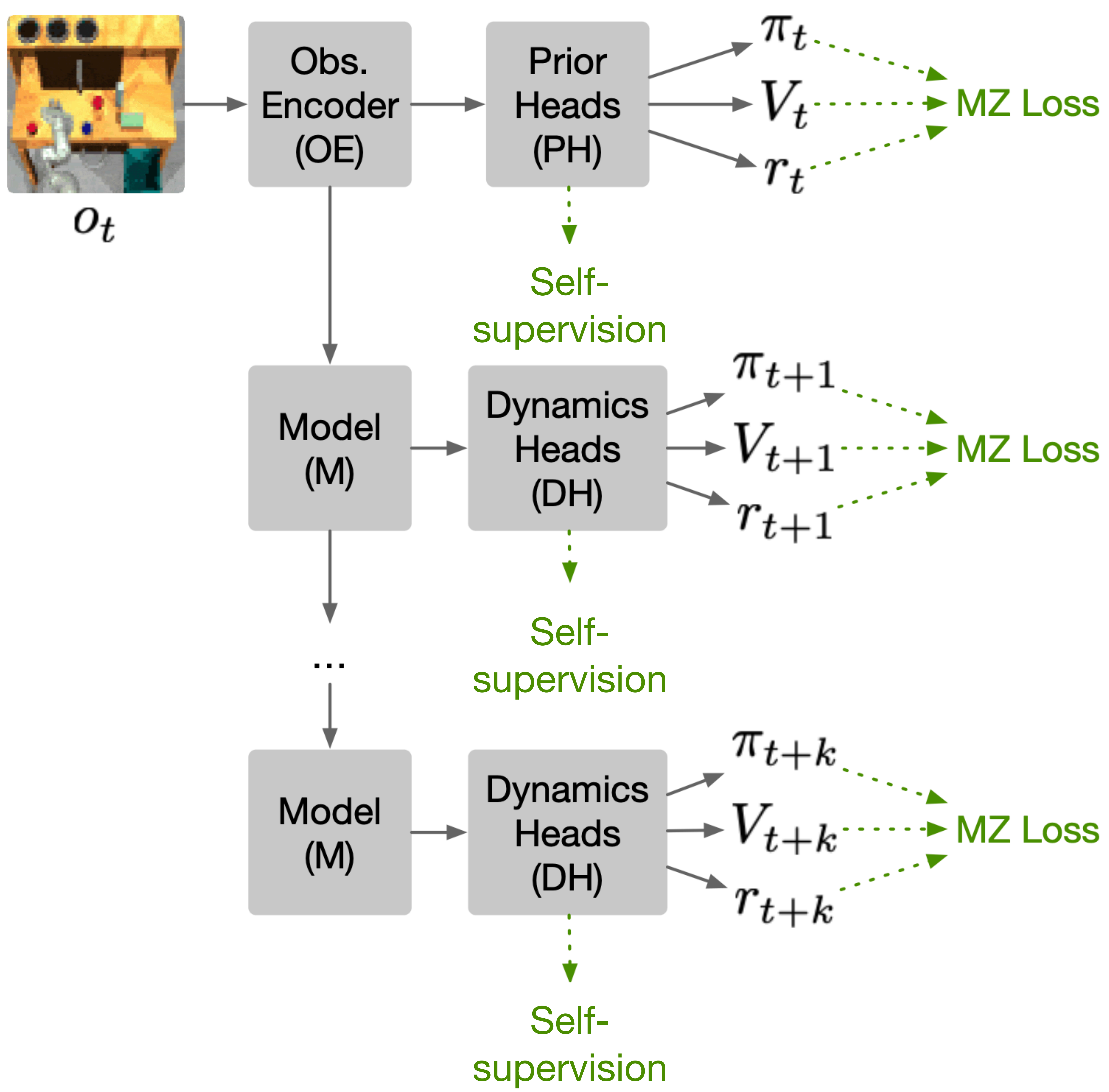
MZ loss: for *k=0...K*

- *Policy*: imitate the search policy at time t+k
- *Value*: predict n-step bootstrapped return, with bootstrapped values estimated via MCTS at time t+k+n
- *Reward*: observed environment reward at time t+k

Self-supervised losses:

Jessica Hamrick - jhamrick@deepmind.com

Improving MuZero with self-supervision



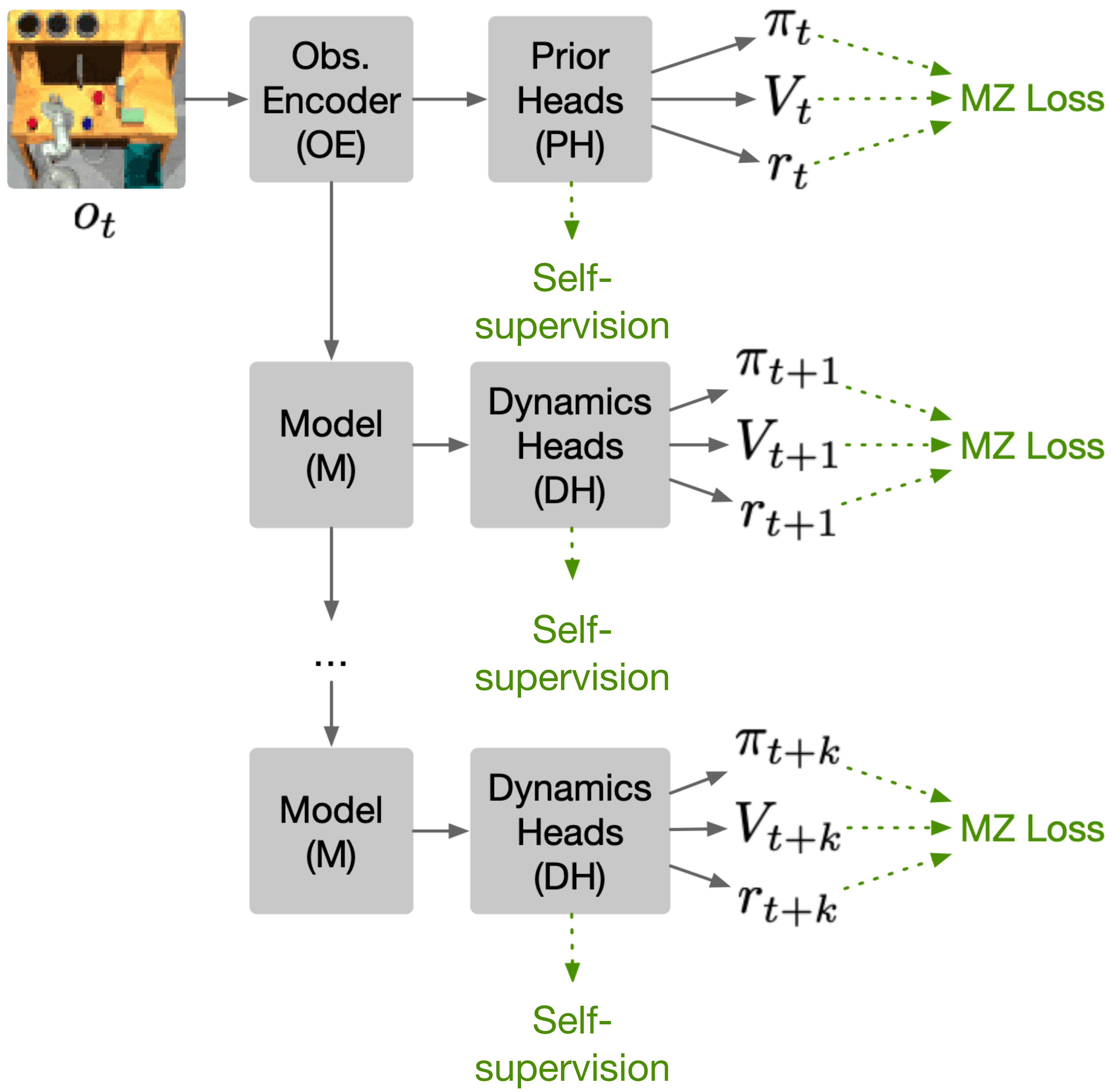
MZ loss: for *k=0...K*

- *Policy*: imitate the search policy at time t+k
- *Value*: predict n-step bootstrapped return, with bootstrapped values estimated via MCTS at time t+k+n
- *Reward*: observed environment reward at time t+k

Self-supervised losses:

- *Reconstruction*: predict the obs. at time *t+k*

Improving MuZero with self-supervision



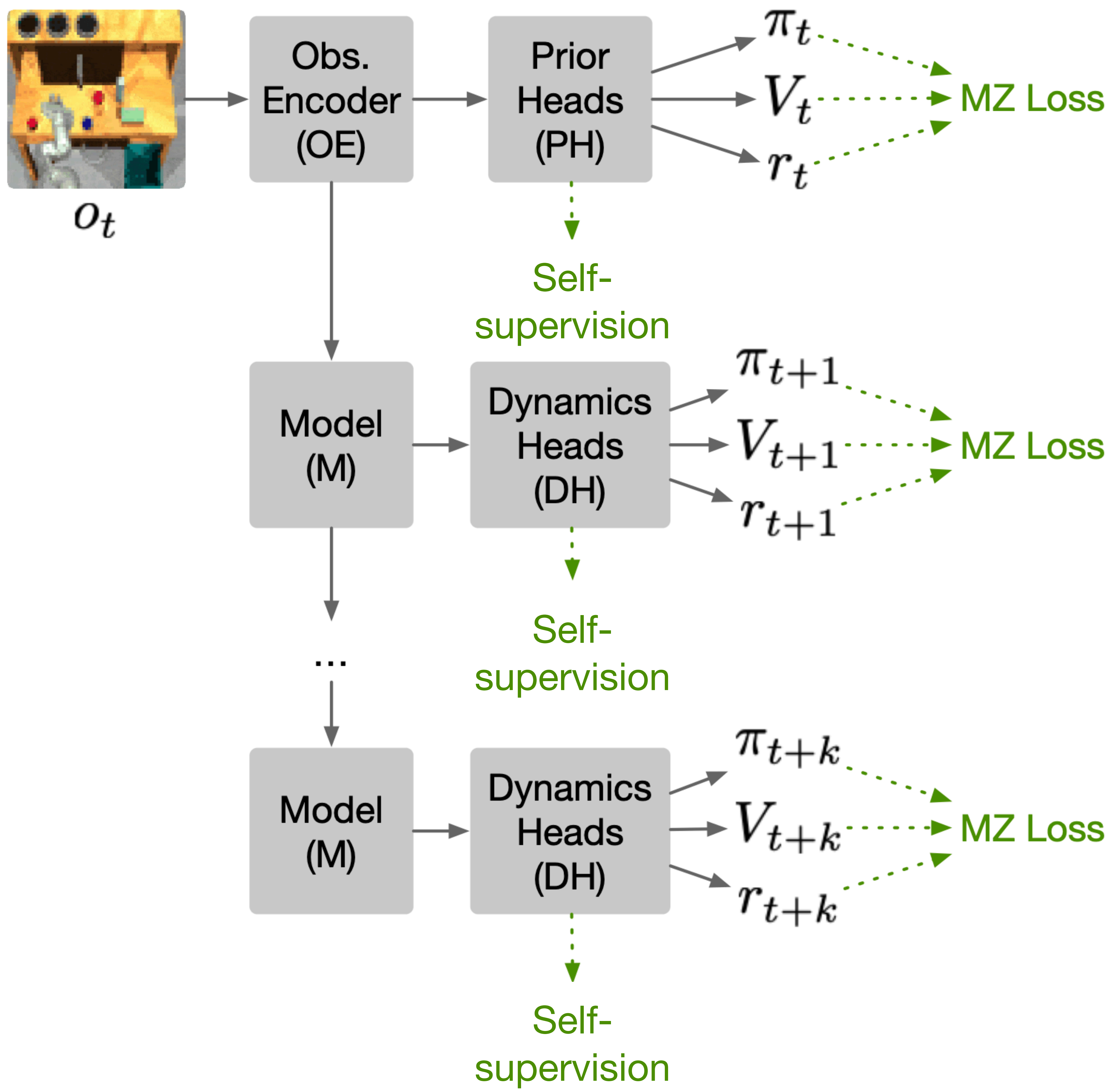
MZ loss: for *k=0...K*

- *Policy*: imitate the search policy at time t+k
- *Value*: predict n-step bootstrapped return, with bootstrapped values estimated via MCTS at time t+k+n
- *Reward*: observed environment reward at time t+k

Self-supervised losses:

- *Reconstruction*: predict the obs. at time *t+k*
- *SPR:* predict the obs. embedding at time *t+k*

Improving MuZero with self-supervision



MZ loss: for *k=0...K*

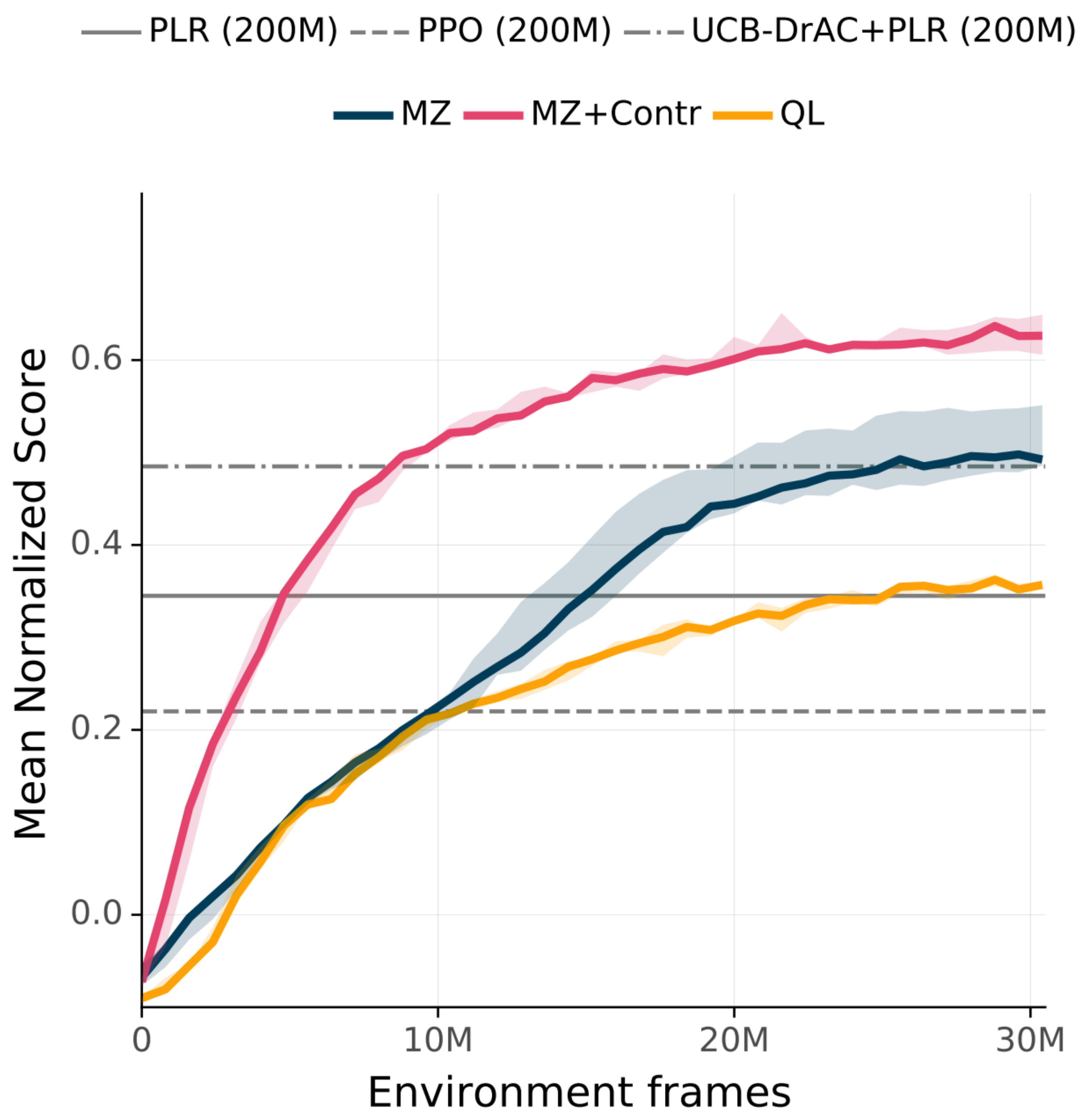
- *Policy*: imitate the search policy at time t+k
- *Value*: predict n-step bootstrapped return, with bootstrapped values estimated via MCTS at time t+k+n
- *Reward*: observed environment reward at time t+k

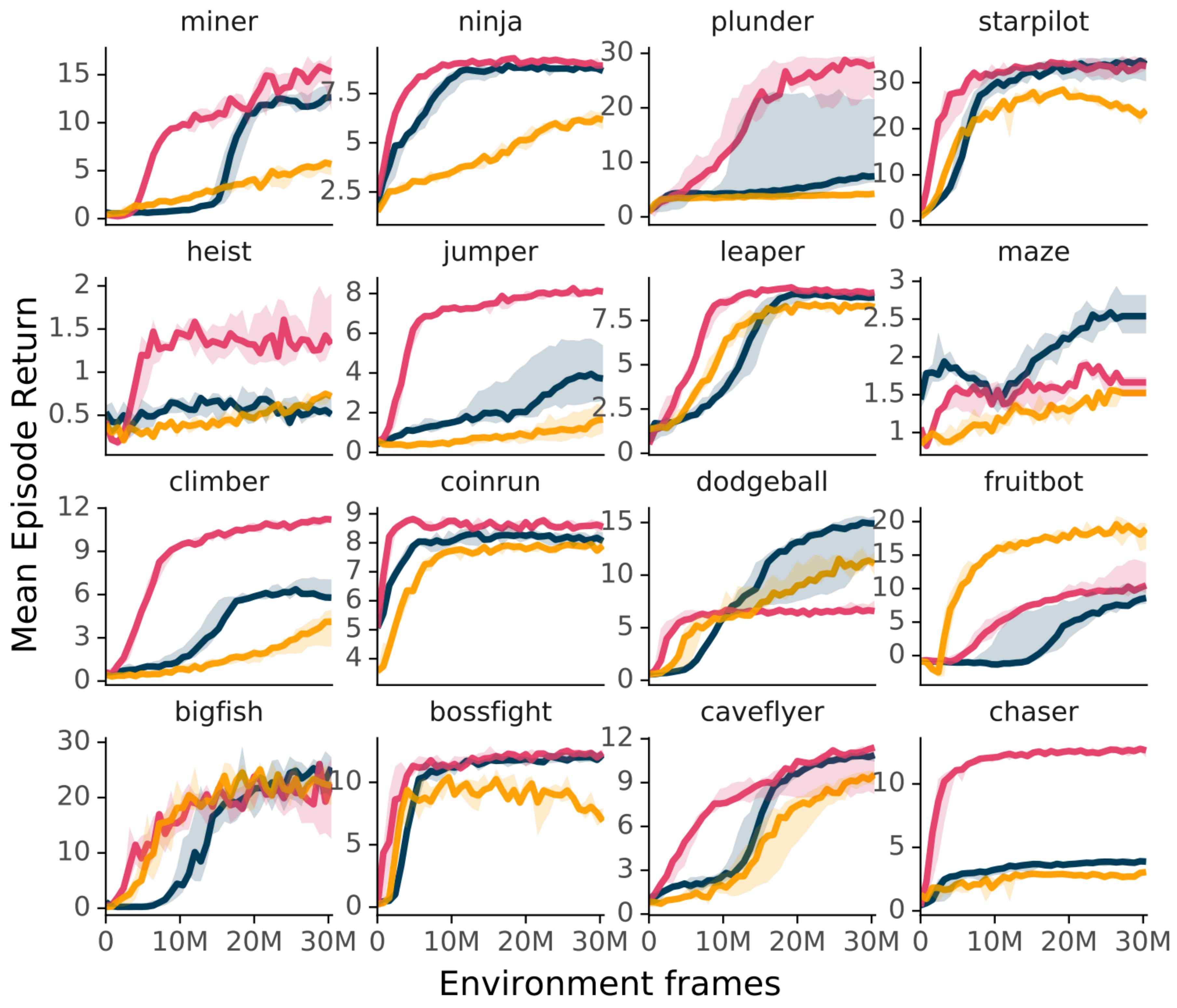
Self-supervised losses:

- *Reconstruction*: predict the obs. at time *t+k*
- *SPR:* predict the obs. embedding at time *t+k*
- *Contrastive:* classify whether a predicted obs. embedding at time *t+k* corresponds to the observation at time *t+i*

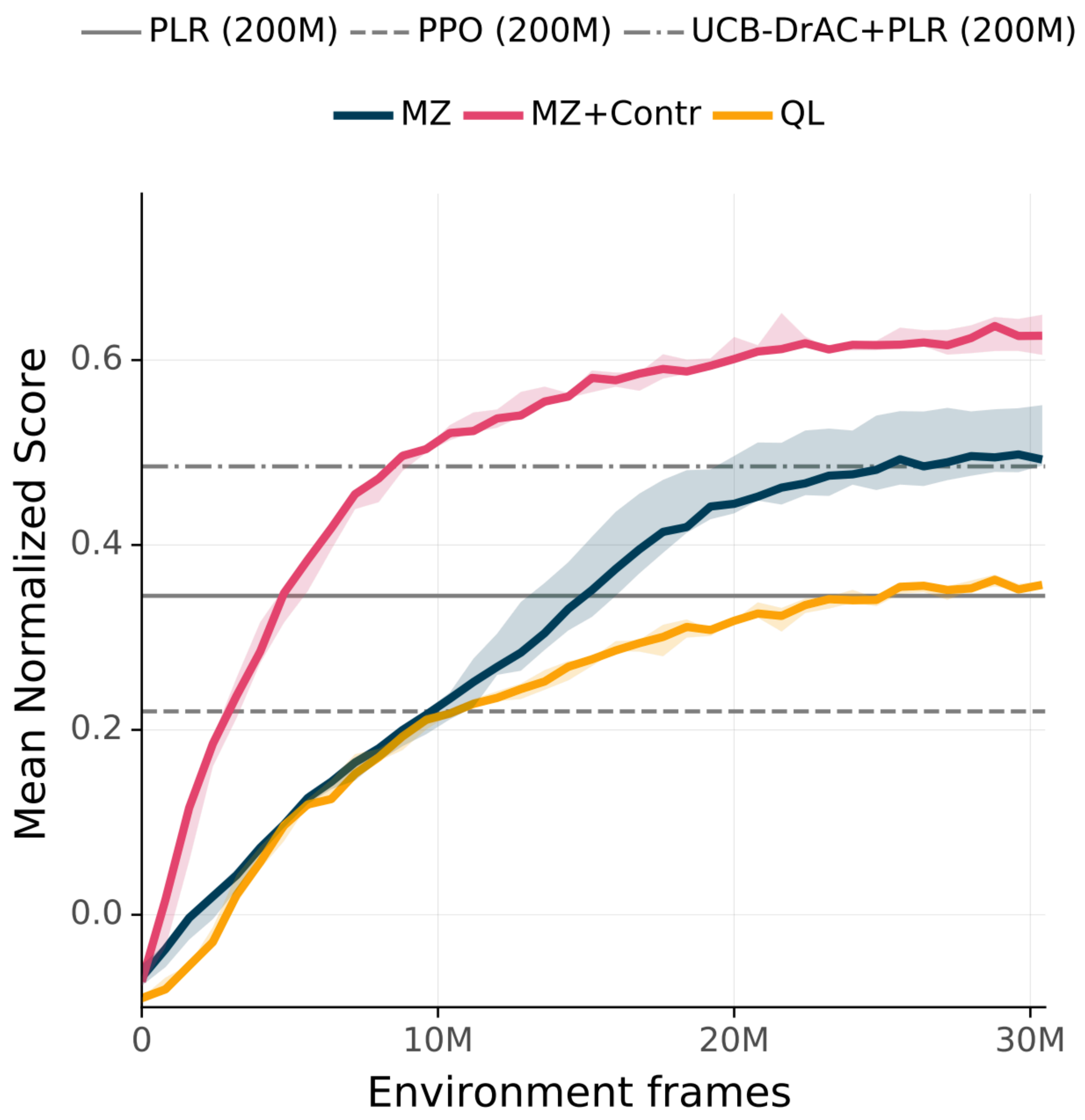
Jessica Hamrick - jhamrick@deepmind.com

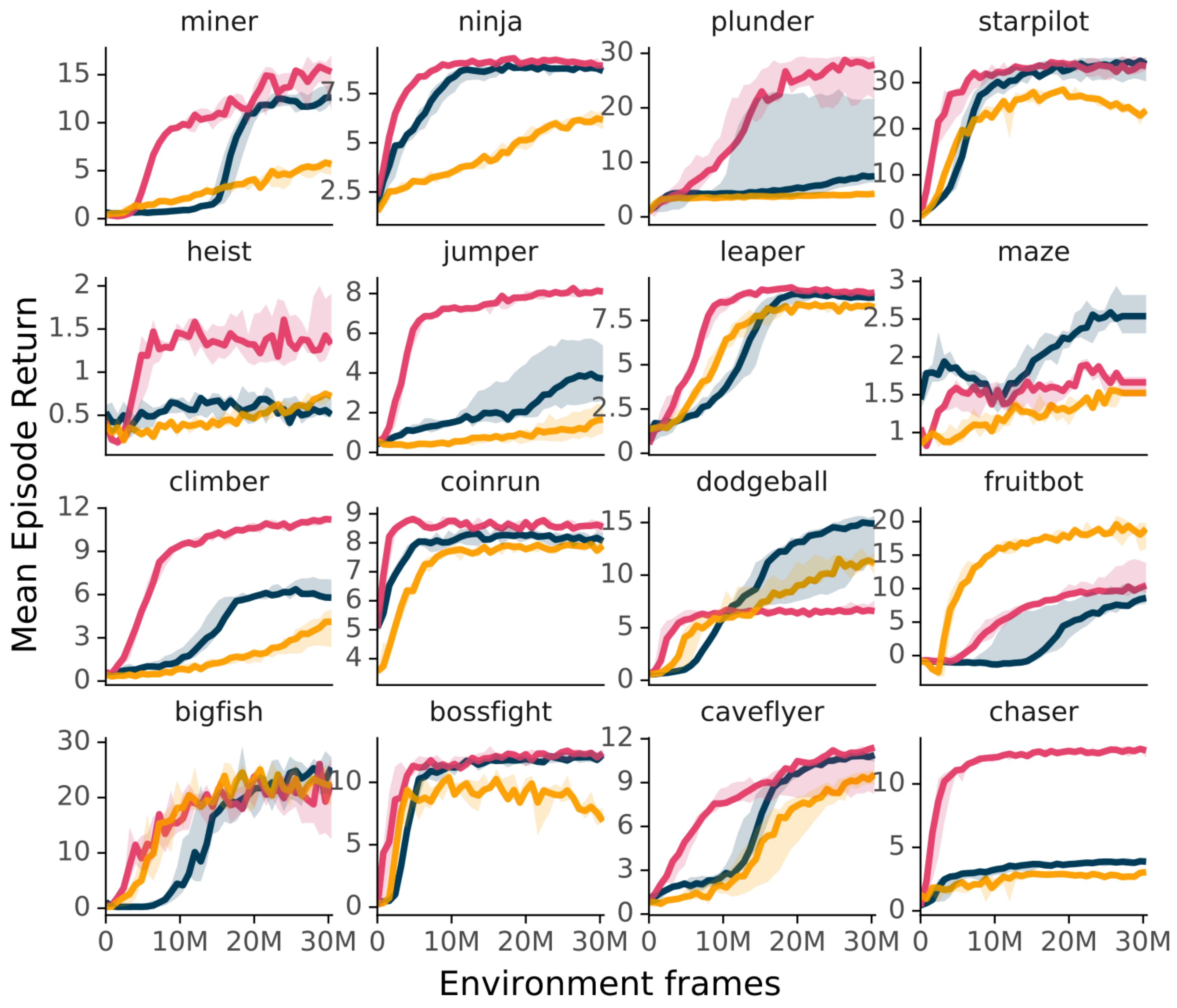
Procgen results (500 levels)





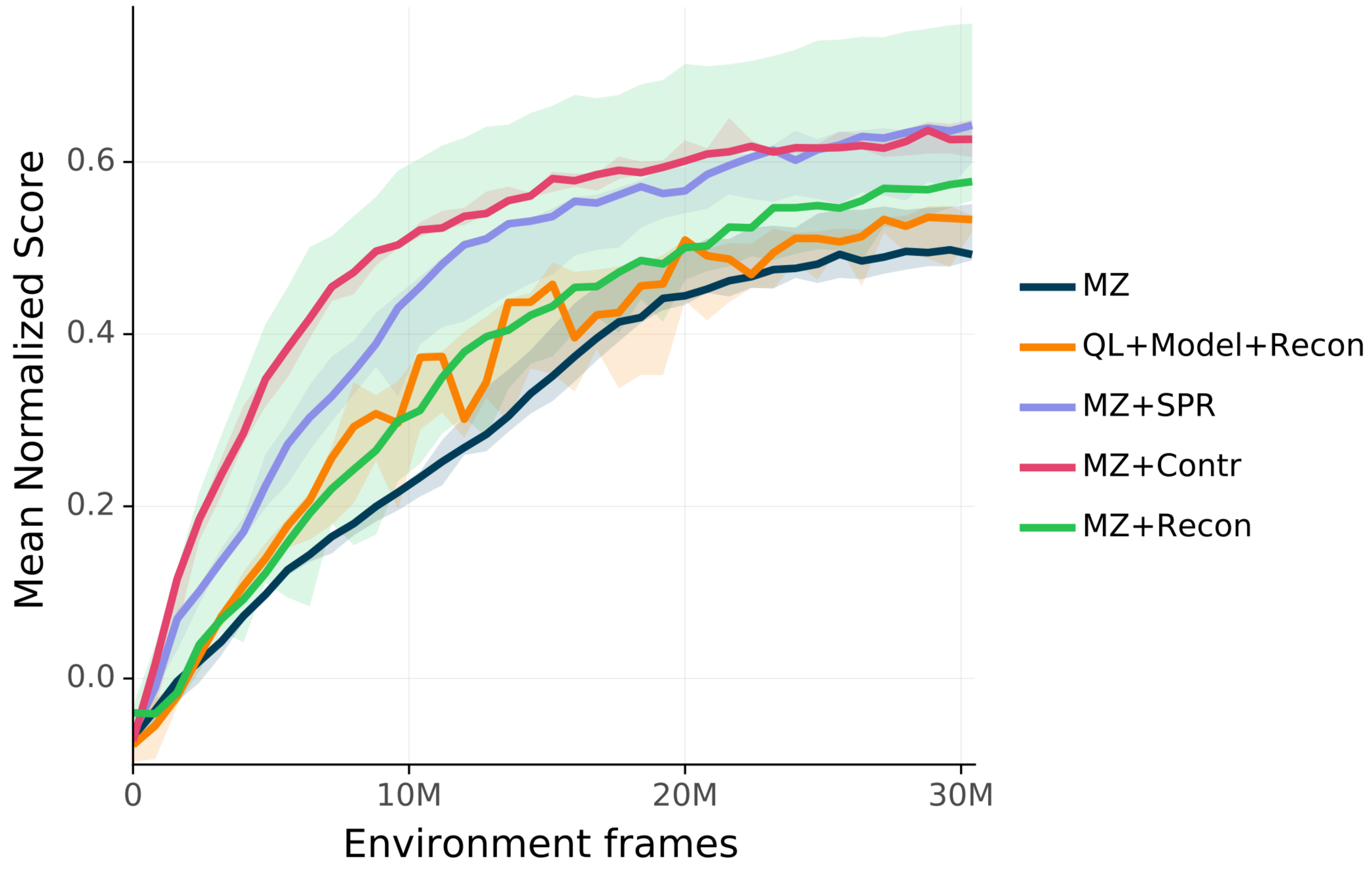
Procgen results (500 levels)





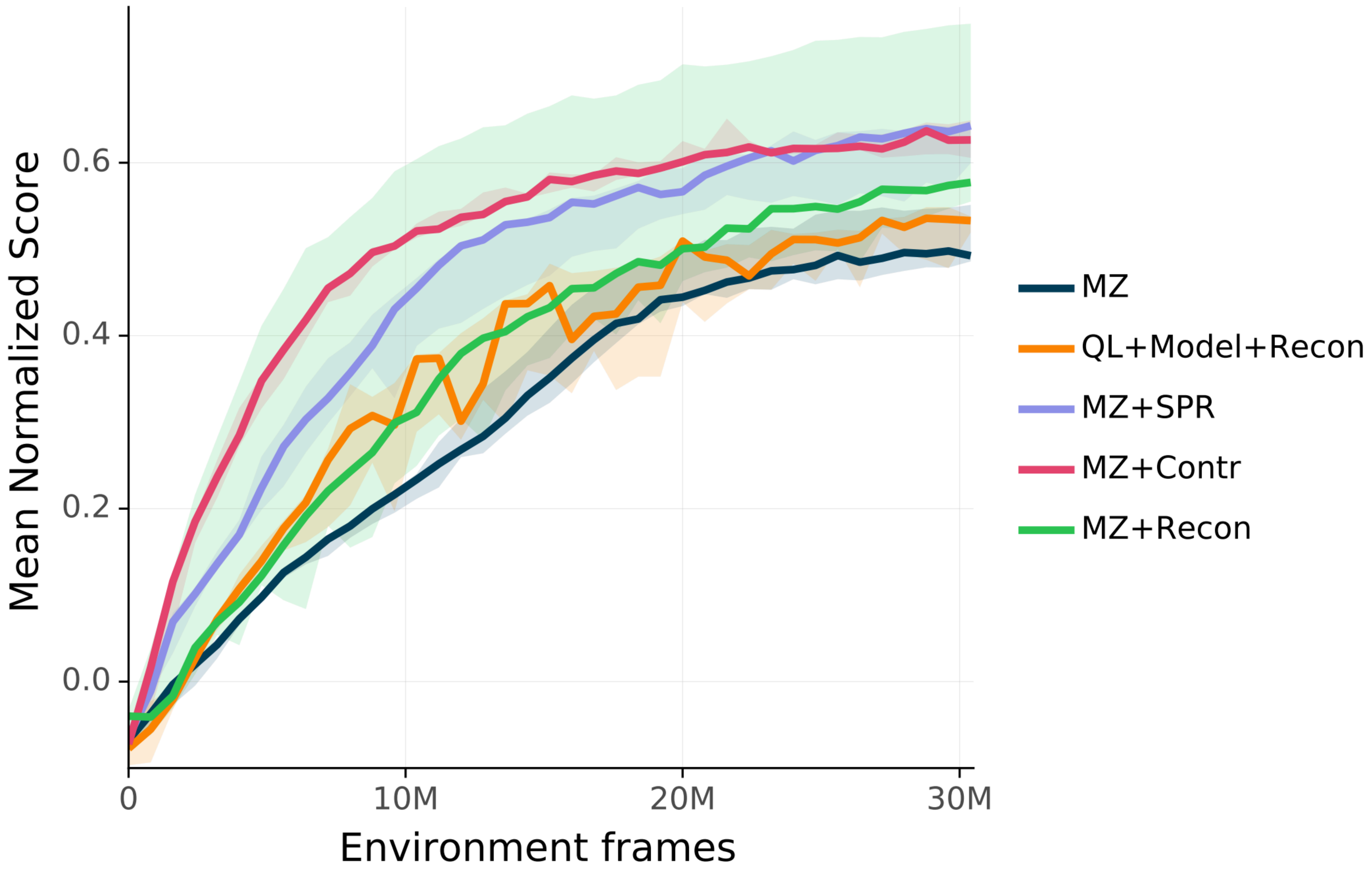
→ Self-supervision has a huge impact on generalization!

Comparing methods of self-supervision



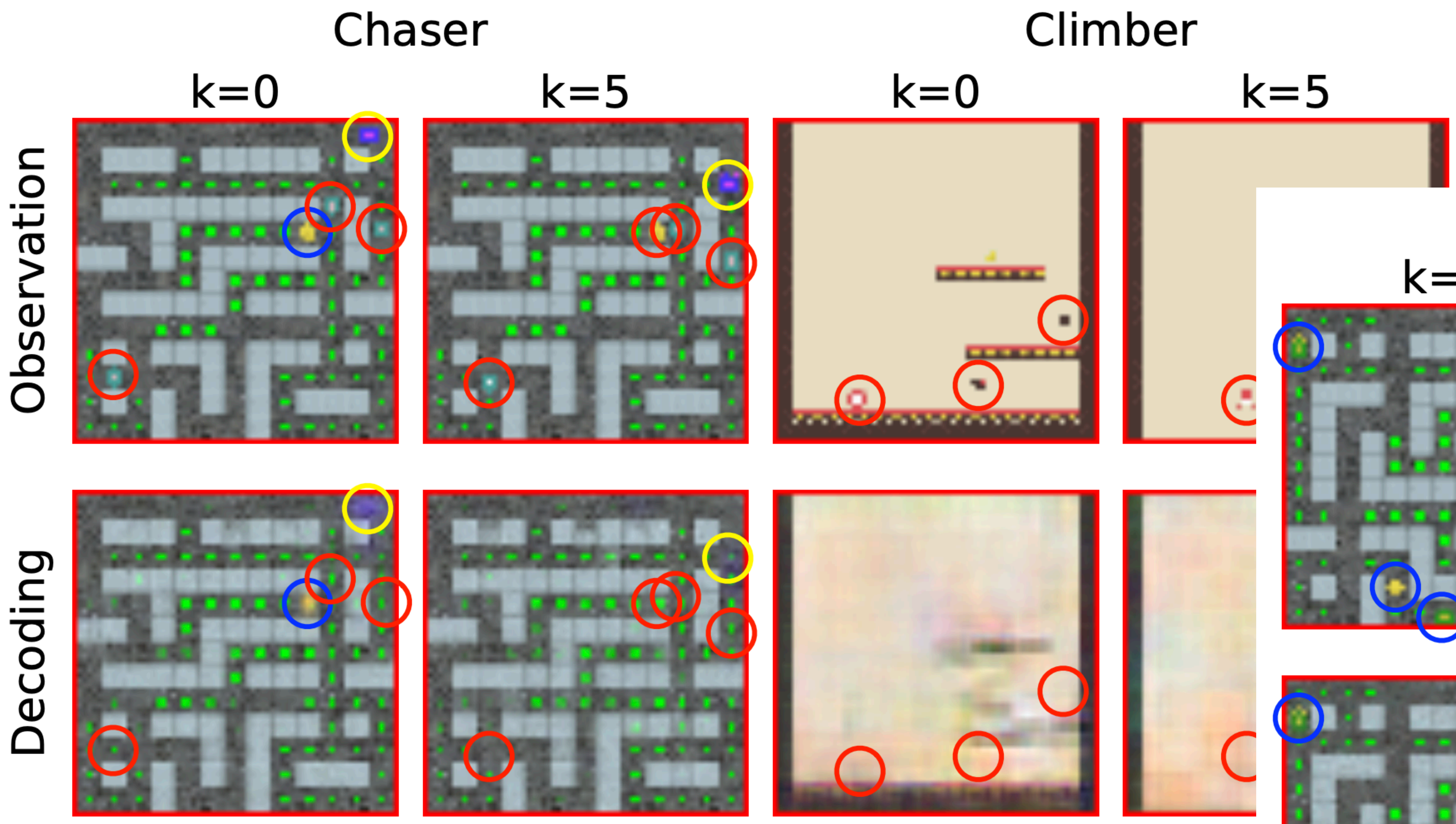
Jessica Hamrick - jhamrick@deepmind.com

Comparing methods of self-supervision

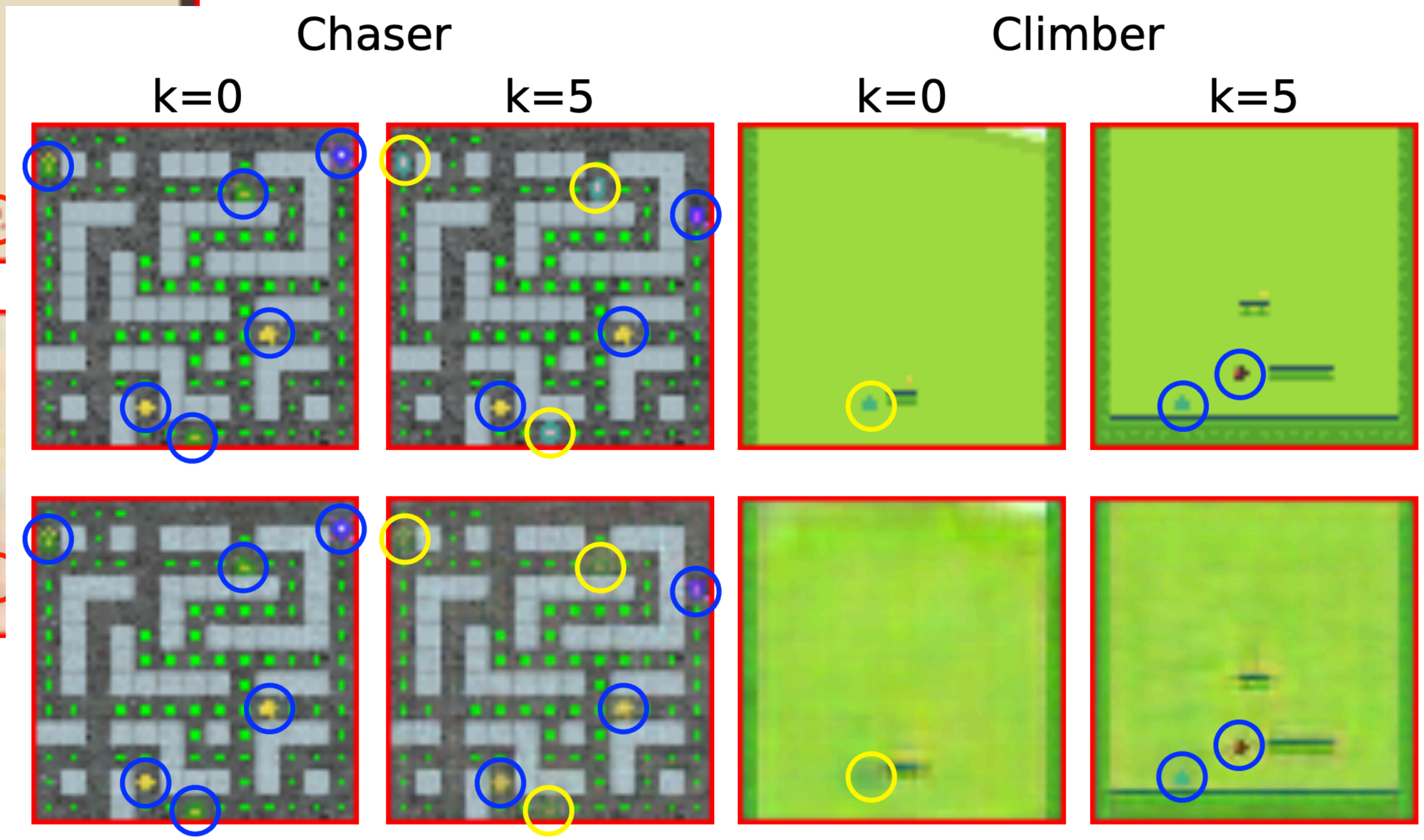


→ All methods of self-supervision are roughly comparable

Improved representations

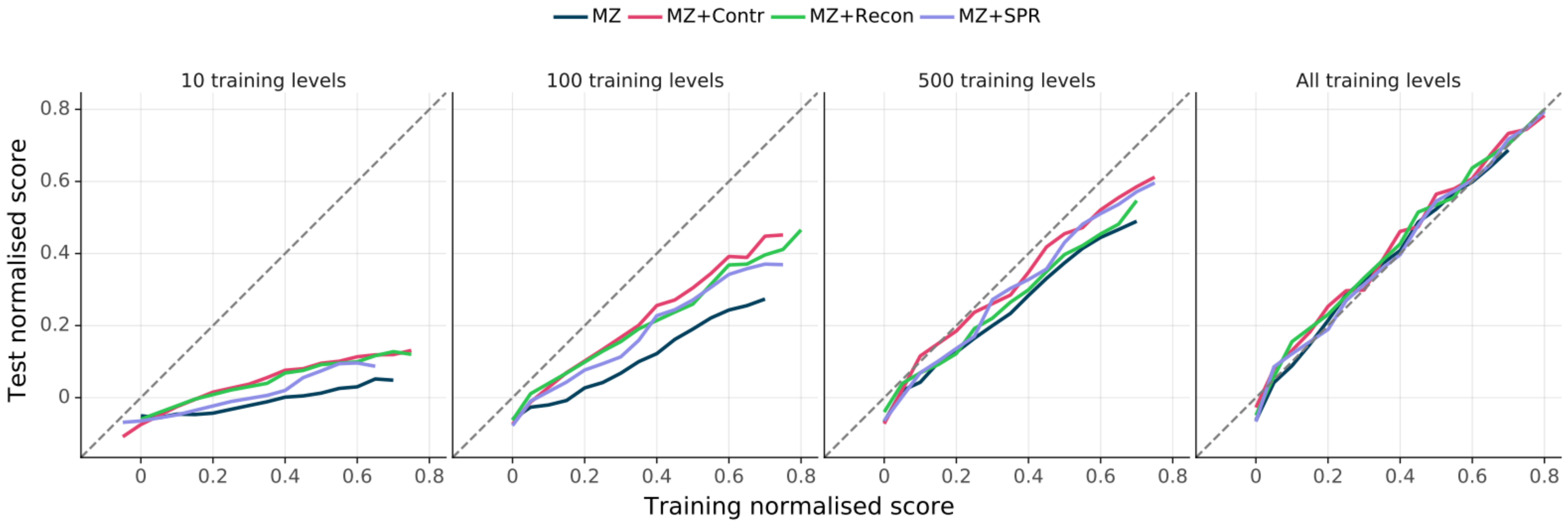


MuZero



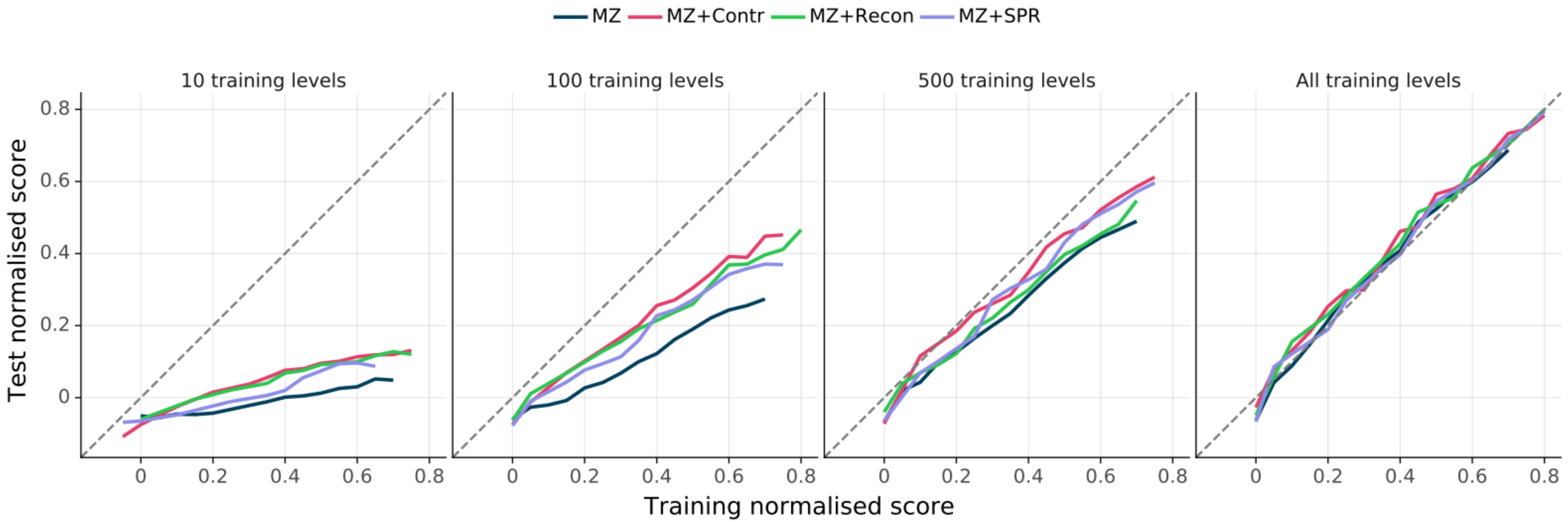
MuZero + Reconstruction

Self-supervision improves generalization



Jessica Hamrick - jhamrick@deepmind.com

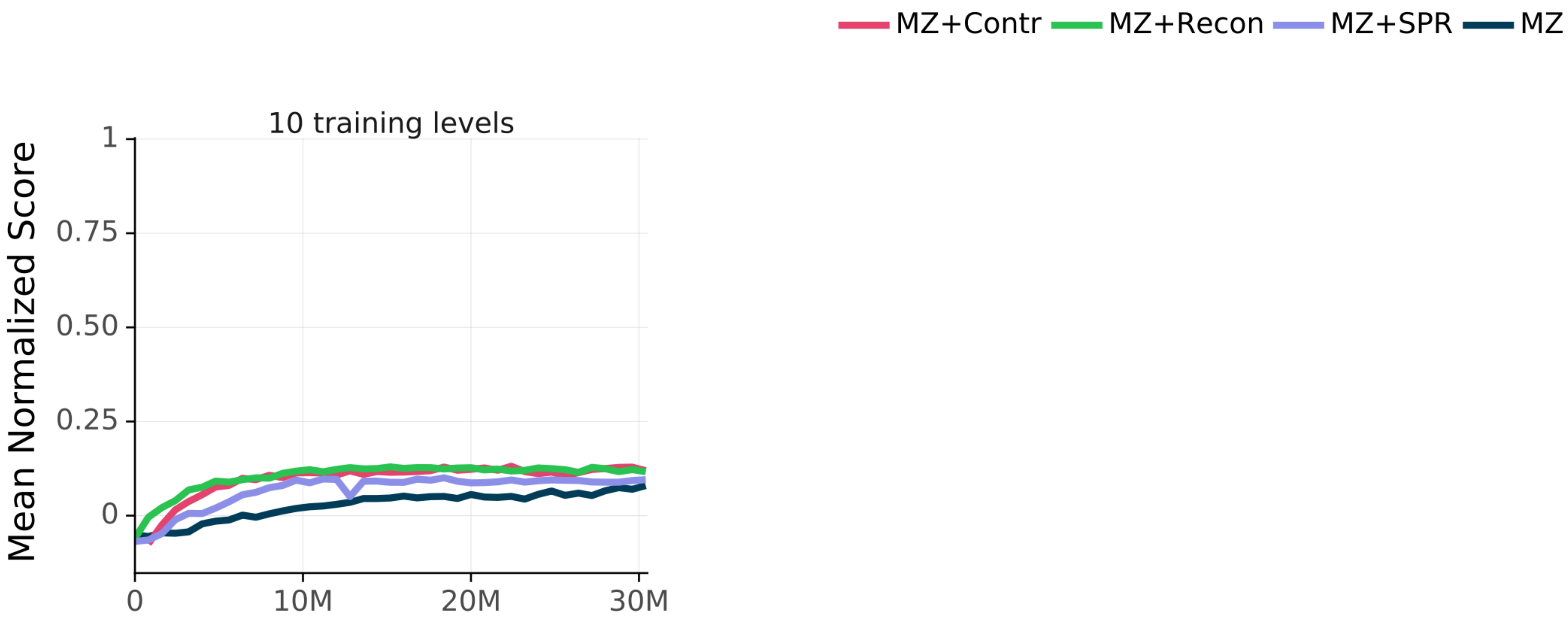
Self-supervision improves generalization



→ Self-supervision improves generalization *even when controlling for training performance*

Jessica Hamrick - jhamrick@deepmind.com

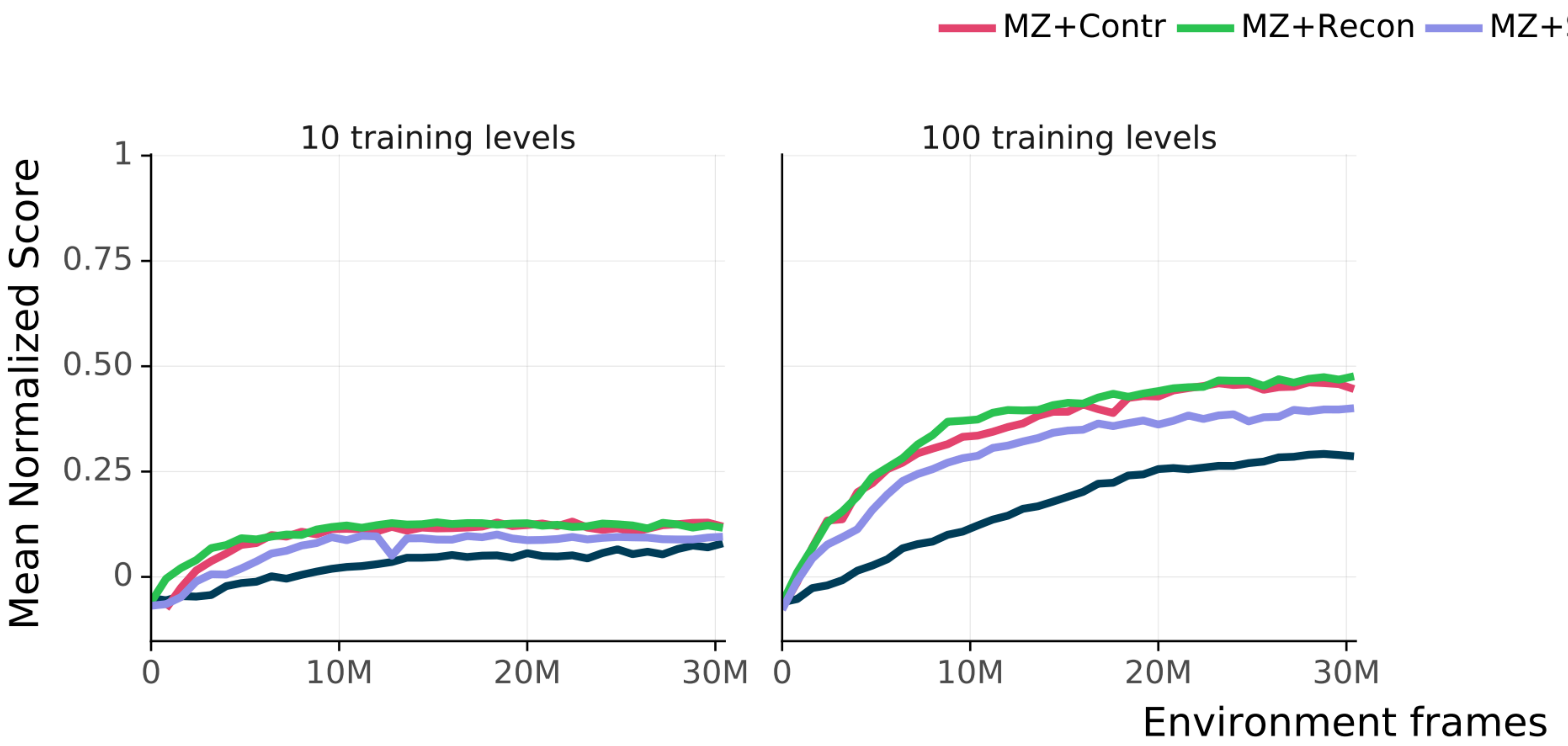
Interaction between self-supervision and dataset size



very little improvement w/ self-supervision

Jessica Hamrick - jhamrick@deepmind.com

Interaction between self-supervision and dataset size

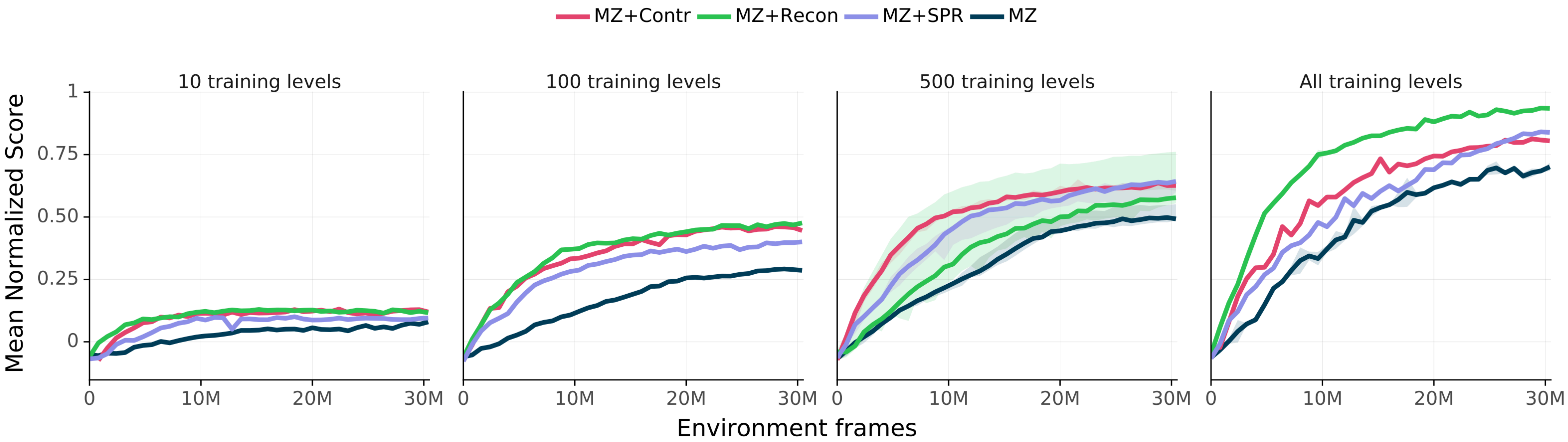


very little improvement w/ self-supervision

Interaction between self-supervision and dataset size

very little improvement w/ self-supervision

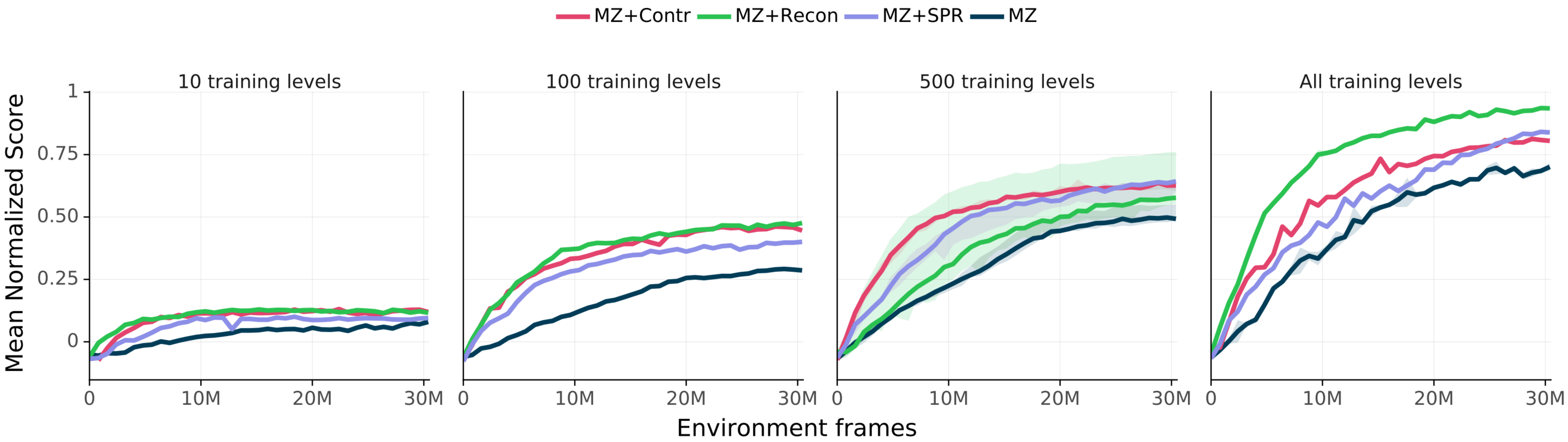
Interaction between self-supervision and dataset size



very little improvement w/ self-supervision

big improvement w/ self-supervision

Interaction between self-supervision and dataset size



very little improvement w/ self-supervision

→ Self-supervision is more useful when training on more environments

big improvement w/ self-supervision

Jessica Hamrick - jhamrick@deepmind.com

Interim Takeaway #3: Generalization requires good representations, which can be improved through any method of self-supervision.

Interim Takeaway #4: Self-supervision interacts positively with the number of environments. We should be wary of drawing conclusions from single-task settings!

Outline



- **Understanding MBRL**
  Hamrick et al. (2021). On the role of planning in model based reinforcement learning. ICLR.

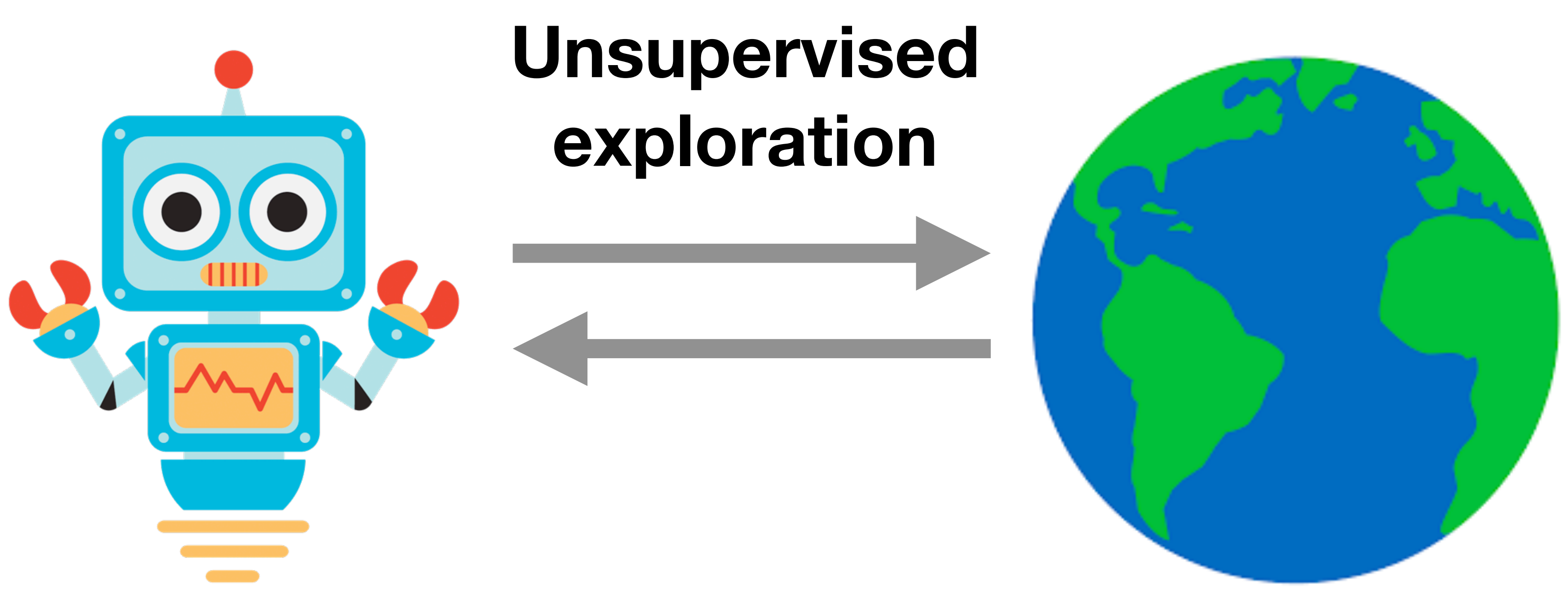
- **Understanding and improving generalization**
  Anand, Walker et al. (2022). Procedural generalization by planning with self-supervised world models. ICLR.

- **Understanding and improving transfer**
  Walker, Vértes, Li et al. (2023). Investigating the role of model-based learning in exploration and transfer. Under review.

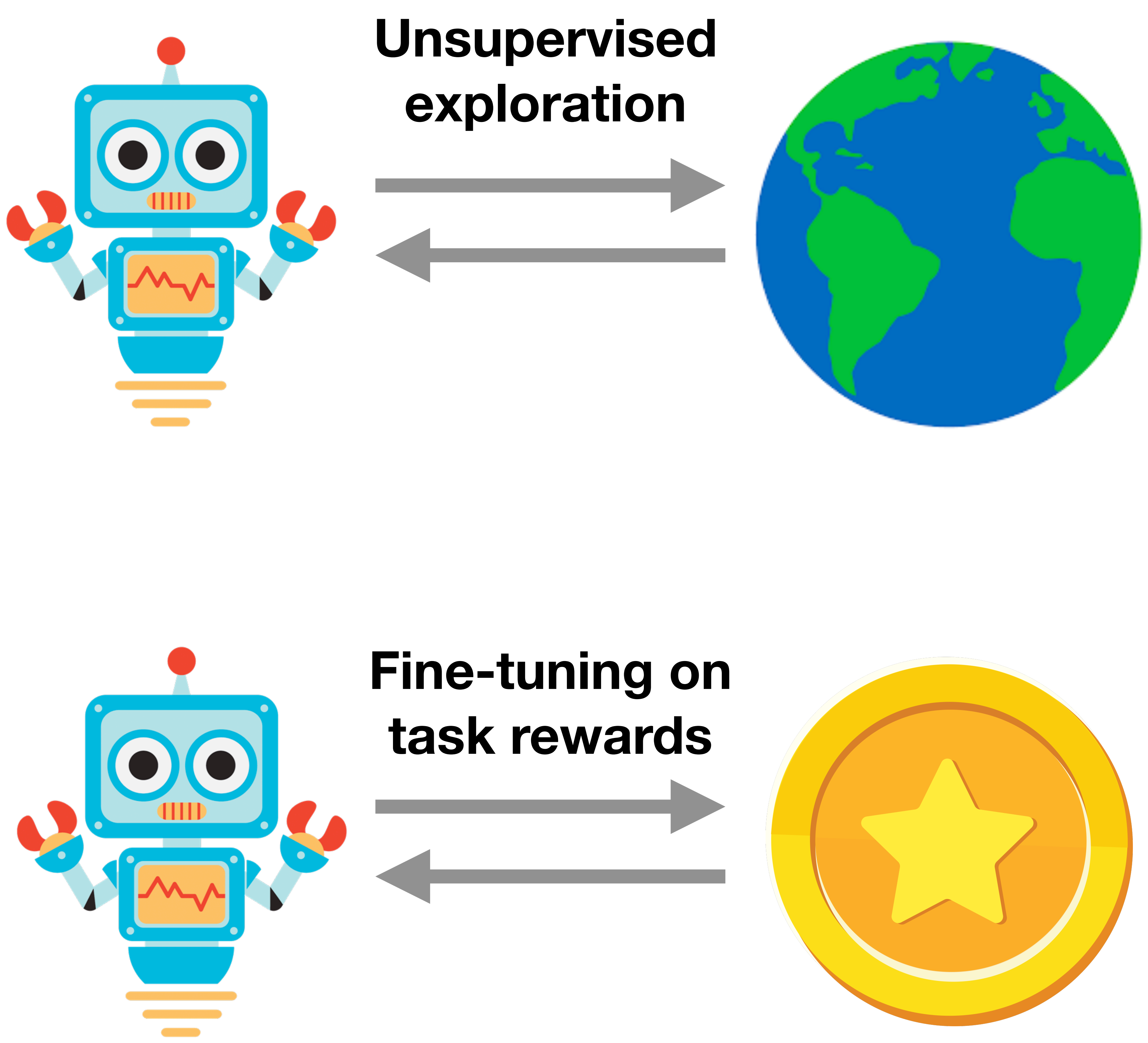
- **The future of MBRL**

Questions regarding transfer

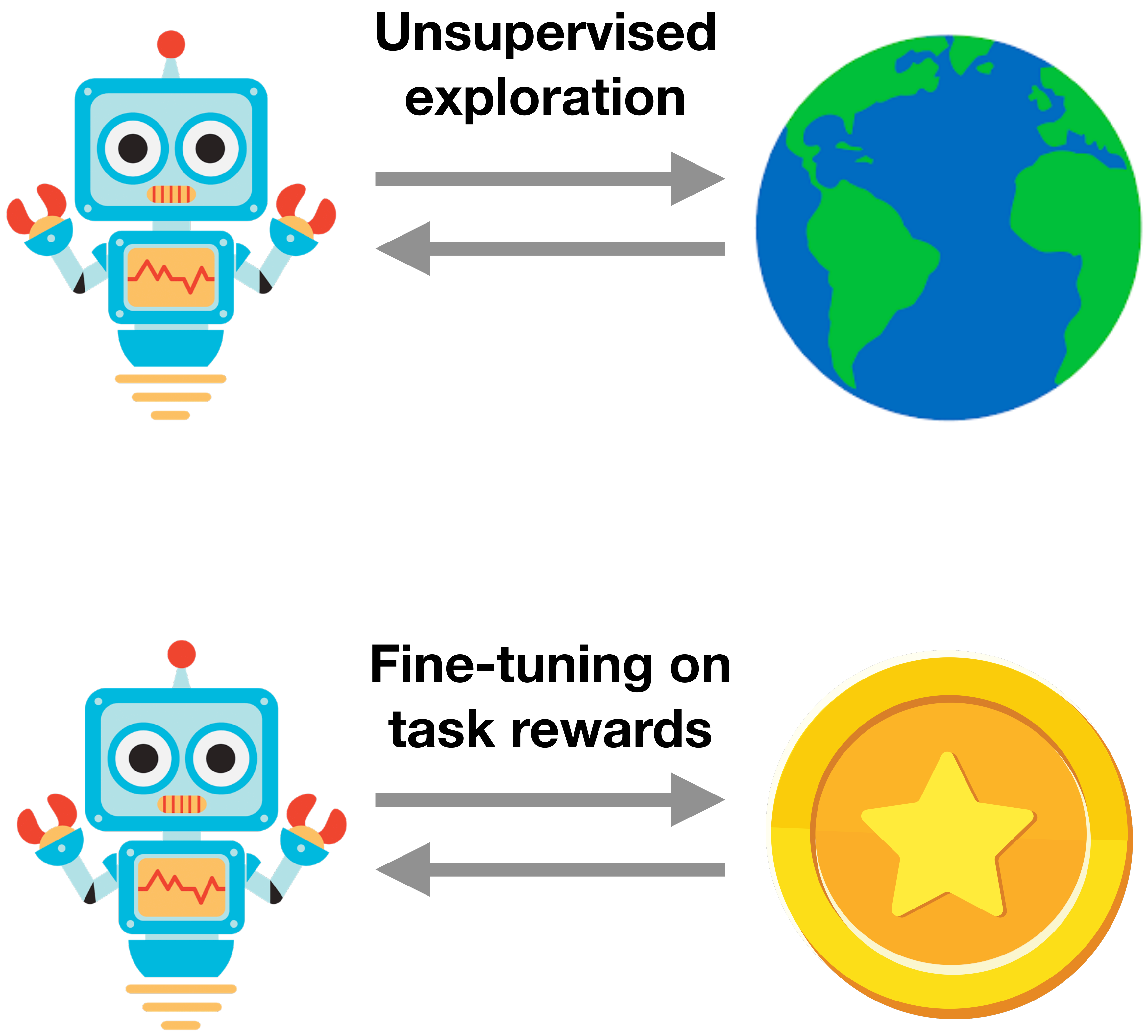
Questions regarding transfer



Questions regarding transfer

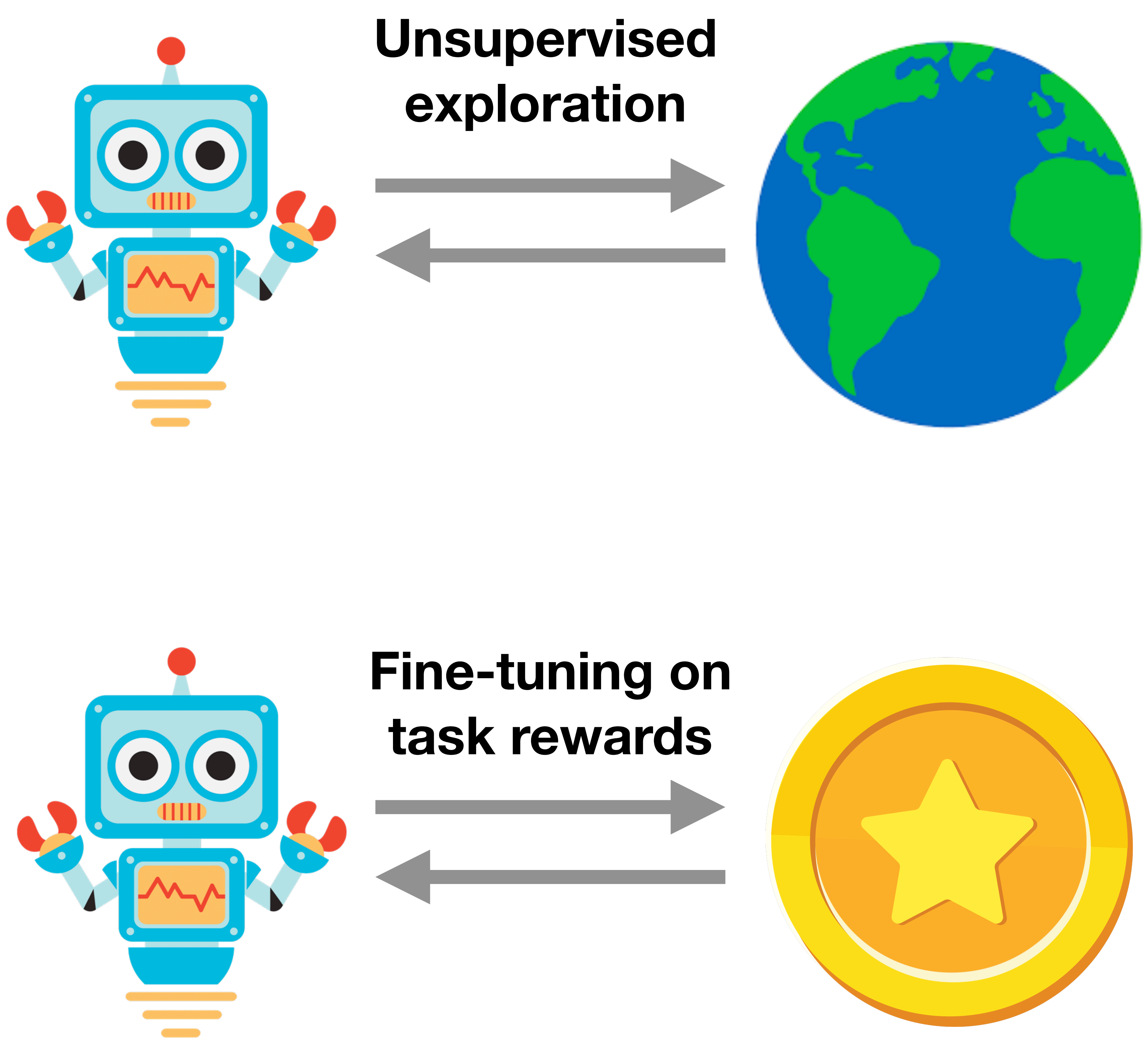


Questions regarding transfer



1. Is there an advantage to an agent being model-based during unsupervised exploration and/or fine-tuning?

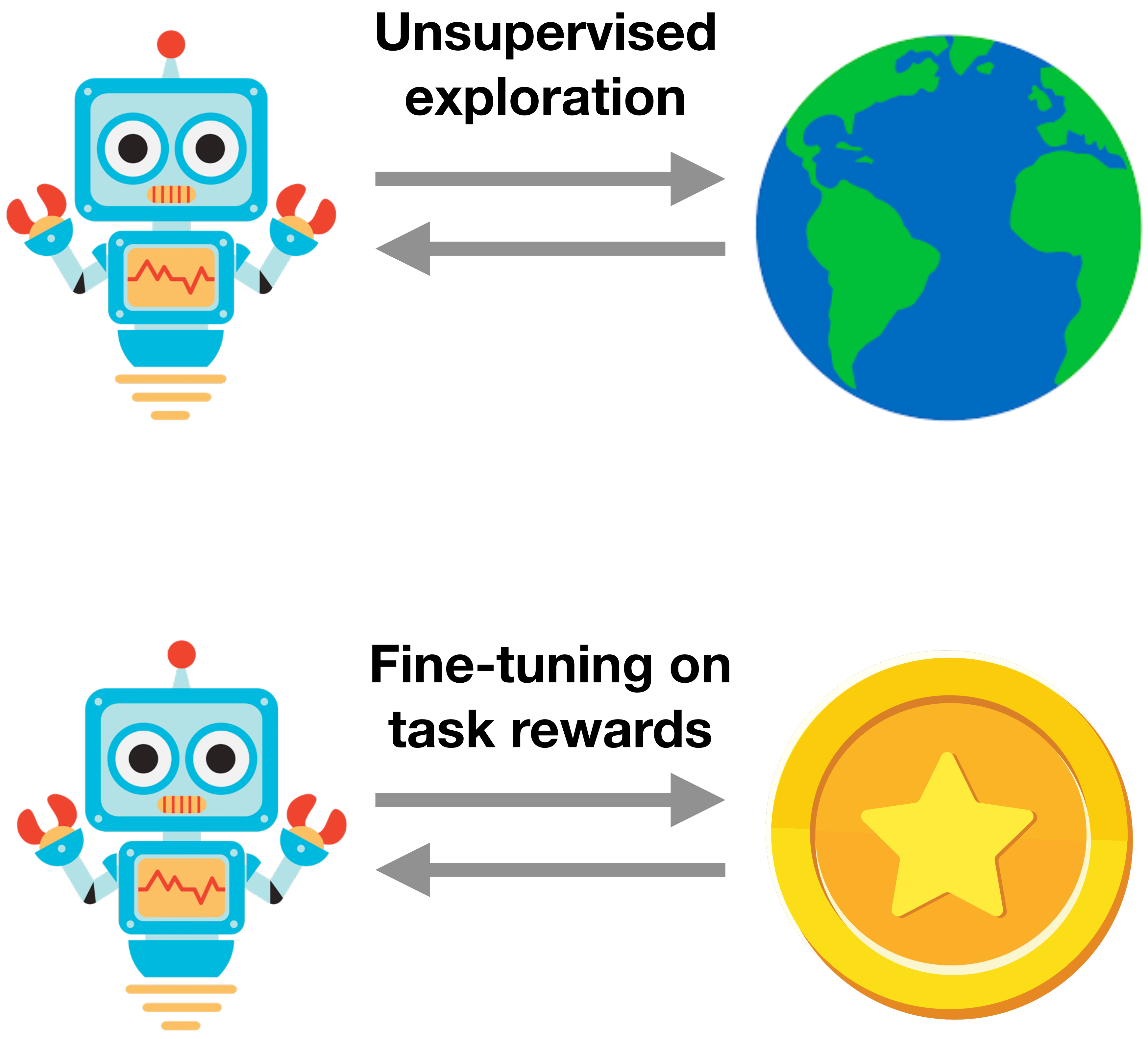
Questions regarding transfer



1. Is there an advantage to an agent being model-based during unsupervised exploration and/or fine-tuning?

2. What are the contributions of each component of a model-based agent for downstream task learning?

Questions regarding transfer



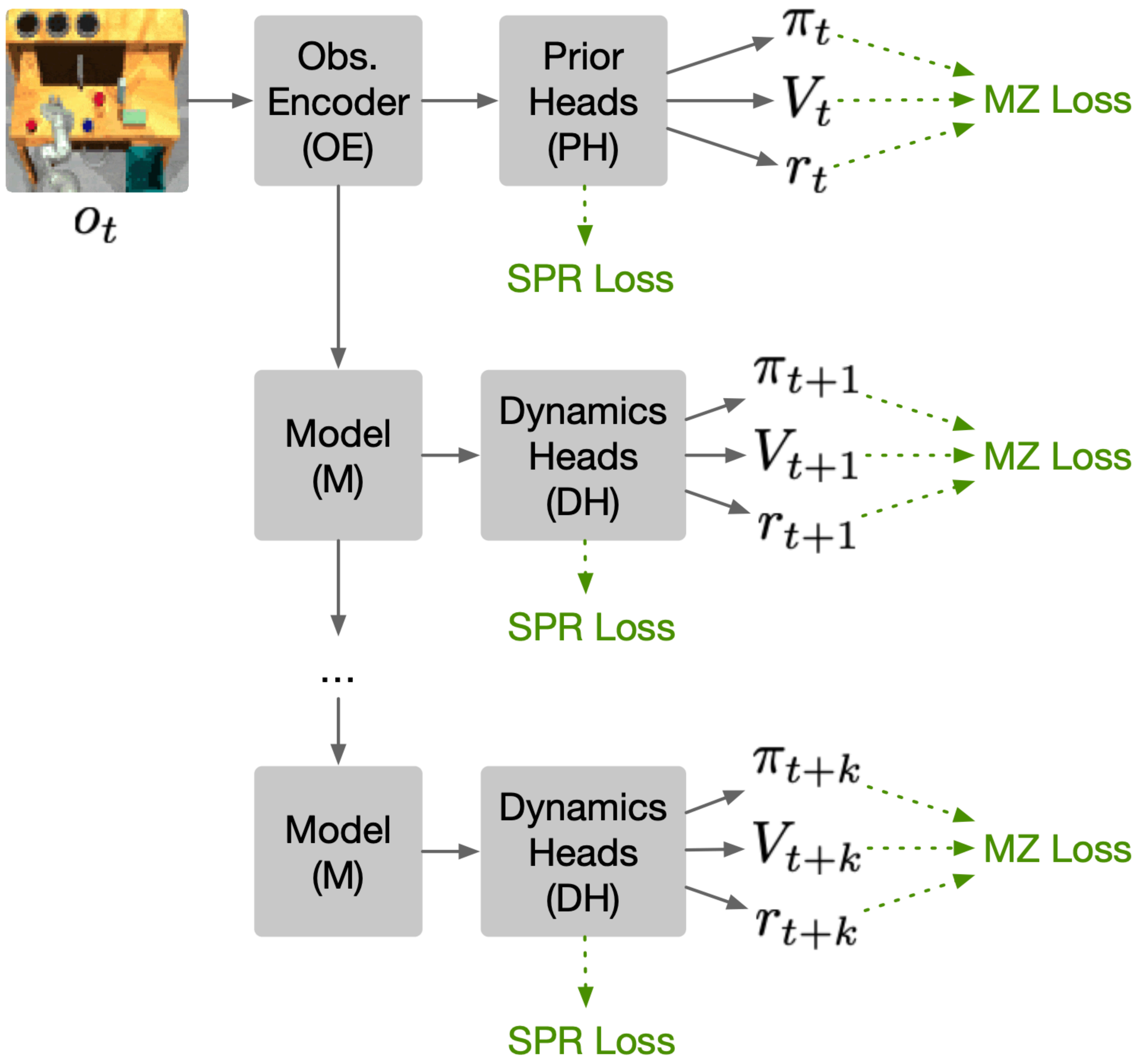
1. Is there an advantage to an agent being model-based during unsupervised exploration and/or fine-tuning?

2. What are the contributions of each component of a model-based agent for downstream task learning?

3. How well does the model-based agent deal with distribution shift between the unsupervised and fine-tuning phases?

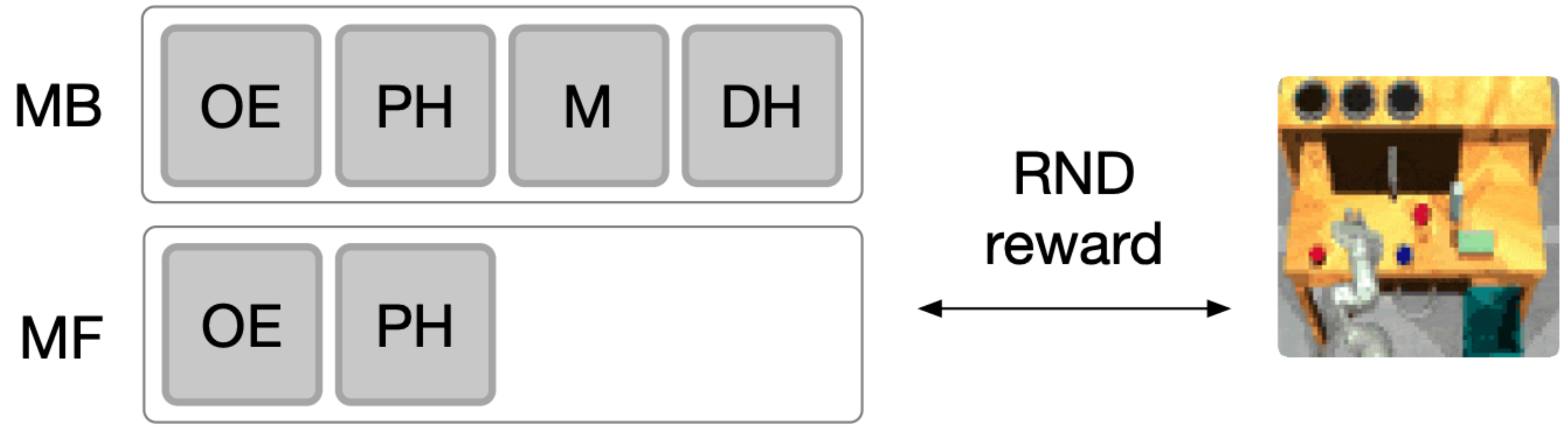
Jessica Hamrick - jhamrick@deepmind.com

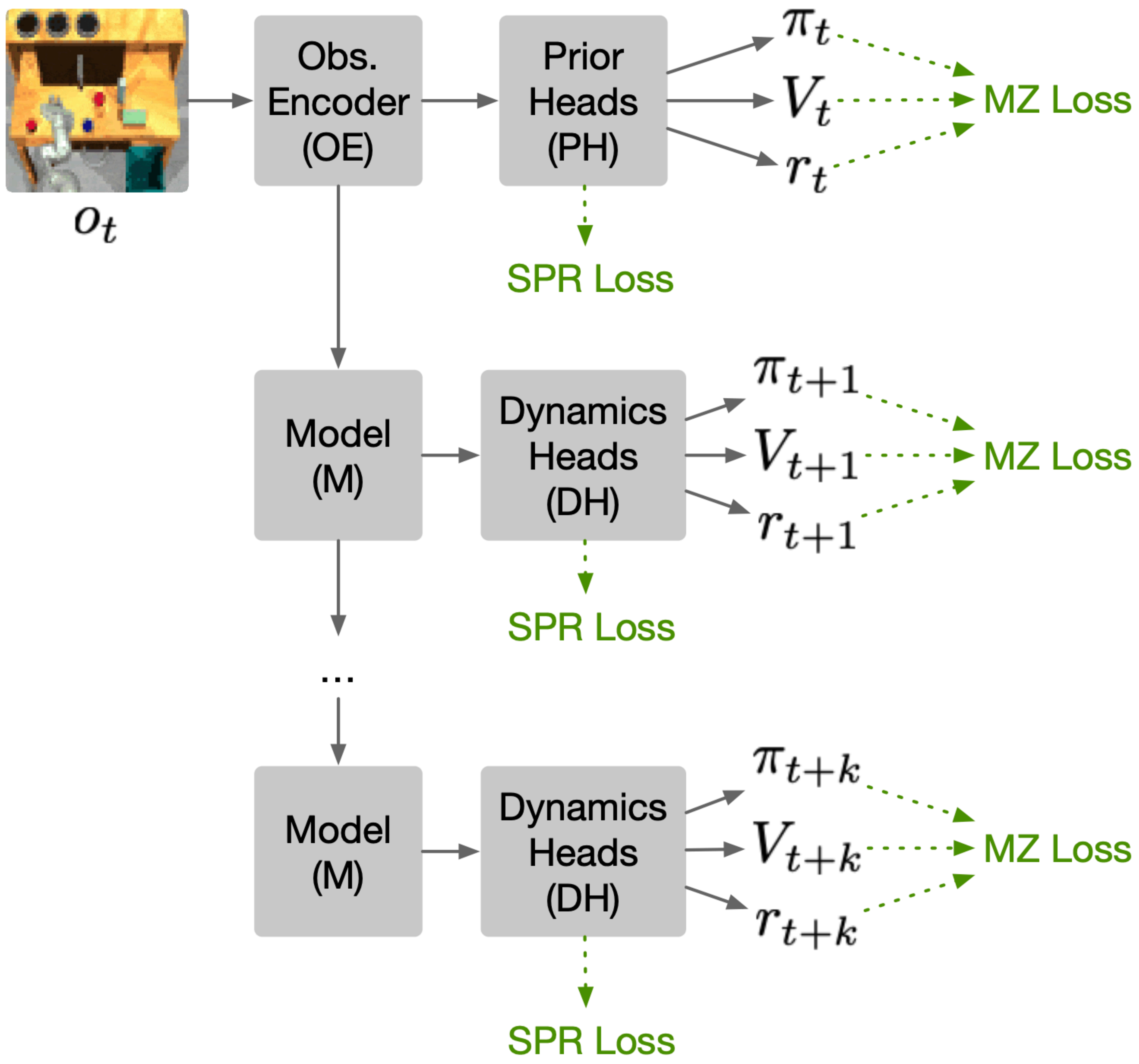
Experimental setup



Experimental setup

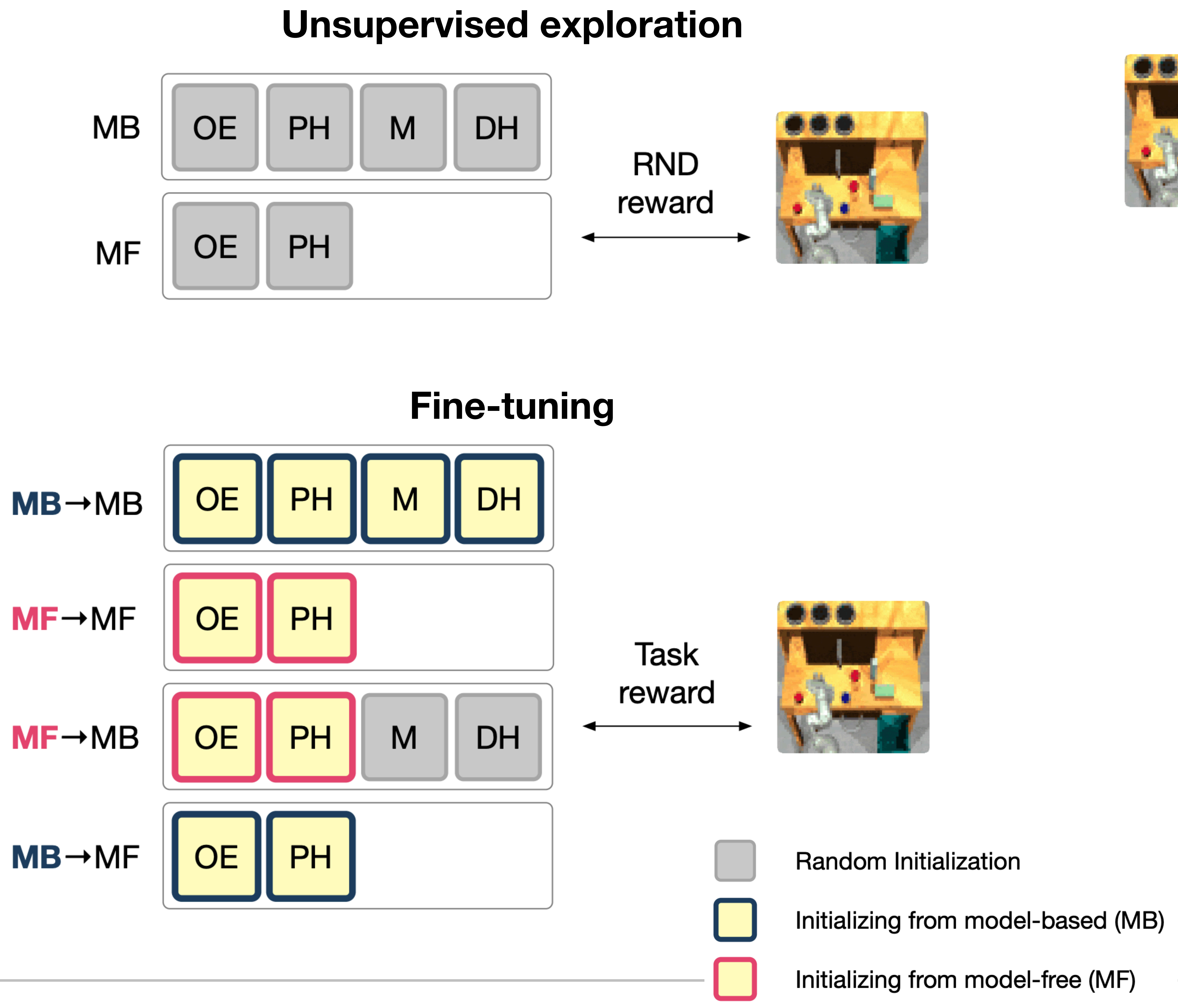
Unsupervised exploration

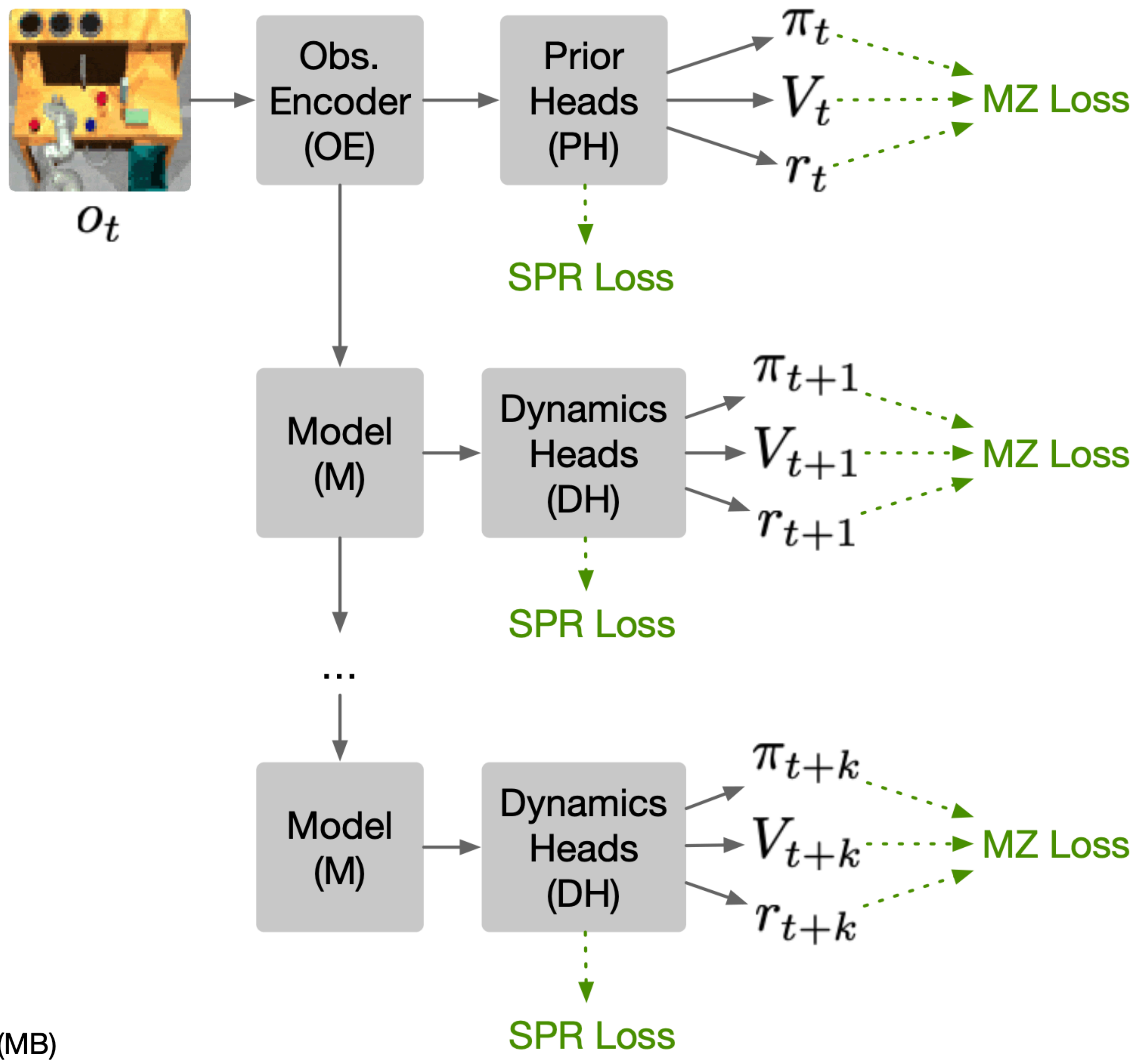




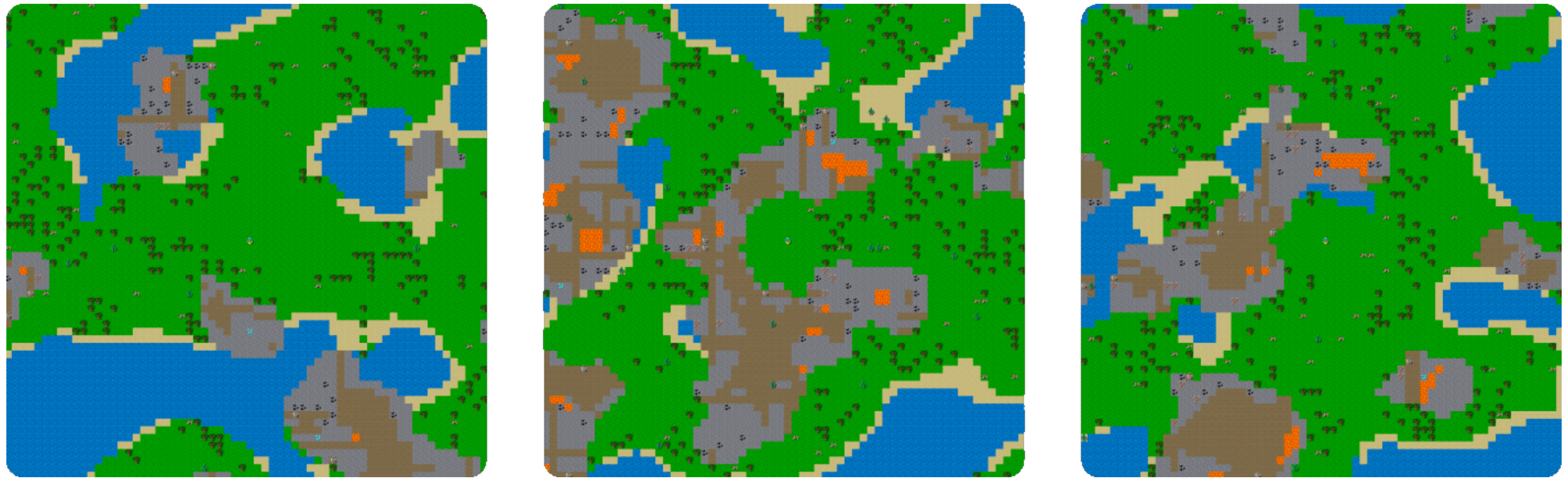
Jessica Hamrick - jhamrick@deepmind.com

Experimental setup





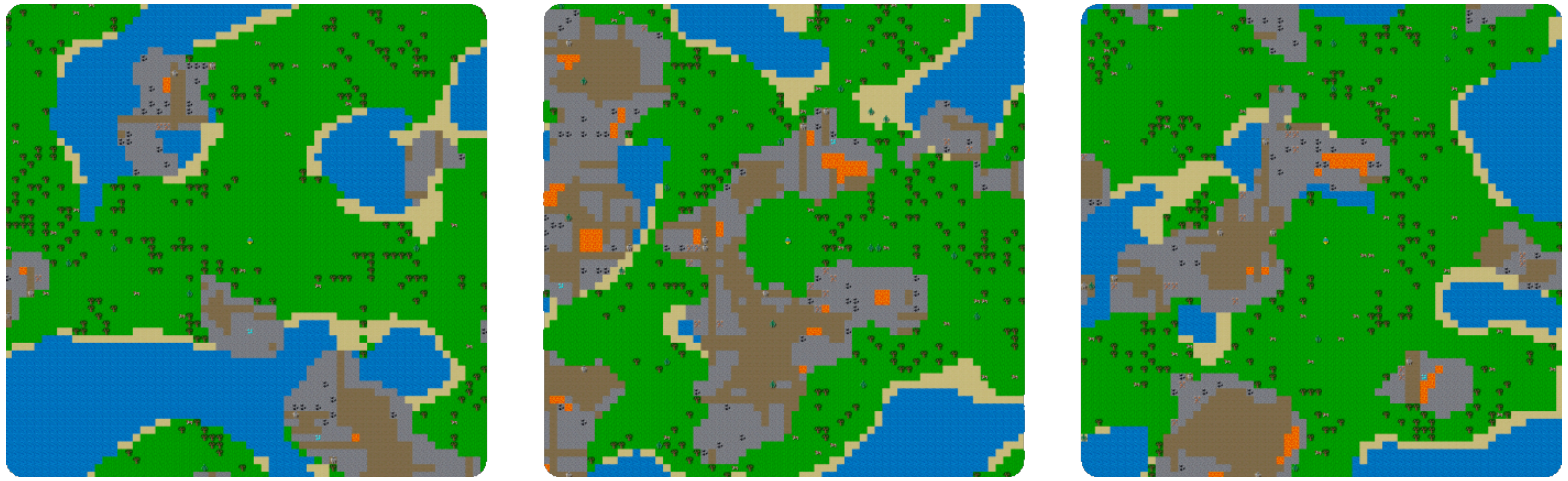
Environments



Crafter (Hafner, 2021)

RoboDesk (Kannan et al., 2021)

Environments

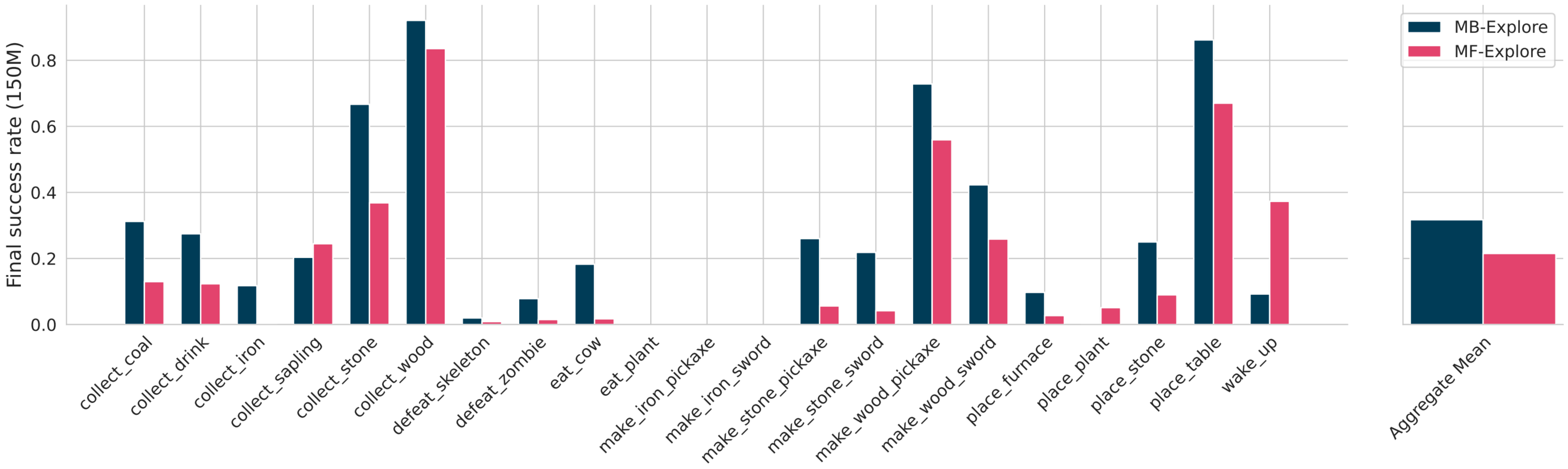


Crafter (Hafner, 2021)

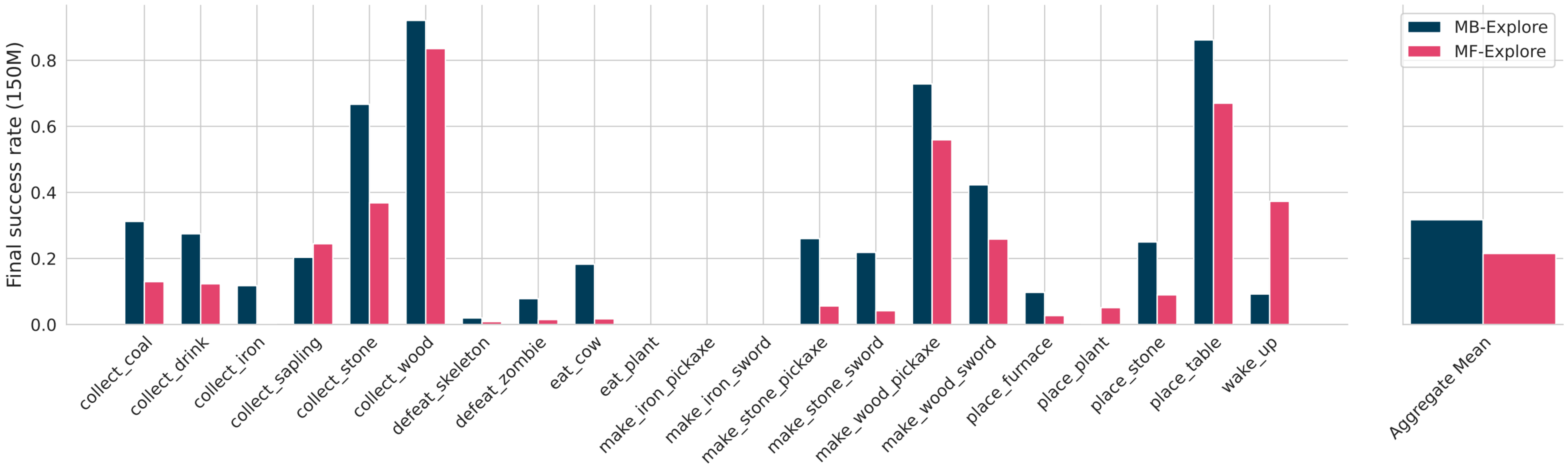
RoboDesk (Kannan et al., 2021)

Jessica Hamrick - jhamrick@deepmind.com

Exploration in Crafter

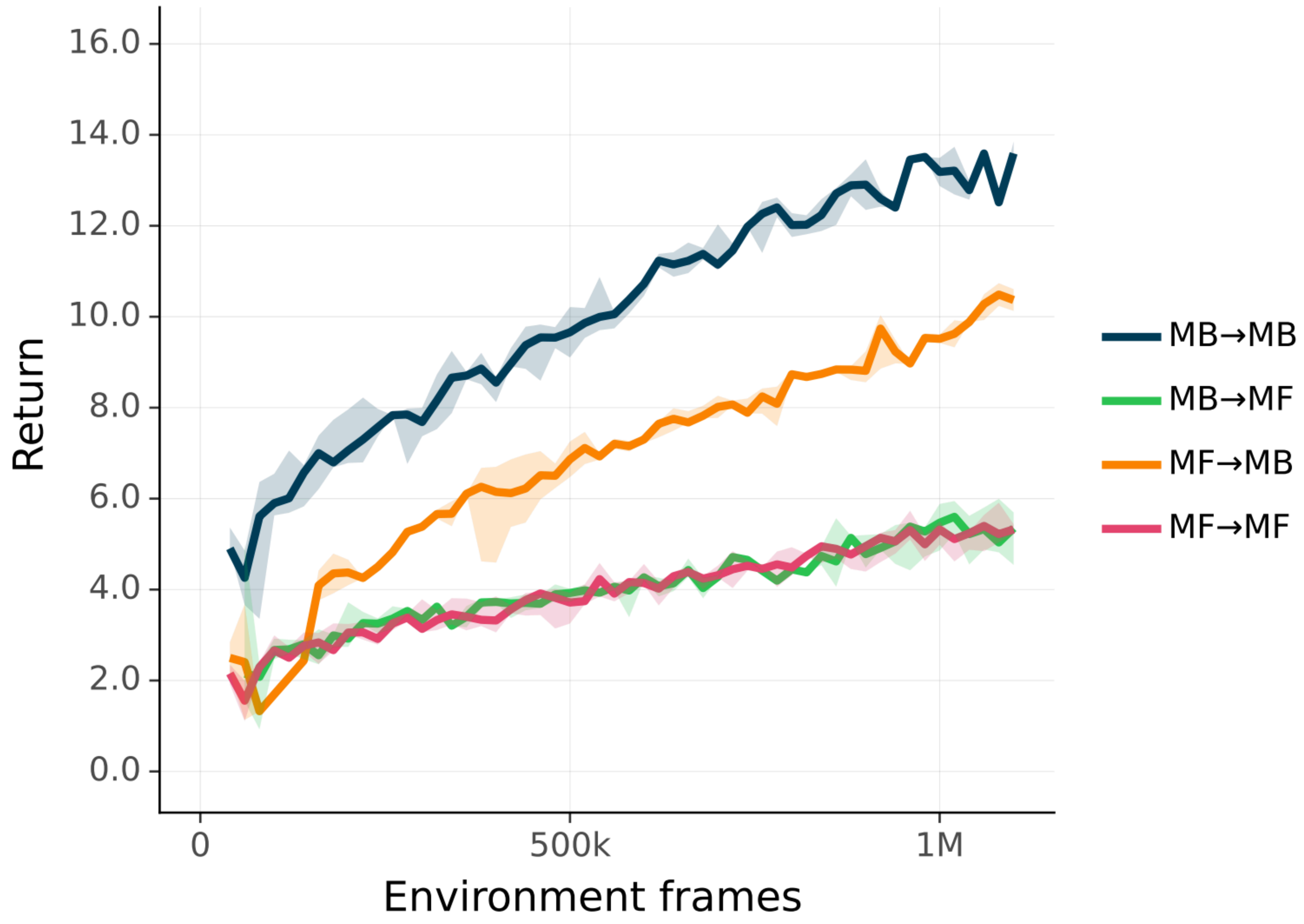


Exploration in Crafter



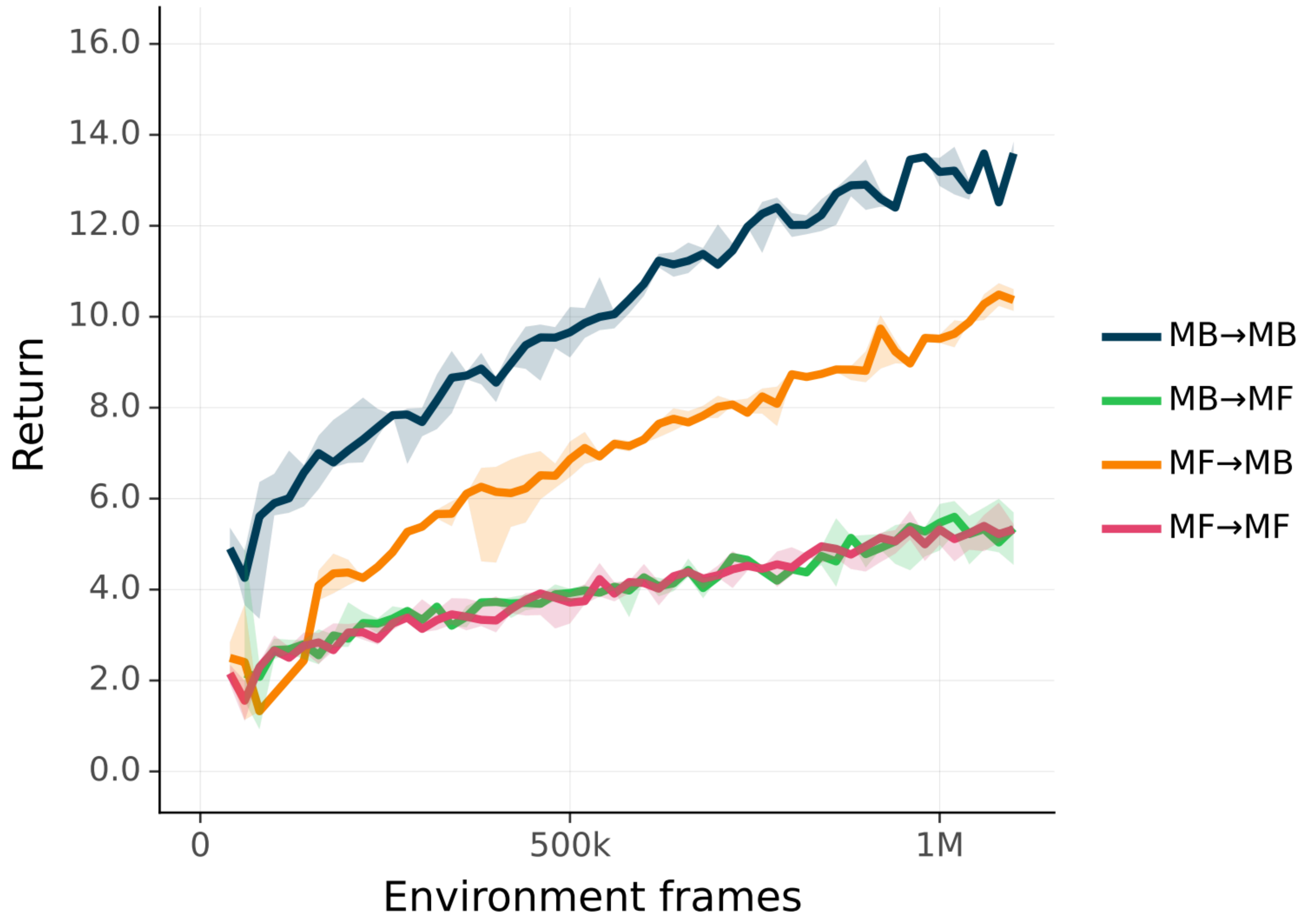
→ MB leads to improved exploration performance

Transfer in Crafter



Method	Score	Reward
Human Experts (Hafner, 2021)	50.5 ± 6.8	14.3 ± 2.3
MB→MB	$\mathbf{16.4 \pm 1.5}$	$\mathbf{12.7 \pm 0.4}$
MB→MF	8.8 ± 0.4	5.0 ± 0.2
MF→MB	6.2 ± 0.5	9.3 ± 0.3
MF→MF	6.7 ± 0.6	6.9 ± 0.2
DreamerV3 (Hafner et al., 2023)	14.5 ± 1.6	11.7 ± 1.9
LSTM-SPCNN (Stanić et al., 2022)	12.1 ± 0.8	-
DreamerV2 (Hafner, 2021)	10.0 ± 1.2	9.0 ± 1.7
MB Scratch	4.4 ± 0.4	8.5 ± 0.1

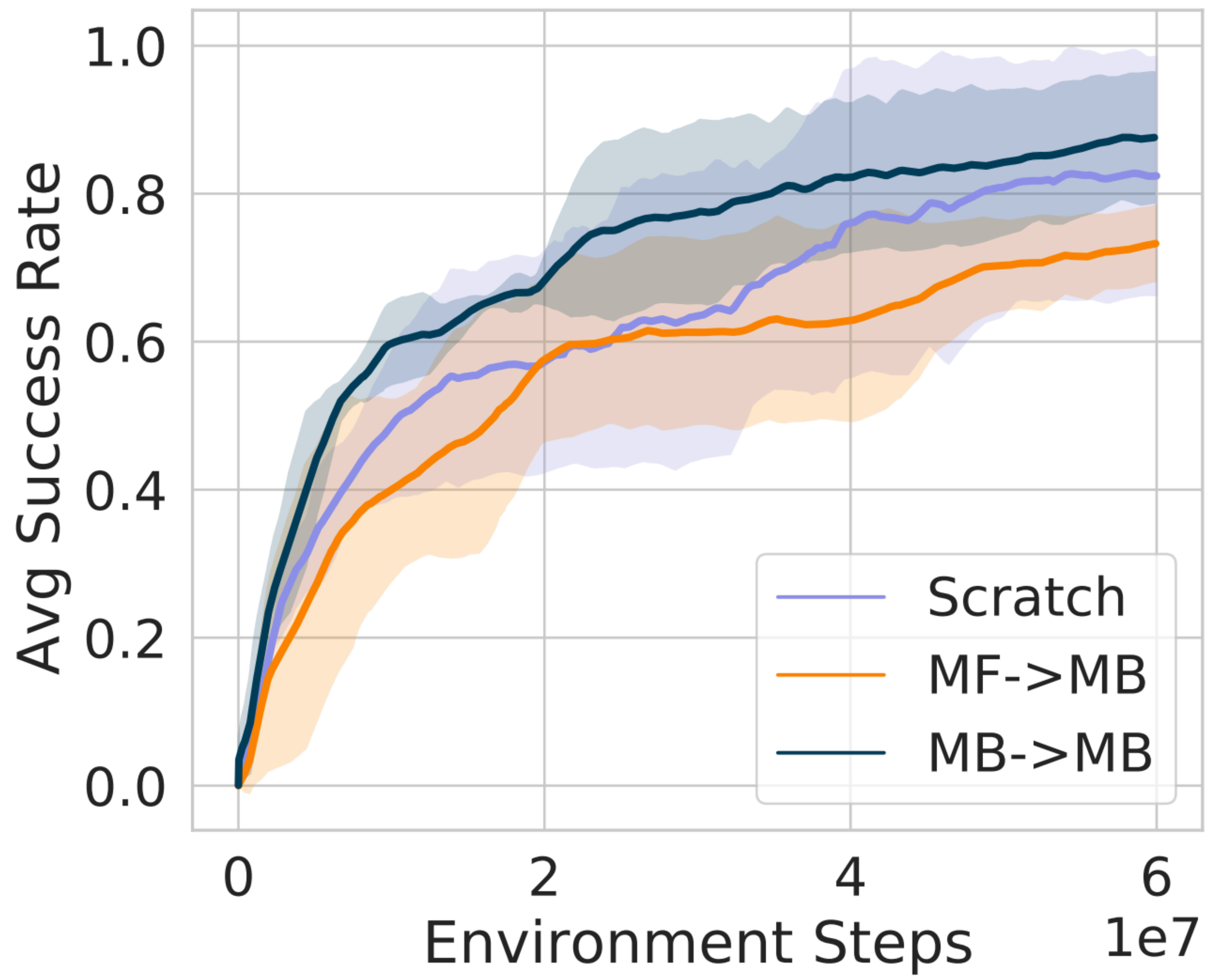
Transfer in Crafter

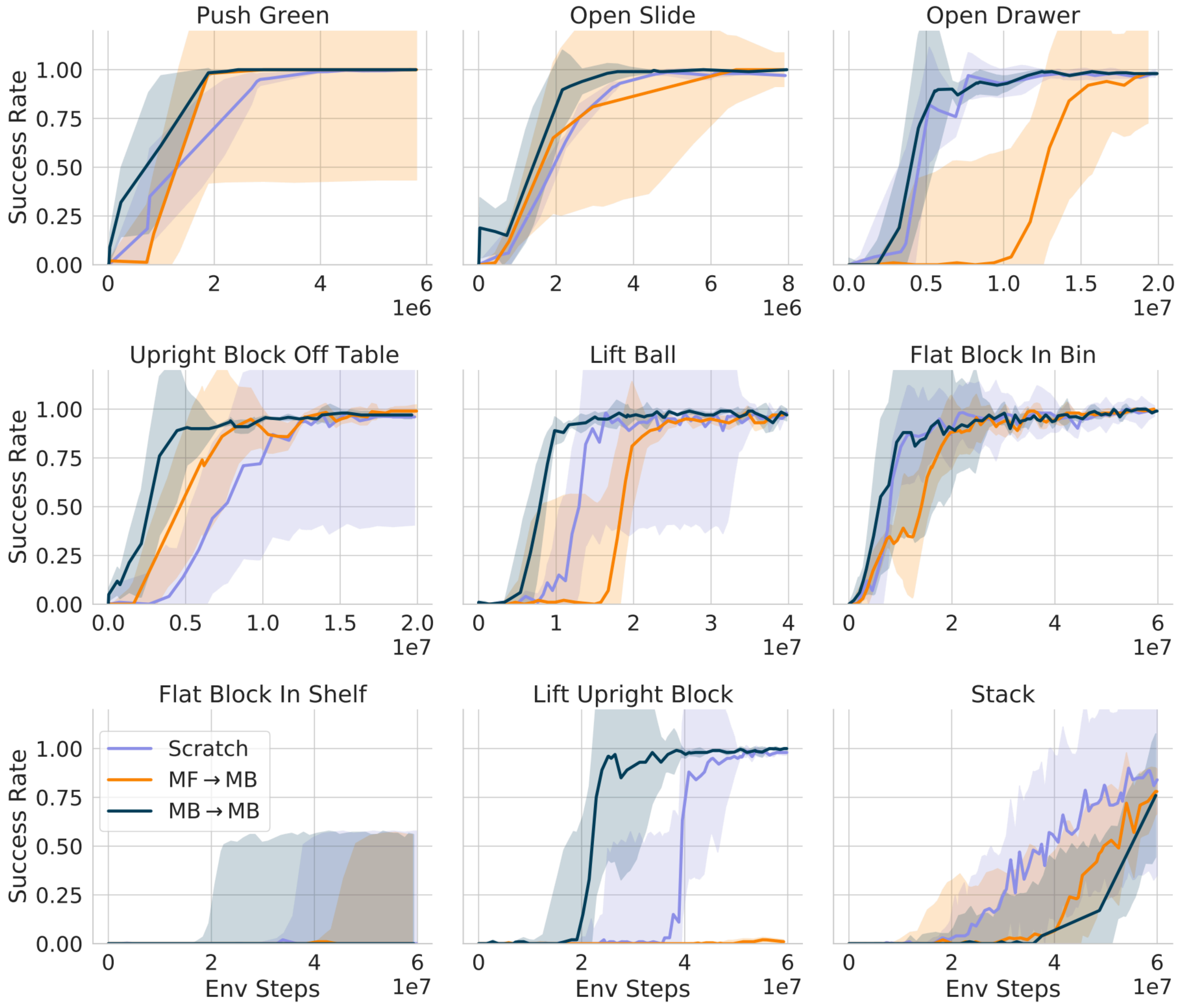


Method	Score	Reward
Human Experts (Hafner, 2021)	50.5 ± 6.8	14.3 ± 2.3
MB→MB	$\mathbf{16.4 \pm 1.5}$	$\mathbf{12.7 \pm 0.4}$
MB→MF	8.8 ± 0.4	5.0 ± 0.2
MF→MB	6.2 ± 0.5	9.3 ± 0.3
MF→MF	6.7 ± 0.6	6.9 ± 0.2
DreamerV3 (Hafner et al., 2023)	14.5 ± 1.6	11.7 ± 1.9
LSTM-SPCNN (Stanić et al., 2022)	12.1 ± 0.8	-
DreamerV2 (Hafner, 2021)	10.0 ± 1.2	9.0 ± 1.7
MB Scratch	4.4 ± 0.4	8.5 ± 0.1

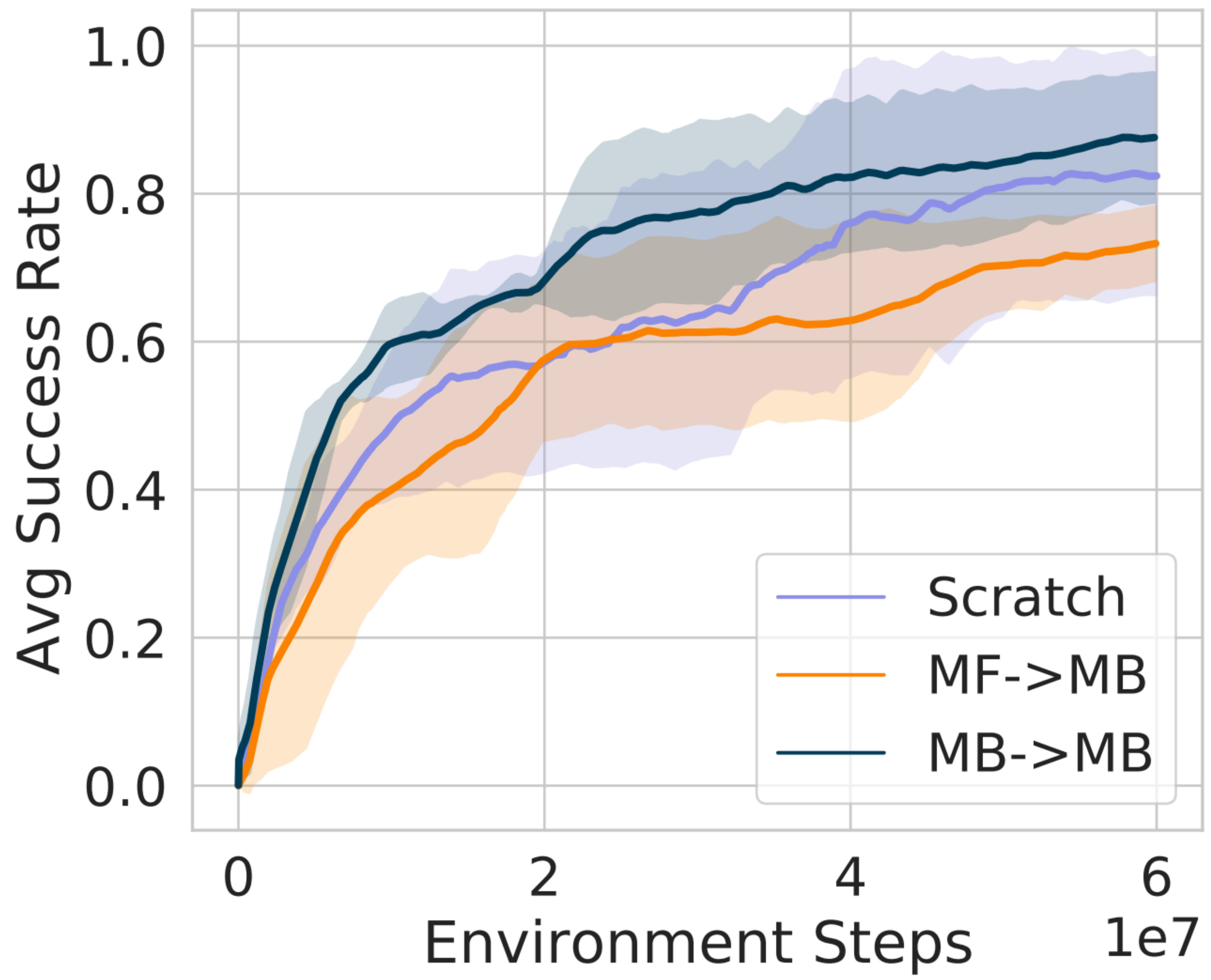
→ MB leads to improved transfer performance,
and matters a lot for finetuning

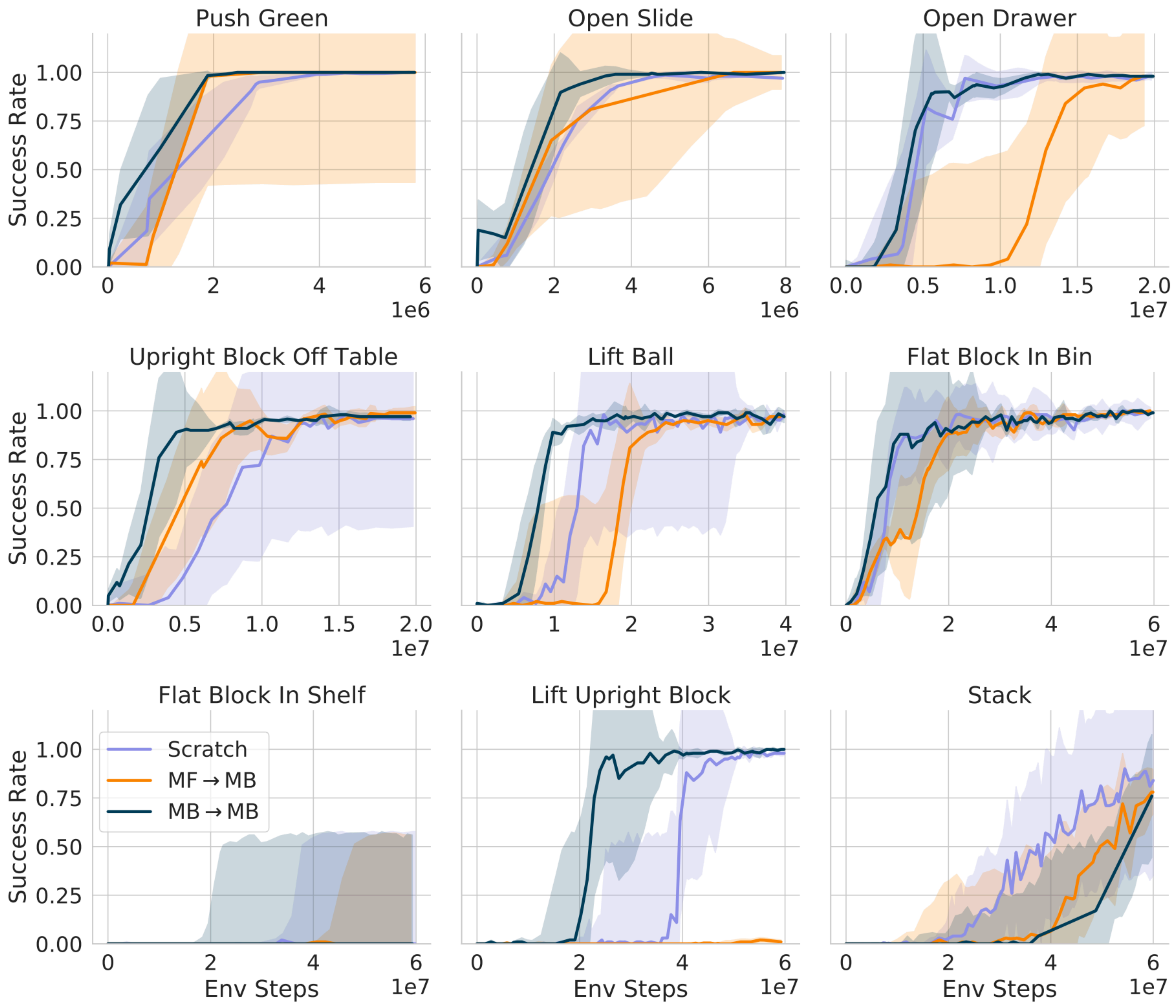
Transfer in Robodesk





Transfer in Robodesk

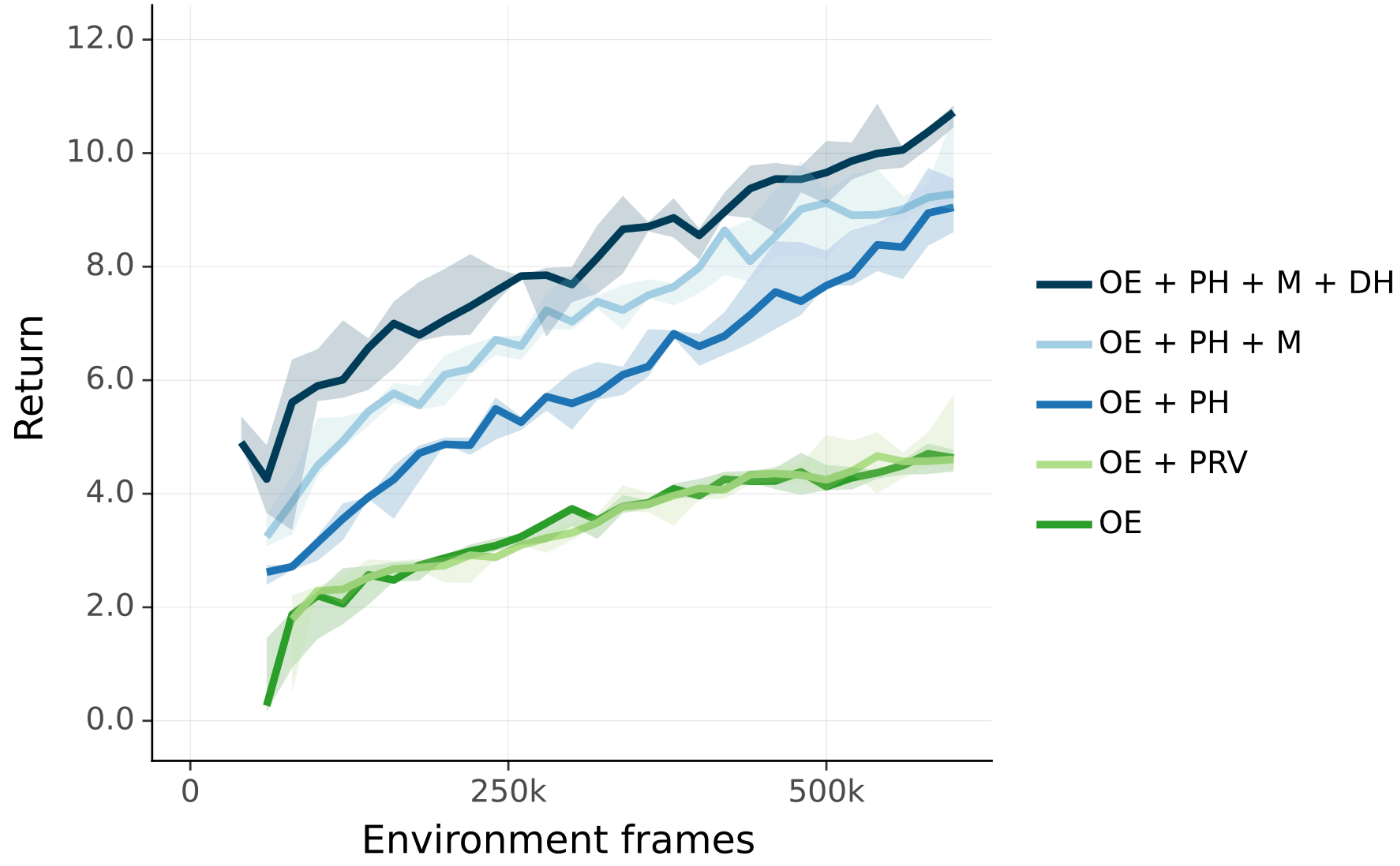


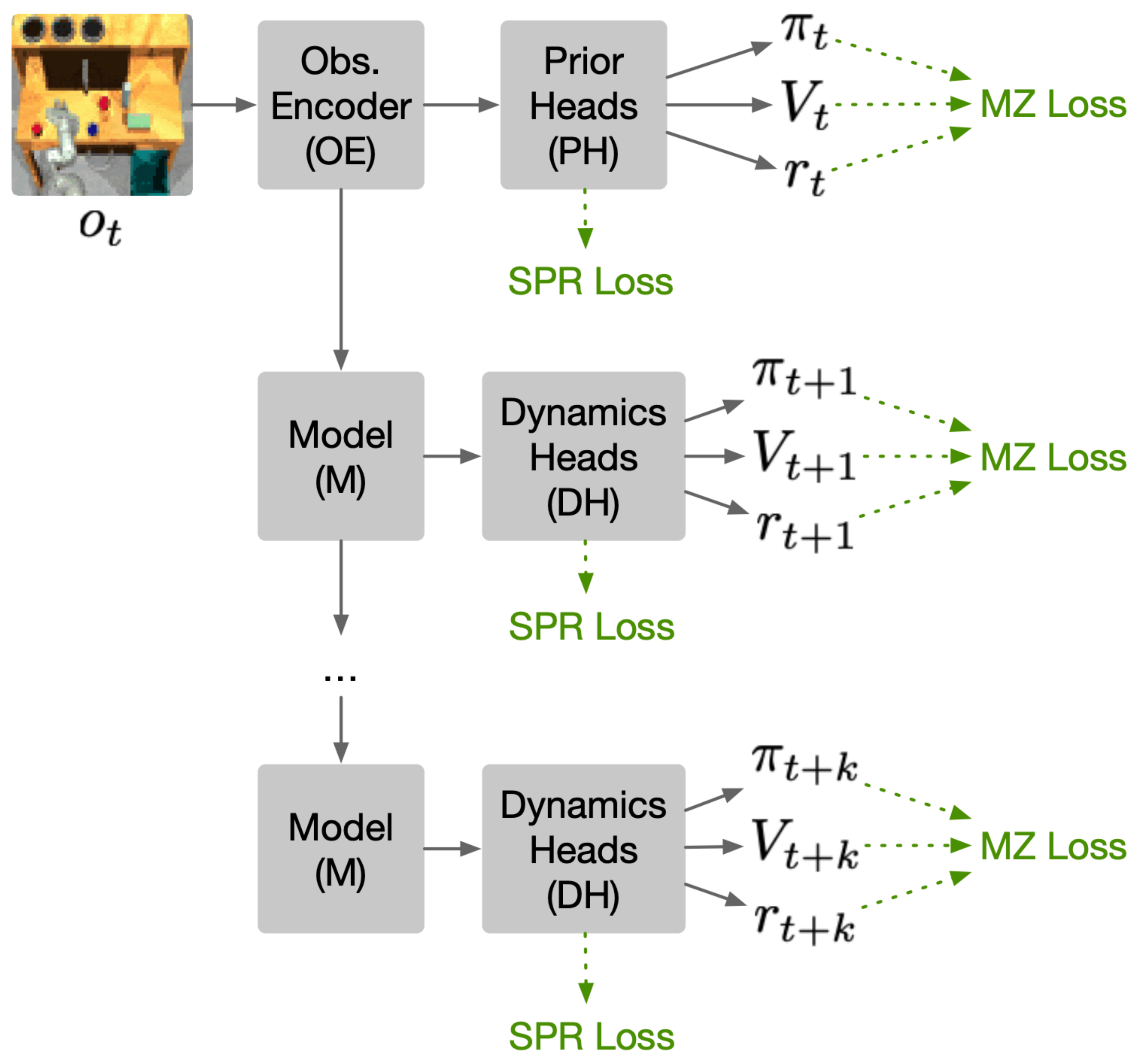


→ MB leads to improved transfer performance

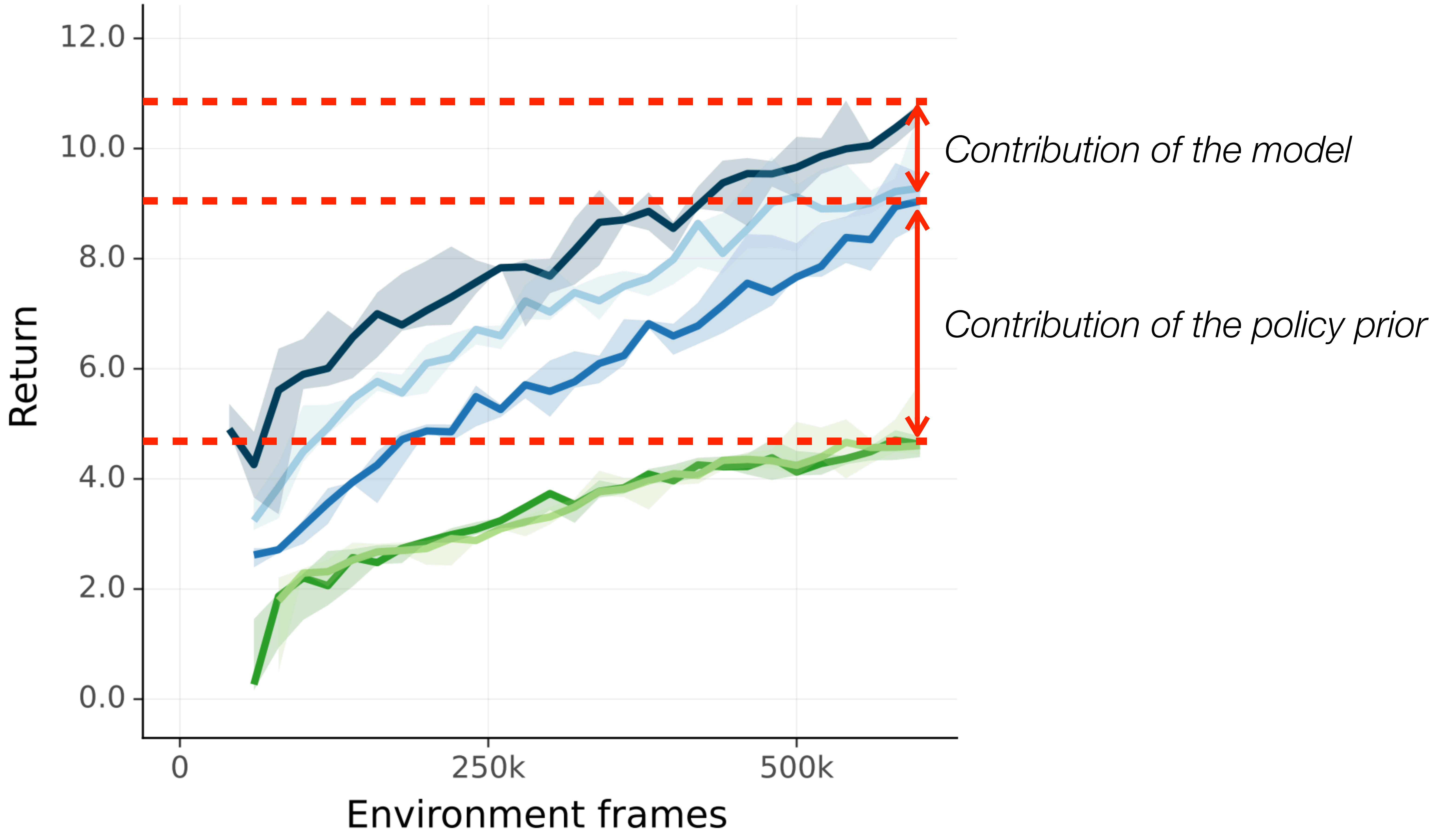
Jessica Hamrick - jhamrick@deepmind.com

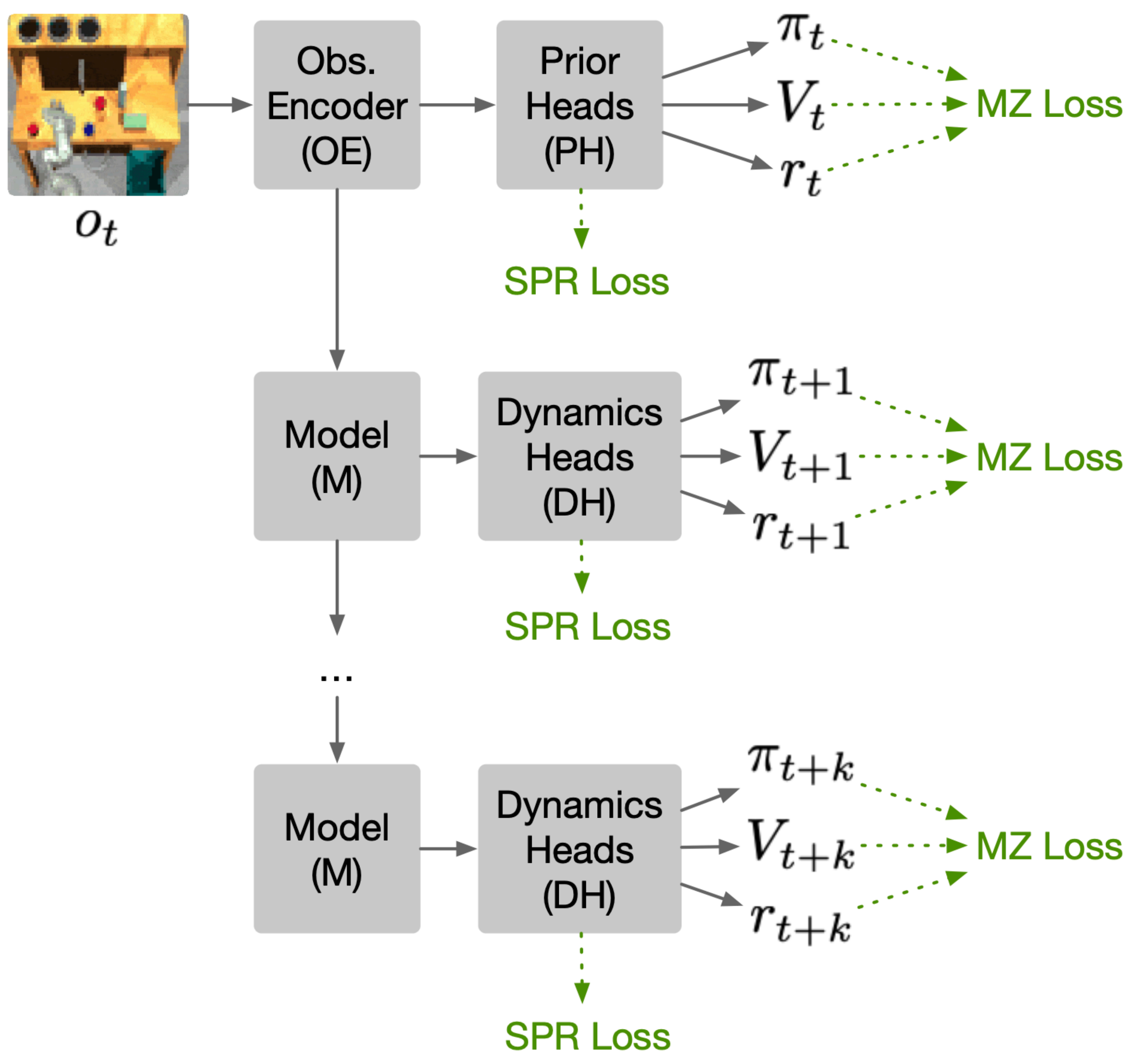
Contribution of different components





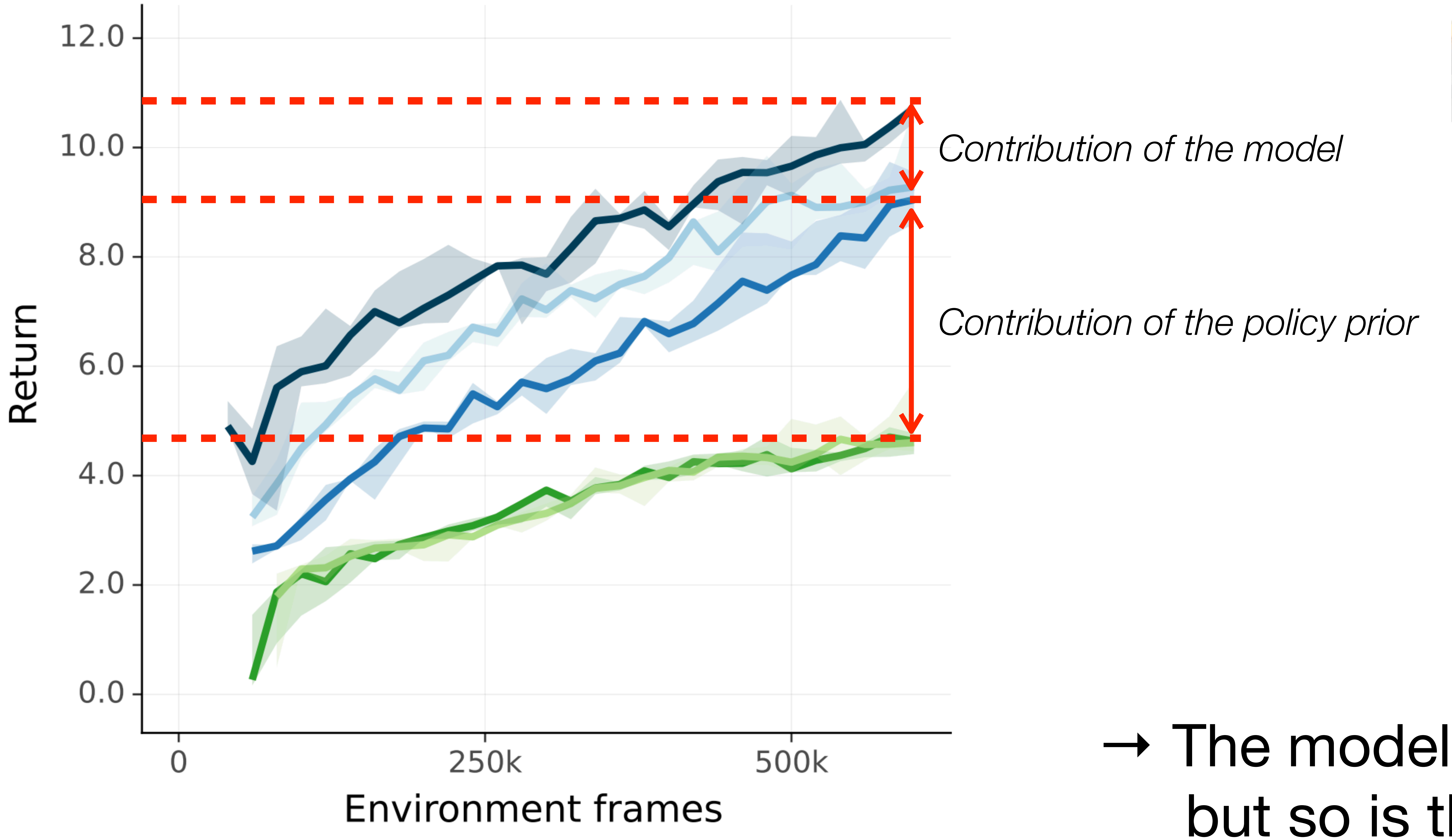
Contribution of different components

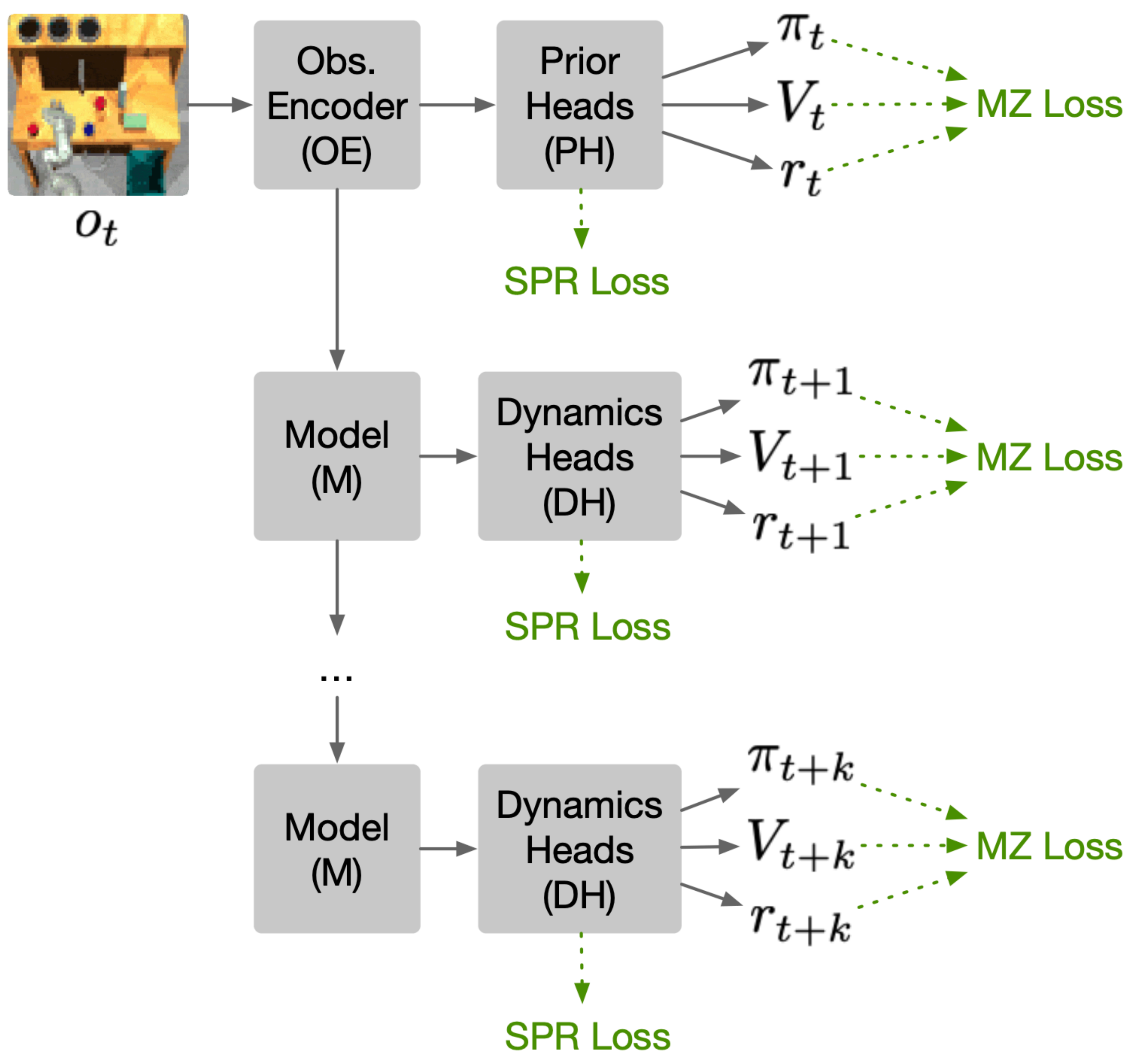




Jessica Hamrick - jhamrick@deepmind.com

Contribution of different components



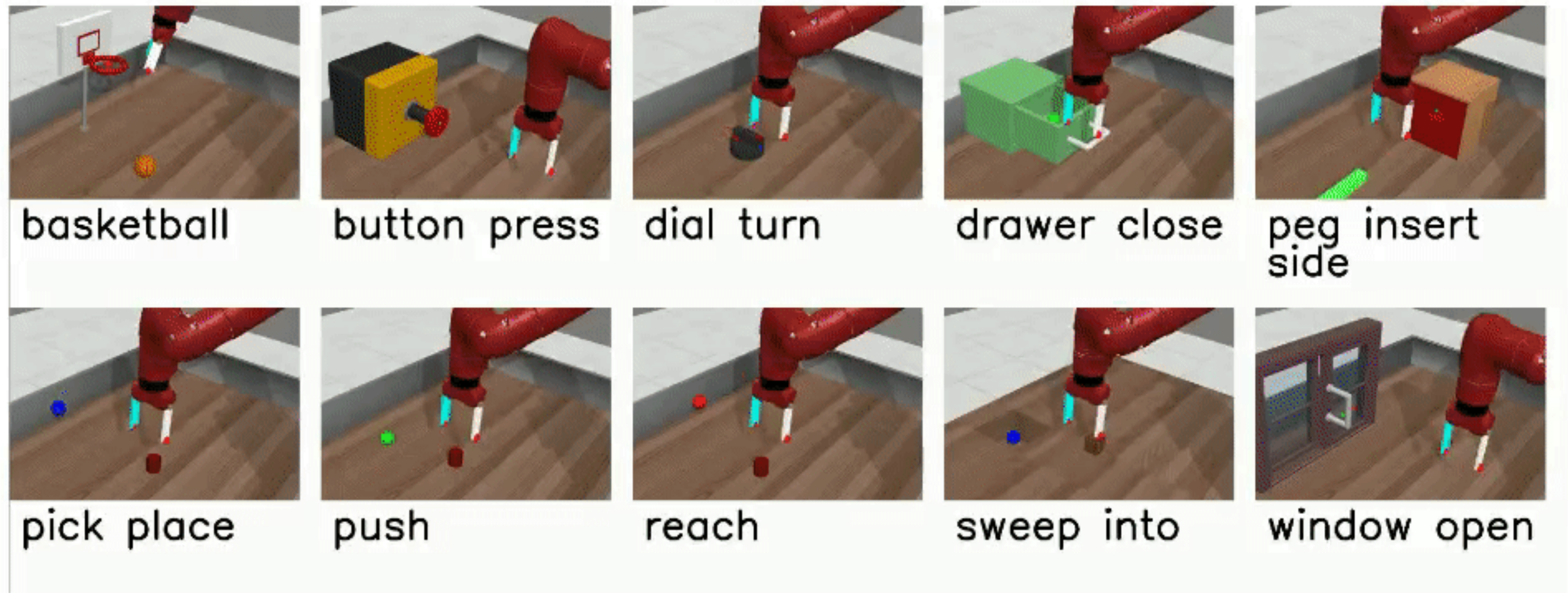


→ The model is important for transfer, but so is the exploration policy!

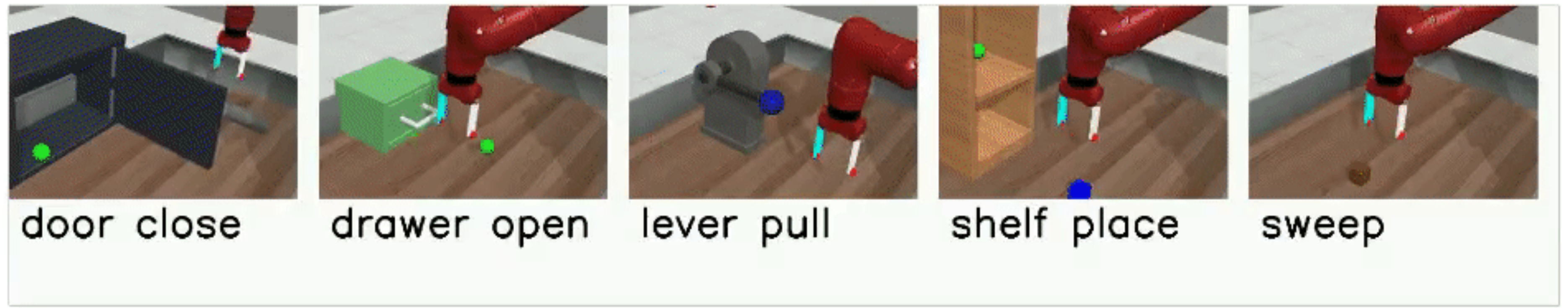
Jessica Hamrick - jhamrick@deepmind.com

Transfer in MetaWorld

Train



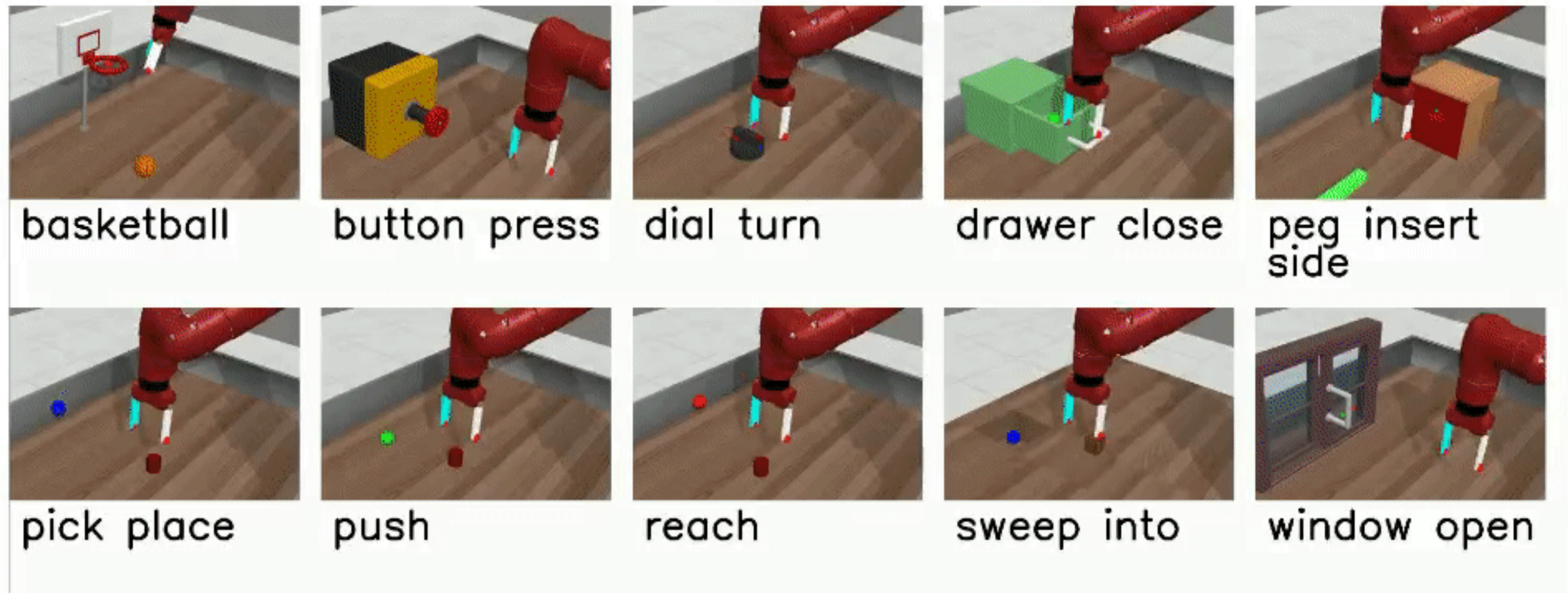
Test



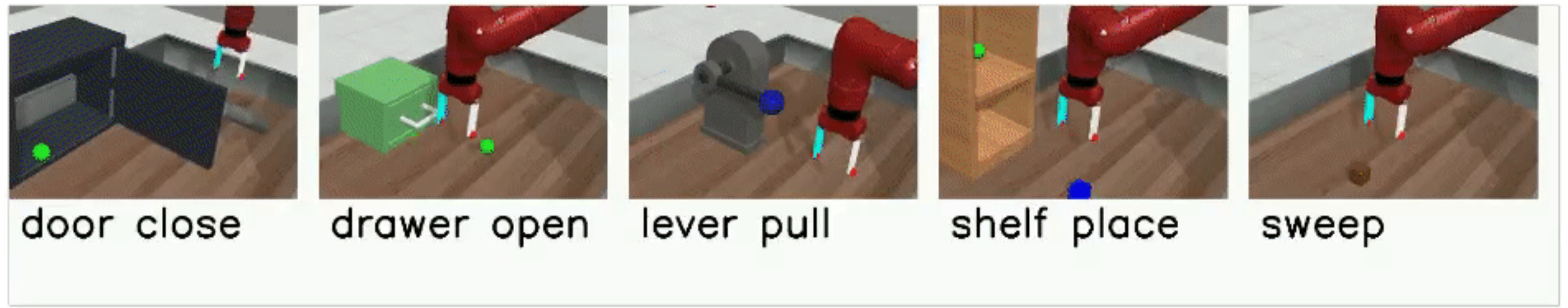
Jessica Hamrick - jhamrick@deepmind.com

Transfer in MetaWorld

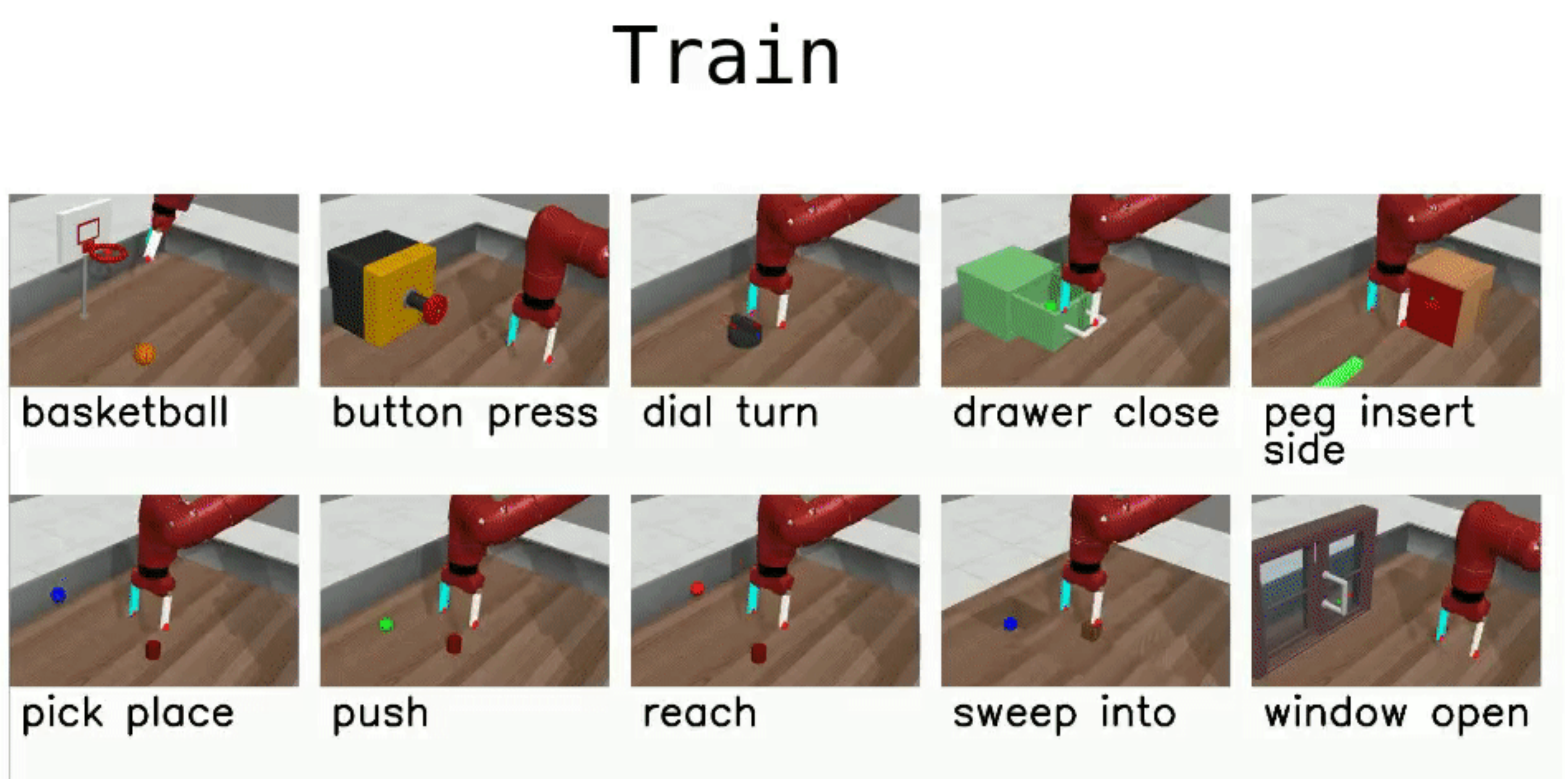
Train

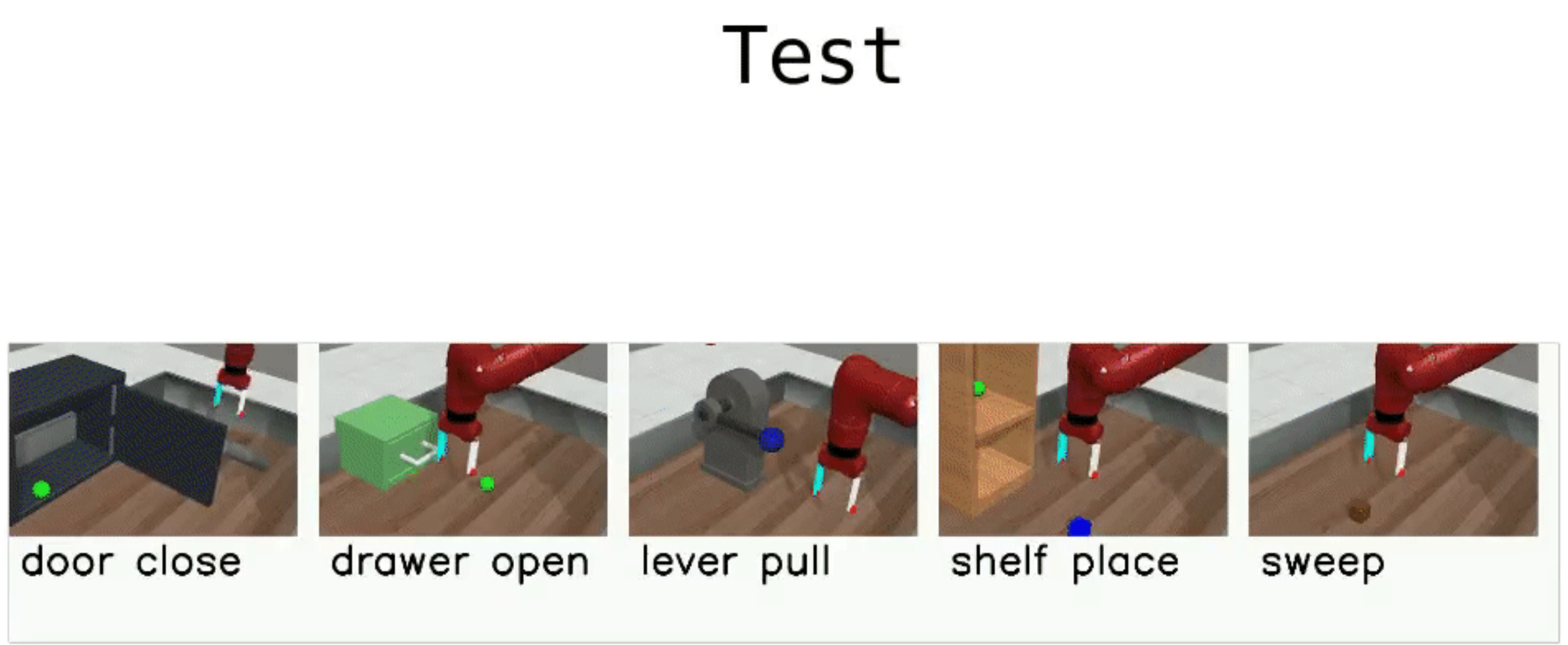


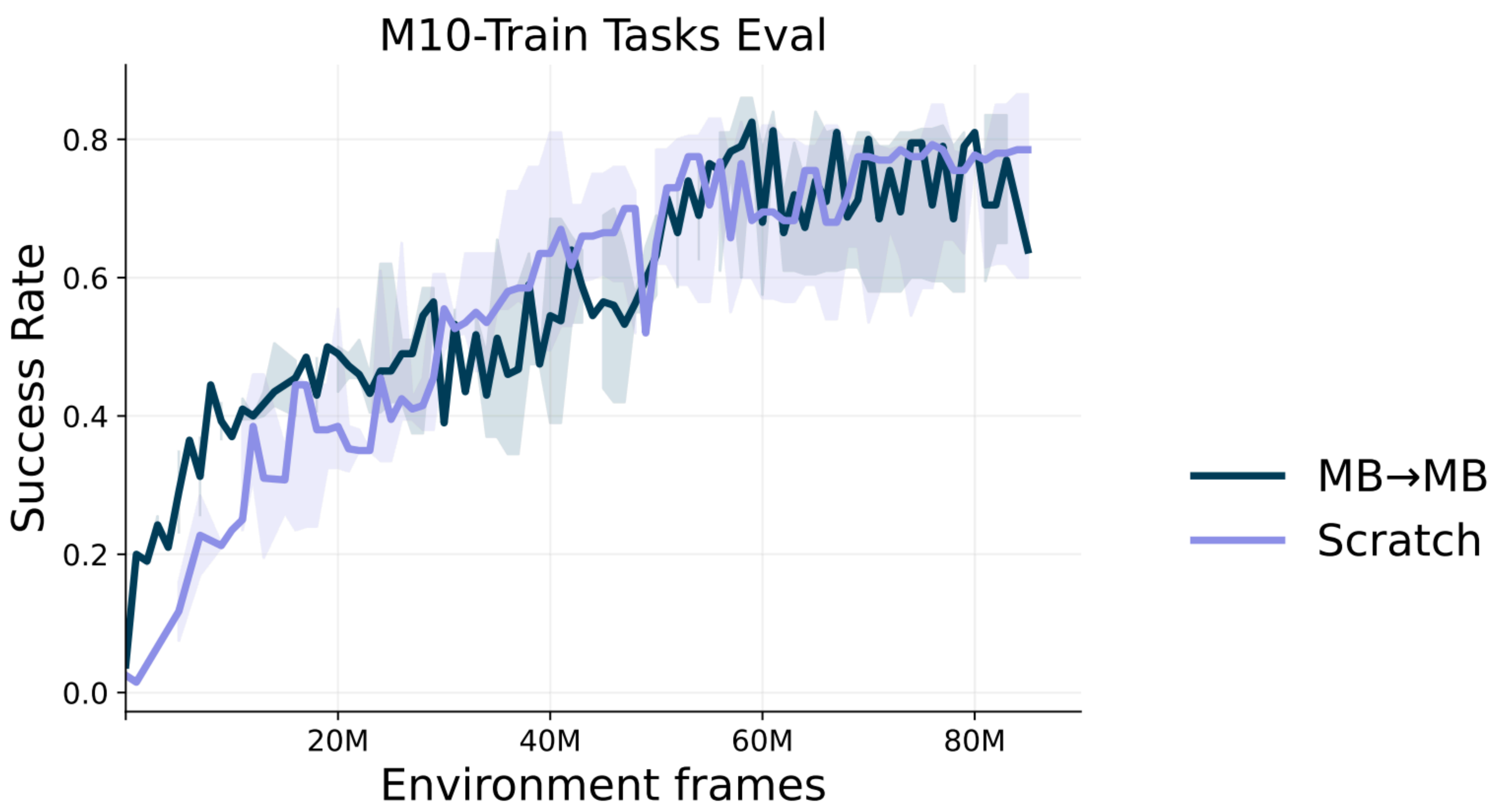
Test



Transfer in MetaWorld





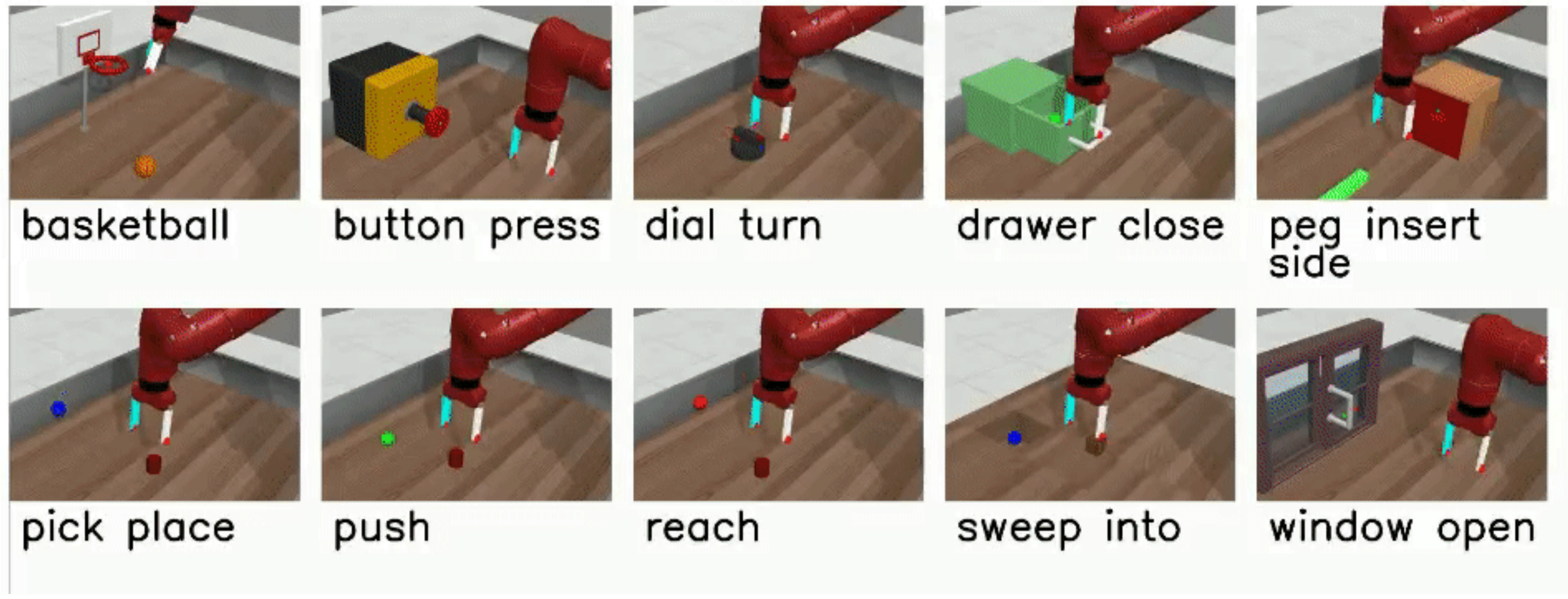


→ MBRL may not substantially improve transfer performance if there is a large environment shift

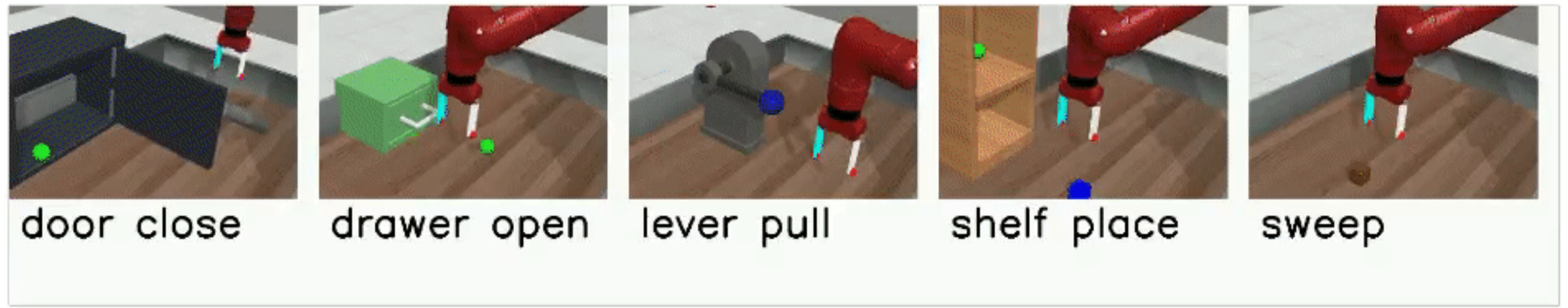
Jessica Hamrick - jhamrick@deepmind.com

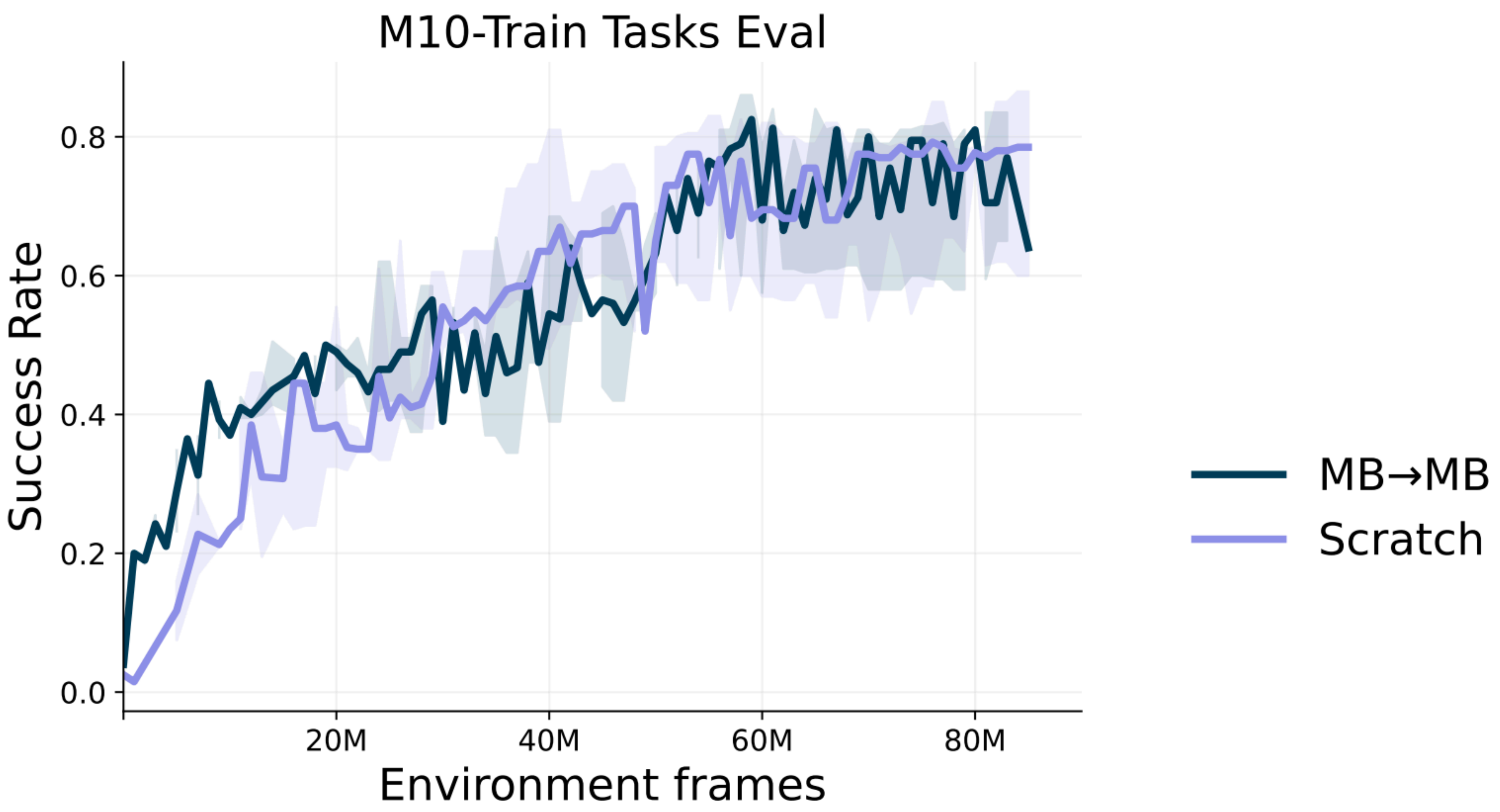
Transfer in MetaWorld

Train



Test





→ MBRL may not substantially improve transfer performance if there is a large environment shift

Interim Takeaway #5: Model-based pre-training and fine-tuning can substantially improve transfer performance, but only if there is minimal distribution shift.

Interim Takeaway #6: Effective transfer requires learning a good policy *and* a good model! (Sounds familiar...)

Outline

- **Understanding MBRL**
  Hamrick et al. (2021). On the role of planning in model based reinforcement learning. ICLR.
- **Understanding and improving generalization**
  Anand, Walker et al. (2022). Procedural generalization by planning with self-supervised world models. ICLR.
- **Understanding and improving transfer**
  Walker, Vértes, Li et al. (2023). Investigating the role of model-based learning in exploration and transfer. Under review.
- **The future of MBRL**

Overall Learnings

Overall Learnings

1. Planning seems to be **most useful during learning** and less so at test time (in most environments).

Overall Learnings

1. Planning seems to be **most useful during learning** and less so at test time (in most environments).

2. Effective planning, generalization, and transfer all depend on **multiple components** (e.g., policies, value functions, models).
    - Improved representations through self-supervision.
    - However, performance still relies on there being minimal distribution shift.

Overall Learnings

1. Planning seems to be **most useful during learning** and less so at test time (in most environments).

2. Effective planning, generalization, and transfer all depend on **multiple components** (e.g., policies, value functions, models).
    - Improved representations through self-supervision.
    - However, performance still relies on there being minimal distribution shift.

3. Self-supervision interacts positively with the **number of environments**. We should be wary of drawing conclusions from single-task settings!

Outline

- **Understanding MBRL**
  Hamrick et al. (2021). On the role of planning in model based reinforcement learning. ICLR.

- **Understanding and improving generalization**
  Anand, Walker et al. (2022). Procedural generalization by planning with self-supervised world models. ICLR.

- **Understanding and improving transfer**
  Walker, Vértes, Li et al. (2023). Investigating the role of model-based learning in exploration and transfer. Under review.

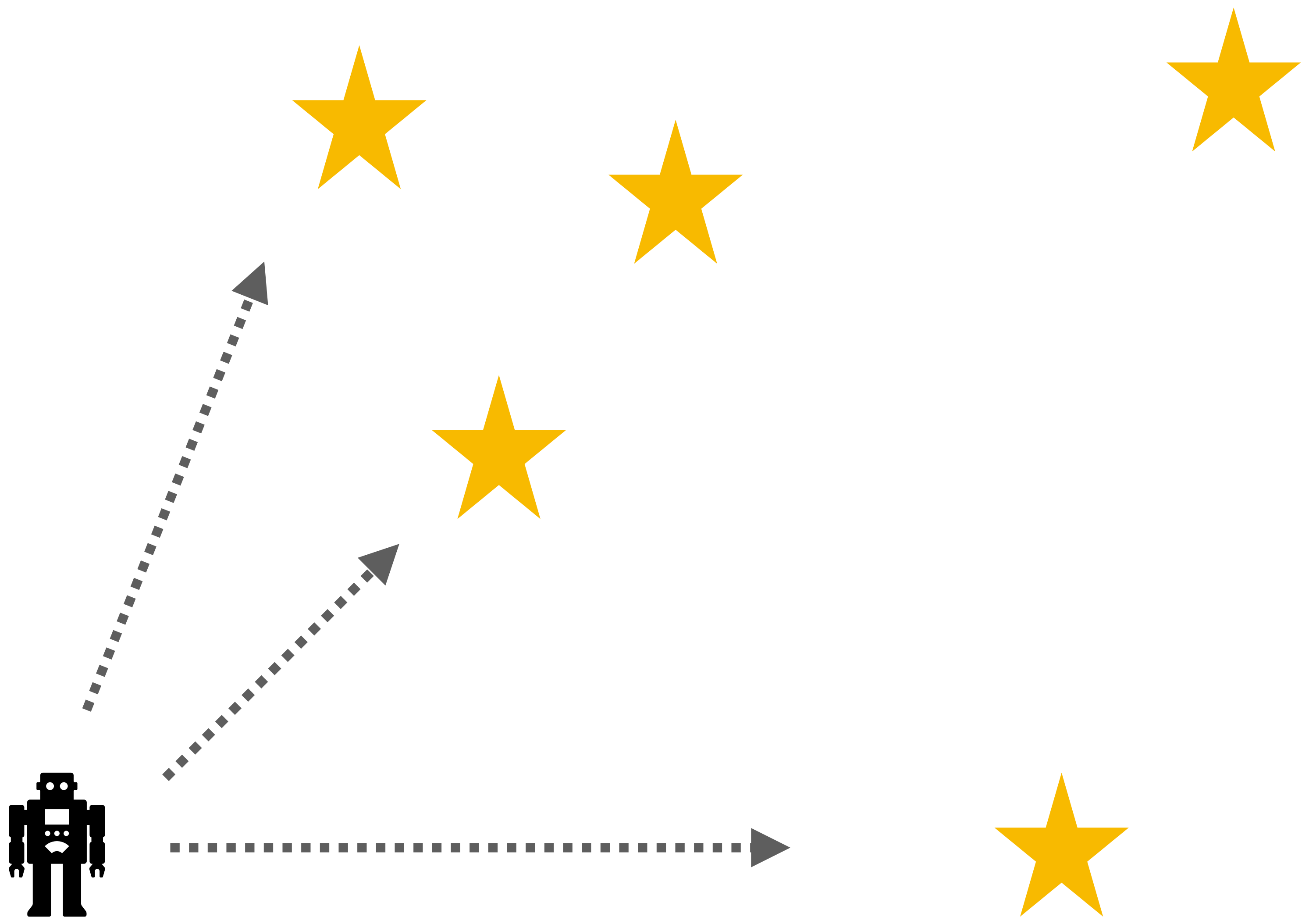
- **The future of MBRL**

Still missing: **deliberative reasoning**

"Model-free algorithms are in turn far from the state of the art in domains that require ***precise and sophisticated lookahead***, such as chess and Go"
-Schrittwieser et al. (2019)

Still missing: **deliberative reasoning**

"Model-free algorithms are in turn far from the state of the art in domains that require ***precise and sophisticated lookahead***, such as chess and Go"
-Schrittwieser et al. (2019)



Still missing: **strong generalization**

"Model-based planning is an essential ingredient of human intelligence, enabling ***flexible adaptation*** to new tasks and goals"
-Lake et al. (2016)

Still missing: **strong generalization**

"Model-based planning is an essential ingredient of human intelligence, enabling ***flexible adaptation*** to new tasks and goals"
-Lake et al. (2016)

Generic world model

Still missing: **strong generalization**

"Model-based planning is an essential ingredient of human intelligence, enabling ***flexible adaptation*** to new tasks and goals"
-Lake et al. (2016)

Generic world model

Generic exploration policy

Still missing: **strong generalization**

"Model-based planning is an essential ingredient of human intelligence, enabling ***flexible adaptation*** to new tasks and goals"
-Lake et al. (2016)

Generic world model

Generic exploration policy

Reward function synthesizer

Thanks!

Ankesh Anand
Victor Bapst
Peter Battaglia
Lars Buesing
Thomas Anthony
Feryal Behbahani
Lars Buesing
Gabriel Dulac-Arnold
Abe Friesen
Arthur Guez
Yazhe Li

Sherjil Ozair
Tobias Pfaff
Alvaro Sanchez-Gonzalez
Julian Schrittwieser
Petar Veličković
Eszter Vértes
Fabio Viola
Jacob Walker
Sims Witherspoon
Theo Weber

Hamrick, Bapst, Sanchez-Gonzalez, Pfaff, Weber, Buesing, & Battaglia (2020). Combining Q-learning and search with amortized value estimates. *ICLR.*

Hamrick, Friesen, Behbahani, Guez, Viola, Witherspoon, Anthony, Buesing, Veličković, & Weber (2021). On the role of planning in model-based deep reinforcement learning. *ICLR.*

Anand*, Walker*, Li, Vértes, Schrittwieser, Ozair, Weber, & Hamrick (2022). Procedural generalization by planning with self-supervised world models. *ICLR.*

Walker*, Vértes*, Li*, Dulac-Arnold, Anand, Weber, & Hamrick (2023). Investigating the role of model-based learning in exploration and transfer. *arXiv.*