IDSIA

Google

Università della Svizzera italiana

KAUST

NEURAL INFORMATION PROCESSING SYSTEMS

Generalization in Planning 2023

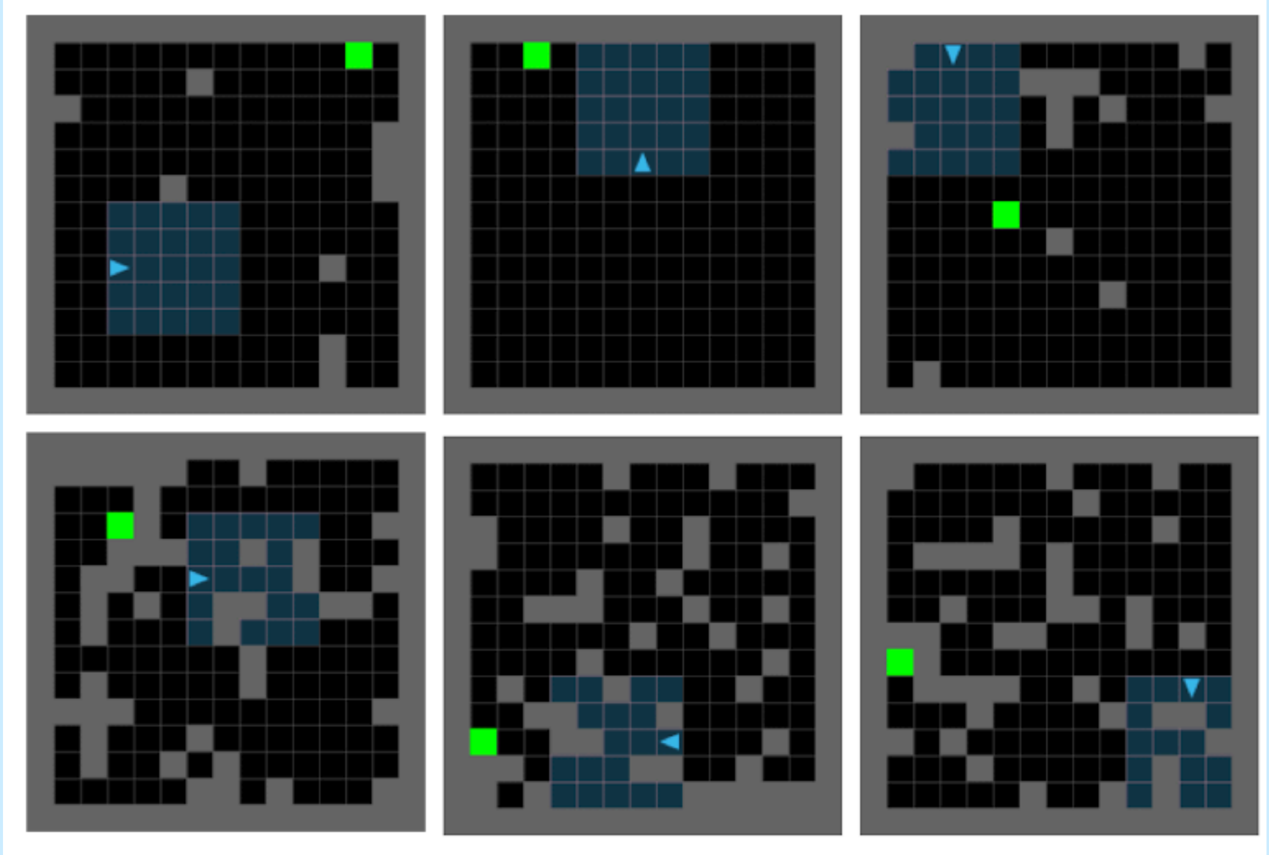
Towards **General-Purpose** In-Context Learning **Agents**

Louis Kirsch, James Harrison, C. Daniel Freeman,
Jascha Sohl-Dickstein, Jürgen Schmidhuber

@LouisKirschAI

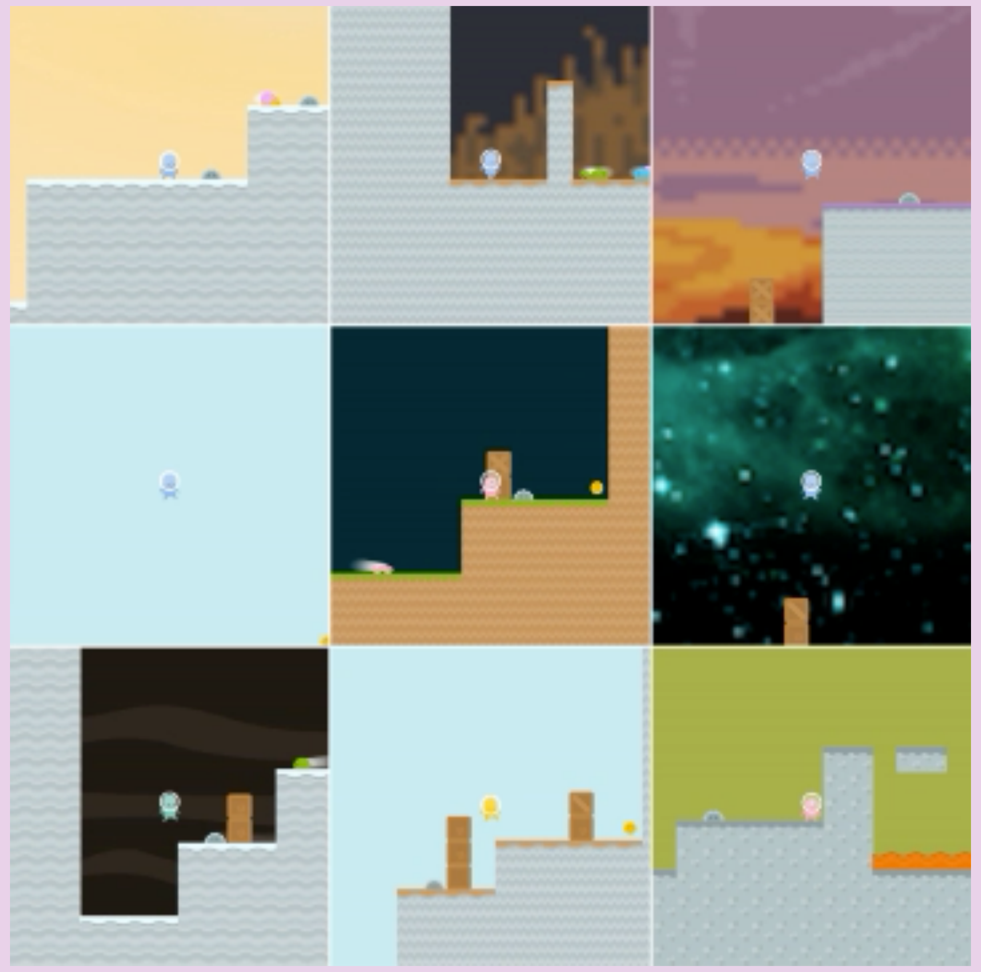
Main focus of Meta-Learning

- Excels in adaptation to unseen but similar tasks



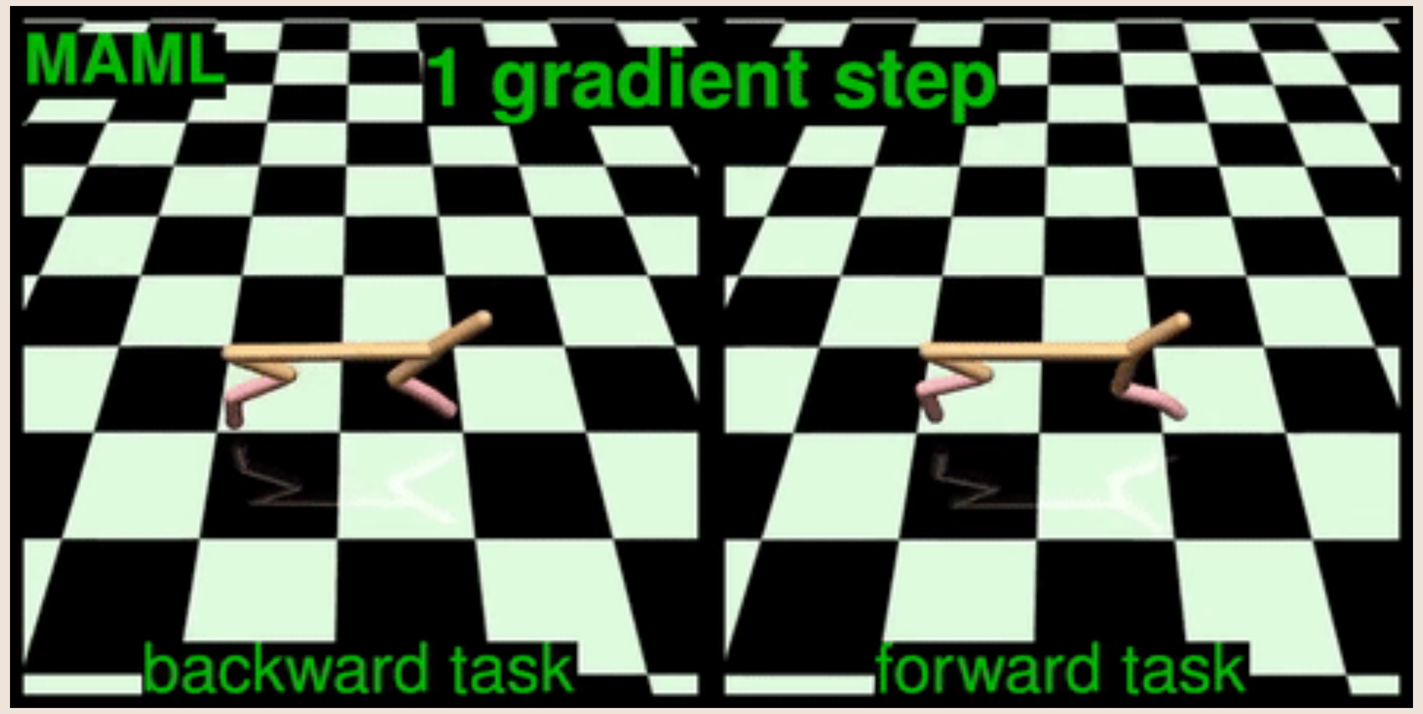
Env Distribution A

Meta Learner A
→ Learning Algorithm A



Env Distribution B

Meta Learner B
→ Learning Algorithm B

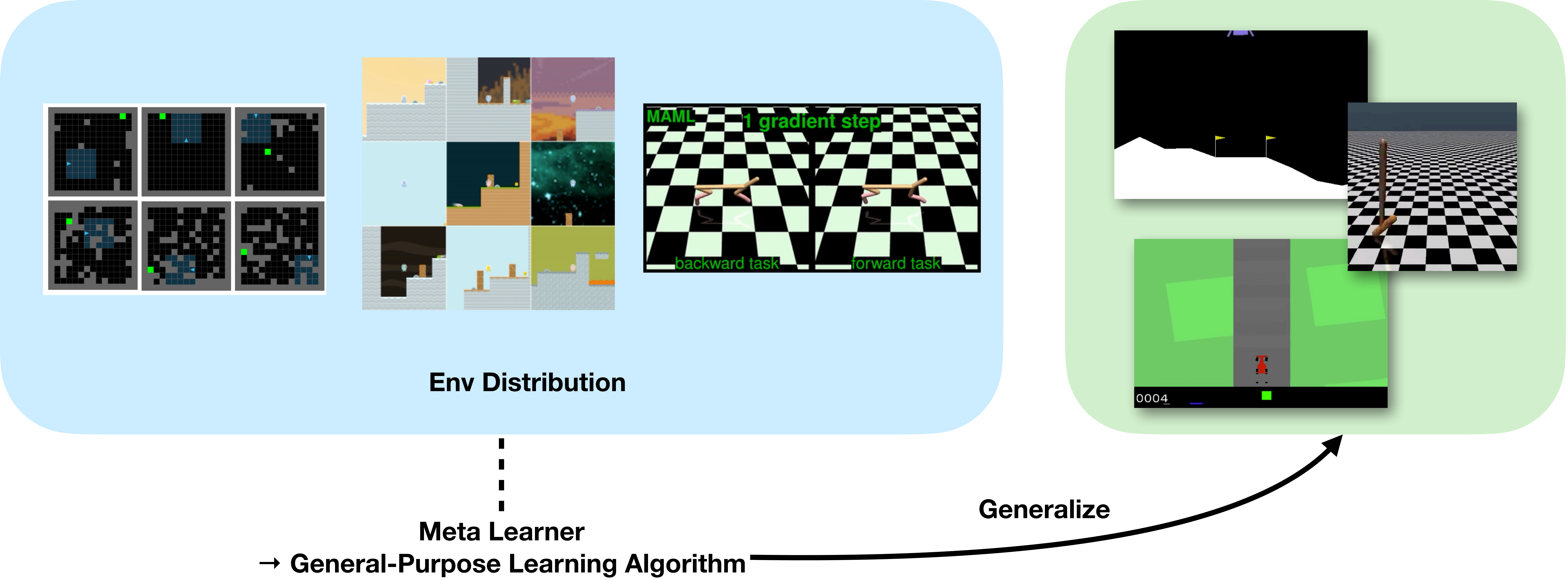


Env Distribution C

Meta Learner C
→ Learning Algorithm C

General-Purpose Meta-Learning

Can we train an agent that can efficiently in-context **learn and act in any environment**?



Env Distribution

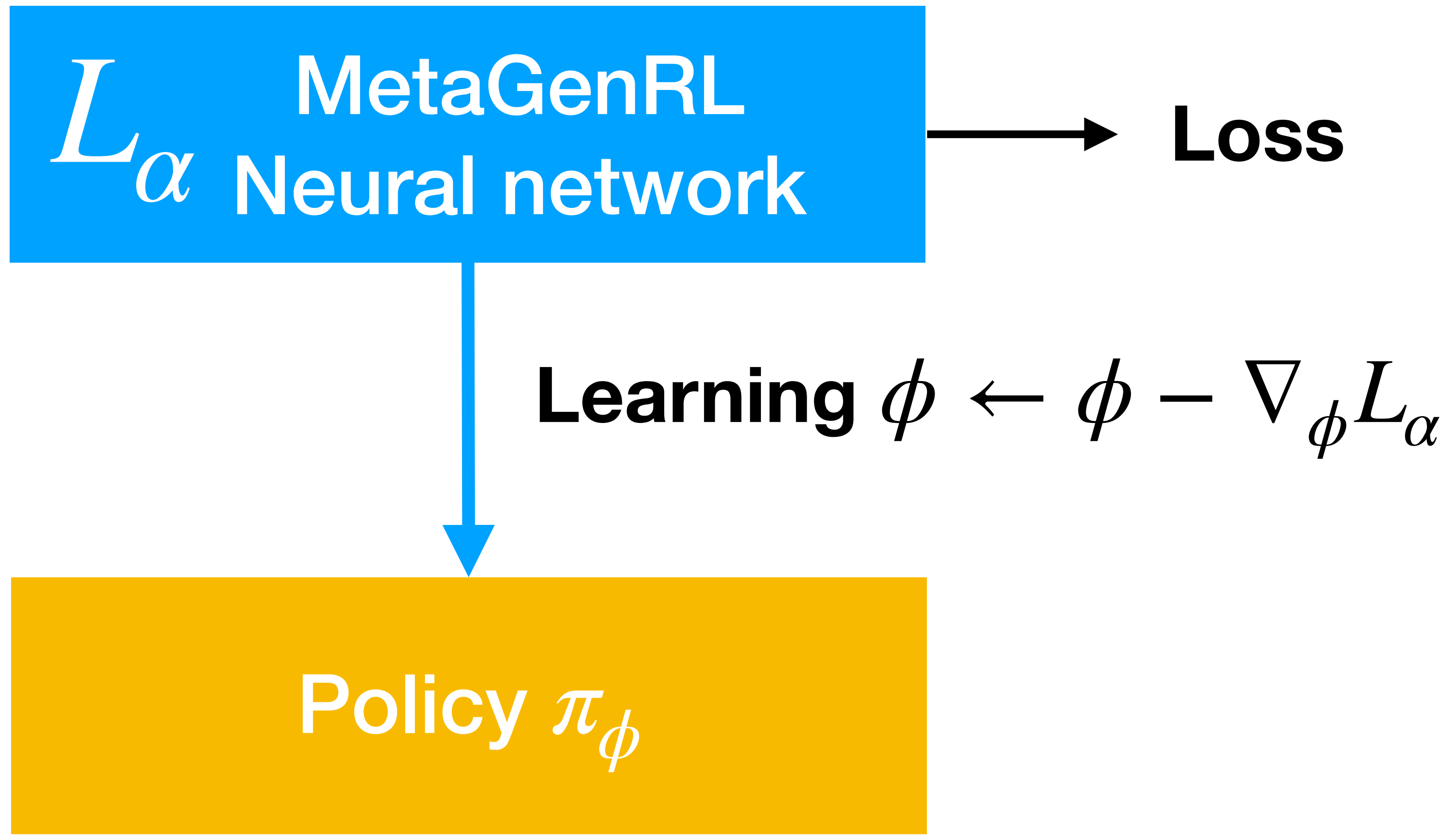
Meta Learner
→ General-Purpose Learning Algorithm

Generalize

From gradient-based learning ...

MetaGenRL

Parameterized Objective / Loss function



L_α MetaGenRL Neural network

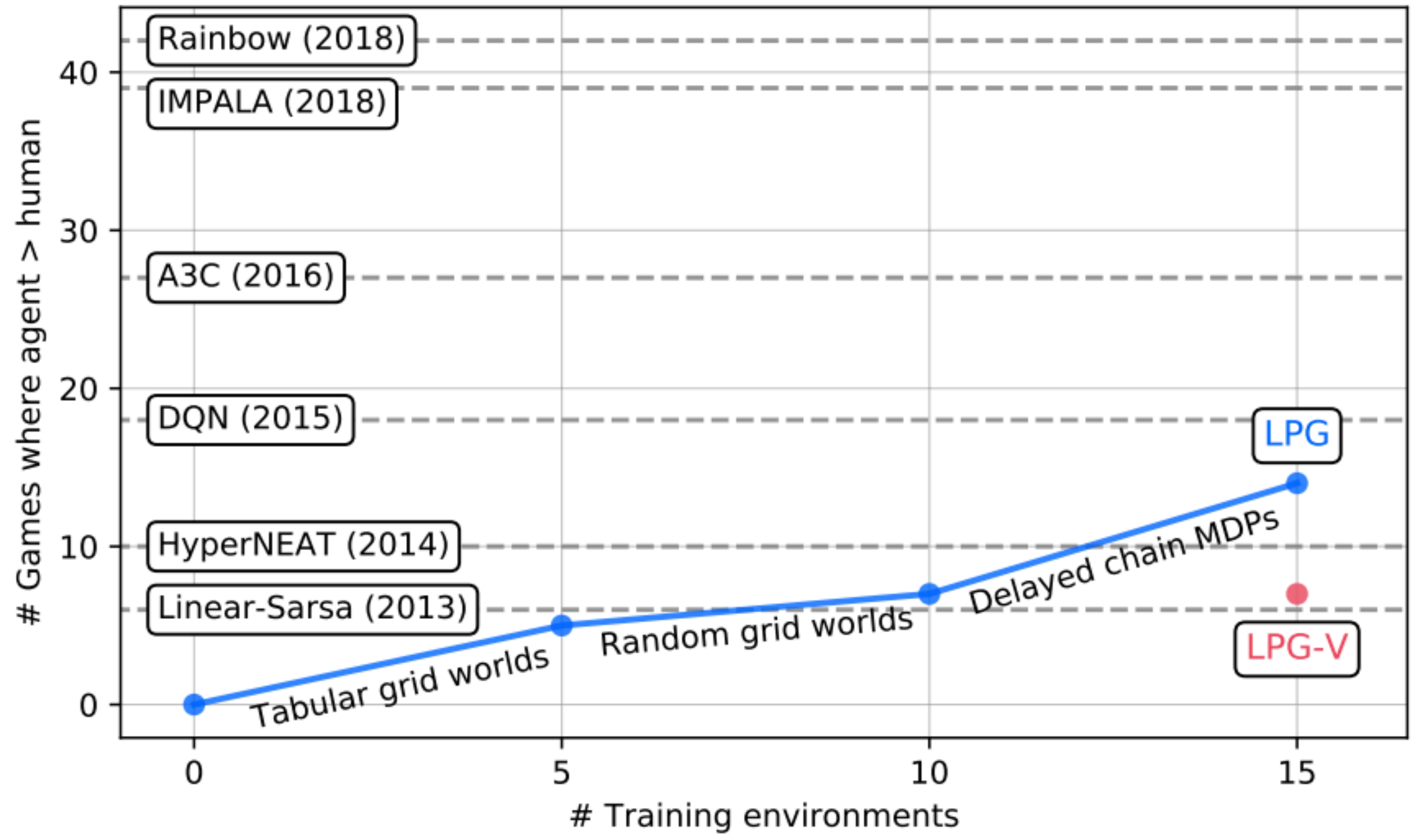
Loss

Learning $\phi \leftarrow \phi - \nabla_\phi L_\alpha$

Policy π_ϕ

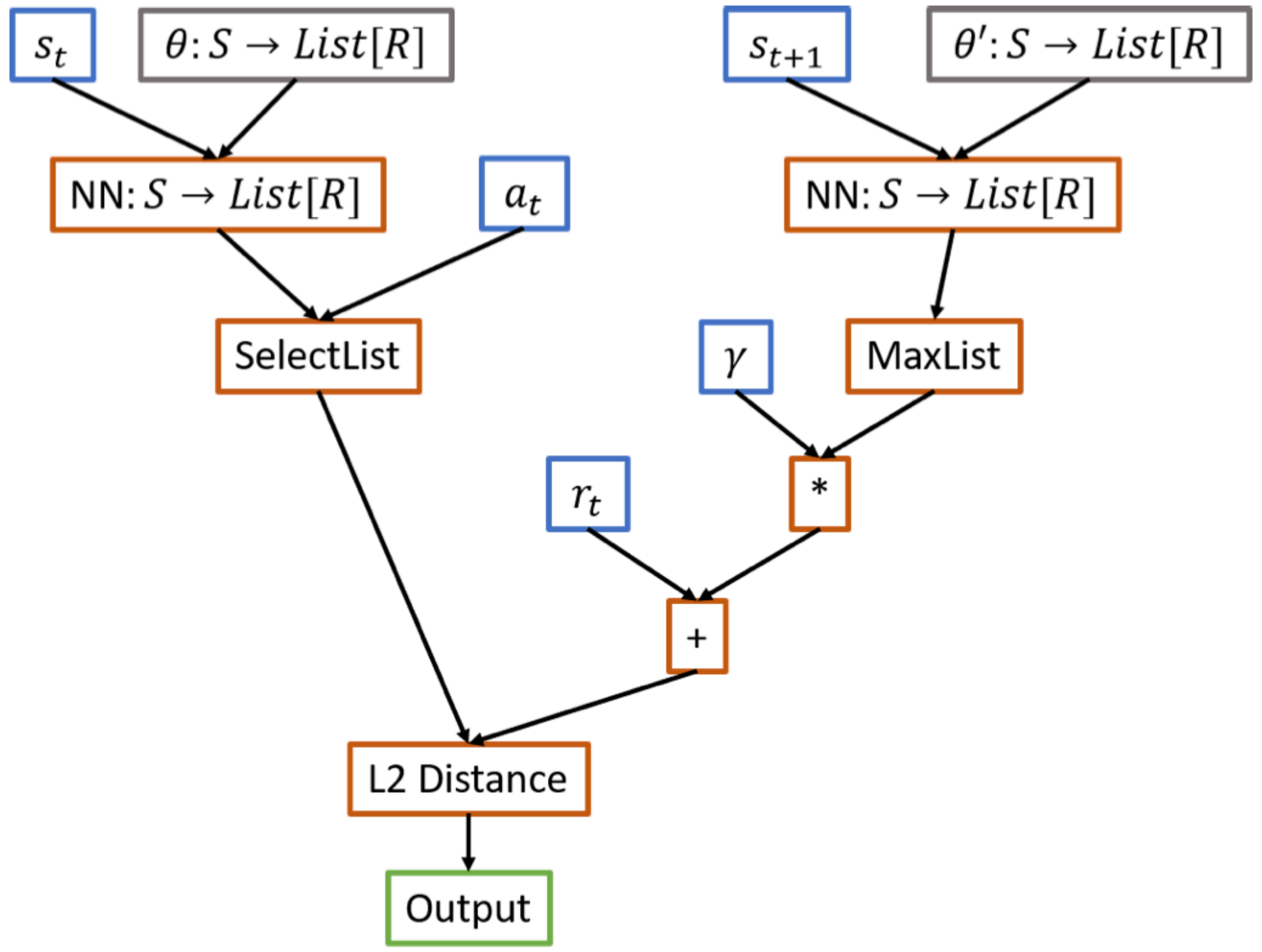
Kirsch et al [ICLR 2020]

Learned Policy Gradient



Oh et al [NeurIPS 2020]

Evolving RL Algorithms

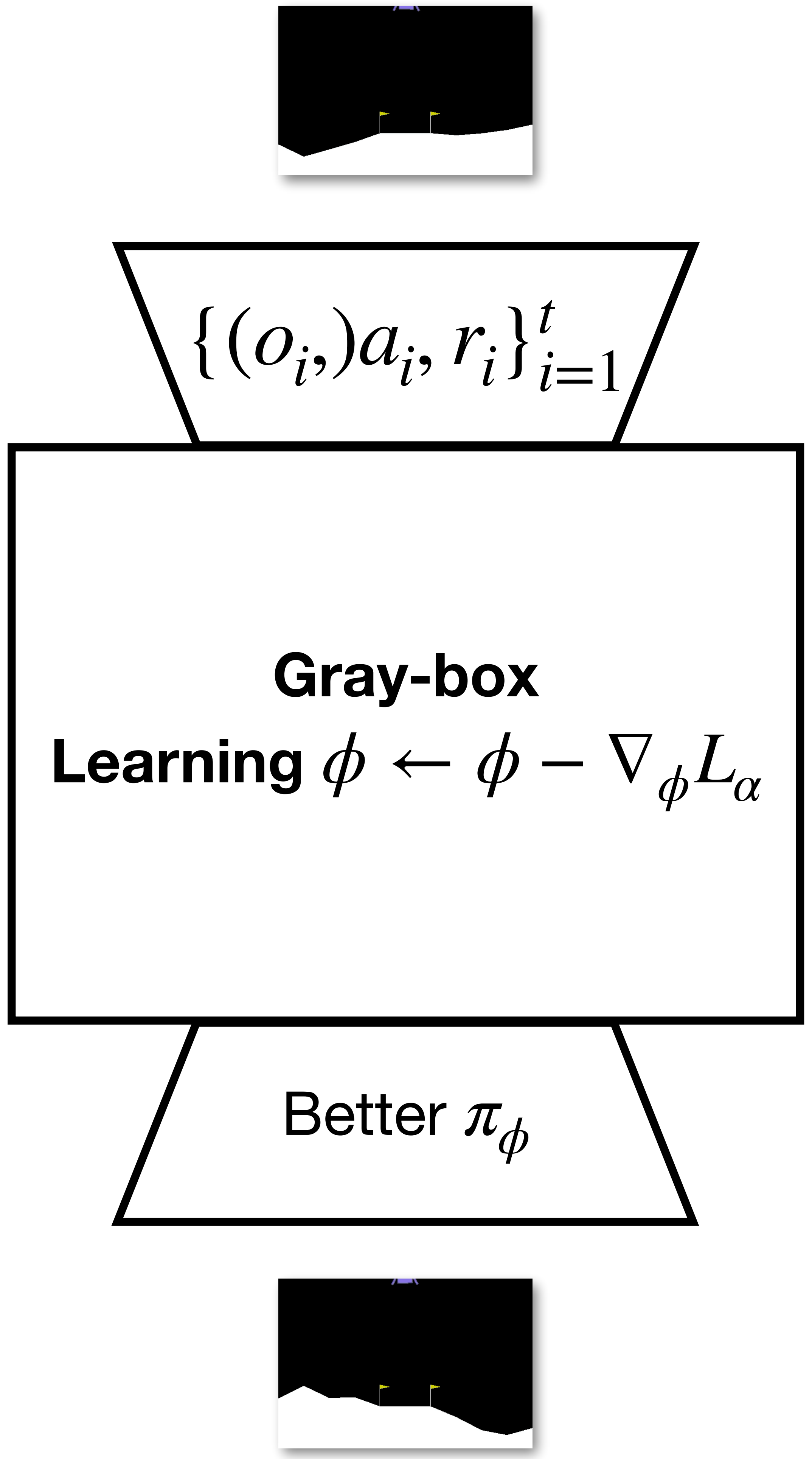

s_t
$\theta : S \rightarrow List[R]$
s_{t+1}
$\theta' : S \rightarrow List[R]$
NN: $S \rightarrow List[R]$
a_t
NN: $S \rightarrow List[R]$
SelectList
γ
MaxList
r_t
*
+
L2 Distance
Output

Co-Reyes et al [ICLR 2021]

… to in-context learning

MetaGenRL, LPG, etc

Good generalization

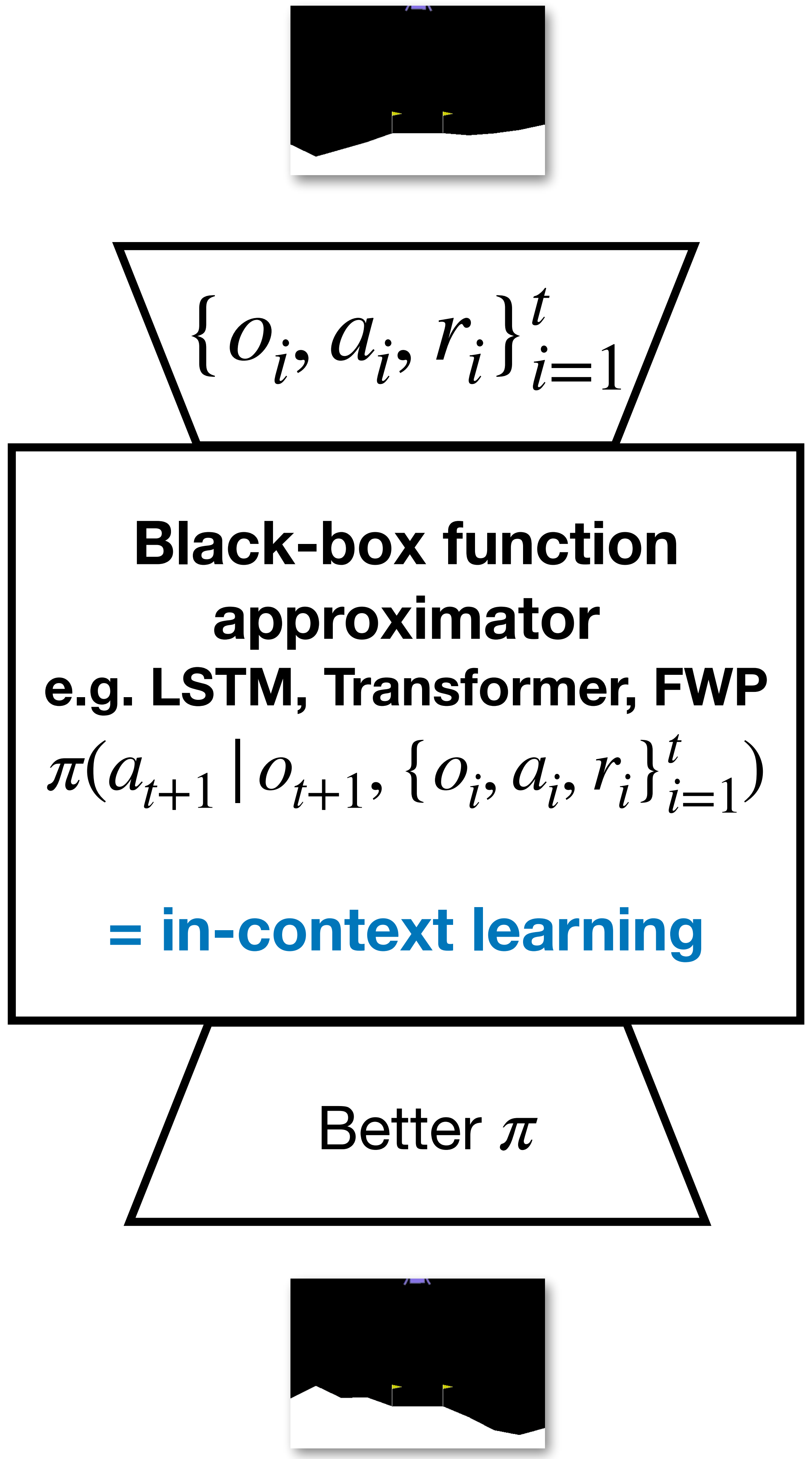


$\{(o_i,)a_i, r_i\}_{i=1}^t$

Gray-box
Learning $\phi \leftarrow \phi - \nabla_\phi L_\alpha$

Better π_ϕ

In-context meta-RL



$\{o_i, a_i, r_i\}_{i=1}^t$

Black-box function approximator
e.g. LSTM, Transformer, FWP
$\pi(a_{t+1} \,|\, o_{t+1}, \{o_i, a_i, r_i\}_{i=1}^t)$

= in-context learning

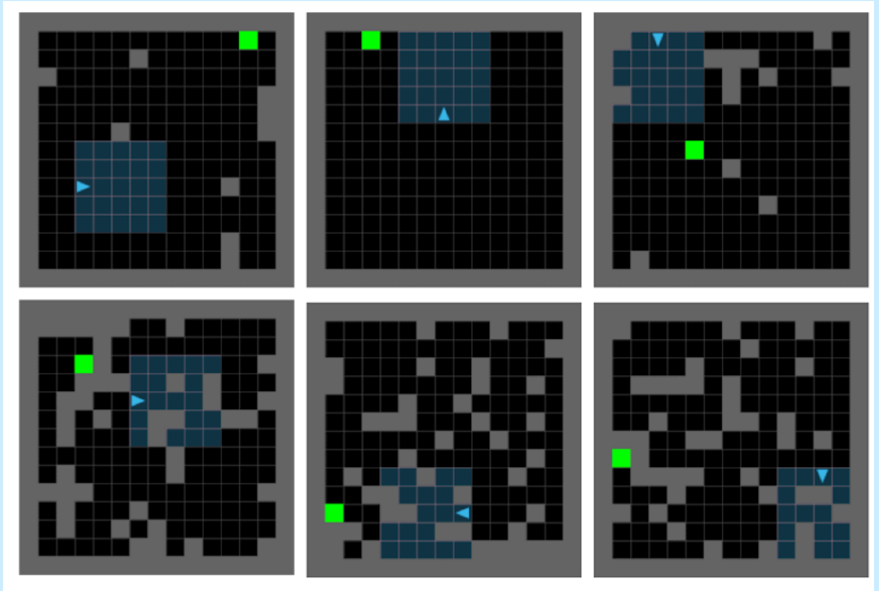
Better π

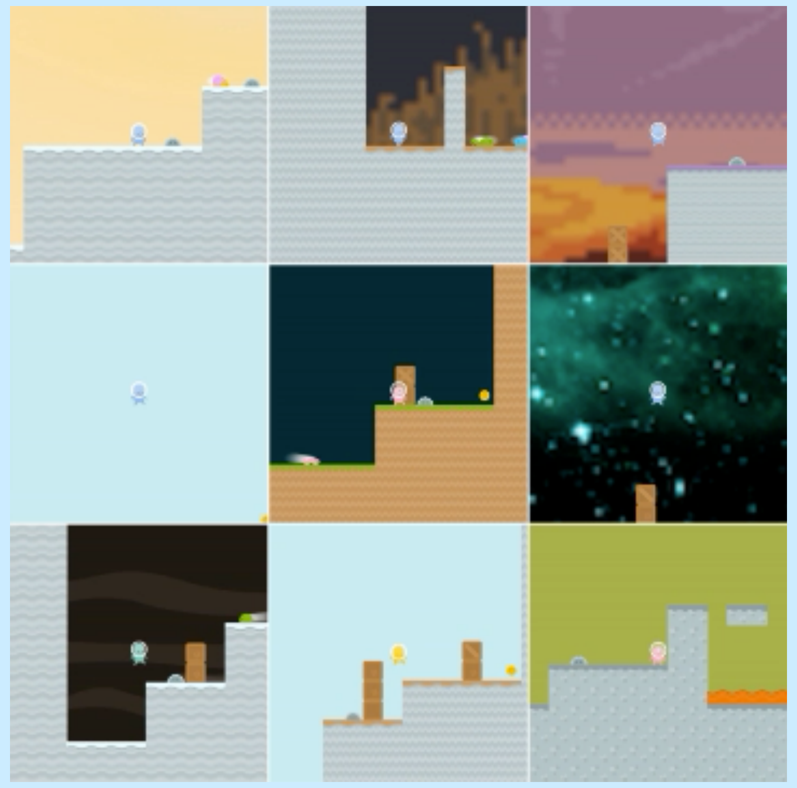
[Schmidhuber since 1992,
RL^2 Duan et al 2016,
Wang et al 2016,
SymLA Kirsch et al 2022,
Adaptive Agents 2023, etc]

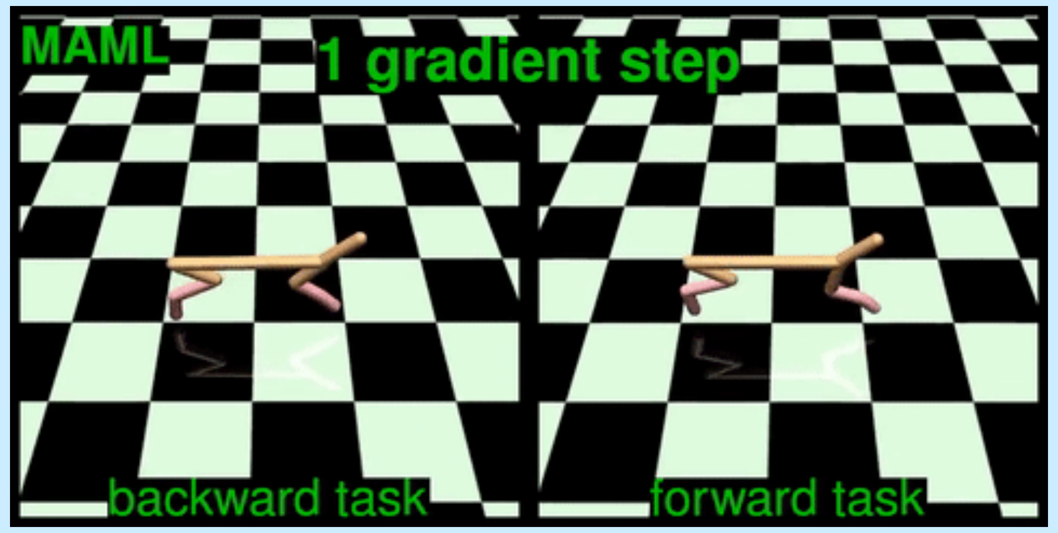
Difficult generalization

This paper:
How to fix this?

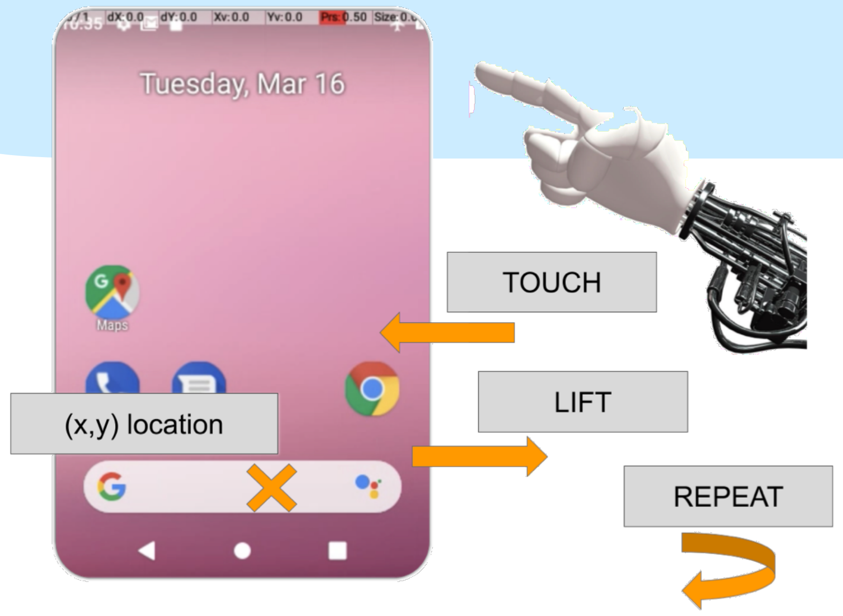
Automated task / environment generation

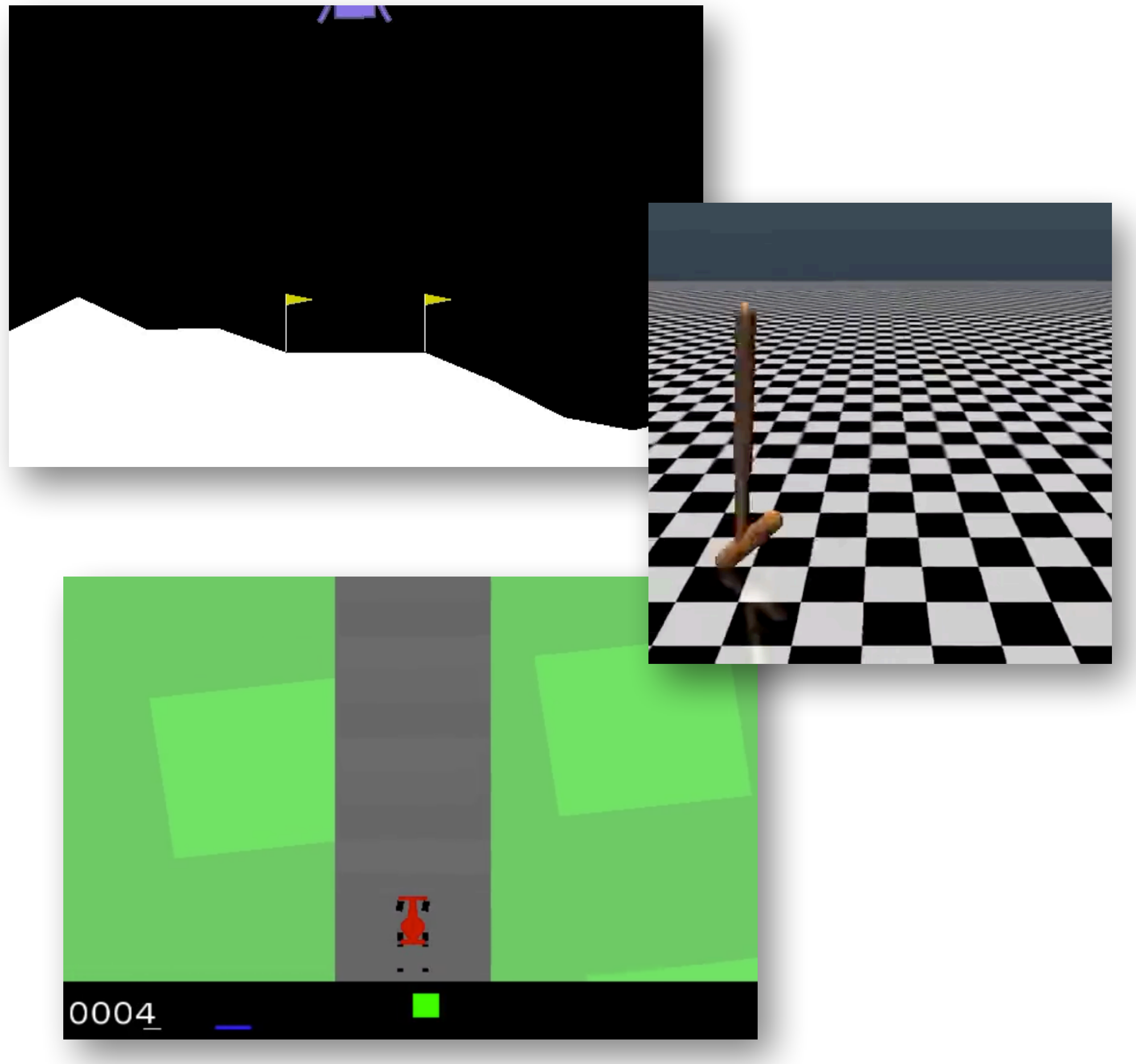




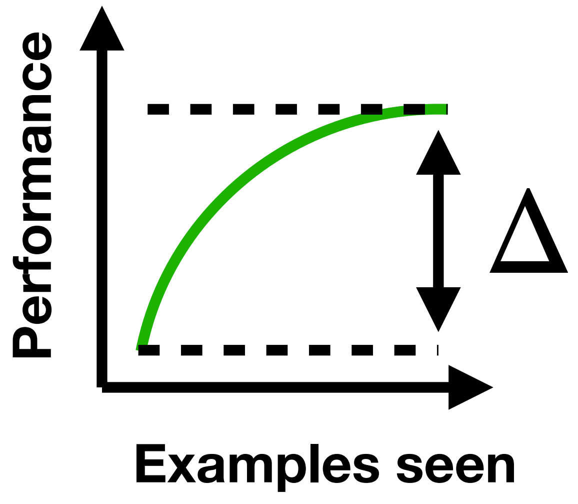


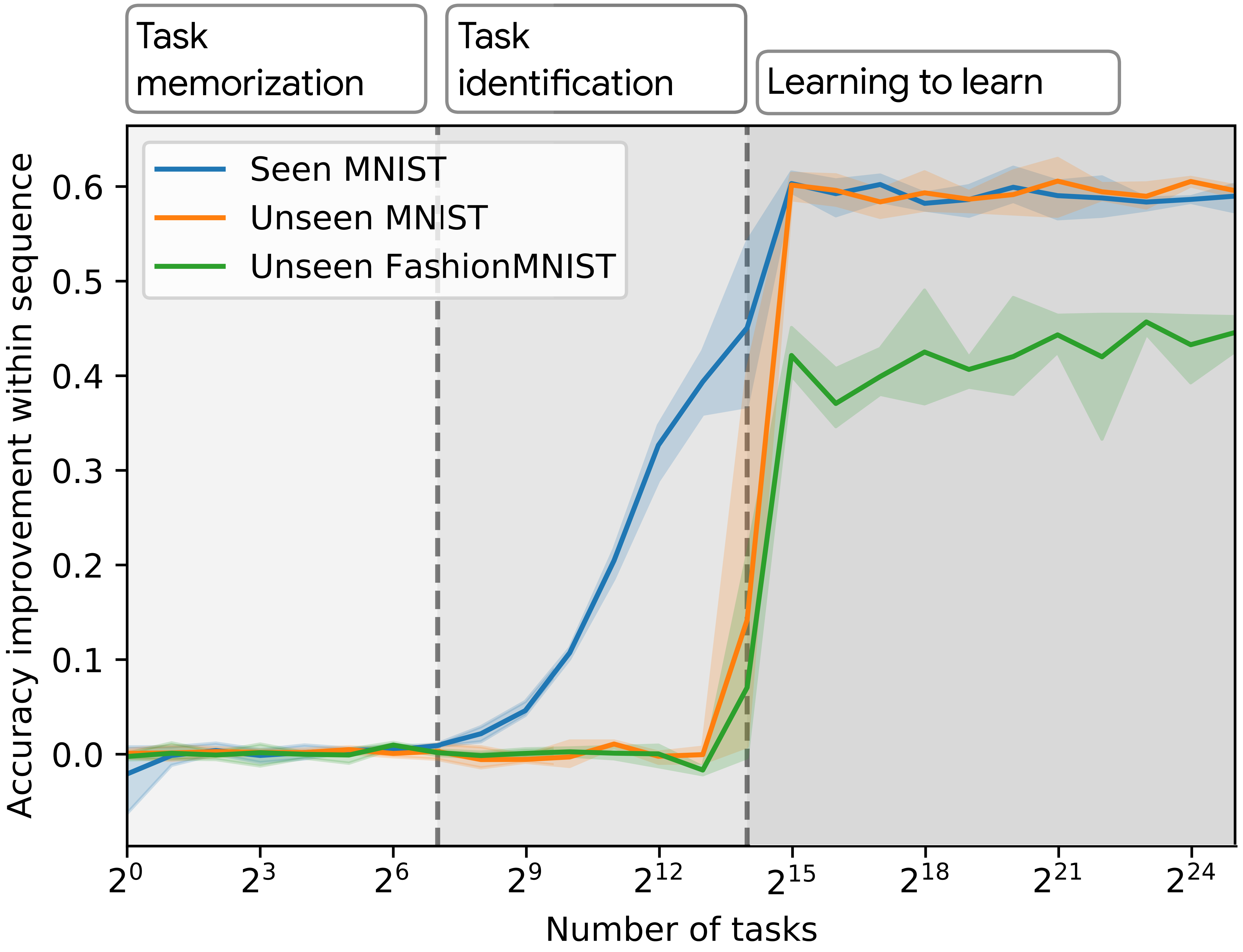
Env Distribution





GPICL: From memorization to general learn-to-learn





Learning	Generalization	Algorithm Description
× No	× No	Task memorization
✓ Yes	× No	Task identification
✓ Yes	✓ Yes	General-purpose learning algorithm

Transformers exhibit three different phases in terms of meta-learned behavior.

Kirsch et al [2022]

From classification to supervised meta-RL

Adaptation to RL

Instead of training on a supervised dataset $\left(\{x_i, y_i\}_{i=1}^{N_D}, x'\right) \mapsto y'$

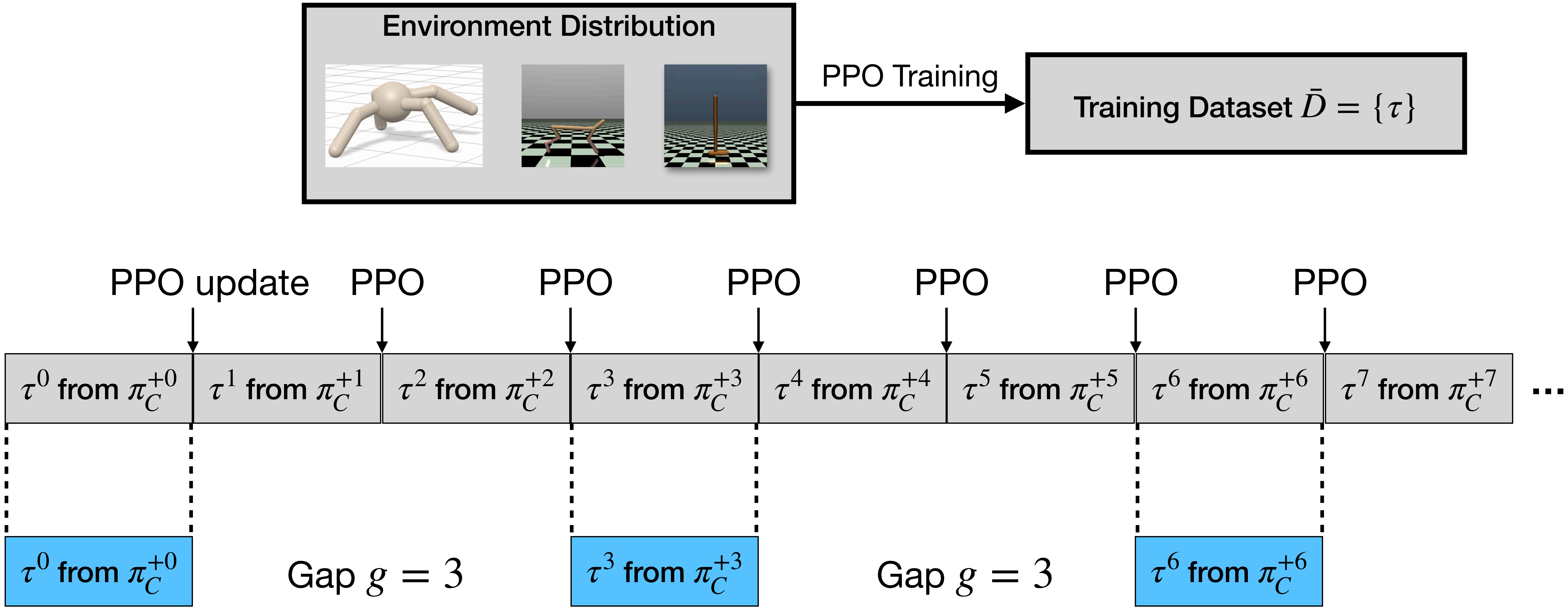
Train on RL data $\left(\{s_i, a_i, r_i, d_i\}_{i=1}^{N_D}, s\right) \mapsto a$

Our recipe: **Offline meta-training + augmented RL data → Generalization**

Training is offline, but agent can learn online!

How it works - PPO Data collection

Data Collection



Environment Distribution

PPO Training

Training Dataset $\bar{D} = \{\tau\}$

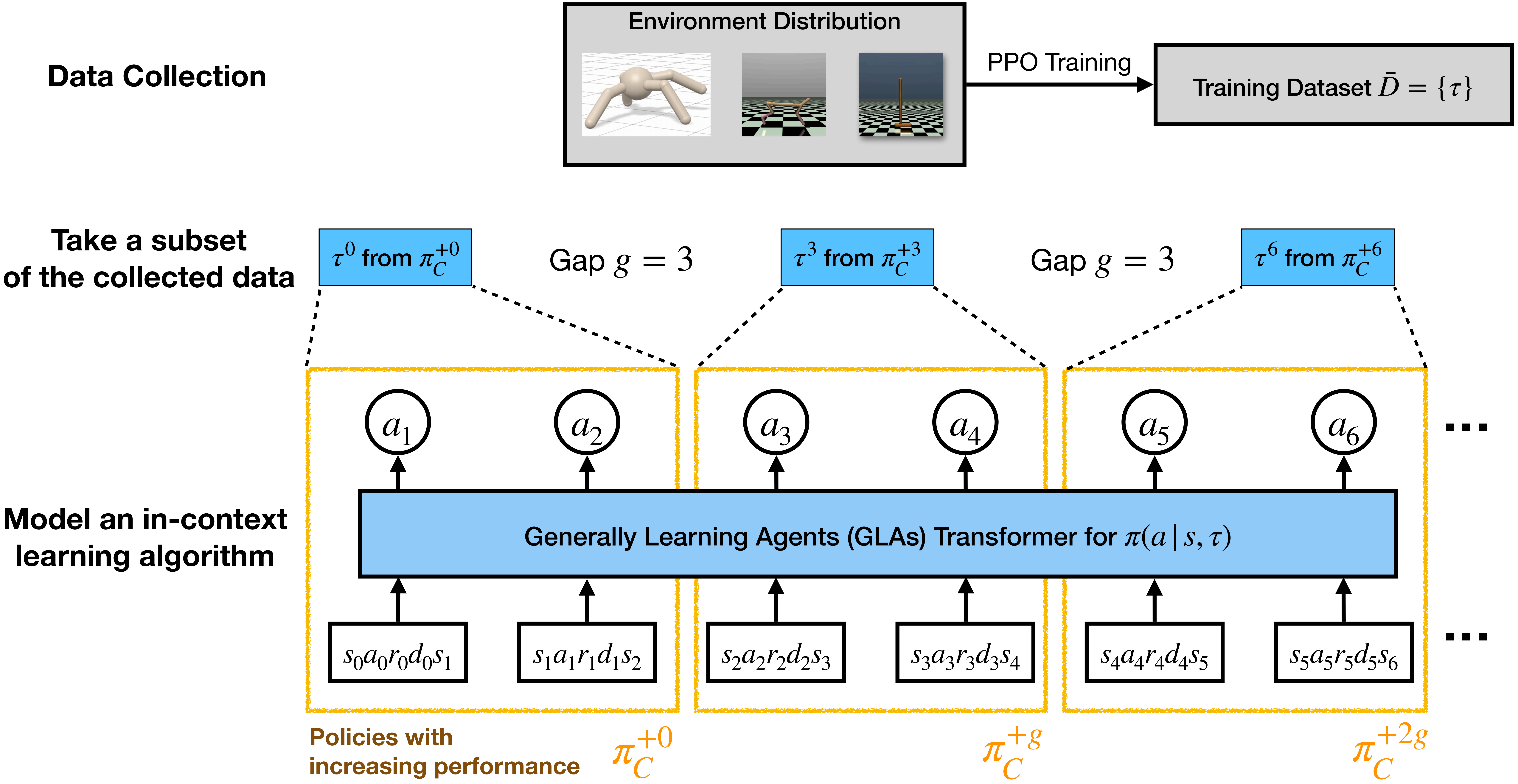
PPO update PPO PPO PPO PPO PPO PPO

τ^0 from π_C^{+0} | τ^1 from π_C^{+1} | τ^2 from π_C^{+2} | τ^3 from π_C^{+3} | τ^4 from π_C^{+4} | τ^5 from π_C^{+5} | τ^6 from π_C^{+6} | τ^7 from π_C^{+7} ...

Take a subset of the collected data

τ^0 from π_C^{+0} Gap $g = 3$ τ^3 from π_C^{+3} Gap $g = 3$ τ^6 from π_C^{+6}

How it works - Model learning algorithm



Data Collection

Environment Distribution

PPO Training

Training Dataset $\bar{D} = \{\tau\}$

Take a subset of the collected data

τ^0 from π_C^{+0}

Gap $g = 3$

τ^3 from π_C^{+3}

Gap $g = 3$

τ^6 from π_C^{+6}

Model an in-context learning algorithm

a_1 a_2 a_3 a_4 a_5 a_6 ...

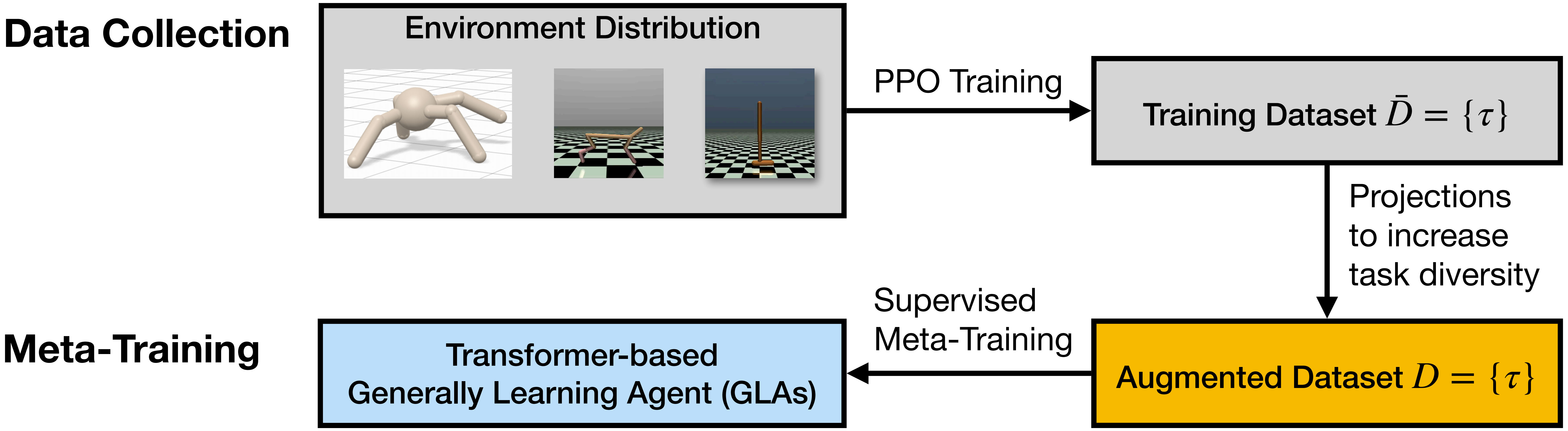
Generally Learning Agents (GLAs) Transformer for $\pi(a \mid s, \tau)$

$s_0a_0r_0d_0s_1$ $s_1a_1r_1d_1s_2$ $s_2a_2r_2d_2s_3$ $s_3a_3r_3d_3s_4$ $s_4a_4r_4d_4s_5$ $s_5a_5r_5d_5s_6$ ...

Policies with increasing performance

π_C^{+0} π_C^{+g} π_C^{+2g}

How it works - Augmentations



Data Collection

Environment Distribution

PPO Training

Training Dataset $\bar{D} = \{\tau\}$

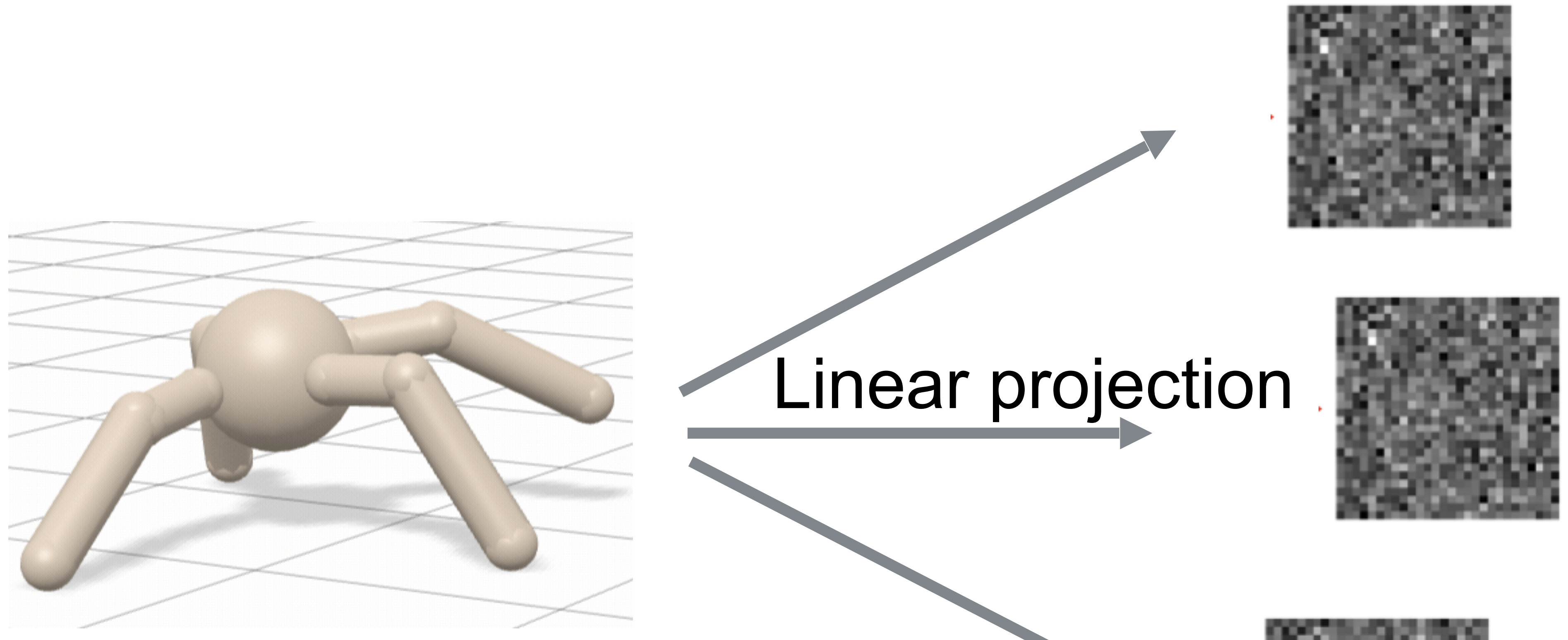
Projections to increase task diversity

Meta-Training

Transformer-based Generally Learning Agent (GLAs)

Supervised Meta-Training

Augmented Dataset $D = \{\tau\}$



Sensor observations & actions

Transition dataset

$$\bar{D} = \{s_i, a_i, r_i, d_i\}$$

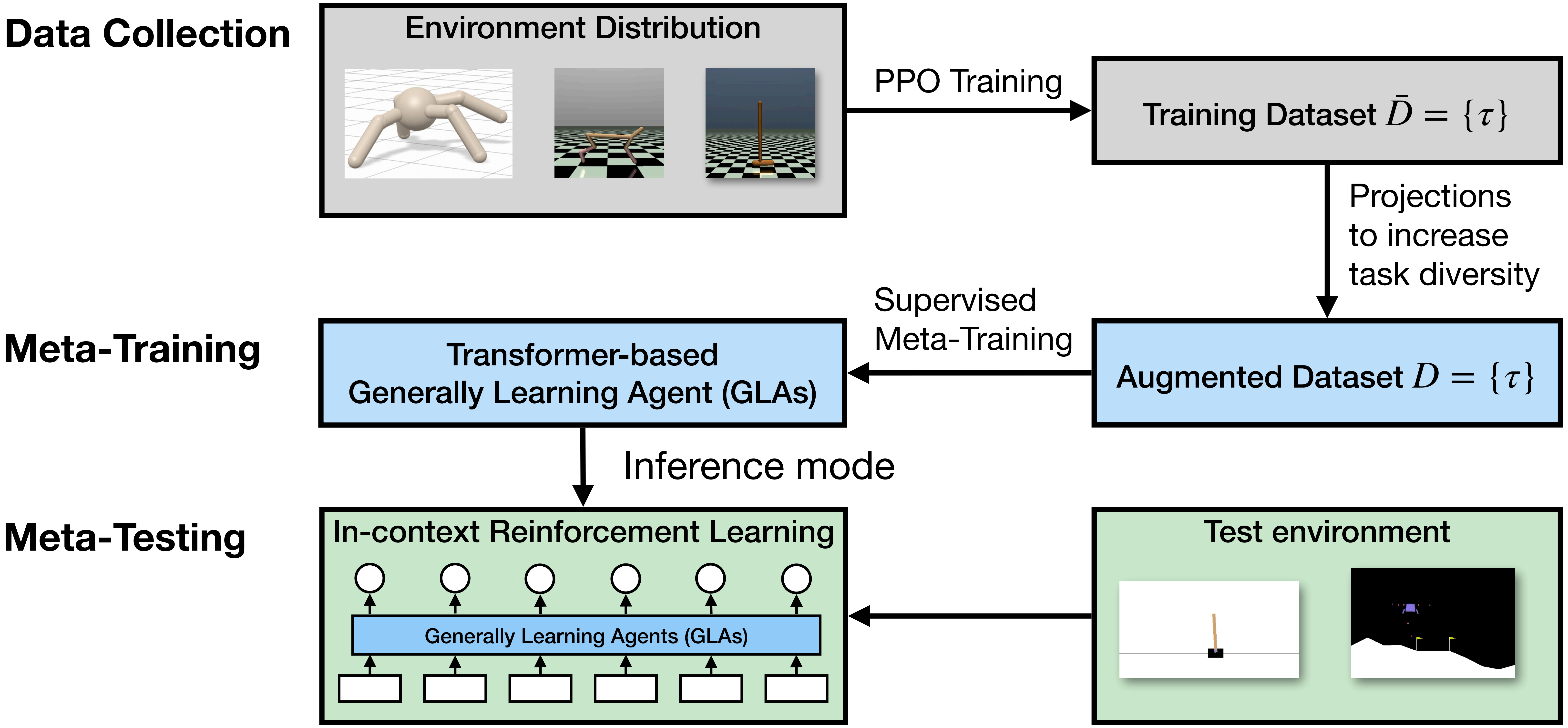
Create n tasks $\{\mathcal{S}_j, \mathcal{A}_j\}_{j=1}^{n}$

$$D_j = \{\mathcal{S}_j s_i, \mathcal{A}_j a_i, r_i, d_i\}$$

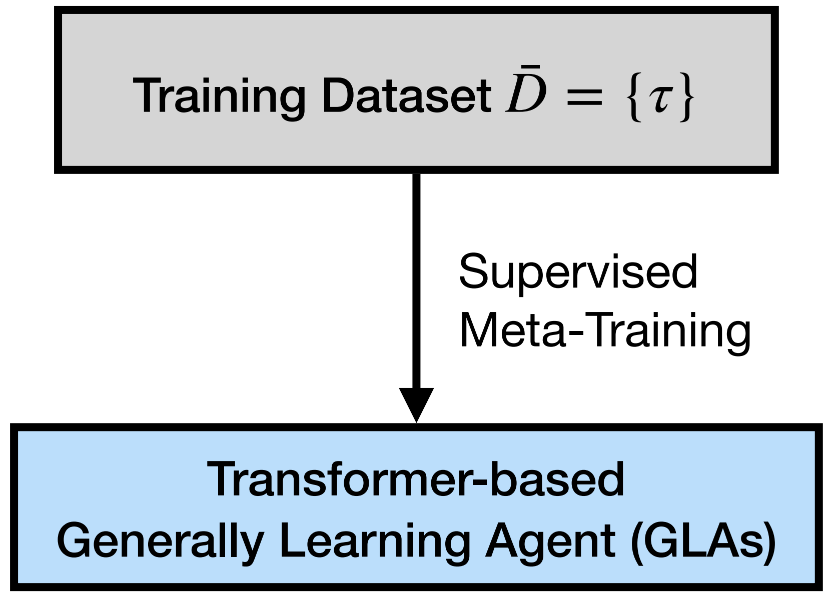
Linear observation projection with $\mathcal{S}_j$

Linear action projection with $\mathcal{A}_j$

How it works - Meta-Testing



Evaluation: Single Environment

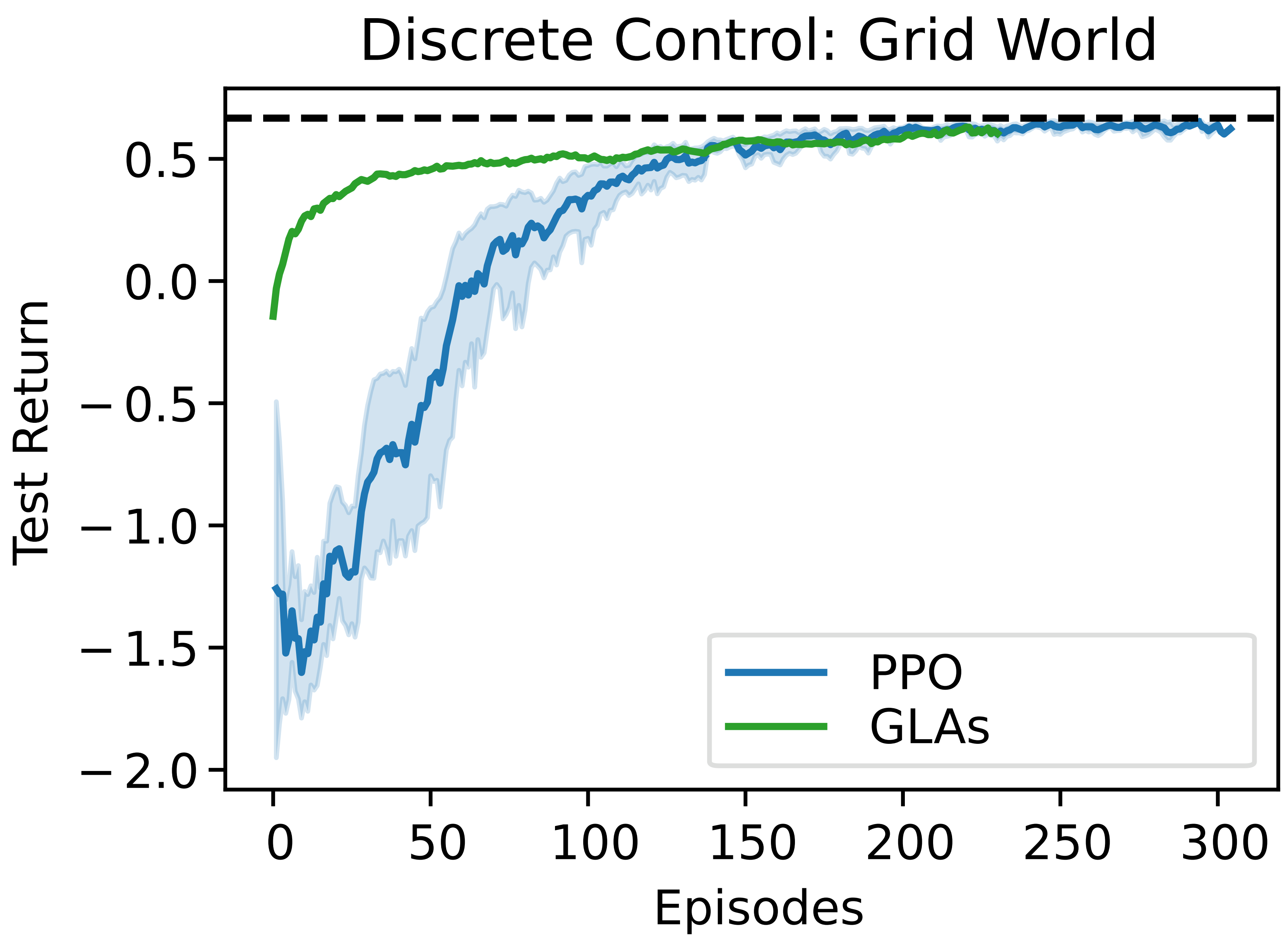


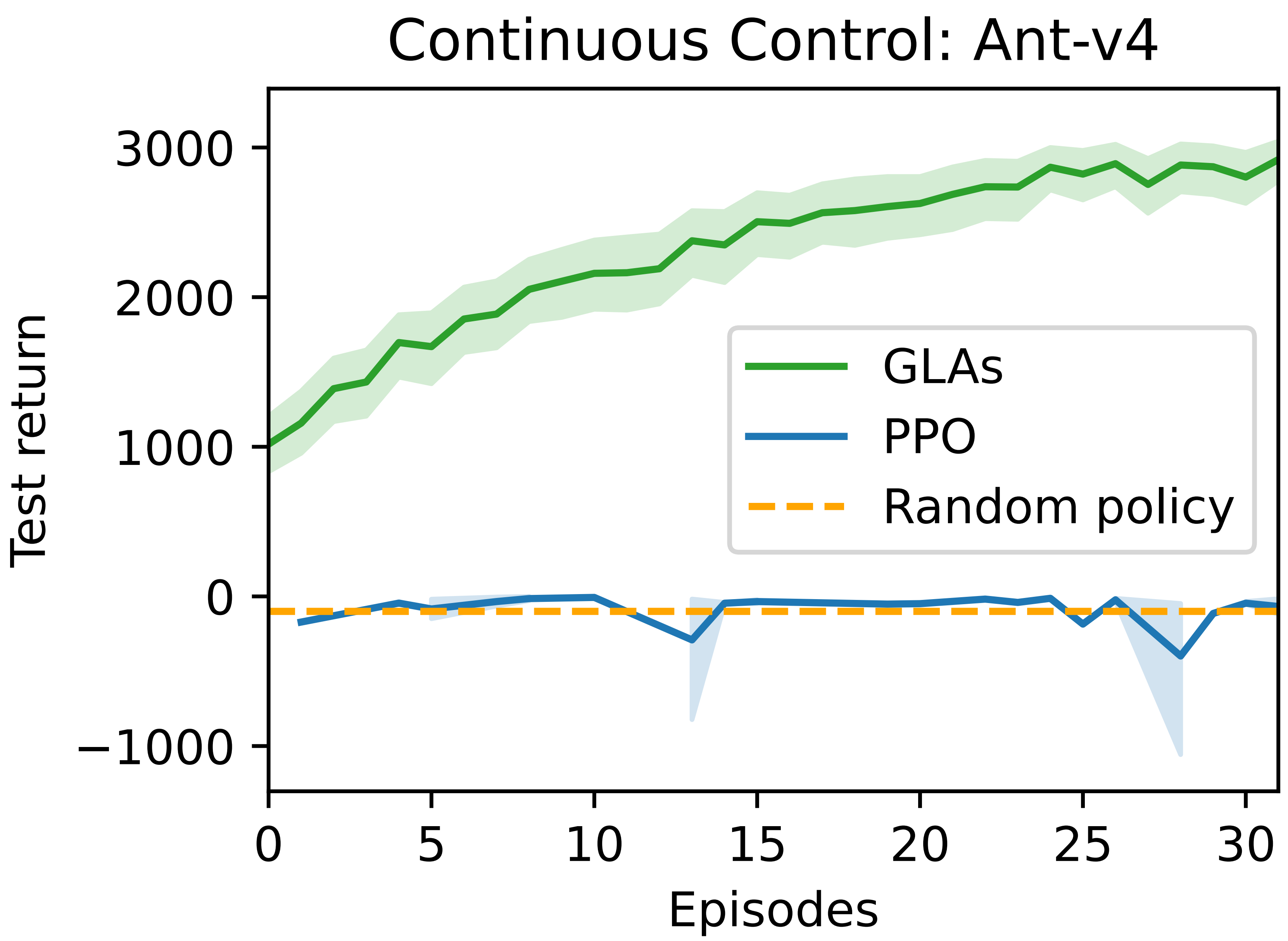
Training Dataset $\bar{D} = \{\tau\}$

Supervised Meta-Training

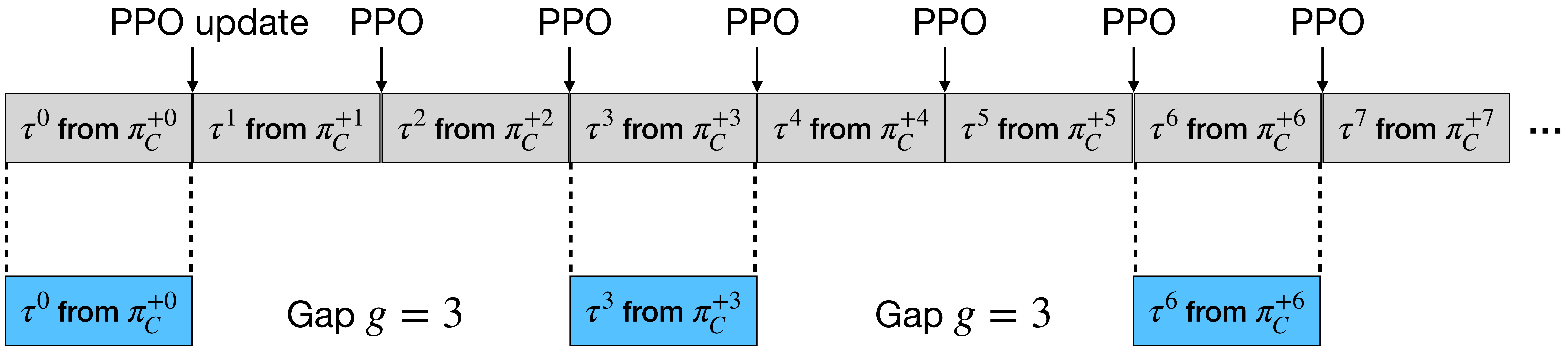
Transformer-based Generally Learning Agent (GLAs)

In-context learning policies that encode a learning algorithm on a specific task





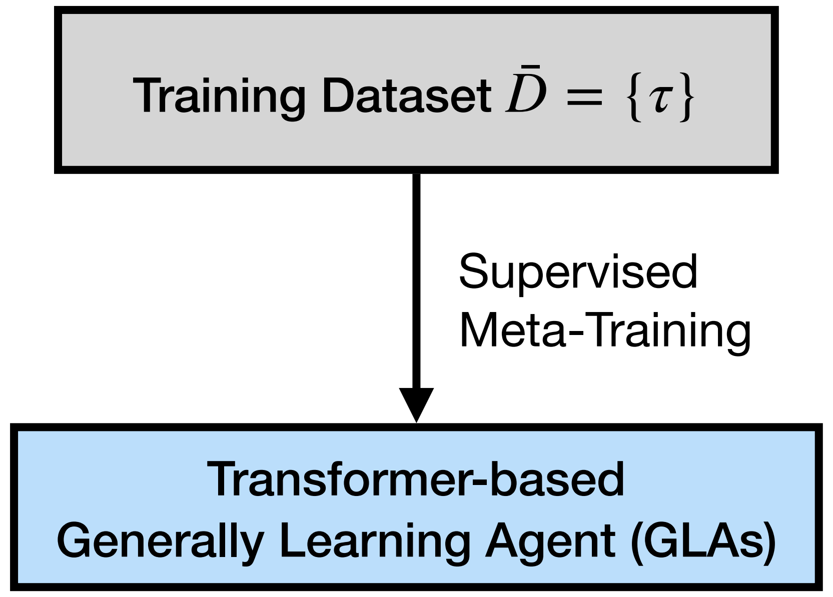
Effect of the gap



PPO update PPO PPO PPO PPO PPO PPO

τ^0 from π_C^{+0} | τ^1 from π_C^{+1} | τ^2 from π_C^{+2} | τ^3 from π_C^{+3} | τ^4 from π_C^{+4} | τ^5 from π_C^{+5} | τ^6 from π_C^{+6} | τ^7 from π_C^{+7} ...

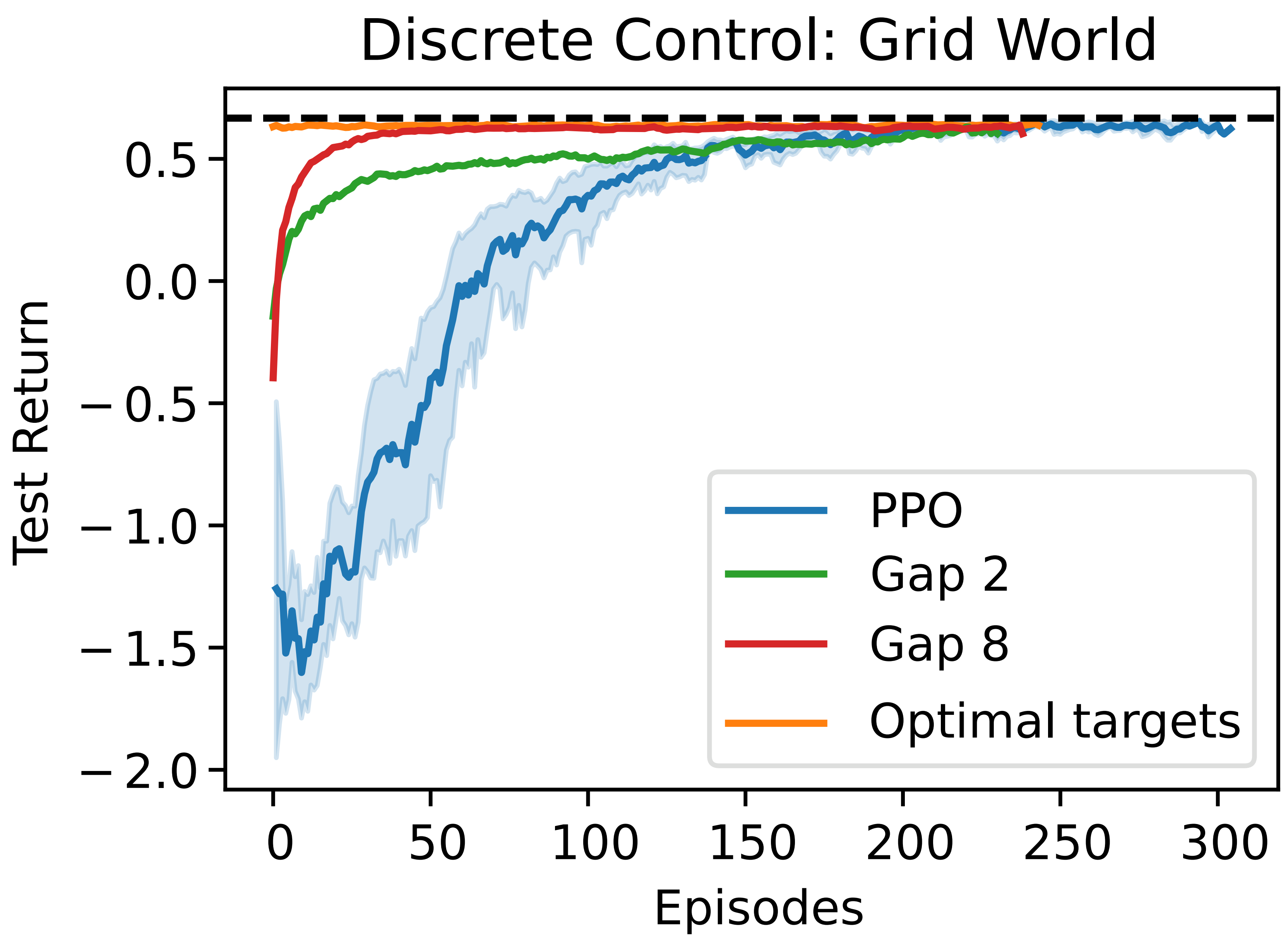
τ^0 from π_C^{+0} Gap $g = 3$ τ^3 from π_C^{+3} Gap $g = 3$ τ^6 from π_C^{+6}



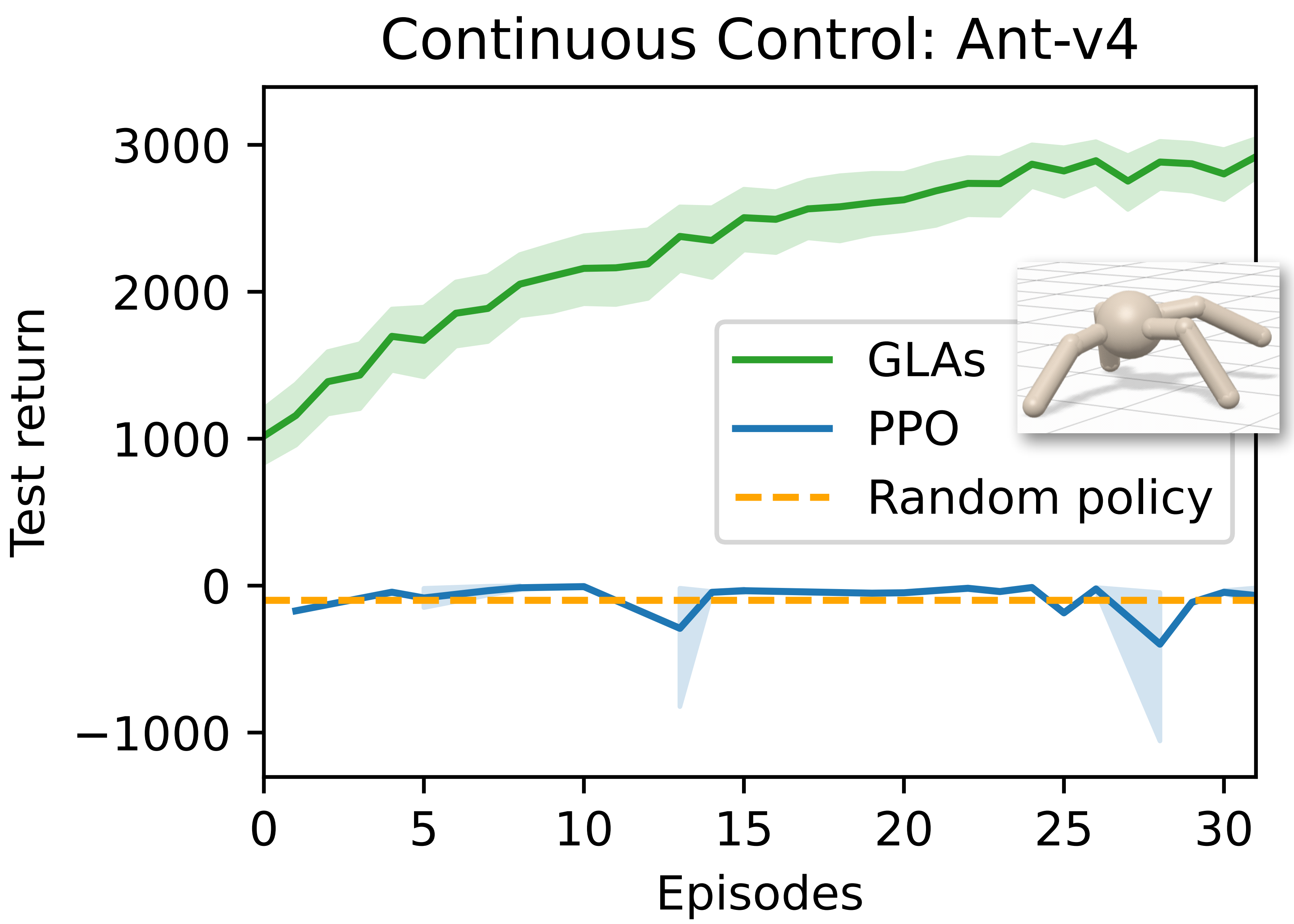
Training Dataset $\bar{D} = \{\tau\}$

Supervised Meta-Training

Transformer-based Generally Learning Agent (GLAs)

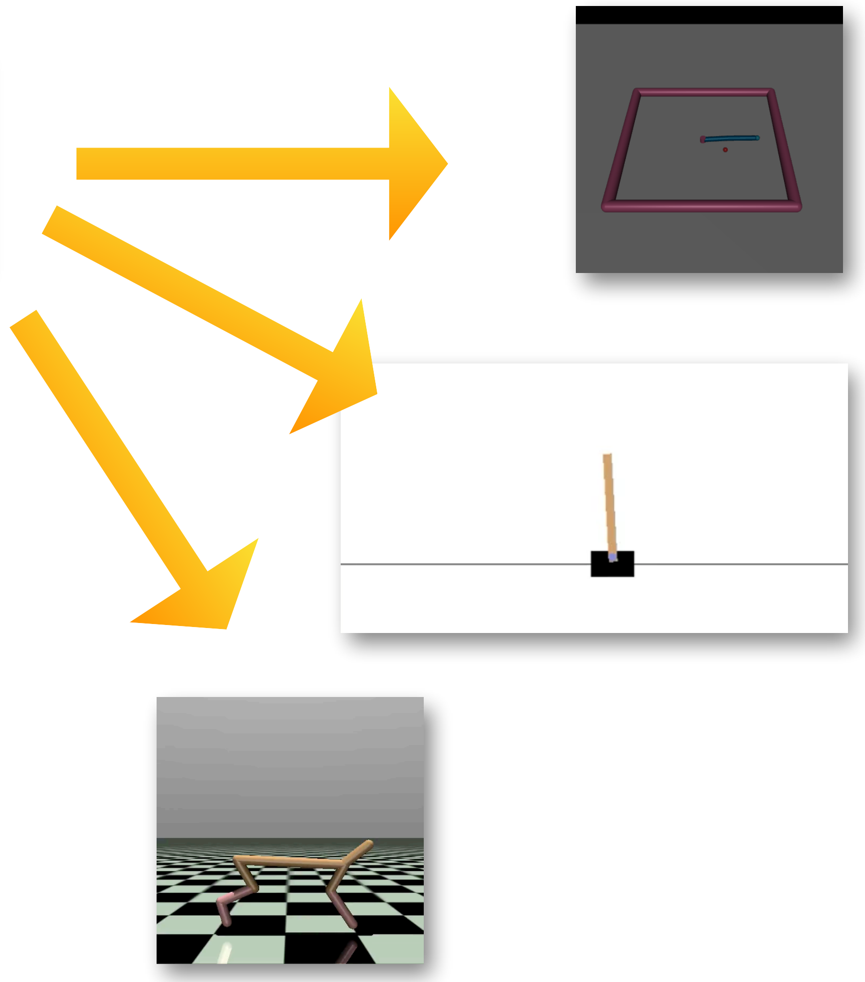


Discrete Control: Grid World



Continuous Control: Ant-v4

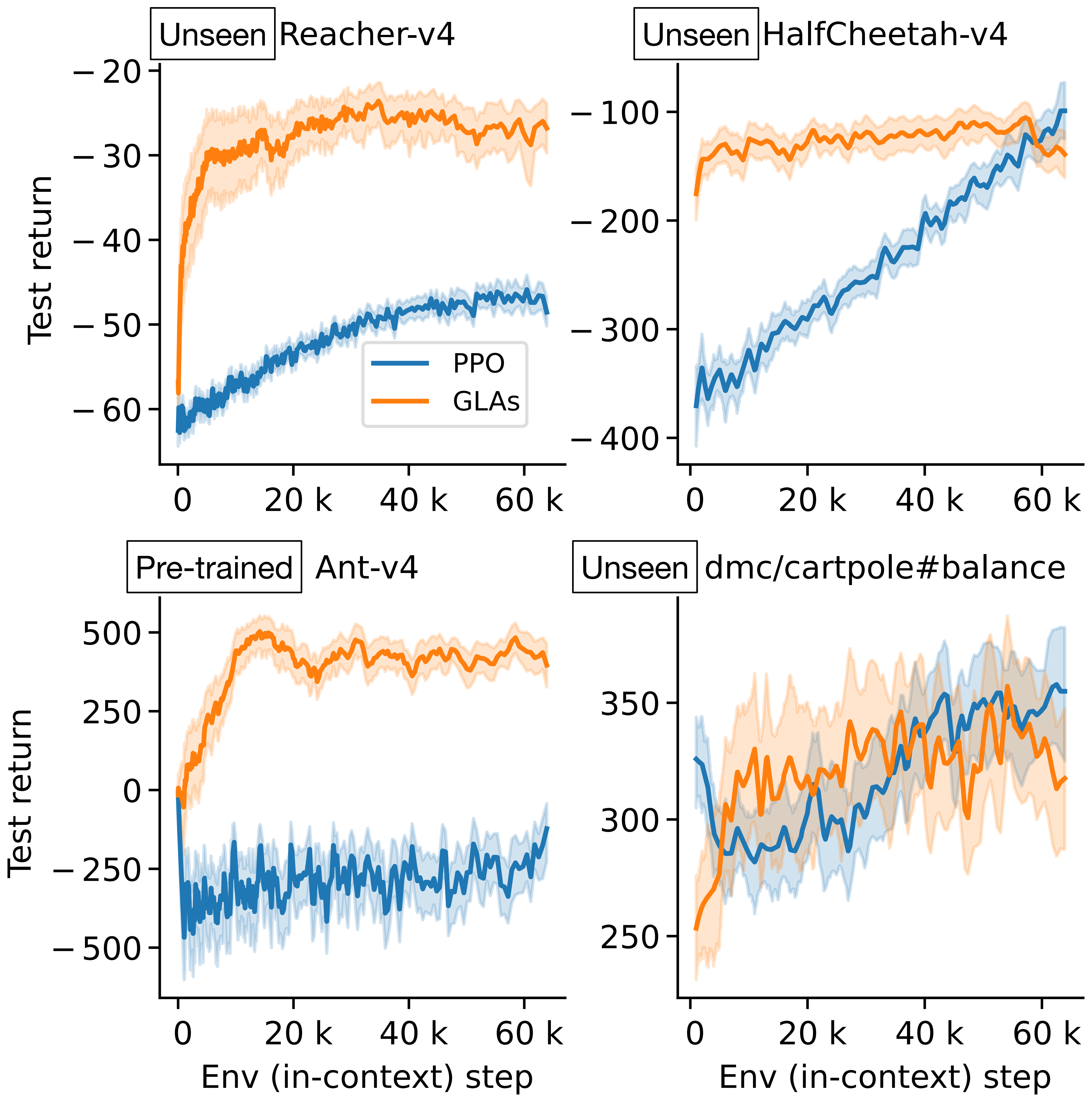
Generalize across environments?

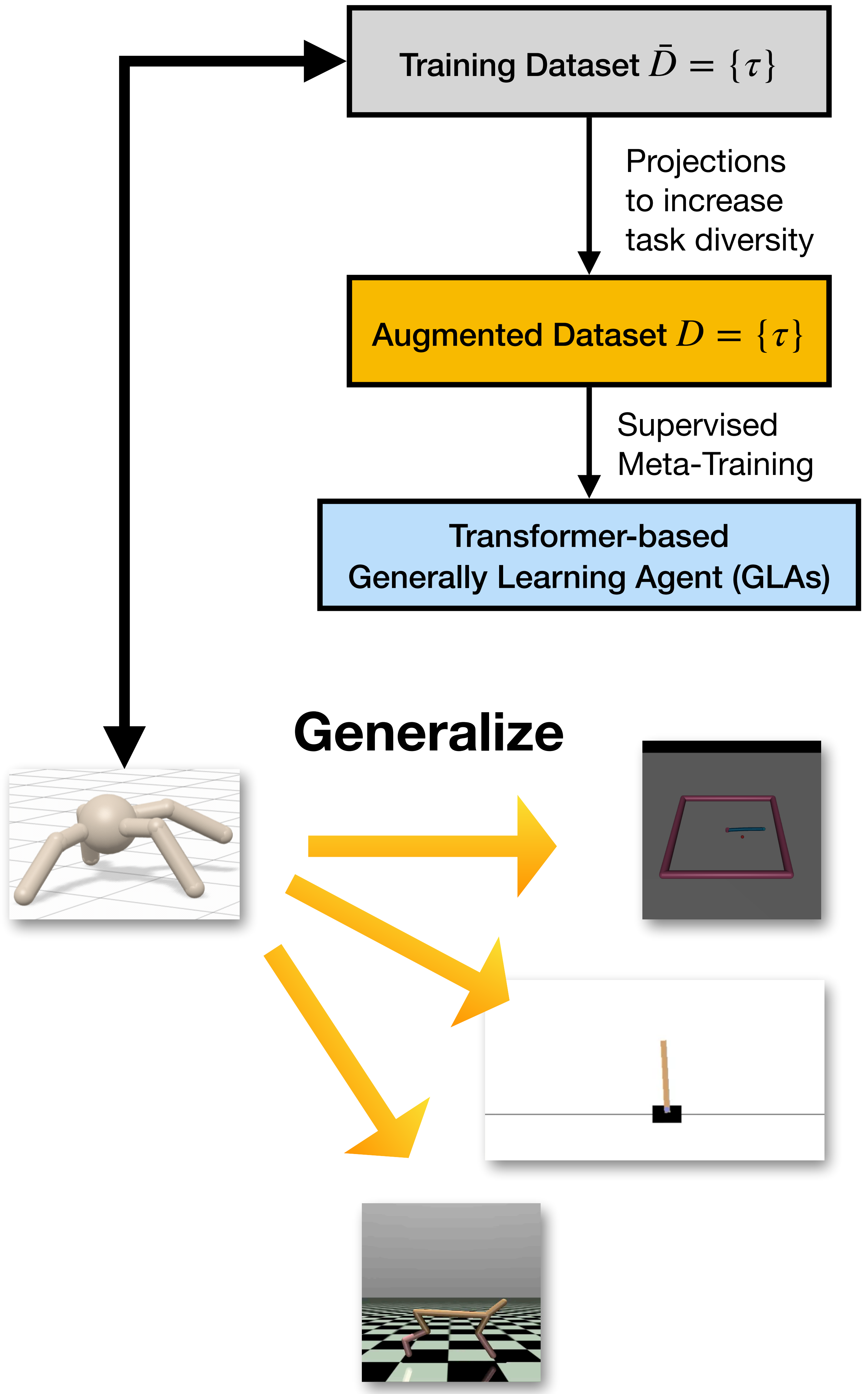


Introducing random projections

Generalize across environments?

Early results!
*** Longer contexts**
*** Training dynamics**





Training Dataset $\bar{D} = \{\tau\}$

Projections to increase task diversity

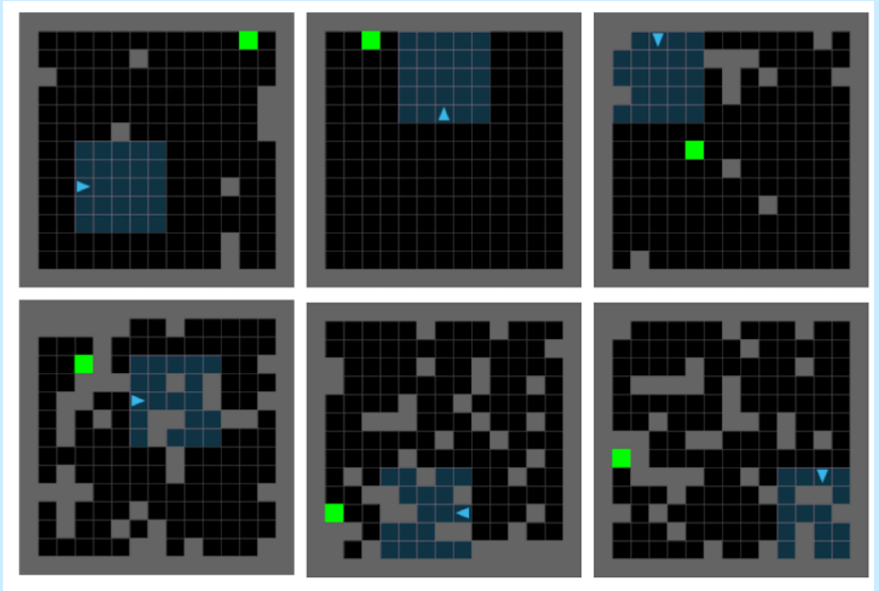
Augmented Dataset $D = \{\tau\}$

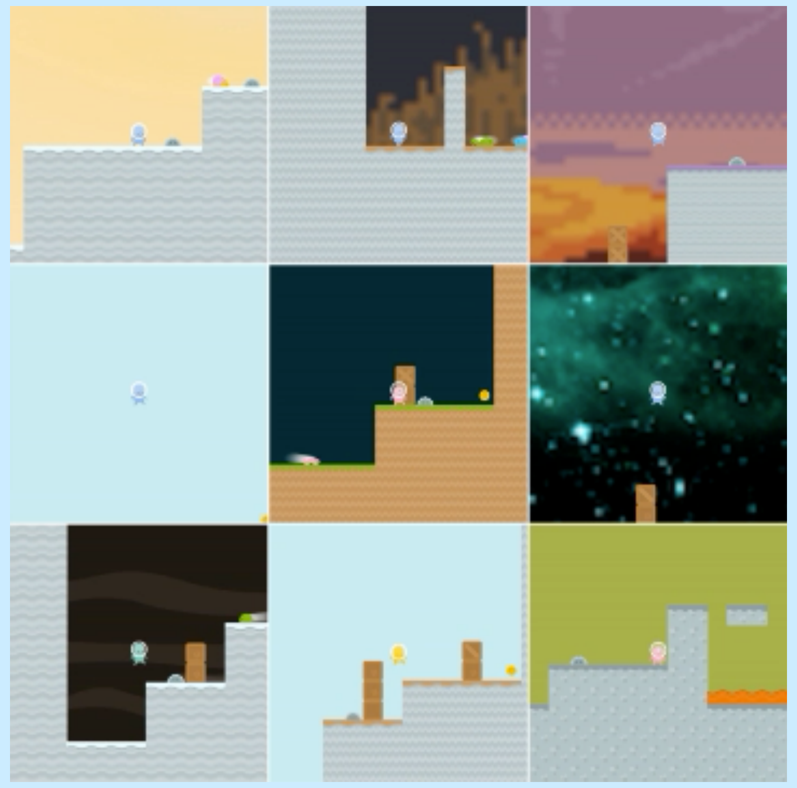
Supervised Meta-Training

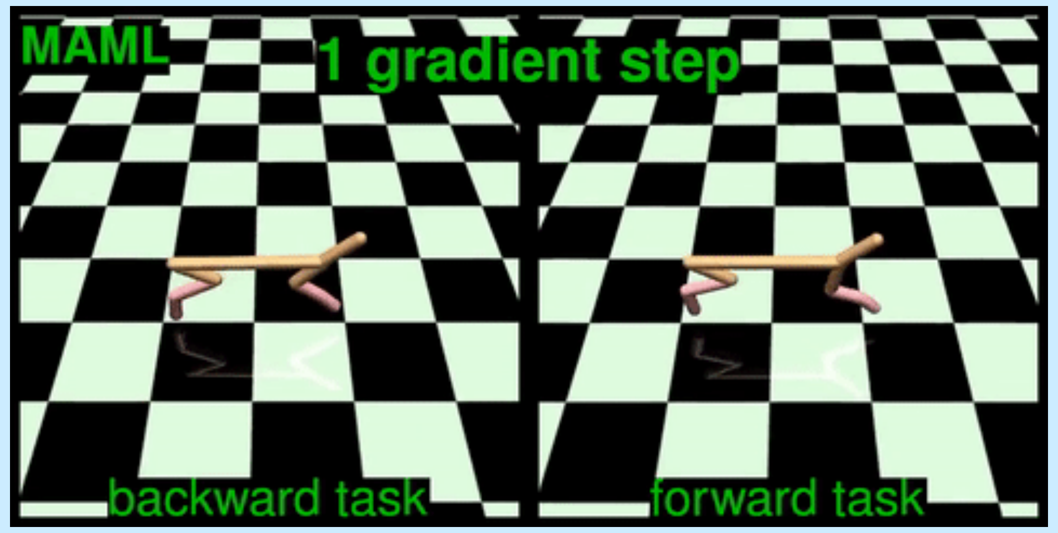
Transformer-based Generally Learning Agent (GLAs)

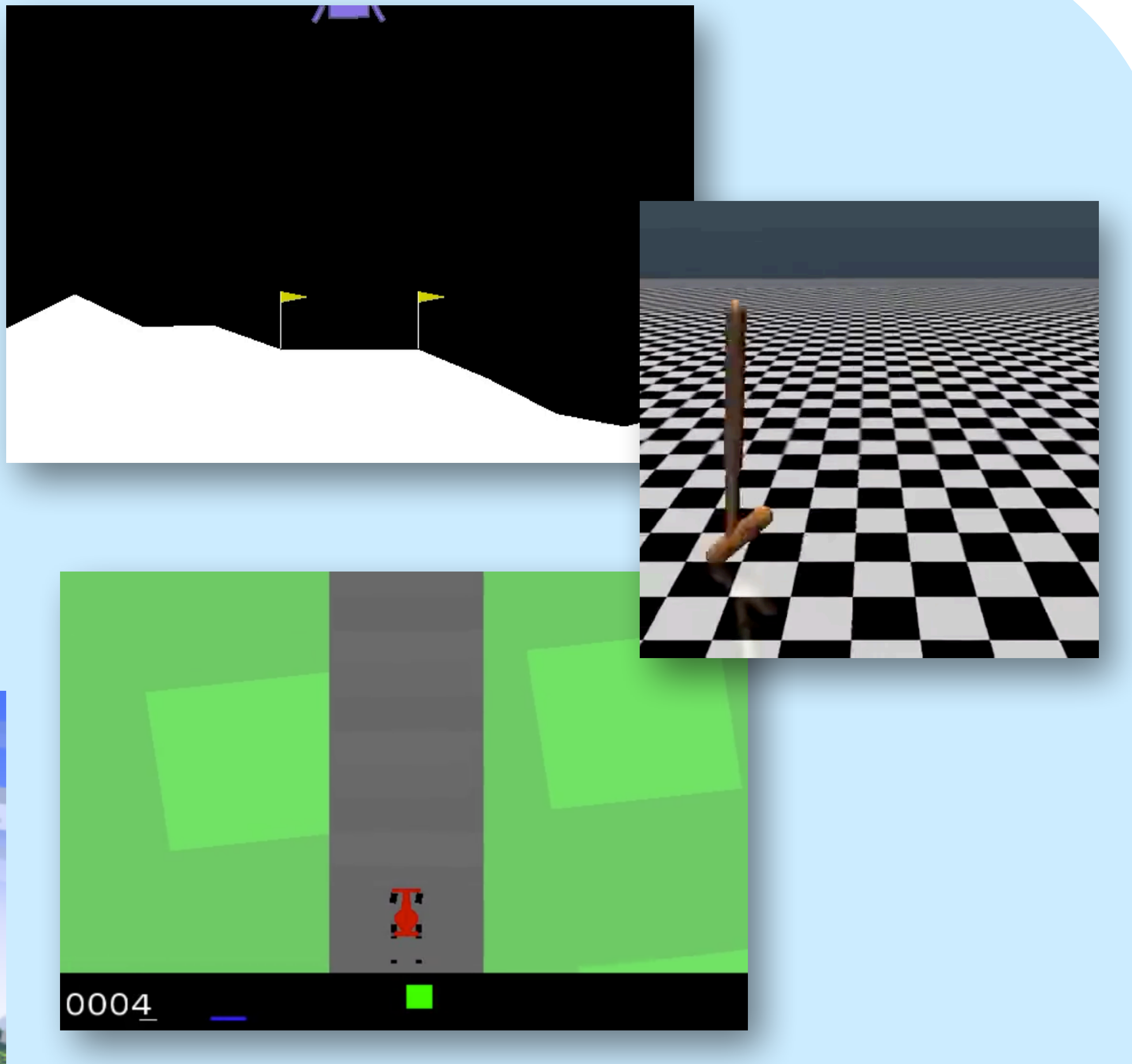
Generalize

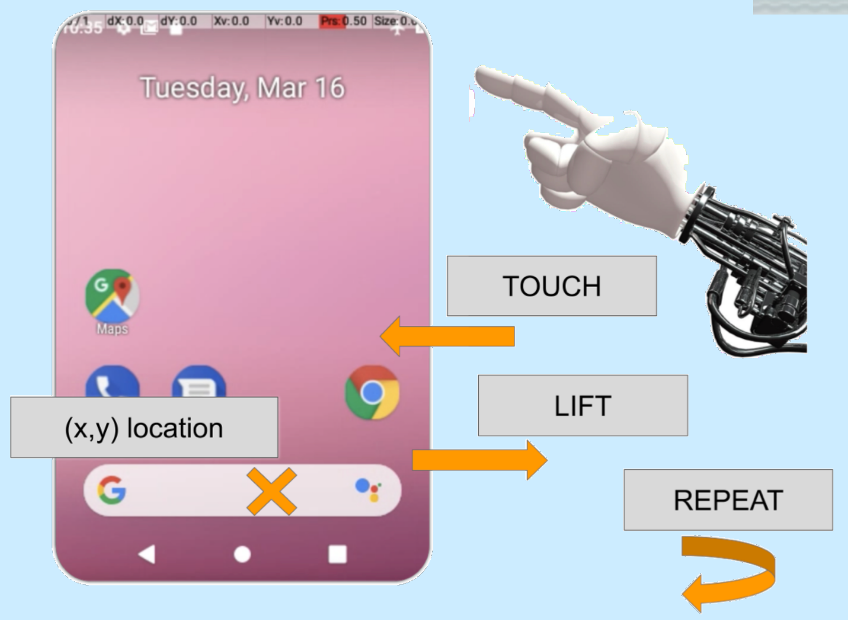
Automated task / environment generation











Env Distribution

→ **General-Purpose Learning Algorithm**

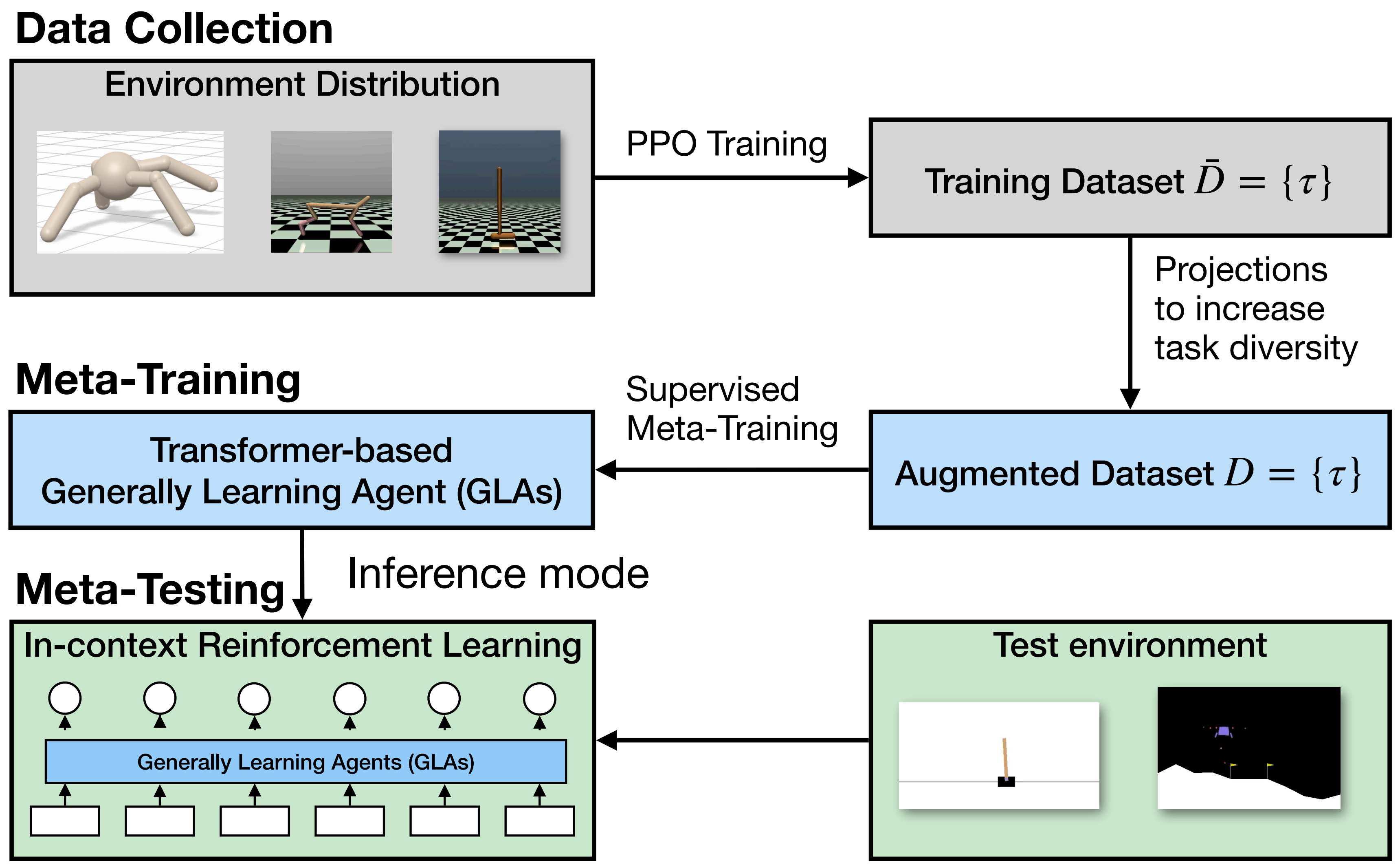
Where to find me

louiskirsch.com

@LouisKirschAI

Summary:

- We distill an accelerated PPO into a Transformer
- Strong data augmentation helps the Transformer implement a general-purpose online RL algorithm generalizing across domains



Data Collection

Environment Distribution

PPO Training

Training Dataset $\bar{D} = \{\tau\}$

Projections to increase task diversity

Meta-Training

Supervised Meta-Training

Transformer-based Generally Learning Agent (GLAs)

Augmented Dataset $D = \{\tau\}$

Meta-Testing

Inference mode

In-context Reinforcement Learning

Generally Learning Agents (GLAs)

Test environment