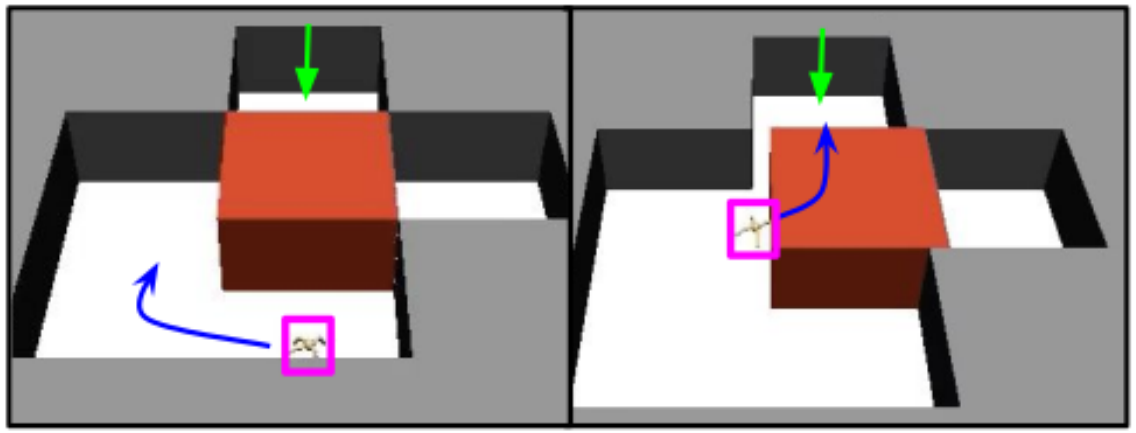
Hierarchical Deep Reinforcement Learning

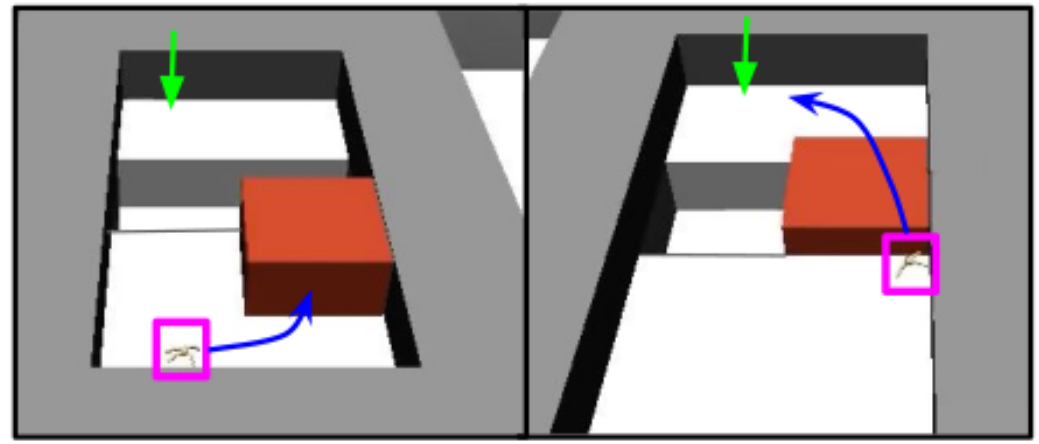
Seminar DRL

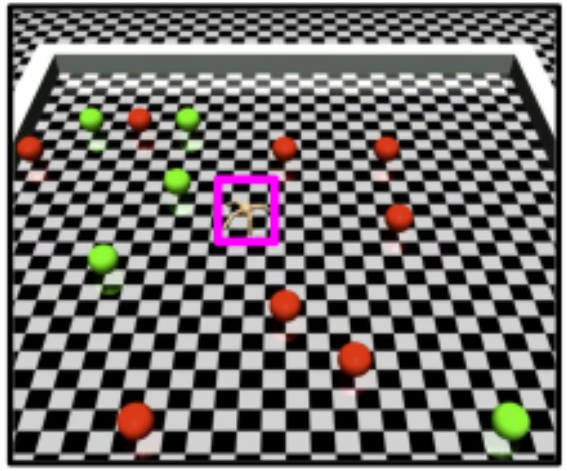
Lucas Brunner 17.03.2020

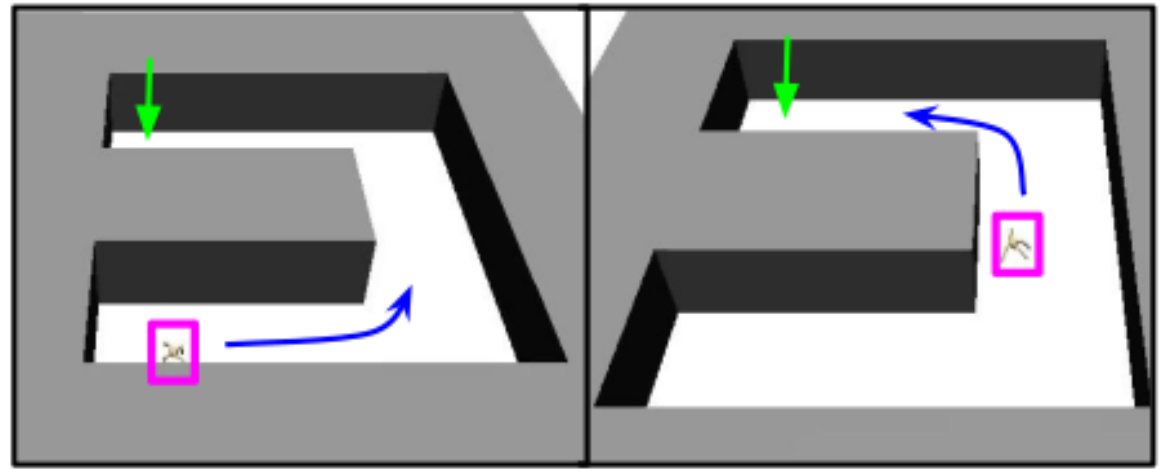
Goals

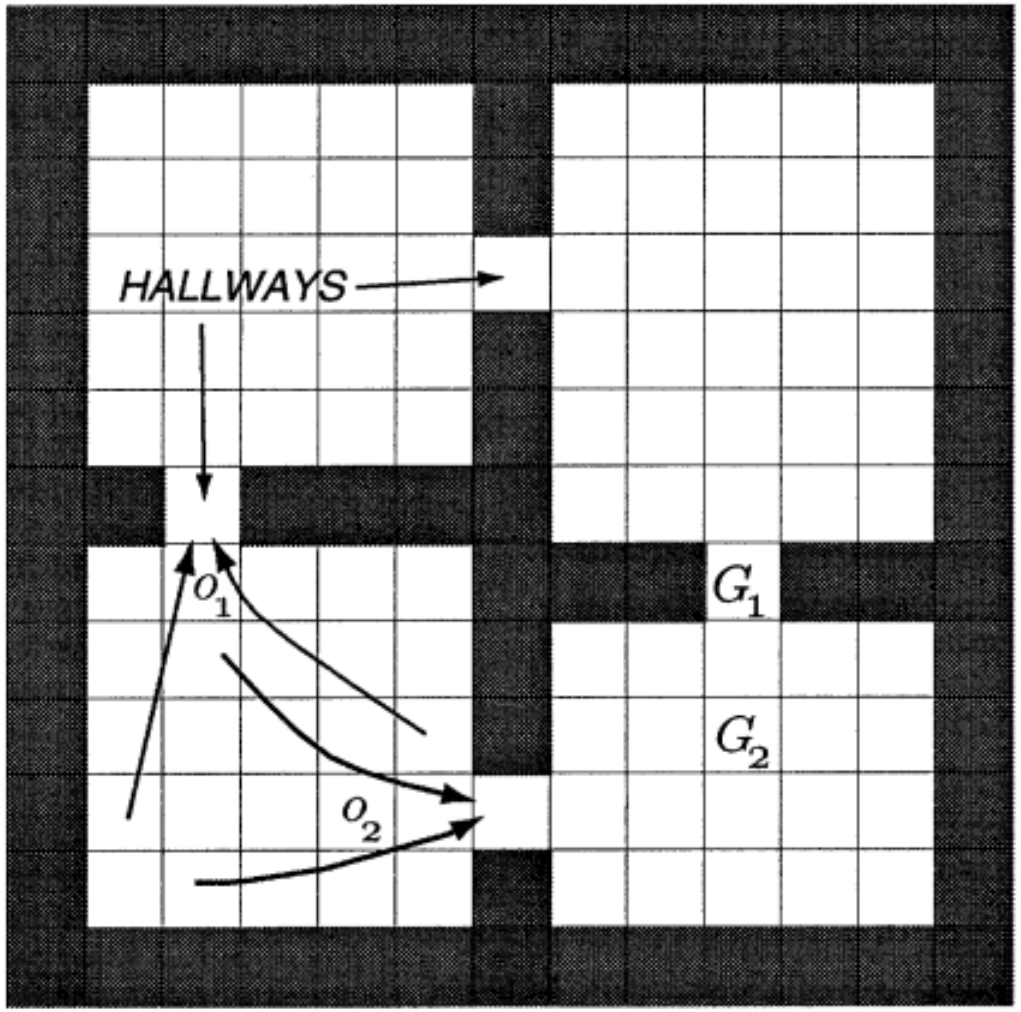
- What problem arise when doing HRL
- How can one solve them
- How are methods connected



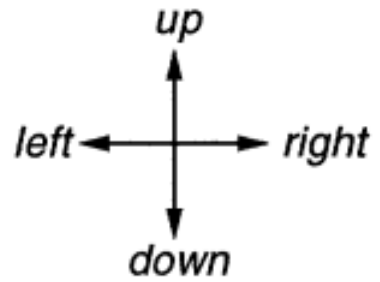








4 stochastic primitive actions



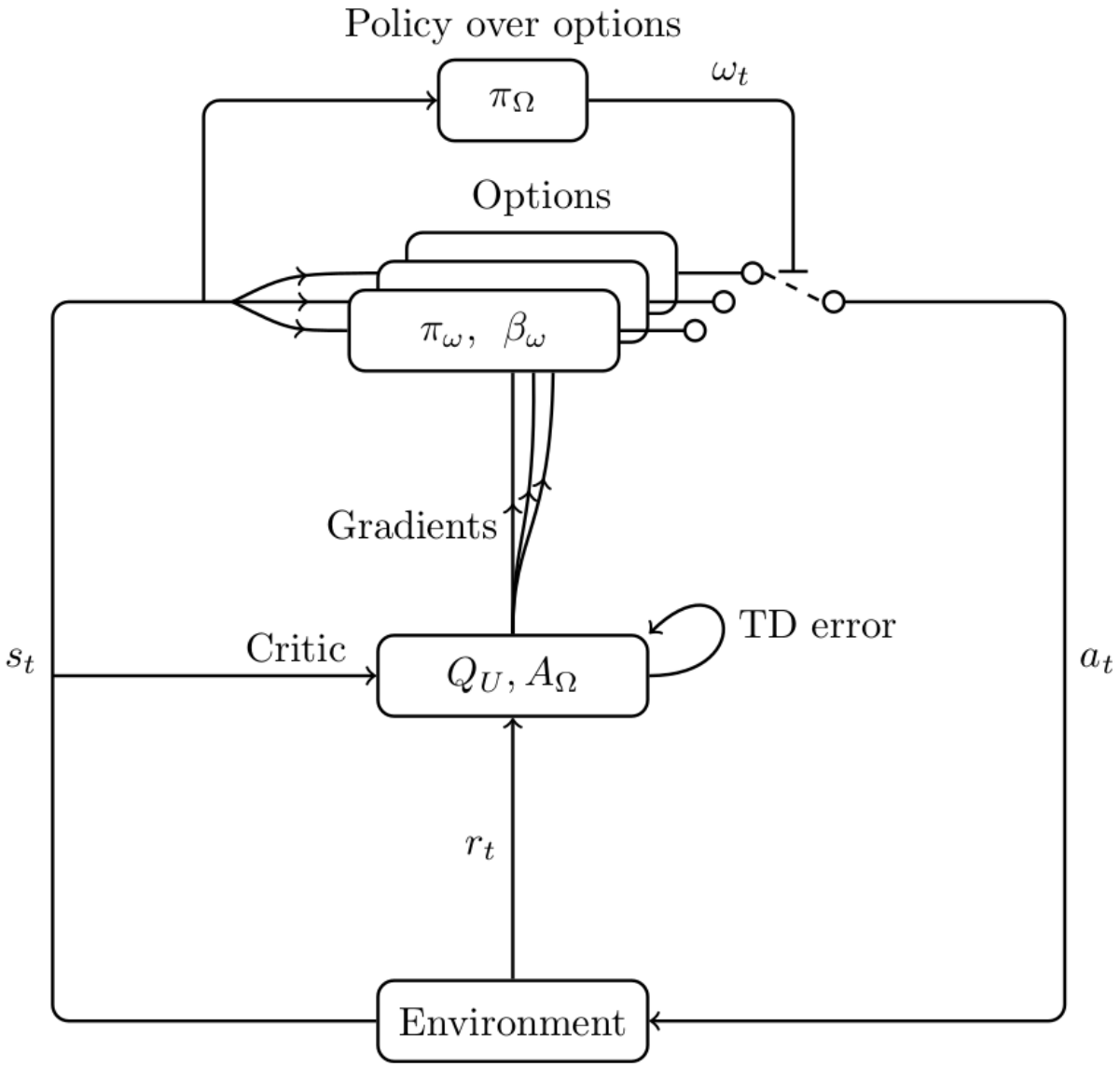
Fail 33% of the time

8 multi-step options

(to each room's 2 hallways)

(OC) The option-critic architecture

Bacon, Pierre-Luc, Jean Harb, and Doina Precup. 2017



Advantage Function

$A(s,w)=Q(s,w)-V(s)$

(“mesures how much better w is compared to alternative actions)

0 if w is optimal action
<0 if w is suboptimal

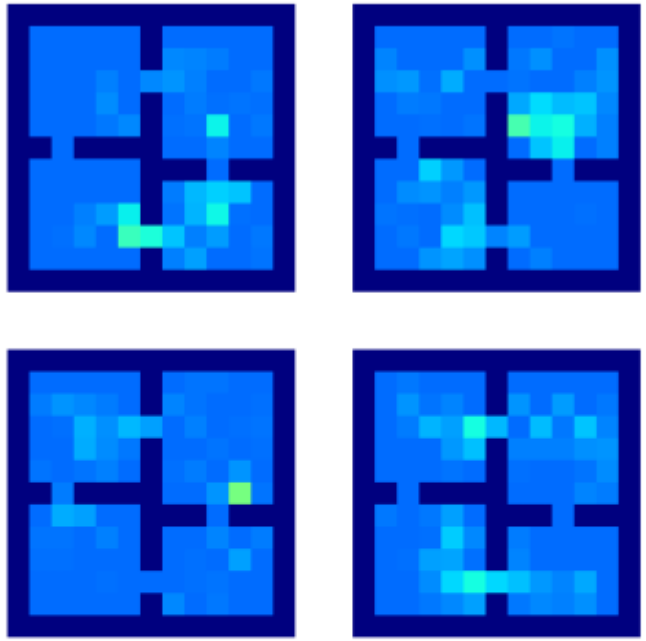
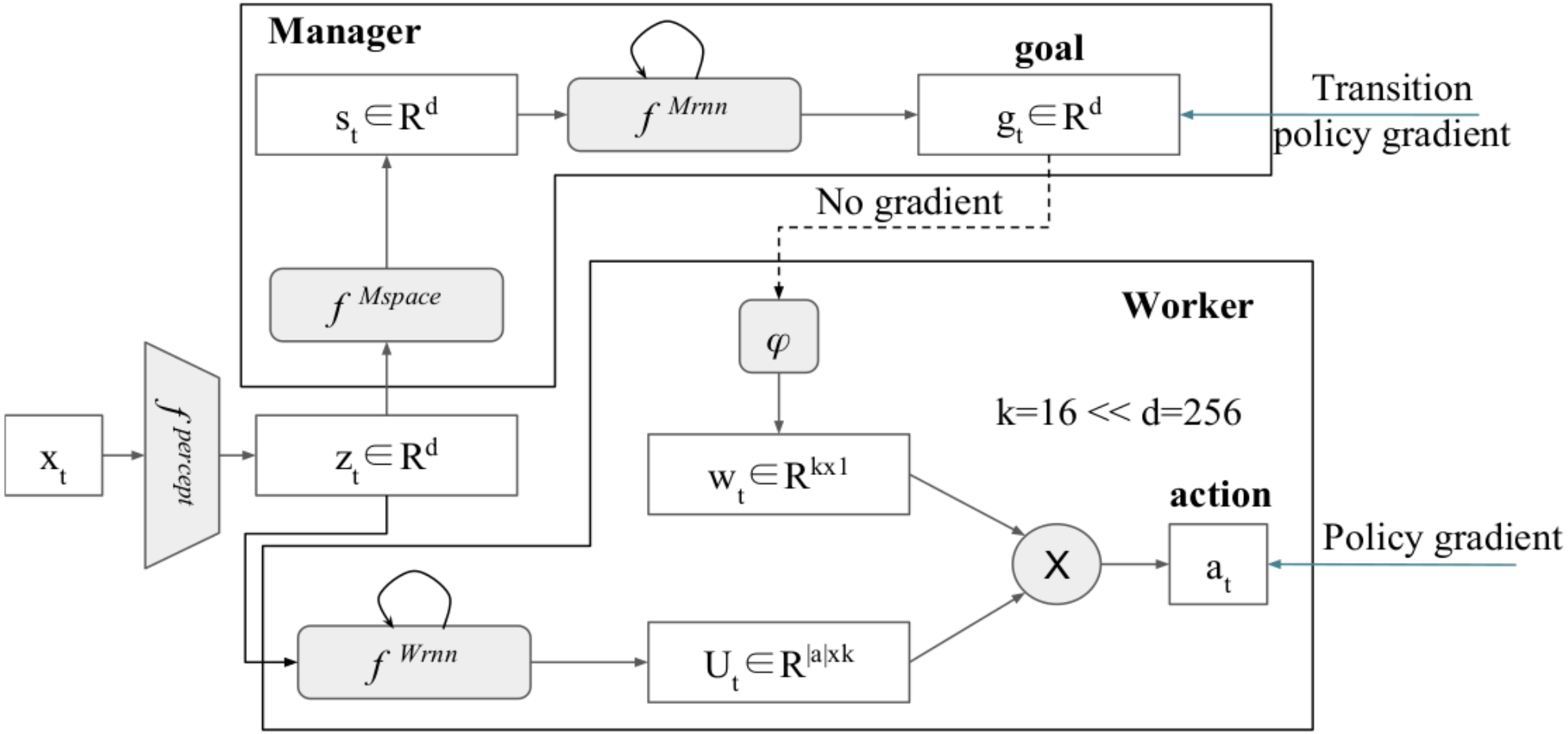


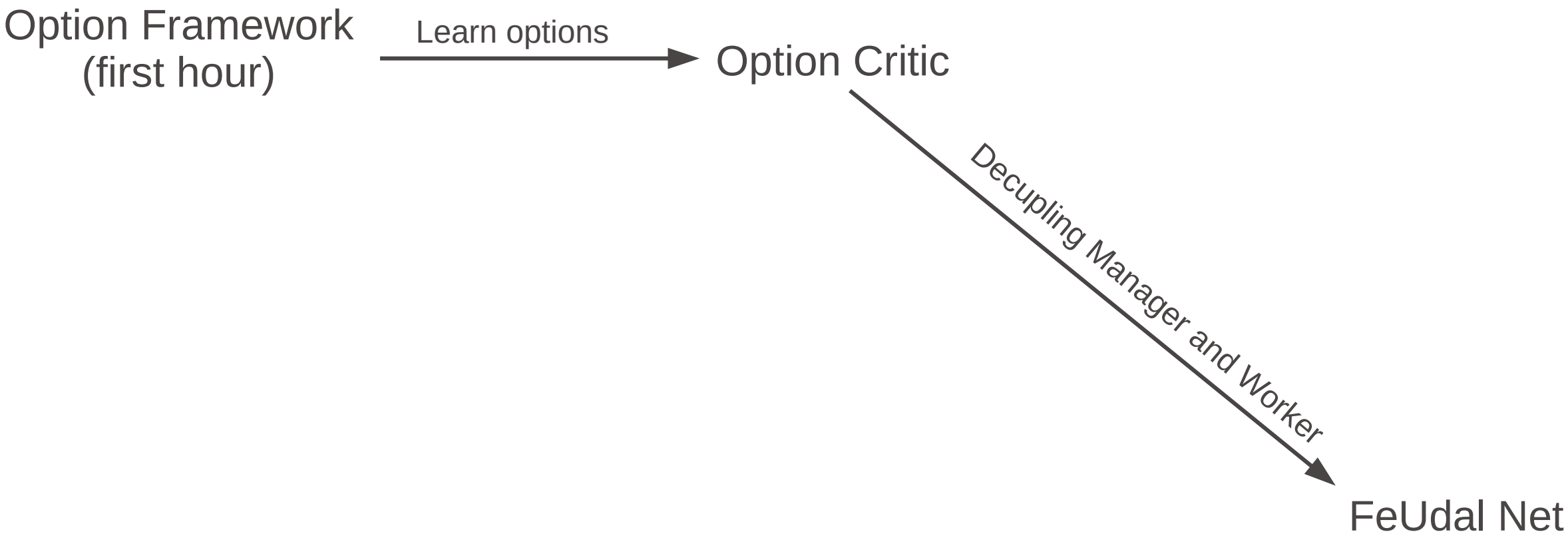
Figure 3: Termination probabilities for the option-critic agent learning with 4 options. The darkest color represents the *walls* in the environment while lighter colors encode higher termination probabilities.

Option Framework
(first hour)

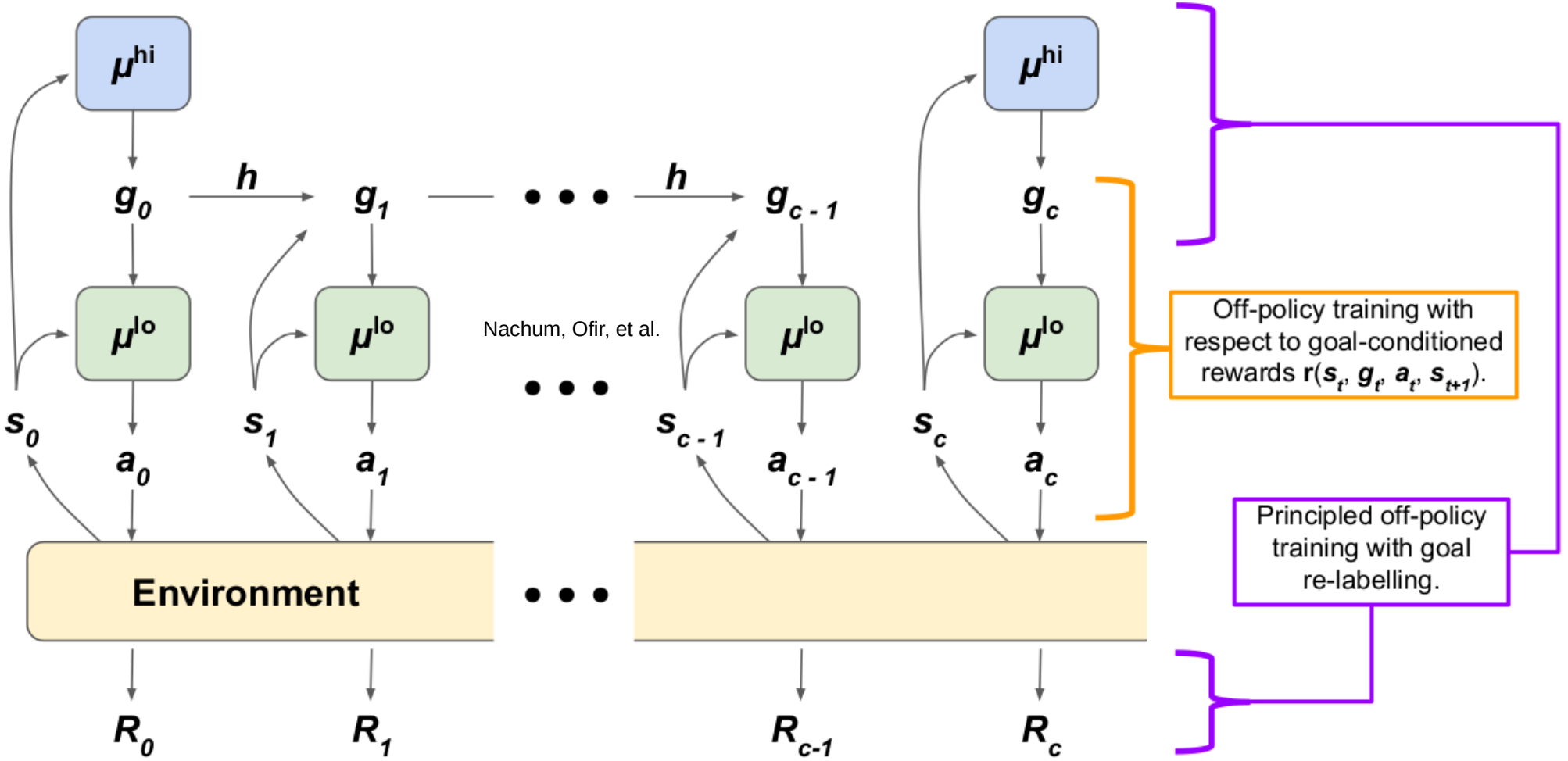
Option Critic

FeUdal Networks for Hierarchical Reinforcement Learning Vezhnevets, Alexander Sasha, et al. 2017

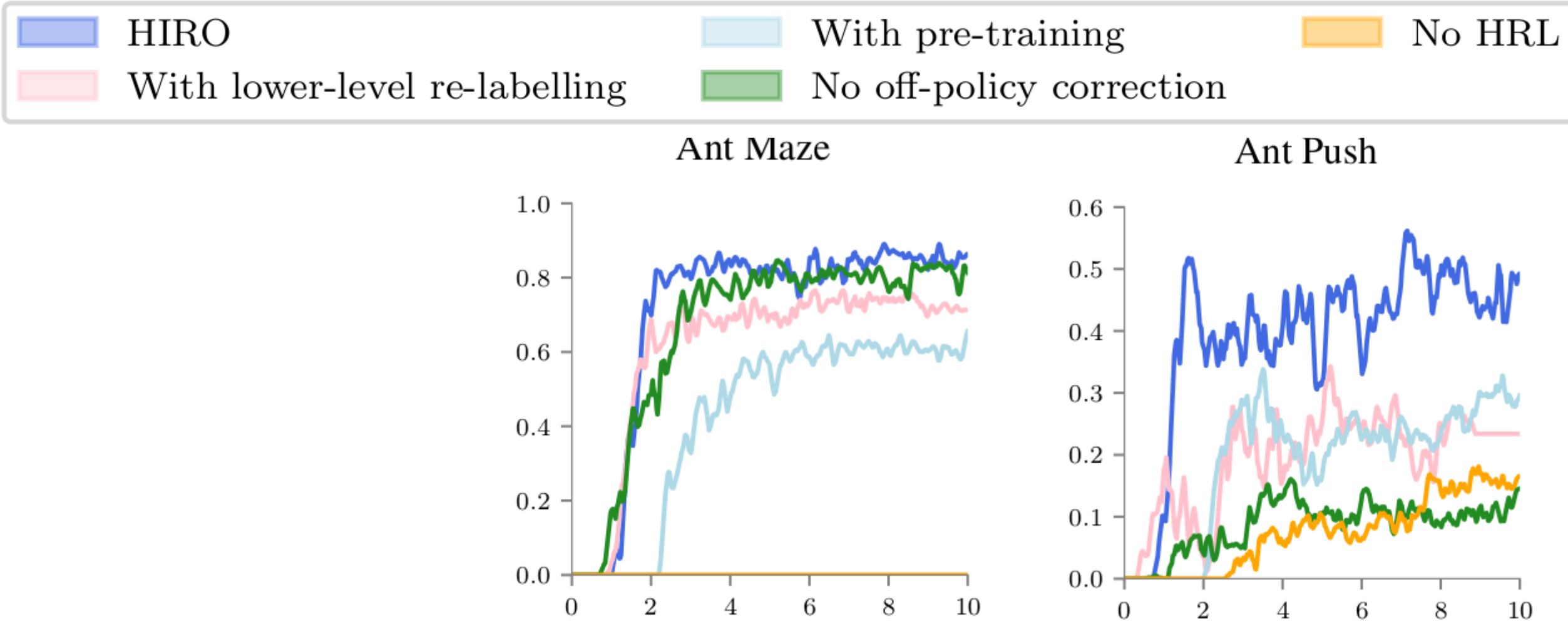


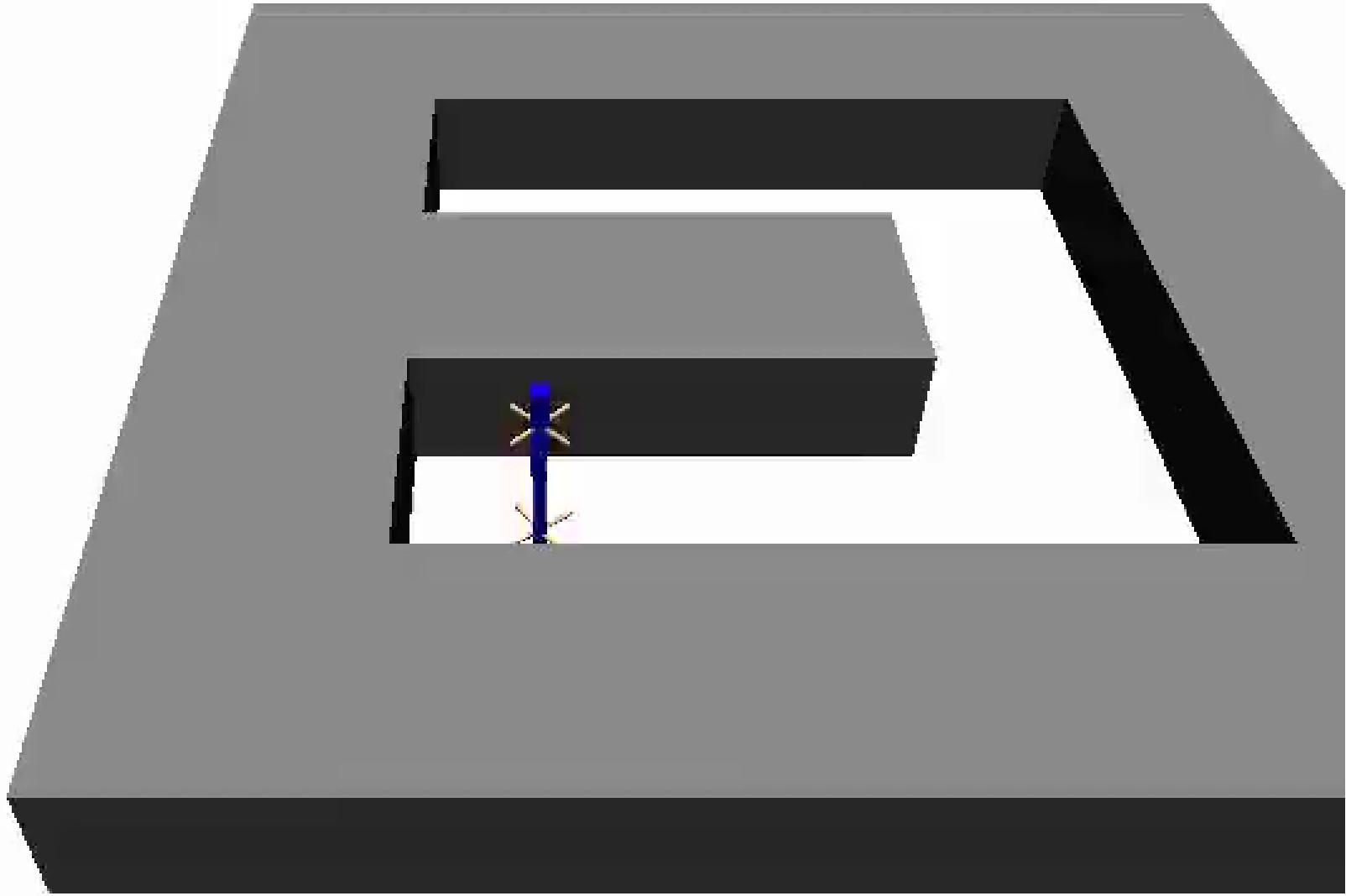

Option Framework
(first hour)
Learn options
Option Critic
Decupling Manager and Worker
FeUdal Net

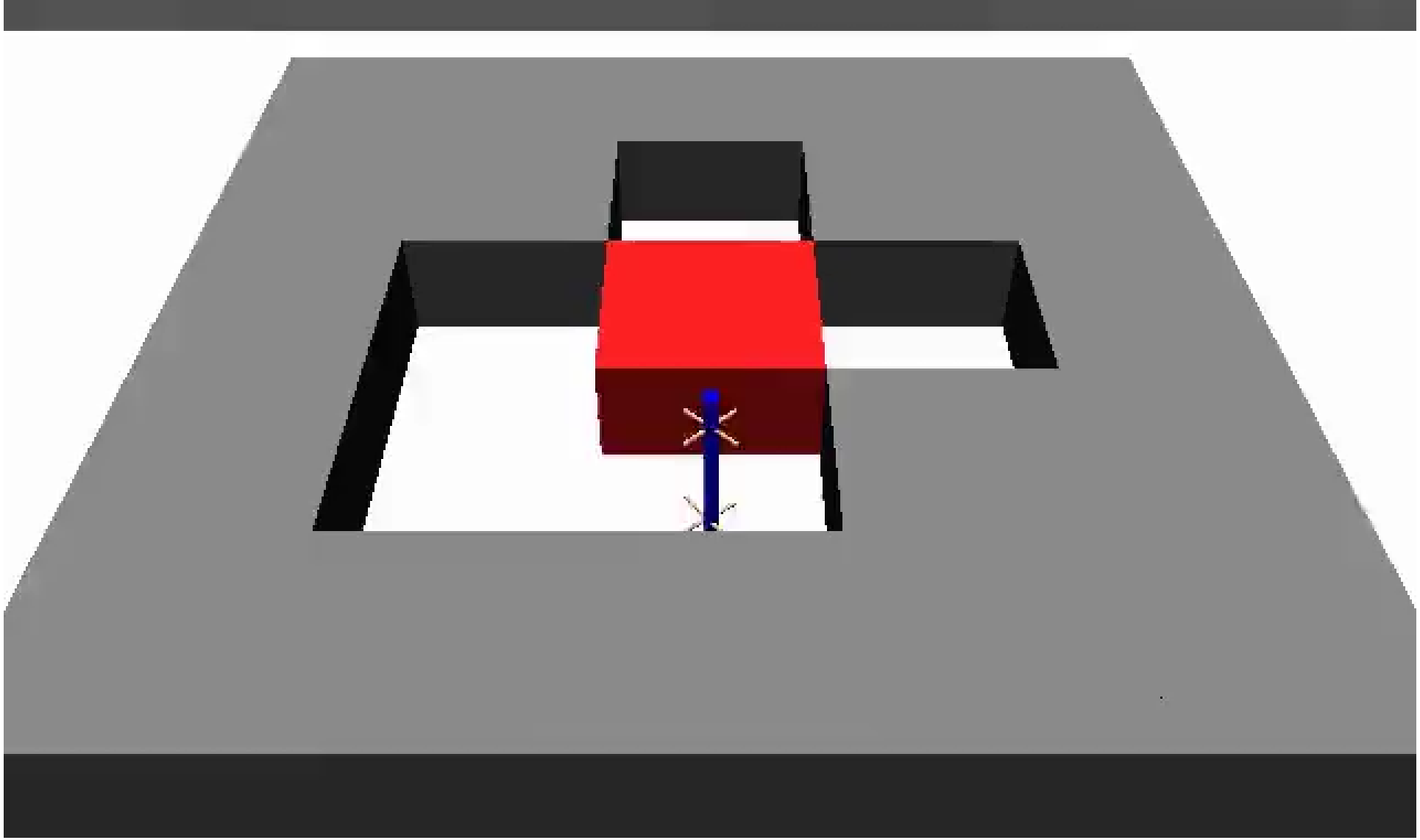
Data-Efficient Hierarchical Reinforcement Learning (HIRO) Nachum, Ofir, et al. 2017

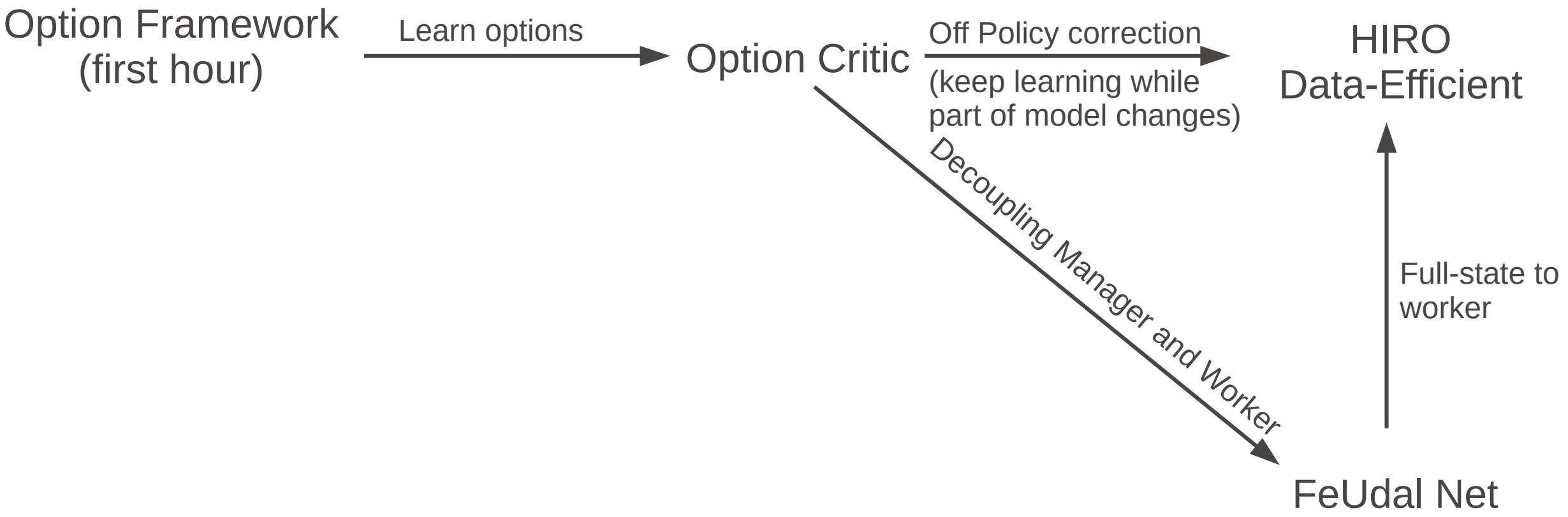


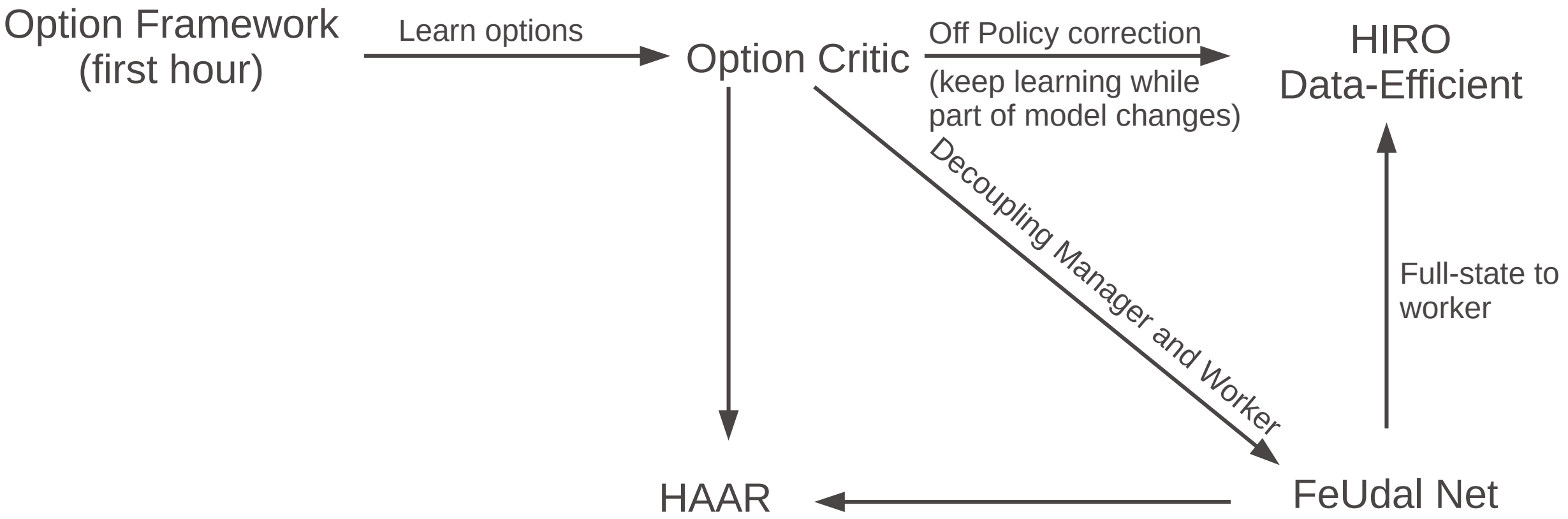
Does it help?


HIRO
With pre-training
No HRL
With lower-level re-labelling
No off-policy correction
Ant Maze
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
Ant Push
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
10



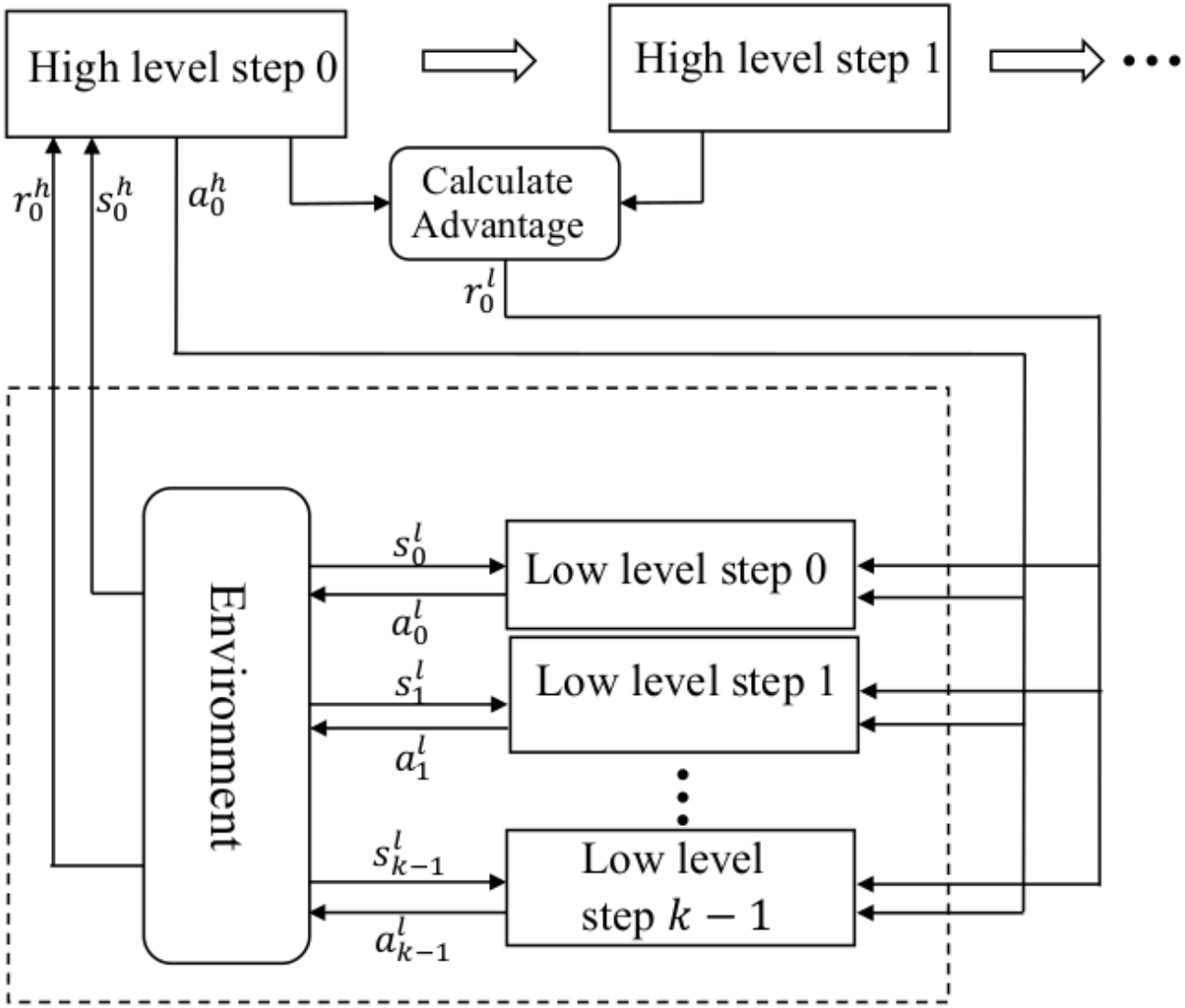


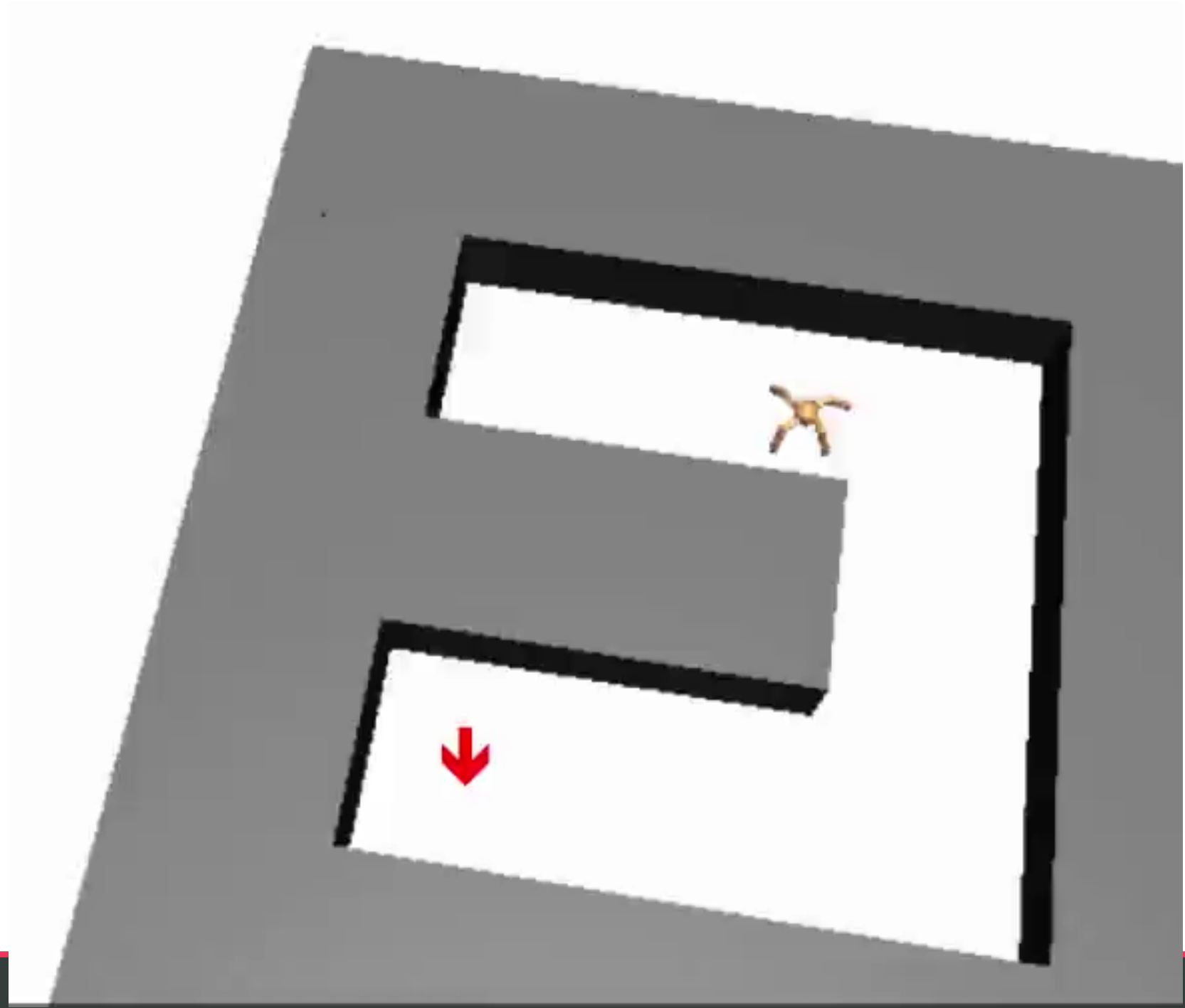

Option Framework
(first hour)
Learn options
Option Critic
Off Policy correction
(keep learning while
part of model changes)
HIRO
Data-Efficient
Decoupling Manager and Worker
Full-state to
worker
FeUdal Net


Option Framework
(first hour)
Learn options
Option Critic
Off Policy correction
(keep learning while
part of model changes)
HIRO
Data-Efficient
Decoupling Manager and Worker
Full-state to
worker
HAAR
FeUdal Net

Hierarchical Reinforcement Learning with Advantage-Based Auxiliary Rewards (HAAR)

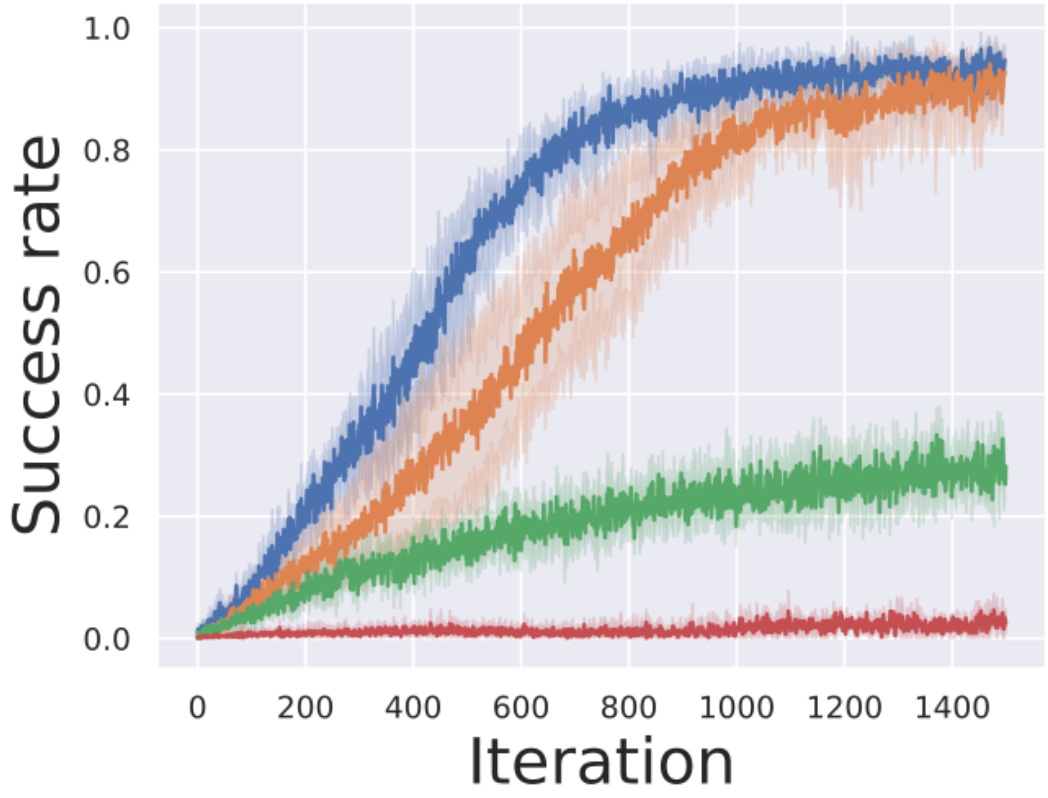
Li, Siyuan, et al. 2019

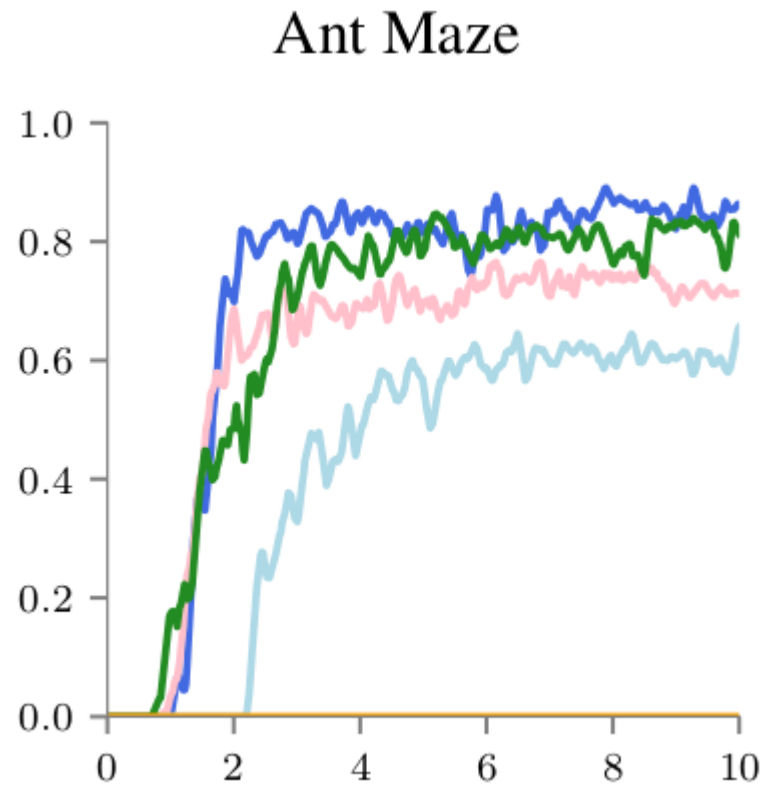




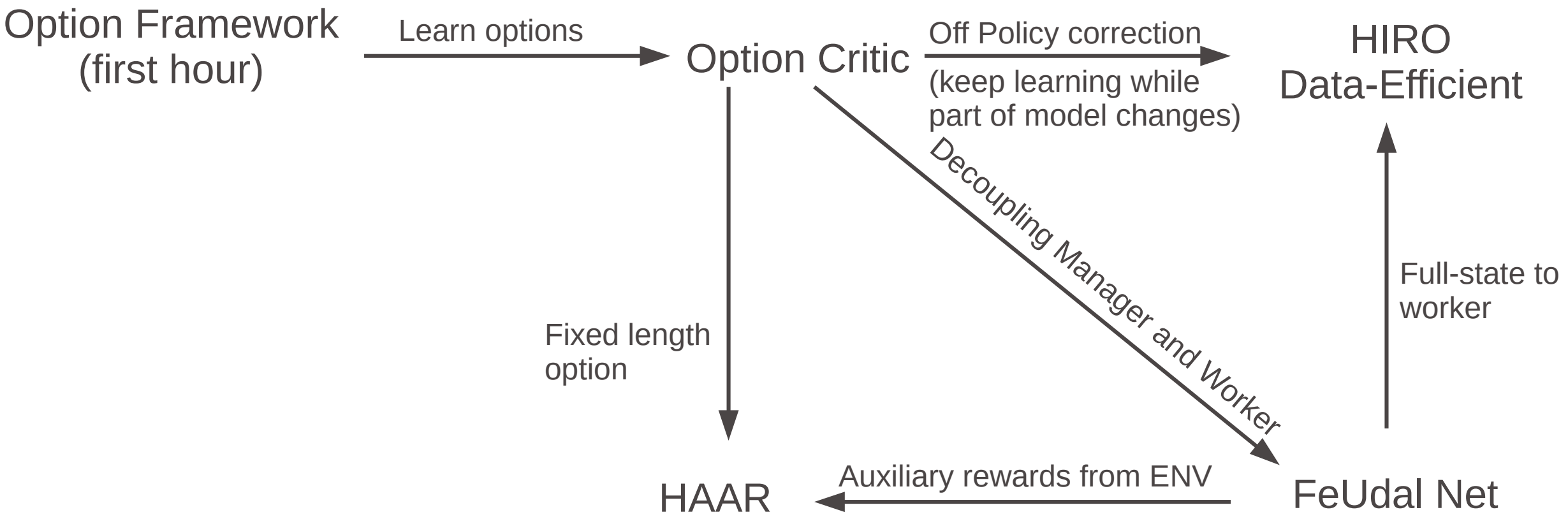
Does it help?

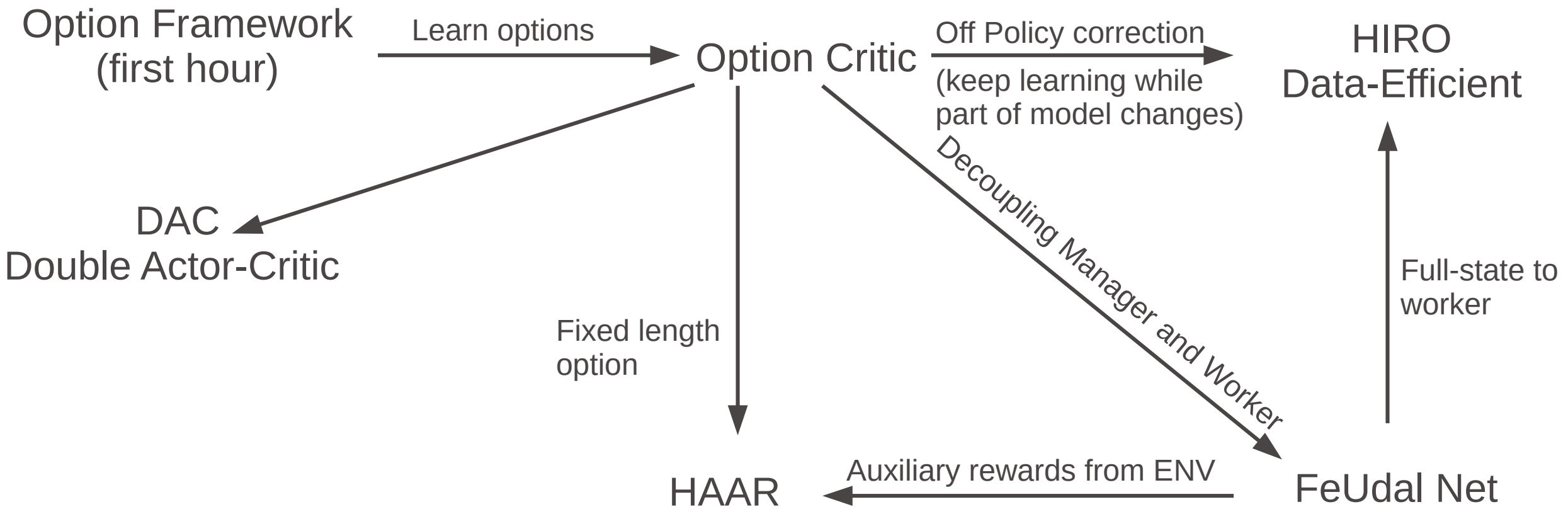
HAAR w/ annealing — HAAR w/o annealing — TRPO





(a) Ant Maze


Option Framework
(first hour)
Learn options
Option Critic
Off Policy correction
(keep learning while
part of model changes)
HIRO
Data-Efficient
Decoupling Manager and Worker
Full-state to
worker
Fixed length
option
HAAR
Auxiliary rewards from ENV
FeUdal Net


Option Framework
(first hour)
Learn options
Option Critic
Off Policy correction
(keep learning while
part of model changes)
HIRO
Data-Efficient
DAC
Double Actor-Critic
Fixed length
option
Decoupling Manager and Worker
Full-state to
worker
HAAR
Auxiliary rewards from ENV
FeUdal Net

Double Actor-Critic (DAC)

Shangtong Zhang, Shimon Whiteson 2019

Algorithm 1: Pseudocode of DAC

Input:

Parameterized $\pi, \{\pi_o, \beta_o\}_{o \in \mathcal{O}}$

Policy optimization algorithms $\mathbb{A}_1, \mathbb{A}_2$

Get an initial state S_0

$t \leftarrow 0$

while *True* **do**

  Sample O_t from $\pi^{\mathcal{H}}\big(\cdot|(O_{t-1}, S_t)\big)$

  Sample A_t from $\pi^{\mathcal{L}}\big(\cdot|(S_t, O_t)\big)$

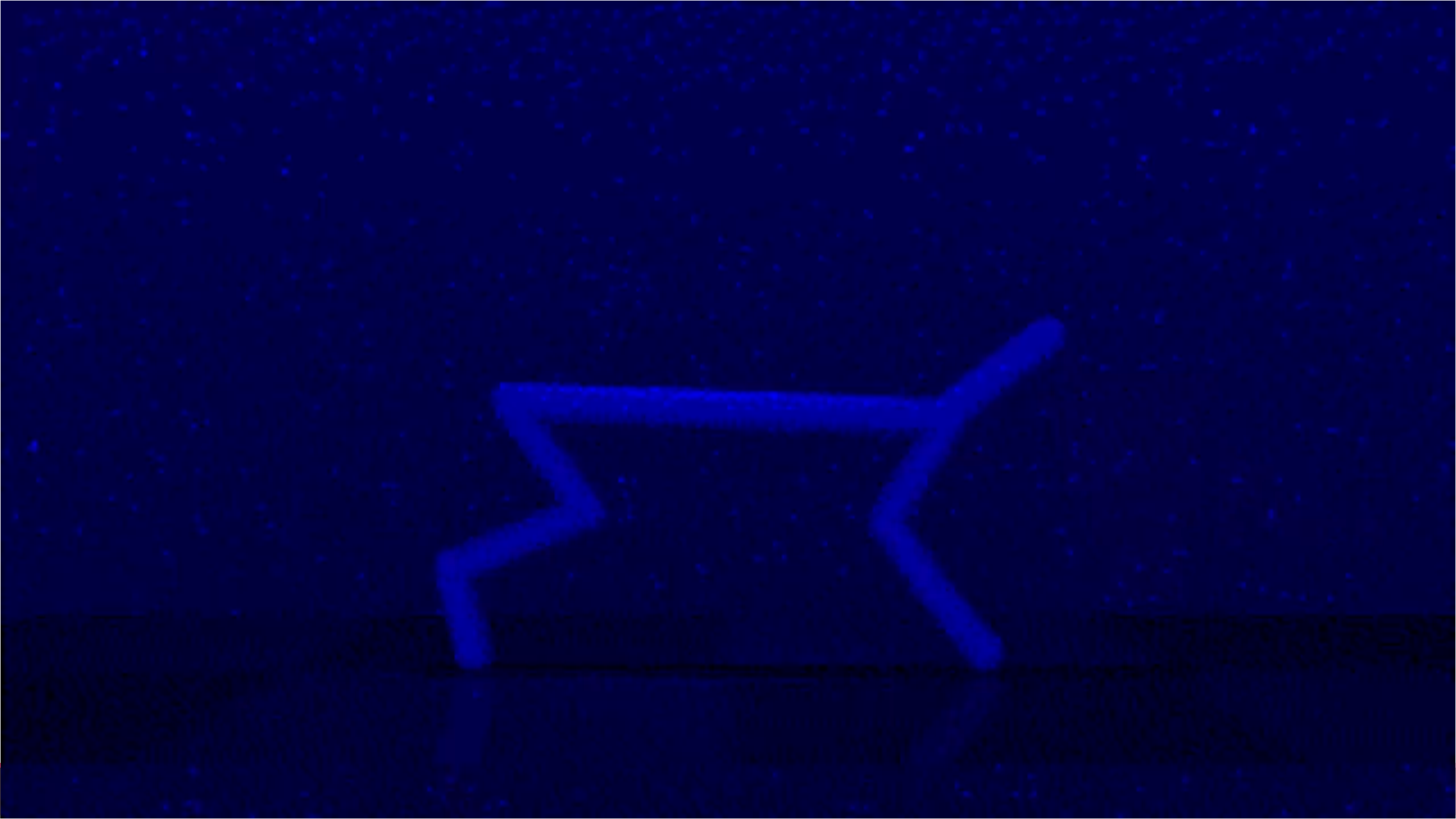
  Execute A_t, get R_{t+1}, S_{t+1}

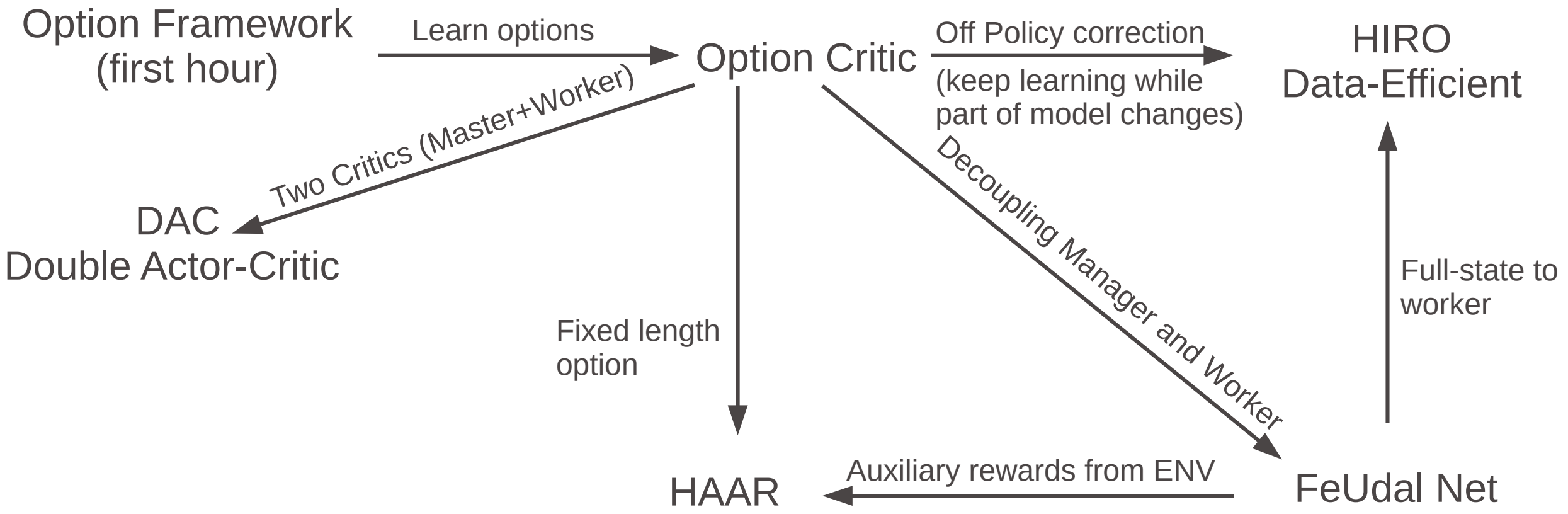
  Optimize $\pi^{\mathcal{H}}$ with $(S_t^{\mathcal{H}}, A_t^{\mathcal{H}}, R_{t+1}, S_{t+1}^{\mathcal{H}})$ and $\mathbb{A}_1$

  Optimize $\pi^{\mathcal{L}}$ with $(S_t^{\mathcal{L}}, A_t^{\mathcal{L}}, R_{t+1}, S_{t+1}^{\mathcal{L}})$ and $\mathbb{A}_2$

  $t \leftarrow t + 1$

end




Option Framework
(first hour)
Learn options
Option Critic
Off Policy correction
(keep learning while
part of model changes)
HIRO
Data-Efficient
Two Critics (Master+Worker)
DAC
Double Actor-Critic
Fixed length
option
Decoupling Manager and Worker
Full-state to
worker
HAAR
Auxiliary rewards from ENV
FeUdal Net

Thank You

Advantage Function

- Do we really ned to compute both