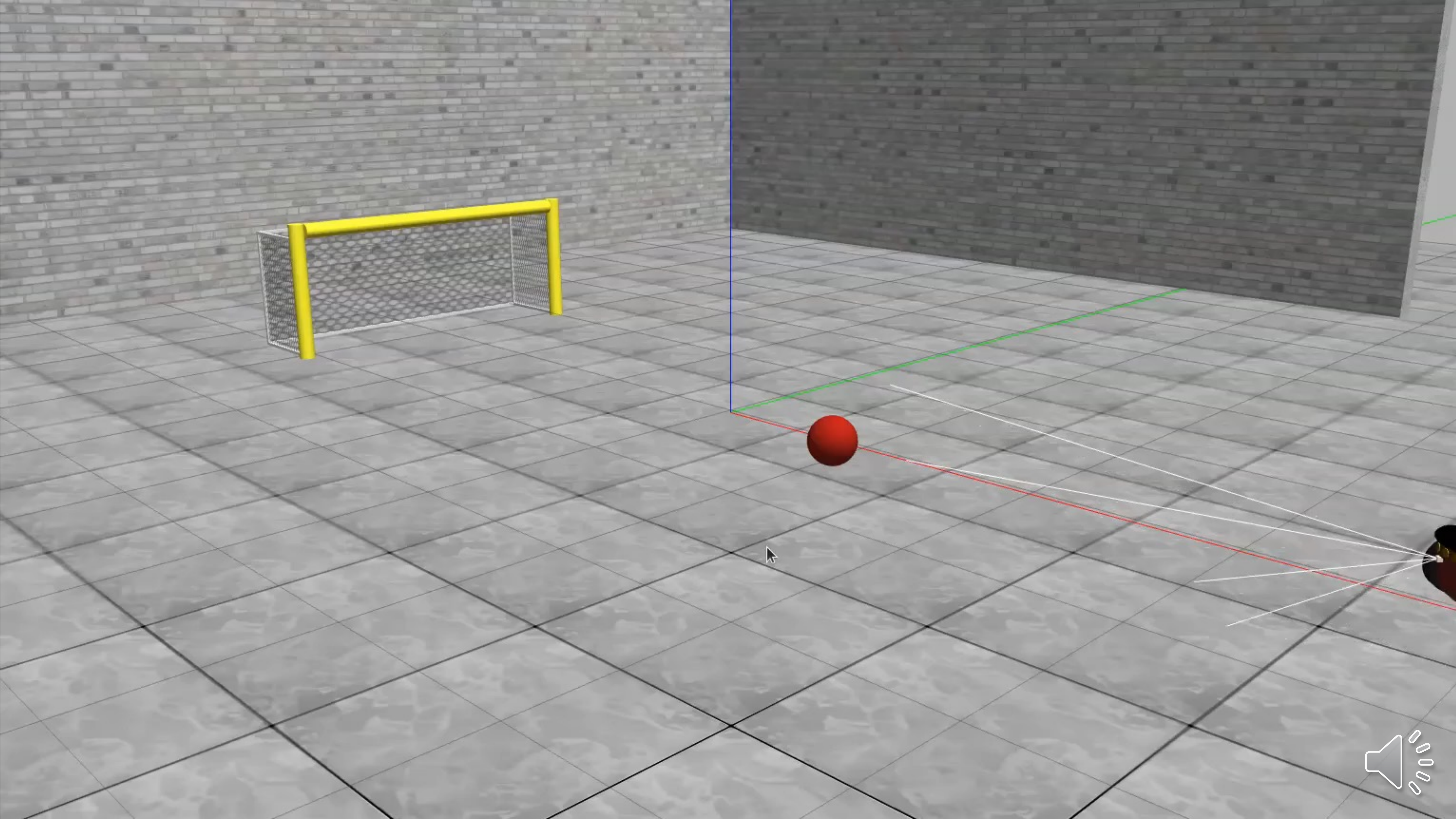
HIDIO: Hierarchical RL by Discovering Intrinsic Options

Jesse Zhang[1]*, Haonan Yu[2]*,
Wei Xu[2]

USC [1]

Horizon Robotics [2]



Motivation

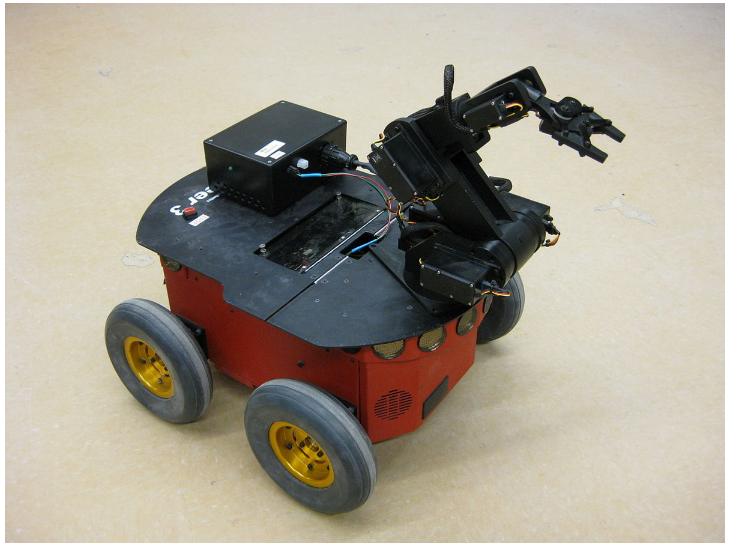
Complex, sparse-reward tasks are difficult with RL!

Hierarchical RL presents an alternative

Easier exploration through increasing temporal abstraction	Lower-level policies can represent useful skills (options) for a given task

Most existing works make assumptions about task structure or required skills for task

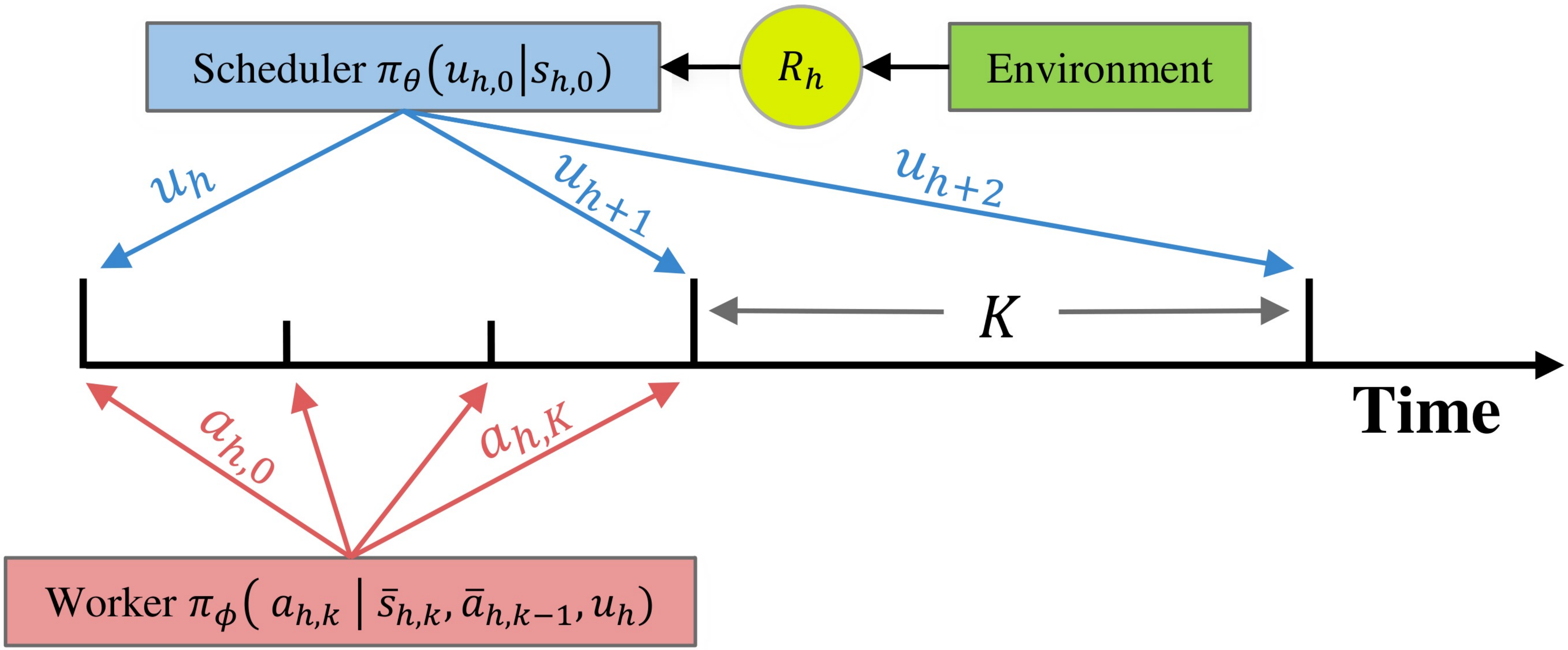
Manual design of task decomposition	Utilizing pre-defined options



Contribution: HIDIO

- Discovers **task-agnostic options** in a self-supervised manner while learning to utilize them to solve the task
- No assumptions about task structure or option type
- Better sample efficiency and final performance than other methods

Contribution: HIDIO



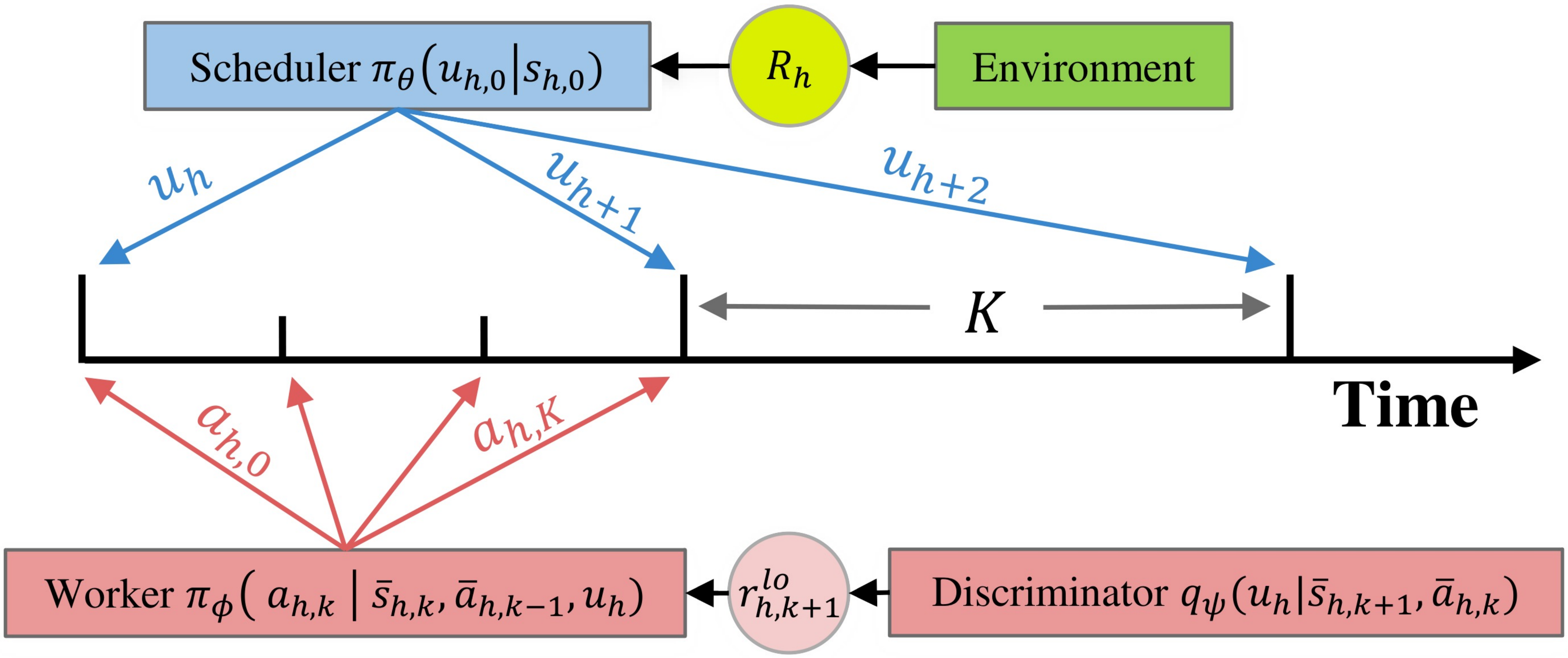
How to learn task agnostic options?

- The worker π_ϕ should help the scheduler *explore*
  - Maximize worker entropy $H(\pi_\phi(a|\bar{s}, \bar{a}, u))$
- Options should be uniquely determined
  - Minimize the entropy of options conditioned on the worker's inputs: $H(p(u|\bar{s}, \bar{a}))$
  - $p(u|\bar{s}, \bar{a})$ intractable, learn a *discriminator* $q_\psi(u|\bar{s}, \bar{a})$ instead

$$\max_{\phi,\psi} H\left(\pi_\phi(a|\bar{s}, \bar{a}, u)\right) - H\left(q_\psi(u|\bar{s}, \bar{a})\right)$$

$$r^{low} := \log q_\psi - \beta \log \pi_\phi$$

Contribution: HIDIO



Discriminator Instantiations

How to learn q_ψ?

$$\max_\psi \log q_\psi(u_t | \bar{s}, \bar{a}) = \max_\psi -|| f_\psi(\bar{s}, \bar{a}) - u_t ||$$

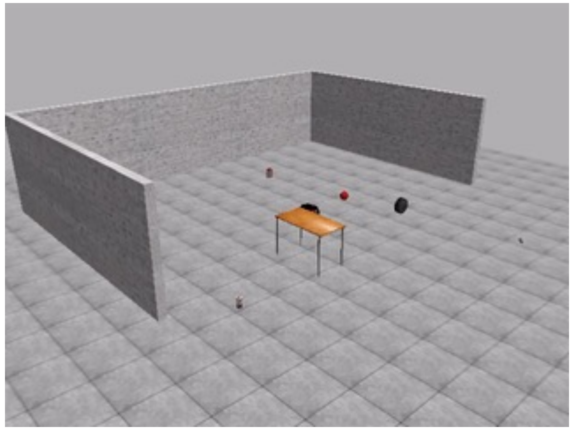
Feature Extractor	Formulation f_ψ (MLP = Multi-Layer Perceptron)	Explanation
State[1]	MLP(s_t)	Current State
Action	MLP$([s_0, a_t])$	Action + first state
StateDiff	MLP$(s_t - s_{t-1})$	Difference between states
StateAction	MLP$([a_{t-1}, s_t])$	Action + current state
StateConcat[2]	MLP$([\bar{s}_{0:t}])$	States so far
ActionConcat	MLP$([s_0, \bar{a}_{0:t-1}])$	Actions so far + first state

[1]Diversity is All You Need, Eysenbach et al. 2018
[2]Unsupervised control through non-parametric discriminative rewards, Warde-Farley et al. 2019; Dynamics-aware unsupervised discovery of skills, Sharma et al. 2019; Variational option discovery algorithms, Achiam et al. 2018

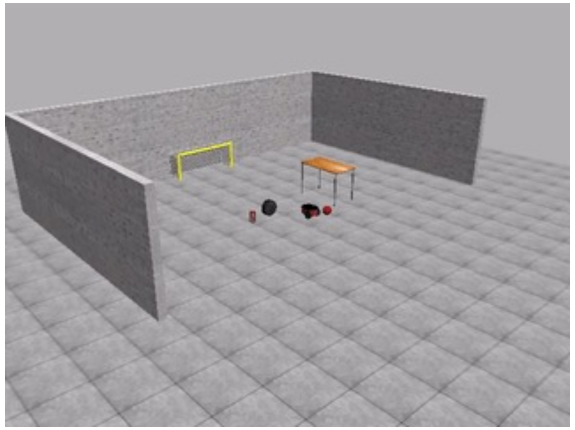
Sparse-Reward Environments

Pusher

Reacher



GoalTask

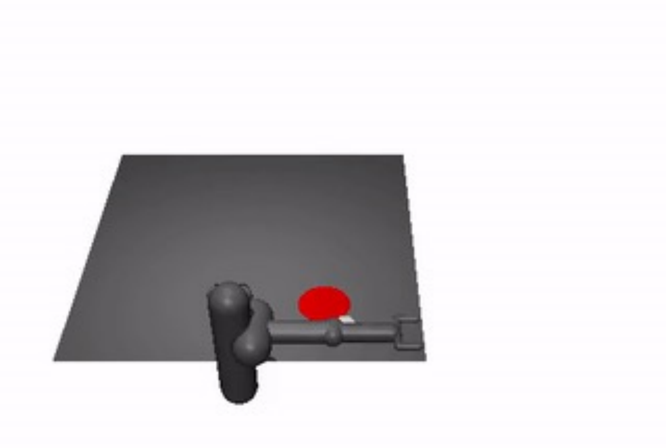


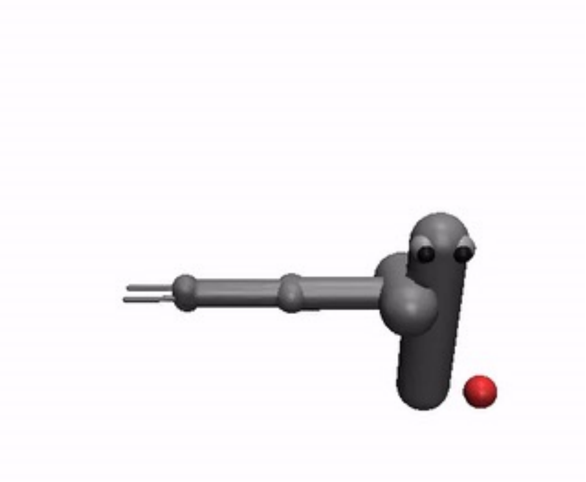
KickBall

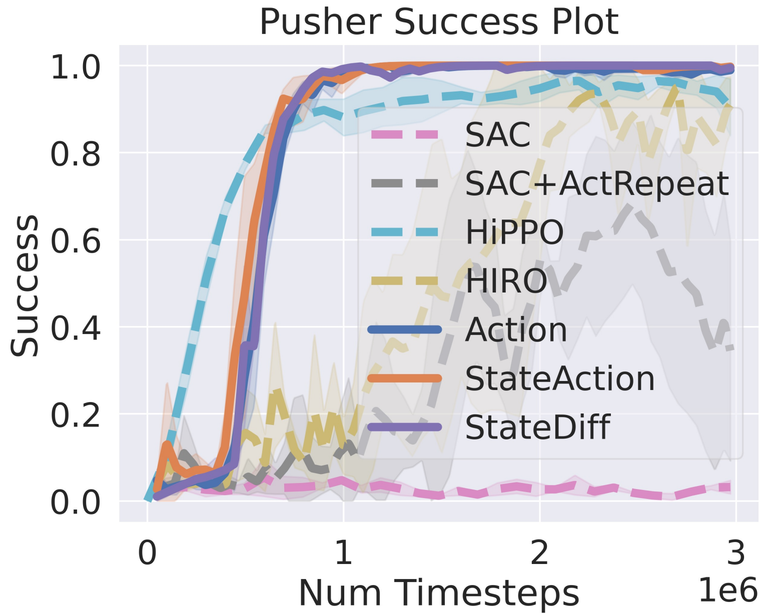
Methods Compared

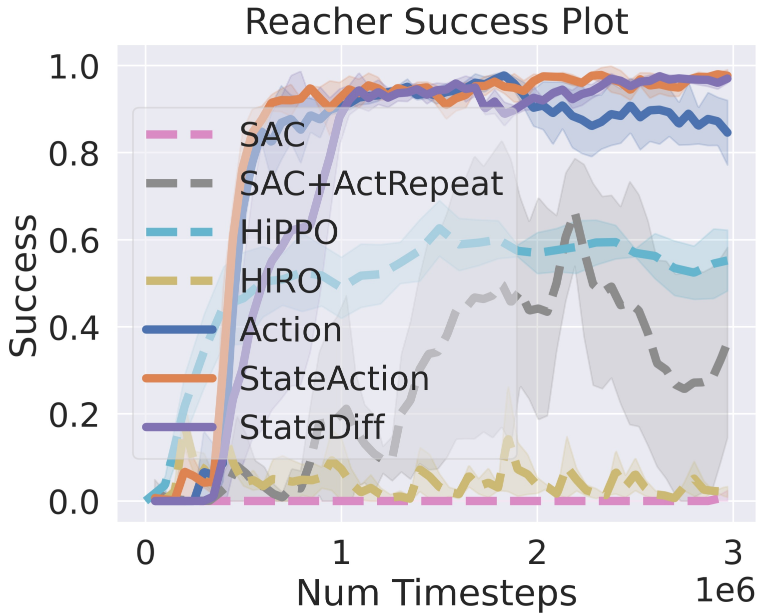
- SAC: Soft Actor-Critic
- SAC+ActRepeat: Soft Actor-Critic with the same temporal abstraction as ours
- HiPPO: Sub-Policy Adaptation for Hierarchical RL
- HIRO: Data-Efficient Hierarchical RL
- Action: HIDIO with Action feature extractor
- StateAction: HIDIO with StateAction feature extractor
- StateDiff: HIDIO with StateDiff feature extractor

Pusher and Reacher

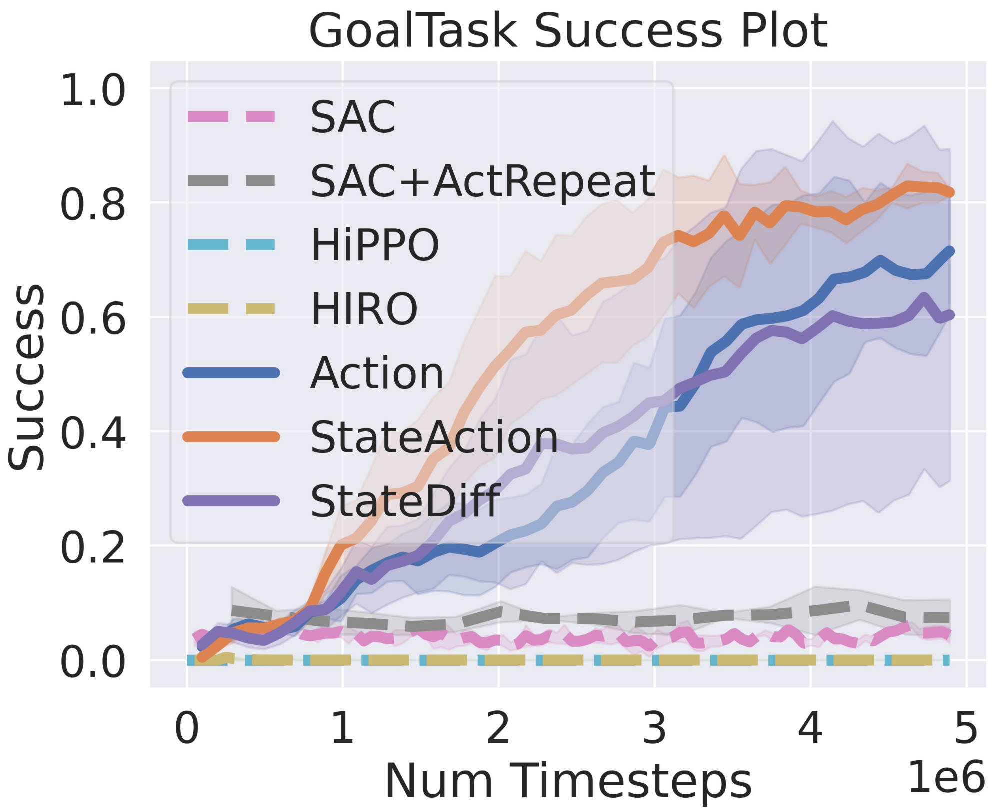


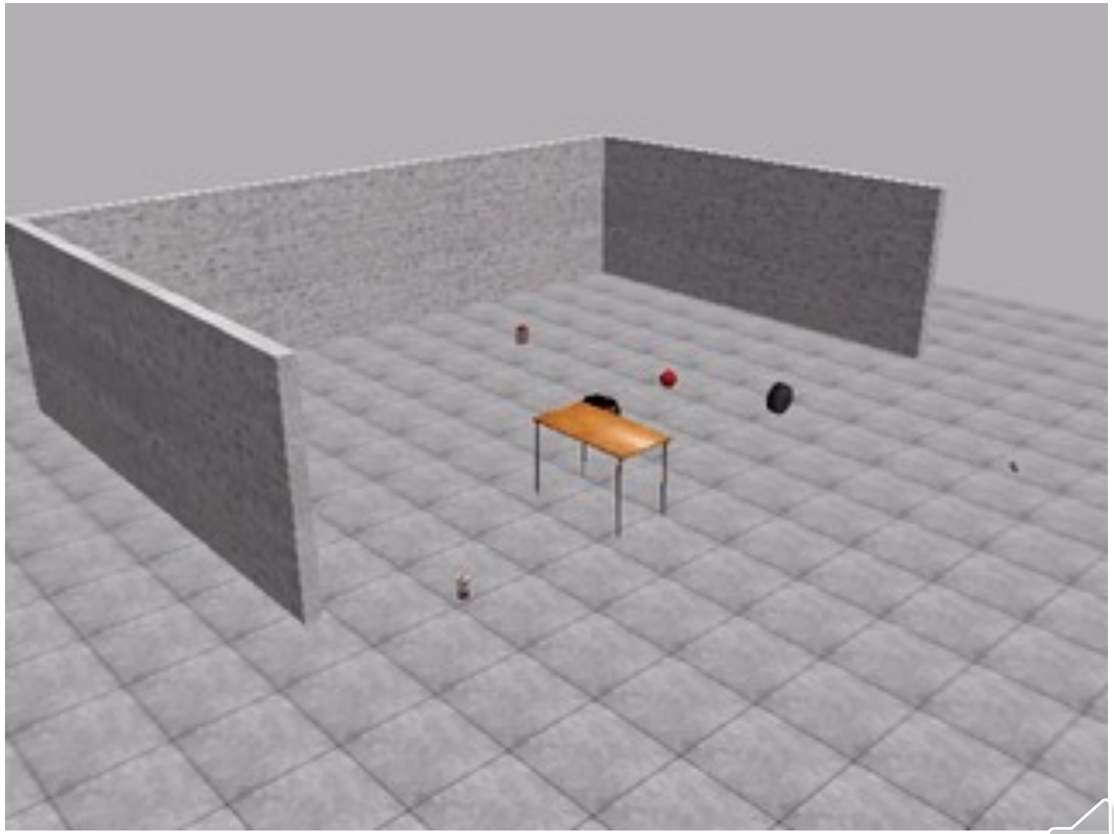




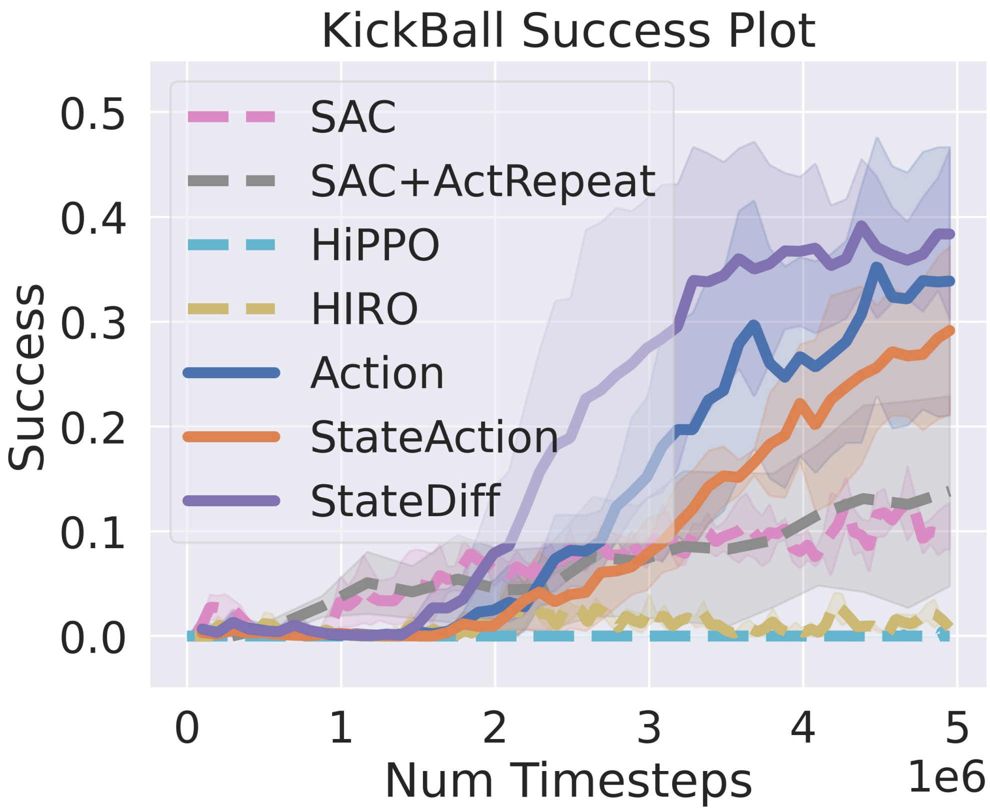


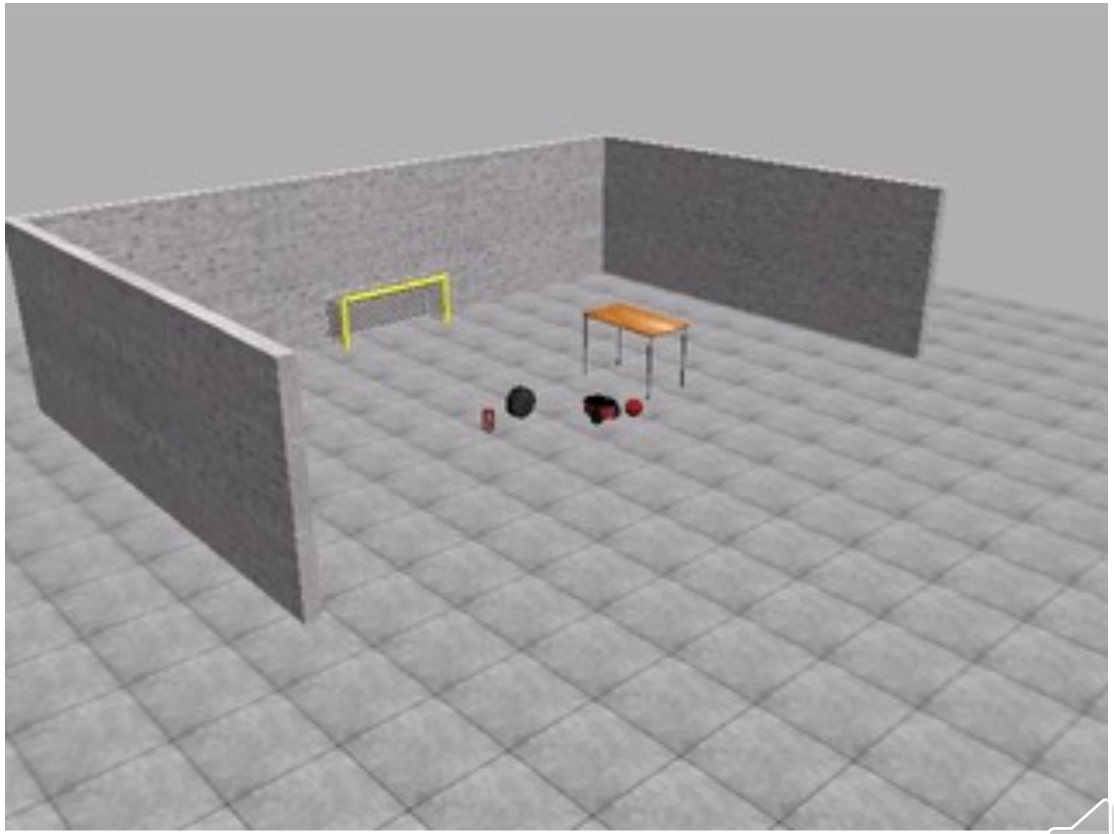
GoalTask


GoalTask Success Plot
1.0
0.8
0.6
0.4
0.2
0.0
Success
0
1
2
3
4
5
Num Timesteps
1e6
SAC
SAC+ActRepeat
HiPPO
HIRO
Action
StateAction
StateDiff



KickBall


KickBall Success Plot
SAC
SAC+ActRepeat
HiPPO
HIRO
Action
StateAction
StateDiff
Success
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
Num Timesteps
1e6



Summary

- HIDIO
  - Discovers diverse options while jointly learning to utilize them to solve a given sparse-reward task
  - Options are **task-agnostic**: no assumptions about task structure
  - Performs better than flat RL and other hierarchical RL methods