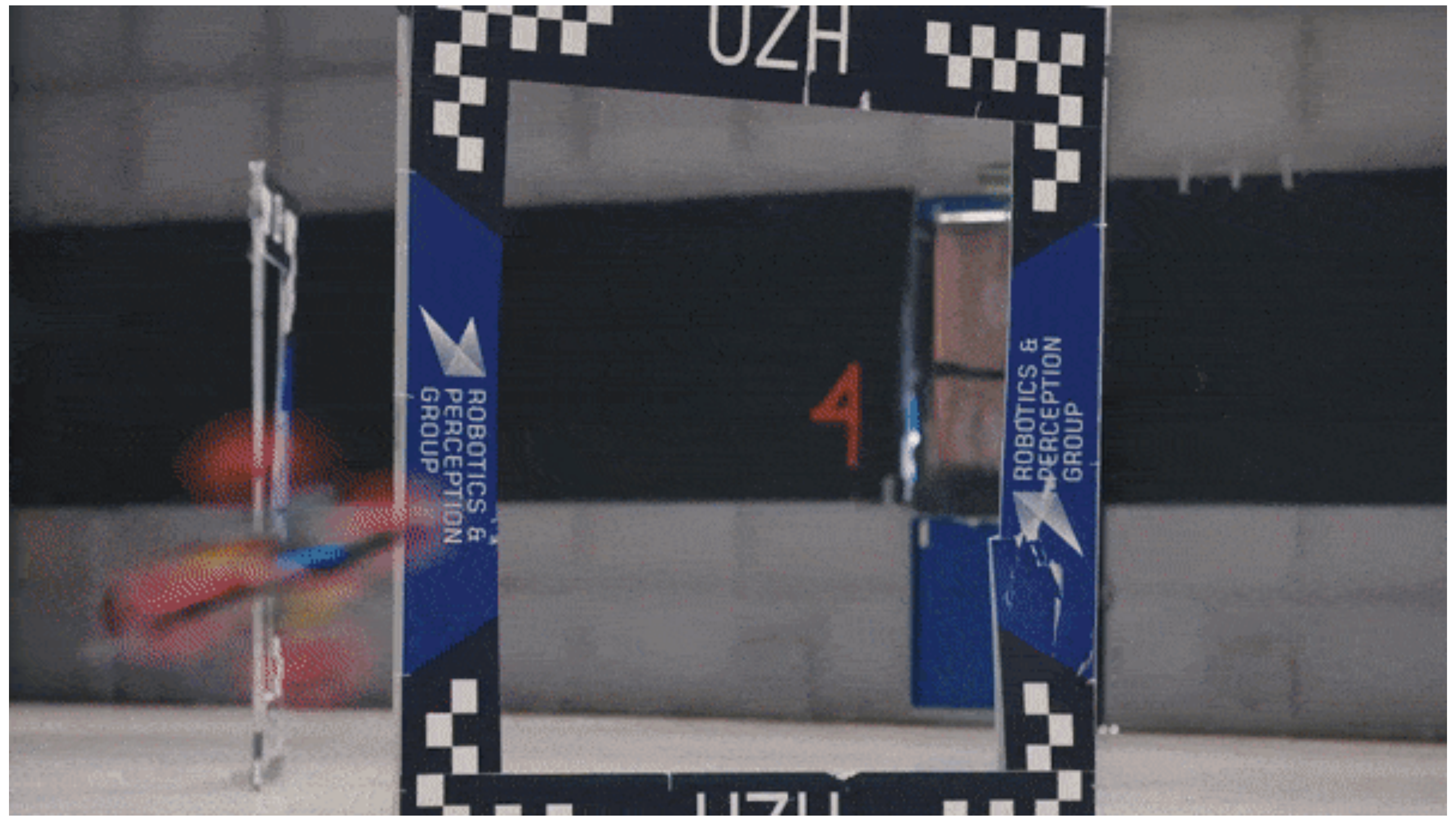
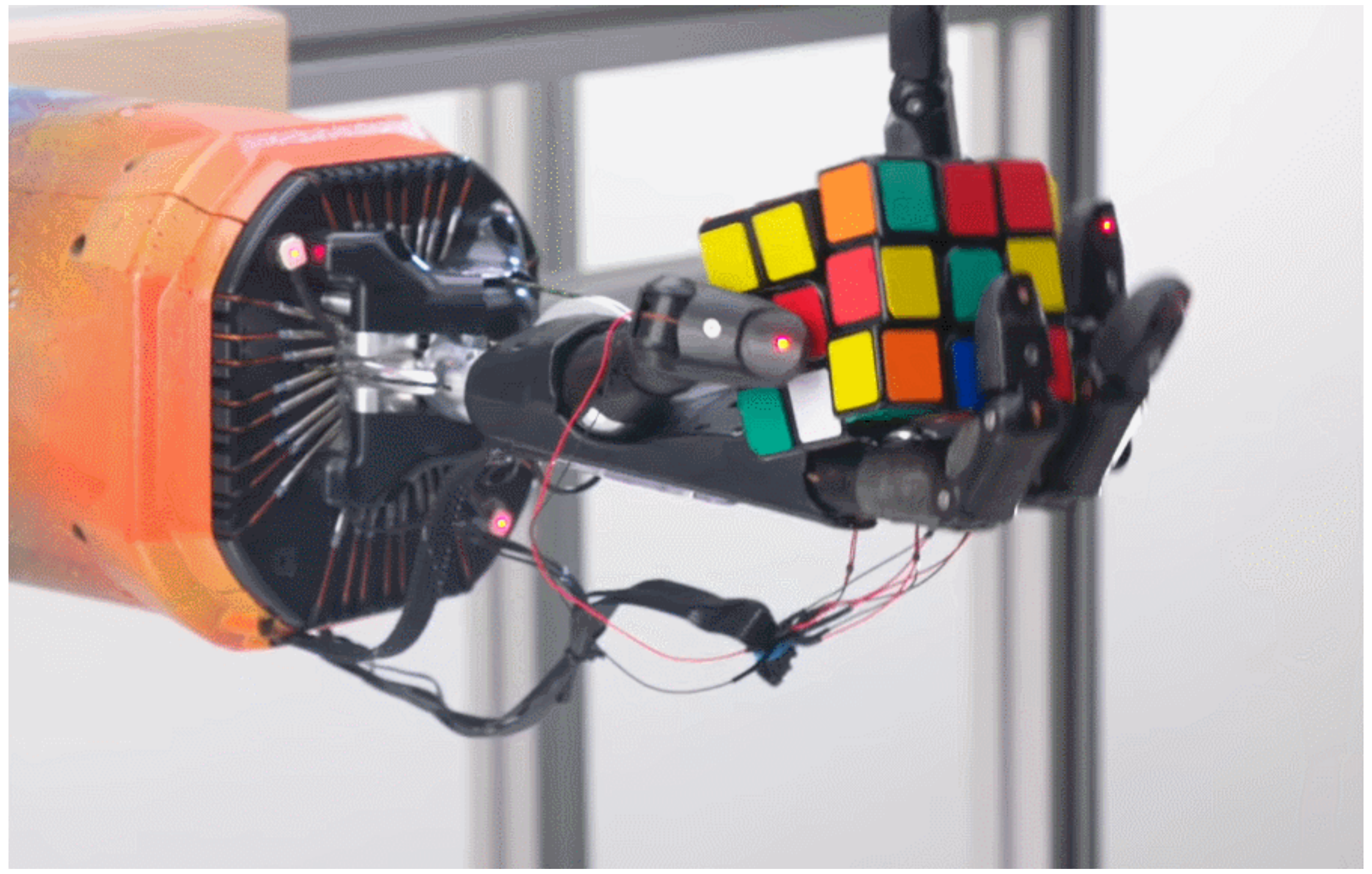


Song et al, Autonomous Drone Racing with Deep Reinforcement Learning (2021)

Rudin et al, Learning To Walk in Minutes (2021)


OpenAI, Solving Rubik's Cube with a Robot Hand (2019)


AHAC (ours)
PPO

Adaptive Horizon Actor-Critic for Policy Learning in Contact-Rich Differentiable Simulators

Ignat Georgiev, Krishnan Srinivasan, Jie Xu, Eric Heiden, Animesh Garg

ICML 2024

Reinforcement Learning

$$\max_{\theta} J(\theta) := \max_{\theta} \mathbb{E}_{\substack{s_1 \sim \rho \\ a_h \sim \pi(\cdot \mid s_h)}} [R_H(s_1)]$$

Zeroth-Order Model-Free

- no assumptions over dynamics
- policy $\pi_\theta(\cdot \mid s_h)$ is trained via the Policy Gradients Theorem

$$\nabla_\theta^{[0]} J(\theta) := \mathbb{E}_{a_h \sim \pi_\theta(\cdot|s_h)}[R_H(s_1) \sum_{h=1}^{H} \nabla_\theta \log \pi_\theta(a_h \mid s_h)]$$

- Works well with little assumptions!
- Widely considered to
  - Solve complex tasks
  - Notoriously sample inefficient

First-Order Model-Based

- Assumes dynamics are known (usually learned)
- Policy $\pi_\theta(\cdot \mid s_h)$ is trained via analytical gradients through the model

$$\nabla_\theta^{[1]} J(\theta) := \mathbb{E}_{a_h \sim \pi_\theta(\cdot|s_h)}[\nabla_\theta R_H(s_1)]$$

- Requires more assumptions
- Widely considered to
  - Have less variance
  - Underperform against model-free

Learning through contact in differentiable simulation

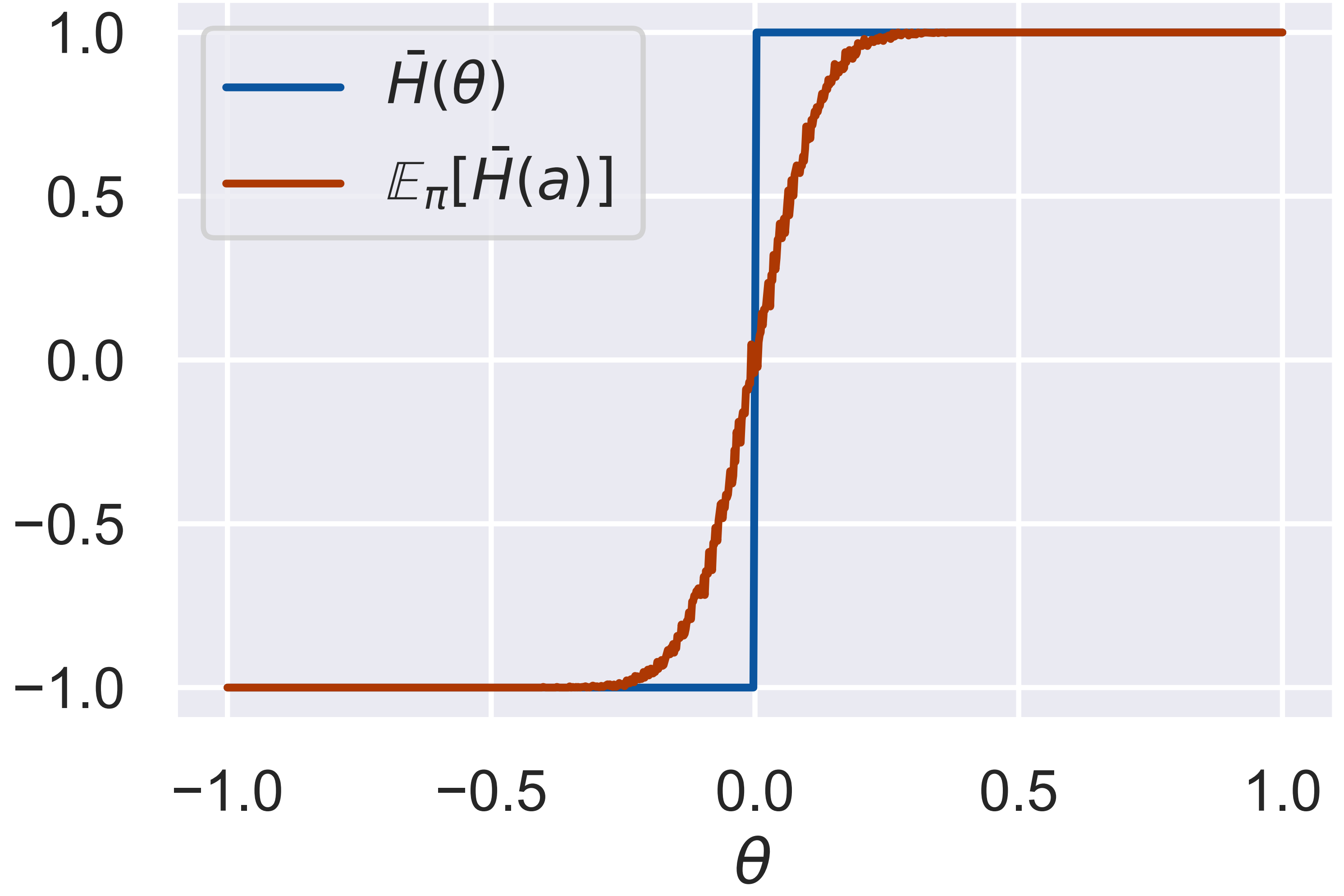
- Differentiable simulators enable differentiating through contact

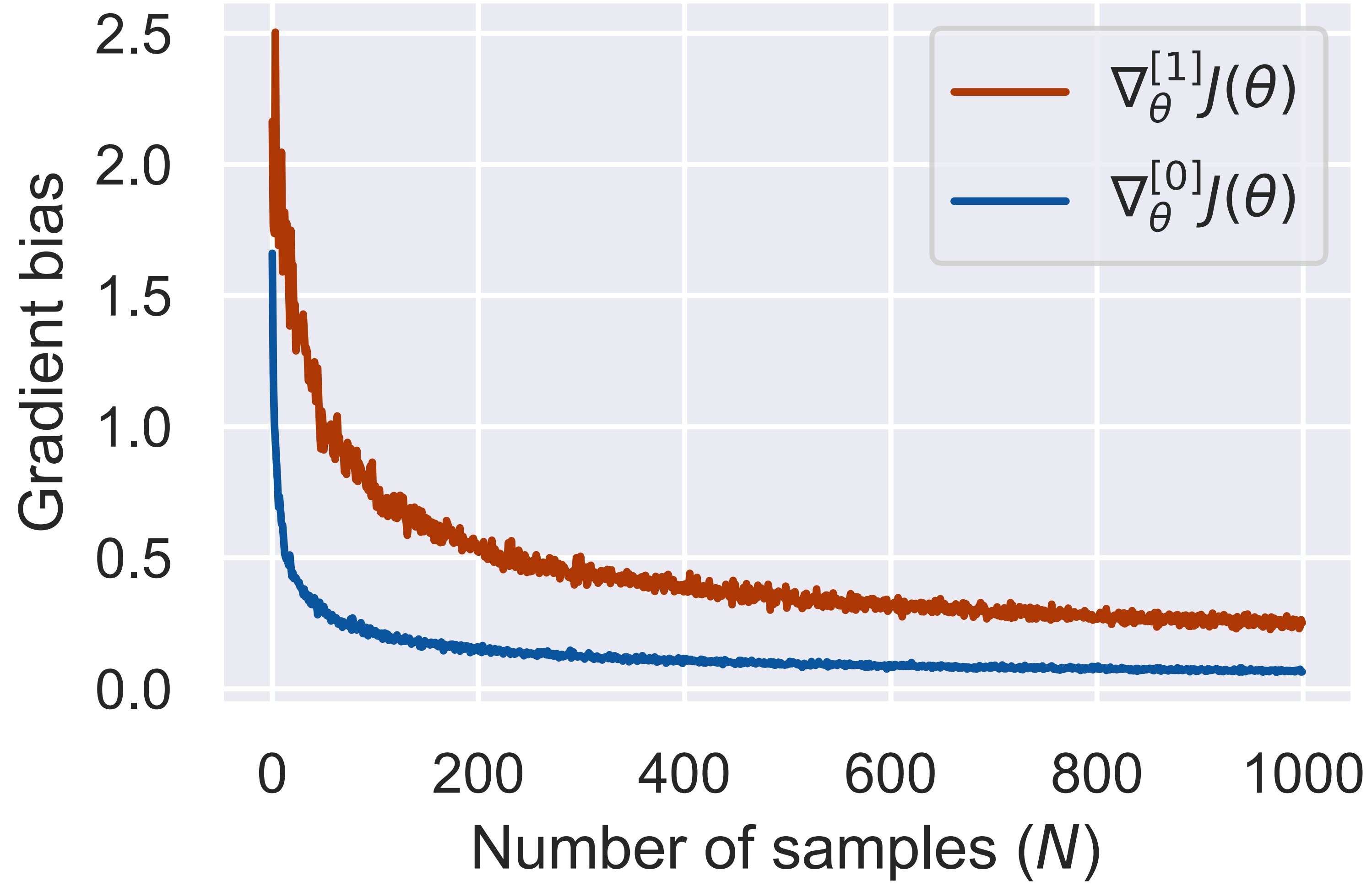
$$\bar{H}(x) = \begin{cases} 1 & x > \nu/2 \\ 2x/\nu & |x| \leq \nu/2 \\ -1 & x < -\nu/2 \end{cases}$$

$$x \sim \pi_\theta(\cdot) = \theta + w \qquad w \sim \mathcal{N}(0,\sigma^2)$$

$\nabla_\theta \mathbb{E}_\pi \bar{H}(a) \neq 0$ at $\theta = 0$

- However, $\nabla_\theta \bar{H}(a) = 0$ with some prob.
- Under finite samples N:
  - First-order grad bias is high with low samples
  - Zeroth-order grad bias is low throughout





Learning through contact

Assumption 2.7. The system dynamics $f(\boldsymbol{s}, \boldsymbol{a})$ and the reward $r(\boldsymbol{s}, \boldsymbol{a})$ are continuously differentiable $\forall \boldsymbol{s} \in \mathbb{R}^n, \forall \boldsymbol{a} \in \mathbb{R}^m$.

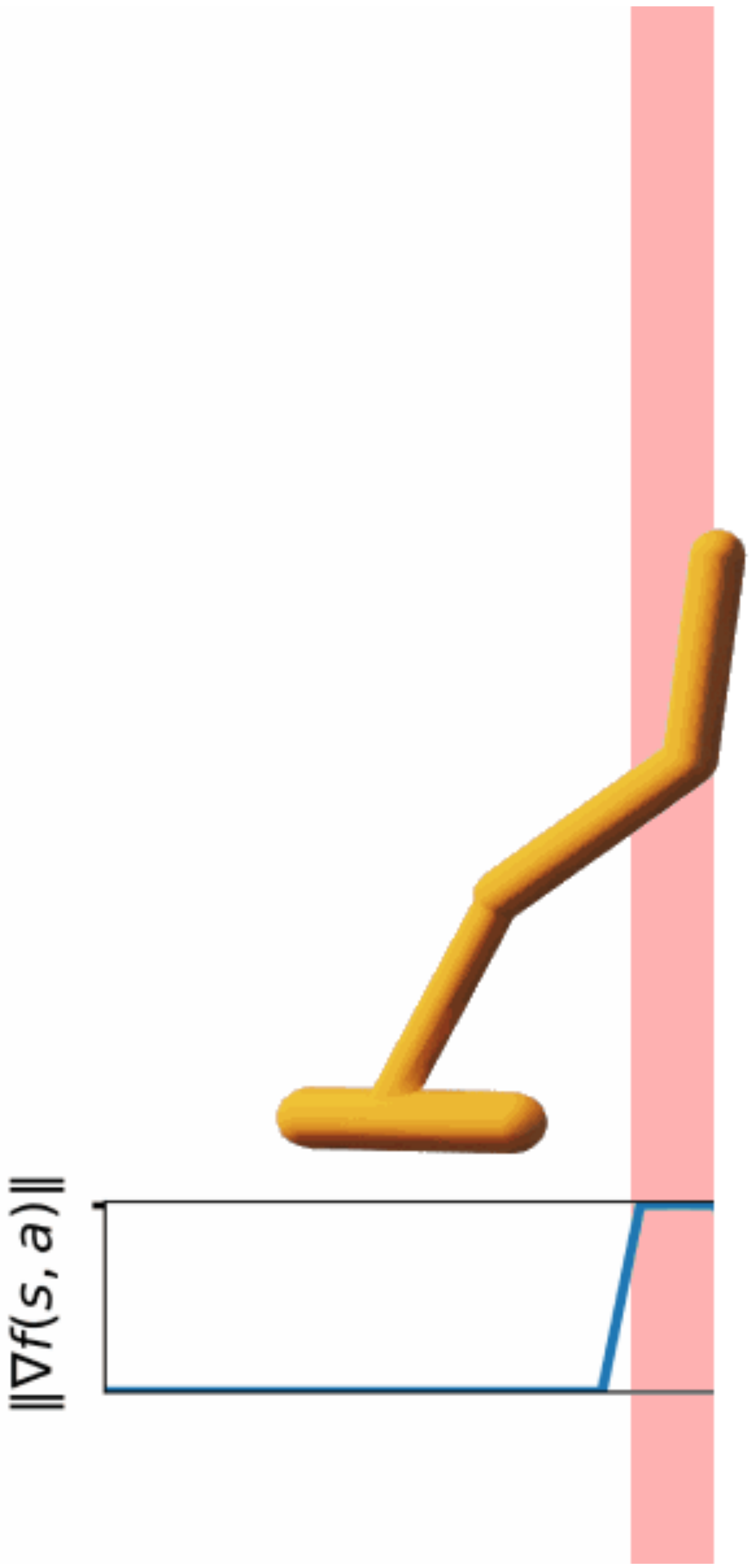
Lemma 3.1. For an H-step stochastic optimisation problem under Assumptions 2.7, which also has Lipshitz-smooth policies $||\nabla \pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})|| \leq B_\pi$ and Lipshitz-smooth and bounded rewards $r(\boldsymbol{s}, \boldsymbol{a}) \leq ||\nabla r(\boldsymbol{s}, \boldsymbol{a})|| \leq B_r$ $\forall \boldsymbol{s} \in \mathbb{R}^n; \boldsymbol{a} \in \mathbb{R}^m; \boldsymbol{\theta} \in \mathbb{R}^d$, then zero-order estimates remain unbiased. However, first-order gradient exhibit bias which is bounded by:

$$\|\mathbb{E}\left[\nabla_\theta^{[1]} J(\theta)\right] - \mathbb{E}\left[\nabla_\theta^{[0]} J(\theta)\right]\| = \leq H^4 B_r^2 B_\pi^2 \, \mathbb{E}_{a \sim \pi} \prod_{t=1}^{H} \|\nabla f(s_t, a_t)\|^2$$

Bias

1. Stiff contact approximation leads to high first-order gradient bias
2. The longer the horizon, the higher the bias

Stop trajectory rollout


‖∇f(s, a)‖

Adaptive Horizon Actor Critic (AHAC)

Building on Short Horizon Actor Critic (SHAC)

Xu et al. Accelerated Policy Learning with Parallel Differentiable Simulation (2022)

Algorithm 1 Adaptive Horizon Actor-Critic

while episode not done **do**
  for $h = 0, 1, .., H$ **do**
    $\boldsymbol{a}_{t+h} \sim \pi_{\boldsymbol{\theta}}(\cdot | \boldsymbol{s}_{t+h})$
    $\boldsymbol{s}_{t+h+1} = f(\boldsymbol{s}_{t+h}, \boldsymbol{a}_{t+h})$
  end for
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \mathcal{L}_\pi(\boldsymbol{\theta}, \boldsymbol{\phi})$
  $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \alpha_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} \mathcal{L}_\pi(\boldsymbol{\theta}, \boldsymbol{\phi})$
  $H \leftarrow H - \alpha_\phi \sum_{t=0}^{H} \phi_t$
  while not converged **do**
    $\psi \leftarrow \psi - \alpha_\psi \nabla_\psi \mathcal{L}_V(\psi)$
  end while
end while

Rollout

Actor Training

Critic Training

$$J(\theta) := \sum_{h=t}^{t+H-1} \gamma^{h-t} r(s_h, a_h) + \gamma^t V_p si(s_{t+H})$$

$$s.t. \quad \|\nabla f(s_t, a_t)\| \le C \quad \forall t \in \{0,..,H\}$$

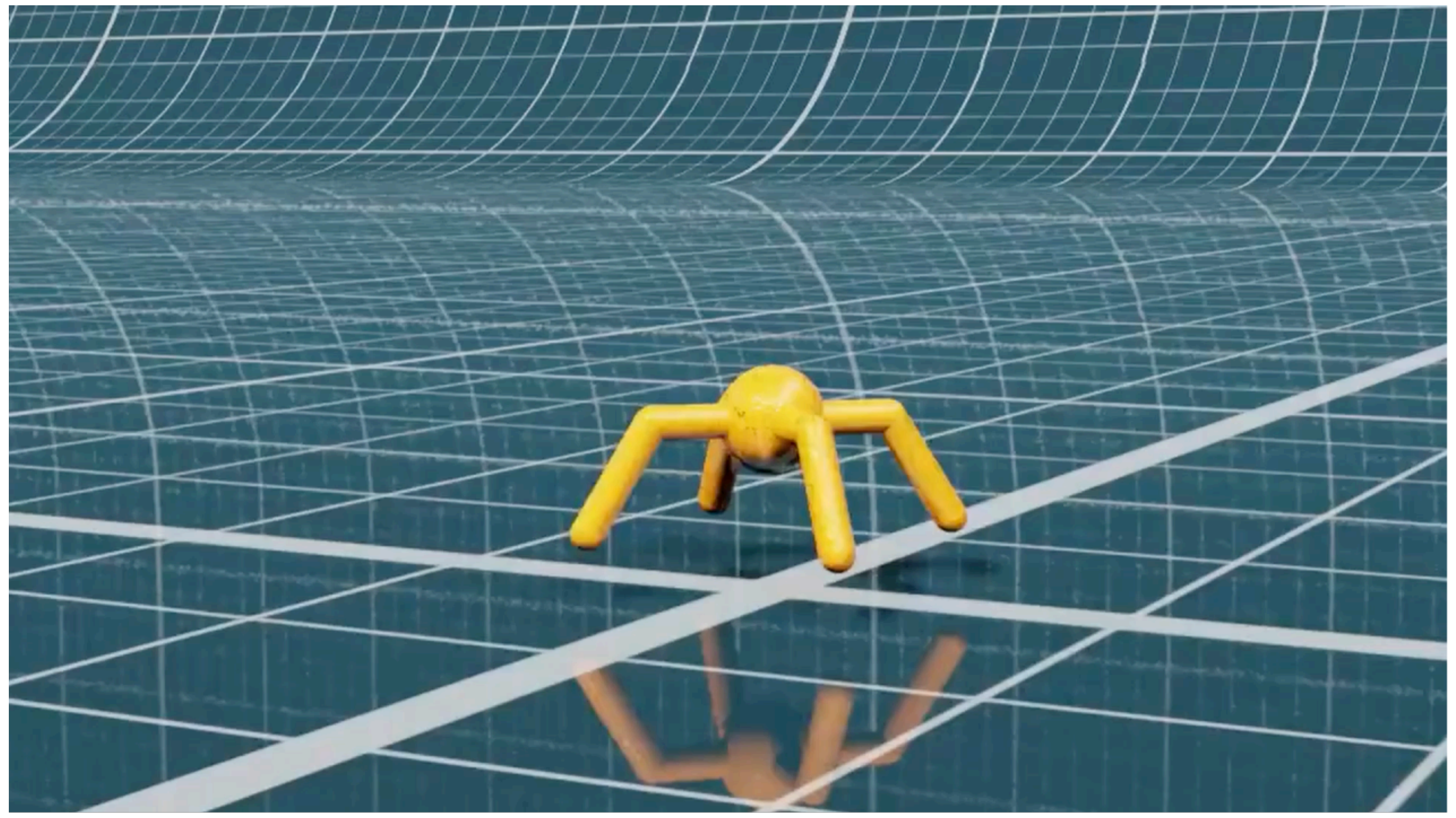
$$\mathcal{L}_\pi(\theta, \phi) = \sum_{h=t}^{t+H-1} \gamma^{h-t} r(s_h, a_h) + \gamma^t V_\psi(s_{t+H})$$

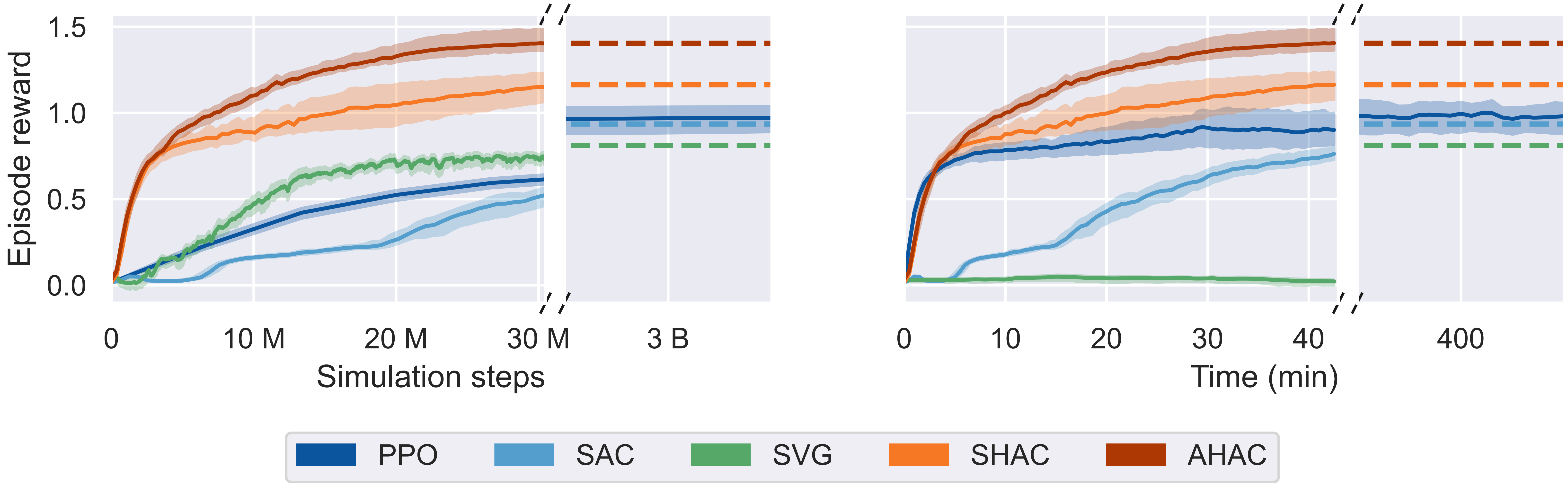
$$+\phi^T \left(\begin{bmatrix} \|\nabla f(s_t, a_t)\| \\ \vdots \\ \|\nabla f(s_{t+H}, a_{t+H})\| \end{bmatrix} - C \right)$$

$$\mathcal{L}_V(\psi) := \sum_{h=t}^{t+H} \|V_\psi(s_h) - \hat{V}(s_h)\|_2^2$$

Asymptotic performance

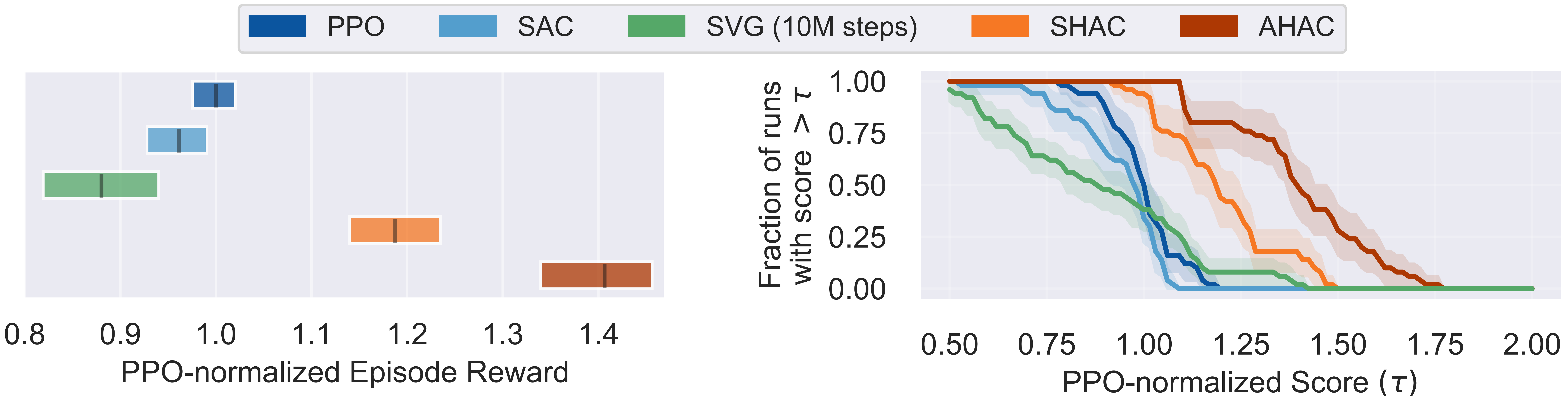
- Standard locomotion benchmarks
- Compare against zeroth-order (PPO and SAC) and first-order (SHAC and SVG) baselines






AHAC (ours)
PPO

Summary results across all tasks



- 50% Interquartile Mean (IQM) with 95% Confidence Interval (CI)
- AHAC achieves 40% higher reward than PPO across all tasks

More at: https://adaptive-horizon-actor-critic.github.io/