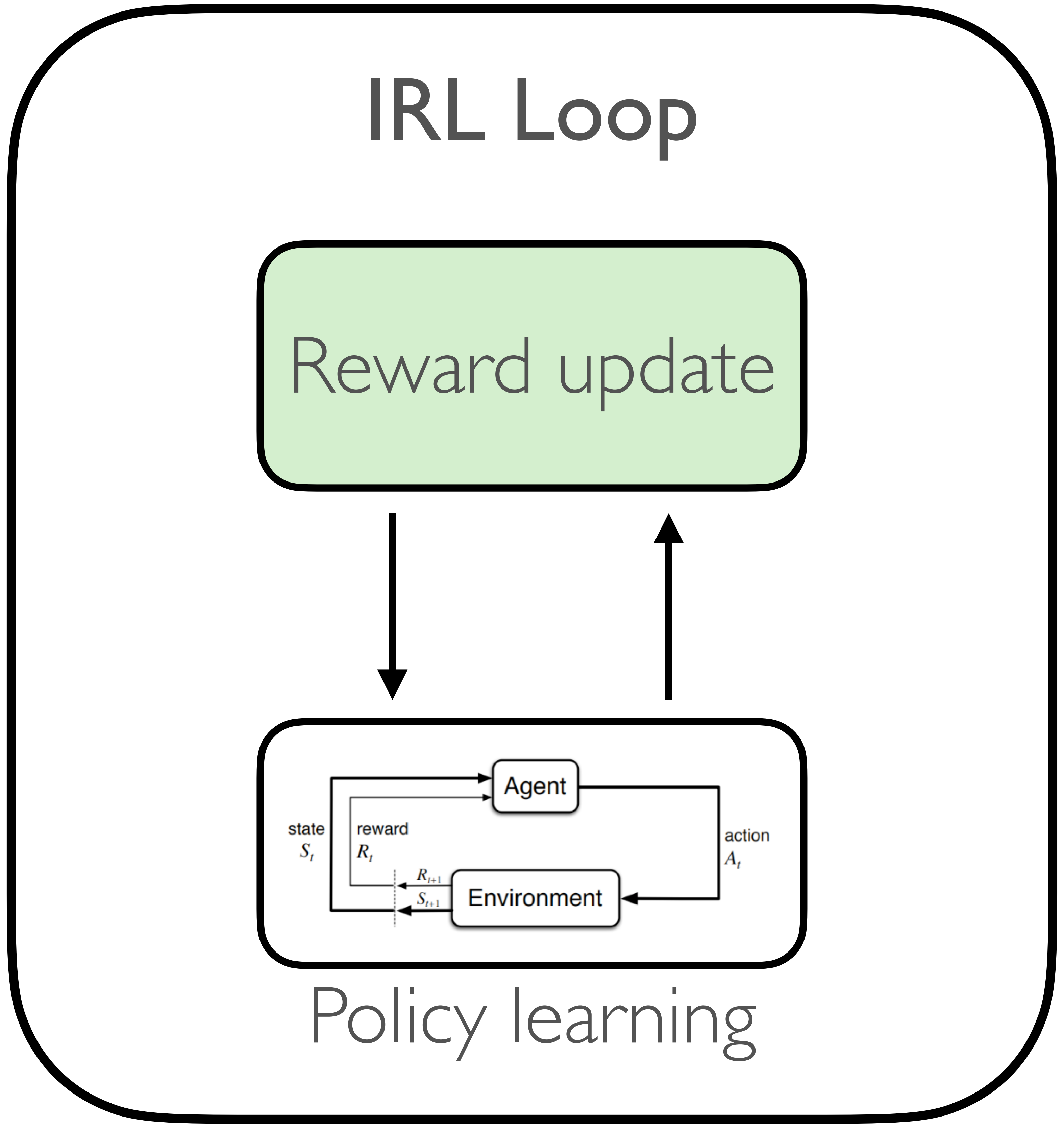
CS 690: Human-Centric Machine Learning

Prof. Scott Niekum

RLHF 1

Problems with standard inverse reinforcement learning

Policy learning in inner loop



Cannot outperform demonstrator

Standard assumption:

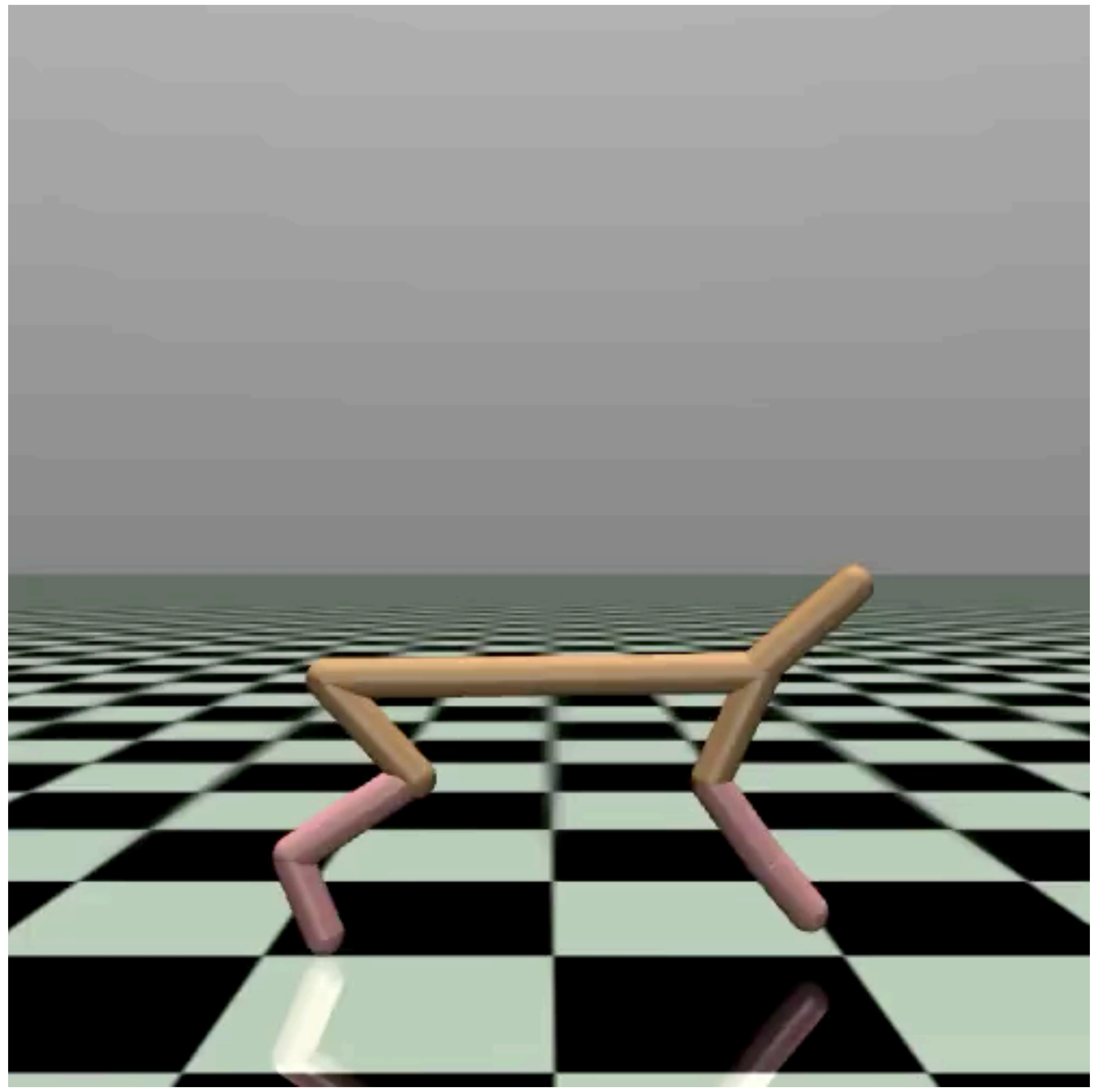
IRL should assume that the expert is near-optimal

↓

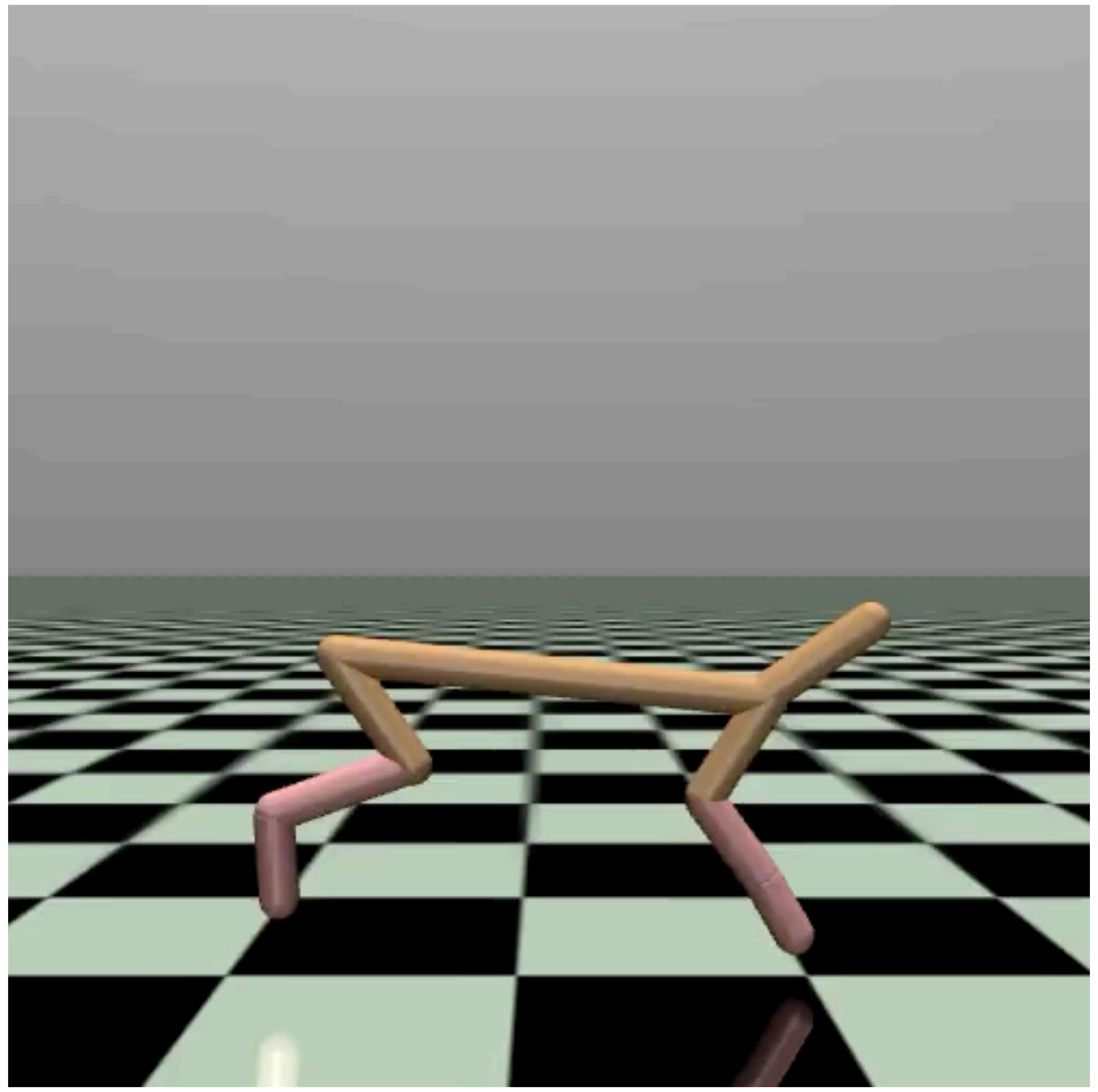
Ranked, suboptimal demonstrations provide significant computational and performance benefits

D.S. Brown, W. Goo, P. Nagarajan, and S. Niekum.
Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations.
International Conference on Machine Learning (ICML), June 2019.

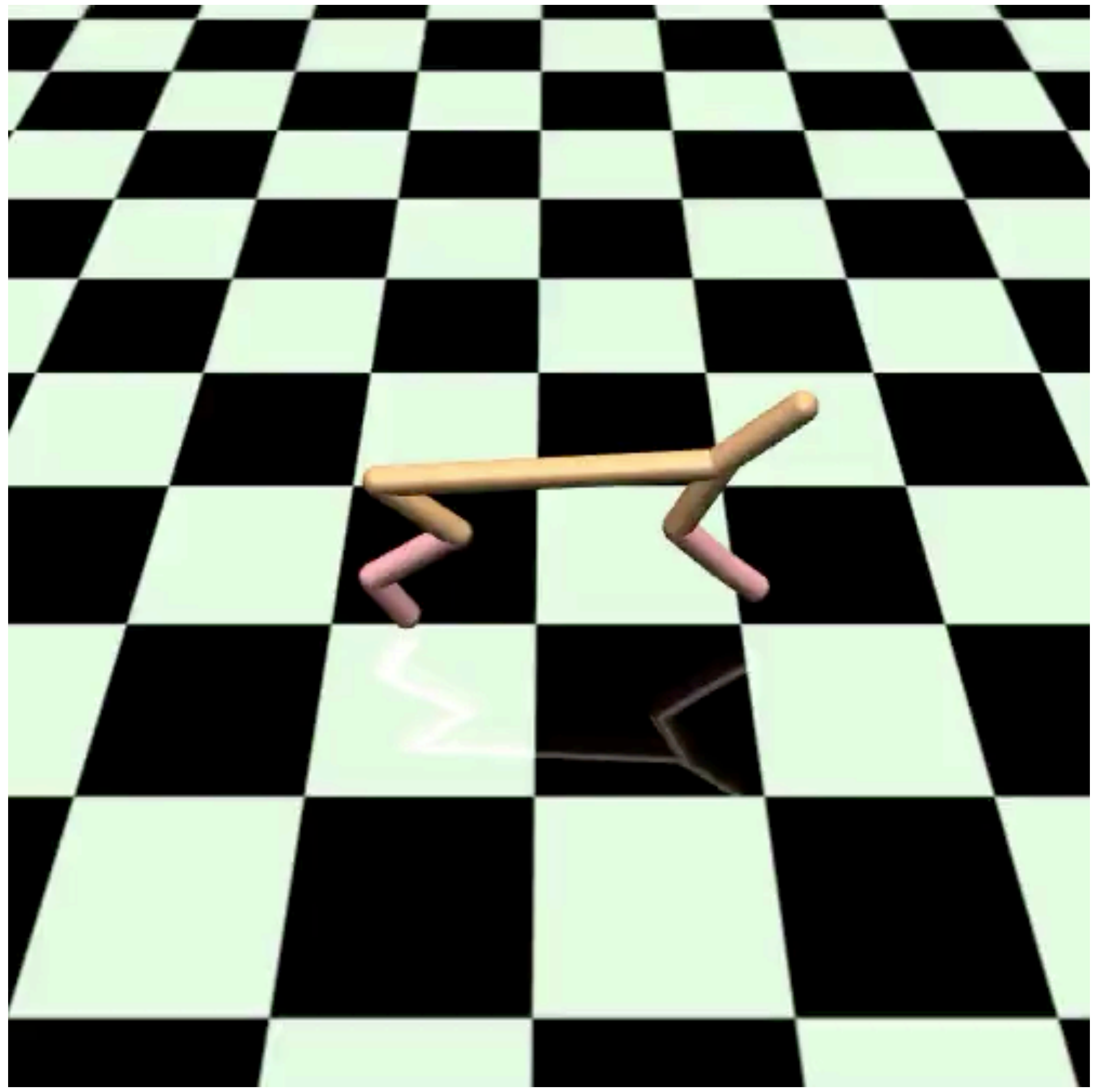
Ranked demonstrations: HalfCheetah



12.52

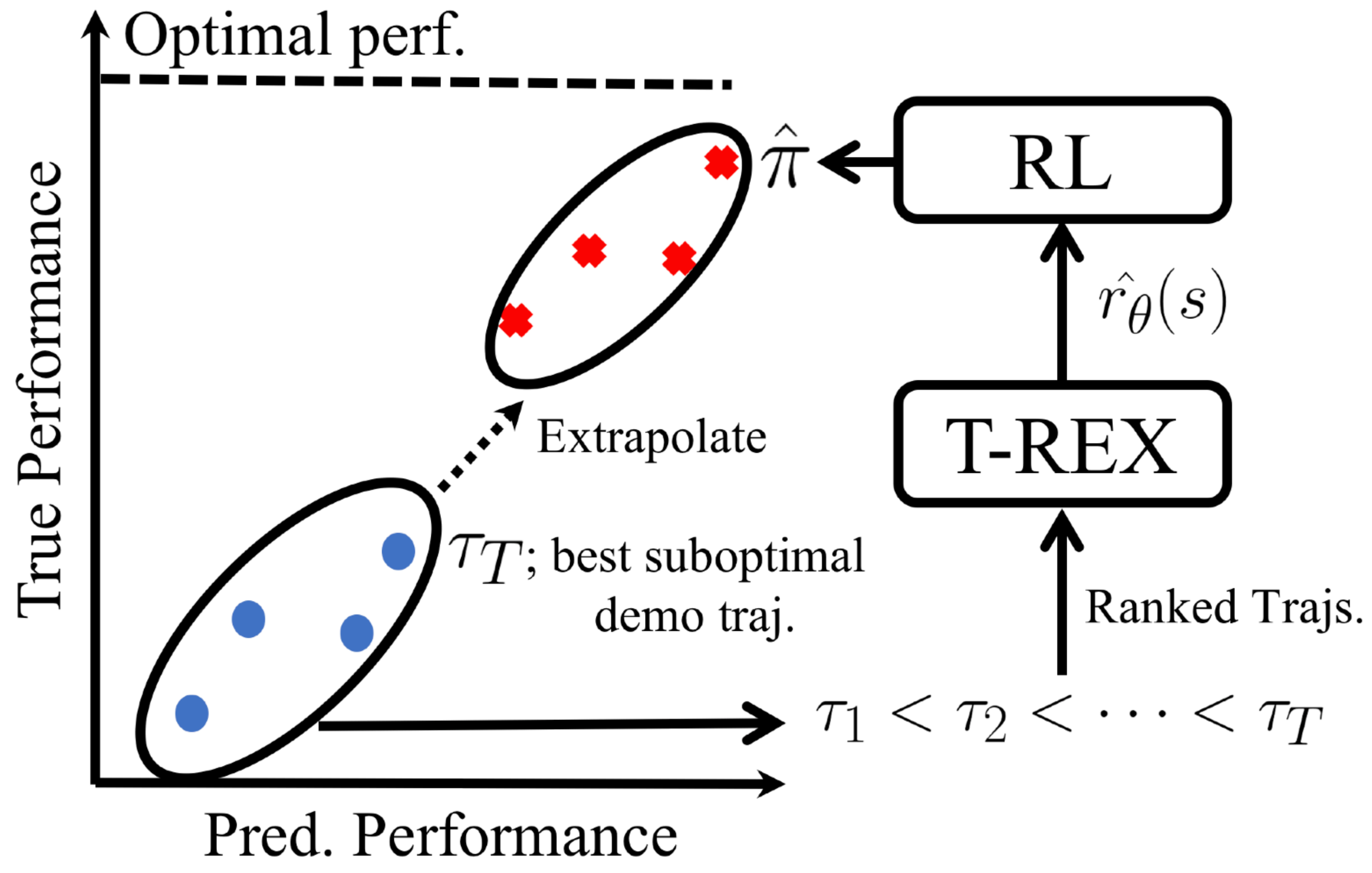


44.98



88.97

T-REX: Trajectory-ranked Reward Extrapolation

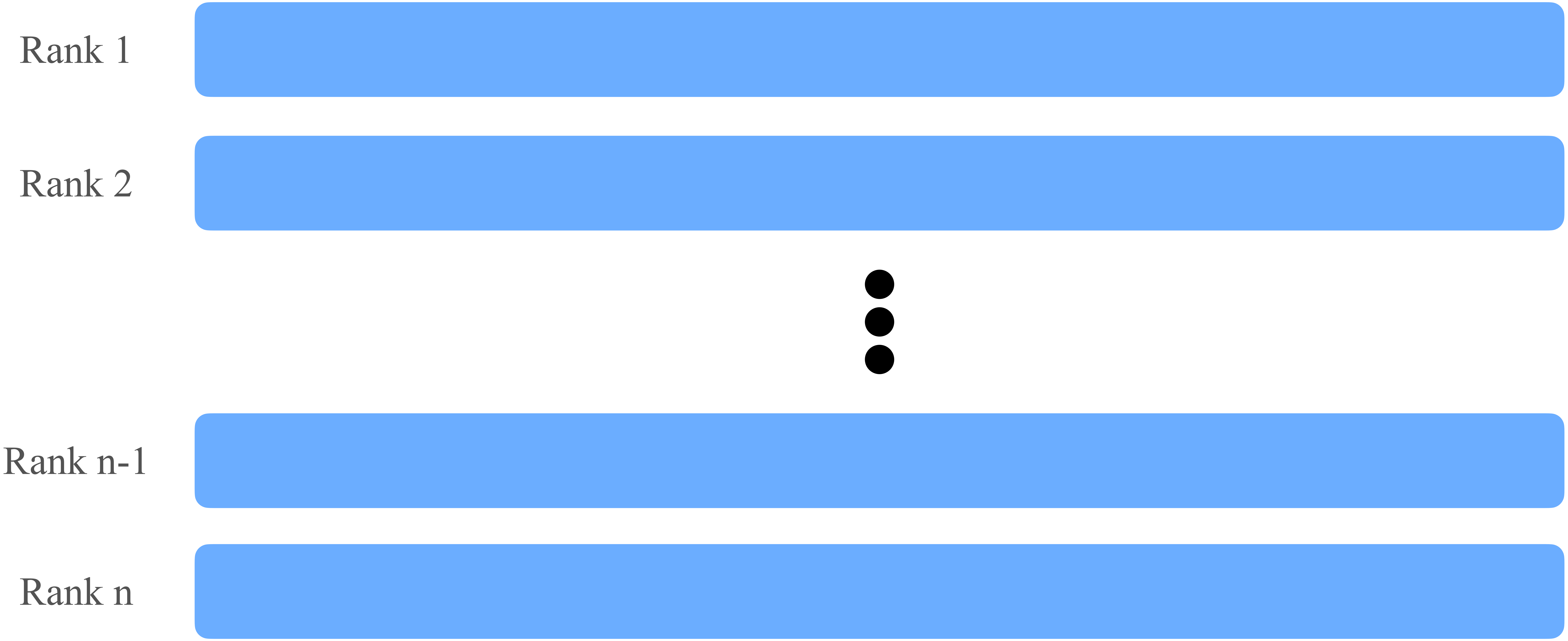


$$\mathcal{L}(\theta) = \mathbf{E}_{\tau_i,\tau_j\sim\Pi}\left[\xi\left(\hat{\mathrm{P}}\left(J_\theta(\tau_i) < J_\theta(\tau_j)\right), \tau_i \prec \tau_j\right)\right]$$

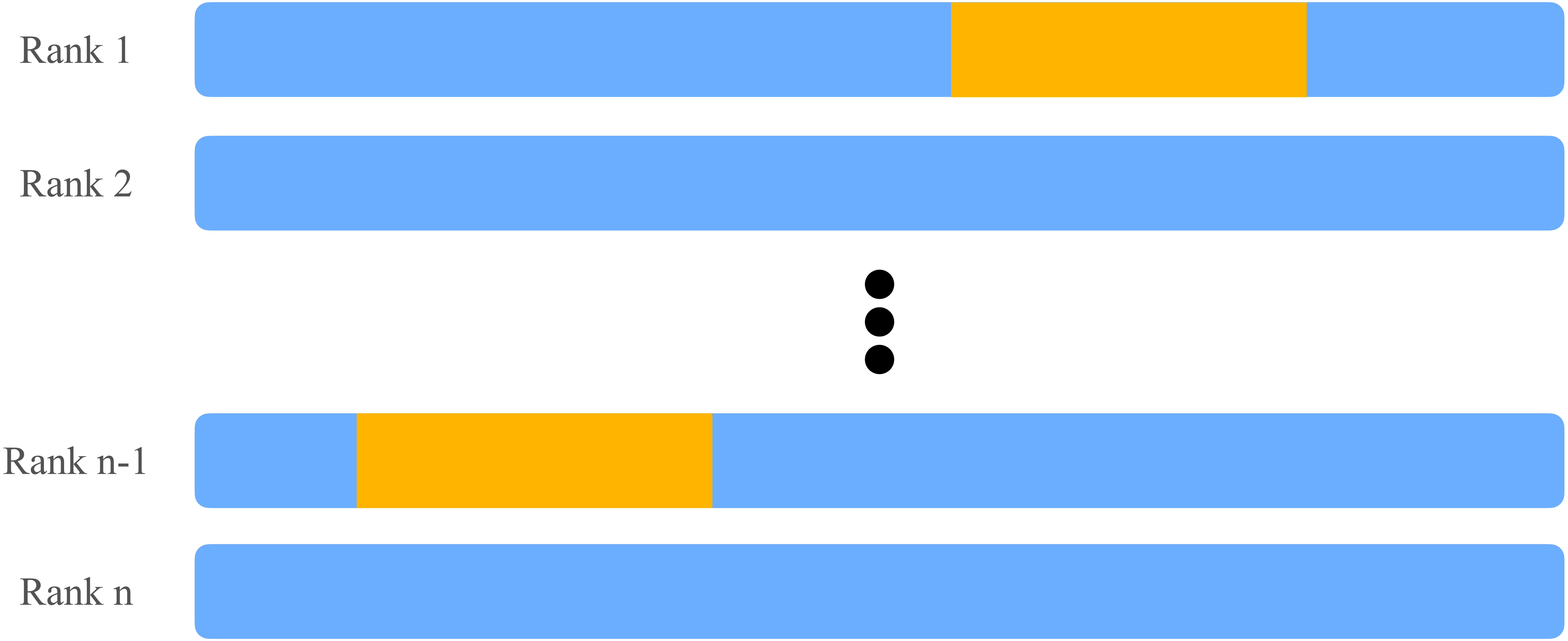
$$\hat{\mathrm{P}}\left(J_\theta(\tau_i) < J_\theta(\tau_j)\right) = \frac{\exp \sum_{s\in\tau_j} \hat{r}_\theta(s)}{\exp \sum_{s\in\tau_i} \hat{r}_\theta(s) + \exp \sum_{s\in\tau_j} \hat{r}_\theta(s)}$$

- Fully supervised — no policy learning
- No action labels required
- Extrapolation potential
- Works on high-dim (e.g. Atari) with ~10 demos

Data augmentation

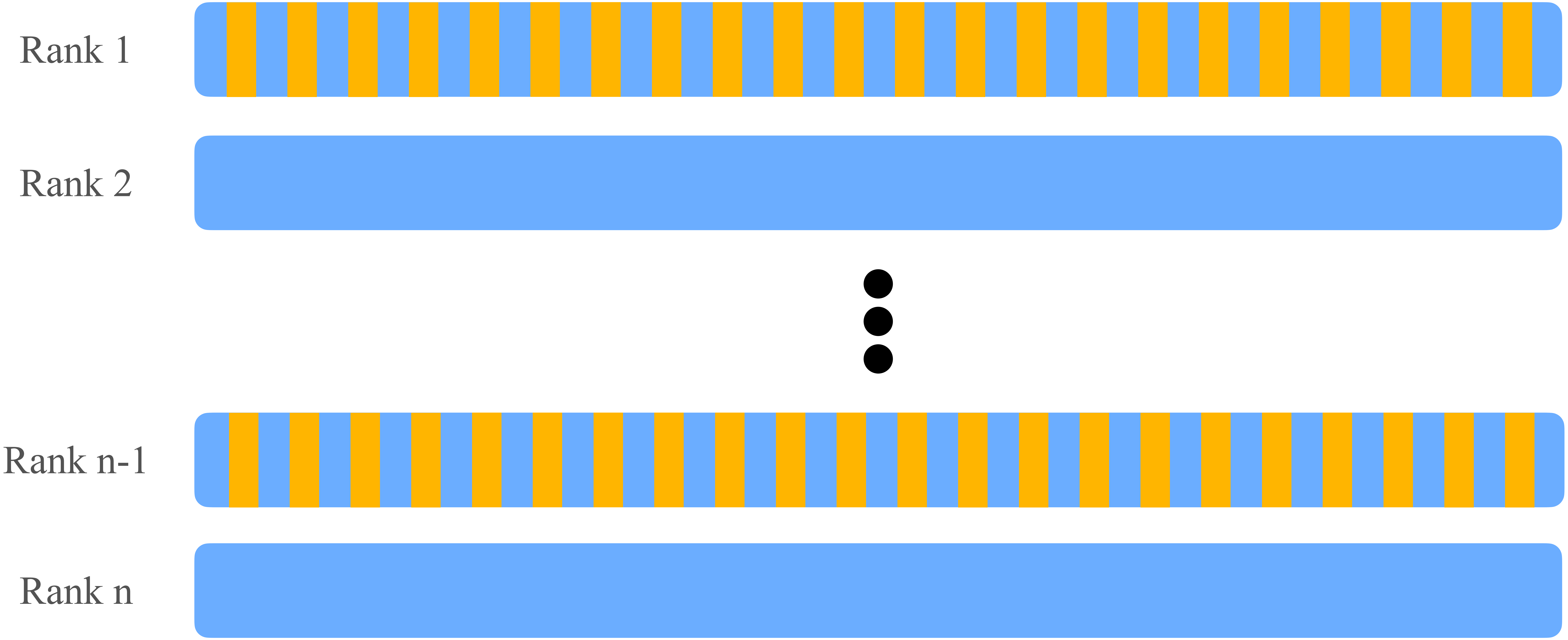


Data augmentation



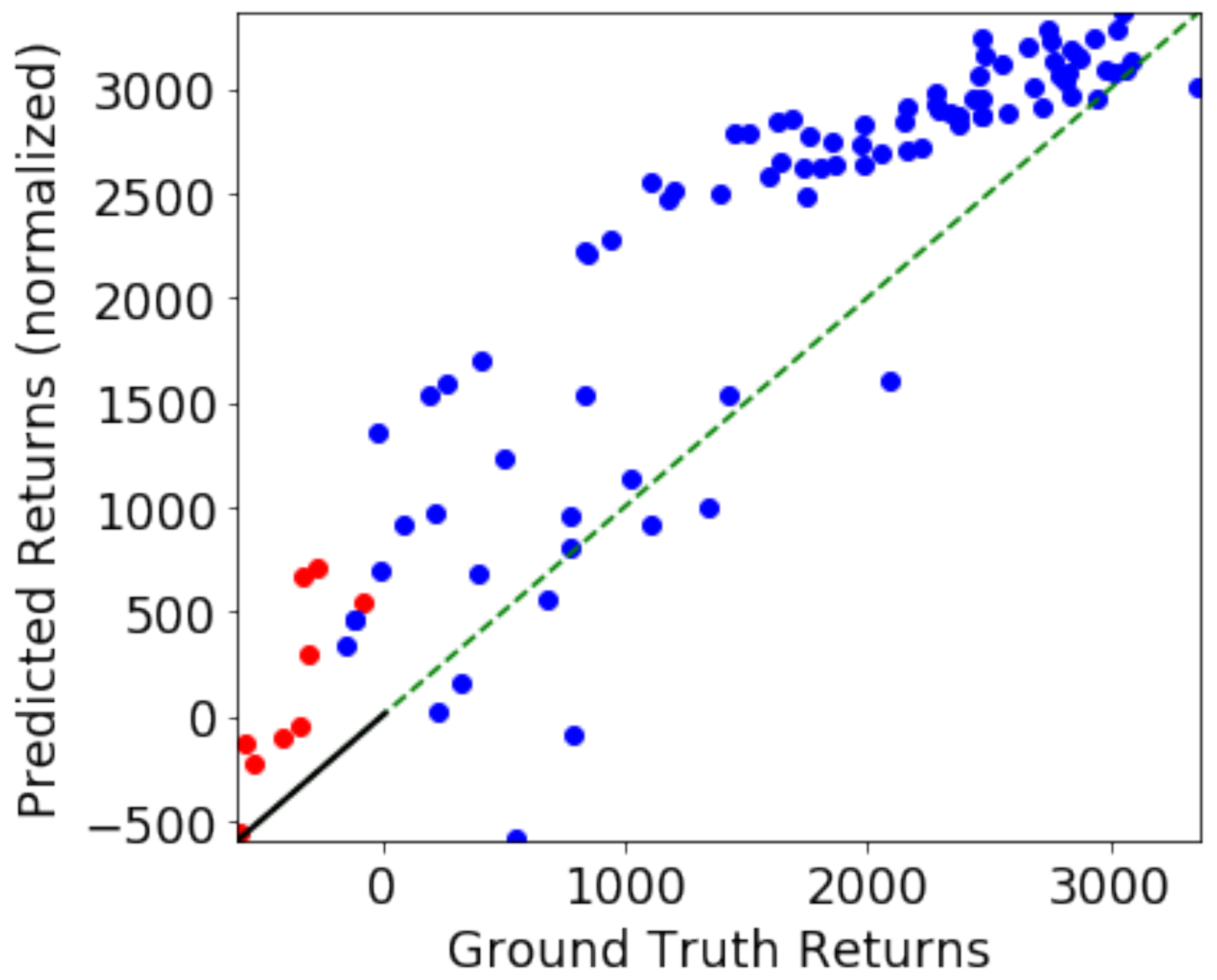
Subsampling

Data augmentation

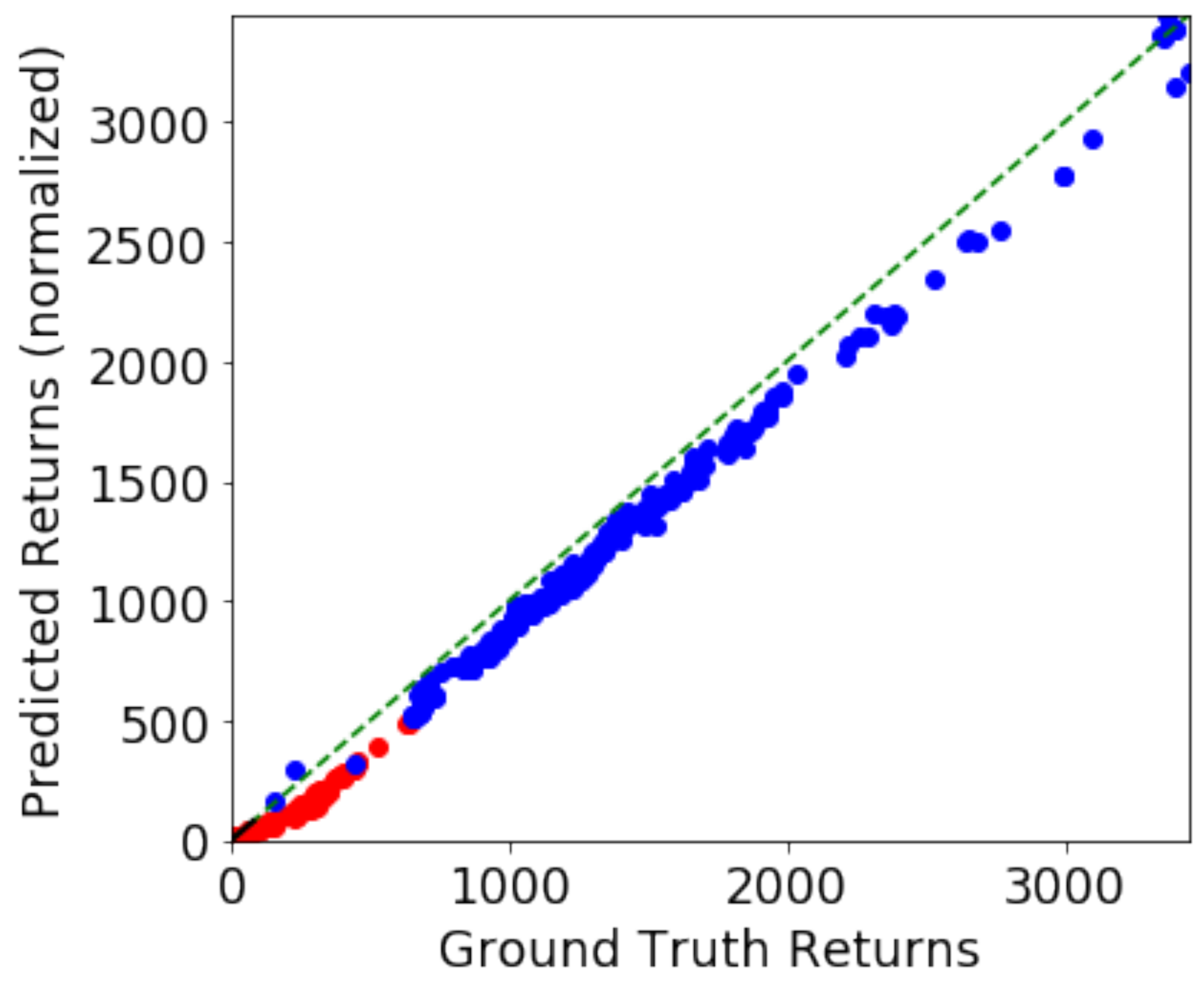


Frame skipping

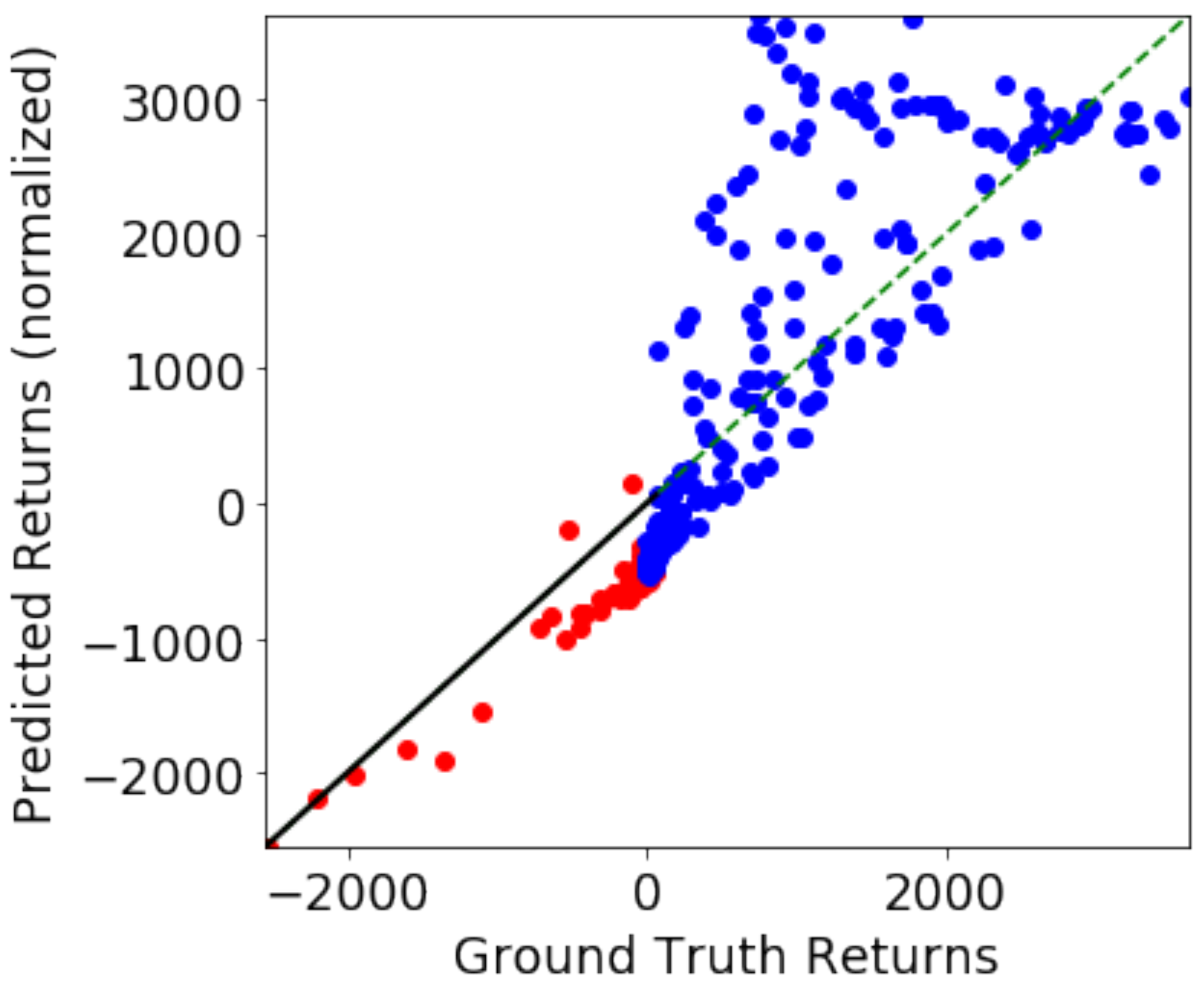
T-REX reward prediction



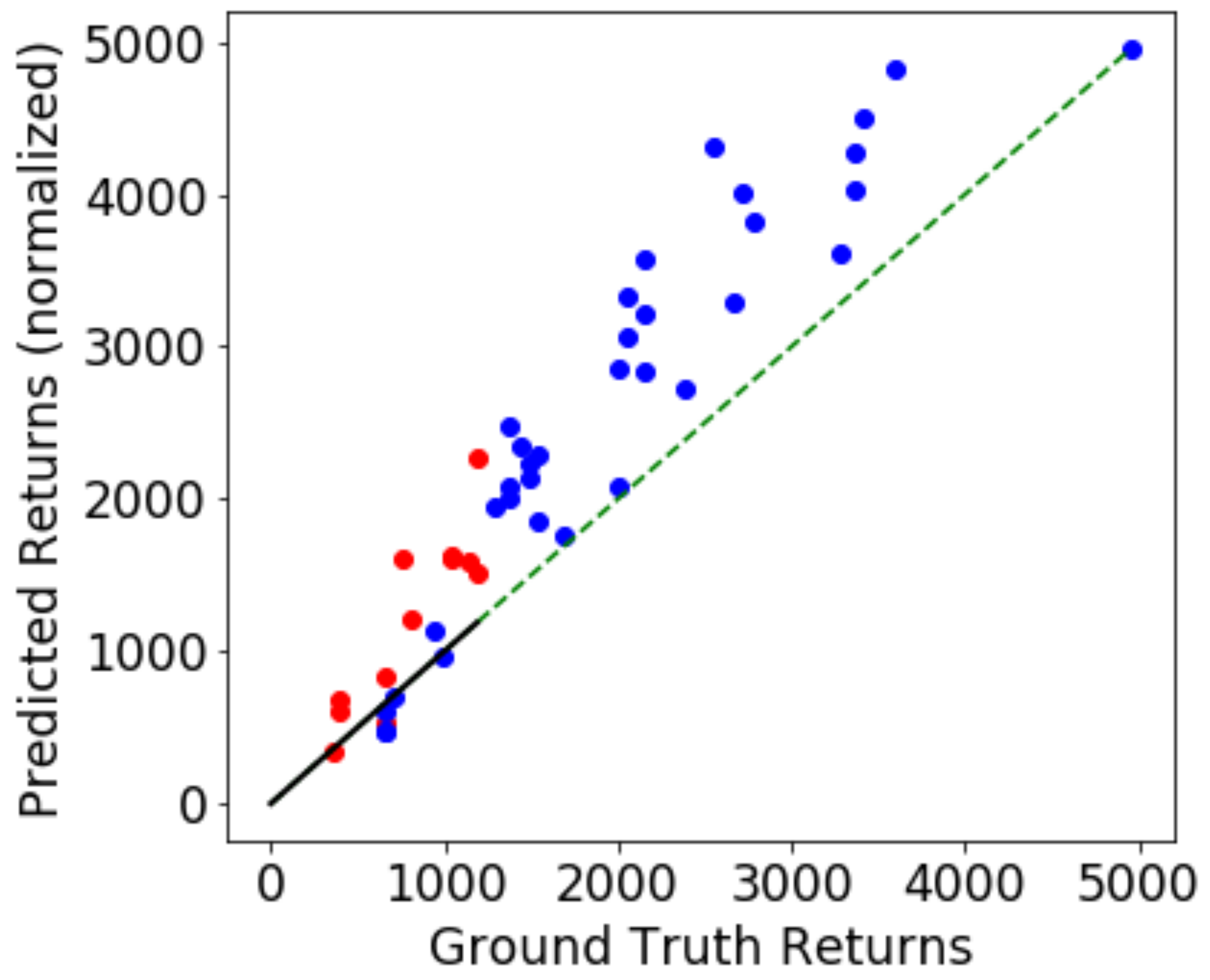
HalfCheetah



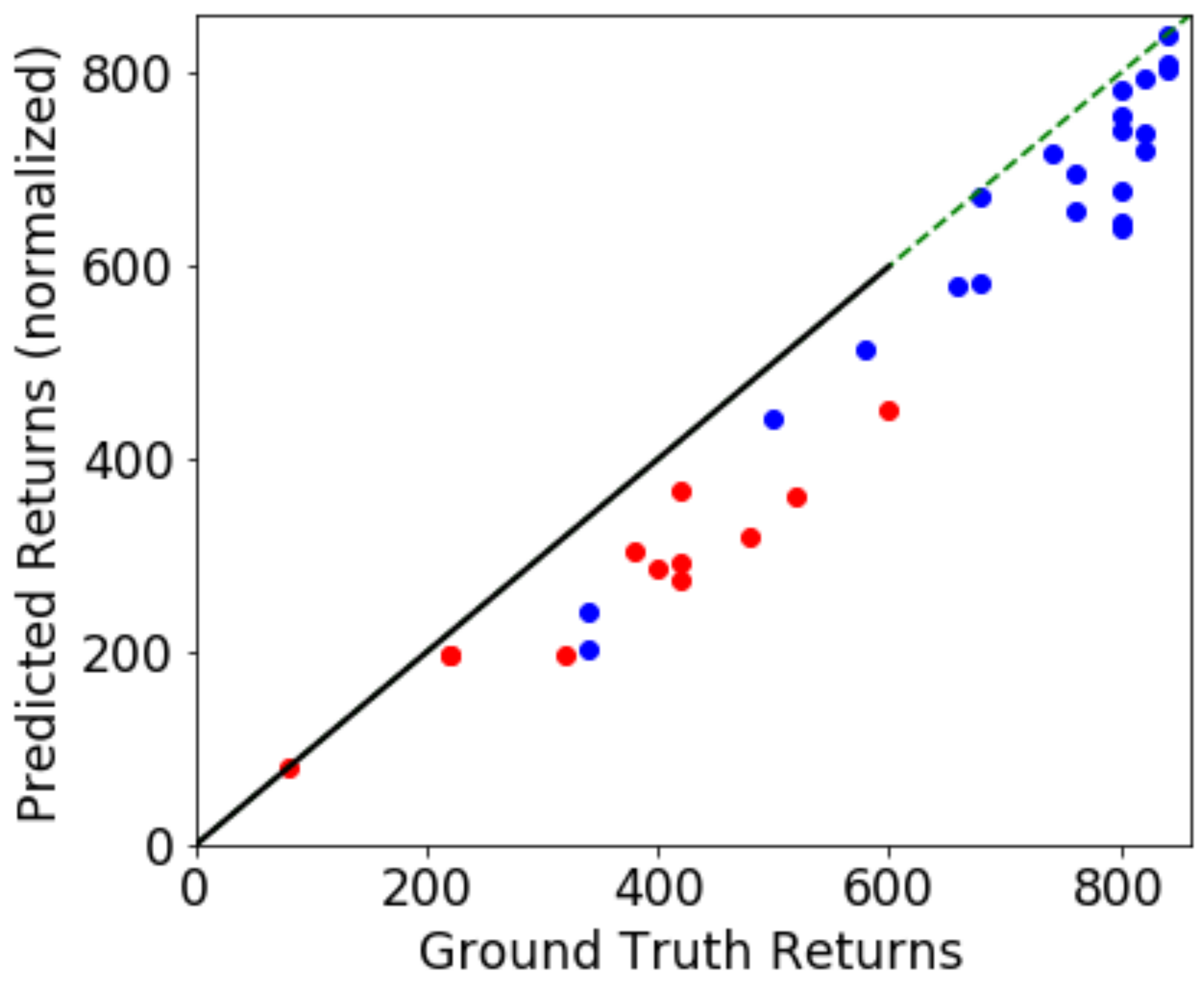
Hopper



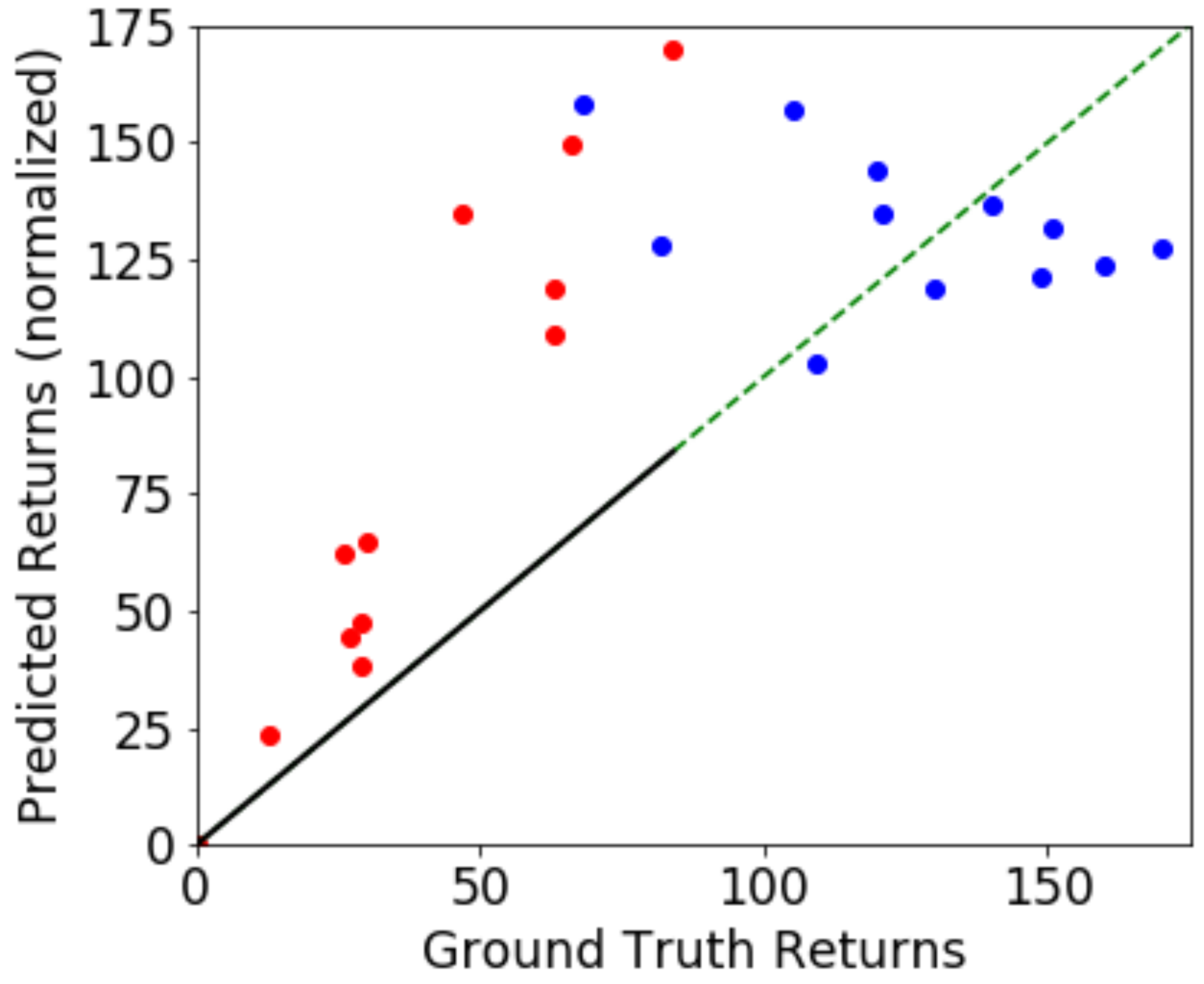
Ant



Beam Rider

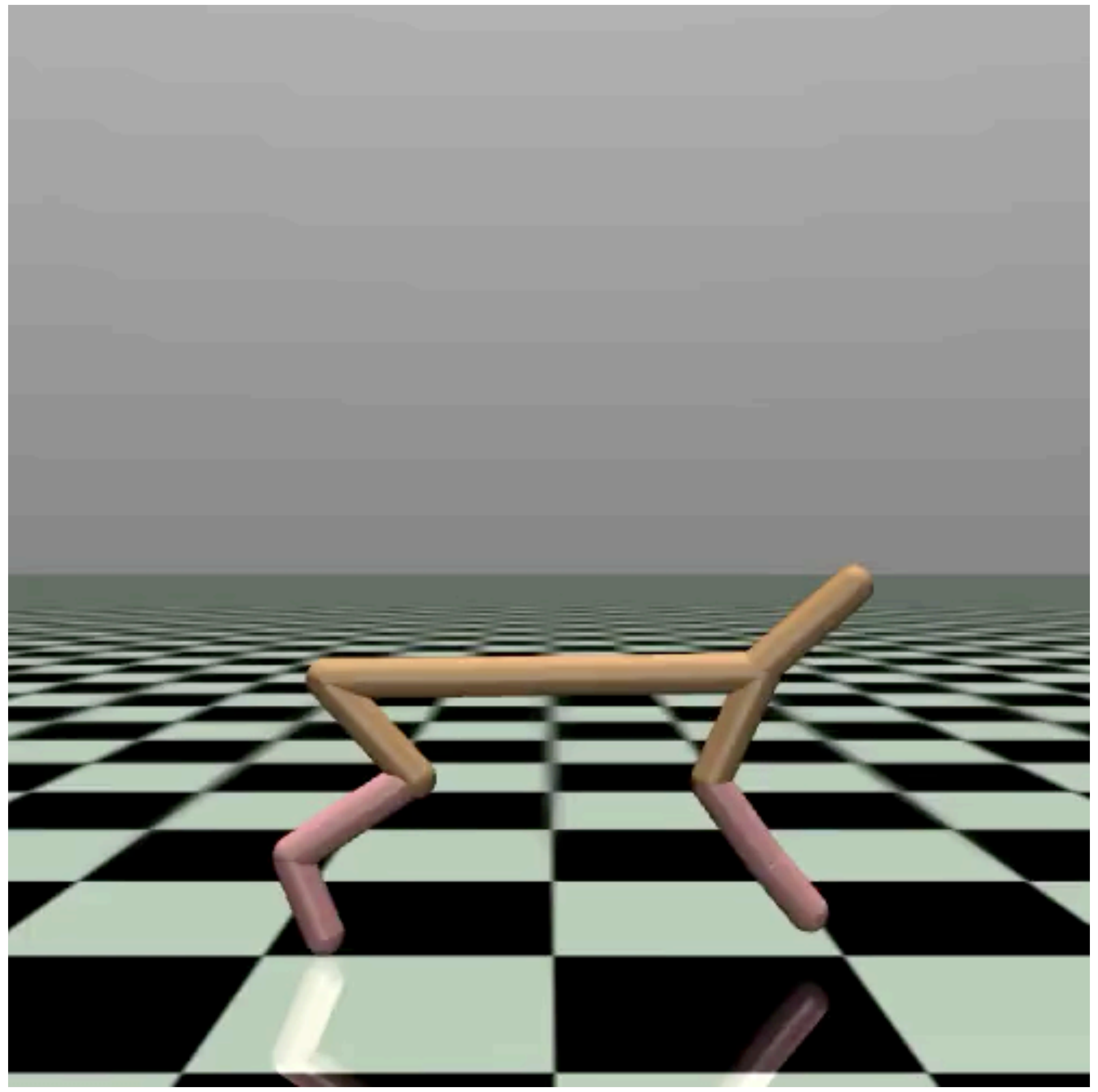


Seaquest

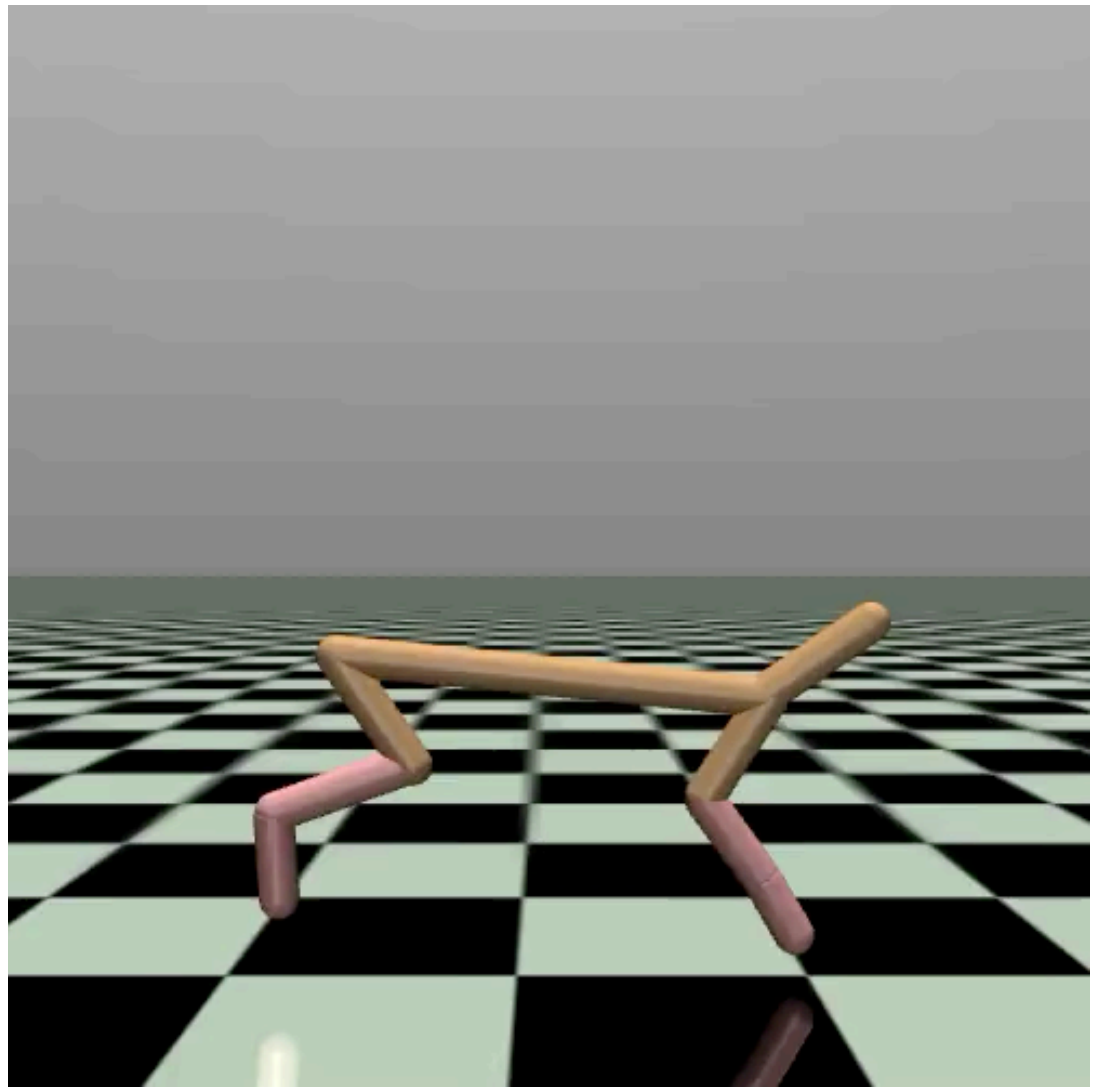


Enduro

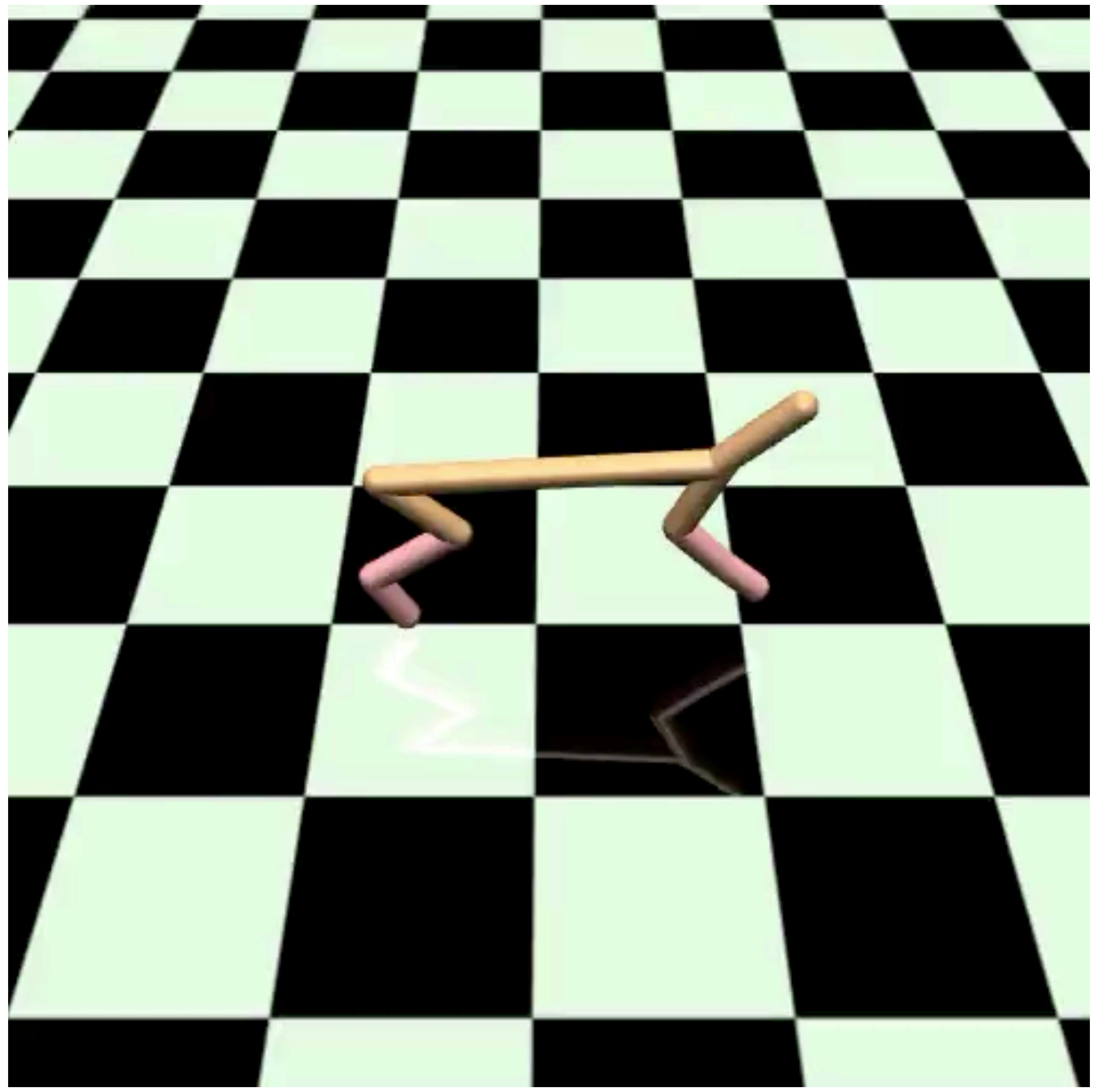
Ranked demonstrations: HalfCheetah



12.52

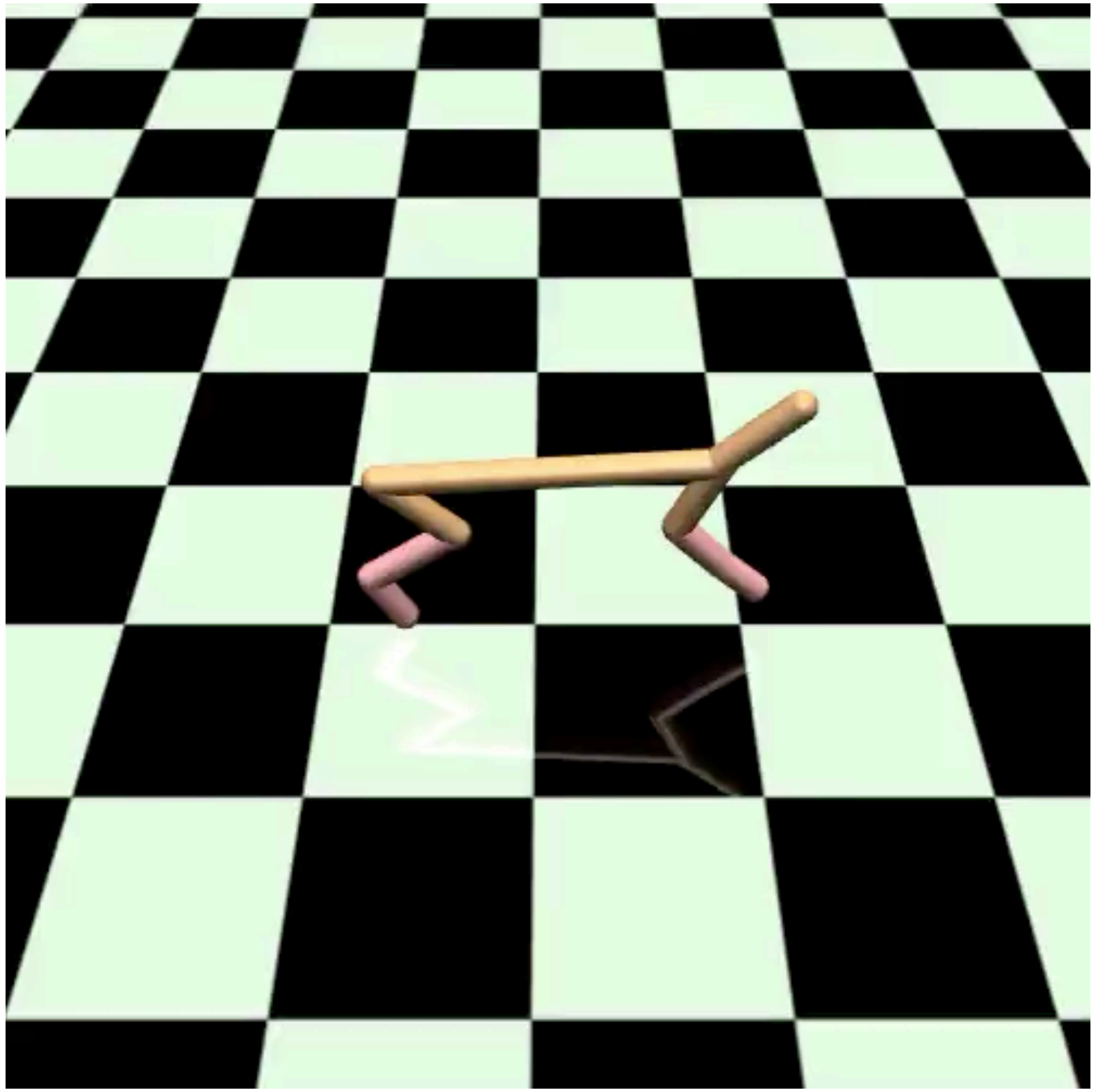


44.98

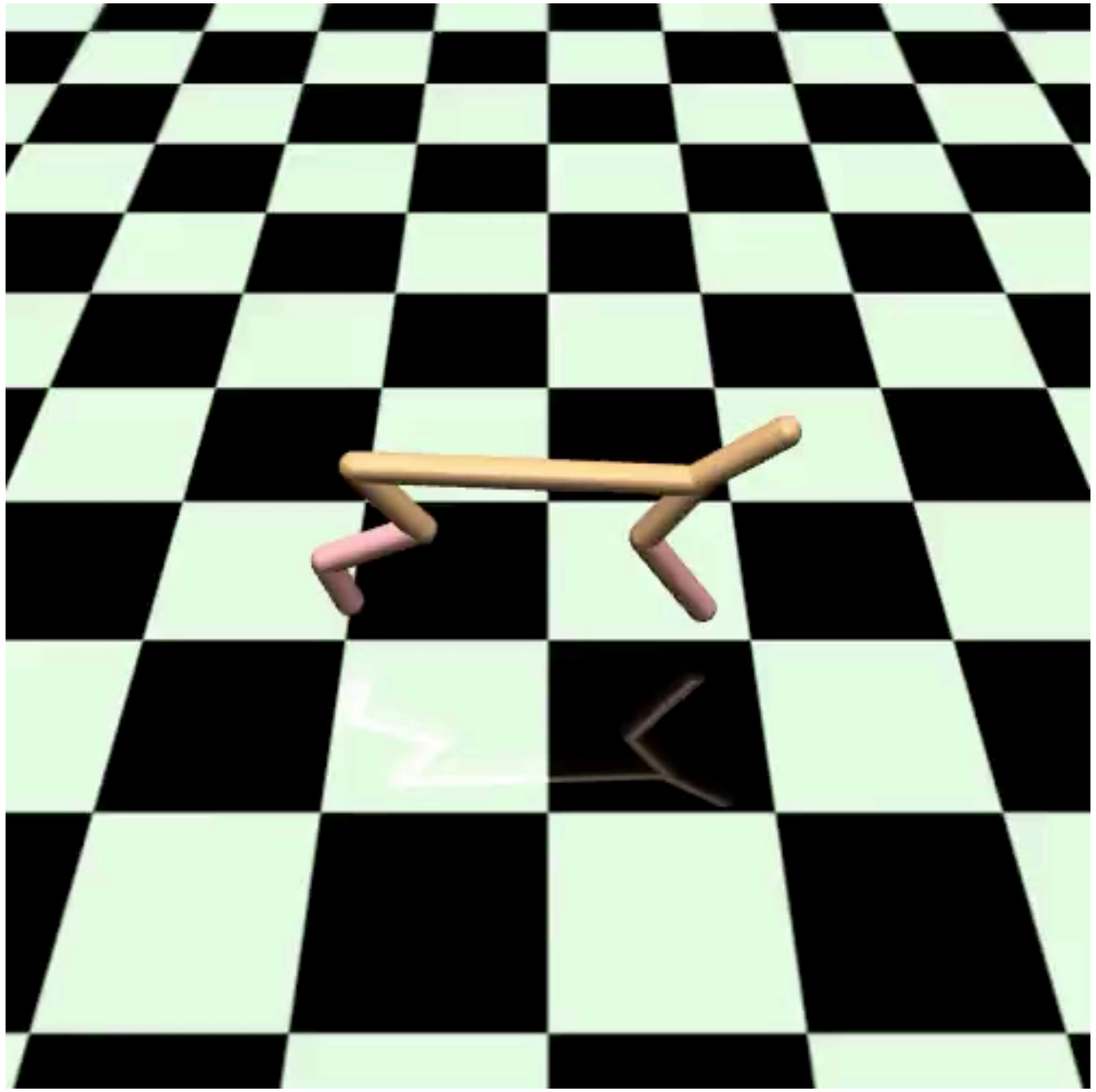


88.97

Results: HalfCheetah

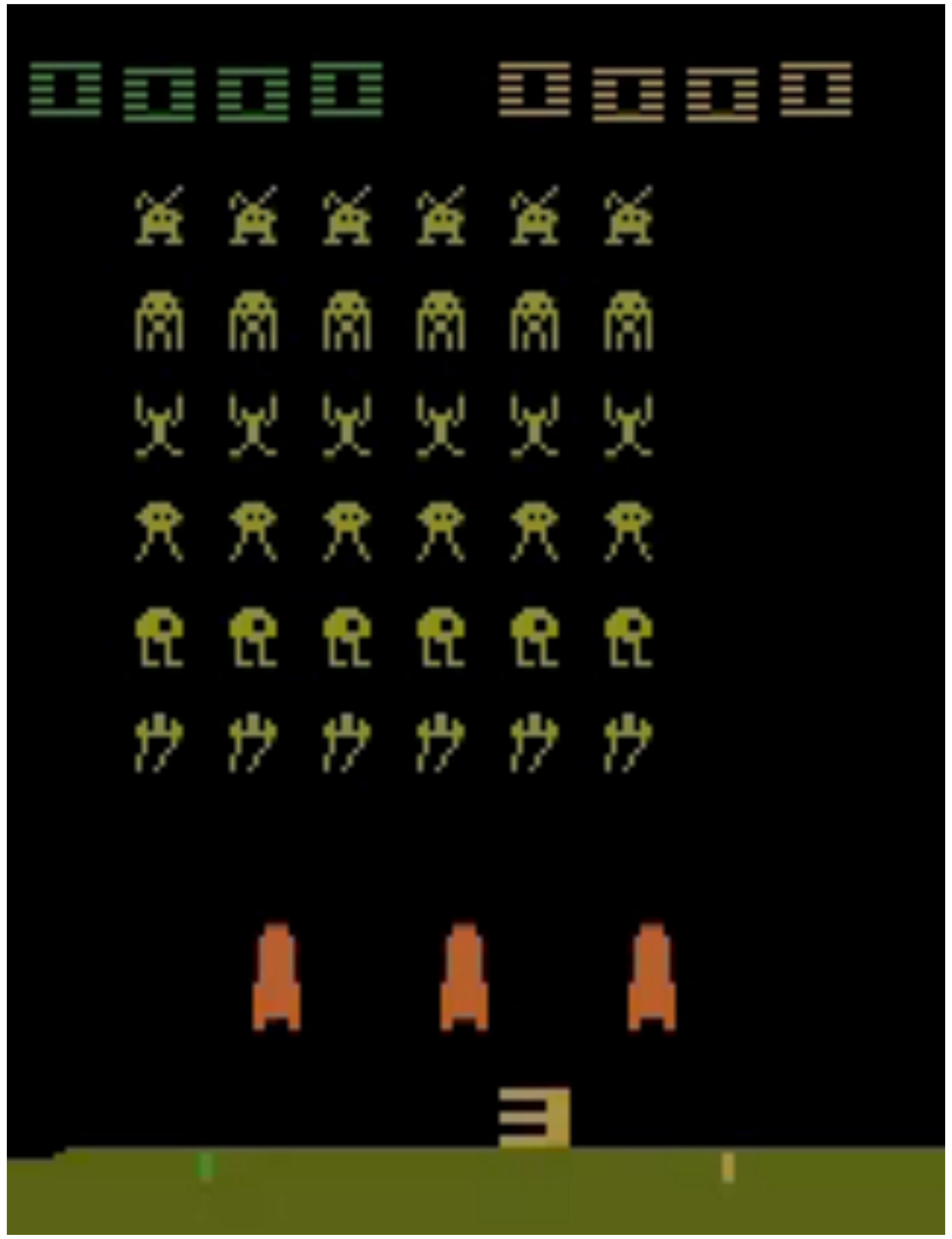


Best demo (88.97)

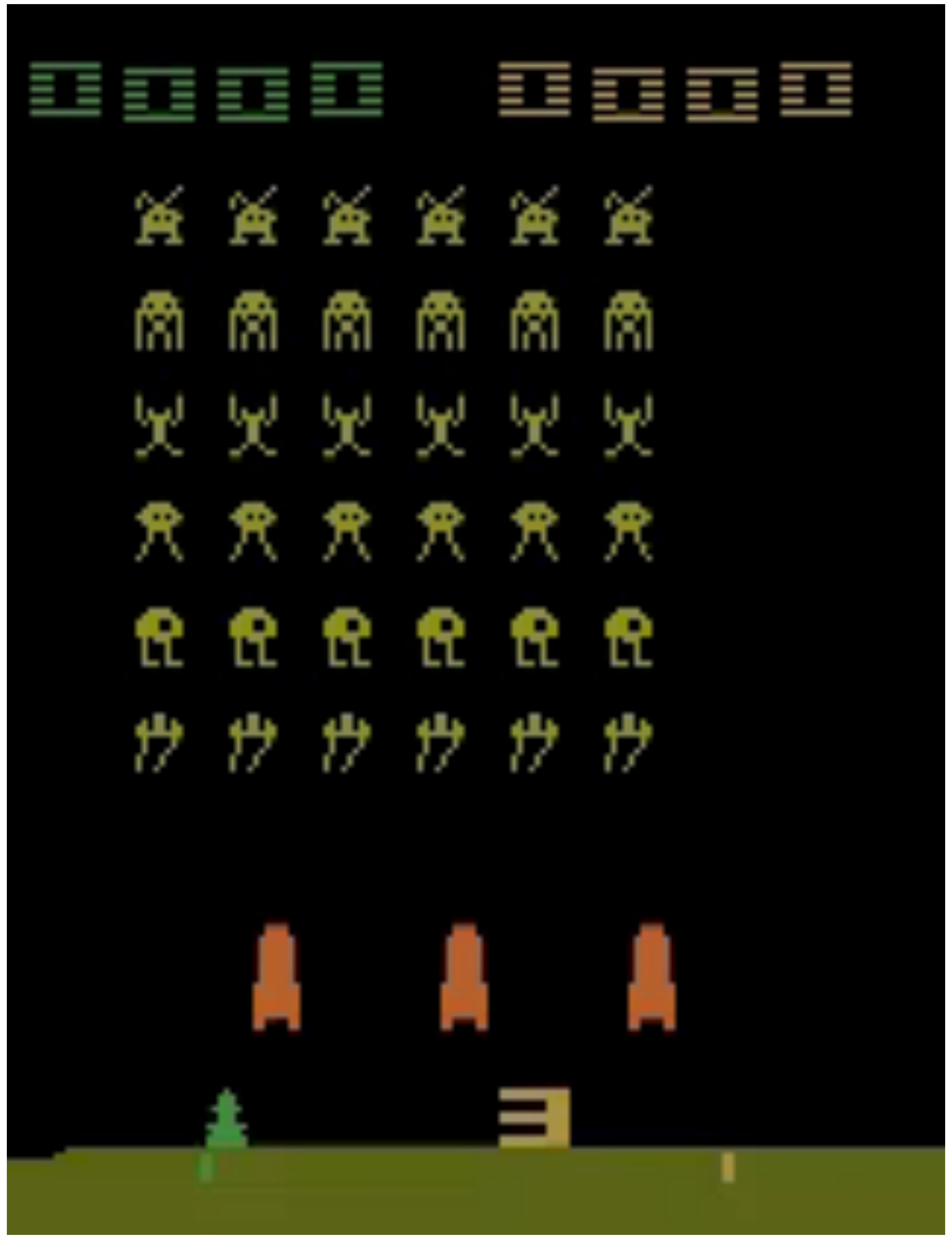


T-REX (143.40)

Results: Atari

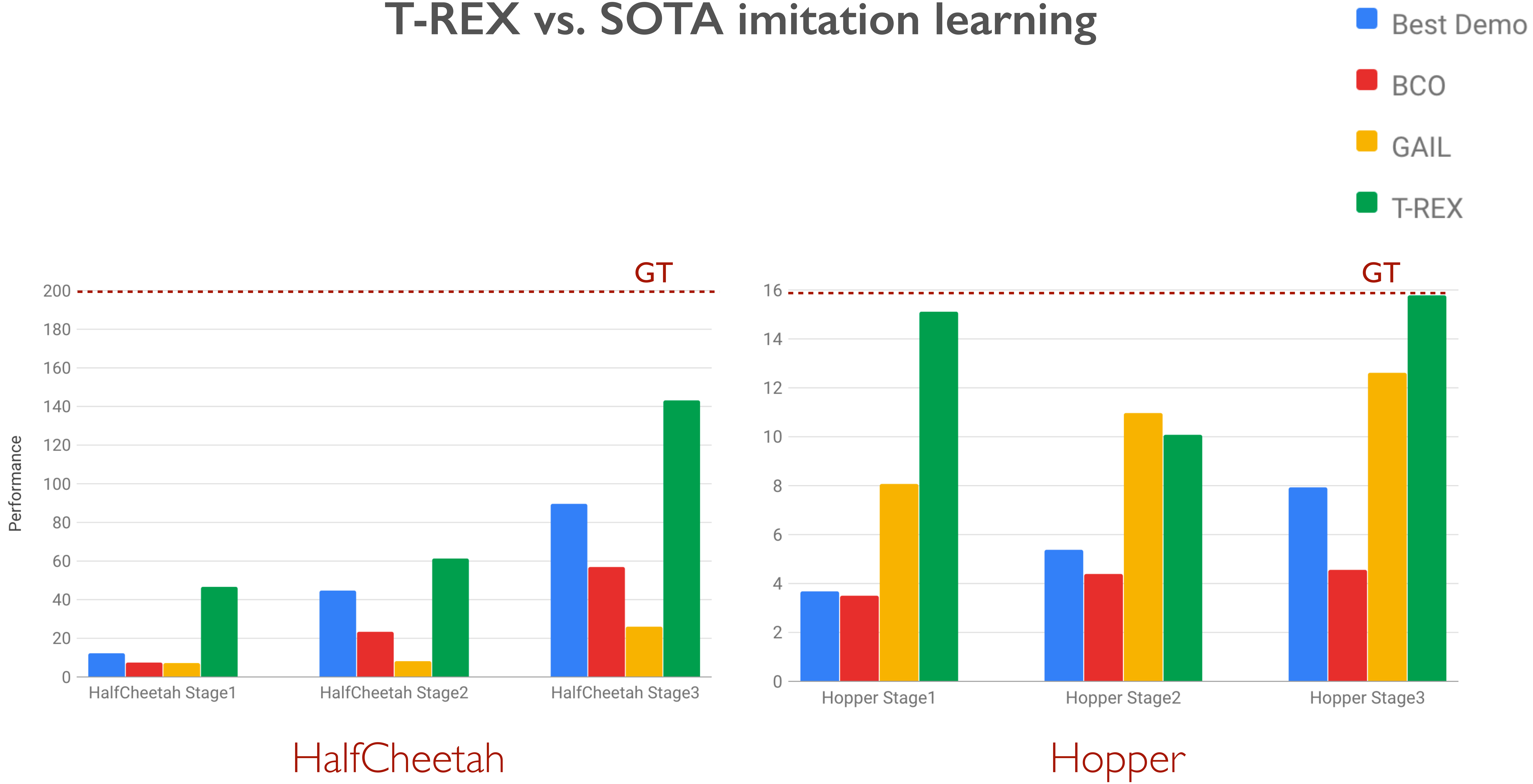


Best demo (600)

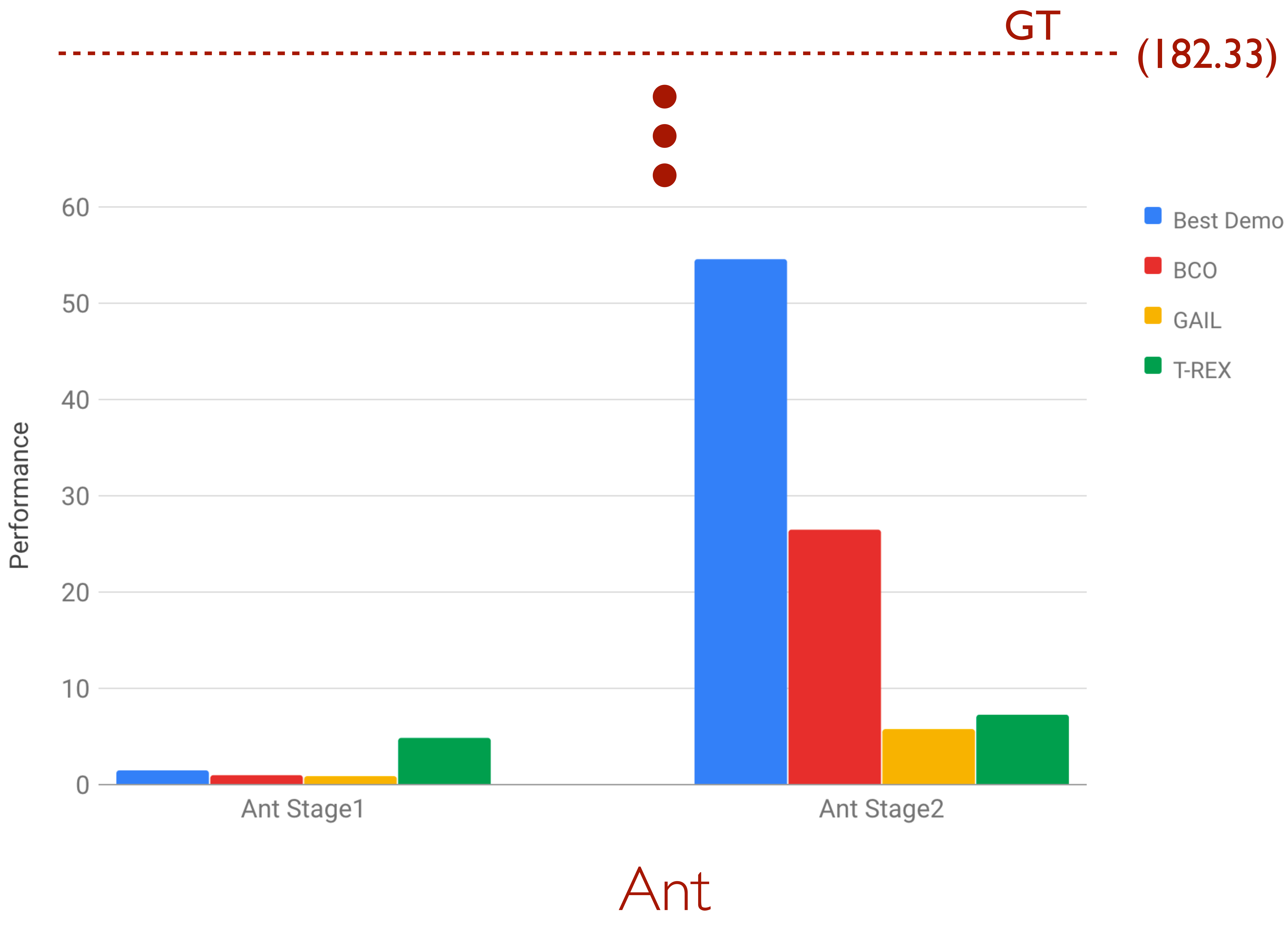


T-REX (1495)

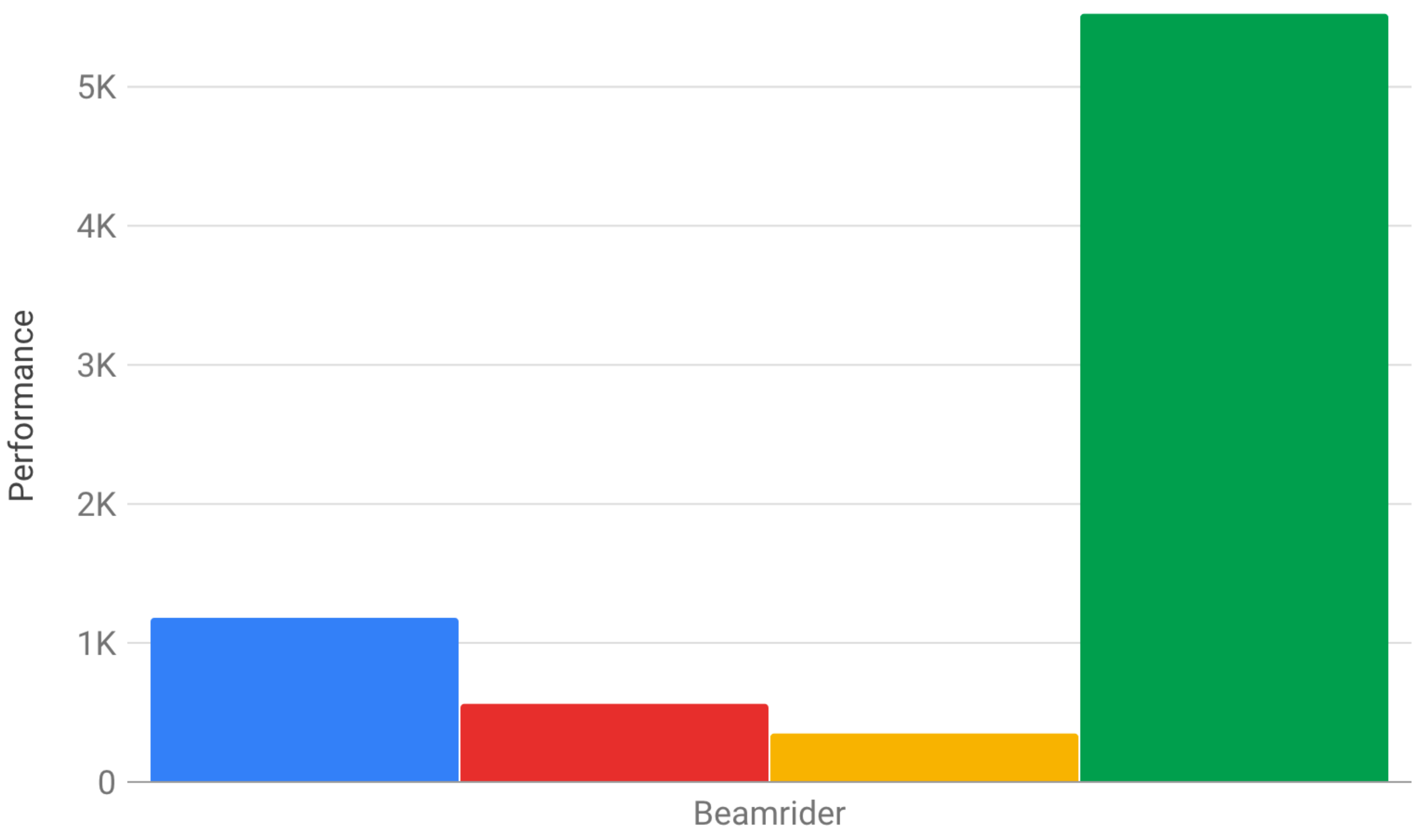
T-REX vs. SOTA imitation learning



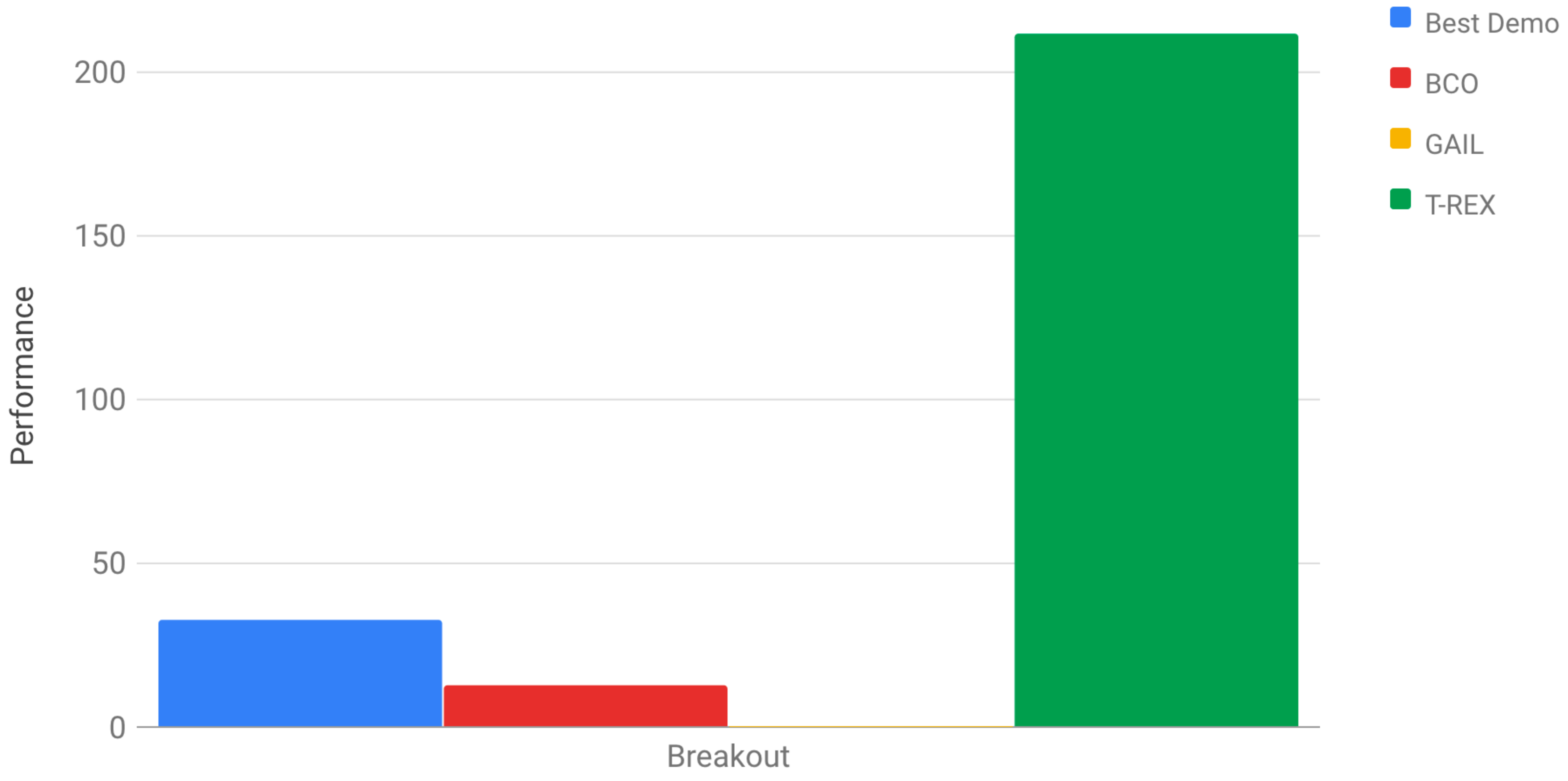
T-REX vs. SOTA imitation learning



T-REX vs. SOTA imitation learning

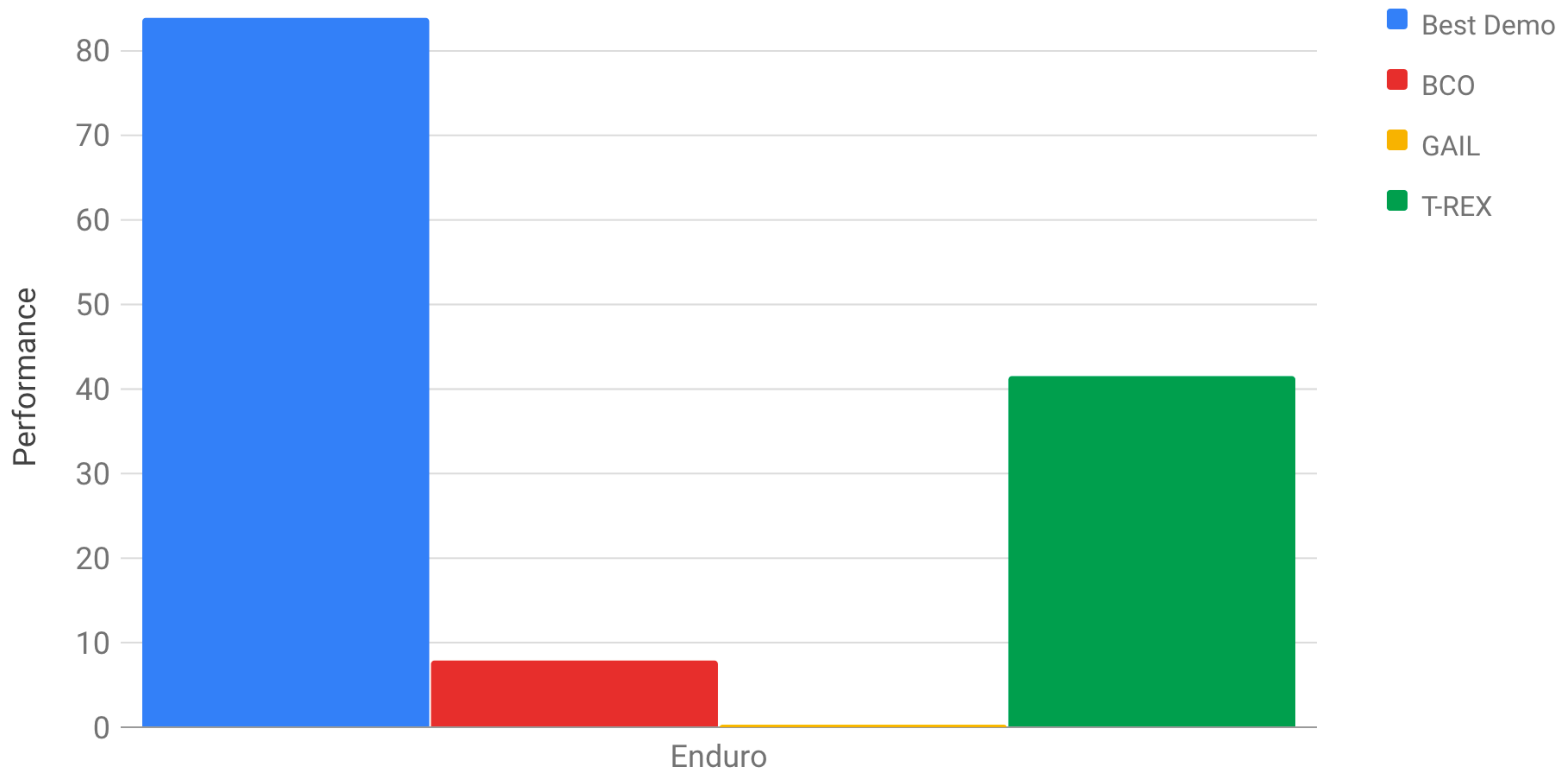


Beamrider



Breakout

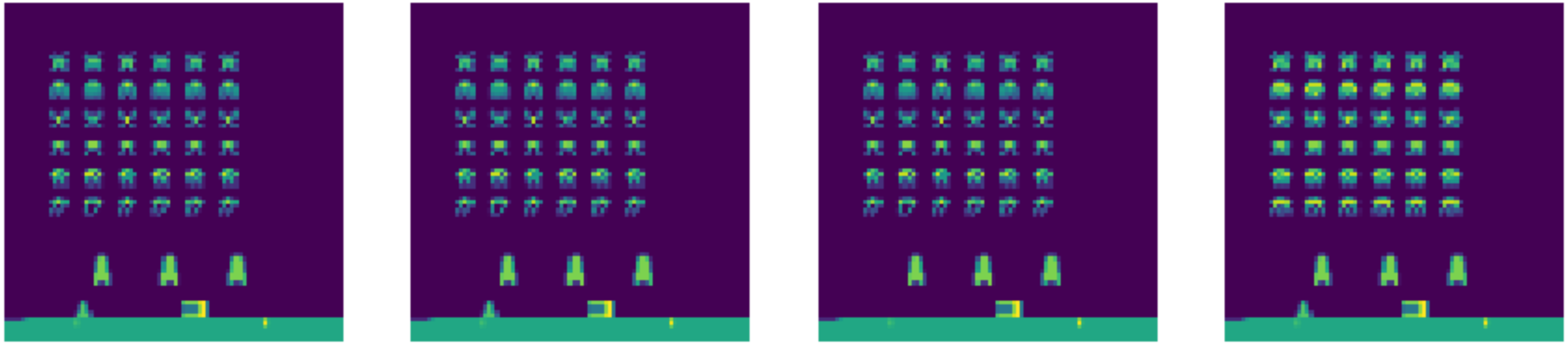
T-REX vs. SOTA imitation learning



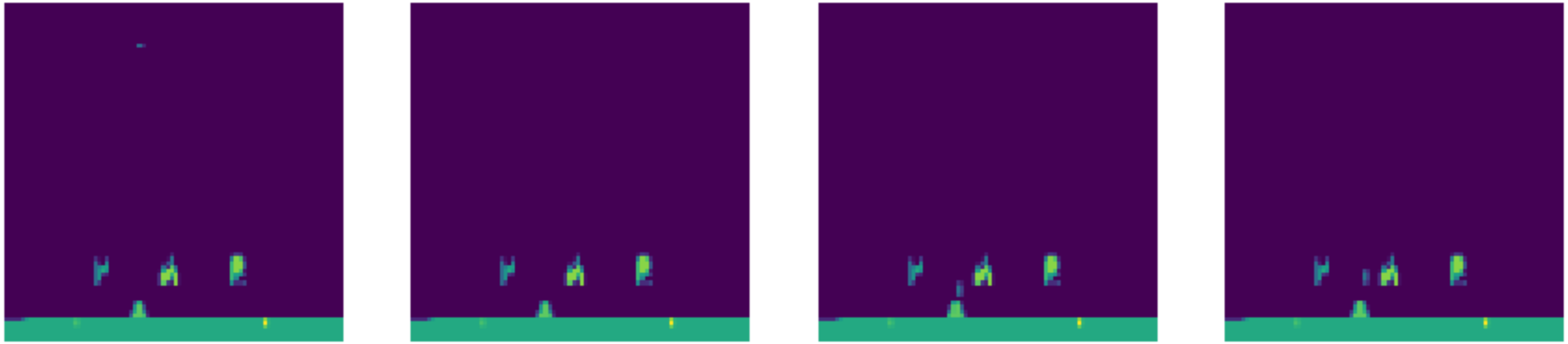
Enduro

Frame stacks: best vs. worst reward

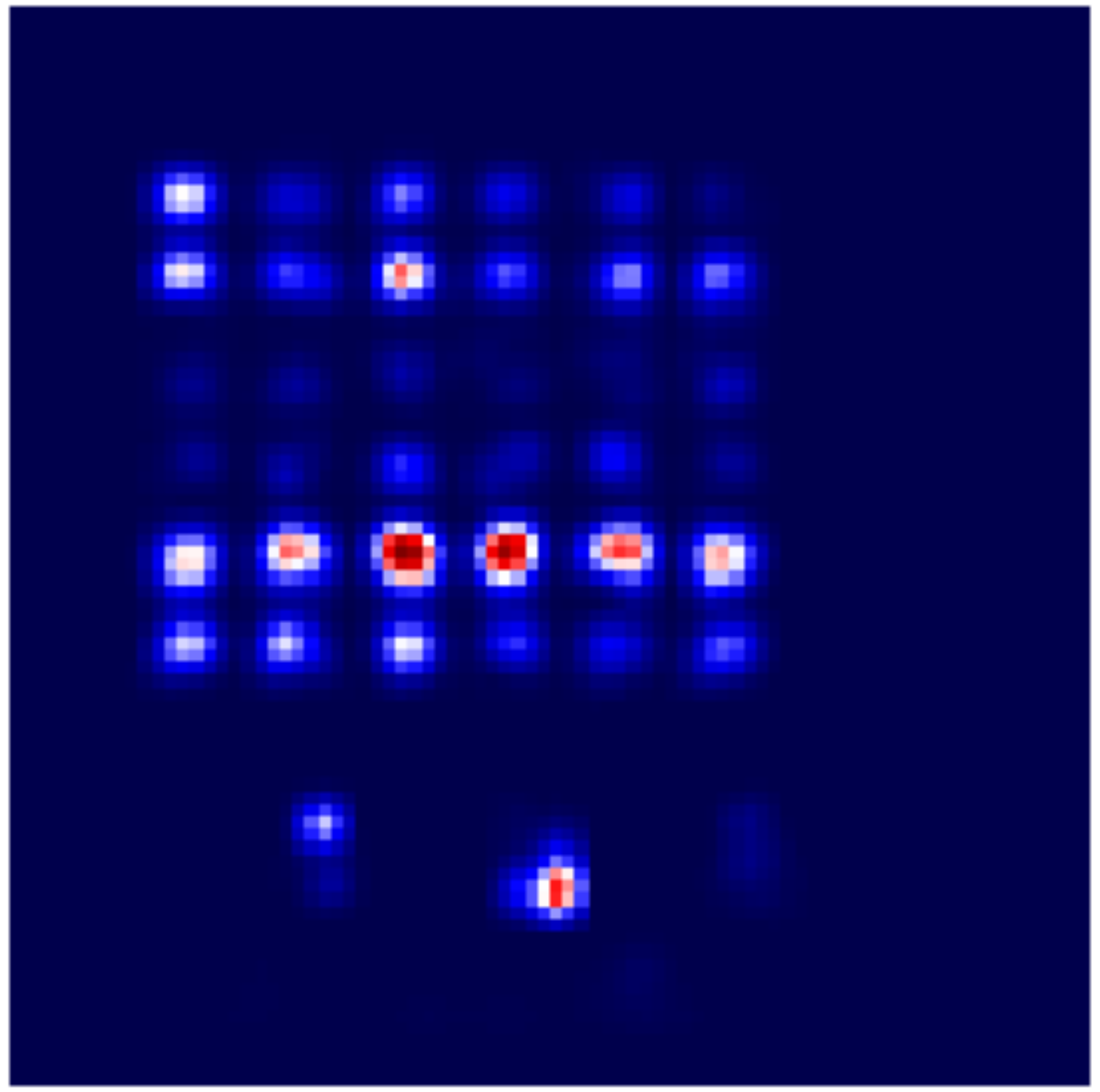
Worst



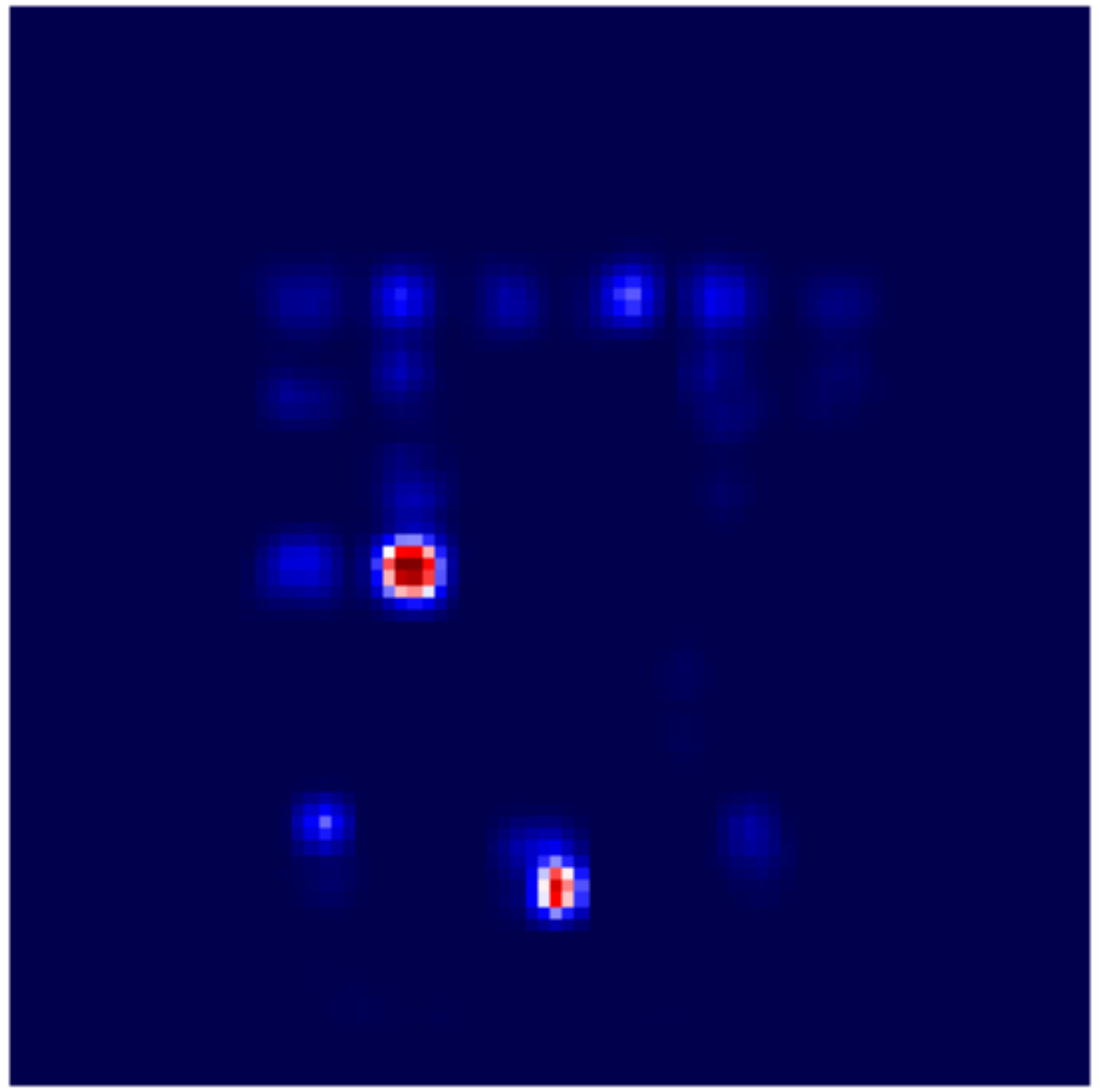
Best



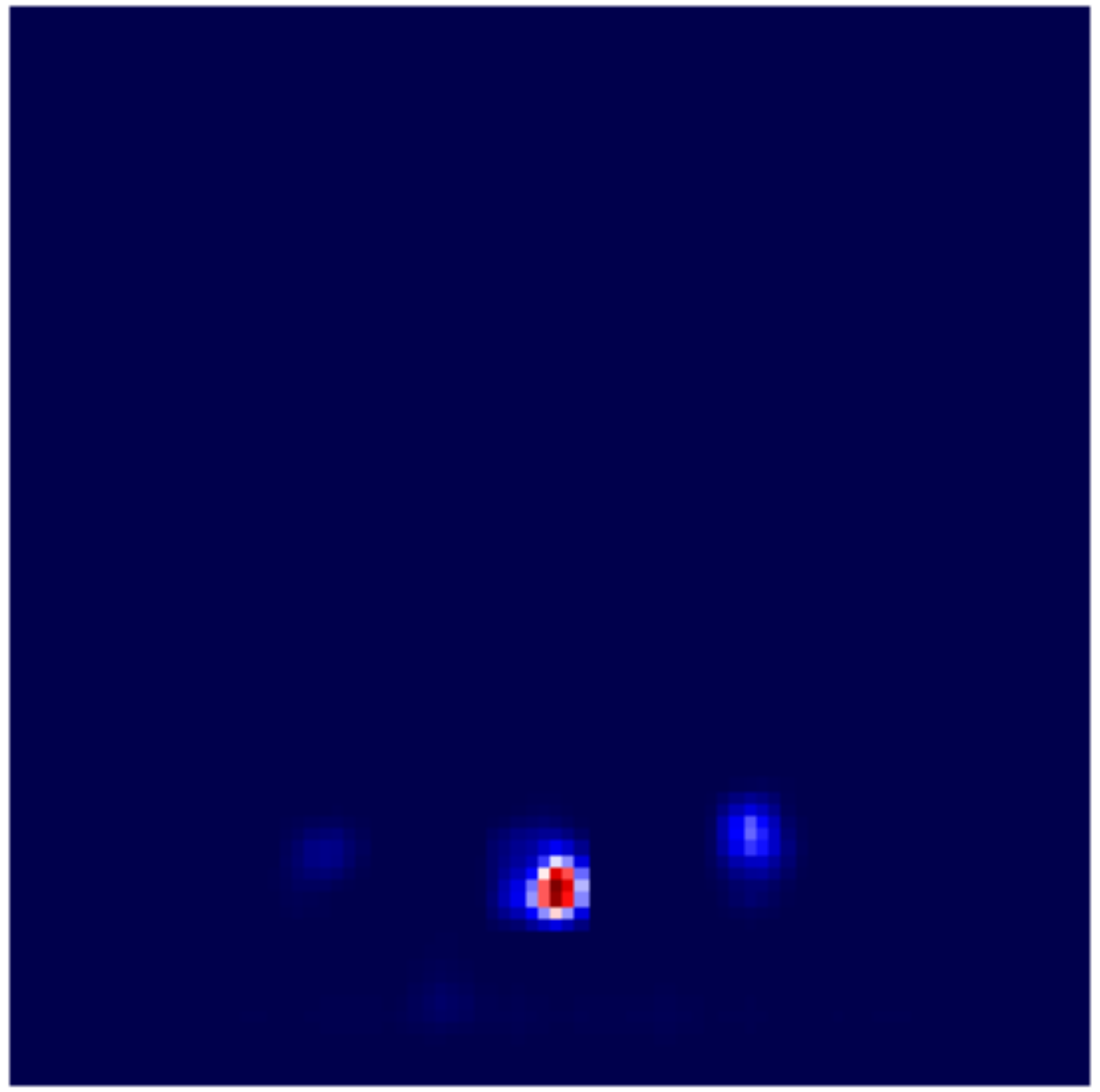
Reward heat maps



Min frame



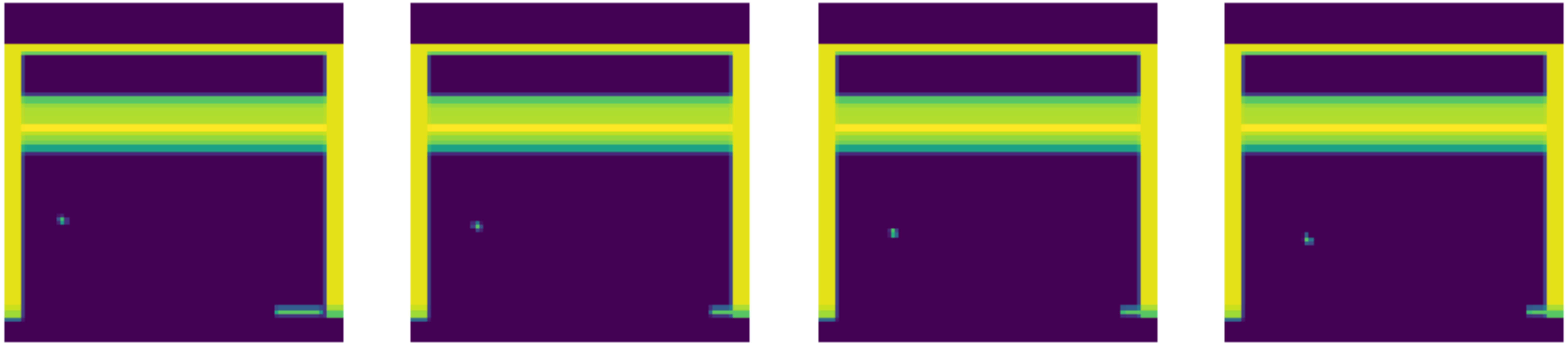
Medium frame



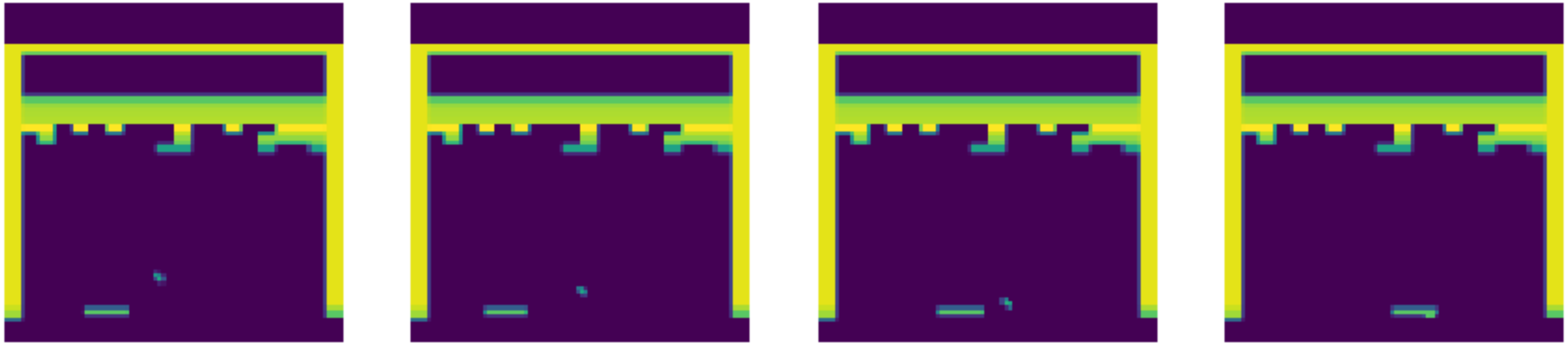
Max frame

Frame stacks: best vs. worst reward

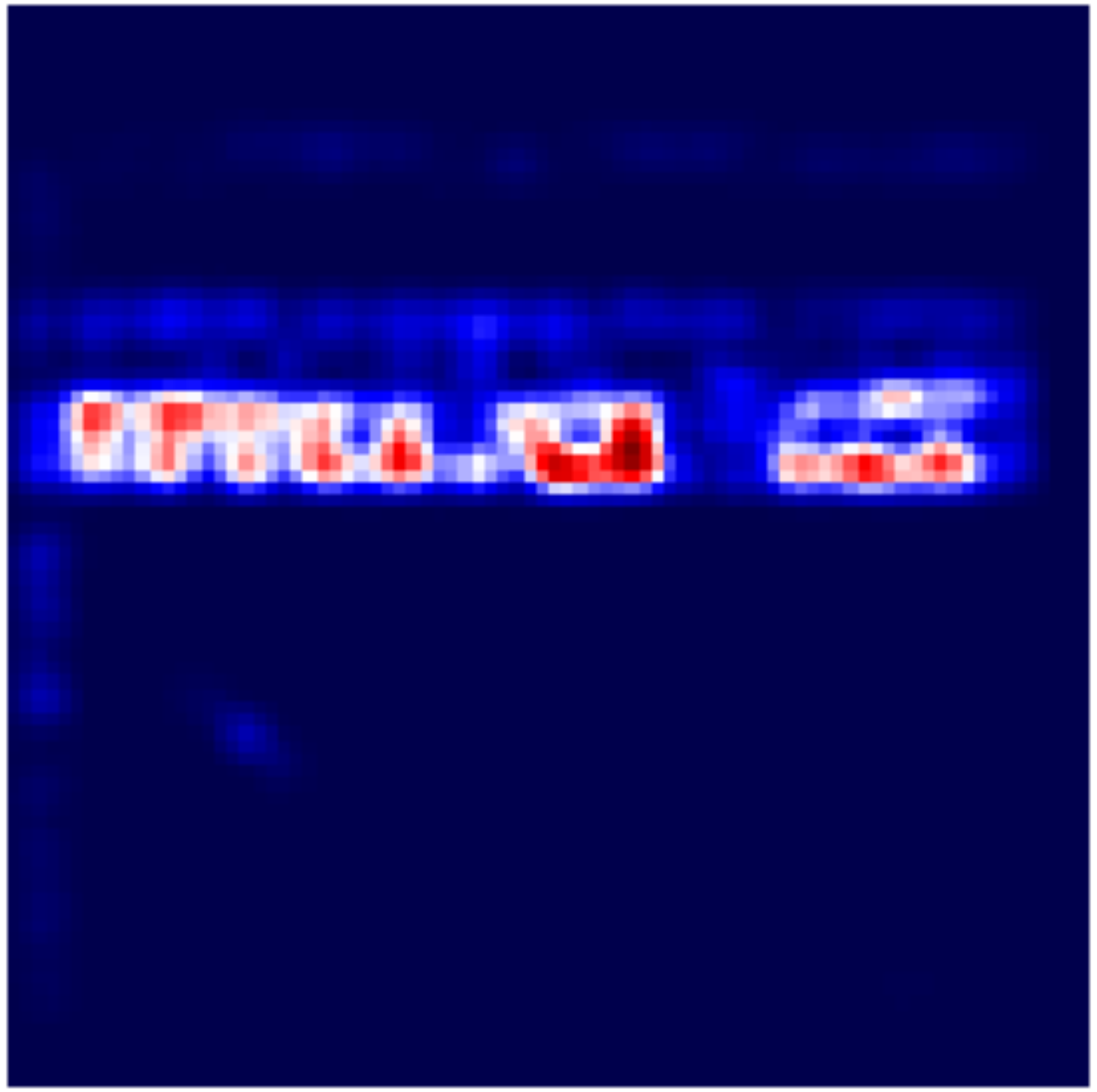
Worst



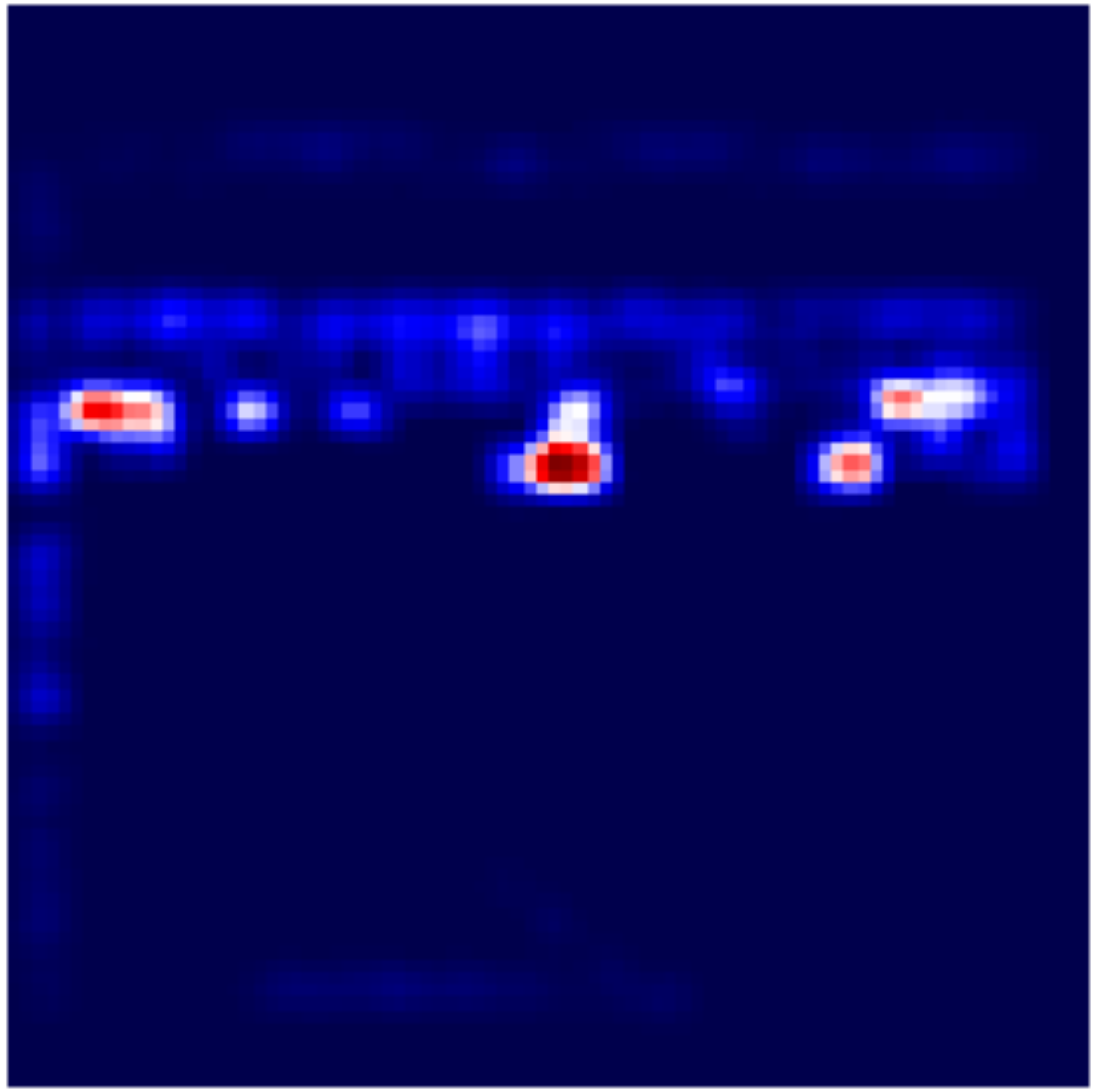
Best



Reward heat maps

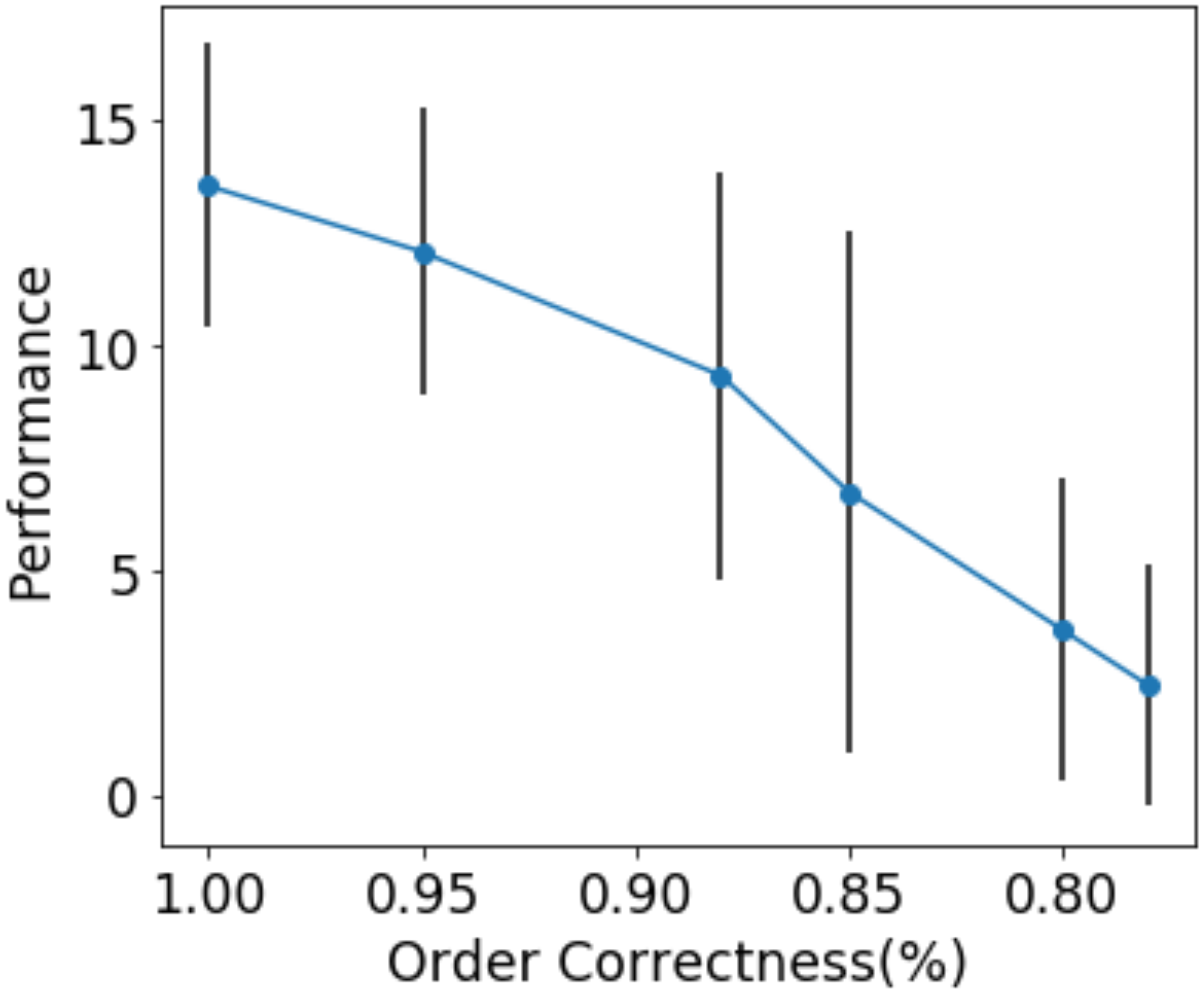


Min frame

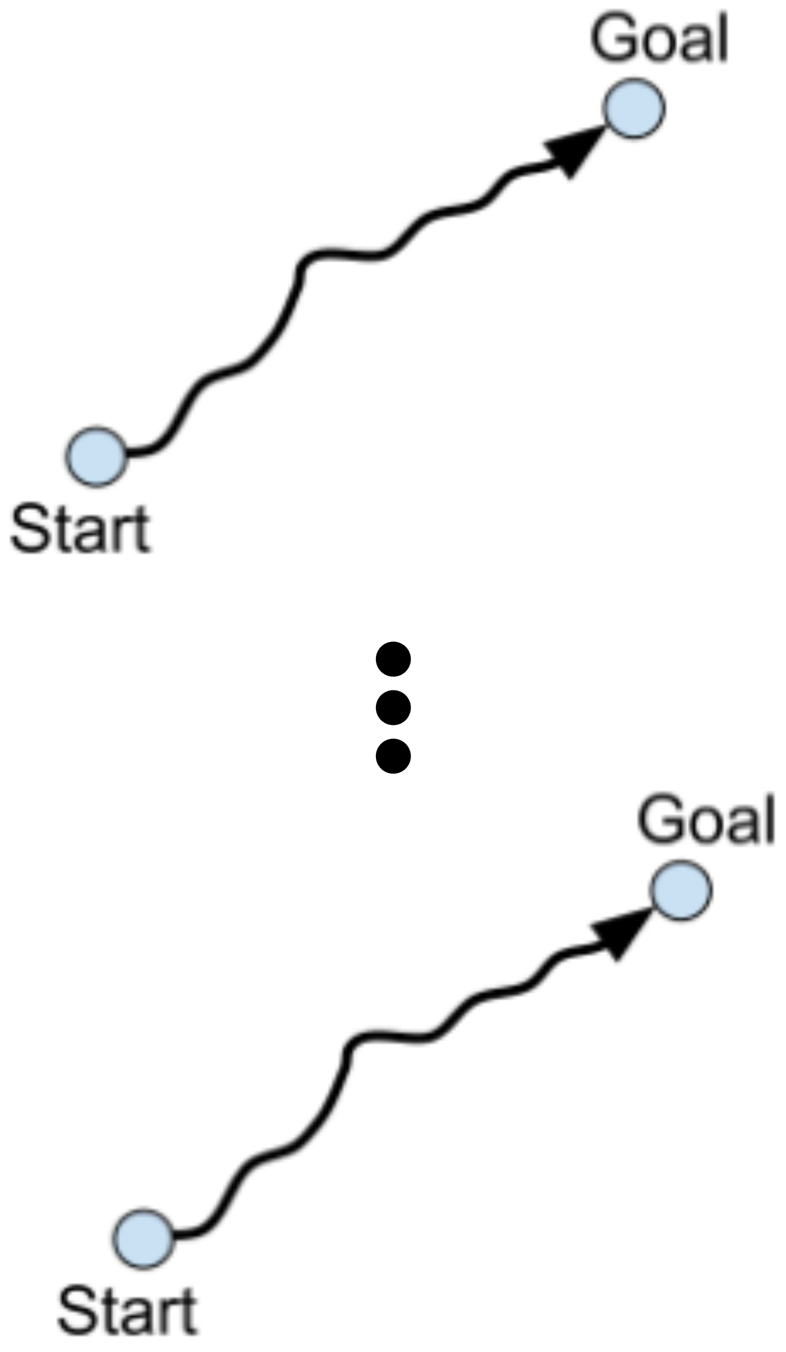


Max frame

Robustness to pairwise ranking noise



D-REX: Auto-generated rankings

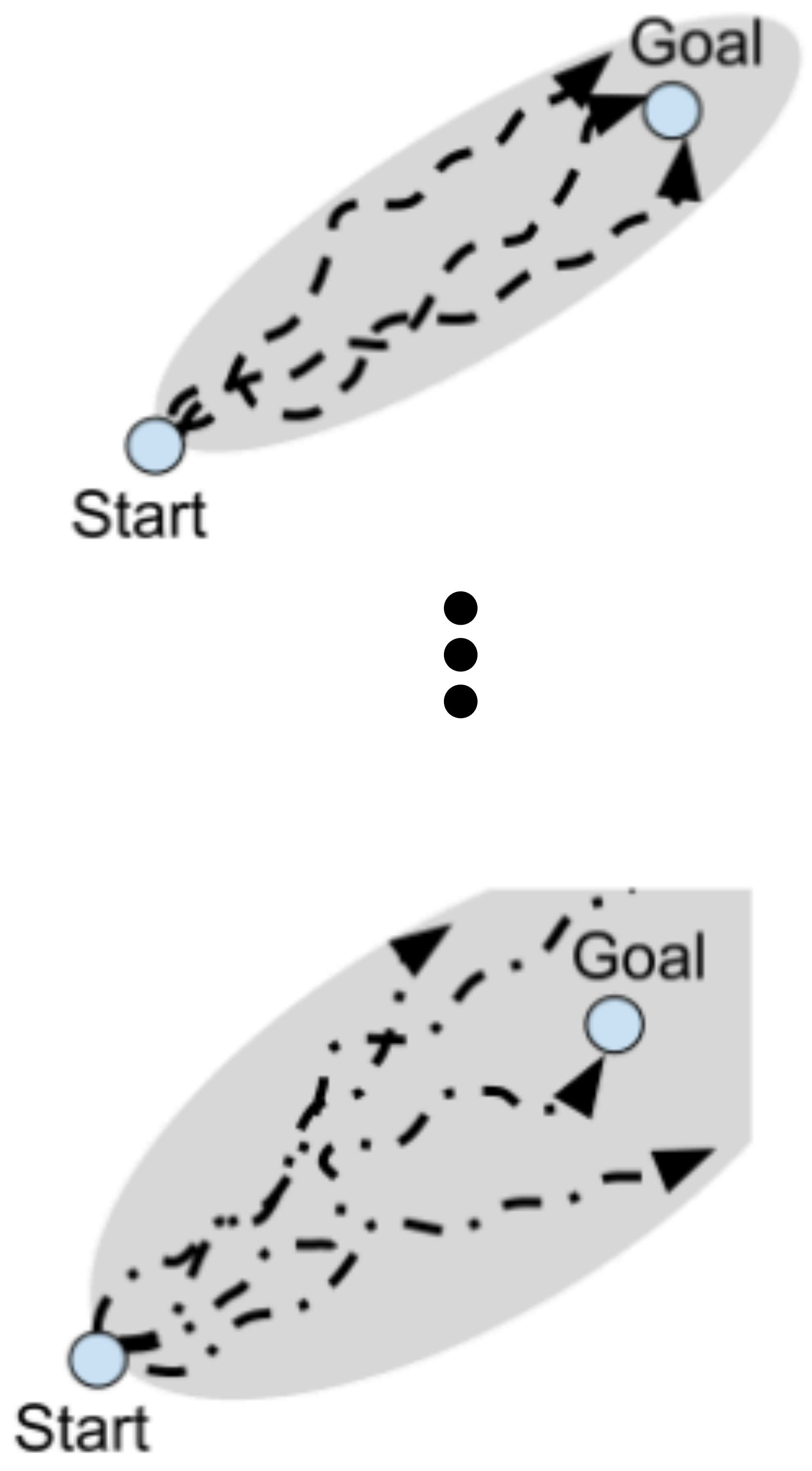


Unranked
Demonstrations

Behavioral
Cloning

$$\pi_{BC}$$

Noise-modified
Rollouts



Increasing
Noise

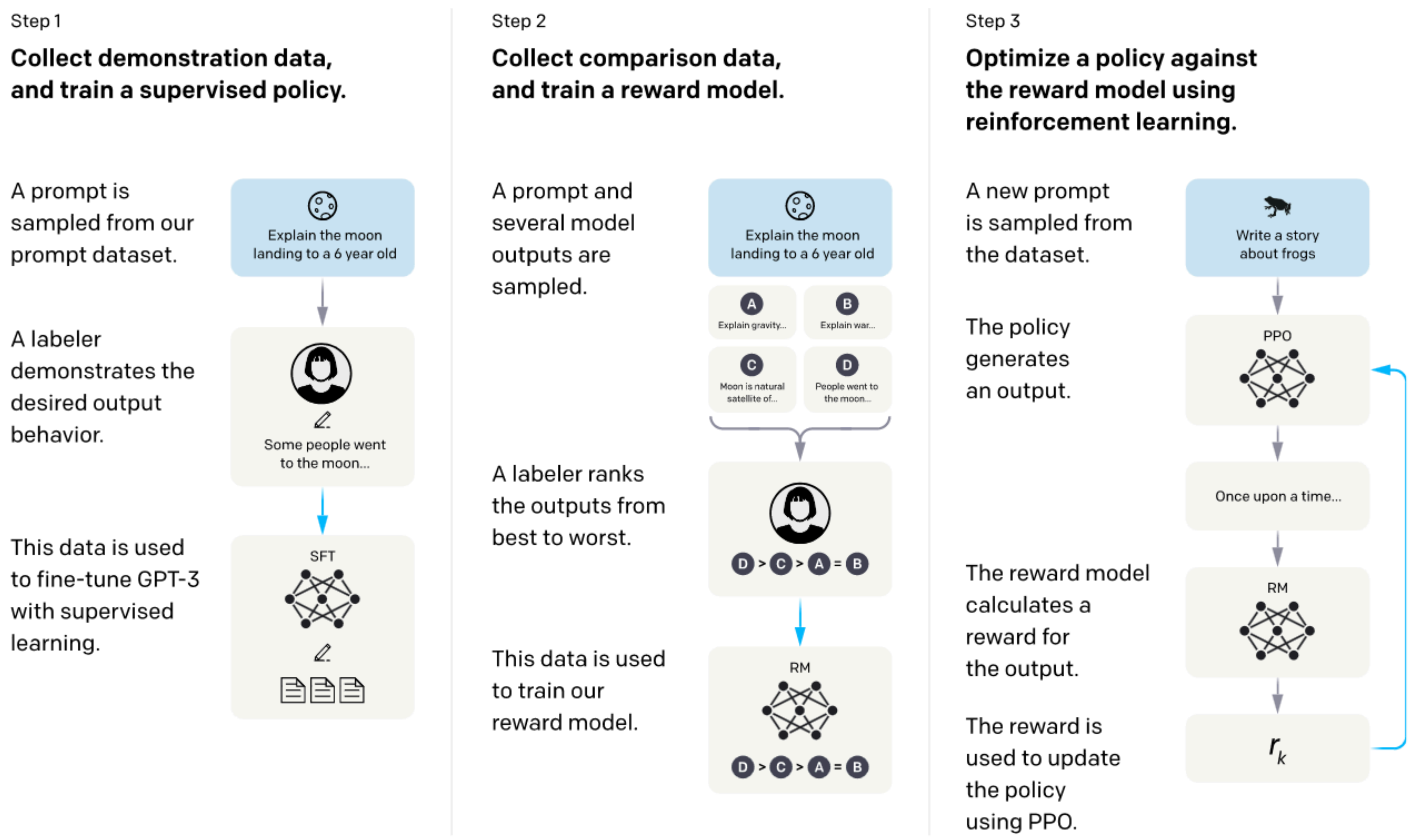
"Ranked"
Trajectories

D. Brown, W. Goo, and S. Niekum.
Ranking-Based Reward Extrapolation without Rankings
Conference on Robot Learning (CoRL), October 2019.

Reinforcement Learning from Human Feedback (RLHF)

- Preference elicitation is a generic mechanism for inferring human goals / desires / priorities
- T-REX operated from an imitation learning perspective, but trajectories can come from anywhere, not just demonstrations
- General recipe for alignment: infer human's reward function and then optimize it with RL
- …or as we'll see later in the course, maybe just learn policies directly from preferences

RLHF for InstructGPT



Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in neural information processing systems* 35 (2022)

Problems with preferences?

How to best align AI with humans?