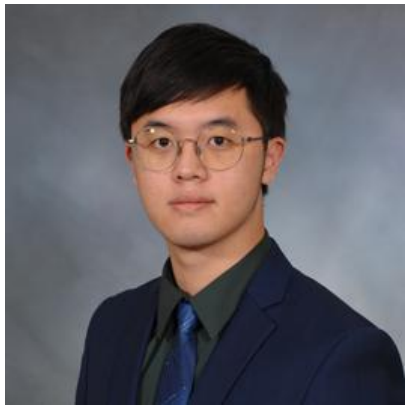
Accelerating Visual-Policy Learning through Parallel Differentiable Simulation

Yale University

NeurIPS 2025 Spotlight

Presenter: Haoxiang You

Haoxiang You



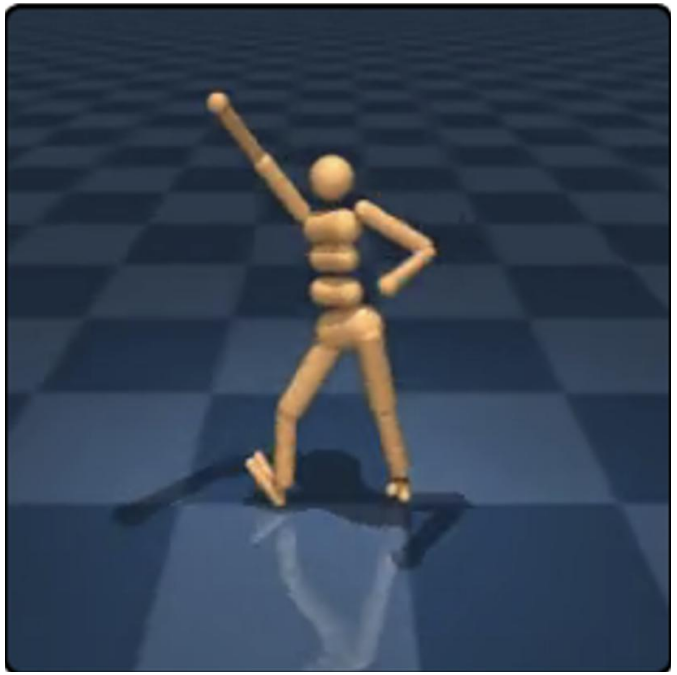
Yilang Liu

Ian Abraham

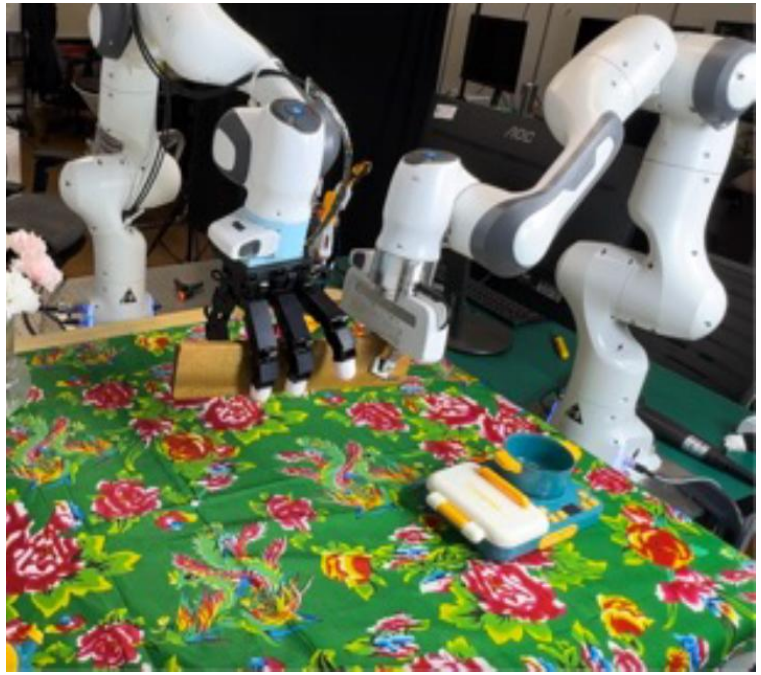
Website: https://haoxiangyou.github.io/Dva_website/

Visual Policy Learning

- Policy learning from **raw pixels** enable control in complex environments
- Visual policy learning method suffer from **slow convergence** and **long training time**



Yarats et. al.2021



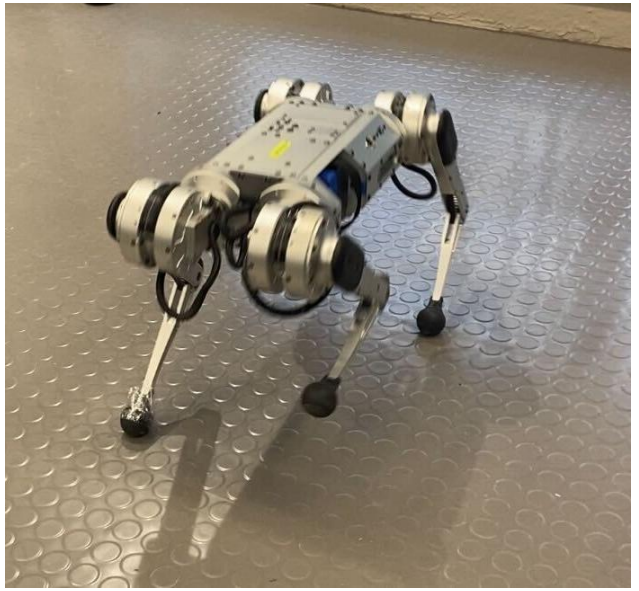
Yuan et. al.2024

Hafner et. al.2025

Differentiable Simulation

- Differentiable simulation enables **first-order** policy gradient estimation, improving training efficiency
- However, current approaches operate on **low-dimensional** state space

Xu et. al.2022

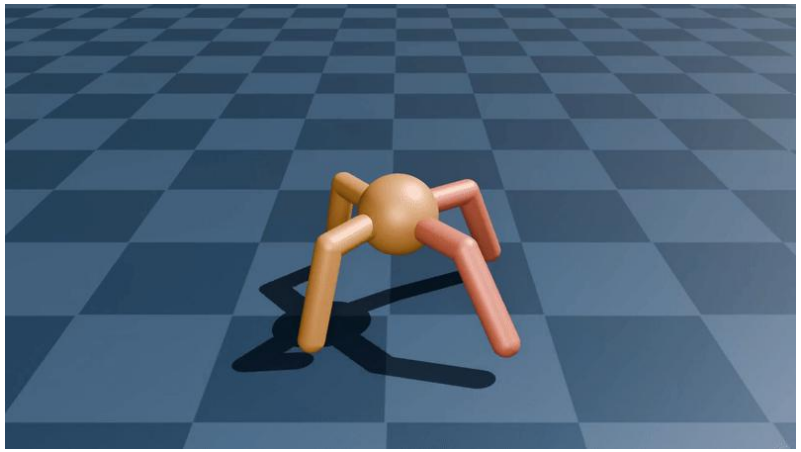


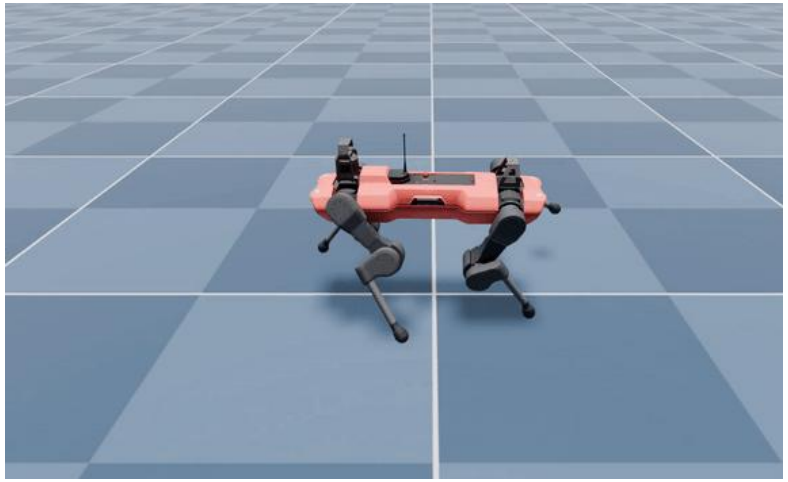
Song et. al.2024

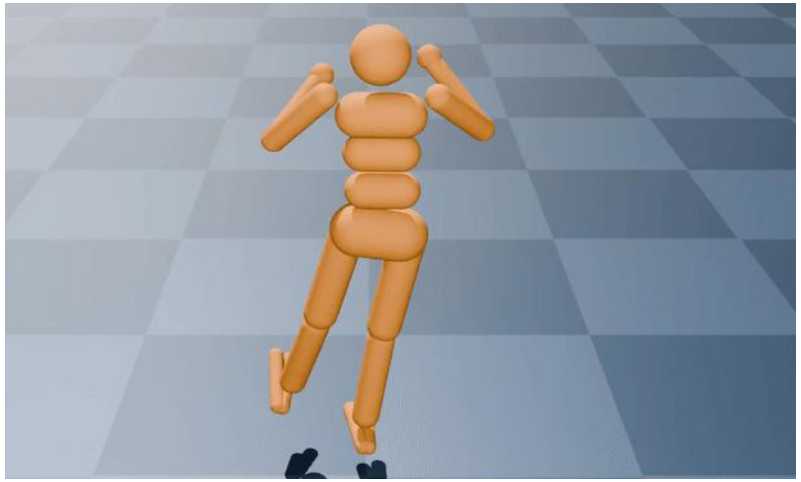
Xing et. al.2025

Our Contribution

D.Va: A **first-order visual RL** method that learns policies directly from pixels, allowing visual-policy learning in **minutes-to-hours** on a personal laptop

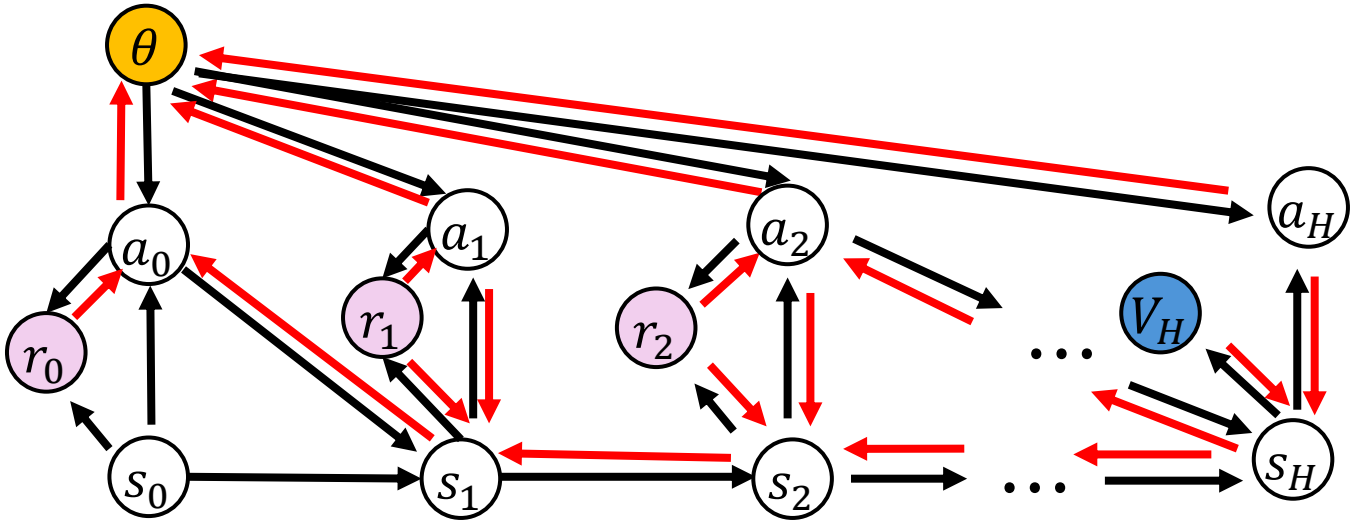






Background on First-order RL Method via Differentiable Simulation

- Construct a computation graph by rolling out trajectories
- Estimate policy gradients by backpropagating through the full trajectories
- Often truncate horizons to avoid gradient explosion and rely on value functions to predict future rewards

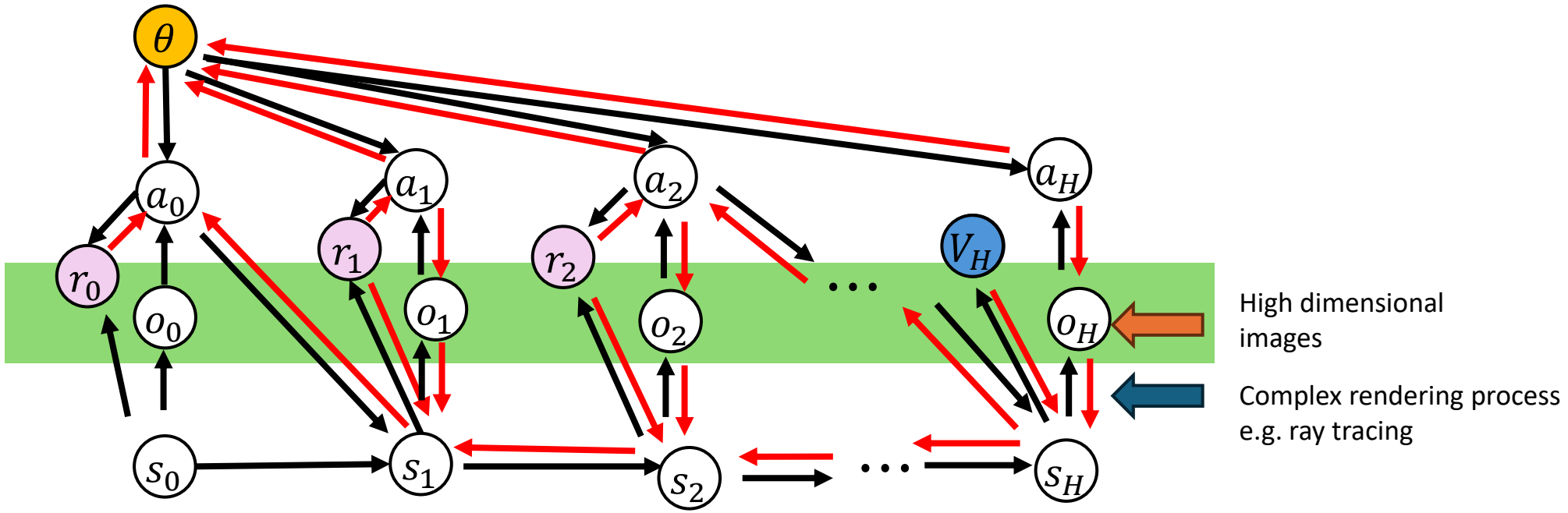


Challenges in Applying First-Order RL to Visual Policy Learning

- Jacobian involving high-dimensional images
  - Require **large memories**
  - Multiply **big matrix is slow**

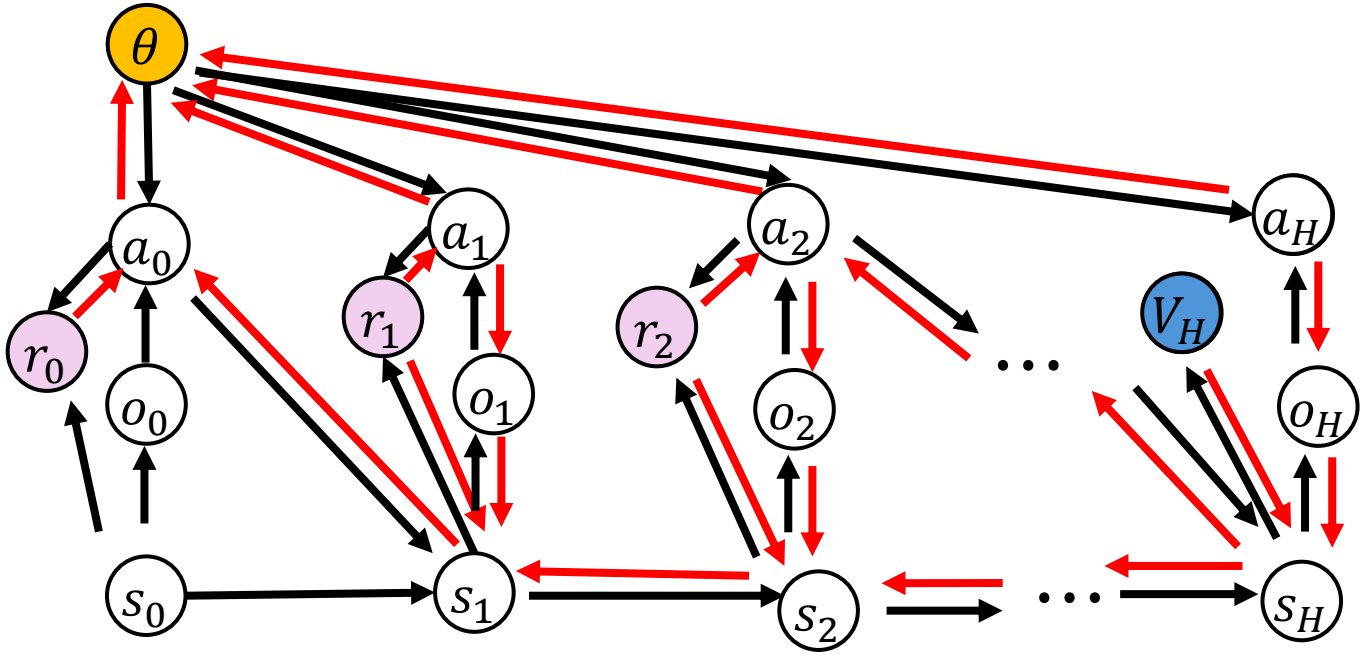
- Most simulation does not support diff-rendering

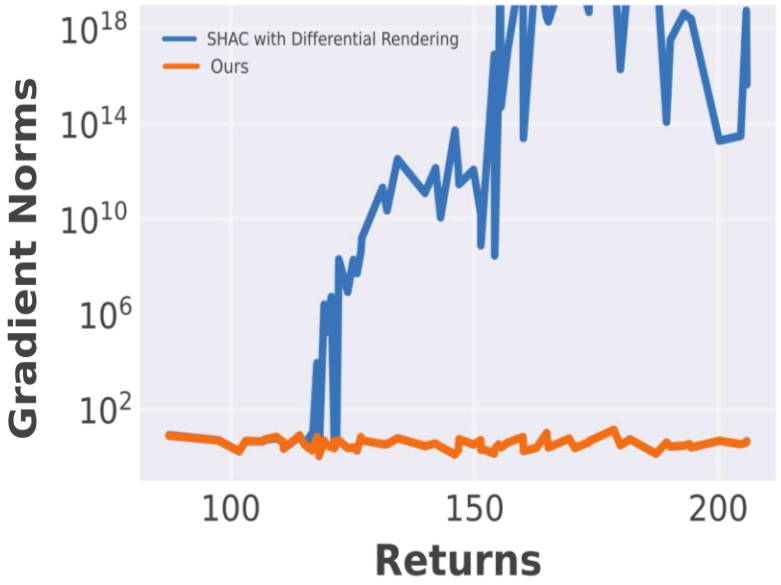
- **Gradient explosion**



Our Method

Our key is to stop gradients through high-dimensional visual inputs, yielding a quasi–policy-gradient estimator that remains **stable** and **informative** to train visual policy

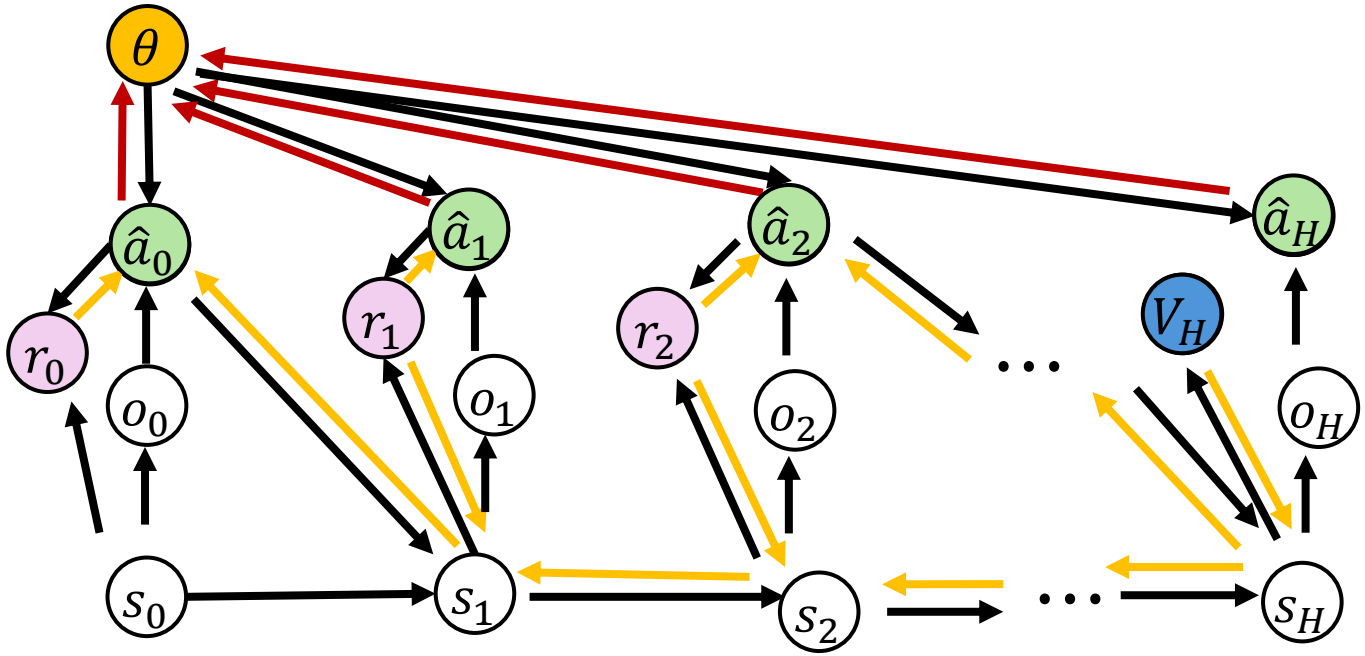




Our Method

We also provide a **conceptual derivation** to deepen understanding of our quasi-policy gradient

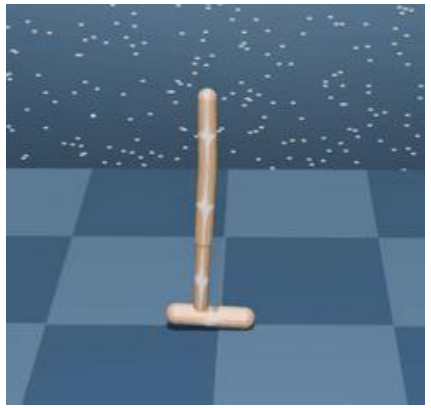
By chain rule, our quasi-policy gradient can be re-factored into two parts: trajectory optimization and policy distillation.



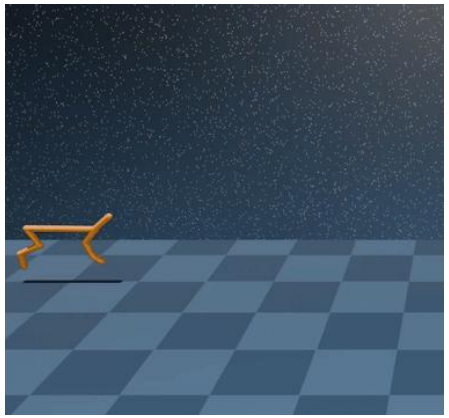
Training Demo

Leveraging differentiable simulation, our method learns a reasonable visual-loco policy in only **5 minutes**, and continues improving with more training.

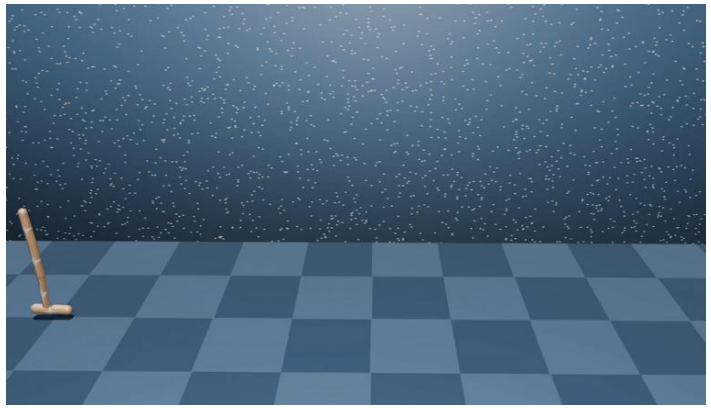
Initial Policy

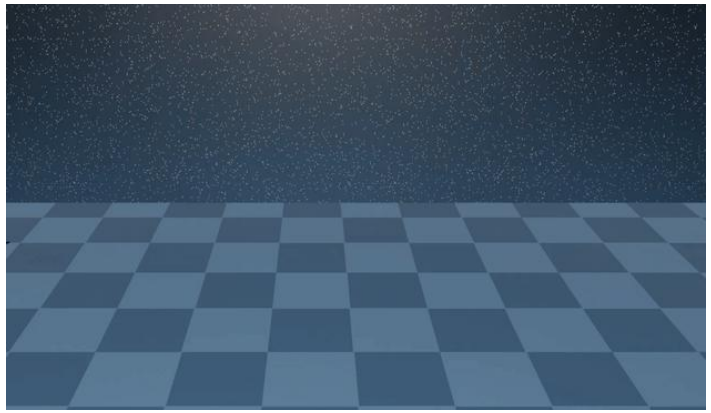


5 mins training



2 hours training

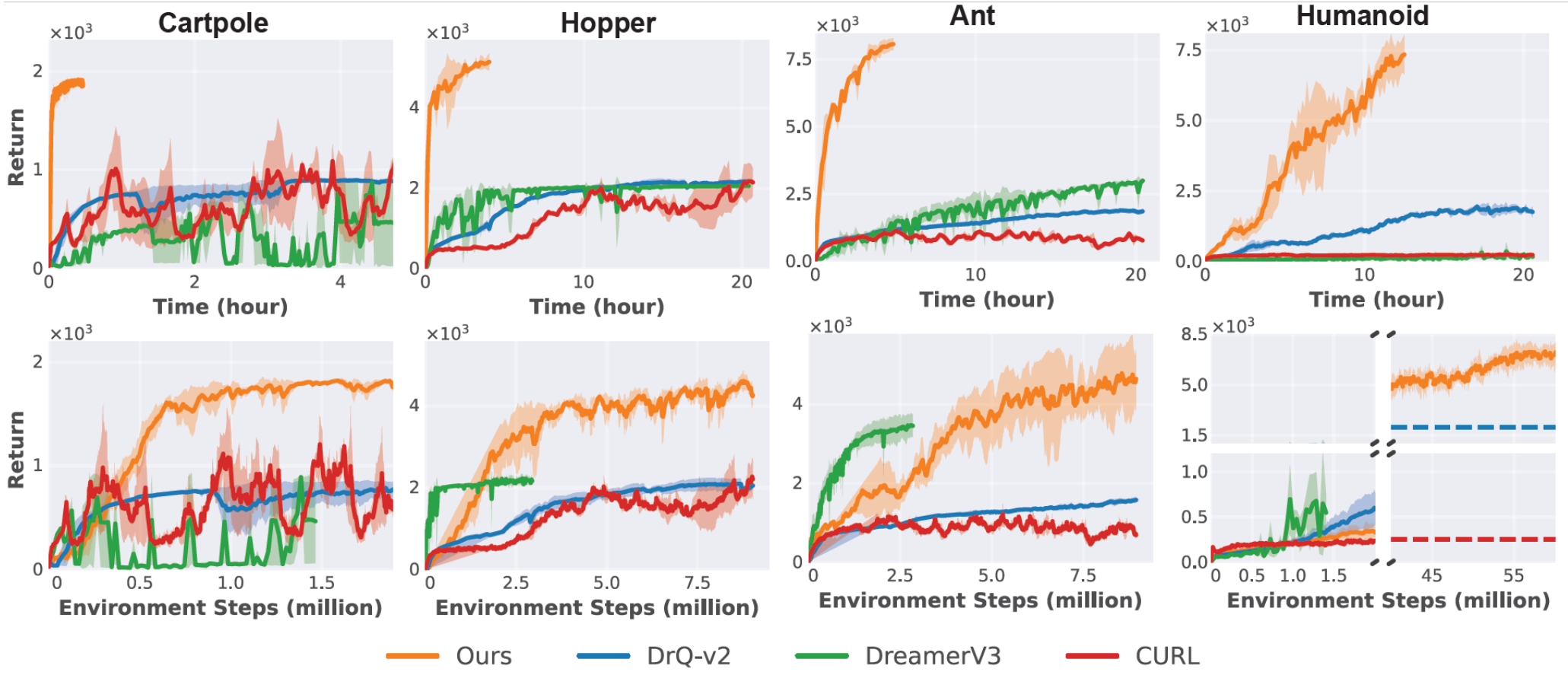




Results

We compare our method against a wide range of strong visual policy learning baselines

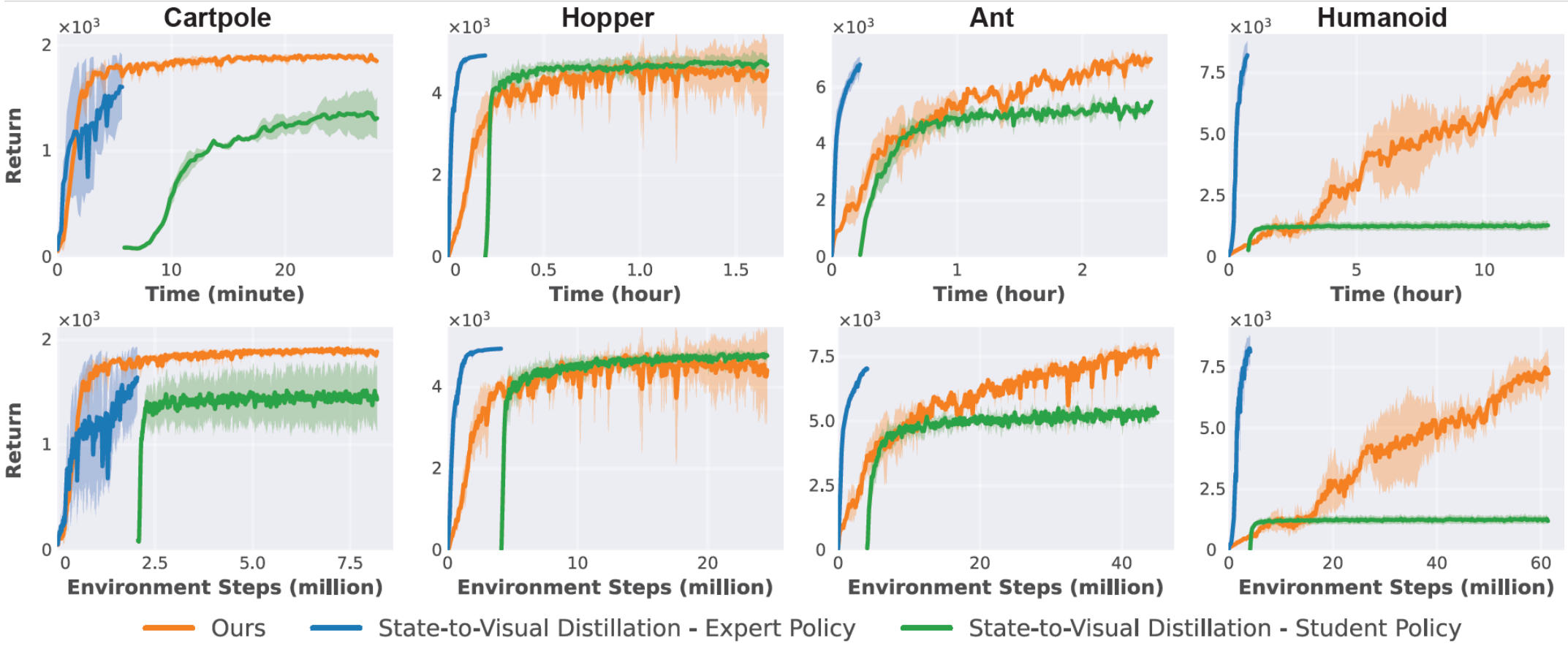
Results show that our method is not only **fast**, but also achieves **higher** rewards than all existing baselines



Results

We compare our method against a wide range of strong visual policy learning baselines

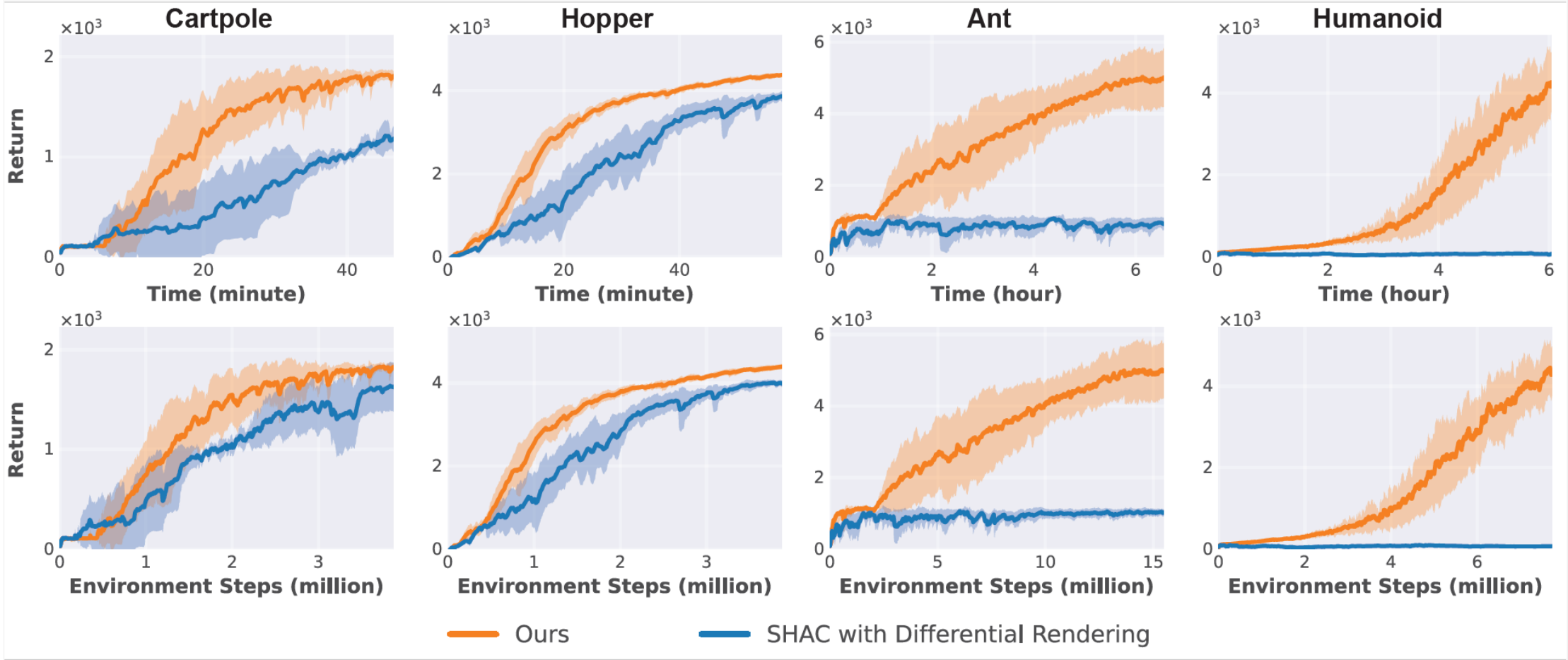
Results show that our method is not only **fast**, but also achieves **higher** rewards than all existing baselines



Results

We compare our method against a wide range of strong visual policy learning baselines

Results show that our method is not only **fast**, but also achieves **higher** rewards than all existing baselines




DreamerV3
Drqv2
Distillation
Ours
Thank you!