2025 IEEE-RAS 24th International Conference on Humanoid Robots
September 30 – October 2, 2025 | COEX, Seoul, Korea

Action Chunking Proximal Policy Optimization for Universal Dexterous Grasping

Sanghyun Hahn and Jonghyun Choi

Introduction

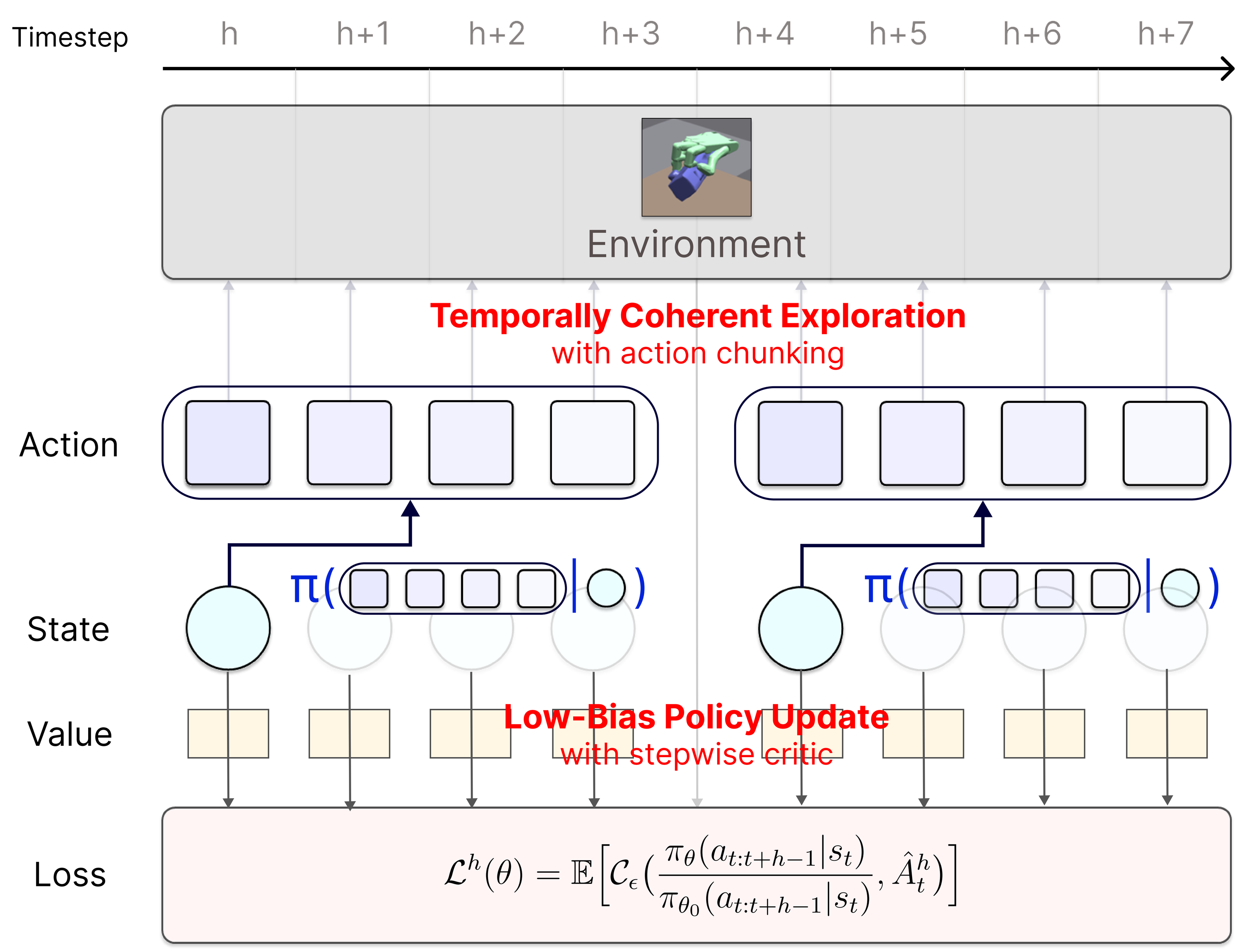
- Universal dexterous grasping is challenging due to the high DoF of the dexterous hand and diverse geometry of target objects.
- While Reinforcement Learning is widely used in this task, the high DoF makes exploration challenging when employing RL.
- **Can we explore a more relevant region of the state-space efficiently?**
  → Reinforcement Learning with Action Chunking!

RL with Action Chunking

- Define a chunked actor $\pi(a_{t:t+h-1}|s_t)$
  → Predicts the next h step of actions from one state
- In standard RL, actions cancel each other out in early exploration.
- In action chunked RL (ACRL), **temporally coherent actions** from the chunked actor provide a wider exploration.

Action Chunking Proximal Policy Optimization

- Prior ACRL methods use $Q(s_t, a_{t:t+h-1})$ as the critic.
- While manageable for small-DoF environments (e.g. 7 DoF robot arm), this dimension becomes intractable for dexterous hands.
  → Instead, use $V(s_t)$ with importance sampling!
- Build everything on top of PPO, the most widely used on-policy RL algorithm based on $V(s_t)$.



- **Action Chunking PPO** directly integrates action chunking into the actor and importance sampling ratio, while avoiding usage of intractable Q-functions.
- ACPPO generates temporally coherent actions, leading to a broader state space exploration which improves the critic V(s).
- ACPPO reduces the bias of Generalized Advantage Estimates (GAE).

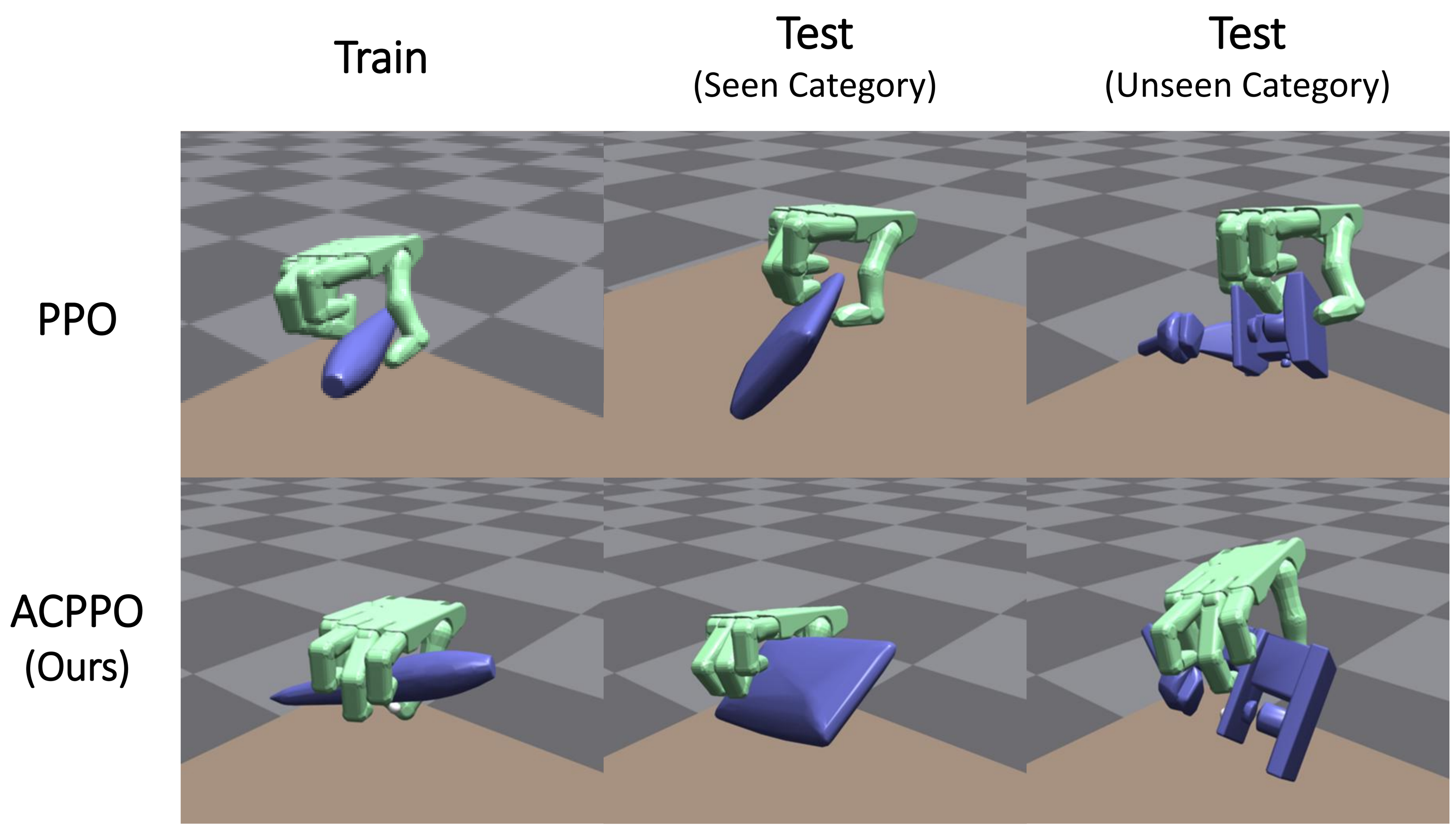
Experiments

- We evaluate our method on the DexGraspNet dataset.
- Training set: 3200 object instances across 99 categories
- Test set
  - Seen Category: 141 new objects from the categories in the training set
  - Unseen Category: 100 objects from novel categories

PPO vs ACPPO

Method	PPO	ACPPO
Importance Sampling	$\rho_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_0}(a_t \mid s_t)}$	$\rho_{t,h}^{ch}(\theta) = \frac{\pi_\theta(a_{t:t+h-1}\|s_t)}{\pi_{\theta_0}(a_{t:t+h-1}\|s_t)}$
Advantage	$Q(s_t, a_t) - V(s_t)$	$\sum_{k=0}^{h-1}\gamma^k r_{t+k} + \gamma^h V^\pi(s_{t+h}) - V^\pi(s_t)$
GAE	$\hat{A}_t^{\mathrm{GAE}(\lambda)} = \delta_t + \underbrace{\sum_{j=1}^{\infty}(\gamma\lambda)^j \delta_{t+j}}_{\text{tail bias}}$	$\hat{A}_t^{\mathrm{GAE}(\lambda)} = \underbrace{\sum_{j=0}^{h-1}(\gamma\lambda)^j \delta_{t+j}}_{\text{inside chunk}} + \underbrace{\sum_{j=h}^{\infty}(\gamma\lambda)^j \delta_{t+j}}_{\text{tail bias}}$
Loss	$\mathbb{E}_{\pi_{\theta_0}}\left[\mathcal{C}_\epsilon\left(\rho_t(\theta), \hat{A}_t\right)\right]$	$\mathbb{E}_t\left[\mathcal{C}_\epsilon\left(\rho_{t,h}^{ch}(\theta), \hat{A}_t\right)\right]$

Qualitative Results



- ACPPO (bottom) generates more stable grasps compared to PPO (top).

Quantitative Results

Method	Train (%)	Test (%)		Training Time (s)
		Seen	Unseen	
ACFQL	79.1	77.2	79.0	**10,126**
PPO	82.3	82.1	82.0	11,702
ACPPO (Ours)	**87.9**	**87.8**	**86.6**	10,600

- ACPPO with action chunk size of 2 outperforms the prior ACRL method (ACFQL) and PPO in grasp success rates.
- ACPPO is also 9.4% faster than PPO in training, benefitting from the reduced forward passes of the policy network.

Conclusion

- ACPPO is an action chunked RL algorithm that exploits the state-value function as its critic, functioning in high DoF environments.
- ACPPO encourages temporally coherent exploration, providing an improved critic which leads to an enhanced policy.

Full Paper