From **Human, Go Beyond:** The Robotic Era

Shanghai Qizhi Institute & Shanghai Jiao Tong University

Kun Lei

13.01.2026

What the robot needs to do ?

- Reliability
- Efficiency
- Robustness

☐ Motivation

– The pre-training and post-training of our human beings

"**Pre-training**" under supervision of parents

"Self-supervised **finetuning**"

Improve generalization ability ?

- **We finetune ourself for a lot of objectives, such as …**

acceleration

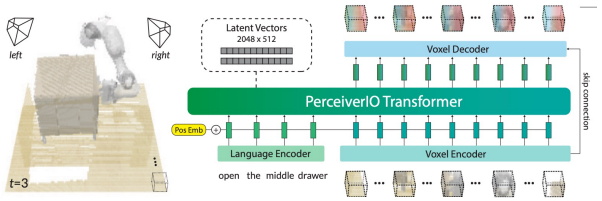
precision

Robustness & safety

☐ Motivation

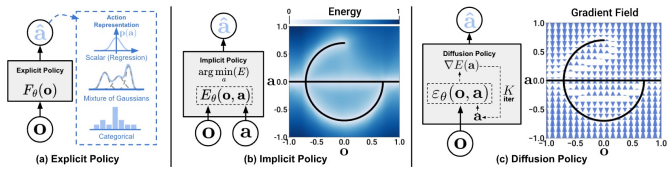
– Learn from demonstration, but also limited by the dataset

PerACT



Diffusion Policy

Visuomotor Policy Learning via Action Diffusion



- ✓ Imitation Learning
- Simple and efficient

😭 Drawbacks:

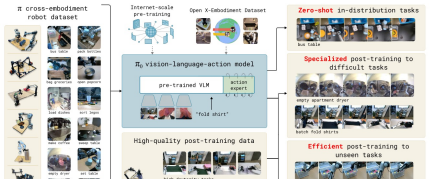
- Expert demonstration
- Poor generalization ability

π_0: A Vision-Language-Action Flow Model for General Robot Control

Physical Intelligence

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, Ury Zhilinsky

https://physicalintelligence.company/blog/pi0



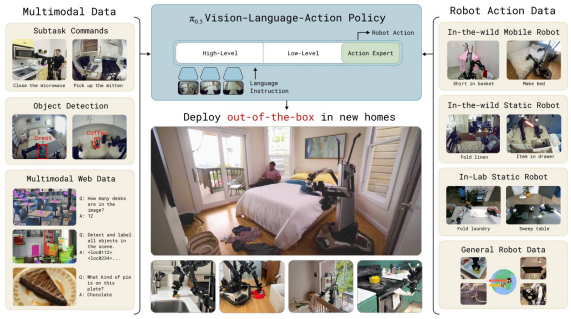
π0

$\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization

Physical Intelligence

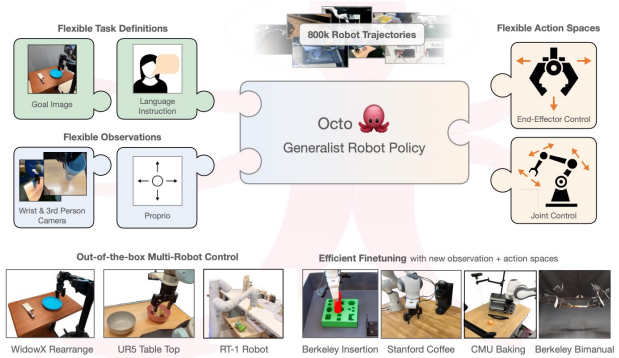
Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, Ury Zhilinsky

https://pi.website/blog/pi05



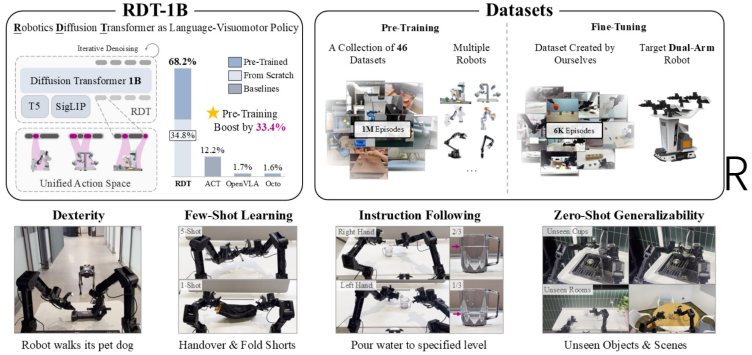
Π0.5

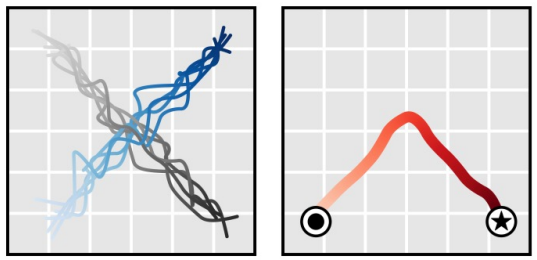
Octo



- ✓ Offline RL
- Learn from Suboptimal data
- Better generalization ability

RDT



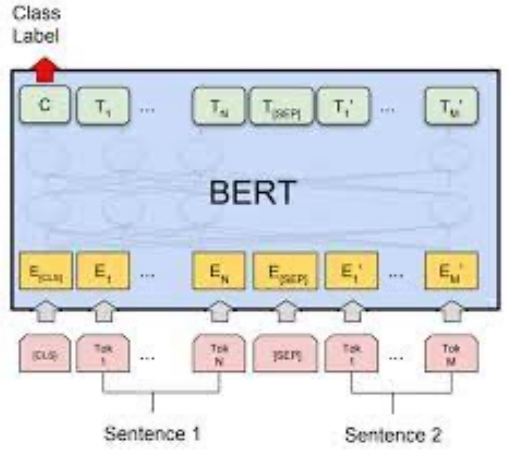


"Stich"

What can we do if the era of "GPT-2" in robotics really comes ?

☐ Motivation

– Supervised Learning is great, but also limited by the dataset

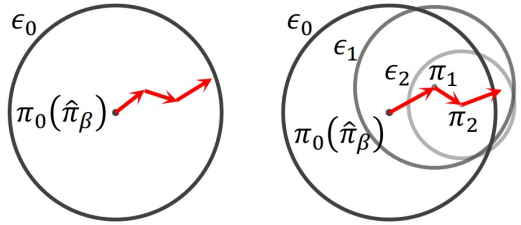


Successes of this paradigm in research areas of CV and NLP

☐ How this paradigm work for robotic?

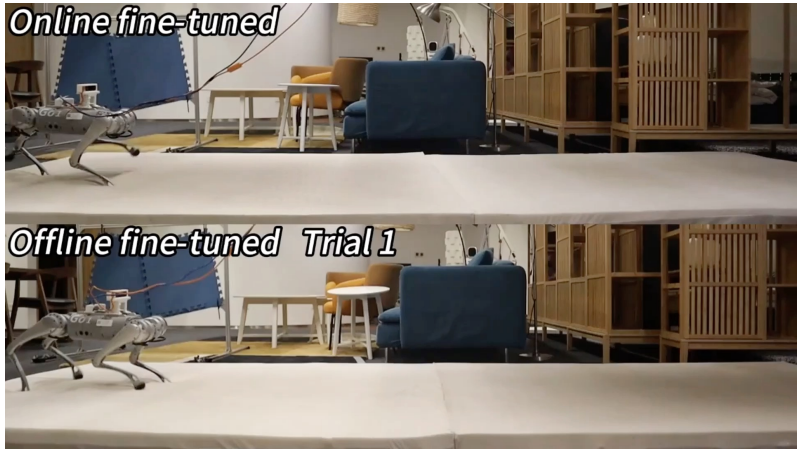
- ✓ Imitation learning
- ✓ Offline reinforcement learning
- ✓ Offline to online RL finetuning

What we did toward this goal

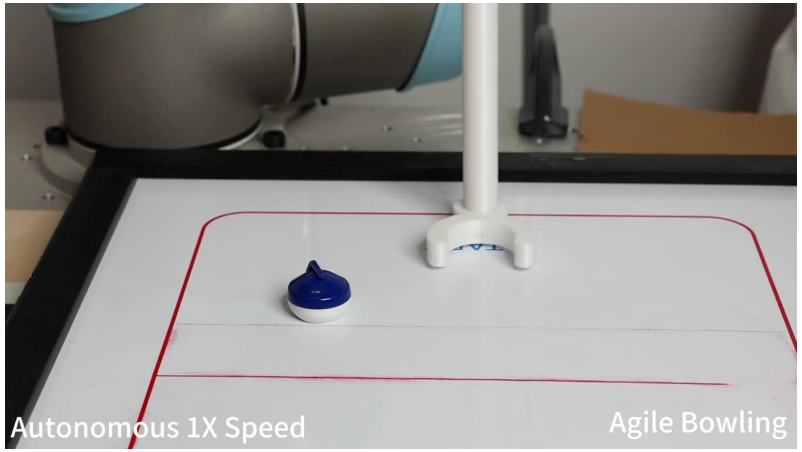


1. Behavior proximal policy optimization
Zifeng Zhuang*, **Kun Lei***, Jinxin Liu, Donglin Wang, Yilang Guo.
International Conference on Learning Representations (ICLR), 2023

2. Uni-O4: Unifying Online and Offline Deep Reinforcement Learning with Multi-Step On-Policy Optimization
Kun Lei, Zhengmao He*, Chenhao Lu*, Kaizhe Hu,
Yang Gao, Huazhe Xu.
International Conference on Learning Representations (ICLR), 2024



3. RL-100: Performant Robotic Manipulation with Real-World Reinforcement Learning
Kun Lei*, Huanyu Li*, Dongjie Yu*, Zhenyu Wei*, Lingxiao Guo,
Zhennan Jiang, Ziyu Wang, Shiyu Liang, Huazhe Xu.
Preparing to submit.



The post-training of robot (foundation) model

Offline reinforcement learning
-Pretraining/Finetuning

Offline-to-online finetuning
- for fast adaptation
- with safety consideration
- for task acceleration

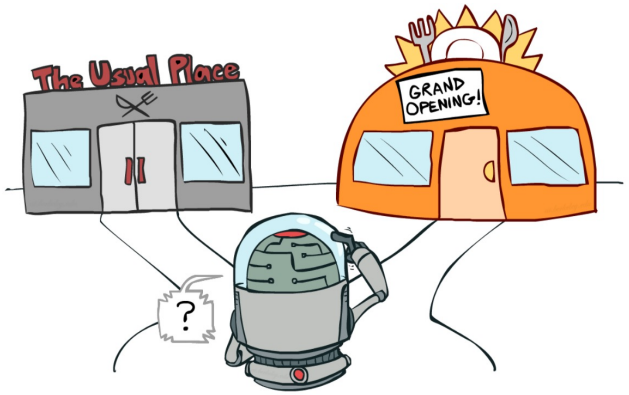
Finetuning from multi-modal perception

Behavior proximal policy optimization

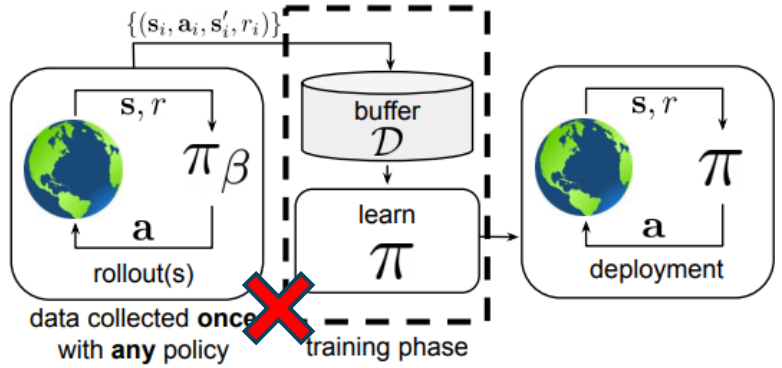
Zifeng Zhuang*, **Kun Lei*,** Jinxin Liu, Donglin Wang, Yilang Guo

Behavior proximal policy optimization (BPPO)

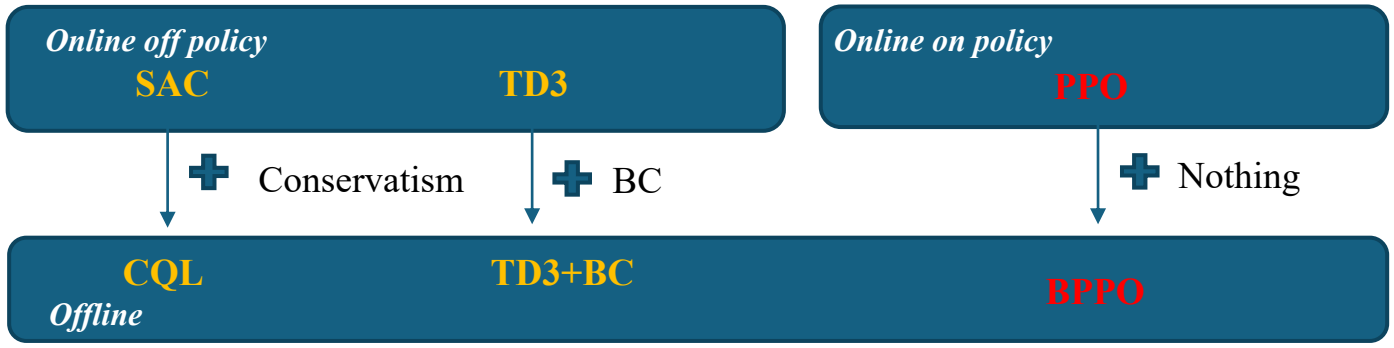
- Online RL
  - ✓ Exploration is crucial



- Offline RL
  - 😭 Exploration is limited



- **Conservative methods**



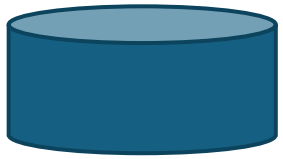
``An offline version of PPO"

Key insight:

The first work paid attention to the policy learning instead of the value learning in offline setting

Method

- Stage 1: Supervised Learning



Offline dataset

Behavior cloning ➡ Behavior policy π_β

Proposition 1. *For offline dataset* $\mathcal{D} = \left\{(s_t, a_t, s_{t+1}, r_t)_{t=1}^N\right\}$ *and policy* $\hat{\pi}_\beta$*, the total variational divergence can be expressed as* $D_{TV}\left(\mathcal{D}\|\hat{\pi}_\beta\right)[s_t] = \frac{1}{2}\left(1 - \hat{\pi}_\beta\left(a_t|s_t\right)\right)$.

Regression (Sarsa style) ➡ Q_{π_β}

Regression (Monte Carlo return) ➡ V_{π_β}

Advantage $A_\pi = \approx Q_{\pi_\beta} - V_{\pi_\beta}$

- Stage 2: Policy improvement using PPO

Offline monotonic improvement over behavior policy

- For two policies π and π', the Performance Difference $J_\Delta(\pi', \pi)$ can be measured by the advantage function:

Online: $$J_\Delta\left(\pi', \pi\right) = \mathbb{E}_{\tau\sim P_{\pi'}(\tau)}\left[\sum_{t=0}^{H}\gamma^t A_\pi(s_t, a_t)\right] = \mathbb{E}_{s\sim\rho_{\pi'}(\cdot), a\sim\pi'(\cdot|s)}\left[A_\pi(s,a)\right]$$

Offline: $$\widehat{J}_\Delta\left(\pi, \hat{\pi}_\beta\right) = \mathbb{E}_{s\sim\rho_{\mathcal{D}}(\cdot), a\sim\pi(\cdot|s)}\left[A_{\hat{\pi}_\beta}(s,a)\right]$$

From TRPO

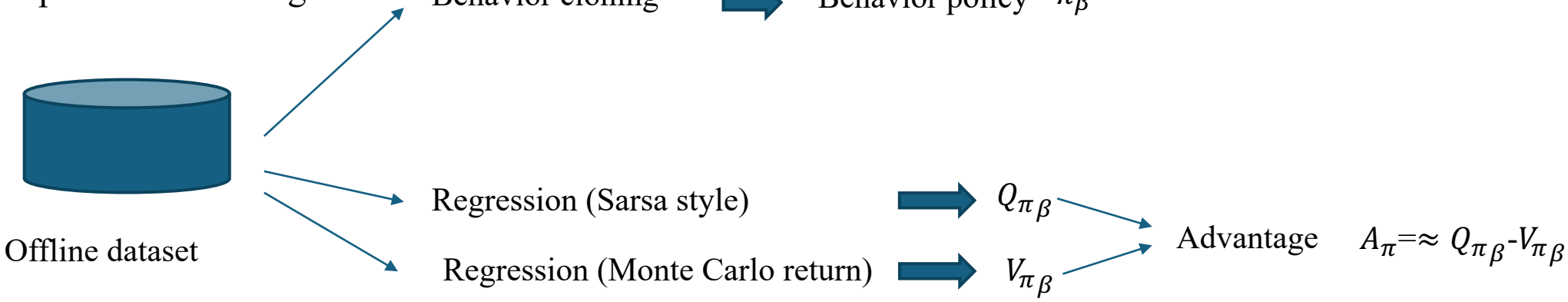
Theorem 2. *Given the distance* $D_{TV}\left(\pi\|\hat{\pi}_\beta\right)[s]$ *and* $D_{TV}\left(\mathcal{D}\|\hat{\pi}_\beta\right)[s] = \frac{1}{2}\left(1 - \hat{\pi}_\beta\left(a_t|s_t\right)\right)$*, we can derive the following bound:*

Only in offline

$$J_\Delta\left(\pi, \hat{\pi}_\beta\right) \geq \widehat{J}_\Delta\left(\pi, \hat{\pi}_\beta\right) - 4\gamma\mathbb{A}_{\hat{\pi}_\beta}\cdot\max_s D_{TV}\left(\pi\|\hat{\pi}_\beta\right)[s]\cdot\mathop{\mathbb{E}}_{s\sim\rho_{\hat{\pi}_\beta}(\cdot)}\left[D_{TV}\left(\pi\|\hat{\pi}_\beta\right)[s]\right]$$
$$- 2\gamma\mathbb{A}_{\hat{\pi}_\beta}\cdot\max_s D_{TV}\left(\pi\|\hat{\pi}_\beta\right)[s]\cdot\mathop{\mathbb{E}}_{s\sim\rho_{\mathcal{D}}(\cdot)}\left[1 - \hat{\pi}_\beta\left(a|s\right)\right],$$

BPPO - Method

- Stage 1: Supervised Learning

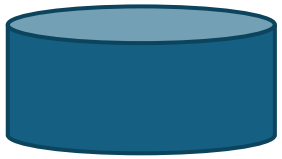


- Stage 2: Policy improvement using PPO

$$L_k(\pi) = \mathbb{E}_{s\sim\rho_{\mathcal{D}}(\cdot),a\sim\pi_k(\cdot|s)}\left[\min\left(\frac{\pi(a|s)}{\pi_k(a|s)}A_{\pi_k}(s,a),\ \text{clip}\left(\frac{\pi(a|s)}{\pi_k(a|s)},1-2\epsilon,1+2\epsilon\right)A_{\pi_k}(s,a)\right)\right]$$

PPO objective with advantage replacement: $\mathrm{A}_\pi(s,a) \approx Q_{\hat{\pi}_\beta} - V_{\hat{\pi}_\beta}$

- RL partitioning
  - ✓ Data interaction patterns



  - ➢ On-policy RL
  - ➢ off-policy RL
  - ➢ Offline RL

➢ A general version of RL policy update forms

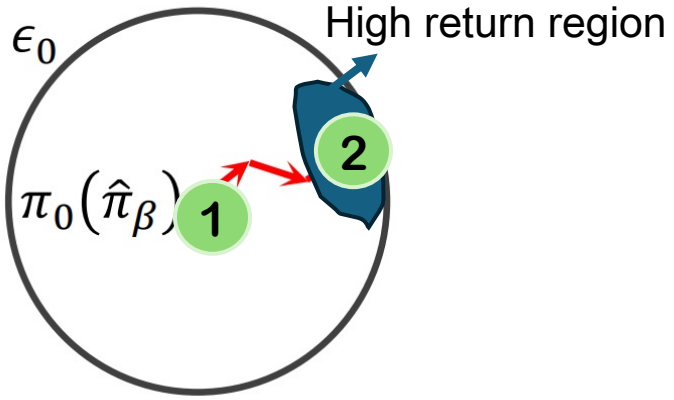
Policy evaluation

$$\hat{Q}^{k+1} = \arg\min_{Q} \mathbb{E}_{(s,a)\sim w}\Big[\Big(\underbrace{r + \gamma\, \mathbb{E}_{s'|s,a}\, \mathbb{E}_{a'\sim\pi^k(\cdot|s')}\big[\hat{Q}(s',a')\big]}_{T^{\pi^k}\hat{Q}} - Q(s,a)\Big)^2\Big].$$ ••• 1

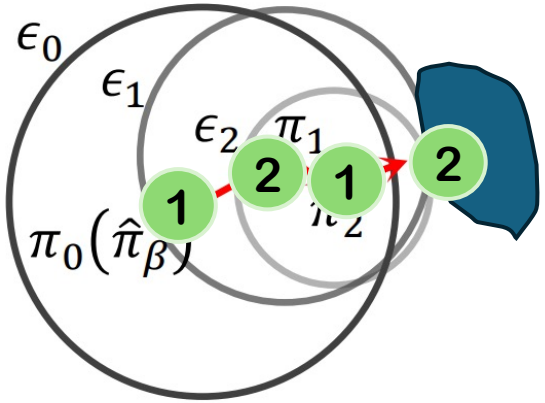
Policy improvement

$$\pi^{k+1} = \arg\max_{\pi} \mathbb{E}_{s\sim\rho,\, a\sim\pi}\big[\hat{Q}^{k+1}(s,a)\big] - \beta\, \mathbb{E}_{s\sim\rho}\big[D_{\mathrm{KL}}(\pi \,\|\, \pi_{\mathrm{ref}})\big].$$ ••• 2

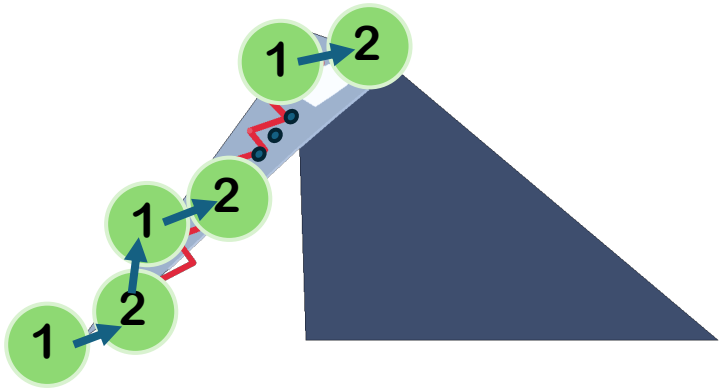
➢ One-step RL



➢ Multi-step RL



➢ Iterative RL

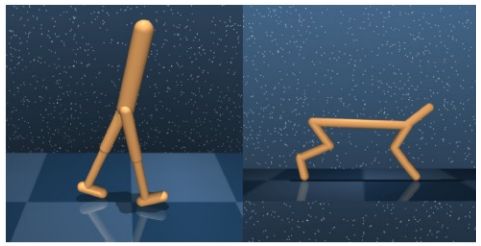


- One-step RL/IQL/IDQL…
- BPPO → Uni-O4/RL-100…
- PPO/SAC/…

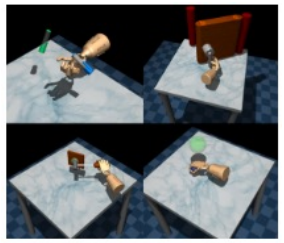
More conservative, but stable

Aggressive exploration, but occasionally crash

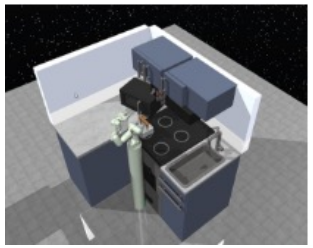
Environments & Main Results



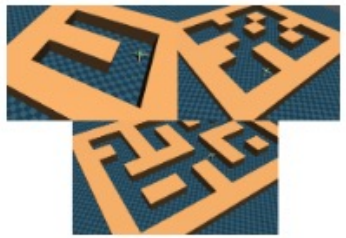
Gym locomotion



Adroit



Kitchen



Antmaze

Suite	Environment	Iterative methods		Onestep methods		BC (Ours)	BPPO (Ours)
		CQL	TD3+BC	Onestep RL	IQL		
Gym	halfcheetah-medium-v2	44.0	**48.3**	**48.4**	**47.4**	43.5±0.1	44.0±0.2
	hopper-medium-v2	58.5	59.3	59.6	66.3	61.3±3.2	**93.9±3.9**
	walker2d-medium-v2	72.5	**83.7**	81.8	78.3	74.2±4.6	**83.6±0.9**
	halfcheetah-medium-replay-v2	**45.5**	**44.6**	38.1	**44.2**	40.1±0.1	41.0±0.6
	hopper-medium-replay-v2	95.0	60.9	**97.5**	94.7	66.0±18.3	92.5±3.4
	walker2d-medium-replay-v2	77.2	**81.8**	49.5	73.9	33.4±11.2	77.6±7.8
	halfcheetah-medium-expert-v2	91.6	90.7	**93.4**	86.7	64.4±8.5	**92.5±1.9**
	hopper-medium-expert-v2	105.4	98.0	103.3	91.5	64.9±7.7	**112.8±1.7**
	walker2d-medium-expert-v2	108.8	110.1	**113.0**	109.6	107.7±3.5	**113.1±2.4**
	Gym locomotion-v2 total	*698.5*	*677.4*	*684.6*	*692.4*	*555.5±57.2*	***751.0±21.8***
Adroit	pen-human-v1	37.5	8.4*	90.7*	71.5	61.6±9.7	**117.8±11.9**
	hammer-human-v1	4.4	2.0*	0.2*	1.4	2.0±0.9	**14.9±3.2**
	door-human-v1	9.9	0.5*	-0.1*	4.3	7.8±3.5	**25.9±7.5**
	relocate-human-v1	0.2	-0.3*	2.1*	0.1	0.1±0.0	**4.8±2.2**
	pen-cloned-v1	39.2	41.5*	60.0	37.3	58.8±16.0	**110.8±6.3**
	hammer-cloned-v1	2.1	0.8*	2.0	2.1	0.5±0.2	**8.9±5.1**
	door-cloned-v1	0.4	-0.4*	0.4	1.6	0.9±0.8	**6.2±1.6**
	relocate-cloned-v1	-0.1	-0.3*	-0.1	-0.2	-0.1±0.0	**1.9±1.0**
	adroit-v1 total	*93.6*	*52.2*	*155.2*	*118.1*	*131.6±31.1*	***291.4±38.8***
Kitchen	kitchen-complete-v0	43.8	0.0*	2.0*	62.5	55.0±11.5	**91.5±8.9**
	kitchen-partial-v0	49.8	22.5*	35.5*	46.3	44.0±4.9	**57.0±2.4**
	kitchen-mixed-v0	51.0	25.0*	28.0*	51.0	45.0±1.6	**62.5±6.7**
	kitchen-v0 total	*144.6*	*47.5*	*65.5*	*159.8*	*144.0±18.0*	***211.0±18.0***
	locomotion+kitchen+adroit	*936.7*	*777.1*	*905.3*	*970.3*	*831.1±106.3*	***1253.4±78.6***

⬆️50% performance based on BC

- How BPPO avoid overestimating:
  - Once policy evaluation
  - CLIP function in PPO loss

- Issues in BPPO
  - The needs of online policy evaluation
  - The performance is highly related to the estimated behavior policy
  - Just offline

Uni-O4

The post-training of robot (foundation) model

Offline reinforcement learning
-Pretraining/Finetuning

Offline-to-online finetuning
- for fast adaptation
- with safety consideration
- for task acceleration

Finetuning from multi-modal perception

Uni-O4: Unifying Online and Offline Deep Reinforcement Learning with Multi-Step On-Policy Optimization

Kun Lei[1] Zhengmao He[14] Chenhao Lu[2] Kaizhe Hu[12] Yang Gao[123] Huazhe Xu[123]

1 Shanghai Qi Zhi Institute. 2 Tsinghua University, IIIS. 3 Shanghai AI Lab.
4 The Hong Kong University of Science and Technology (Guangzhou).

ICLR

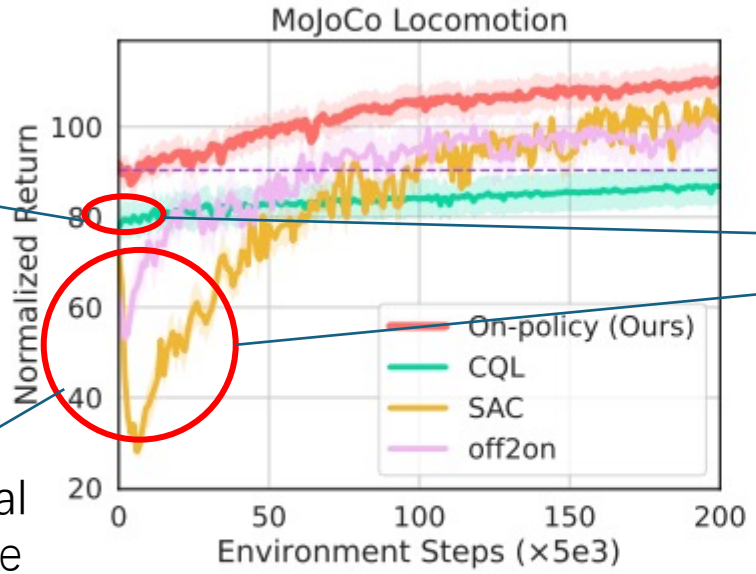
TEA
Tsinghua Embodied AI Laboratory

Uni-O4: Unifying Online and Offline Deep Reinforcement Learning with Multi-Step On-Policy Optimization

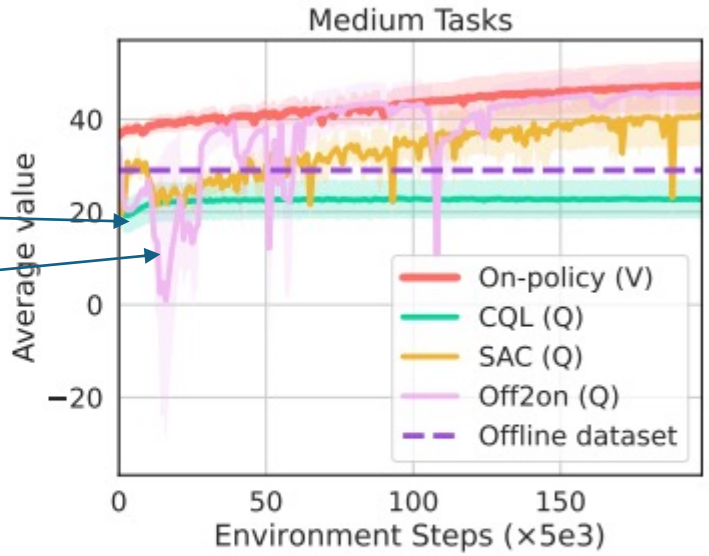
Offline-to-online issues: *distribution shift* due to the *conservatism* used in offline phase.

Over-conservative

``Crash" in the initial offline2online stage



(a) Normalized Return

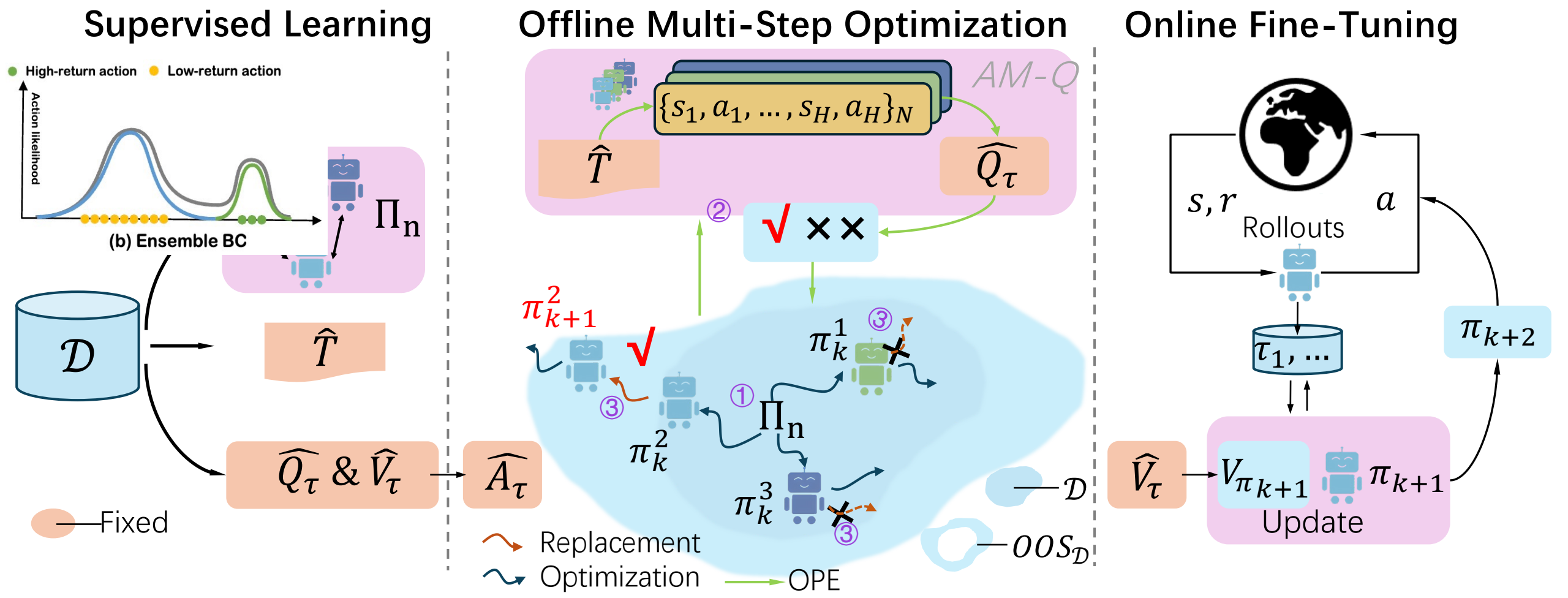


(b) Average value

Can we avoid introducing the conservatism term during offline training and eliminate the need for off-policy evaluation during offline-to-online fine-tuning? → On-policy: Uni-O4 [1]

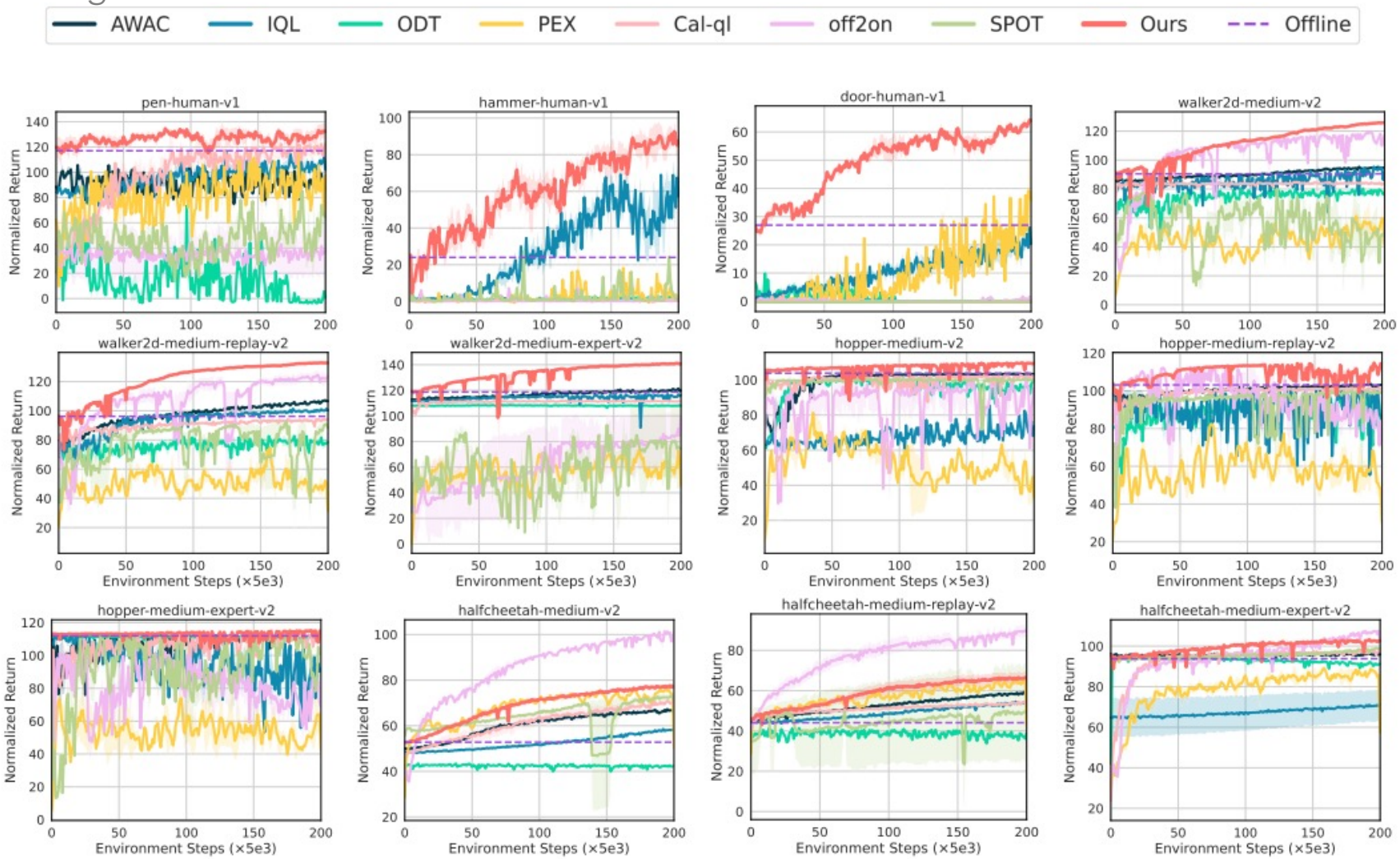
[1] Uni-O4: Unifying Online and Offline Deep Reinforcement Learning with Multi-Step On-Policy Optimization. Kun Lei, Zhengmao He, Chenhao Lu, Kaizhe Hu, Yang Gao, Huazhe Xu. ICLR 2024

Offline-to-online framework

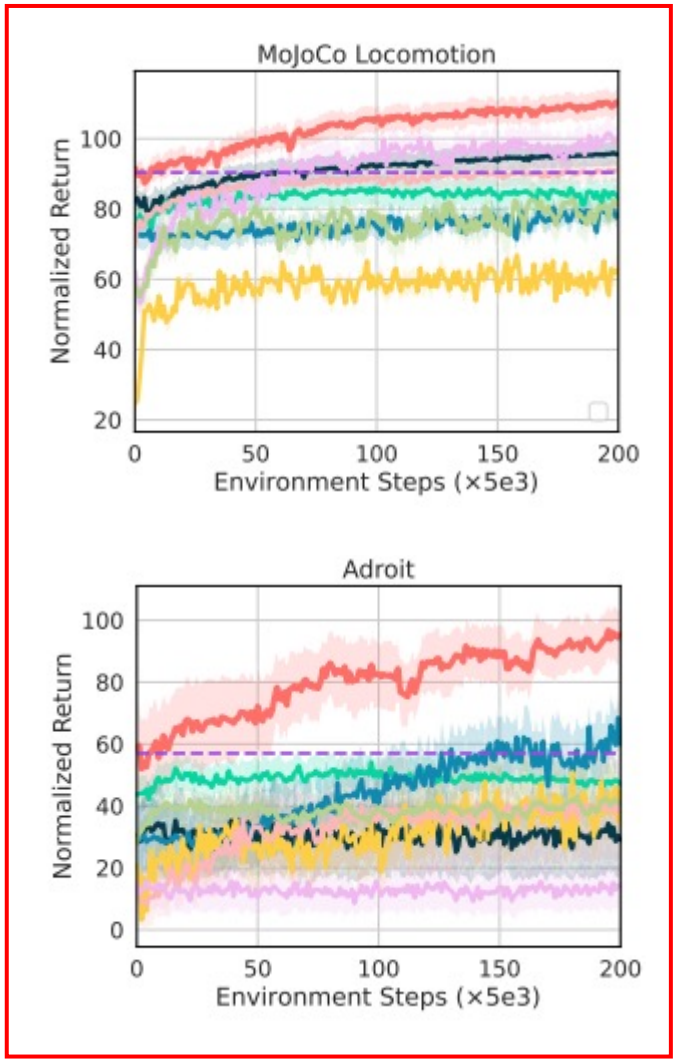


◘ Simulated Results

- Offline phase, offline-to-online phase, real robot setting



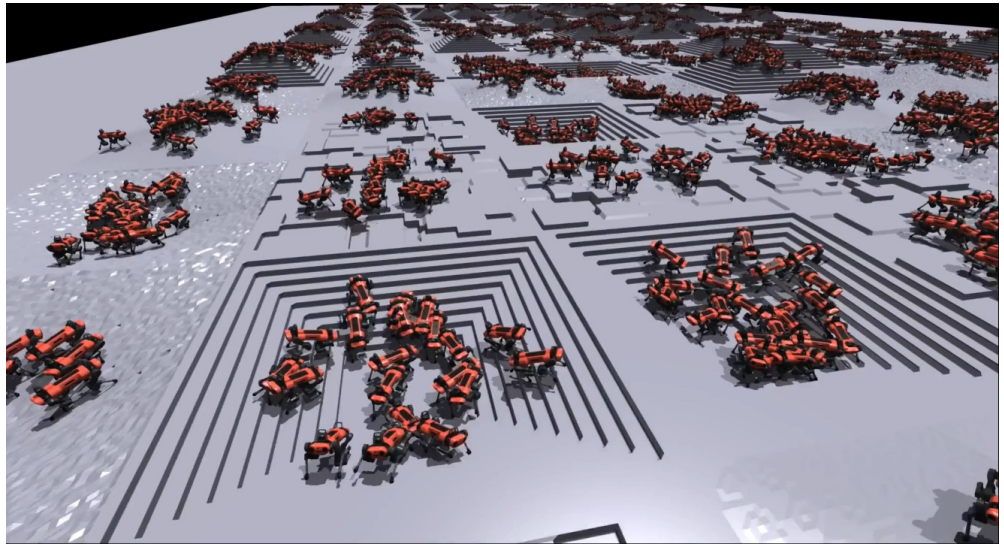
Results over single task



Average results over domains

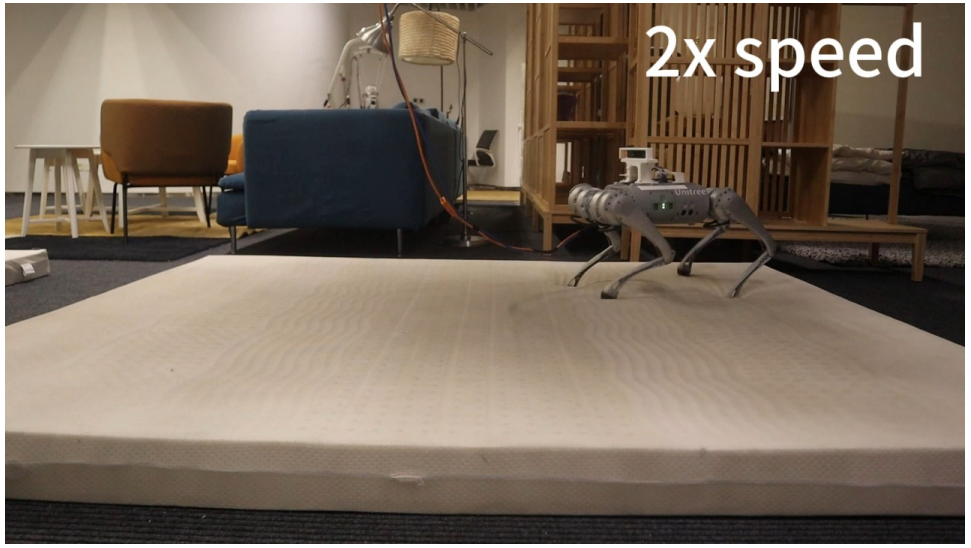
➢ Offline phase, offline-to-online phase, real robot setting

Online Phase (Sim)



Train quadruped robot For several minutes

Deploy in Real-World



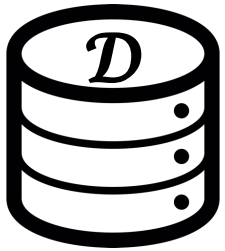
$\tau_1, \tau_2, \ldots$

Data collection

$\mathcal{D}$

Offline fine-tuning Phase (Real-world)

Online Phase (Real-world)

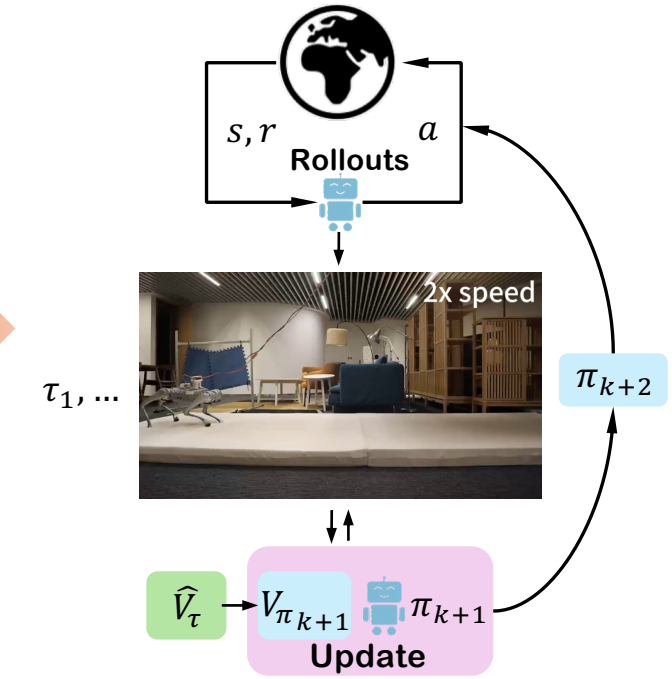


Supervised Learning

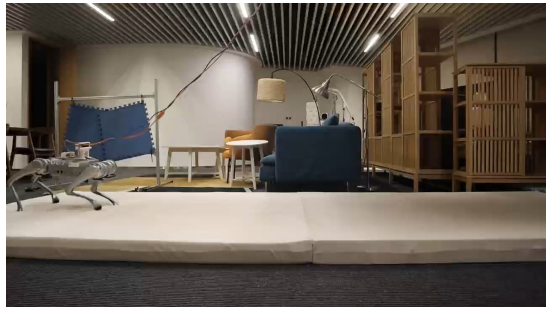
$\hat{T}$

$\widehat{Q_\tau}$ & $\widehat{V_\tau}$

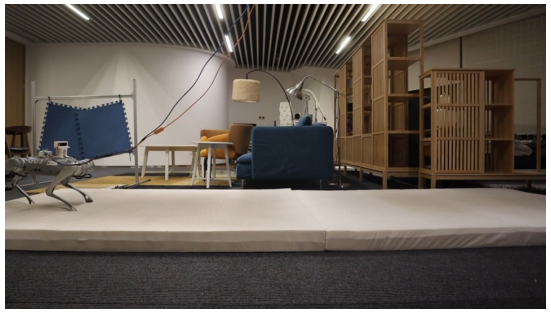
Offline Multi-Step Optimization



Performance Improved on
soft & deformable terrain



But still not satisfactory
when speed is fast



Faster and more robust

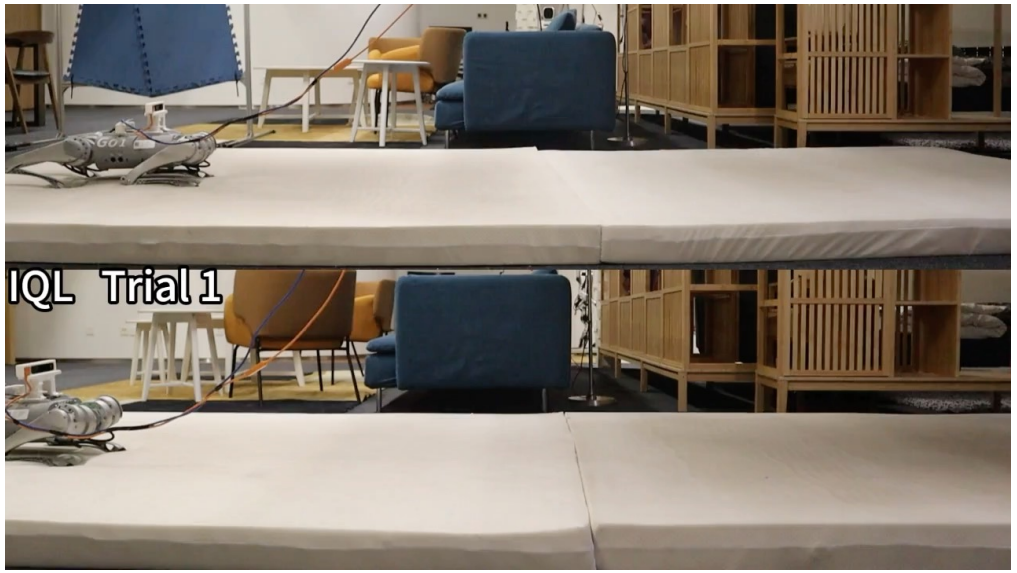
Online Phase (Sim)

Offline Phase (Real-world)

Online Phase (Real-world)

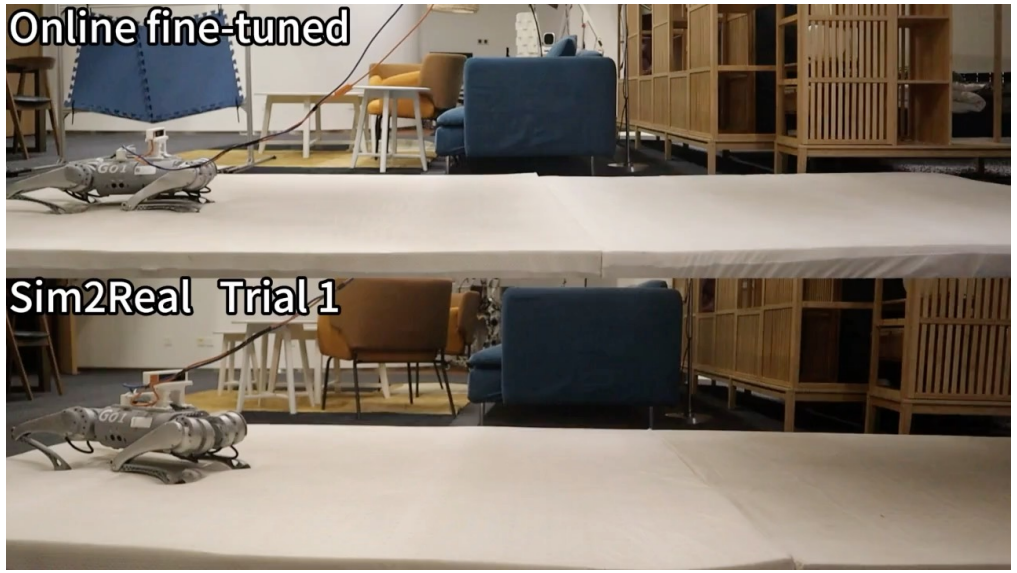
All in one with PPO!

Offline and online fine-tuning comparison



Uni-O4 vs. IQL [1]

Offline-to-online baseline



Uni-O4 vs. Walk these ways [2]

[1] Kostrikov I, Nair A, Levine S. Offline reinforcement learning with implicit q-learning[J]. arXiv preprint arXiv:2110.06169, 2021.
[2] Margolis, Gabriel B., and Pulkit Agrawal. "Walk these ways: Tuning robot control for generalization with multiplicity of behavior." *Conference on Robot Learning*. PMLR, 2023.

Key insight of Uni-O4:

- On-policy RL can unify offline and online setting
- Offline RL could work as a finetuning paradigm

Only explored locomotion tasks.

The post-training of robot (foundation) model

Offline reinforcement learning
-Pretraining/Finetuning

Offline-to-online finetuning
- for fast adaptation
- with safety consideration
- for task acceleration

Finetuning from multi-modal perception

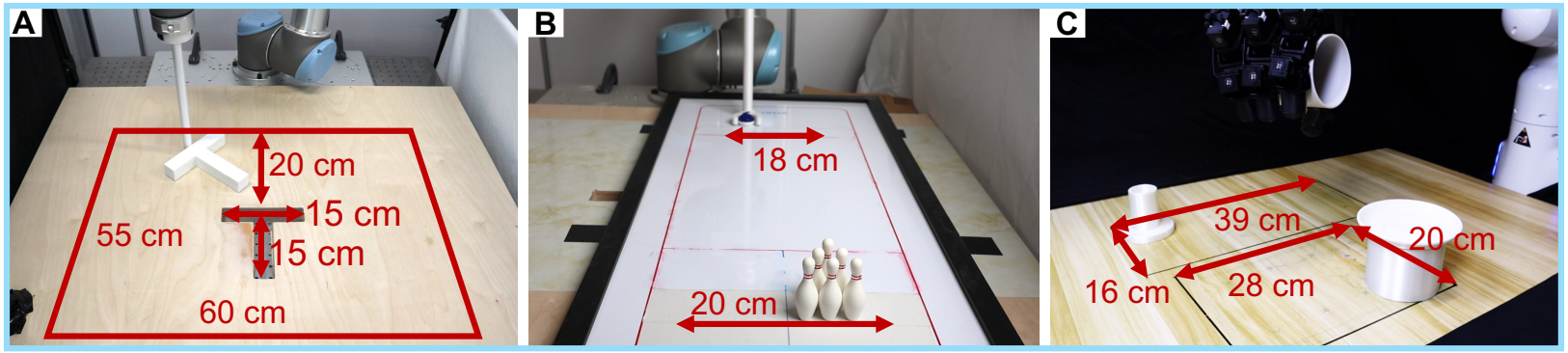
RL-100: Performant Robotic Manipulation with Real-World Reinforcement Learning

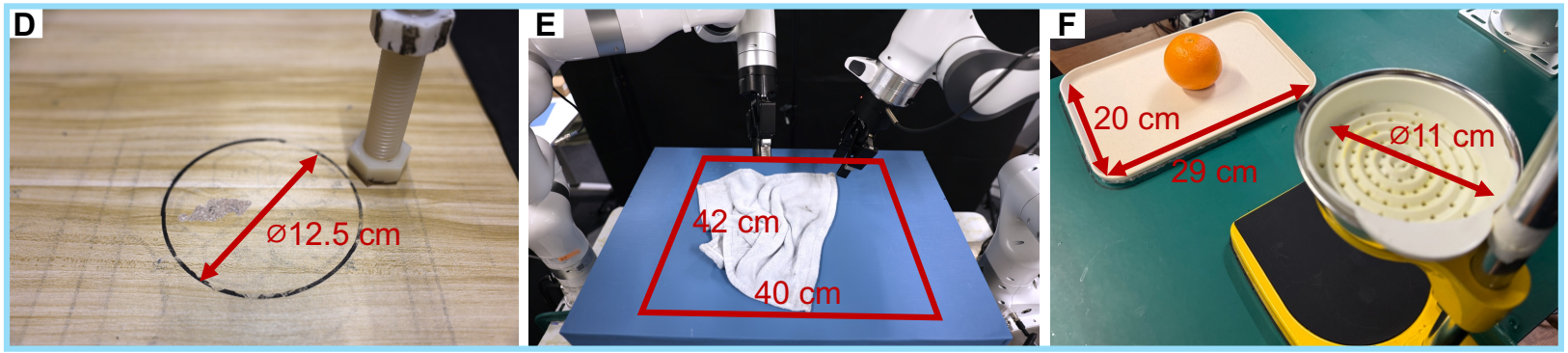
Kun Lei*, Huanyu Li*, Dongjie Yu*, Zhenyu Wei*, Lingxiao Guo, Zhennan Jiang, Ziyu Wang, Shiyu Liang, Huazhe Xu.

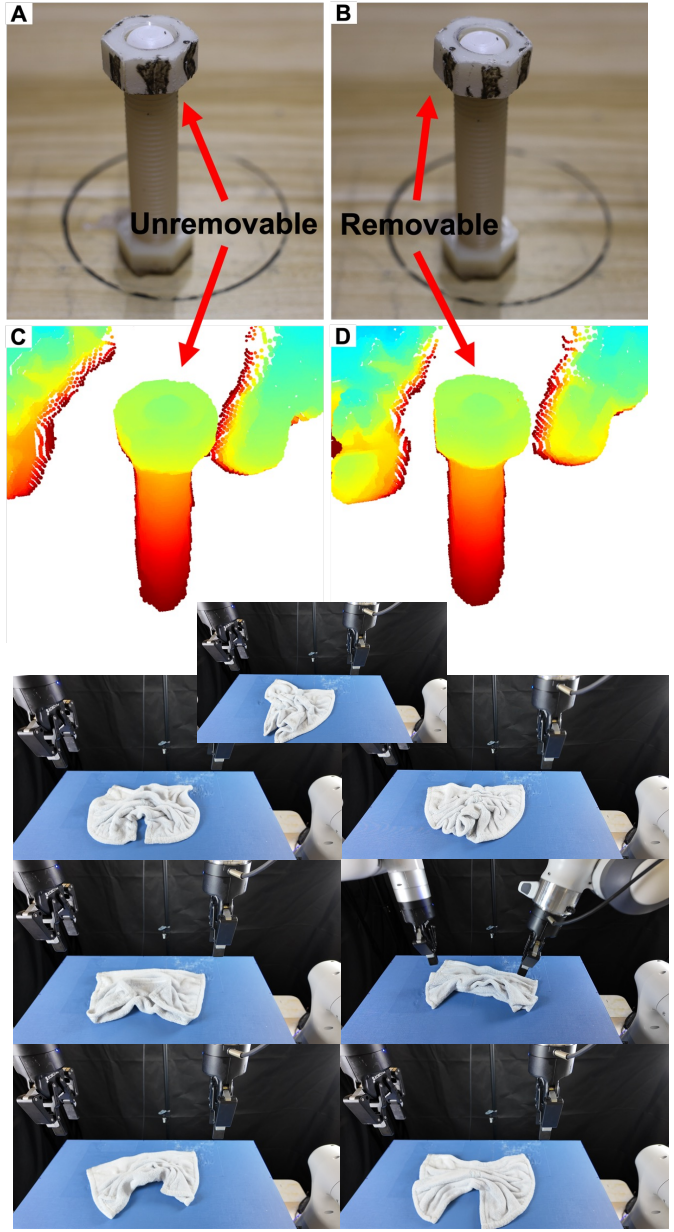
Preparing to submit.

What we did?

- **7** real robot tasks, **900/900** successes. Up to **250** consecutive trials in one task, running **2 hours** nonstop without failure.
- High success rate against physical disturbances, zero-shot, and few-shot adaptation

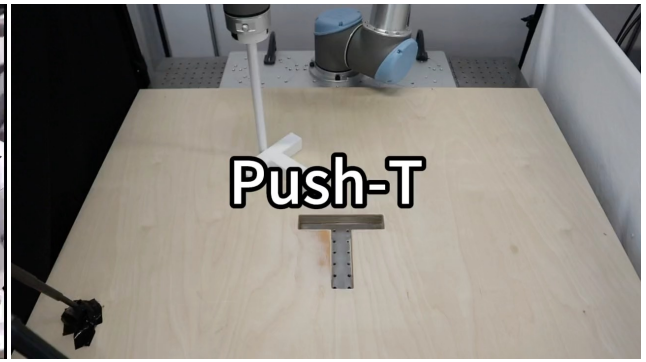




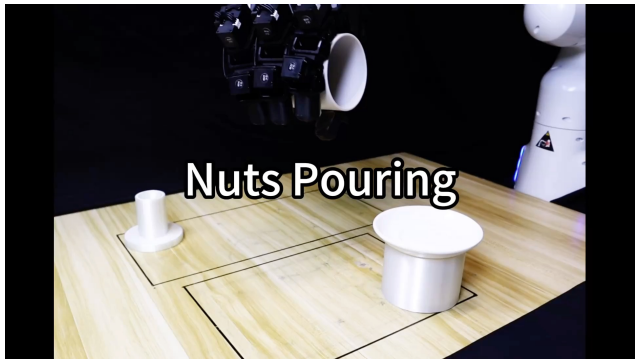


What we did?

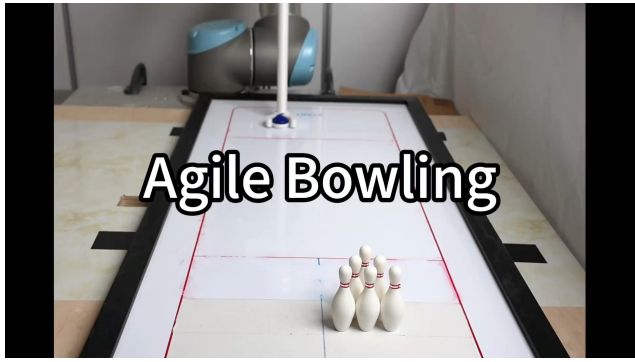
- Folding
  - Dual-arm
  - Deformable



- Dynamic Push-T
  - Rigid-body dynamics



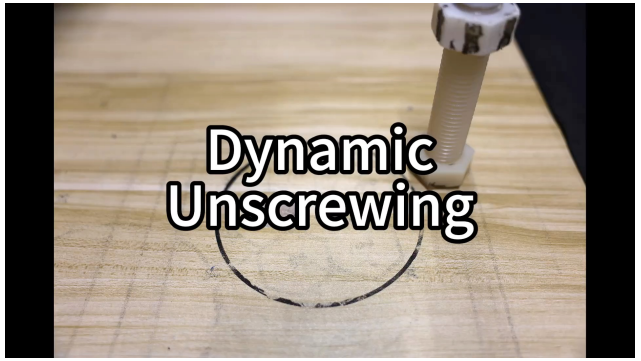
- Pour
  - Fluids / granular



- Bowling
  - agile

- Juicing-stage 1
  - Diff. Size
  - Various inclination

- Juicing-stage 2
  - Deformable
  - Confined-space



- Unscrewing
  - Dynamic

- Serve for 7 hours
  - Zero-shot

■ Key Words: ➢ 250/250 ➢ 2 Hours ➢ Efficient ➢ 7 hours outdoor serving

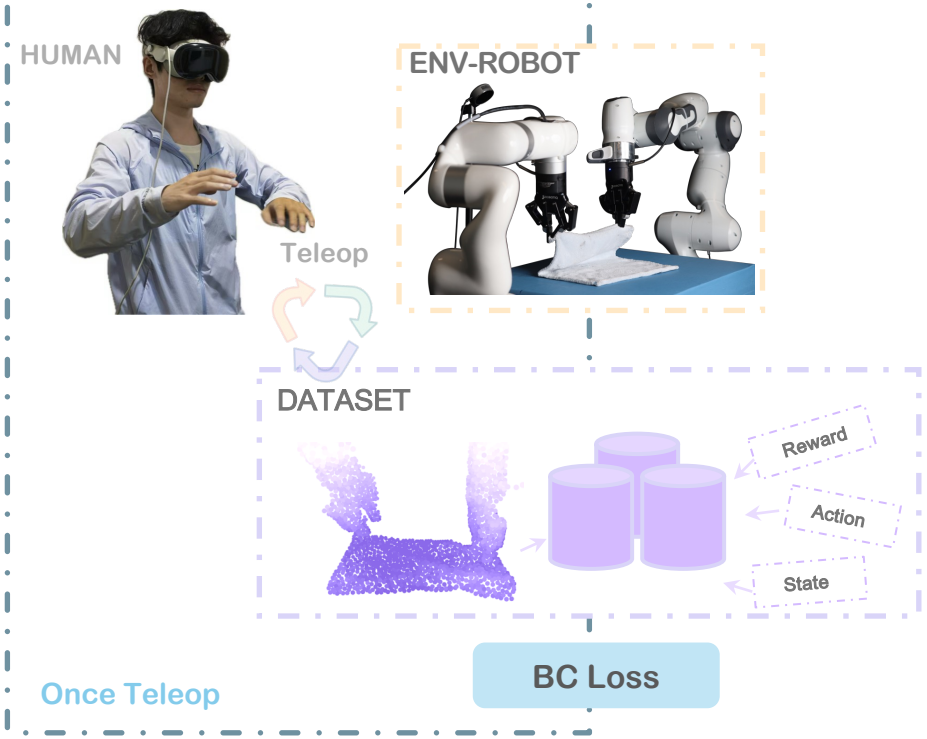
How to do it

Training Pipeline

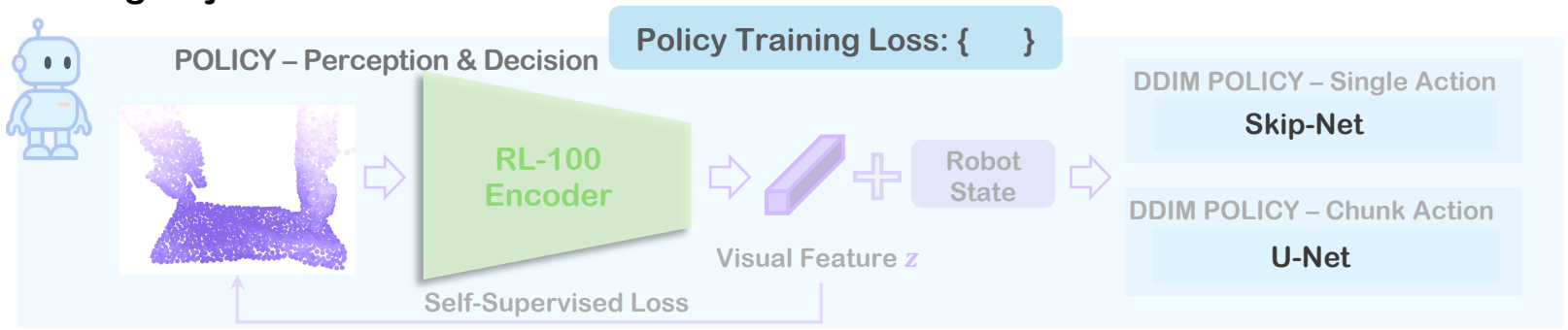
Learn from Human Priors

Post-Training – Iterative Offline

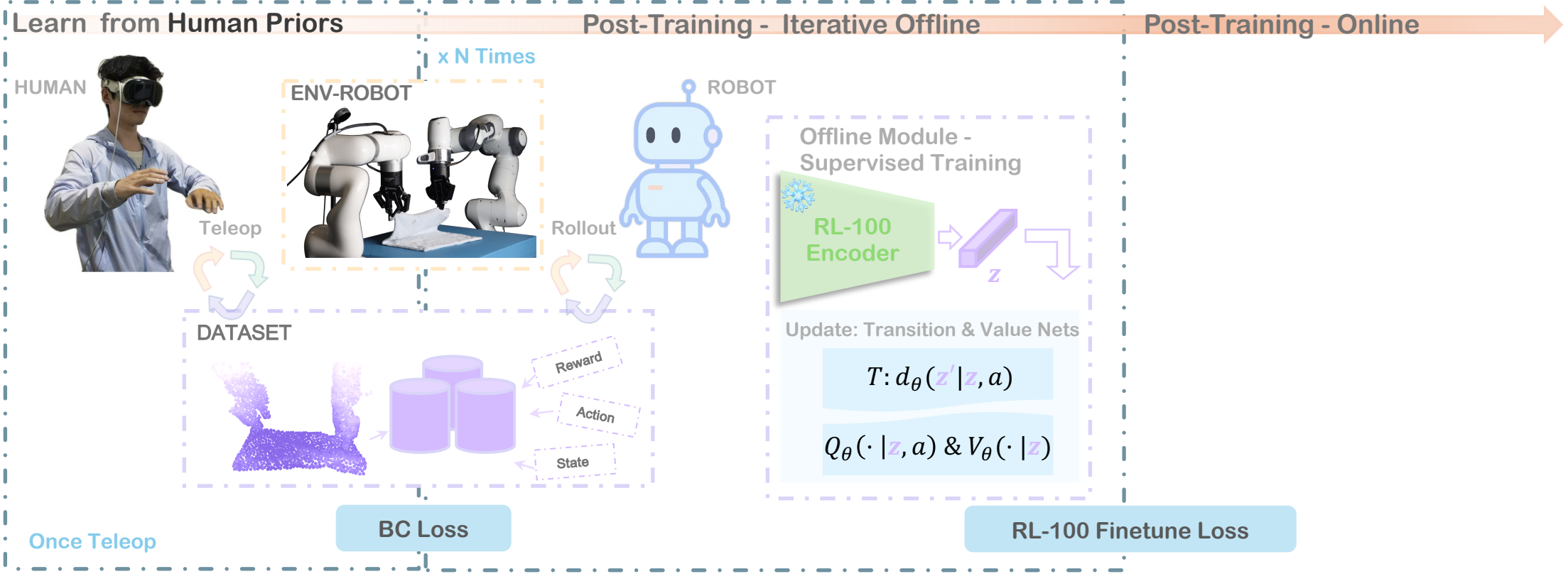
Post-Training - Online



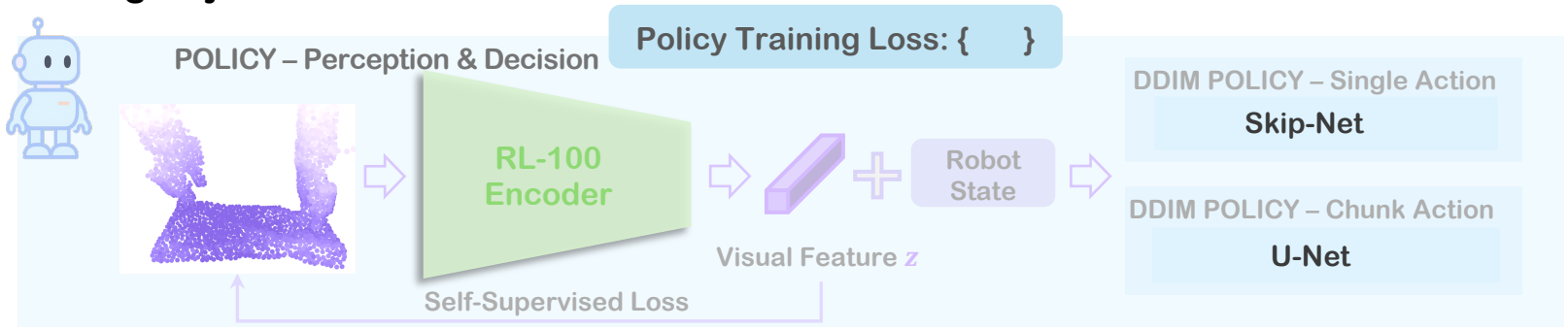
Training Objective



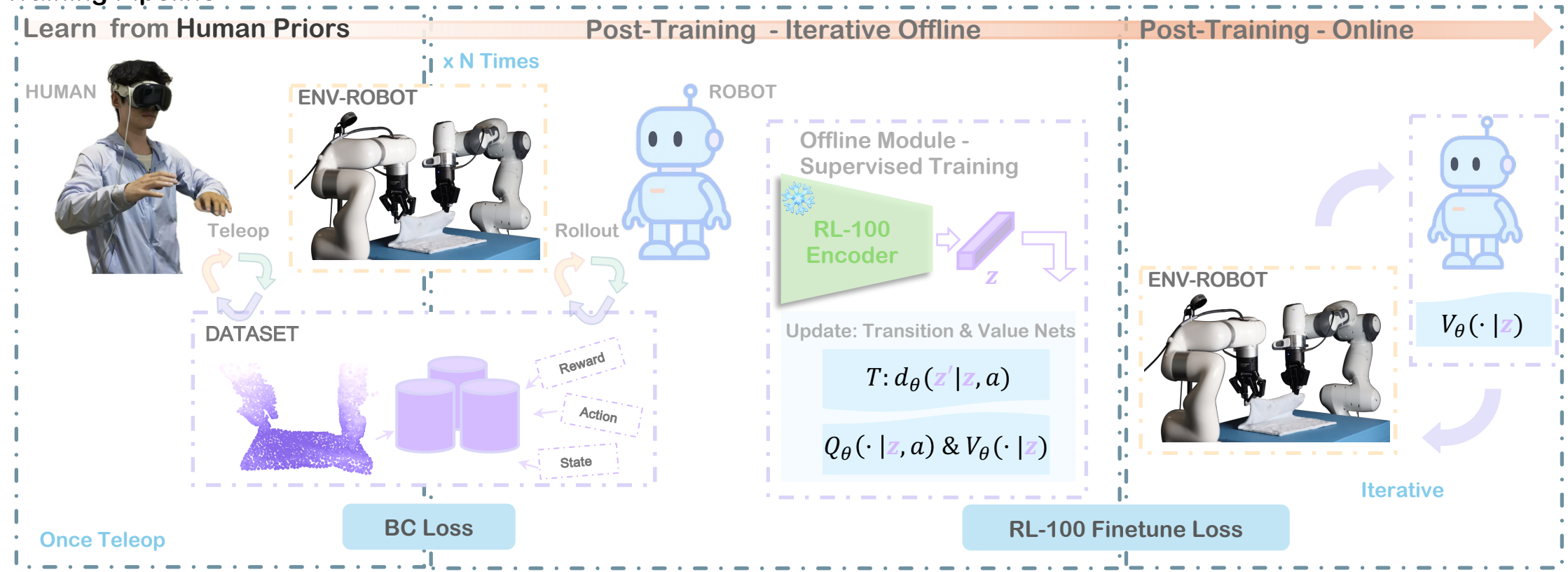
Training Pipeline



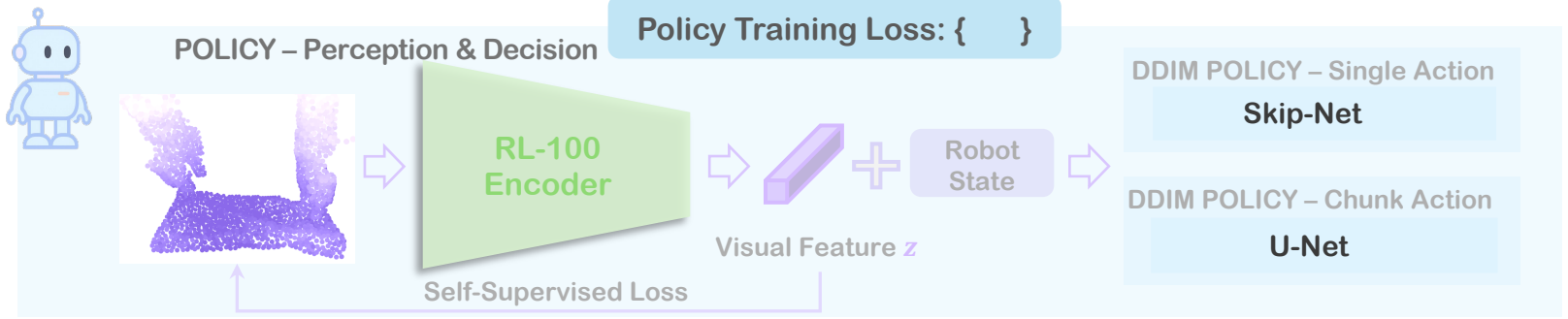
Training Objective



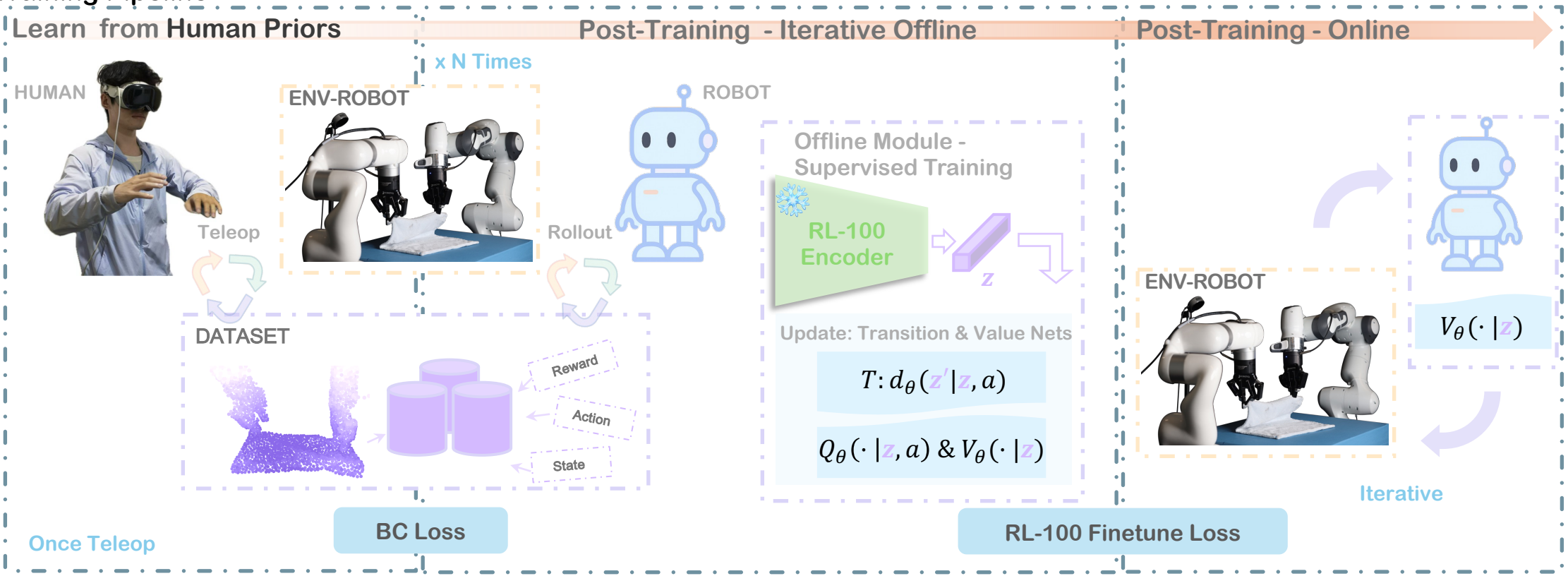
Training Pipeline



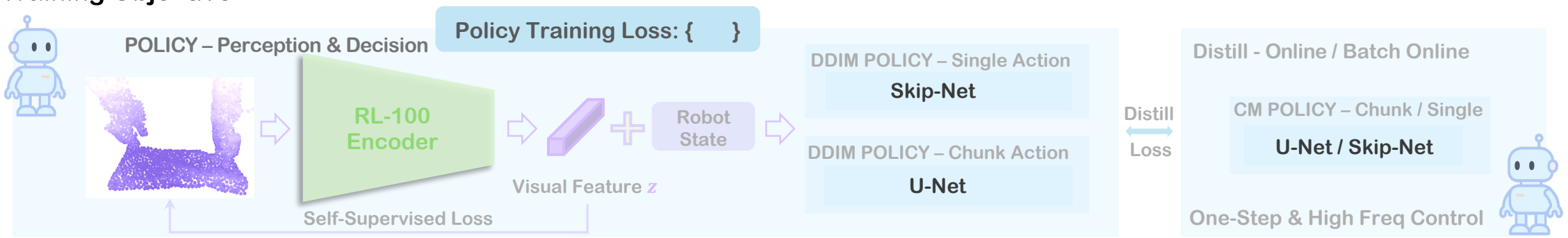
Training Objective



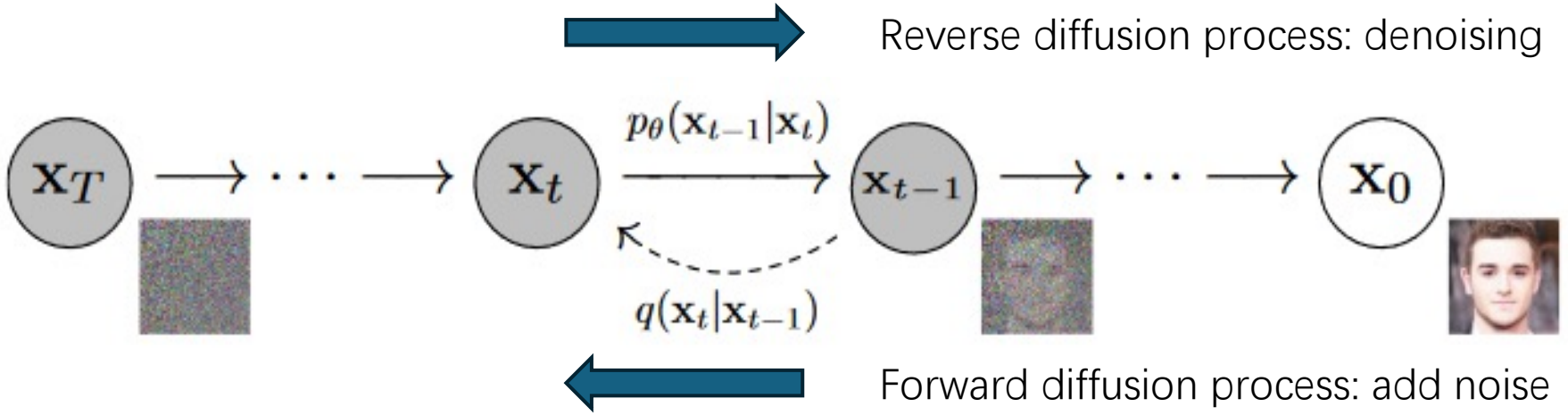
Training Pipeline



Training Objective



- Two-level MDP with DDIM sampling with stochastic form



Policy gradient (PG) loss: $\nabla_\theta J = \mathbb{E}_\pi[\nabla_\theta \log p_\theta(a|s)\, R_\pi]$

PG loss with multi step sampling (DDIM): $\nabla_\theta J = \mathbb{E}_\pi[\sum_{k=0}^{K} \nabla_\theta \log p_\theta\left(x_{\tau_{k-1}}|x_{\tau_k}, s\right) R_\pi]$

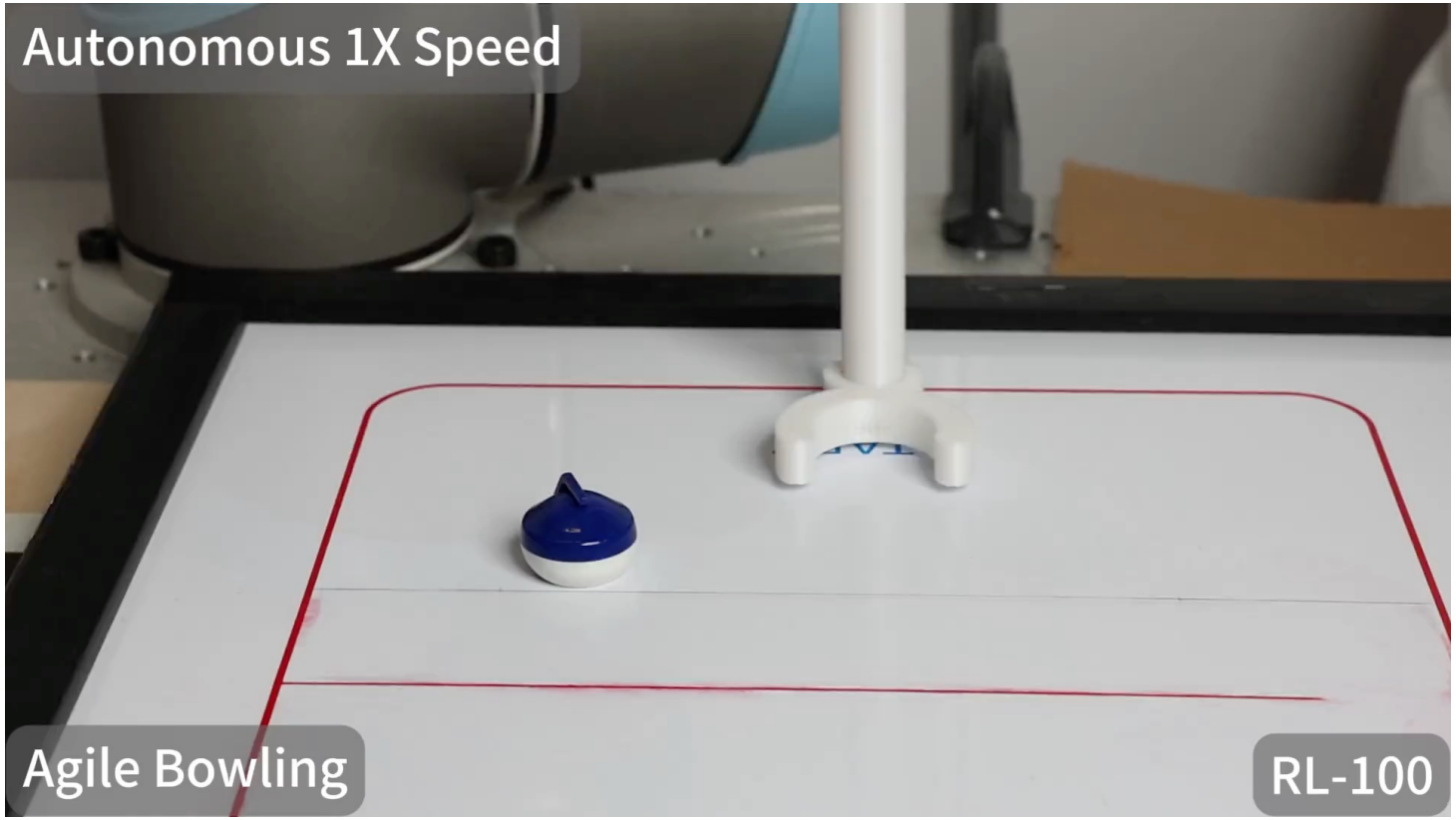
PG loss with multi step sampling and importance sampling: $\nabla_\theta J = \mathbb{E}_\pi[\sum_{k=0}^{K} \nabla_\theta \log \frac{p_\theta\left(x_{\tau_{k-1}}|x_{\tau_k}, s\right)}{p_{\theta_{old}}\left(x_{\tau_{k-1}}|x_{\tau_k}, s\right)} A_\pi]$

- $A_\pi = Q - V$ offline
- $A_\pi = GAE$ online

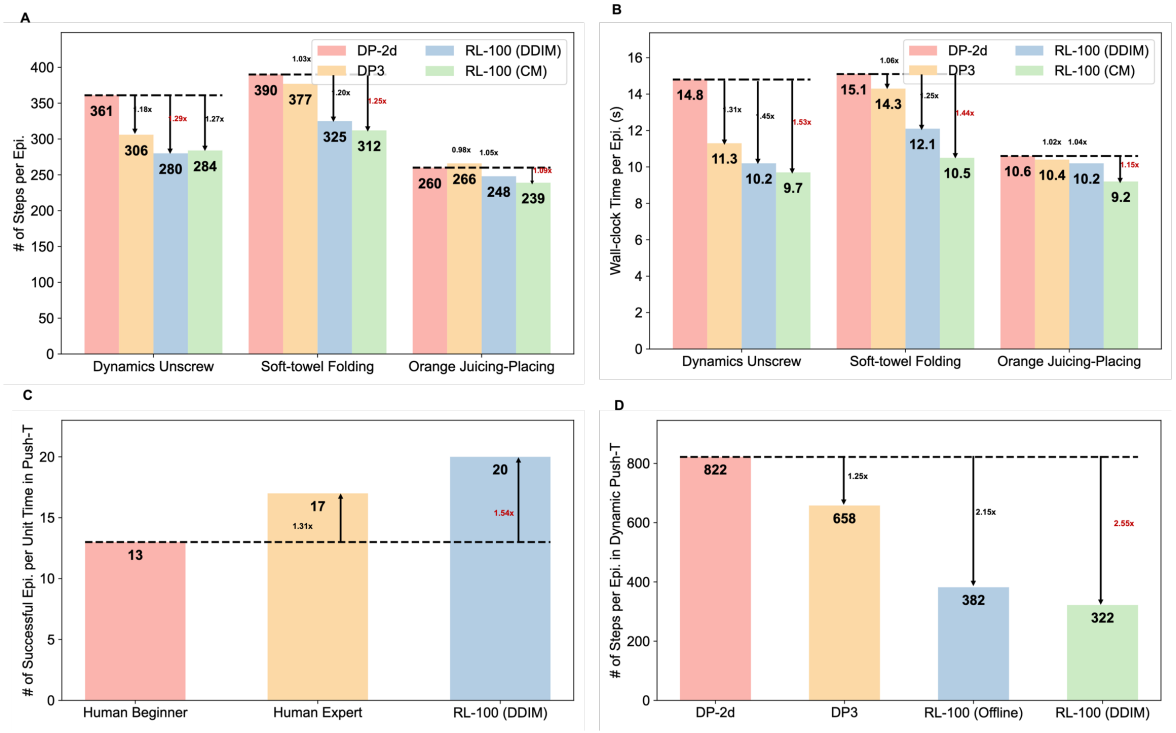
One-step consistency distillation: $\mathcal{L}_{\mathrm{CD}}(\theta) = \mathbb{E}_{x_0,\tau,\varepsilon}\left[\left\|C_\theta(x^\tau, \tau) - \mathrm{sg}\left[\Psi_\varphi(x^\tau, \tau \to 0)\right]\right\|_2^2\right]$

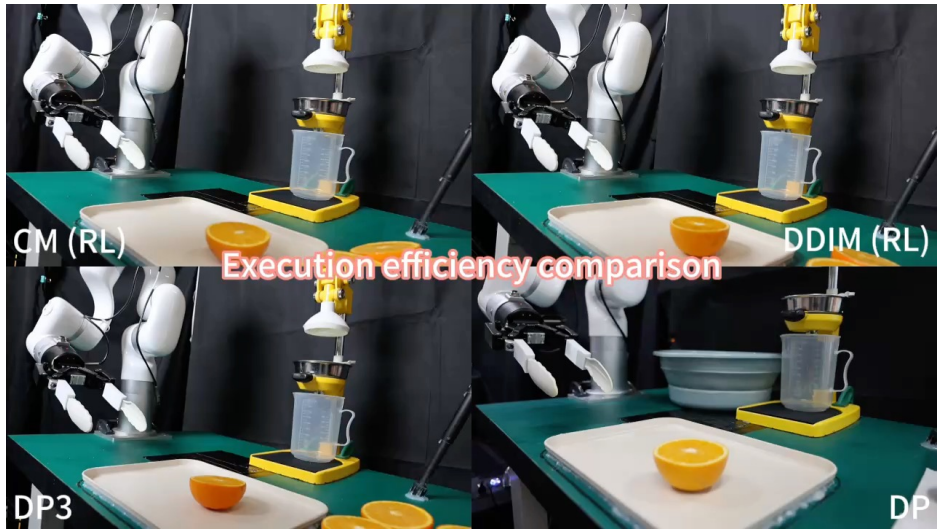
Overall finetune loss: $\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{RL}} + \lambda_{\mathrm{CD}} \cdot \mathcal{L}_{\mathrm{CD}}$

- Robustness, Zero-Shot & Few-shot Generalization



Execution efficiency

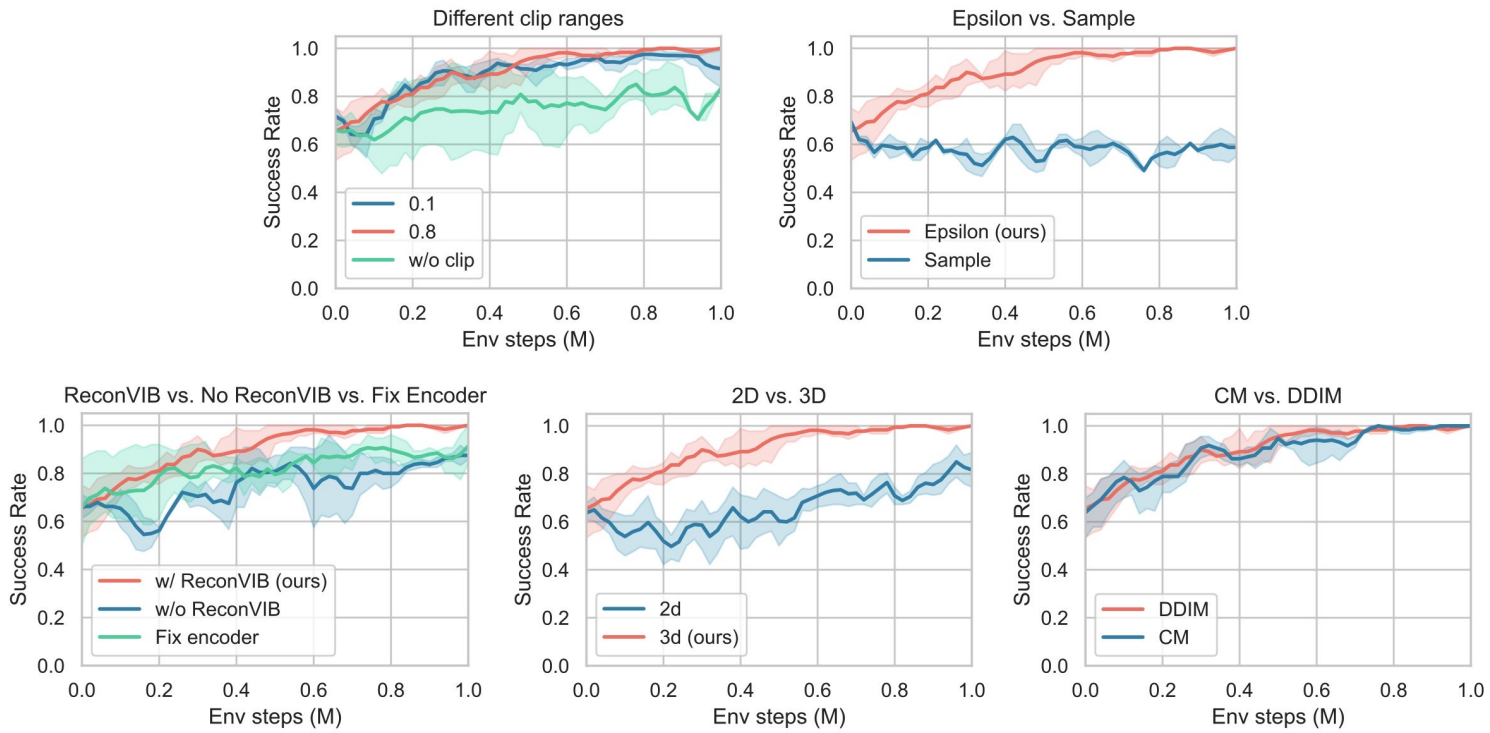




- Data usage

Task	Human Demonstration		Iterative Offline RL		Online RL	
	# of epi.	Collection time (h)	# of epi.	Collection time (h)	# of epi.	Collection time (h)
Dynamic Push-T	100	2	821	8	763	7.5
Agile Bowling	100	2	249	2	213	2.5
Pouring	64	1	741	6.8	129	1.5
Soft-towel Folding	400	5	896	11	654	8.5
Dynamic Unscrew	31	0.5	467	4.5	288	3
Orange Juicing – Placing	80	1.5	642	10.5	750	12.5
Orange Juicing – Removal	29	0.5	149	2.5	240	4
Average	**115**	**1.8**	**566**	**6.5**	**434**	**5.6**

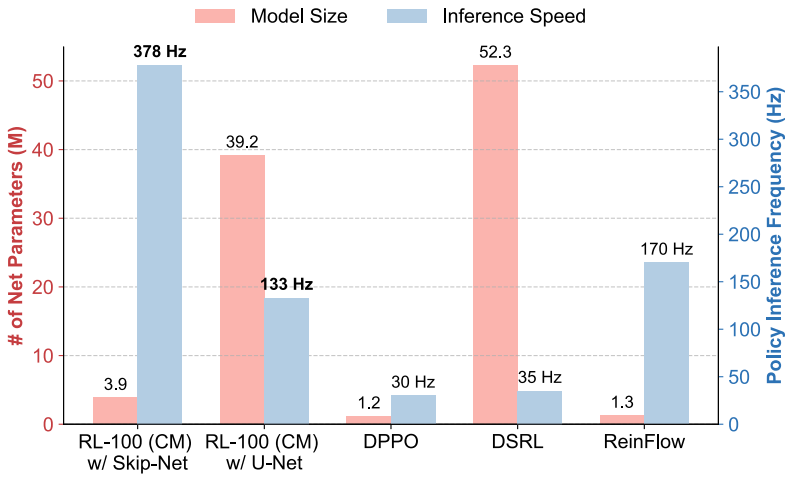
Ablation study

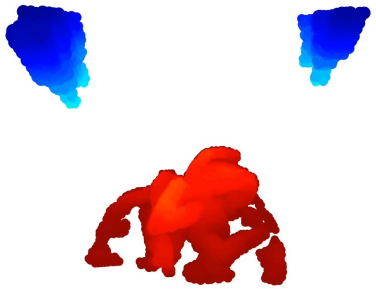


Takeaway:

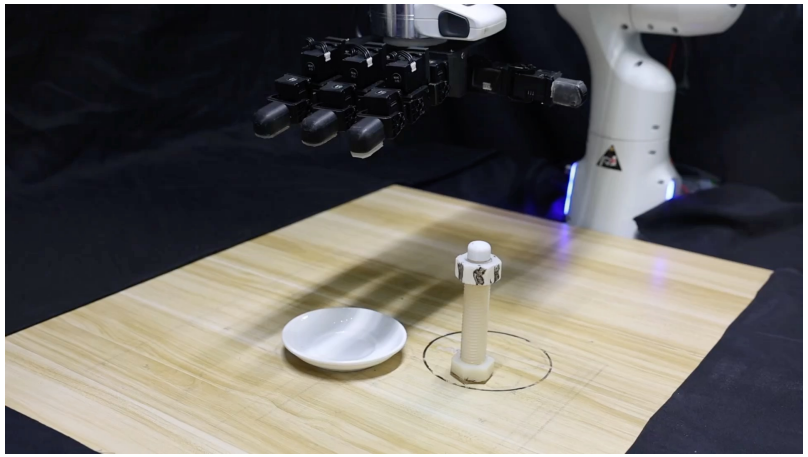
1) Variance clipping is valid for stable exploration - variance clipping in the stochastic DDIM sampling process.
2) Epsilon prediction is more suitable for RL: large noise schedule for exploration
3) Reconstruction is crucial for visual robotic manipulation RL as it mitigates representational drift and improves sample efficiency.
4) On a relatively clean scene, the 3D variant learns faster and attains a higher final success rate.
5) CM effectively compresses the iterative denoising process without sacrificing control quality, enabling high-frequency deployment.

Ablation study





Folding



Next move - Liberate productive forces: robot helps

Single task

Multi task:
The same series
More data-more robots

Understand humans‘ instructions

Thanks!

Project page

Wechat