Abstract

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

Contribution : Tsinghua TEA Lab

Backbone： diffusion-based

针对问题 : **纯粹的模仿学习继承了人类的偏见与低效**　**纯粹的真机强化学习很危险(HIL-SERL)**

追求目标 : 高效利用先验数据的同时，超越遥操作的水平

新思路 : 1.**模仿学习（遥操作数据）** 2.**迭代离线强化学习** 3.**On-policy的真机强化学习**

小创新点 : 1.**多模态输入** 简单切换编码器实现其余网络不变情况下同时支持3D点云和2D图像输入

2.**统一RL 和 diffusion 的策略梯度** 提升不同框架训练衔接的柔顺性

3.**压缩K步扩散策略为1步** 提升实时性

4.**提升任务泛化性** 对不同控制频率需求的任务提供单步和chunk两种模式

Preliminaries

三个重要（参数）：时间步t，噪声ε，数据x

DDPM

$$P(x_{t-1} \mid x_t) \sim N(\tilde{\mu}, \beta_t^2) \longrightarrow \tilde{\mu} = f(x_t, \widetilde{x_0}) \longrightarrow \widetilde{x_0} = f(x_t, \epsilon) \longrightarrow \epsilon_\theta = \epsilon$$

训练

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$\boxed{\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2}$
6: **until** converged

推理

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\boxed{\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}}$
5: **end for**
6: **return** $\mathbf{x}_0$

针对**DDPM采样慢**的问题，开发出了一种能“**跳步**”的采样方法

新的方法不遵循马尔可夫公式，但遵循x_t与x_0之间的公式

DDIM

启示

DDPM训练 + DDIM采样

Preliminaries

Consistency Policy(Ilya)

引入一个去噪教师，把去噪生成过程做成一步

关于以上3块的详细知识，推荐B站一个up主 Nik_Li

Diffusion-based RL

$$a^{\tau_{k-1}} = f_\theta(a^{\tau_k}, \tau_k \,|\, o)\,, \qquad k = K, \ldots, 2,$$

给定观察的conditional扩散生成

将去噪过程看作为一个sub-MDP

$$a_t := a^{\tau_0}$$

Initial state: $s^K = (a^{\tau_K}, \tau_K, o)$ with $a^{\tau_K} \sim \mathcal{N}(0, \mathbf{I})$.

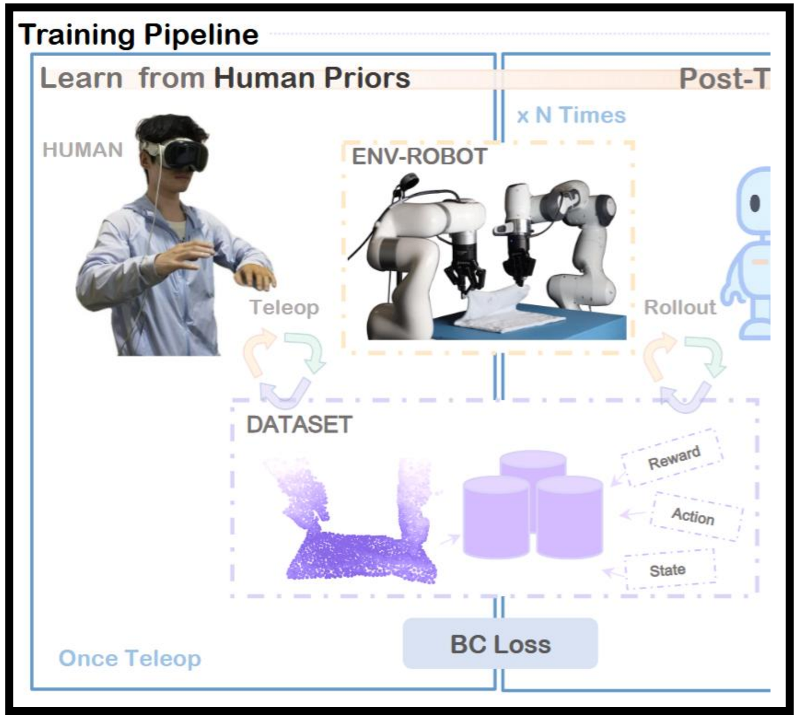
State: $s^k = (a^{\tau_k}, \tau_k, o)$, $k = K, \ldots, 1$.

Action: $u^k = a^{\tau_{k-1}}$ drawn from the denoising sub-policy $\pi_\theta(u^k \mid s^k) = \mathcal{N}(\mu_\theta(a^{\tau_k}, \tau_k, o),\ \sigma^2_{\tau_k}\mathbf{I})$.

Transition: $s^{k-1} = (u^k, \tau_{k-1}, o)$.

Reward: this sub-MDP only receives terminal reward $R(a^{\tau_0})$ from the upper environment MDP.

Imitation Learning



Backbone： conditional diffusion

数据来源： 人类遥操作数据 $\{o_t, q_t, a_t\}_{t=1}^{T_e}$

训练过程

$$c_t = [\phi(o_i, q_i)]_{i=t-n_o+1}^{t}$$

t时刻去噪目标

$$a_t^{\tau_0} = u_t \in \mathbb{R}^{d_a}$$

$$a_t^{\tau_0} = [u_t, \dots, u_{t+n_c-1}] \in \mathbb{R}^{n_c d_a}$$

扩散模型基本公式

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{2a}$$

$$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s. \tag{2b}$$

$$\mathcal{L}_{\mathrm{IL}}(\theta) = \mathbb{E}_{(a^{\tau_0}, c_t)\sim\mathcal{D},\, \tau,\, \varepsilon}\left[\, \|\varepsilon - \varepsilon_\theta(a^\tau, \tau, c_t)\|_2^2 \,\right]$$

本文输出的是关节角增量

Delta空间

视觉与感知编码器

RGB输入：ViT
Point Clouds：DP3

$$\mathcal{L}_{\mathrm{recon}} = \beta_{\mathrm{recon}}\left(d_{\mathrm{Chamfer}}(\hat{o}, o) + \|\hat{q} - q\|_2^2\right)$$ 对齐空间

$$\mathcal{L}_{\mathrm{KL}} = \beta_{\mathrm{KL}}\, \mathrm{KL}\left(\phi(z|o, s)\, \|\, \mathcal{N}(0, I)\right)$$ probabilistic encoder

两个损失(RL微调时减小权重)

$$\hat{a}_t^{\tau_0} \leftarrow \mathrm{DDIM}_K(\varepsilon_\theta(\cdot, \cdot, c_t))$$

推理过程DDIM采样

Unified RL Fine-tuning

Offline RL（过程中视觉编码器冻结）

Environment MDP Iteration i

Denoising MDP timestep t

先用IQL在数据集上学习Critics

加和去噪每一步的PPO目标，作为迭代i的PPO目标

$$J_i(\pi) = \mathbb{E}_{s_t \sim \rho_\pi,\, a_t \sim \pi_i}\left[\sum_{k=1}^{K} \min\left(\boxed{r_k(\pi)}\, A_t,\ \mathrm{clip}\left(r_k(\pi),\, 1-\epsilon,\, 1+\epsilon\right) A_t\right)\right],$$

$$r_k^{\mathrm{off}}(\pi) = \frac{\pi(a^{\tau_{k-1}} \mid s^k)}{\pi_i(a^{\tau_{k-1}} \mid s^k)}$$

OPE门控 (offline policy evaluation)

$$\widehat{J}^{\text{AM-Q}}(\pi) = \mathbb{E}_{(s,a)\sim(\hat{T},\pi)}\left[\sum_{t=0}^{H-1} Q_\psi(s_t, a_t)\right]$$

$$\widehat{J}^{\text{AM-Q}}(\pi) - \widehat{J}^{\text{AM-Q}}(\pi_i) \geq \delta$$

$$\text{set} \quad \delta = 0.05 \cdot |\widehat{J}^{\text{AM-Q}}(\pi_i)|$$

评估是否更新策略的门控机制

先用一个T去学后续可能序列，求序列的价值，若大，则更新

Online RL 整体与离线一致

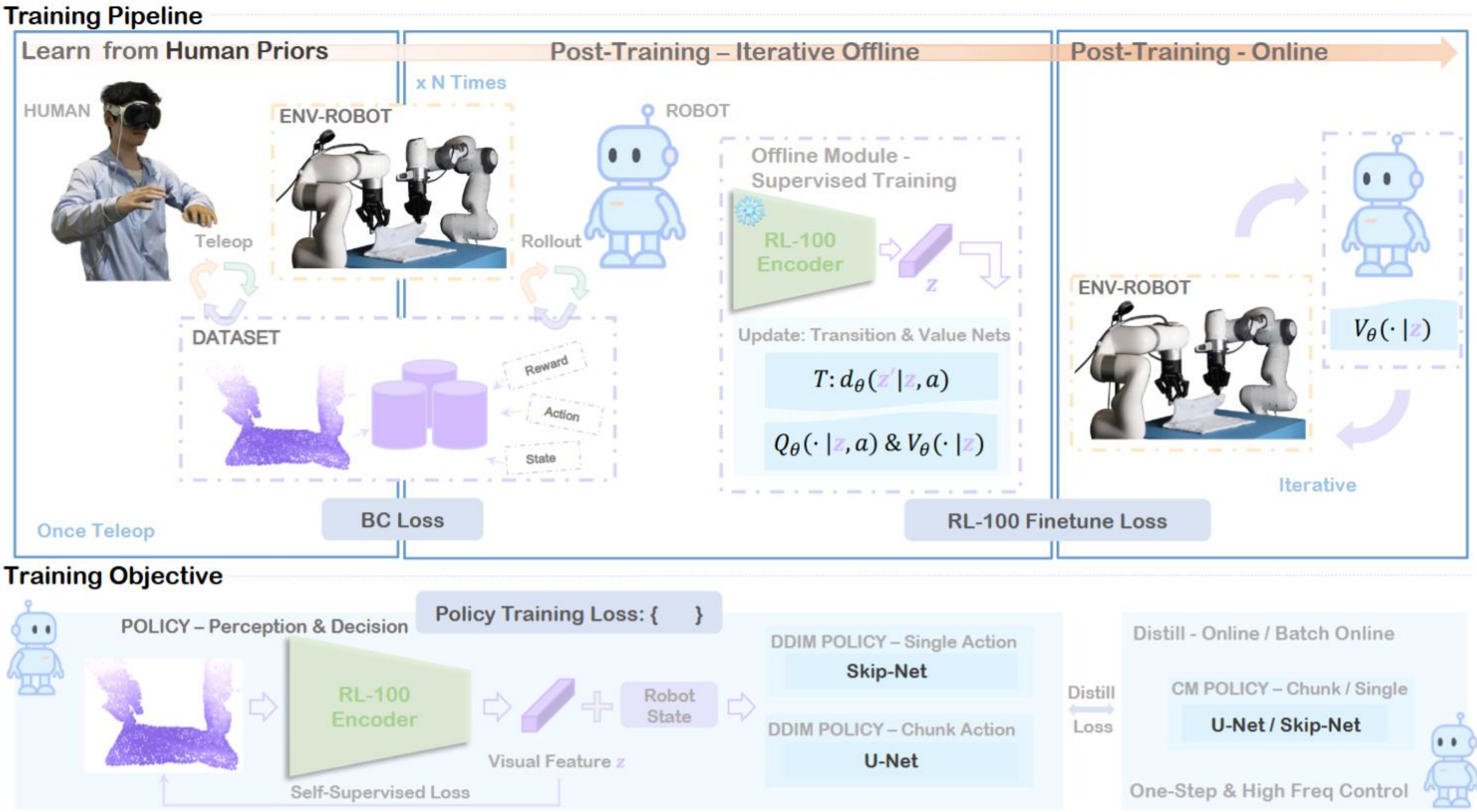
$$A_t^{\text{on}} = \mathrm{GAE}(\lambda, \gamma; r_t, V_\psi)$$

优势函数用GAE形式

$$\mathcal{L}_{\text{RL}}^{\text{on}} = -J_i(\pi) + \lambda_V\, \mathbb{E}\left[\left(V_\psi(s_t) - \hat{V}_t\right)^2\right]$$

在损失函数中加入Critic

流程图

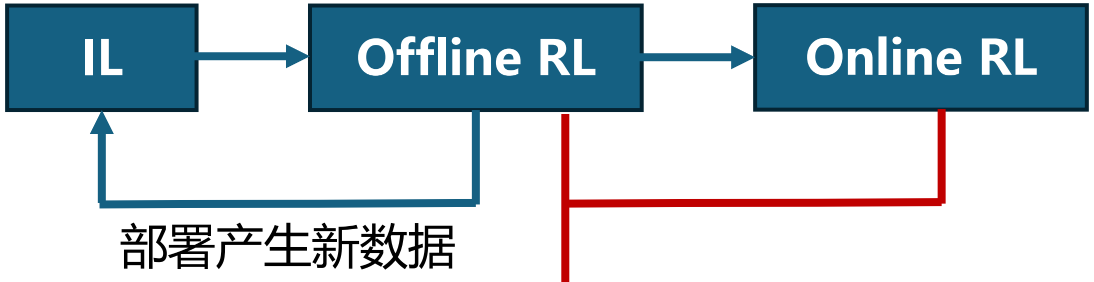

Training Pipeline
Learn from Human Priors
Post-Training – Iterative Offline
Post-Training - Online
x N Times
HUMAN
ENV-ROBOT
ROBOT
Offline Module - Supervised Training
RL-100 Encoder
z
Teleop
Rollout
DATASET
Reward
Action
State
Update: Transition & Value Nets
$T: d_\theta(z'|z,a)$
$Q_\theta(\cdot|z,a)$ & $V_\theta(\cdot|z)$
ENV-ROBOT
$V_\theta(\cdot|z)$
Iterative
Once Teleop
BC Loss
RL-100 Finetune Loss
Training Objective
POLICY – Perception & Decision
Policy Training Loss: { }
RL-100 Encoder
Robot State
Visual Feature z
Self-Supervised Loss
DDIM POLICY – Single Action
Skip-Net
DDIM POLICY – Chunk Action
U-Net
Distill
Loss
Distill - Online / Batch Online
CM POLICY – Chunk / Single
U-Net / Skip-Net
One-Step & High Freq Control

Pseudocode

Algorithm 1 RL-100 training pipeline

Input: Demonstrations $\mathcal{D}_0$, iterations M
Initialize: $\pi_0^{\text{IL}} \leftarrow \text{ImitationLearning}(\mathcal{D}_0)$
for iteration $m = 0$ to $M - 1$ **do**
  // Offline RL improvement
  Train critics: $(Q_{\psi_m}, V_{\psi_m}) \leftarrow \text{IQL}(\mathcal{D}_m)$
  Train transition: $T_{\theta_m}(s'|s, a)$
  Optimize:
    $\pi_m^{\text{ddim}}, \pi_m^{\text{cm}} \leftarrow \text{OfflineRL}(\pi_m^{\text{IL}}, Q_{\psi_m}, V_{\psi_m}, T_{\theta_m})$
  // Data expansion
  Deploy: $\mathcal{D}_{\text{new}} \leftarrow \text{Rollout}(\pi_m^{\text{ddim}}$ or $\pi_m^{\text{cm}})$
  Merge: $\mathcal{D}_{m+1} \leftarrow \mathcal{D}_m \cup \mathcal{D}_{\text{new}}$
  // IL re-training on expanded data
  $\pi_{m+1}^{\text{IL}} \leftarrow \text{ImitationLearning}(\mathcal{D}_{m+1})$
end for
// Final online fine-tuning
$\pi_{\text{ddim}}^{\text{final}}, \pi_{\text{cm}}^{\text{final}} \leftarrow \text{OnlineRL}(\pi_{M-1}, V_{\psi_{M-1}})$
Output: $\pi_{\text{ddim}}^{\text{final}}, \pi_{\text{cm}}^{\text{final}}$

正循环过程



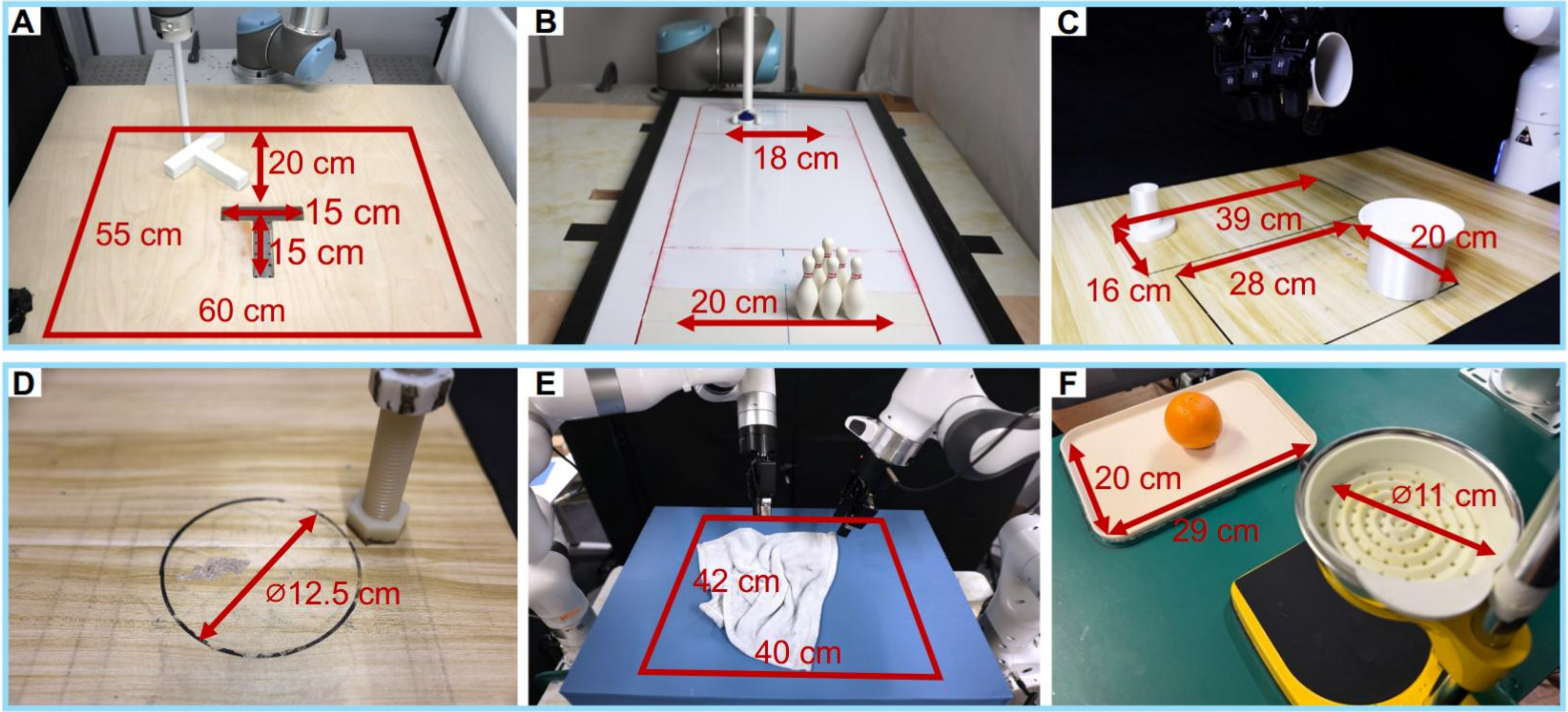
Insight：不断提升数据质量，自监督学习

方差限制(variance clipping)

$$\tilde{\sigma}_k = \text{clip}(\sigma_k, \sigma_{\min}, \sigma_{\max})$$

主要保证Offline RL稳定，也兼具一点探索性

Experiment


A
20 cm
15 cm
55 cm
15 cm
60 cm
B
18 cm
20 cm
C
39 cm
20 cm
16 cm
28 cm
D
ø12.5 cm
E
42 cm
40 cm
F
20 cm
ø11 cm
29 cm

Success rate

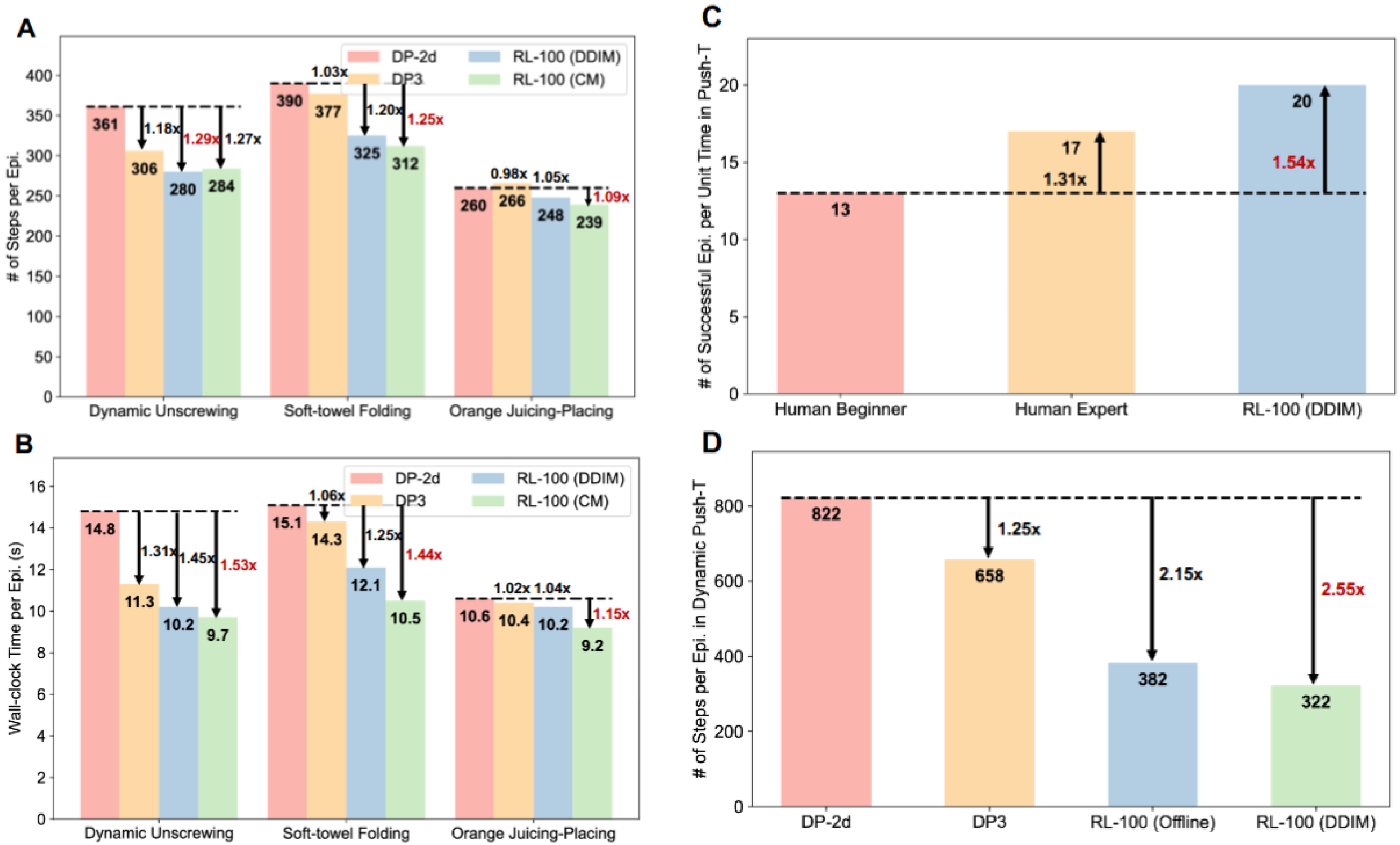
Task	Imitation baselines		RL-100 (ours)		
	DP-2D	DP3	Iterative Offline RL	Online RL (DDIM)	Online RL (CM)
Dynamic Push-T	40 (20/50)	64 (32/50)	90 (45/50)	100 (50/50)	100 (50/50)
Agile Bowling	14 (7/50)	80 (40/50)	88 (44/50)	100 (50/50)	100 (50/50)
Pouring	42 (21/50)	48 (24/50)	92 (46/50)	100 (50/50)	100 (50/50)
Soft-towel Folding	46 (23/50)	68 (34/50)	94 (47/50)	100 (50/50)	100 (250/250)
Dynamic Unscrewing	82 (41/50)	70 (35/50)	94 (47/50)	100 (50/50)	100 (50/50)
Orange Juicing – Placing	78 (39/50)	88 (44/50)	94 (47/50)	100 (100/100)	100 (50/50)
Orange Juicing – Removal	48 (24/50)	76 (38/50)	86 (43/50)	100 (50/50)	—
Mean (unweighted)	**50.0**	**70.6**	**91.1**	**100.0**	**100.0**†

Task Variation	Zero-shot Success Rate (%)
Pouring (Water)	90
Push-T (Changed surface)	100
Push-T (Interference Objects)	80
Bowling (Changed Surface)	100
Folding (unseen shape)	80
Average	**90.0**

Task & Disturbance Stage	Success Rate (%)
Folding (Stage 1: Grasping)	90
Folding (Stage 2: Pre-folding)	90
Unscrewing	100
Push-T (Whole stage)	100
Average	**95.0**

外界干扰

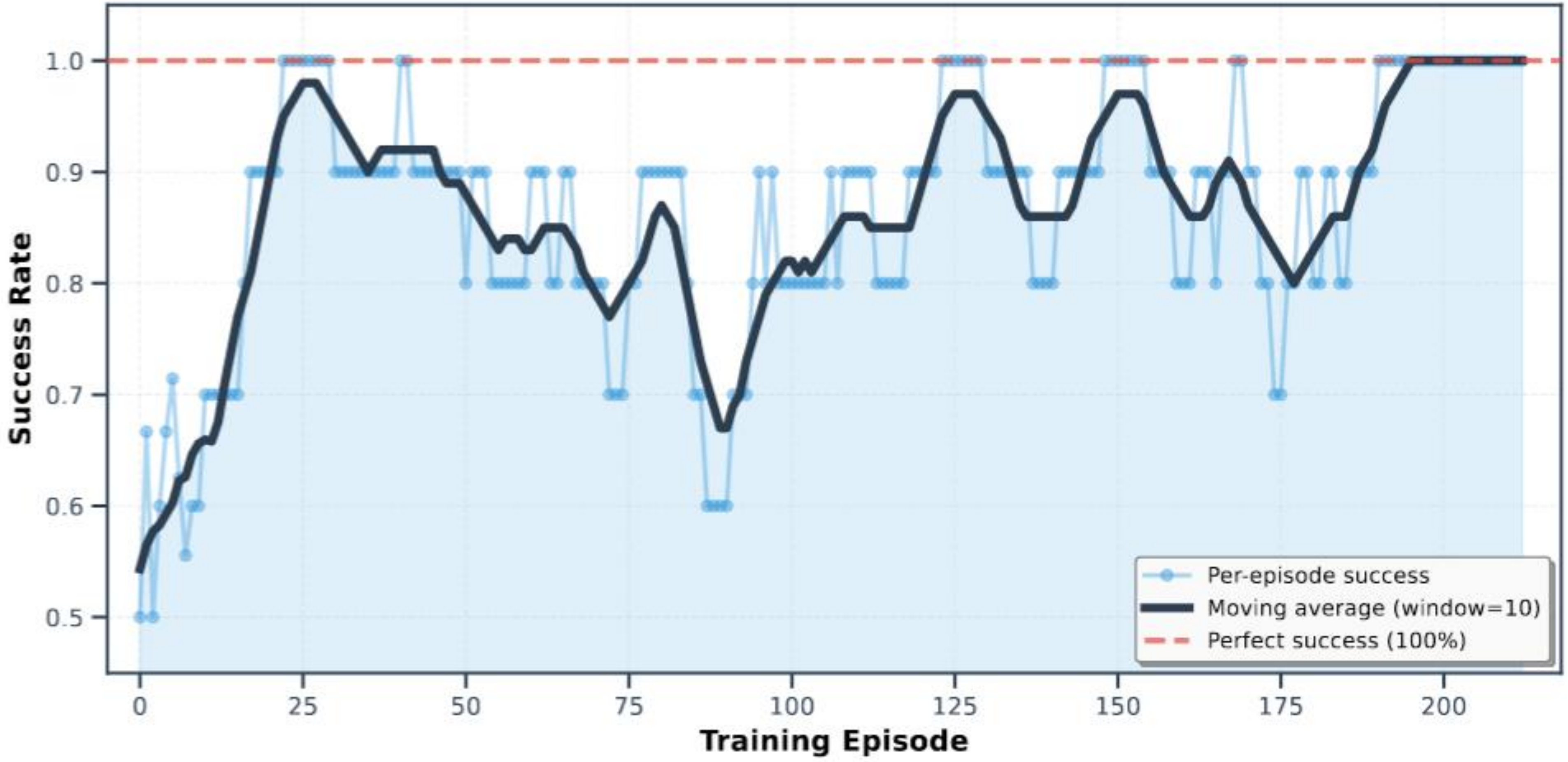
Efficiency



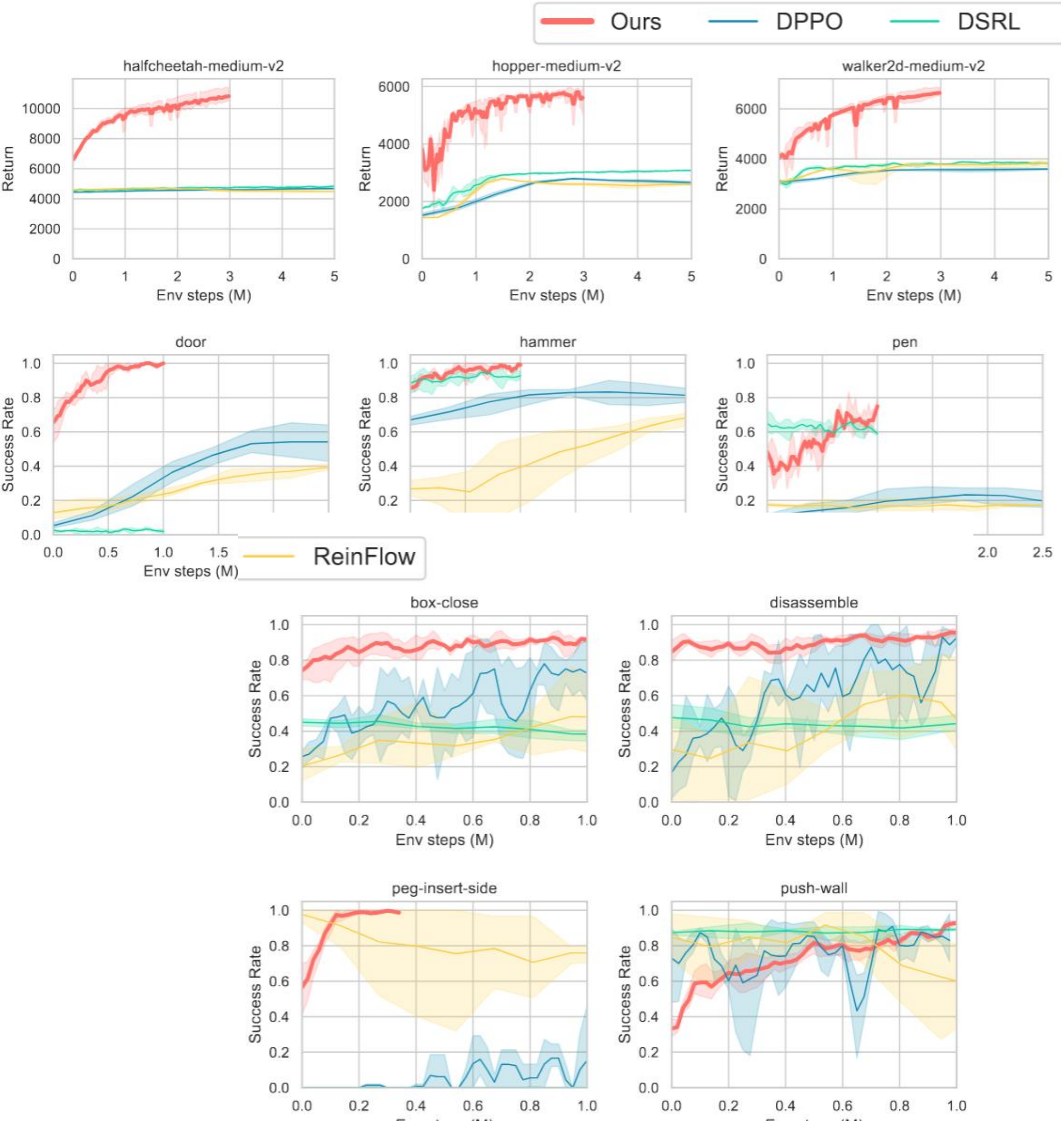
A
B
C
D
DP-2d
DP3
RL-100 (DDIM)
RL-100 (CM)
# of Steps per Epi.
Dynamic Unscrewing
Soft-towel Folding
Orange Juicing-Placing
361
306
280
284
1.18x
1.29x
1.27x
390
377
325
312
1.03x
1.20x
1.25x
260
266
248
239
0.98x
1.05x
1.09x
Wall-clock Time per Epi. (s)
14.8
11.3
10.2
9.7
1.31x
1.45x
1.53x
15.1
14.3
12.1
10.5
1.06x
1.25x
1.44x
10.6
10.4
10.2
9.2
1.02x
1.04x
1.15x
# of Successful Epi. per Unit Time in Push-T
Human Beginner
Human Expert
RL-100 (DDIM)
13
17
20
1.31x
1.54x
# of Steps per Epi. in Dynamic Push-T
DP-2d
DP3
RL-100 (Offline)
RL-100 (DDIM)
822
658
382
322
1.25x
2.15x
2.55x

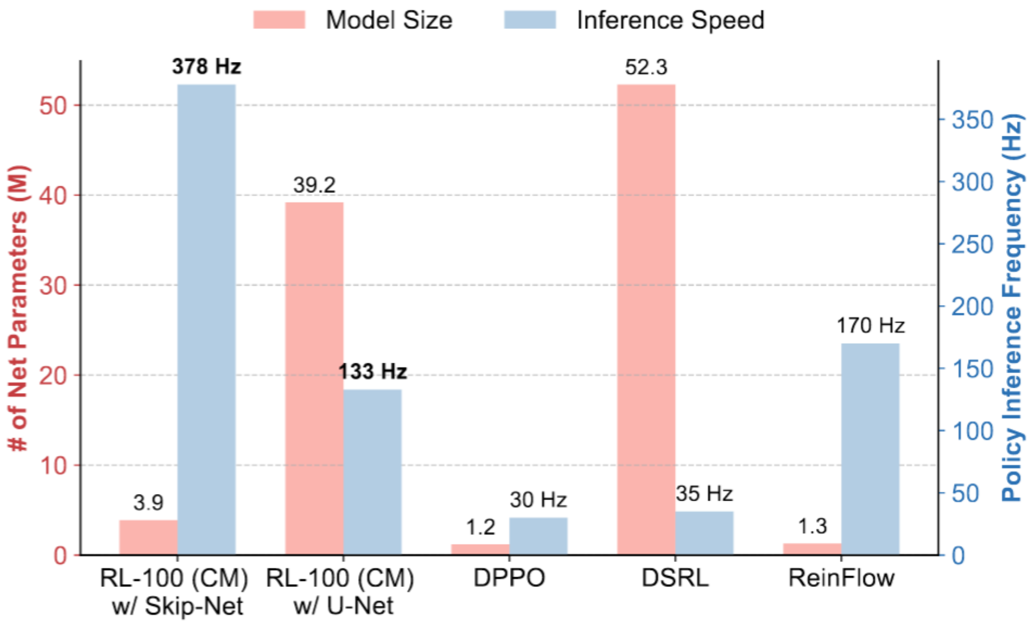
Training Efficiency

在线微调

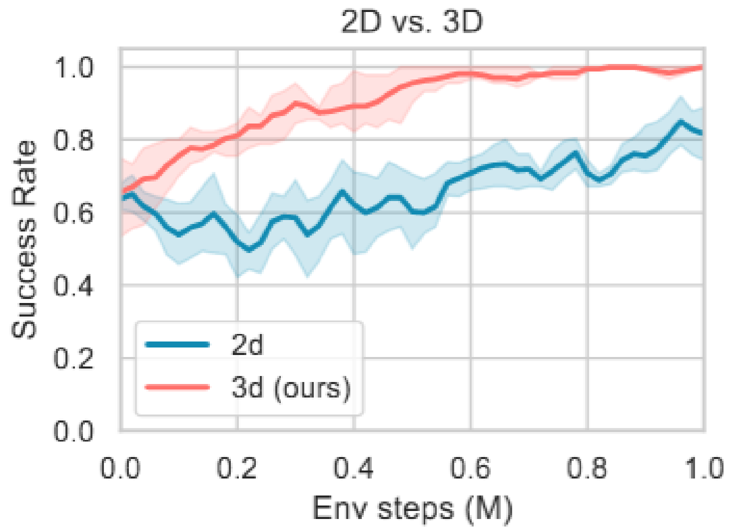


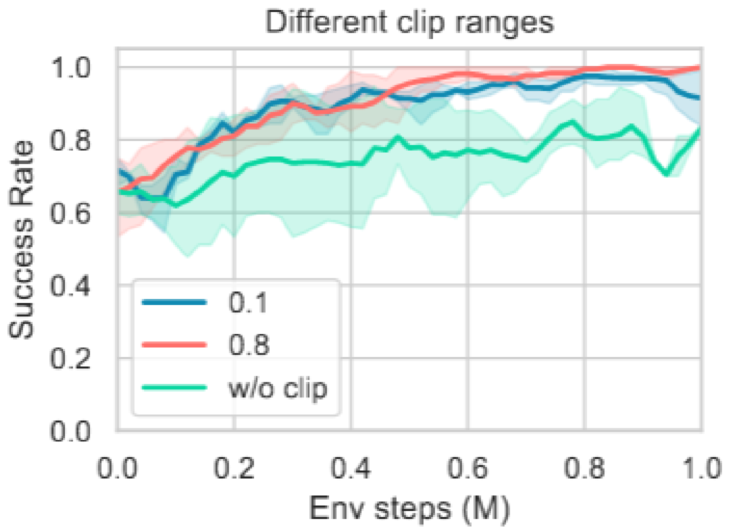
Simulation Experiment





Ablation





$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\, f_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \qquad (\epsilon\text{–prediction}),$$

$$\hat{x}_0 = f_\theta(x_t, t) \qquad (x_0\text{–prediction}).$$