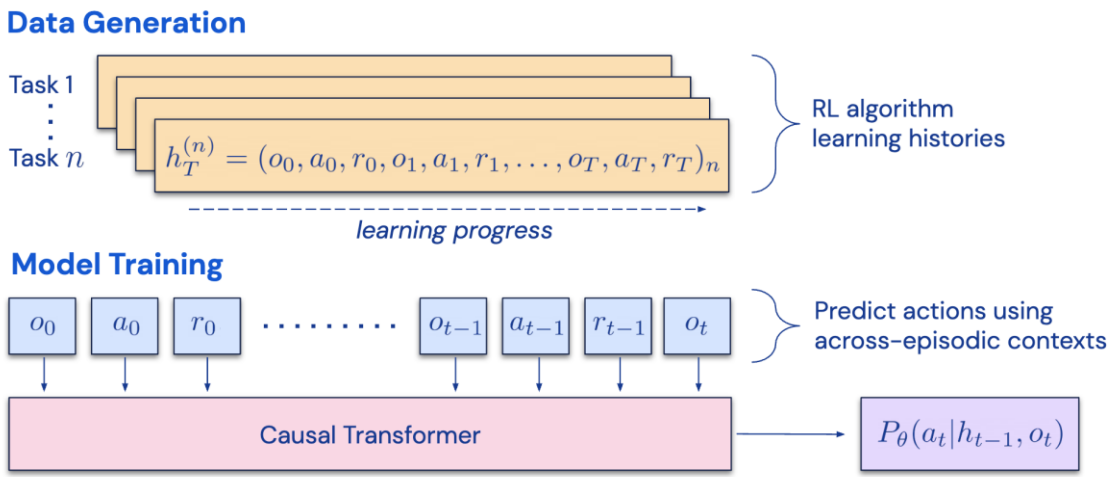
Distilling Reinforcement Learning Algorithms for In-Context Model-Based Planning

Jaehyeon Son, Soochan Lee, Gunhee Kim

SEOUL NATIONAL UNIV.
VISION & LEARNING

Prior Works on In-Context RL

- Prior works collect *learning histories* of specific RL algorithm.
- Then, they feed them to Transformer to model policy improvement process.
- Transformer mimics the *exploration-exploitation behavior* of the source algorithm.



<Algorithm Distillation> Laskin et al., 2023

Limitations of Previous Approaches

- Previous approaches inherit suboptimal behaviors of source algorithms.
- RL algorithms deliberately prevent abrupt changes.

$$\underset{\theta}{\text{maximize}}\, \mathbb{E}_{s\sim\rho_{\theta_{\text{old}}},a\sim q}\left[\frac{\pi_\theta(a|s)}{q(a|s)}Q_{\theta_{\text{old}}}(s,a)\right].$$
$$\text{subject to}\ \mathbb{E}_{s\sim\rho_{\theta_{\text{old}}}}\left[D_{\text{KL}}\left(\pi_{\theta_{\text{old}}}(\cdot|s)\|\pi_\theta(\cdot|s)\right)\right]\leq\delta.$$

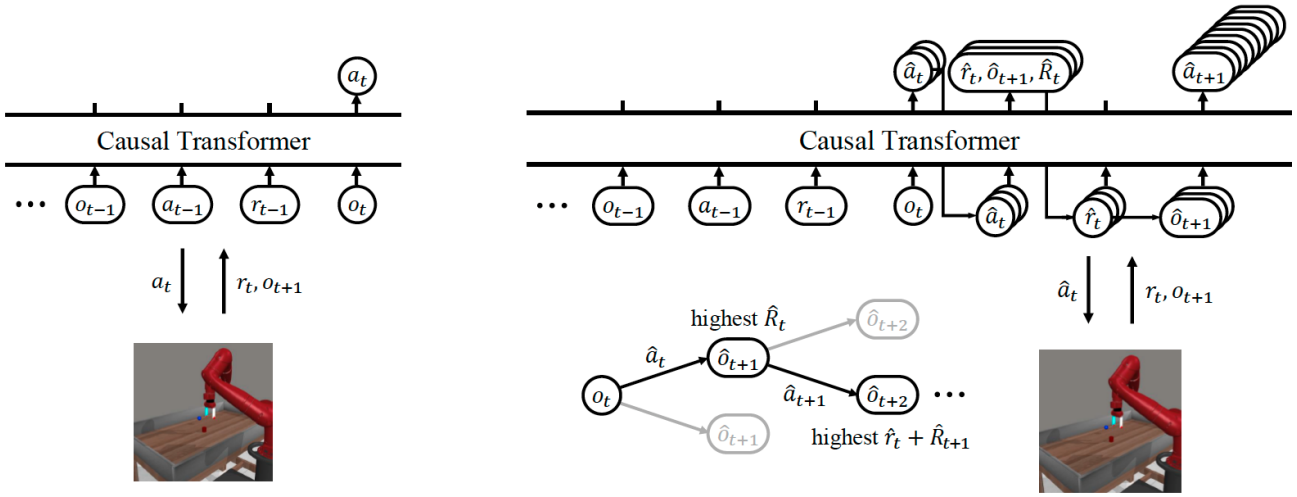
<TRPO> Schulman et al., 2015

$$r_t(\theta)=\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$$
$$L^{\text{CLIP}}(\theta)=\hat{\mathbb{E}}_t\left[\min\left(r_t(\theta)\hat{A}_t,\text{clip}(r_t(\theta),1-\epsilon,1+\epsilon)\hat{A}_t\right)\right],$$
$$L_t^{\text{CLIP+VF+S}}(\theta)=\hat{\mathbb{E}}_t\left[L_t^{\text{CLIP}}(\theta)-c_1L_t^{\text{VF}}(\theta)+c_2S\left[\pi_\theta\right](s_t)\right],$$

<PPO> Schulman et al., 2017

Distillation for In-Context Model-Based Planning (DICP)

- DICP predict not only actions, but also the outcome of the actions:
  - rewards, next observations, and return-to-go.
- Using this world model, DICP simulates the future before taking actions, without interactions.
- DICP constructs world model solely in-context even in novel tasks.

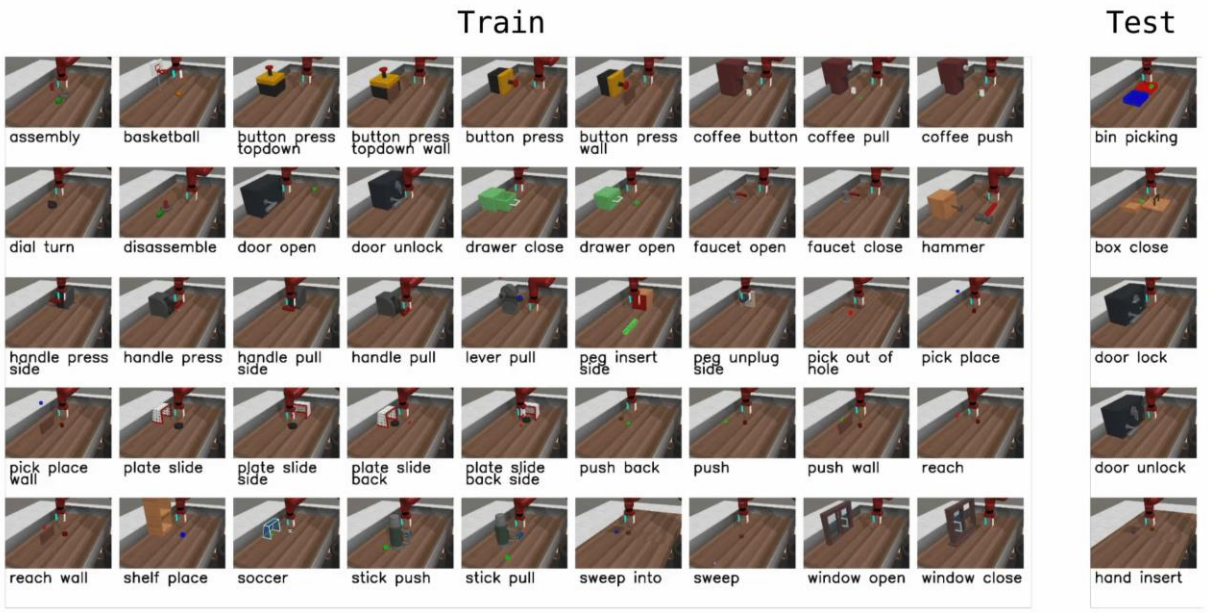


(a) Previous approaches

(b) Distillation for In-Context Planning (Ours)

Experiments: Benchmarks

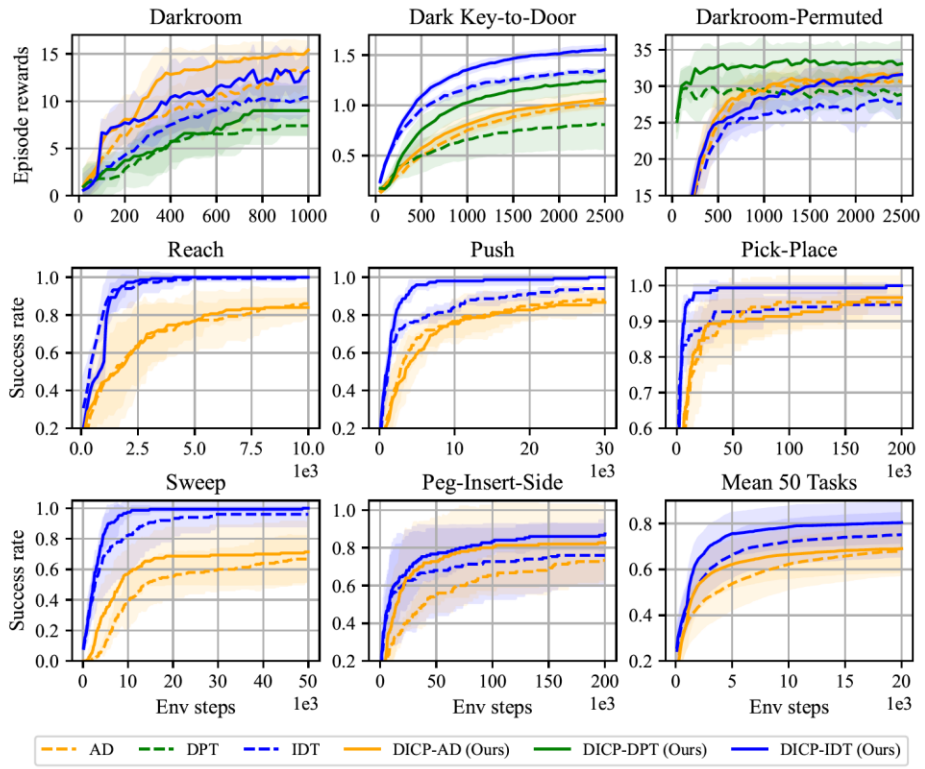
- Darkroom variants
  - Darkroom, Dark Key-to-Door, Darkroom-Permuted
- Meta-World benchmark suite



<Meta-World> Yu et al., 2019

Yu et al., *Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning*, CoRL 2019

Experiments: Learning Curves & Main Table



Method	Reach	Push	Pick-Place	Sweep	Peg-Insert-Side	Max Steps
RL^{2*}	**100**	96	98	–	–	300M
MAML*	**100**	94	80	–	–	300M
PEARL*	68	44	28	–	–	300M
MACAW†	–	–	–	4	0	5K
FOCAL†	–	–	–	38	10	5K
MuZero	**100**	**100**	**100**	–	–	10M
MoSS	86	**100**	**100**	–	–	40M
BoREL†	–	–	–	0	0	5K
IDAQ†	–	–	–	59	30	5K
AD	86	88	96	67	73	200K
IDT	**100**	94	95	96	76	200K
DICP-AD (Ours)	84	87	97	71	83	200K
DICP-IDT (Ours)	**100**	**100**	**100**	**100**	**87**	200K

Conclusion

- We introduced an in-context model-based RL framework.
- Our approach effectively addresses the limitations of previous approaches to in-context RL.
- Our approach demonstrated superior performance across various environments.