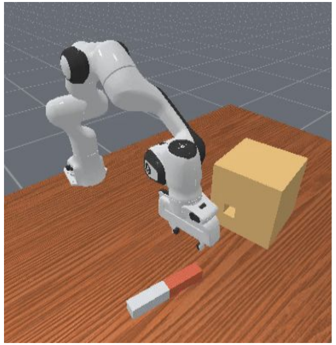
UC San Diego
SU Lab

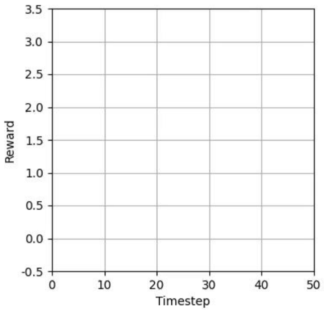
ICML
International Conference
On Machine Learning
2025

DEMO³

Multi-Stage Manipulation with Demonstration-Augmented Reward, Policy, and World Model Learning

Adrià López Escoriza, Nicklas Hansen, Stone Tao, Tongzhou Mu, Hao Su



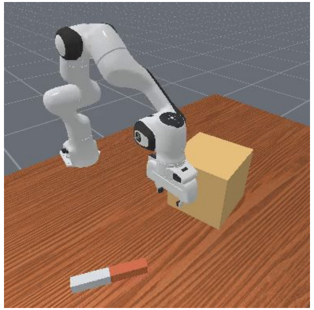


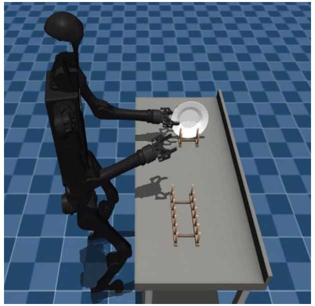
Long-horizon manipulation

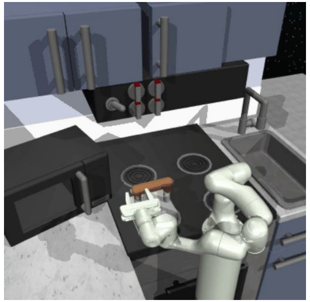
Hard problem for RL

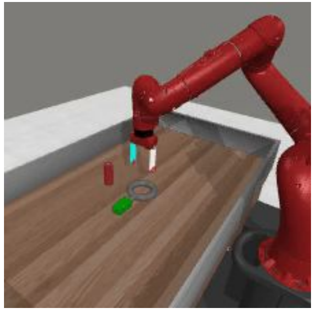
- **High precision** tasks
- **Complex reward** design
- **Large exploration** space

RL becomes too **inefficient**







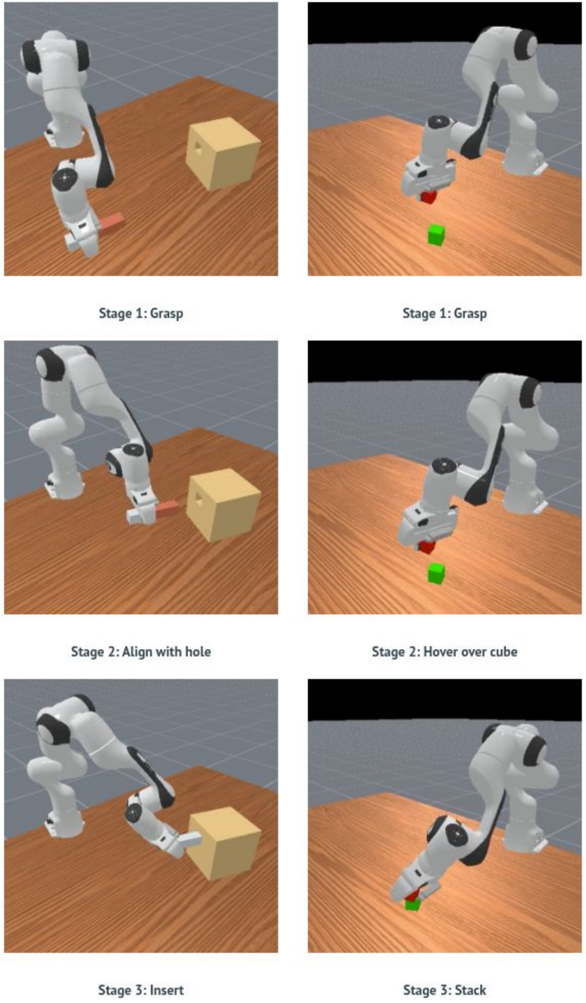


Multi-stage feedback

Long-horizon tasks have a **multi-stage structure**

Key idea:
Stage feedback + **demos** can guide learning

DEMO3:
Sample efficient RL for **multi-stage manipulation**



How do we achieve this?

Demonstration to simultaneously learn:

1. **Online reward function**
2. **Policy pre-training**
3. **World Model for planning**

How do we achieve this?

From images observations!

Demonstration to simu[illegible]

1. Online reward function

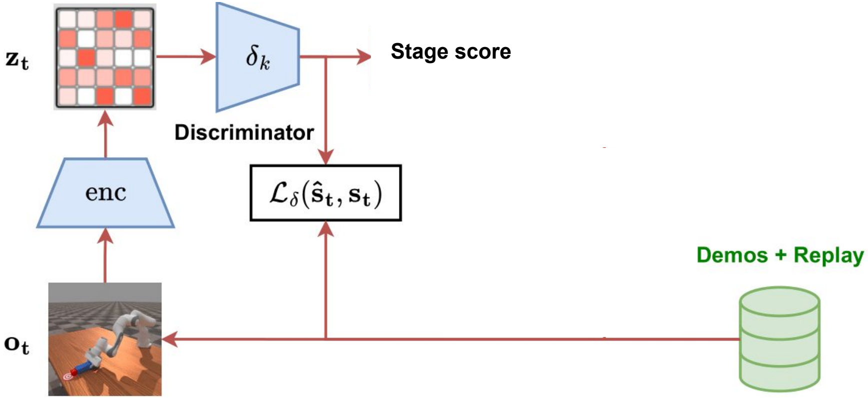
2. Policy pre-training

3. World Model for planning

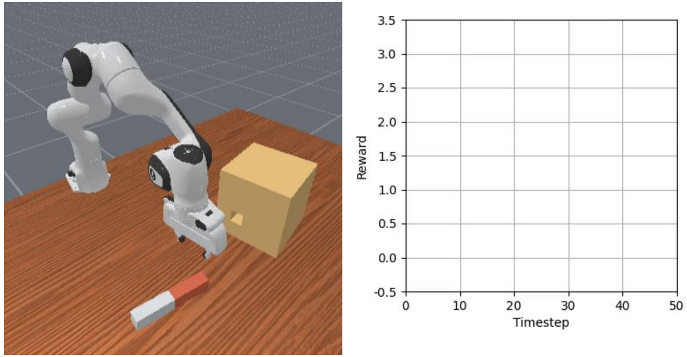
Online reward learning

Use demos for **online reward learning**:

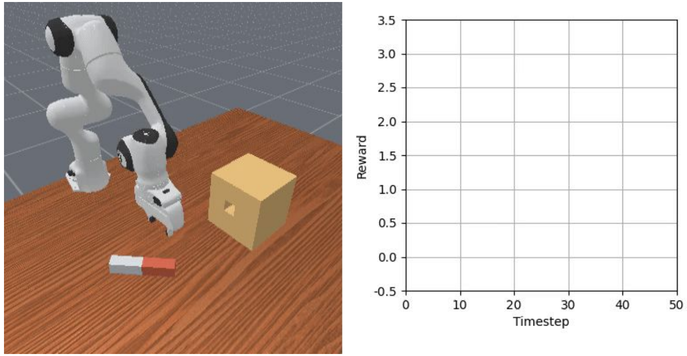
- Dataset: Demos + Replays
- Discriminator separates success / fail frames
- Policy trained on learned dense reward



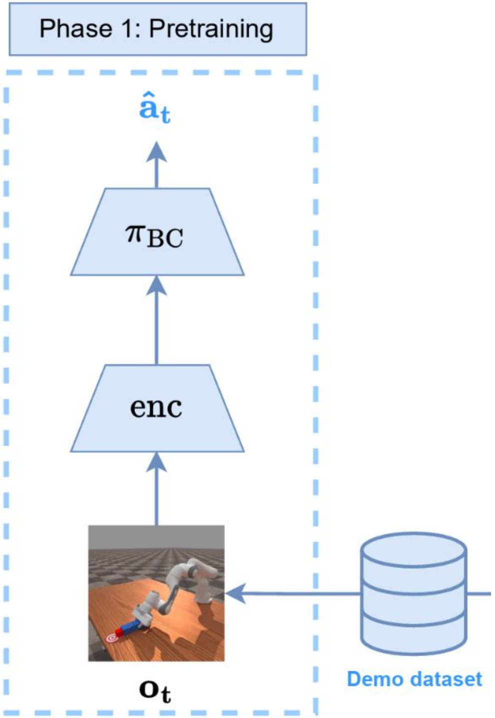
Successful rollout



Failed rollout



Policy pre-training

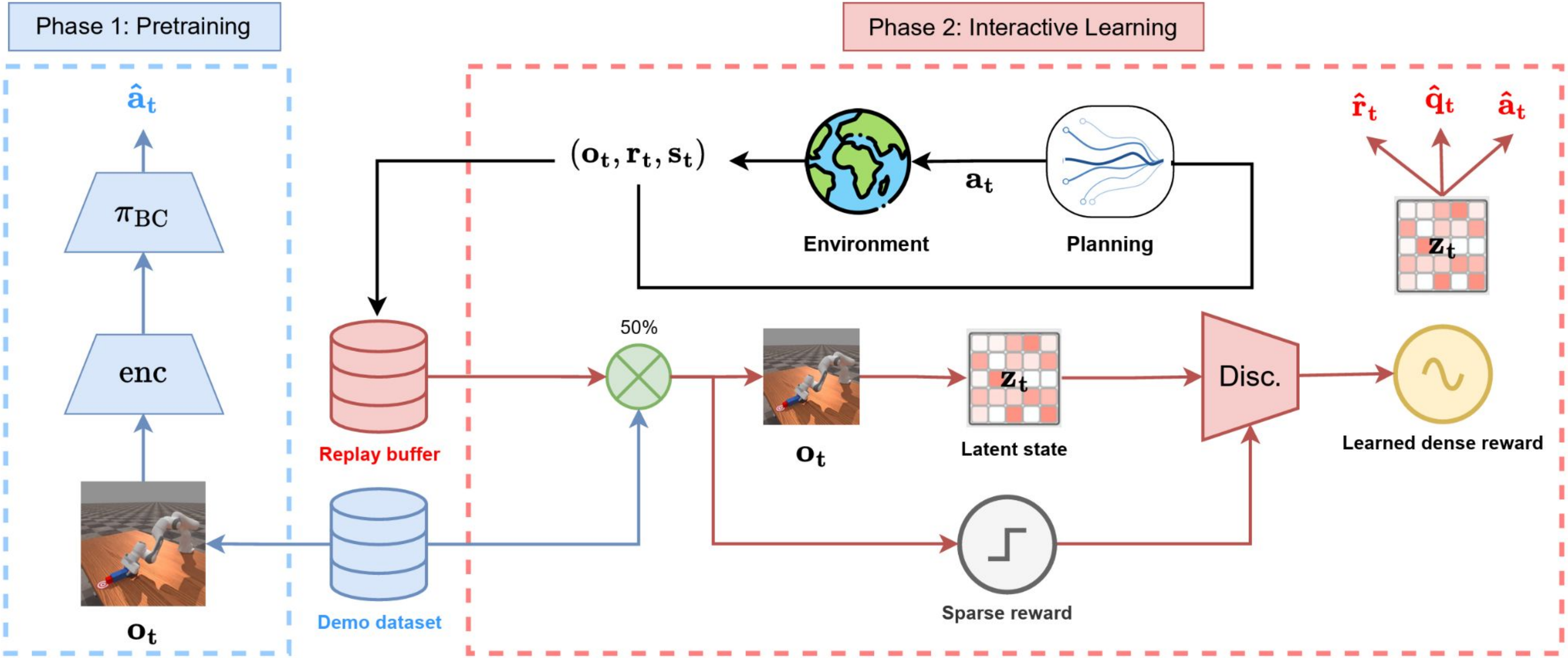


Behavioral cloning warm-starts **policy** and **encoder**

We sample from BC policy in **early stages**

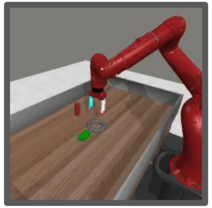
Progressively transition to RL policy

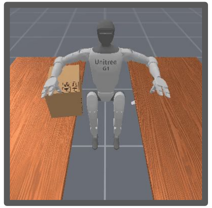
World Model Learning

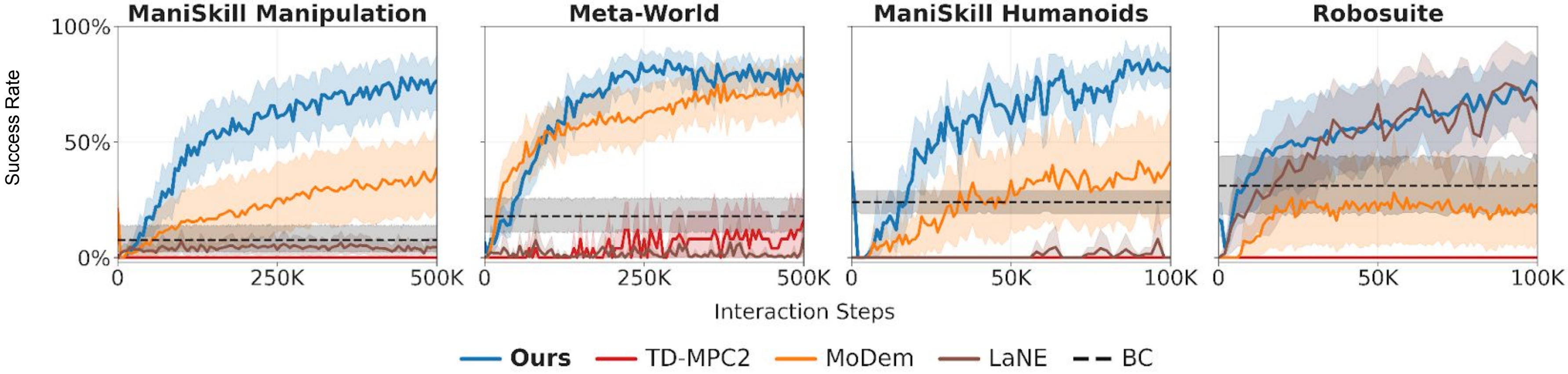


DEMO^3: Multi-Stage Manipulation with Demonstration-Augmented **Reward**, **Policy** and **World-Model** Learning

Results: Sample Efficiency





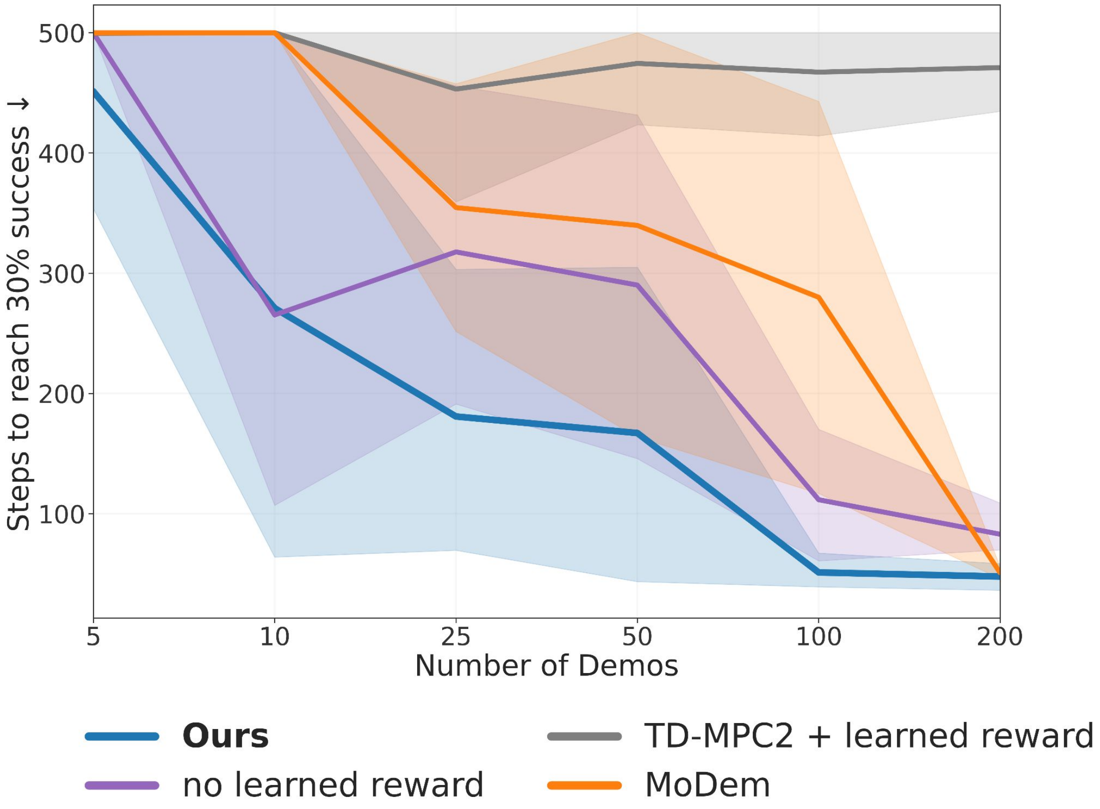


We solve hard tasks in < 100K Steps from **only images and sparse rewards**

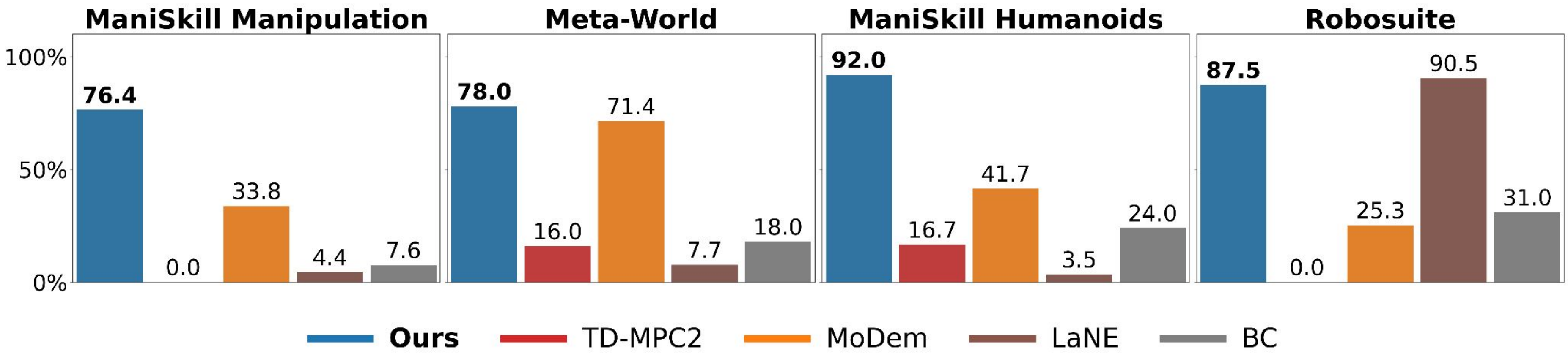
Results: Demonstration Efficiency

DEMO3 excels at **demonstration efficiency**

Hard tasks solved with **only 5 demos**



Thank you!


ManiSkill Manipulation
Meta-World
ManiSkill Humanoids
Robosuite
100%
50%
0%
76.4
0.0
33.8
4.4
7.6
78.0
16.0
71.4
7.7
18.0
92.0
16.7
41.7
3.5
24.0
87.5
0.0
25.3
90.5
31.0
Ours
TD-MPC2
MoDem
LaNE
BC