Sim-to-Real Contact-Rich Pivoting via Optimization-Guided RL with Vision and Touch
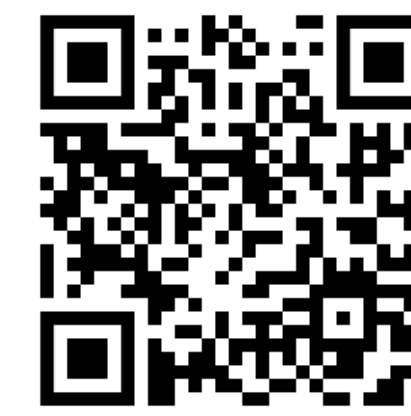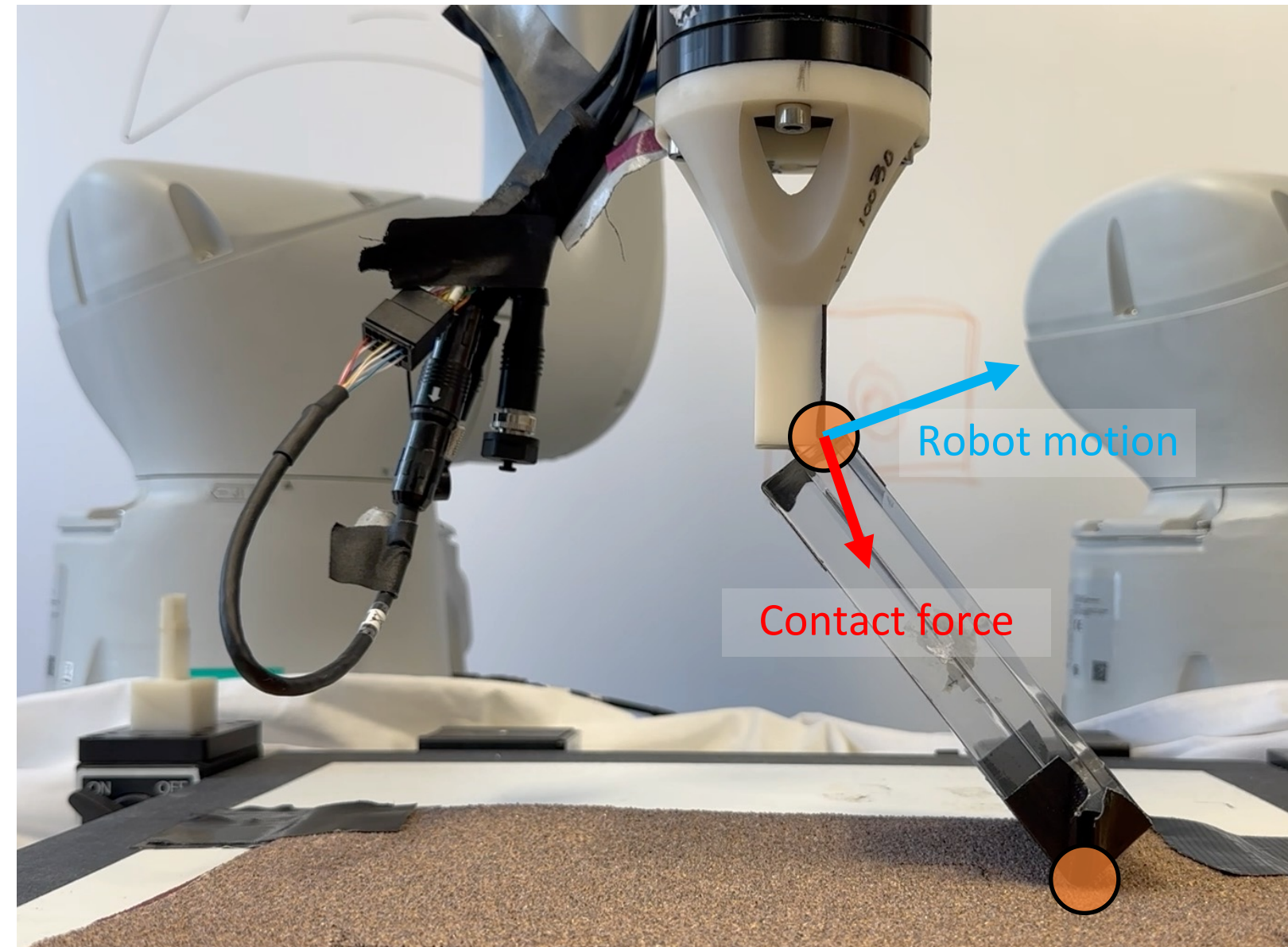
Yuki Shirai, Kei Ota, Devesh Jha, Diego Romeres

MITSUBISHI ELECTRIC
MITSUBISHI ELECTRIC RESEARCH LABORATORIES

NEURAL INFORMATION PROCESSING SYSTEMS

TL;DR: Contact-Aware Demos + RL => Sample Efficient Training & Zero-Shot Sim2Real Transfer for Non-Prehensile Manipulation
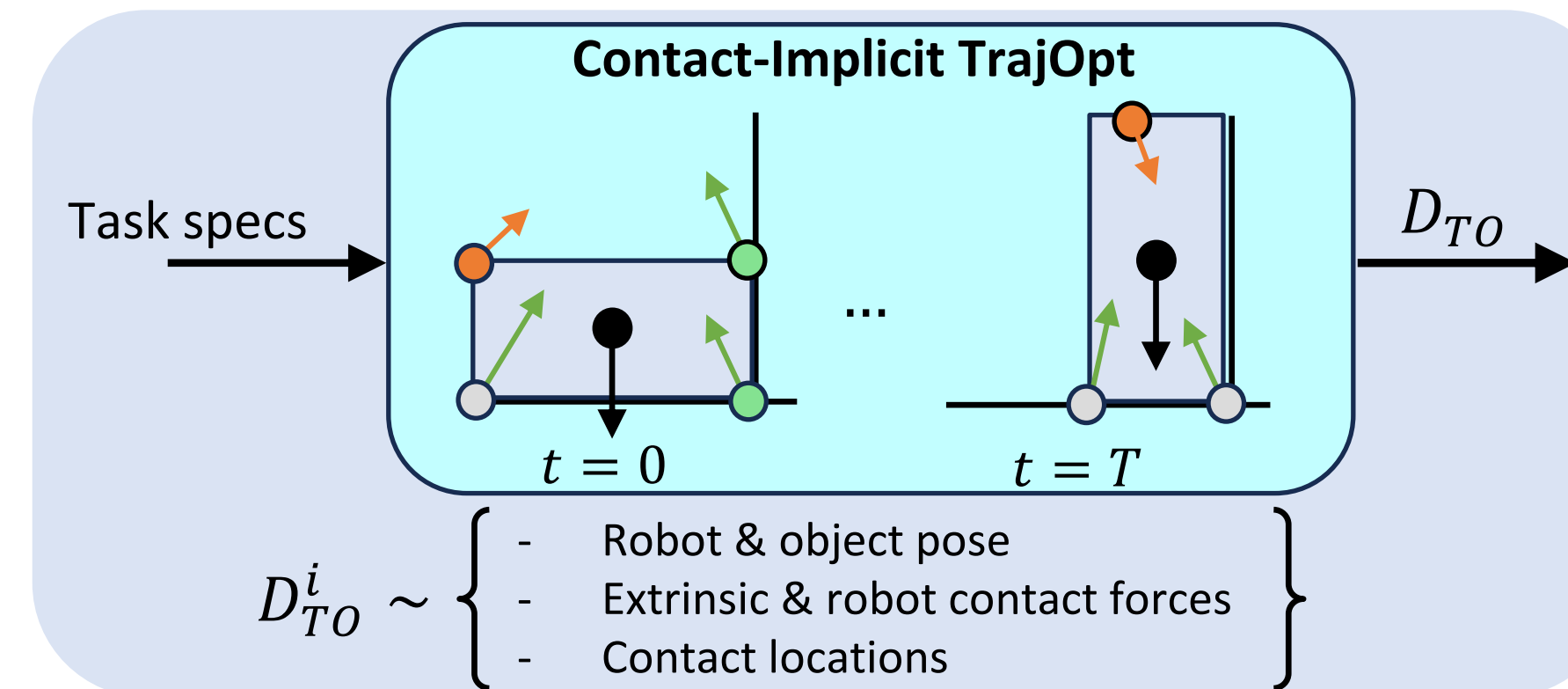
Motivation

Robot motion
Contact force

- Designing closed-loop controller for manipulation is difficult

- MPC: Requires privileged information (object pose, size, mass)
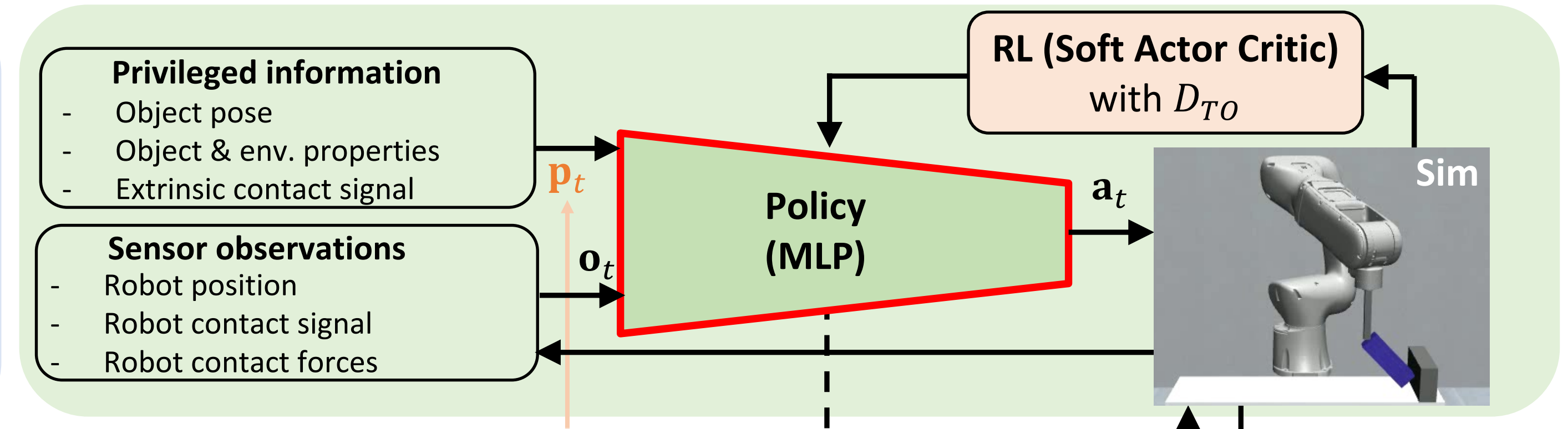- RL: Lots of data
-> Can we do better?

Contributions:
1. Framework for learning closed-loop non-prehensile manipulation by leveraging demonstrations generated by Contract TrajOpt
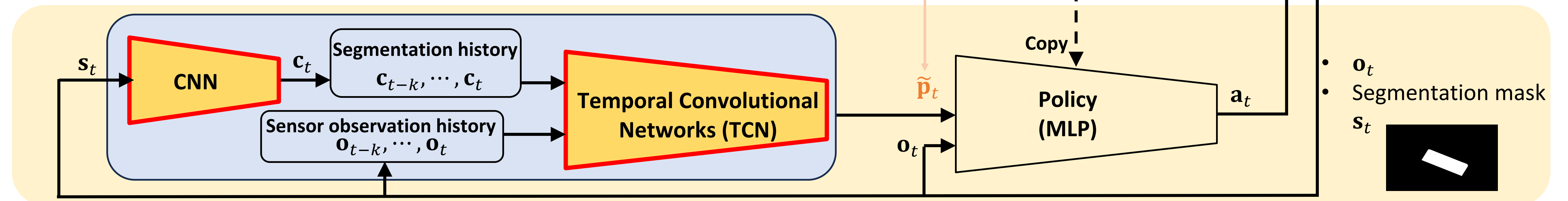2. Sim2Real transfer using temporal history of visual and force sensing
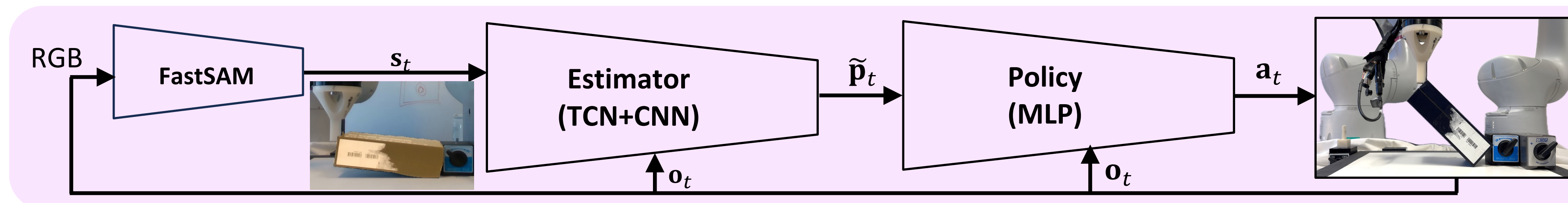
Method

Step 1: Data collection

Contact-Implicit TrajOpt

Task specs
$t = 0$ ... $t = T$ D_{TO}

$D_{TO}^i \sim$ { - Robot & object pose
- Extrinsic & robot contact forces
- Contact locations }

Step 2: Policy Training

Privileged information
- Object pose
- Object & env. properties
- Extrinsic contact signal
$\mathbf{p}_t$

Sensor observations
- Robot position
- Robot contact signal
- Robot contact forces
$\mathbf{o}_t$

Policy (MLP) $\mathbf{a}_t$

RL (Soft Actor Critic) with D_{TO}

Sim

Step 3: Privileged Information Estimator Training

$\mathbf{s}_t$ → CNN → $\mathbf{c}_t$ → Segmentation history $\mathbf{c}_{t-k}, \cdots, \mathbf{c}_t$

Sensor observation history $\mathbf{o}_{t-k}, \cdots, \mathbf{o}_t$ → **Temporal Convolutional Networks (TCN)**

$\widetilde{\mathbf{p}}_t$ → Policy (MLP) → $\mathbf{a}_t$

Regress
Copy
$\mathbf{o}_t$

- $\mathbf{o}_t$
- Segmentation mask
- $\mathbf{s}_t$

Step 4: Deployment in Real-World for Zero-Shot Sim2Real

RGB → FastSAM → $\mathbf{s}_t$ → **Estimator (TCN+CNN)** → $\widetilde{\mathbf{p}}_t$ → **Policy (MLP)** → $\mathbf{a}_t$

$\mathbf{o}_t$ $\mathbf{o}_t$

Assumptions.
- 2D
- Quasistatic
- Rigid objects

Result

Do Demonstrations by TrajOpt Facilitate Learning?

Vanilla RL
Kinematics-conditioned RL
Dynamics-conditioned RL

With wall / Without wall
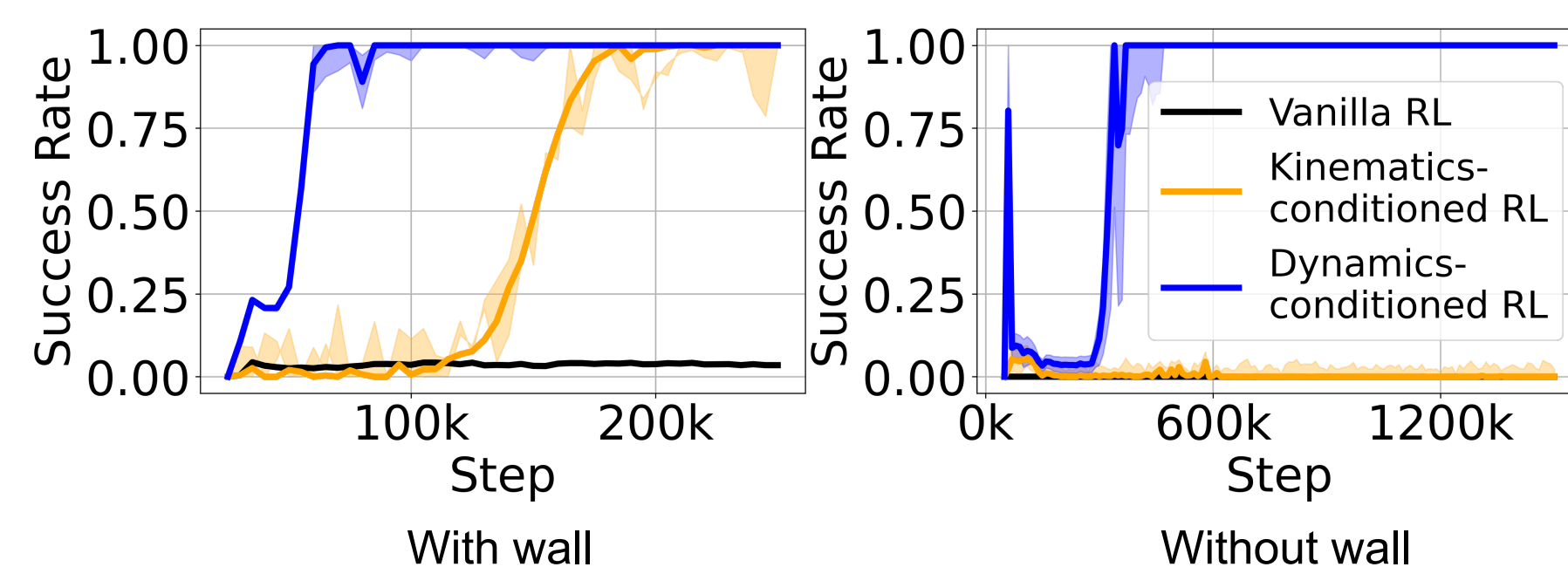
- Learning curves for different RL training runs over 3 seeds.
1. RL considering dynamics-conditioned demonstrations shows the best performance because of tight feasible action space

Comparison against Baselines

Task	MPC	BC	Dynamics-conditioned RL
With wall	98 / 100	2 / 100	100 / 100
Without wall	81 / 100	0 / 100	100 / 100

| Mass	Kinematics-conditioned RL	Dynamics-conditioned RL
50 g	2 / 5	**5 / 5**
110 g	**5 / 5**	**5 / 5**
300 g	0 / 5	**5 / 5**

- Number of successful attempts.
1. Our RL shows the highest success rate
2. Ours is more robust than RL only considering kinematics demos

Estimator Prediction Performance

Ground Truth
Student Estimate
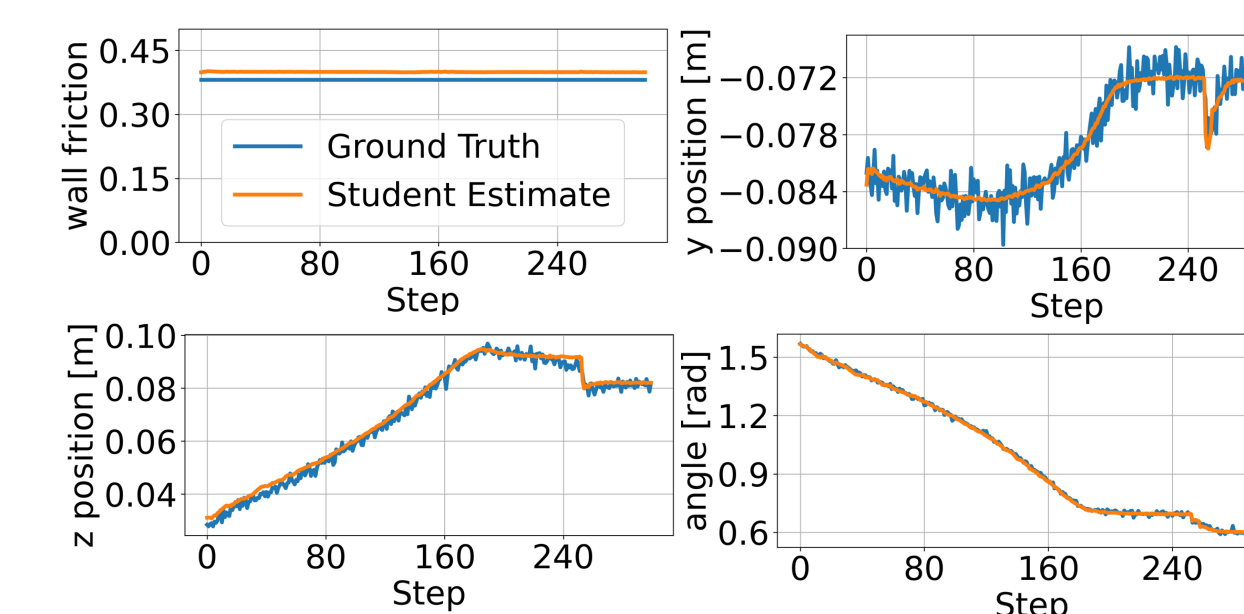
- Deploy the trained policy in MuJoCo
1. Our estimator successfully predicts the privileged information with reasonable accuracy.

Sim2Real Gap

Sim (mean)
Sim (95% CI)
Real (mean)
Real (95% CI)

With wall / Without wall
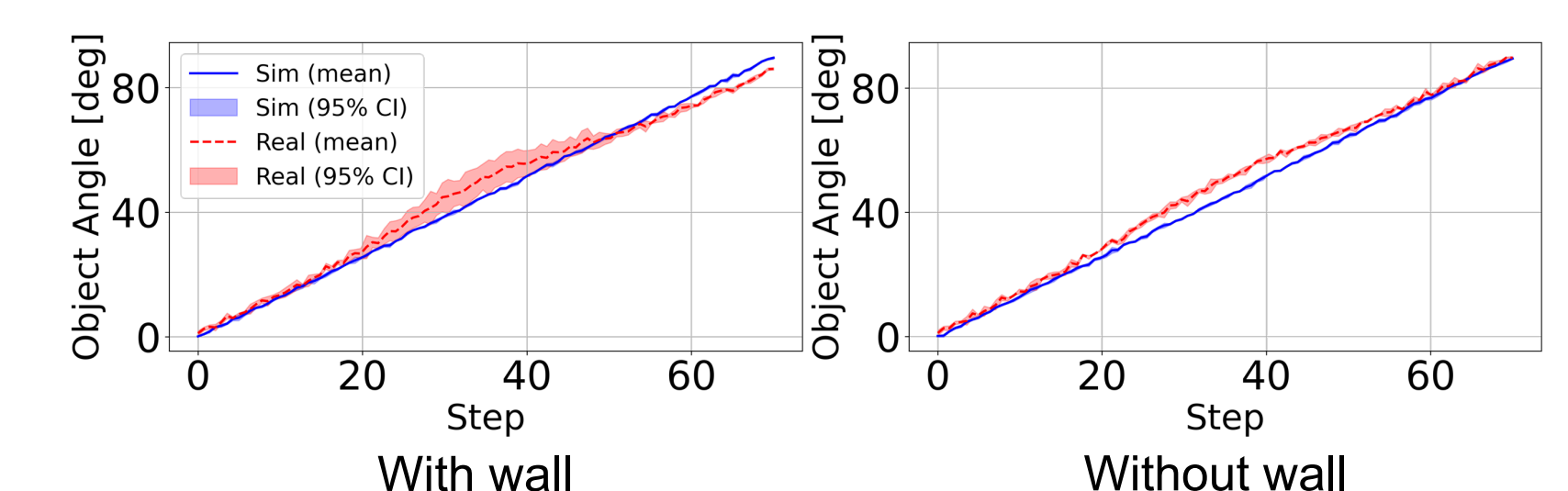
- Trajectory of object orientation by deploying the policy on simulation and hardware over 3 trials
1. Larger sim2real gap for the pivoting with external wall. This is because the object induces the sliding contacts, which are challenging to model in MuJoCo.