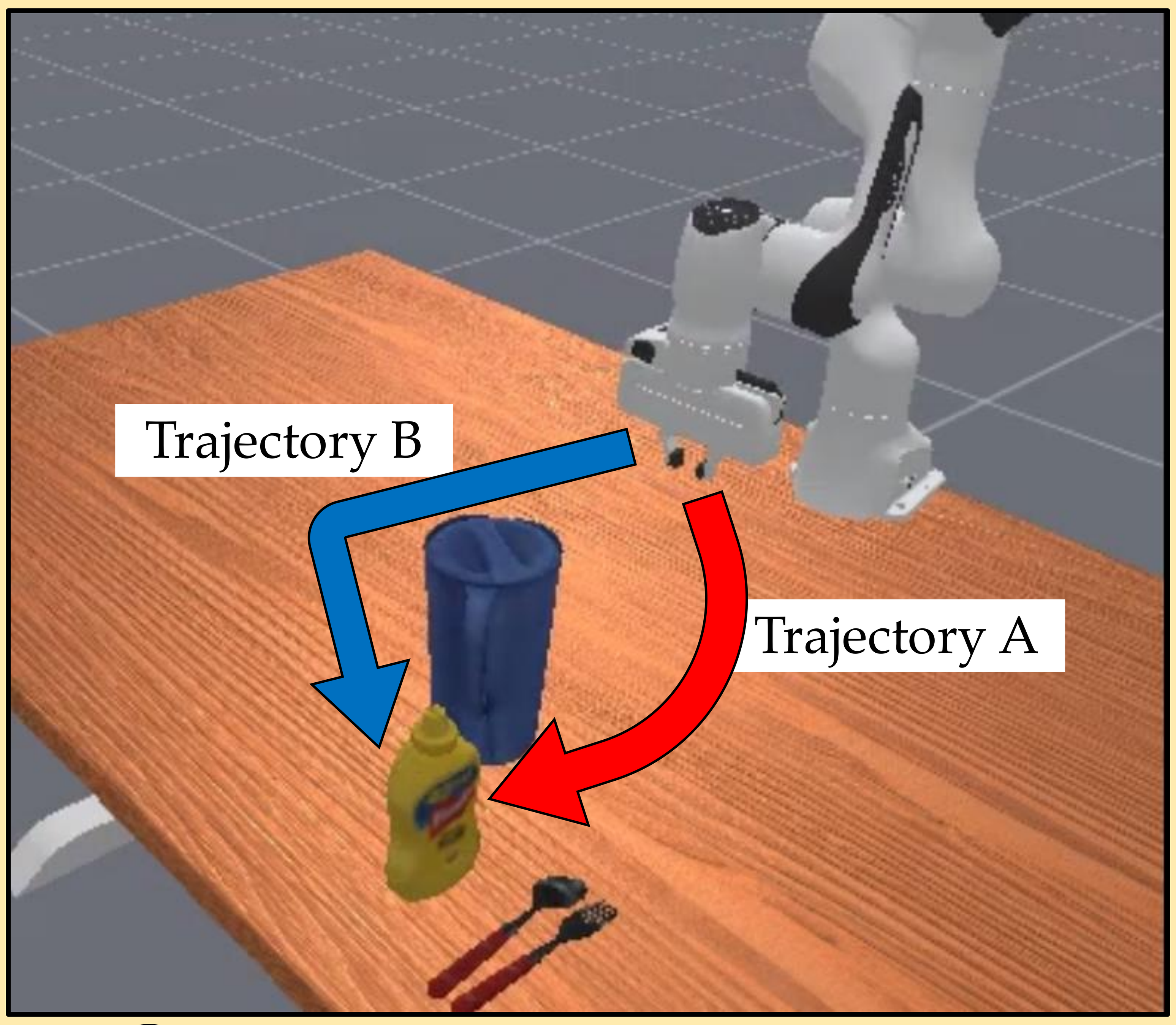
Causally Robust Reward Learning from Reason-Augmented Preference Feedback

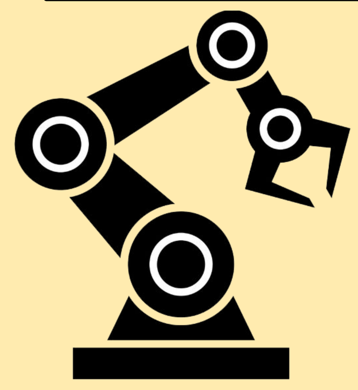
Minjune Hwang, Yigit Korkmaz, Daniel Seita†, Erdem Bıyık† (†: equal advising)
University of Southern California, Viterbi School of Engineering

Motivation

- **Preference-based reinforcement learning** shapes agent behaviors from user's binary preferences, which can leave it vulnerable to **causal confusion**.
- The learned reward can latch onto **spurious features** that cooccur with preferred trajectories, collapsing when those correlations disappear at test time.

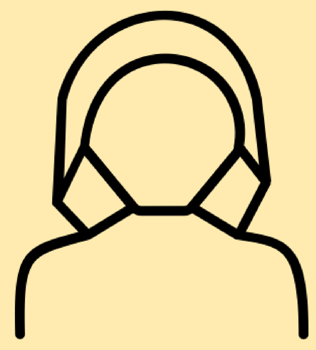
Insight: A natural language reason can (1) clarify true causal signals behind preference and (2) improve generalization beyond spurious features.





Could you tell me which trajectory you prefer?

I prefer Trajectory A, **because it smoothly avoids the obstacle.**



How can we use this information, i.e., **reasoning**?

Method: ReCouPLe

- We learn a trajectory encoder ϕ that maps a trajectory $\boldsymbol{\tau}$ into the pretrained language encoder's embedding space.
- We represent each task's reward given its task description ℓ_{task} as **an inner product** of $\phi(\boldsymbol{\tau})$ and task embeddings $\boldsymbol{\theta} = \text{LM}(\ell_{\text{task}})$:

$$r(\boldsymbol{\tau}, \ell_{\text{task}}) = \phi(\boldsymbol{\tau})^\top \boldsymbol{\theta}$$

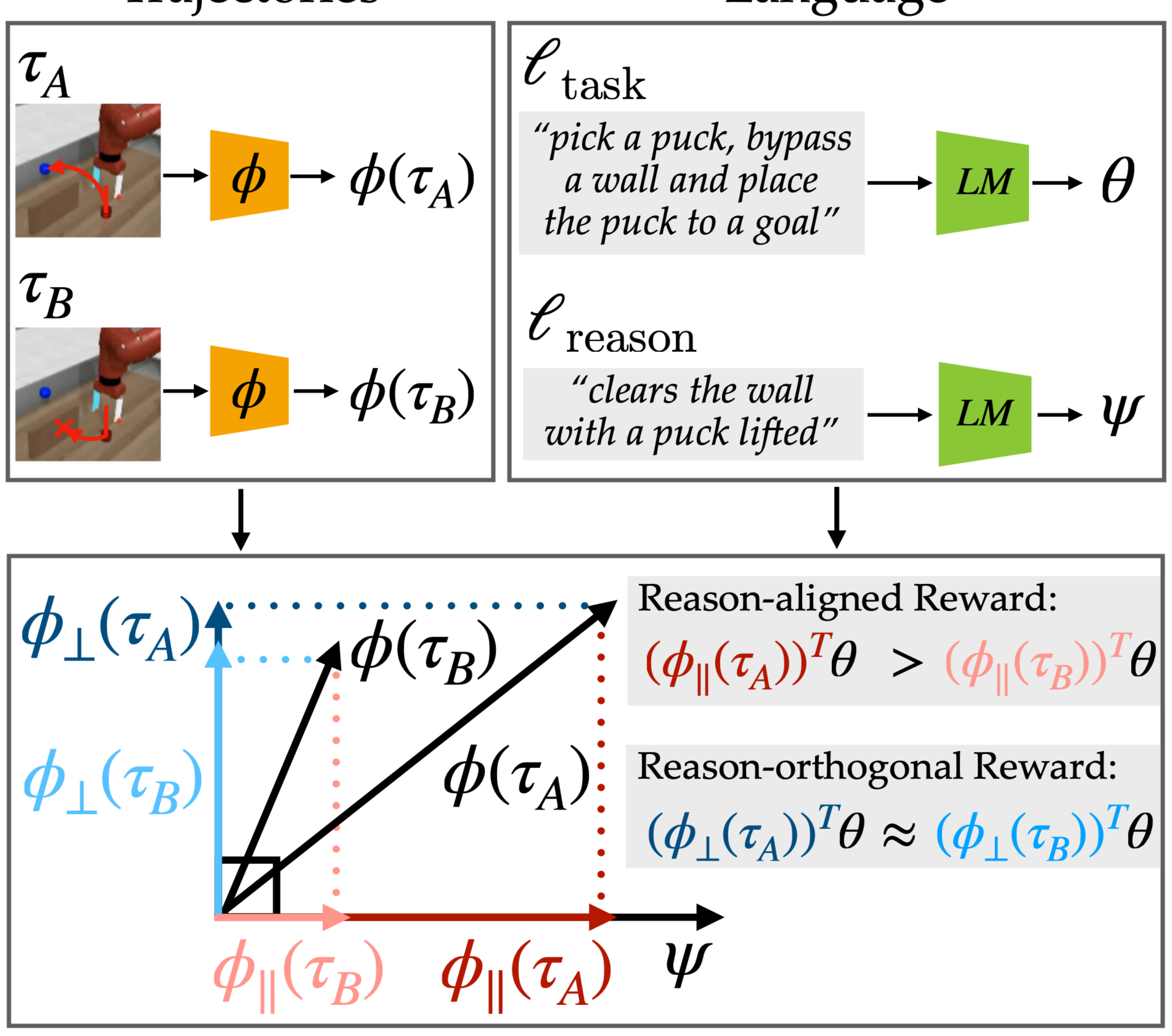
ReCouPLe treats the rationale embedding ψ as **a projection axis**, splitting the trajectory representation into reason-aligned and reason-orthogonal parts:

$$\phi_{\parallel}(\boldsymbol{\tau}) = \left(\frac{\phi(\boldsymbol{\tau})^\top \psi}{\|\psi\|_2^2}\right)\psi, \quad \phi_{\perp}(\boldsymbol{\tau}) = \phi(\boldsymbol{\tau}) - \phi_{\parallel}(\boldsymbol{\tau})$$

Then, $r(\boldsymbol{\tau}, \ell_{\text{task}}) = \phi_{\parallel}(\boldsymbol{\tau})^\top \boldsymbol{\theta} + \phi_{\perp}(\boldsymbol{\tau})^\top \boldsymbol{\theta}$

$$= \underbrace{r_{\parallel}(\boldsymbol{\tau})}_{\text{reason-aligned}} + \underbrace{r_{\perp}(\boldsymbol{\tau})}_{\text{reason-orthogonal}}$$

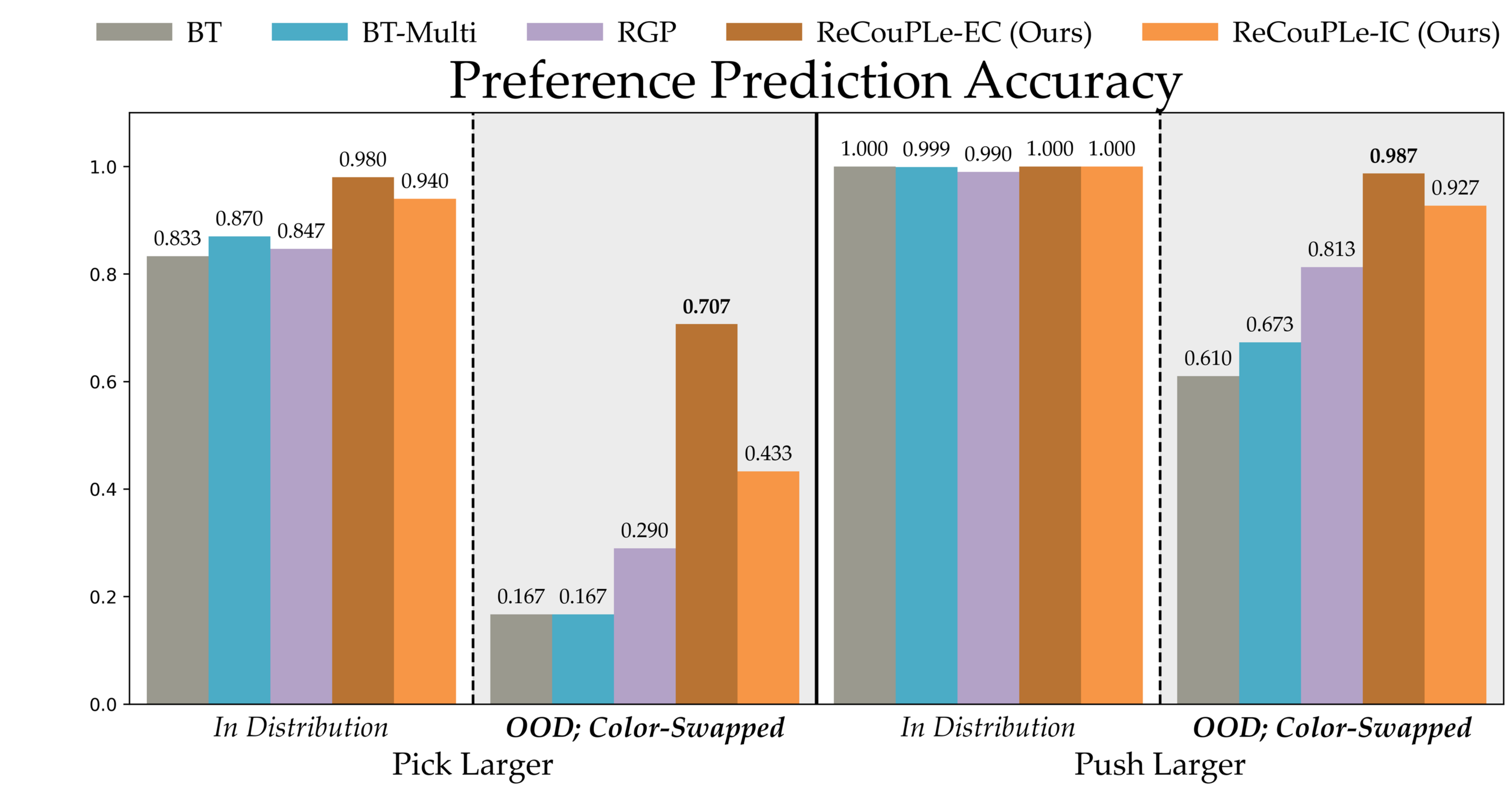
Should mainly determine preference | Should not affect preference as much



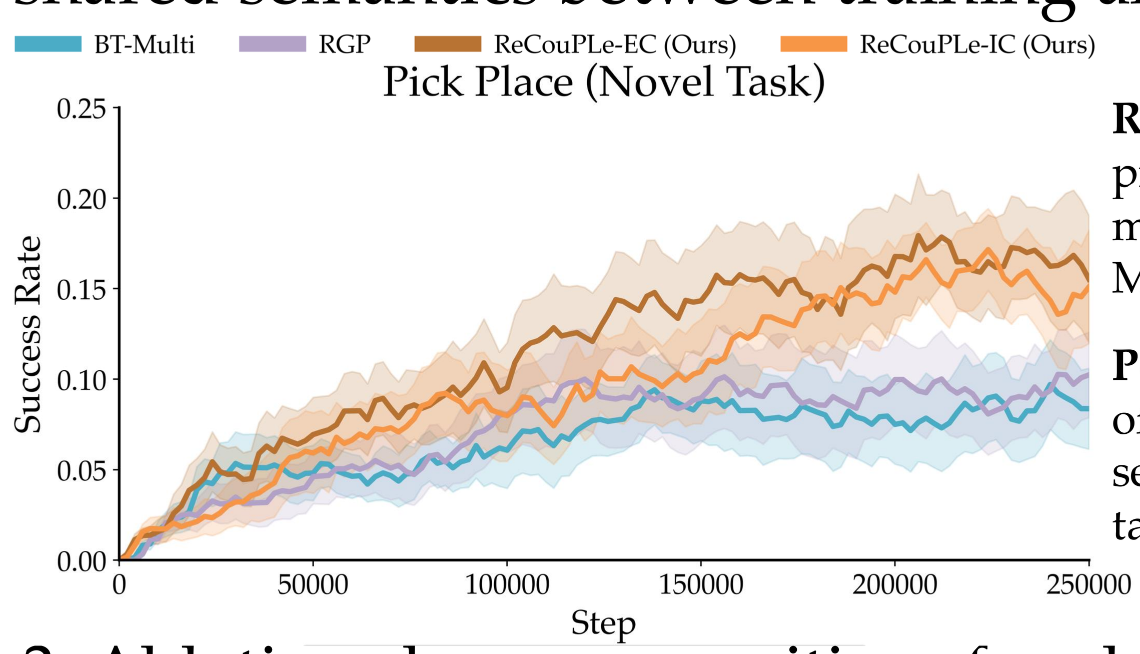
Joint Representation Space

Results

1. **ReCouPLe** robustly predicts user's preference under the distribution shifts.



2. **ReCouPLe** transfers to a novel task, thanks to shared semantics between training and unseen tasks.



Reward: trained with preference data of three manipulation tasks in Meta-world benchmark.

Policy: trained on the offline play data of a semantically similar, novel task (result shown left).

3. Ablation shows necessities of each component.

Model	2-task		4-task	
	ID	*OOD*	*ID*	*OOD*
ReCouPLe	**0.995**	**0.872**	**1.000**	**0.878**
ReCouPLe-no-consistency	0.980	0.726	0.977	0.745
ReCouPLe-no-consistency-no-ratio	0.987	0.727	0.990	0.730