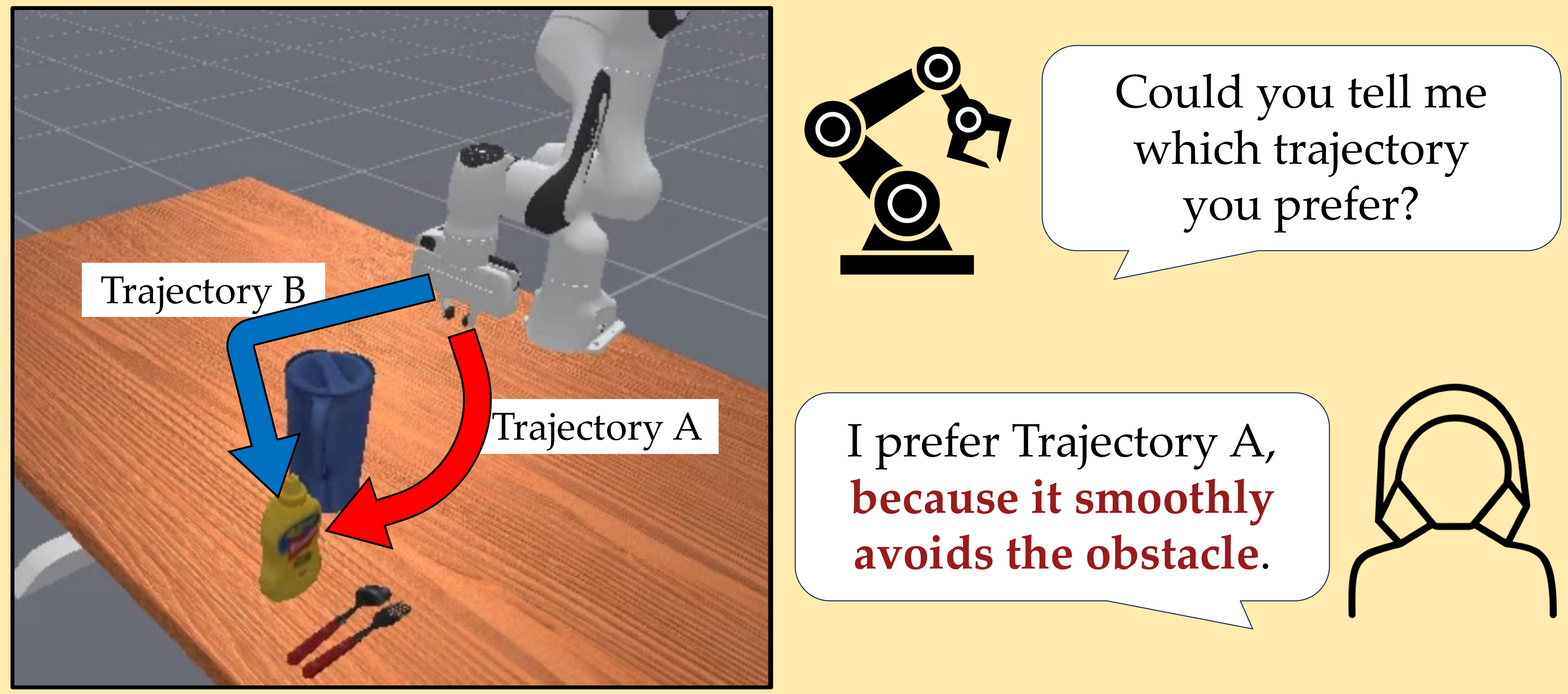
Causally Robust Preference Learning with Reasons

Minjune Hwang, Yigit Korkmaz, Daniel Seita†, Erdem Bıyık† (†equal advising)
{minjuneh, ykorkmaz, seita, biyik}@usc.edu

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 rationale can (1) clarify true causal signals behind preference and (2) improve generalization beyond spurious features.



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

Mitigating Causal Confusion with Reasons

Preliminary:

- We learn a trajectory encoder ϕ that maps a trajectory τ 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 $\phi(\tau)$ and task embeddings $\theta = \mathrm{LM}(\ell_{\text{task}})$:

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

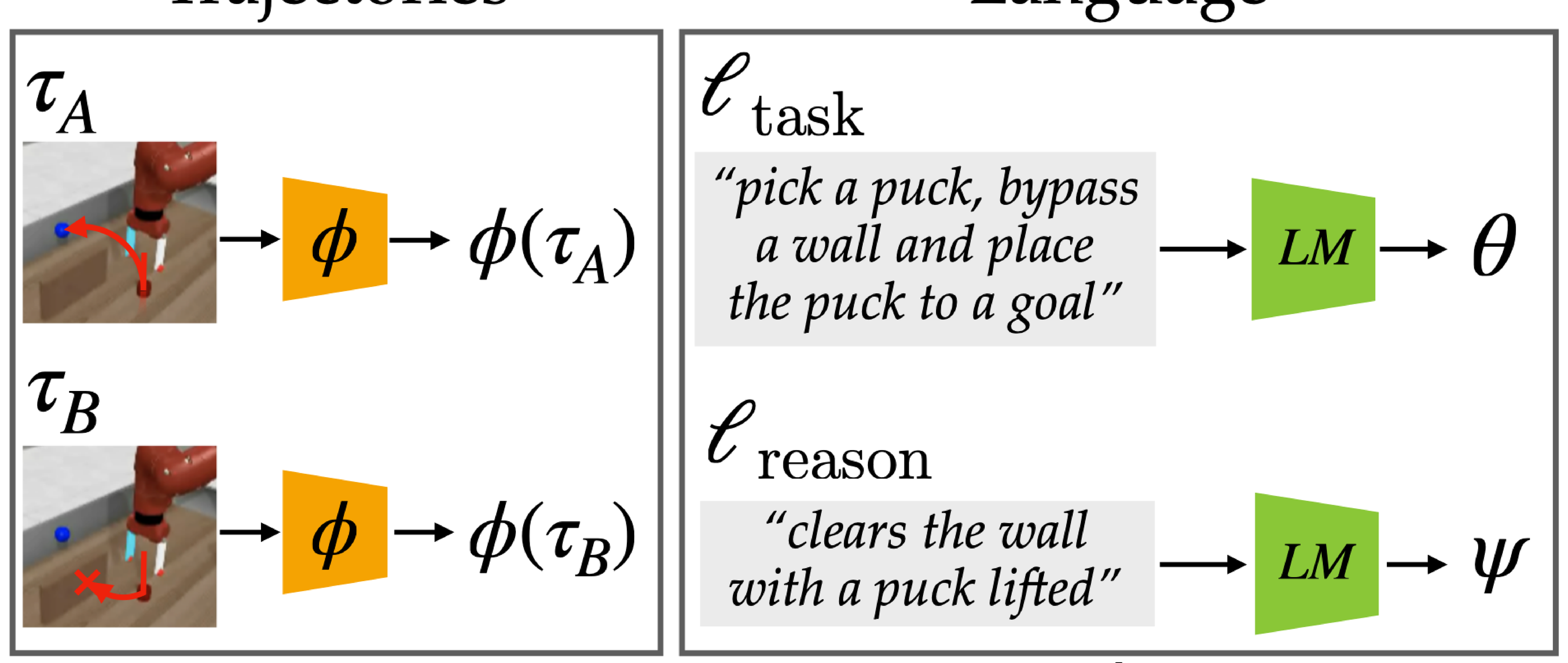
ReCouPLe treats the rationale embedding ψ as **a projection axis**, splitting the trajectory representation into reason-aligned and reason-orthogonal parts:

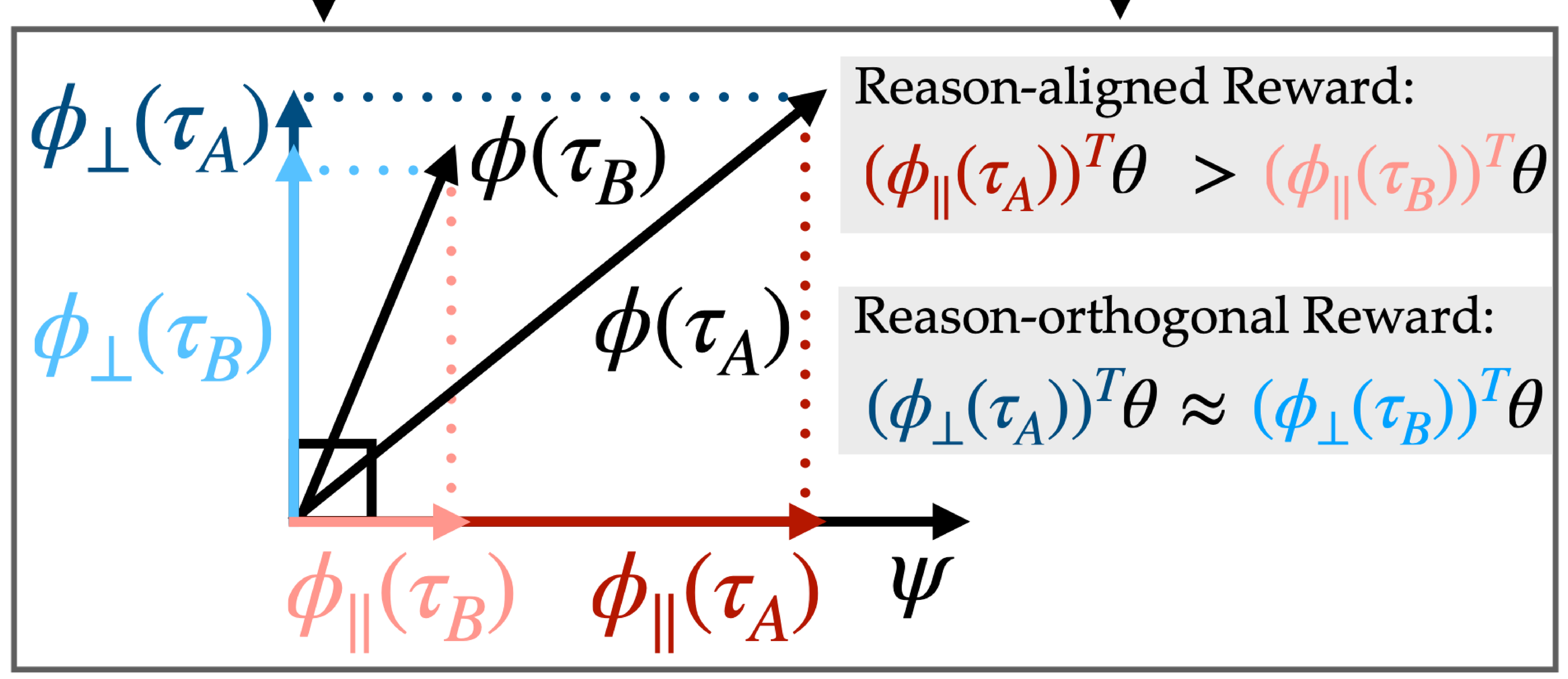
$$\phi_{\parallel}(\tau) = \left(\frac{\phi(\tau)^\top \psi}{\|\psi\|_2^2}\right)\psi, \quad \phi_{\perp}(\tau) = \phi(\tau) - \phi_{\parallel}(\tau)$$

Then, we can decompose the reward into two components:

$$r(\tau, \ell_{\text{task}}) = \phi_{\parallel}(\tau)^\top \theta + \phi_{\perp}(\tau)^\top \theta = \underbrace{r_{\parallel}(\tau)}_{\text{reason-aligned}} + \underbrace{r_{\perp}(\tau)}_{\text{reason-orthogonal}}$$

Should mainly determine preference | Should not affect preference as much

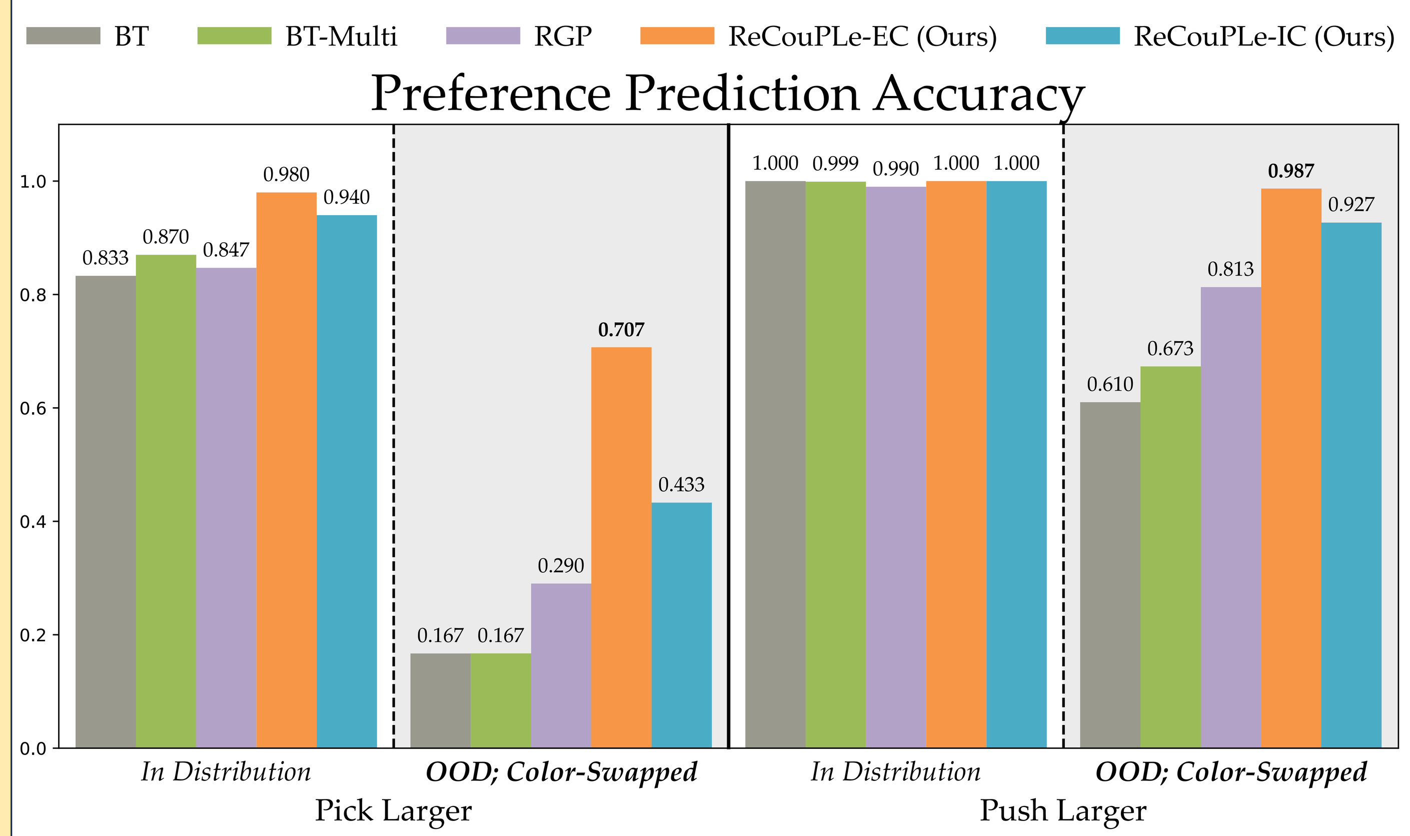




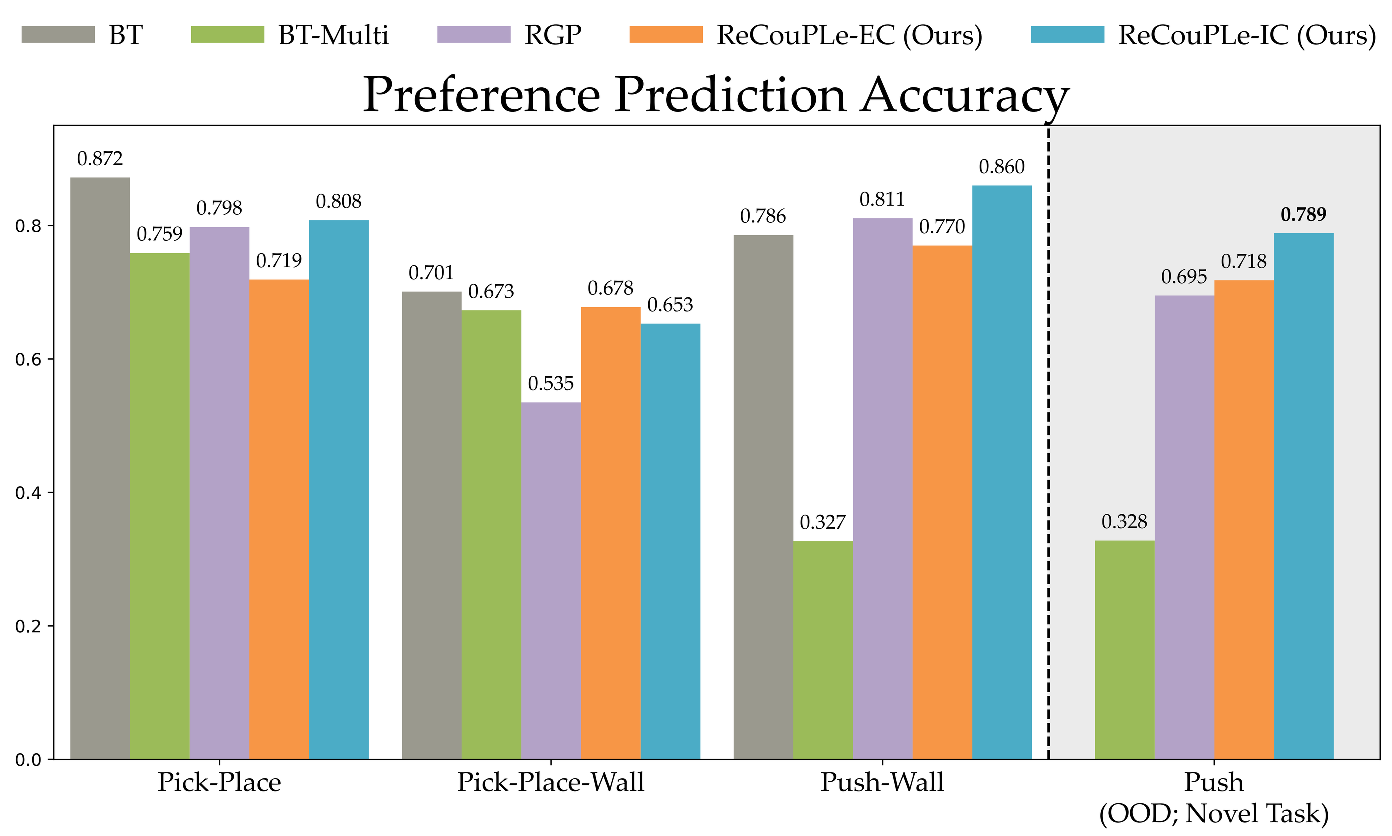
Joint Representation Space

Results

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



2. **ReCouPLe** transfers to a novel task without additional preference queries.



Future Work

- Policy learning with learned rewards to validate effectiveness in downstream tasks.
- Experiments on physical robot tasks.
- Active querying of a rationale for data efficiency.

References

[1] Tien, J., He, J.Z.Y., Erickson, Z., Dragan, A. and Brown, D.S., Causal Confusion and Reward Misidentification in Preference-Based Reward Learning. In The Eleventh International Conference on Learning Representations.

[2] Yang, Z., Jun, M., Tien, J., Russell, S., Dragan, A. and Biyik, E., Trajectory Improvement and Reward Learning from Comparative Language Feedback. In 8th Annual Conference on Robot Learning.