Causally Robust Preference Learning with Reasons

Author Names Omitted for Anonymous Review. Paper-ID [add your ID here]

***Abstract*—Preference-based reward learning is widely used for shaping agent behavior to match a user's preference, yet its sparse binary feedback makes it especially vulnerable to causal confusion. The learned reward often latches onto spurious features that merely co-occur with preferred trajectories during training, collapsing when those correlations disappear or reverse at test time. We introduce ReCouPLe, a lightweight framework that uses natural language rationales to provide the missing causal signal. Each rationale is treated as a guiding projection axis in embedding space, training the model to score trajectories by features aligned with that axis while de-emphasizing context that is unrelated to the stated reason. Because identical rationales can arise across multiple tasks (e.g., "it avoids collisions with a fragile object", "it correctly picks the tool I prefer"), ReCouPLe naturally reuses the same causal direction whenever tasks share semantics, and transfers preference knowledge to novel tasks without extra data or language-model fine-tuning. Our learned reward model can ground preferences on the articulated reason, aligning better with user intent and generalizing beyond spurious features.**

I. Introduction

Designing reward functions that faithfully capture human intent is one of the central obstacles to deploying learning agents in the real world. Preference-based reinforcement learning (PbRL) removes the need for hand-crafted rewards by asking a human which of two trajectories they prefer ([3, 1]). Unfortunately, this binary feedback conveys at most a single bit of information and leaves the reward model free to explain the preference with any correlating feature in its observation space. Under the presence of non-causal distractor features that are spuriously correlated with preference labels, reward models often learn to rely on such features. These features, however, are irrelevant to the task success. When those cues disappear or change at test time, the agent can suffer from reward misidentification and fail to generalize [13]. Since each comparison supplies at most one bit of information, it leaves many causal explanations indistinguishable. Without extra guidance, the learner cannot tell whether users prefer a trajectory for its smoothness, its speed, or some spurious cue in the background.

For instance, consider we want to train a robotic arm to pick up a box that is large enough to put toys (Fig: 1). During data collection, every preference query shows a large red box and a small blue box, and the annotator always prefers the former. Because size and color are perfectly correlated in these comparisons, a reward model can reach zero training error by attending to the color cue instead of true size. At test time, when it encounters a large blue box next to a small red box, the learned reward could mistakenly favor the small red box.

A natural solution is to supply richer feedback. Prior work has begun to augment pairwise comparisons with natural-language descriptions of how two trajectories differ. Following advancements in natural language processing, recent works in robot learning employed language for task planning, policy learning, and reward shaping [16, 4, 5, 11, 6]. Shi et al. [11] employ language-conditioned behavior cloning (LCBC) for corrective language commands and improving policies. Cui et al. [4] introduce an approach to use human language feedback to correct robot manipulation in real-time via shared autonomy. Dai et al. [5] propose a data generation pipeline that automatically augments expert demonstrations with failure recovery trajectories and fine-grained language annotations for training recovery policies. In the domain of preference learning, Yang et al. [14] learn a shared latent space for trajectories and comparative language like "move farther from the stove", showing that language can make reward learning faster and more intuitive. These existing approaches treat language as an additional input to the reward model, without exploiting its compositional structure or underlying rationale. A recent work by Peng et al. [9] also introduces an approach to incorporate feature-wise preference learning framework to enrich the informative signals with *why an example is preferred*. However, their study assumes agents have access to structured, task-relevant features for state abstraction, and their experiments are confined to a linear bandit setting.

We claim that short natural language *rationales* carries exactly the causal signal the model is missing. "*I prefer this trajectory because it avoids collisions*" tells the learner which feature matters for the user's preference. We present ReCouPLe (**Re**ason-based **Co**nf**u**sion Mitigation in **P**reference **Le**arning), a lightweight framework that treats each rationale as a directional guide in a shared trajectory–language representation space. We design a simple loss that encourages the preference to be based on the direction of the reason specified by the user, rather than on incidental correlations in a pair of trajectories or other unspecified factors. The language encoder remains frozen, retaining the same semantics across tasks. Subsequently, this decoupling of reason components enables us to exploit shared rationales appearing across tasks, achieving a reward function that generalizes from one task to another, potentially without any additional preference query.

In summary, our contributions are:

1) We observe that pairwise preferences can provide limited information when non-causal distractor features exist, which makes reward models vulnerable to causal confusion. By augmenting each comparison with a natural language rationale, we supply complementary causal cues that help disambiguate the true preference signal.
2) We propose ReCouPLe, a projection-and-regularization scheme that injects causal structure through language.

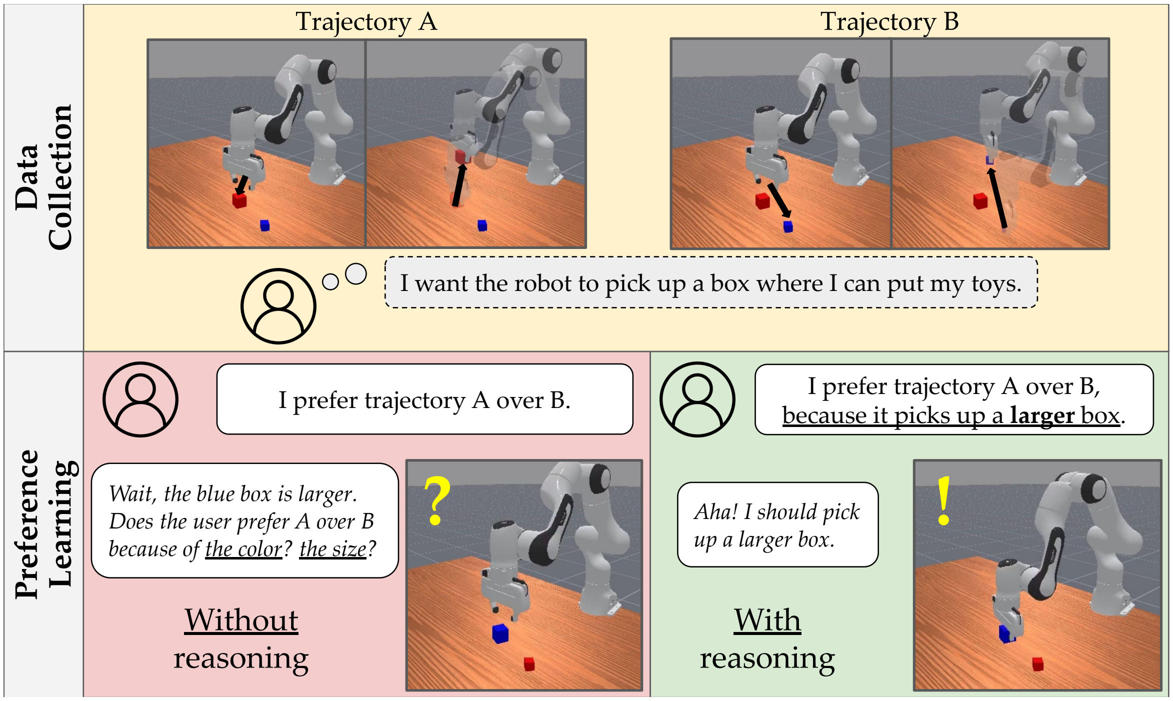
Data Collection
Trajectory A
Trajectory B
I want the robot to pick up a box where I can put my toys.
Preference Learning
I prefer trajectory A over B.
Wait, the blue box is larger. Does the user prefer A over B because of the color? the size?
?
Without reasoning
I prefer trajectory A over B, because it picks up a larger box.
Aha! I should pick up a larger box.
!
With reasoning

Fig. 1: Preference learning can be susceptible to causal confusion, especially with the presence of non-causal distractor features that merely co-occur with preferred trajectories. In the example above, the reward model struggles to identify the exact feature of a trajectory that determined user's preference. By providing reasoning, the agent can identify the causal feature.

3) We show that adding a rationale to each comparison yields transferable reward models that mitigate causal confusion compared to other baselines.

II. Problem Definition

We consider a collection of tasks modeled as finite-horizon Markov decision processes (MDPs) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, T)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $P(s_{t+1} \mid s_t, a_t)$ the transition kernel, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function, $\gamma \in [0, 1)$ the discount factor, and T the maximum time horizon. However, unlike standard RL, the reward function r is unknown and it must be inferred from the user's pairwise preference feedback.

A. Reward Learning from Preference Data

We assume access to preference data in the form of binary comparisons. Given a pair of trajectory segments (τ_A, τ_B) of horizon $H \leq T$, a human user provides preference label y:

$$y = \begin{cases} 1 & \text{if } \tau_A \succ \tau_B, \\ 0 & \text{otherwise.} \end{cases}$$

where $\tau = (s_0, a_0, \dots, s_H, a_H)$. Following the Bradley-Terry model [2], the probability that the trajectory τ_A is preferred over the trajectory τ_B is given by:

$$P_r(\tau_A \succ \tau_B) = \frac{\exp(r(\tau_A))}{\exp(r(\tau_A)) + \exp(r(\tau_B))} \tag{1}$$

In order to estimate the true reward, prior works in preference-based RL train the reward function $\hat{r}_\omega : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ parameterized by ω by minimizing the binary cross-entropy loss with the Bradley-Terry model:

$$\mathcal{L}_{\text{BT}} = -\sum_{(A,B)} \left[y_{AB} \log P_{\hat{r}_\omega}(\tau_A \succ \tau_B) \right. + \\ \left. (1 - y_{AB}) \log(1 - P_{\hat{r}_\omega}(\tau_A \succ \tau_B)) \right] \tag{2}$$

B. Language Interfaces

In addition to the standard MDP transitions, each task comes with a **task description** ℓ_{task}, a short sentence such as "pick up the cup" or "push the cube." We assume each task language label should contain semantic information regarding the corresponding task and it can imply the task's reward function. Additionally, each preference label has an optional **reason** ℓ_{reason} that explains why one trajectory is preferred over the other (e.g., "because it avoids collisions"). A frozen language encoder LM maps these strings to fixed embeddings of dimension d: $\theta = \text{LM}(\ell_{\text{task}}) \in \mathbb{R}^d$ and $\psi = \text{LM}(\ell_{\text{reason}}) \in \mathbb{R}^d$.

III. Methods

Preference-based reinforcement learning typically fits a single-task reward by maximizing the likelihood of observed comparisons (Eq. (1)). We extend this framework to the multi-task setting, where each task is identified by its language description. Specifically, we model the reward as the inner product between the trajectory representation and the task embedding:

$$\begin{aligned} r(\tau, \ell_{\text{task}}) &= \phi(\tau)^\top \text{LM}(\ell_{\text{task}}) \\ &= \phi(\tau)^\top \theta, \end{aligned} \tag{3}$$

where the trajectory encoder $\phi : \tau \rightarrow \mathbb{R}^d$ is the only trainable component, as the task embedding $\theta = \mathrm{LM}(\ell_{\text{task}})$ is frozen (Sec. II-B). We use this reward formulation across all methods for consistency. Although linear in structure, the nonlinearity of both the trainable trajectory encoder and the frozen language model allows this simple form to capture complex, task-specific reward structures.

We study three methods that share the same multi-task reward formulation but differ only in the loss terms used to train the trajectory encoder ϕ:

- **BT-Multi (baseline)**: Uses the standard Bradley–Terry loss on the multi-task reward, without reasons.
- **RFP (baseline)**: Adds an additional Bradley–Terry loss on the reason–trajectory dot product.
- **ReCouPLe (ours)**: Decomposes trajectory features into reason-aligned and orthogonal components and regularizes them.

In our experiments, we use the pretrained T5 [10] language model encoder as LM.

A. Multi-task Bradley-Terry Baseline (BT-Multi)

The BT-Multi baseline learns ϕ by minimizing the binary cross-entropy loss $\mathcal{L}_{\text{BT}}$ (Eq. (2)) across all tasks, using the shared reward definition above and ignoring the rationale ℓ_{reason}. It therefore serves as the baseline without reason inputs.

B. Reason-Feature-Preference Baseline (RFP)

Our RFP baseline follows *Pragmatic Feature Preferences* (PFP) [9]. PFP assumes that each state can be represented by an *explicit, hand-designed feature vector*. Humans (i) specify which of those features are relevant to the task and (ii) give pairwise labels that compare *each relevant feature* across two items. The algorithm then fits a linear reward with a Bradley-Terry (BT) loss for every feature-level comparison, as well as for the trajectory-level comparison.

In our setting, such manually chosen features do not exist. While PFP relies on manually chosen task-relevant features that humans compare one-by-one, we treat the frozen rationale embedding $\psi = \mathrm{LM}(\ell_{\text{reason}})$ as a single *implicit* feature direction in the representation space. Besides the shared reward $r(\tau, \ell_{\text{task}})$ from Eq. 3, we also define a **reason score** q:

$$\begin{aligned} q(\tau, \ell_{\text{reason}}) \;&=\; \phi(\tau)^\top \mathrm{LM}(\ell_{\text{task}}) \\ &=\; \phi(\tau)^\top \psi. \end{aligned}$$

Training minimizes the standard task BT loss from Eq. 2 (same as BT-Multi) and an additional auxiliary BT term for the specified reason's score, with its weight λ_{q}:

$$\mathcal{L}_{\text{RFP}} = \mathcal{L}_{\text{BT}}\big(r_A, r_B\big) + \lambda_{\text{q}}\, \mathcal{L}_{\text{BT}}\big(q_A, q_B\big),$$

where $r_A = r(\tau_A, \ell_{\text{task}})$ and $q_A = q(\tau_A, \ell_{\text{reason}})$.

Limitations of RFP. RFP adds a reason score to better reflect the rationale behind a preference, but it lacks two structural safeguards.

- **No built-in separation**: It gives the model no explicit signal to tell apart the dimensions that should explain the stated reason from the rest of the embedding.
- **No neutrality constrains**: It offers no mechanism to keep the leftover dimensions from sneaking into the preference signal or to stop the whole trajectory embedding from collapsing onto the reason direction.

Without these safeguards, the encoder can still ignore the task vector θ, rely on incidental cues, and generalize poorly.

C. ReCouPLe

Key idea. A sentence such as *"I prefer this path because it avoids collisions"* pinpoints the causal feature. ReCouPLe treats the rationale embedding ψ as a projection axis, splitting the trajectory representation into reason-aligned and reason-orthogonal parts.

Explicit geometric split. This projection induces two disjoint subspaces:

$$\phi(\tau) = \underbrace{\phi_{||}(\tau)}_{\text{reason-aligned}} + \underbrace{\phi_{\perp}(\tau)}_{\text{reason-orthogonal}}, \quad \phi_{||}^\top \phi_{\perp} = 0,$$

which is achieved by

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

Correspondingly, the reward term (Eq. 3) decomposes as

$$r(\tau, \ell_{\text{task}}) = \underbrace{r_{||}(\tau)}_{\text{explained by rationale}} + \underbrace{r_{\perp}(\tau)}_{\text{residual task signal}},$$

where

- $r_{||}(\tau) = \phi_{||}^\top \theta$ is the *causal* component explicitly justified by the user's rationale;
- $r_{\perp}(\tau) = \phi_{\perp}^\top \theta$ is the *orthogonal* component that captures any task-relevant information the rationale overlooks (e.g., shaping rewards or domain priors) but is prevented from influencing pairwise preferences by ReCouPLe's orthogonal constraints.

By forcing reward differences to depend solely on $r_{||}$ while holding $r_{\perp}$ neutral, ReCouPLe grounds decisions in the stated reason and prevents the model from relying on incidental correlations.

Training losses. Given the decomposed reward, we train our trajectory presentation by three reward terms:

1) **Reason BT loss**: BT loss $\mathcal{L}_{\text{BT}}$ (Eq. (2)) on $r_{||}$ *only*, enforcing that preferences are explained through the stated causal feature.
2) **Orthogonal consistency loss**: Consistency loss ensuring that the preference to be **not** explained by the reason-orthogonal component. There are two variants for this term:
   a) **ReCouPLe-EC**: **Equality** constraint $r_{\perp}(\tau_A) \approx r_{\perp}(\tau_B)$ for every comparison, ensuring $\phi_{\perp}$ carries no preference signal: $\mathcal{L}_{\text{eq}} = \big(r_{\perp}(\tau_A) - r_{\perp}(\tau_B)\big)^2$.
   b) **ReCouPLe-IC**: **Inequality** (BT) constraint encouraging the difference between $r_{||}$ is greater than $r_{\perp}$: $\mathcal{L}_{\text{ineq}} = \mathcal{L}_{\text{BT}}\Big(r_{||}(\tau_A) - r_{||}(\tau_B),\; r_{\perp}(\tau_A) - r_{\perp}(\tau_B)\Big)$.

   ReCouPLe-EC imposes a strict condition, requiring the reason-orthogonal components to be identical for compared

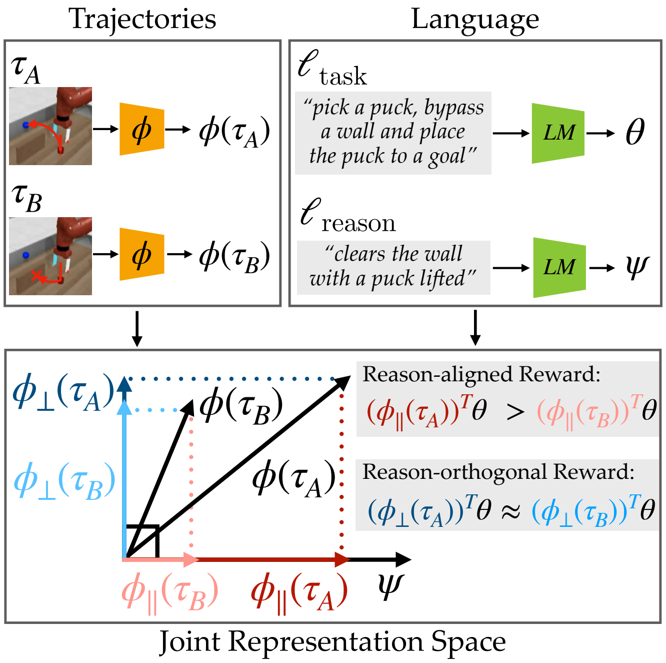
Trajectories
Language
τ_A
ϕ
$\phi(\tau_A)$
τ_B
ϕ
$\phi(\tau_B)$
ℓ_{task}
"pick a puck, bypass a wall and place the puck to a goal"
LM
θ
ℓ_{reason}
"clears the wall with a puck lifted"
LM
ψ
$\phi_\perp(\tau_A)$
$\phi(\tau_B)$
$\phi_\perp(\tau_B)$
$\phi(\tau_A)$
Reason-aligned Reward:
$(\phi_{||}(\tau_A))^T\theta > (\phi_{||}(\tau_B))^T\theta$
Reason-orthogonal Reward:
$(\phi_\perp(\tau_A))^T\theta \approx (\phi_\perp(\tau_B))^T\theta$
$\phi_{||}(\tau_B)$
$\phi_{||}(\tau_A)$
ψ
Joint Representation Space

Fig. 2: ReCouPle decomposes the task reward by orthogonally projecting the trajectory representation to the reason language embedding and decomposing the representation into reason-aligned and reason-orthogonal components. This allows the reward model to isolate the causal feature specified in the rationale to explain the user's preference, while preserving auxiliary task-relevant signals that do not influence the pairwise preference in the orthogonal component.

trajectories. In contrast, ReCouPLe-IC is less restrictive, incentivizing differences in the reason-aligned component to dominate differences in total task rewards.

3) **Reward-ratio regularizer** $\mathcal{L}_{\text{ratio}}$ for keeping the magnitude of $r_{||}$ below a fraction α of the total reward magnitude $(r_{||}+r_\perp)$, preventing trivial collapse into the causal subspace: $\mathcal{L}_{\text{ratio}} = \text{ReLU}\left(\frac{|r_{||}|}{|r_{||}|+|r_\perp|+\epsilon} - \alpha\right)$, where small constant ϵ is included in the denominator to prevent division by zero.

The final objective is the following:

$$\mathcal{L}_{\text{ReCouPLe}} = \mathcal{L}_{\text{BT}}(r_{||}) + \lambda_{\text{ratio}}\mathcal{L}_{\text{ratio}} + \begin{cases} \lambda_{\text{eq}}\mathcal{L}_{\text{eq}} & \text{(ReCouPLe-EC)}, \\ \lambda_{\text{ineq}}\mathcal{L}_{\text{ineq}} & \text{(ReCouPLe-IC)}. \end{cases}$$

IV. Experiments

We evaluate ReCouPLe on two complementary suites that probe distinct facets of the method. The first suite focuses on causal robustness in a single visuomotor task whose visual cues are deliberately confounded; the second investigates cross-task generalization in a multi-task manipulation benchmark. Together they address two research questions:

- **RQ1** (Robustness against causal confusion): Can ReCouPLe maintain preference accuracy when the covariate distribution shifts in a way that exposes spurious correlations?
- **RQ2** (Task transfer): Does the reason-aligned subspace learned on a small set of tasks transfer to novel tasks without additional preference queries?

We test **RQ1** with a set of custom ManiSkill environments for visuomotor tasks where distribution shift in non-causal visual features can easily yield causal confusion [8]. We assess **RQ2** with a set of Meta-World [17] tasks that are widely used to test few-shot/zero-shot transfer.

For both experiments, we let $\phi(\tau)$ be the sum of per-step state-action embeddings. We use a neural network encoder $e : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ to encode every state-action pair of a trajectory into the corresponding per-step embedding; we use a convolutional encoder similar to DrQ-v2 [15] for the ManiSkill visual control tasks and a fully-connected network for Meta-world state-based control tasks.

A. ManiSkill Task Suite for RQ1

Task design. To test whether our proposed method can mitigate causal confusion in preference-based learning, we design a set of object manipulation tasks in ManiSkill [12] that can easily induce causal confusion under distribution shifts. Each scene has two cubes of different sizes on a tabletop, and the agent must manipulate the **larger** cube. We have 4 total tasks: *MS-Pick-Larger*, *MS-Push-Larger*, *MS-Place-Larger*, and *MS-Pull-Larger*. During training, the larger cube is always a fixed color for each task, creating a perfect correlation between color and the correct behavioral choice. For *MS-Pick-Larger* and *MS-Pull-Larger*, the larger cube is always red and the smaller cube is always blue. In contrast, for *MS-Push-Larger* and *MS-Place-Larger*, the larger cube is always blue and the smaller cube is always red. At test time we swap the colors so that the distribution shift induces the classic "shortcut" failure: a model that latches onto color will choose the wrong object.

Data generation. We design motion planning solutions that manipulate either the larger or the smaller cube for each task, where initial cube poses are randomized. We then collect 500 synthetic preference queries for each task by pairing trajectories that manipulate the larger and smaller cubes, respectively. Each trajectory is randomly sub-sampled to a segment of length 64. The preference label selects the trajectory handling the larger cube. The accompanying rationale ℓ_{reason} is *"(because) the cube is larger"*. The task label ℓ_{task} is simply *"[manipulating verb] the larger cube"*.

Metric. After training a reward model, we evaluate the model using preference accuracy, defined as the proportion of held-out preference queries where the model correctly predicts which of the two trajectories is preferred. We run evaluation both on the in-distribution (ID) validation set and on the color-swapped, out-of-distribution (OOD) set. We test baselines and ReCouPLe on a 2-task setting with *MS-Pick-Larger* and *MS-Push-Larger* tasks, and on a 4-task setting with all tasks.

B. Meta-World Task Suite for ***RQ2***

Task design. We select three training tasks from Meta-world: *Pick-Place*, *Pick-Place-Wall*, and *Push-Wall*. We re-

TABLE I: Reward accuracy comparison for ManiSkill 2-task setting (**RQ1**), averaged over 3 seeds.

Model	In Distribution		Color Swapped	
	Pick	Push	Pick	Push
Single Task				
BT (pick)	0.833	-	0.167	-
BT (push)	-	1.000	-	0.610
Multi-Task				
BT-Multi	0.870	0.999	0.167	0.673
Multi-Task with Reasons				
RFP	0.847	0.990	0.290	0.813
ReCouPLe-EC ($\lambda_{\text{ratio}} = 0.2$)	**0.987**	**1.000**	**0.733**	**1.000**
ReCouPLe-EC ($\lambda_{\text{ratio}} = 0.4$)	0.980	**1.000**	0.707	0.987
ReCouPLe-IC ($\lambda_{\text{ratio}} = 0.2$)	0.980	**1.000**	0.560	0.653
ReCouPLe-IC ($\lambda_{\text{ratio}} = 0.4$)	0.940	**1.000**	0.433	0.927

serve *Push*, a variant of *Push-Wall* task without a wall that parallels the structural difference between *Pick-Place* and *Pick-Place-Wall*. Each task's ground-truth reward is linearly decomposed into interpretable components (grasp, lift, collision avoidance waypoints, etc.) provided by the benchmark.

Data generation. We first collect trajectories by rolling out policies with different levels of optimality and Gaussian noise, similar to data collection procedure in Hejna and Sadigh [7]. Then, for each query, we randomly sample two trajectory segments τ_A and τ_B and generate the preference label based on their total reward $\sum_i r(s_i, a_i)$, where Meta-World's predefined environment reward can be linearly decomposed into feature components $\{f_i\}$: $r(s,a) = \sum_i w_i f_i(s,a)$. Without loss of generality, suppose τ_A is preferred over τ_B. Now, we synthetically generate the reason label by computing component-wise advantages $\Delta_i = w_i(f_i(\tau_A) - f_i(\tau_B))$ and convert them to a softmax-human distribution from which we sample the reason behind the preference:

$$\mathbf{Pr}(\text{choose reason i}) = \frac{\exp(\Delta_i)}{\sum_j \exp(\Delta_j)}$$

Each sampled reason is a free-form sentence such as "keeps a firm grasp while steering toward the goal." We generate 4000 preference–rationale pairs for each training task (12000 total).

V. Results

A. ManiSkill Task Suite for RQ1

Tables I and II summarize preference prediction accuracy for all models before and after we swap the distracting color cue that is perfectly correlated with object size during training.

Under the two-task setting, single-task BT baselines appear competent while the training correlation holds (83.3% and 100%; in-distribution), yet their accuracy drops once the colors are swapped (16.7% and 61.0%). Sharing visual features across tasks with the multi-task baseline (BT-Multi) helps *Push-Larger* but leaves *Pick-Larger* just as brittle (0.167). Adding natural language rationales without our projection (i.e., RFP) raises OOD accuracy only slightly, confirming that naively adding the auxiliary BT term for reasons alone provides an informative but insufficient signal. ReCouPLe significantly improves its accuracy under distribution shift, demonstrating its robustness against causal confusion. A similar pattern persists for the four-task setting, yet overall performances of all baselines and our method are significantly improved as additional data across a more diverse set of tasks provide better generalization signal.

For all tasks under different settings, ReCouPLe exhibits highest out-of-distribution generalization performance. It persistently outperforms the RFP baseline, showing how our reason-guided reward decomposition method helps learn more **robust reward models against causal confusion**, compared to the simple addition of BT loss for reason features.

Another finding is that ReCouPLe with the equality constraint appears slightly more effective than the inequality constraint variant. We hypothesize that this is a result of our data collection scheme: for each task, both the preferred and suboptimal trajectories are generated using motion planning solutions that follow identical action sequences, differing only in which cube (larger or smaller) is manipulated and its randomized initial position. All other aspects of the trajectories, such as path smoothness, timing, and waypoints, nearly remain identical. Thus, the equality constraint for reason-orthogonal reward components can provide stronger and more accurate regularization.

Lastly, our method is robust against the λ_{ratio} hyperparameter choices. We observe that values in $[0.2, 0.4]$ do not significantly affect the results, and are sufficient to prevent embedding collapse.

B. Meta-World Task Suite for RQ2

Table III demonstrates preference prediction accuracy for the Meta-World task suite, which assesses whether each method can generalize to an unseen task that shares some level of similarities with training tasks. In our experiment, we evaluate its preference prediction accuracy on the held-out *Push* preference dataset, as well as on the validation dataset with training tasks.

We first observe that augmenting preference queries with reasons improves task generalization. The RFP baseline achieves better accuracy on the novel task (69.5%) compared to the BT-Multi baseline, demonstrating that reason features provide helpful signals that generalize across tasks with shared features and semantics. However, ReCouPLe further improves generalization performance, especially the ReCouPLe-IC variant, which achieves the highest accuracy (78.9%) on the unseen *Push* task. This supports our hypothesis that projecting and regularizing preference explanations through causal directions allows the model to transfer reward structure more effectively.

Among our variants for orthogonal consistency loss, ReCouPLe-IC mostly outperforms ReCouPLe-EC in both in-distribution and novel tasks, only excluding *Pick-Place-Wall* tasks. Unlike our Maniskill experiment, in which preference queries consist of a pair of trajectories with a minimal difference in features other than the stated reasons, datasets in Meta-World contain noisy trajectories with different levels of

TABLE II: Reward accuracy comparison for ManiSkill 4-task setting (**RQ1**), averaged over 3 seeds.

Model	**In Distribution**				**Color Swapped**			
	Pick	**Push**	**Place**	**Pull**	**Pick**	**Push**	**Place**	**Pull**
Single Task								
BT (pick)	0.833	-	-	-	0.167	-	-	-
BT (push)	-	1.000	-	-	-	0.610	-	-
BT (place)	-	-	0.980	-	-	-	0.460	-
BT (pull)	-	-	-	1.000	-	-	-	0.053
Multi-Task								
BT-Multi	0.867	**1.000**	0.987	1.000	0.533	0.867	0.833	0.587
Multi-Task with Reasons								
RFP	0.867	**1.000**	0.993	**1.000**	0.807	0.967	0.947	0.833
ReCouPLe-EC ($\lambda_{\text{ratio}} = 0.2$)	**0.993**	**1.000**	0.993	**1.000**	**0.960**	**1.000**	**1.000**	0.973
ReCouPLe-EC ($\lambda_{\text{ratio}} = 0.2$)	0.980	**1.000**	0.993	**1.000**	**0.960**	**1.000**	**1.000**	0.980
ReCouPLe-IC ($\lambda_{\text{ratio}} = 0.2$)	0.973	**1.000**	**1.000**	**1.000**	0.940	**1.000**	0.993	**0.987**
ReCouPLe-IC ($\lambda_{\text{ratio}} = 0.4$)	0.960	**1.000**	**1.000**	**1.000**	**0.960**	**1.000**	0.993	**0.987**

TABLE III: Reward accuracy comparison for Meta-world setting (**RQ2**), averaged over 3 seeds.

Model		**Training Tasks**		**Novel Task**
	Pick-Place	***Pick-Place-Wall***	***Push-Wall***	***Push***
Single Task				
BT (*Pick-Place*)	0.872	-	-	-
BT (*Pick-Place-Wall*)	-	0.701	-	-
BT (*Push-Wall*)	-	-	0.786	-
Multi-Task				
BT-Multi	0.759	0.673	0.327	0.328
Multi-Task w/ Reasons				
RFP	0.798	0.535	0.811	0.695
ReCouPLe-EC	0.719	**0.678**	0.770	0.718
ReCouPLe-IC	**0.808**	0.653	**0.860**	**0.789**

optimality. Also, each query has a different reason behind its preference. Thus, it is less realistic to assume that reason-orthogonal components should remain identical across compared trajectories. In this setting, the strict equality constraint enforced by ReCouPLe-EC may overly penalize legitimate differences unrelated to the stated reason, thereby harming its performance. As shown in its performance, this makes ReCouPLe-IC more suited for real-world, noisy datasets with diverse reasons behind preferences.

VI. Conclusion and Future Work

We introduced ReCouPLe, a lightweight yet powerful framework that turns free-form natural language rationales into causal projection axes for preference-based reward learning. Across two complementary evaluations, ReCouPLe consistently mitigated causal confusion and exhibited strong zero-shot transfer to novel tasks where prior methods collapsed. These findings support two core takeaways:

1) **Causally robust preference learning with rationale.** Attaching a one-sentence rationale to each comparison supplies the missing causal signal, enabling the model to ground preferences in task-relevant features rather than incidental correlations. Especially, our method, ReCouPLe, uses the rationale as a guiding projection direction to separate out the part of the trajectory that explains the preference, ensuring the model focuses on the feature that actually matters. This leads to more robust preference learning under distribution shifts.
2) **Compositional rewards transfer across tasks.** Since the same reason can arise in multiple tasks, ReCouPLe leverages shared causal structure to transfer reward signals without additional preference data or language model fine-tuning.

Future directions. Our current work focuses on reward inference. A natural next step is to close the loop by training policies with our learned reward model with ReCouPLe. This would allow end-to-end evaluation in more complex manipulation tasks, both in simulation and in real-world settings. Another promising direction is to extend ReCouPLe to dialog-style rationales which could support more nuanced forms of causal supervision. Lastly, future work will also explore active querying strategies that can selectively query for a rationale only when causal uncertainty is high, improving data efficiency and human alignment.

References

[1] Erdem Bıyık, Malayandi Palan, Nicholas C Landolfi, Dylan P Losey, and Dorsa Sadigh. Asking easy questions: A user-friendly approach to active reward learning. In *Proceedings of the 3rd Conference on Robot Learning*, 2019.

[2] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[3] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[4] Yuchen Cui, Siddharth Karamcheti, Raj Palleti, Nidhya Shivakumar, Percy Liang, and Dorsa Sadigh. No, to the right: Online language corrections for robotic manipula-

tion via shared autonomy. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 93–101, 2023.

[5] Yinpei Dai, Jayjun Lee, Nima Fazeli, and Joyce Chai. Racer: Rich language-guided failure recovery policies for imitation learning. In *International Conference on Robotics and Automation (ICRA)*, 2025.

[6] Prasoon Goyal, Scott Niekum, and Raymond J. Mooney. Using natural language for reward shaping in reinforcement learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

[7] Joey Hejna and Dorsa Sadigh. Inverse preference learning: Preference-based rl without a reward function. *Advances in Neural Information Processing Systems*, 36: 18806–18827, 2023.

[8] Jongjin Park, Younggyo Seo, Chang Liu, Li Zhao, Tao Qin, Jinwoo Shin, and Tie-Yan Liu. Object-aware regularization for addressing causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 34:3029–3042, 2021.

[9] Andi Peng, Yuying Sun, Tianmin Shu, and David Abel. Pragmatic feature preferences: Learning reward-relevant preferences from human input. In *International Conference on Machine Learning (ICML)*, 2024.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[11] Lucy Xiaoyang Shi, Zheyuan Hu, Tony Z. Zhao, Archit Sharma, Karl Pertsch, Jianlan Luo, Sergey Levine, and Chelsea Finn. Yell at your robot: Improving on-the-fly from language corrections. In *Robotics: Science and Systems*, 2024.

[12] Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-kai Chan, et al. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*, 2024.

[13] Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca Dragan, and Daniel S Brown. Causal confusion and reward misidentification in preference-based reward learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[14] Zhaojing Yang, Miru Jun, Jeremy Tien, Stuart Russell, Anca Dragan, and Erdem Bıyık. Trajectory improvement and reward learning from comparative language feedback. In *8th Annual Conference on Robot Learning*, 2024.

[15] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. *arXiv preprint arXiv:2107.09645*, 2021.

[16] J-Anne Yow, Neha Priyadarshini Garg, Manoj Ramanathan, and Wei Tech Ang. Extract – explainable trajectory corrections from language inputs using textual description of features. *arXiv preprint arXiv:2401.03701*, 2024.

[17] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*. PMLR, 2020.