PRIMT: Preference-based Reinforcement Learning with Multimodal Feedback and Trajectory Synthesis from Foundation Models

Ruiqi Wang, Dezhong Zhao, Ziqin Yuan, Tianyu Shao,

Guohua Chen, Dominic Kao, Sungeun Hong, Byung-Cheol Min

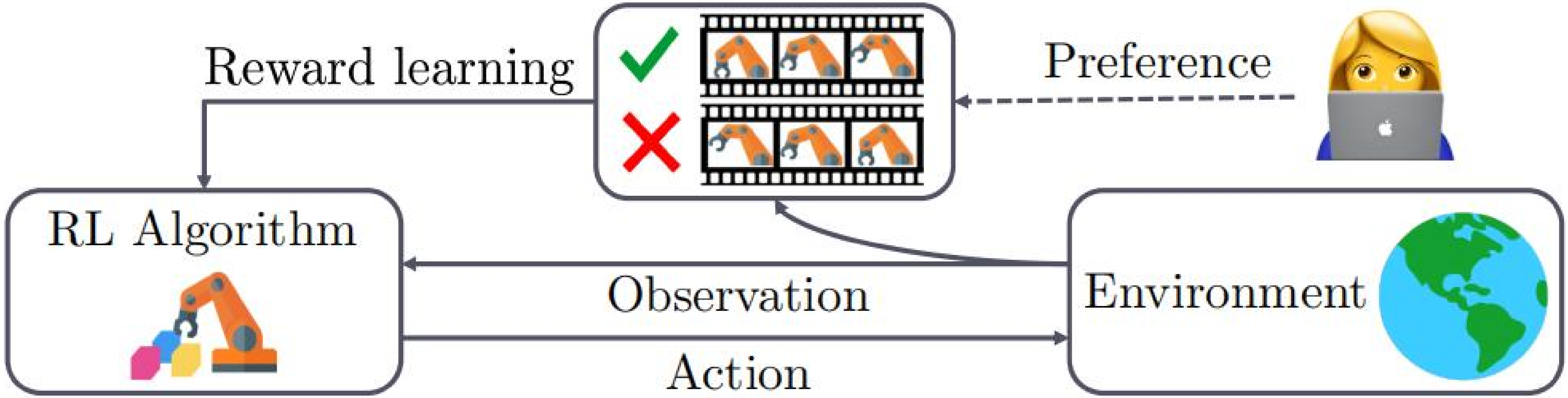
Purdue University, Indiana University Bloomington,

UIUC, BUCT, Sungkyunkwan University

NEURAL INFORMATION PROCESSING SYSTEMS

Preference-based Reinforcement Learning

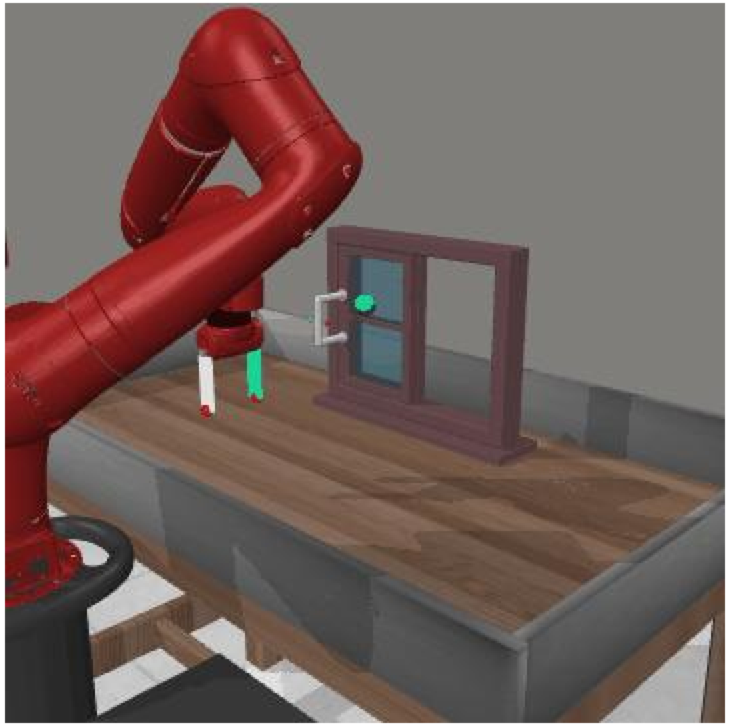
Preference-based RL (PbRL) has emgered as a promising paradigm for teaching robots complex behaviors without reward engineering.



(Lee et al., 2022)

Preference-based Reinforcement Learning

However, the reliance on extensive human labels prevents its scalability

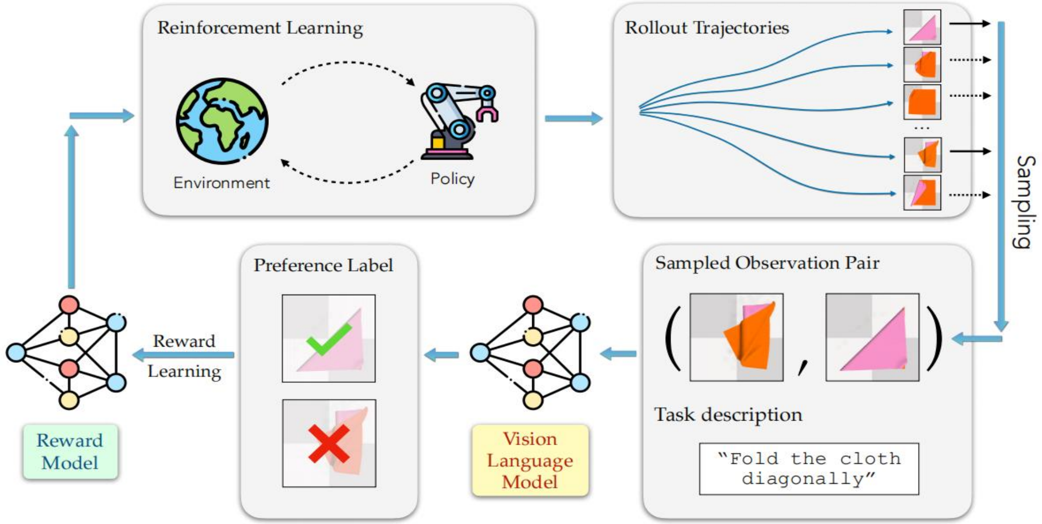


~10000 rounds of human preference are needed!

~84 hours

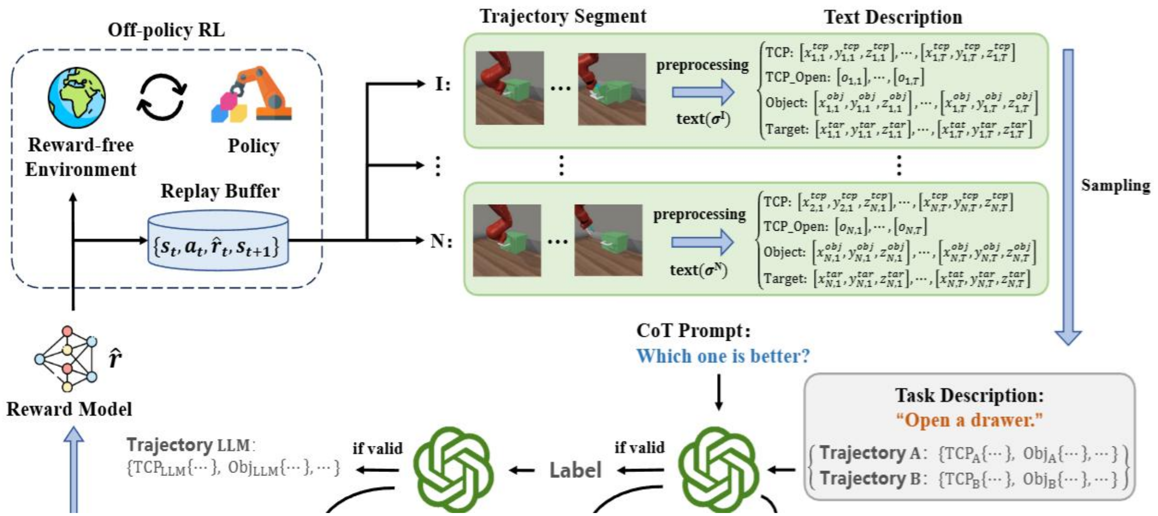
Foundation Models for Synthetic Feedback

VLM agents giving preference labels by analyzing state images



(Wang et al., 2024)

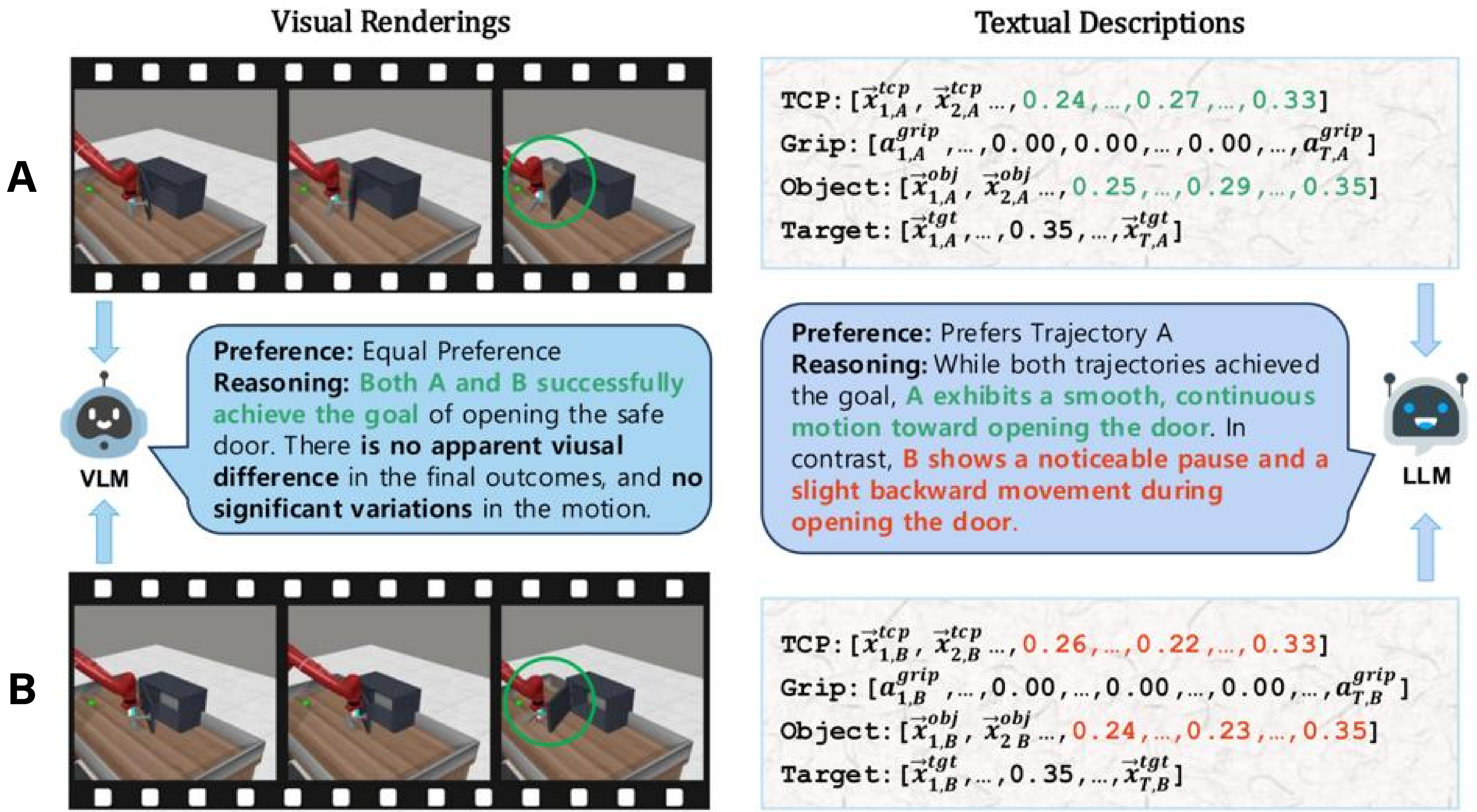
LLM agents giving preference label by analyzing textual trj descriptions



(Tu et al., 2025)

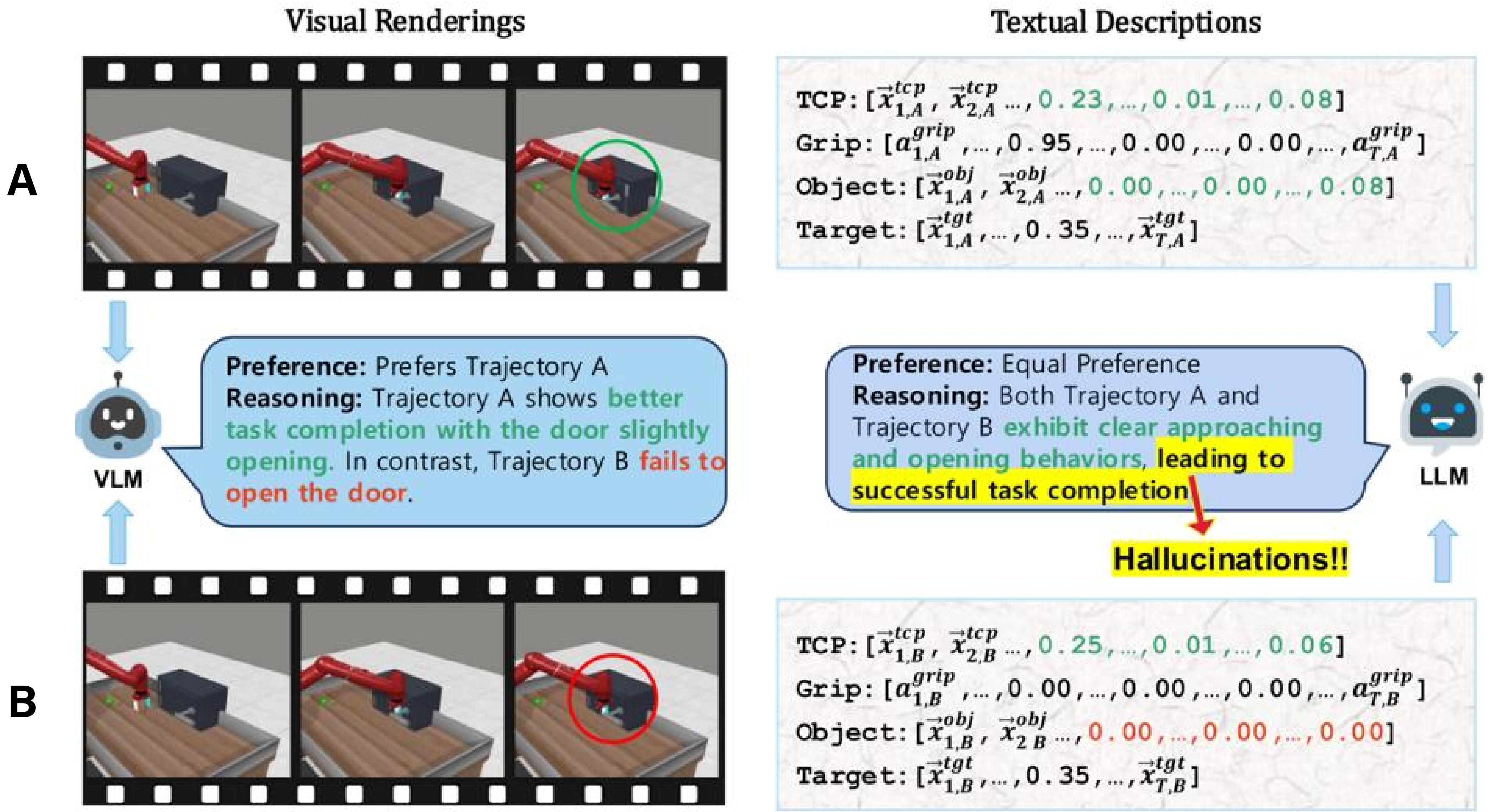
However, existing work still faces challenges in ensuring reliable synthetic feedback due to the **single-modal evalution patterns.**

Limitations of Single-Modal Evaluation



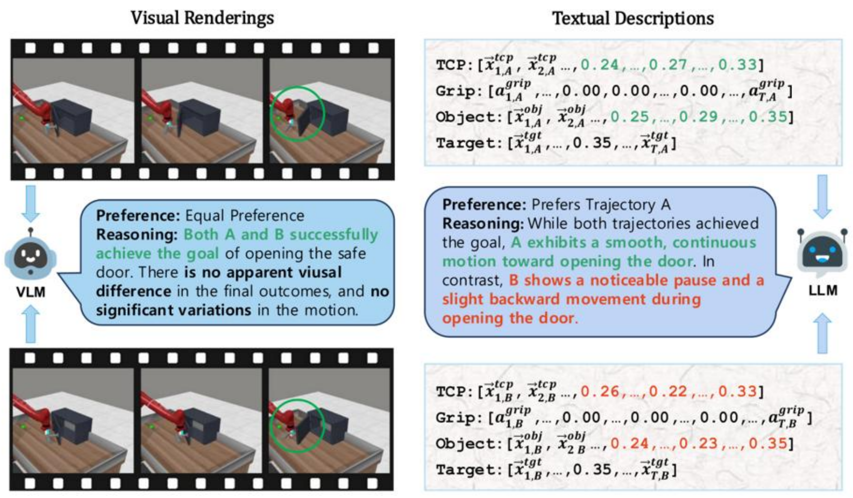
(Answers are from gpt-4o)

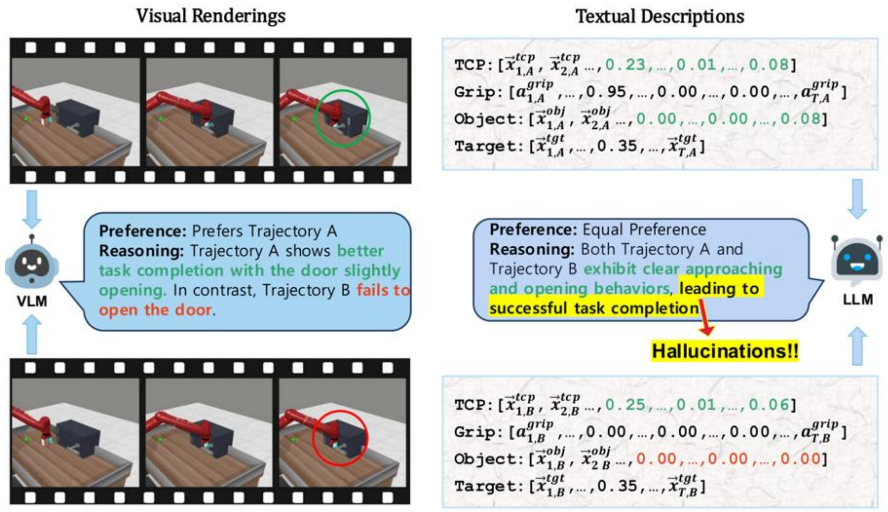
Limitations of Single-Modal Evaluation



(Answers are from gpt-4o)

Limitations of Single-Modal Evaluation





Visual-modal-only (VLMs) → reliable **spatial grounding and key events assessment**, but limited ability to interpret temporal progression or subtle motion dynamics.

Textual-modal-only (LLMs) → good **temporal and logical reasoning**, but often hallucinate or miss fine-grained spatial interactions and key events.

Calls for **multimodal evaluation** methods for reliable synthetic feedback!

PbRL Inherent Challenges

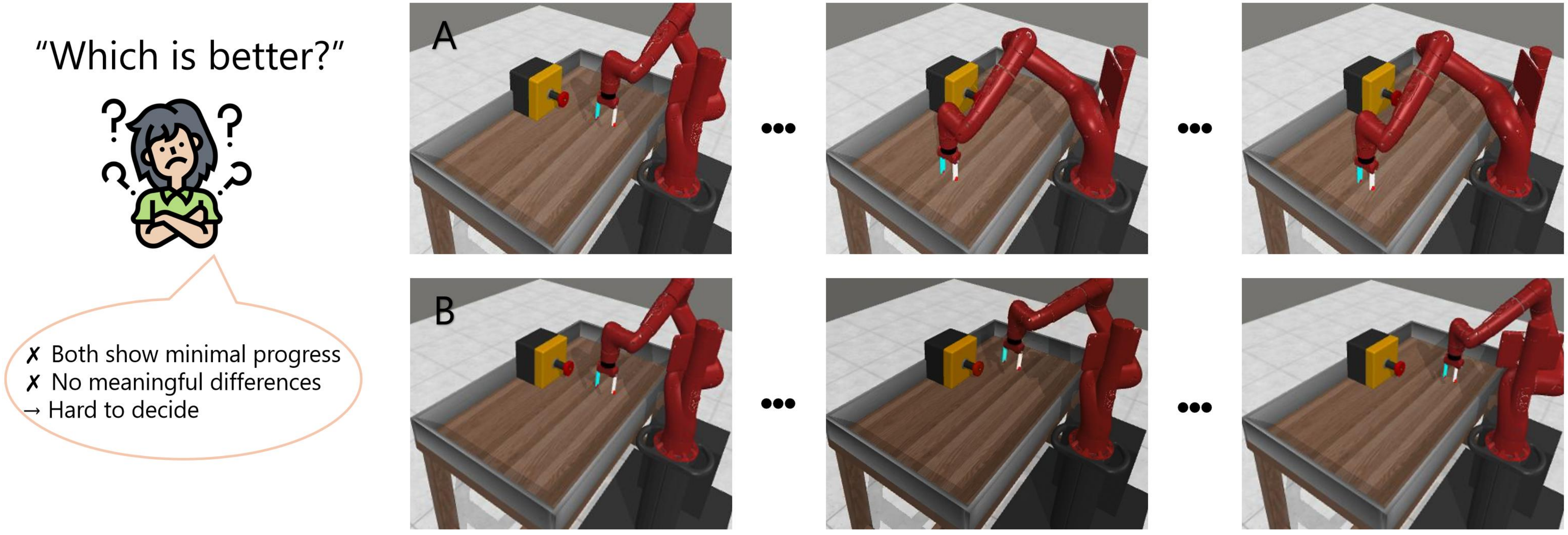
Even if feedback from FMs reaches human-expert-level quality, PbRL still faces two inherent challenges in reward learning.

- Query Ambiguity
- Preference Credit Assignment Uncertainty

PbRL Inherent Challenges - Query Ambiguity

Early trajectories from randomly initialized policies are uniformly low quality lacking meaningful task differences

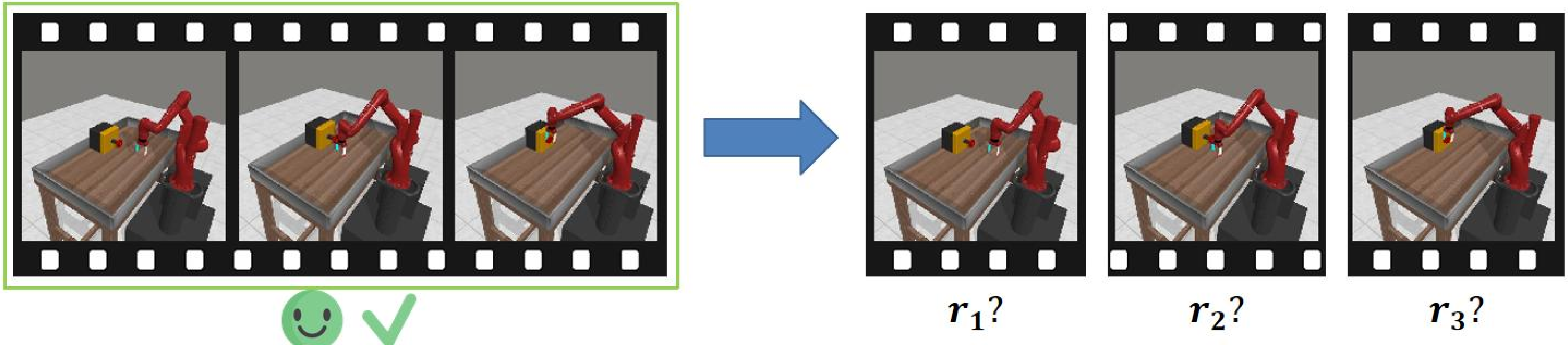
→ cannot provide informative comparisons



PbRL Inherent Challenges - Preference Credit Assignment Uncertainty

Preferences are given at **trajectory level**, but desired reward models operate at **state-action level**

→ uncertainty in attributing preference credit to specific steps in the trajectory, impairing the alignment of the learned reward model

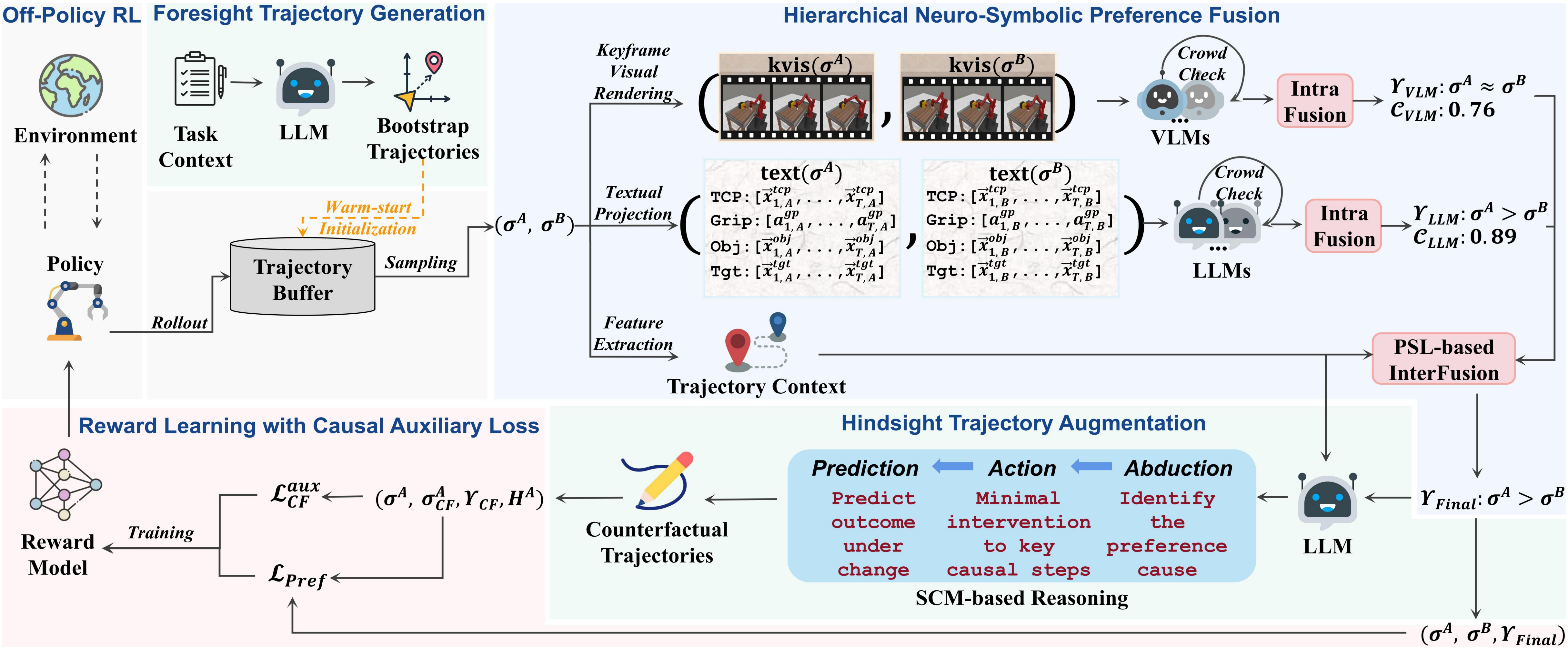


Our Goal

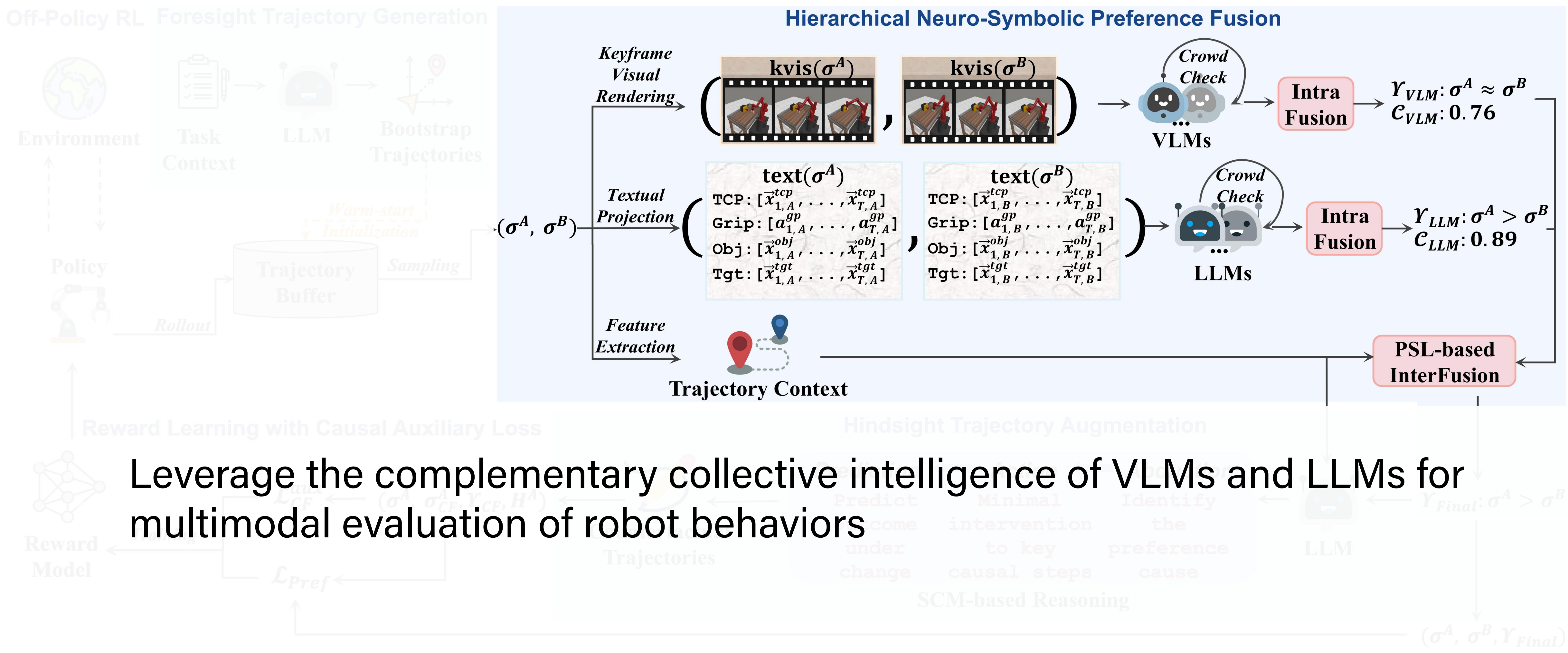
Enable scalable, efficient, and zero-shot PbRL with FMs

- Improve the quality and reliability of synthetic feedback
- Reduce query ambiguity
- Enhance preference credit assignment

Method - PRIMT



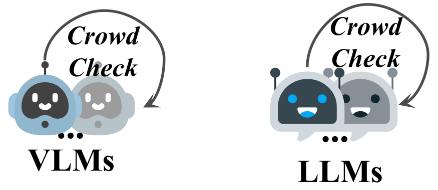
PRIMT - Multimodal Preference Generation and Fusion



Leverage the complementary collective intelligence of VLMs and LLMs for multimodal evaluation of robot behaviors

PRIMT - Hierarchical Preference Fusion

Intra-modal Fusion



i) query LLM or VLM multiple times with randomly permuted trajectory orderings

ii) get final intra-modal labels via major voting

$$\Upsilon_M = \underset{l \in \{-1,0,1\}}{\operatorname{argmax}} \sum_{k=1}^{K} \mathbb{I}(\Upsilon_M^{(k)} = l)$$

iii) calculate label confidence

$$\bar{\mathcal{C}}_M = \frac{1}{N} \sum_{k=1}^{K} \mathcal{C}_M^{(k)} \cdot \mathbb{I}(\Upsilon_M^{(k)} = \Upsilon_M); \;\; \dot{\mathcal{C}}_M = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}(\Upsilon_M^{(k)} = \Upsilon_M)$$

$$\mathcal{C}_M = \alpha \cdot \bar{\mathcal{C}}_M + (1 - \alpha) \cdot \dot{\mathcal{C}}_M$$

PRIMT - Hierarchical Preference Fusion

Inter-modal Fusion

Consider intra-modal uncertainty, inter-modal conflicts, and trajectory context that reflects the relative difficulty of visual v.s. textual evaluation.

To model latent dependencies across these factors, we employ the Probabilistic Soft Logic (PSL) framework

PRIMT - Hierarchical Preference Fusion

Defined Rules in PSL:

i) Agreement Rule:

$$\forall \Upsilon, M: \ \texttt{IsAgree}(\Upsilon) \land \texttt{ConfHigh}(M) \rightarrow \texttt{FinalLabel}(\Upsilon)$$

ii) Conflict Resolution Rules:

$$\forall \Upsilon: \ \neg\texttt{IsAgree}(\Upsilon) \land \texttt{VLMLabel}(\Upsilon) \land \texttt{ConfHigh}(\mathrm{VLM}) \land \texttt{VDHigh} \land \rightarrow \texttt{FinalLabel}(\Upsilon)$$

$$\forall \Upsilon: \ \neg\texttt{IsAgree}(\Upsilon) \land \texttt{LLMLabel}(\Upsilon) \land \texttt{ConfHigh}(\mathrm{LLM}) \land \texttt{TDHigh} \rightarrow \texttt{FinalLabel}(\Upsilon)$$

iii) Indecision Rule:

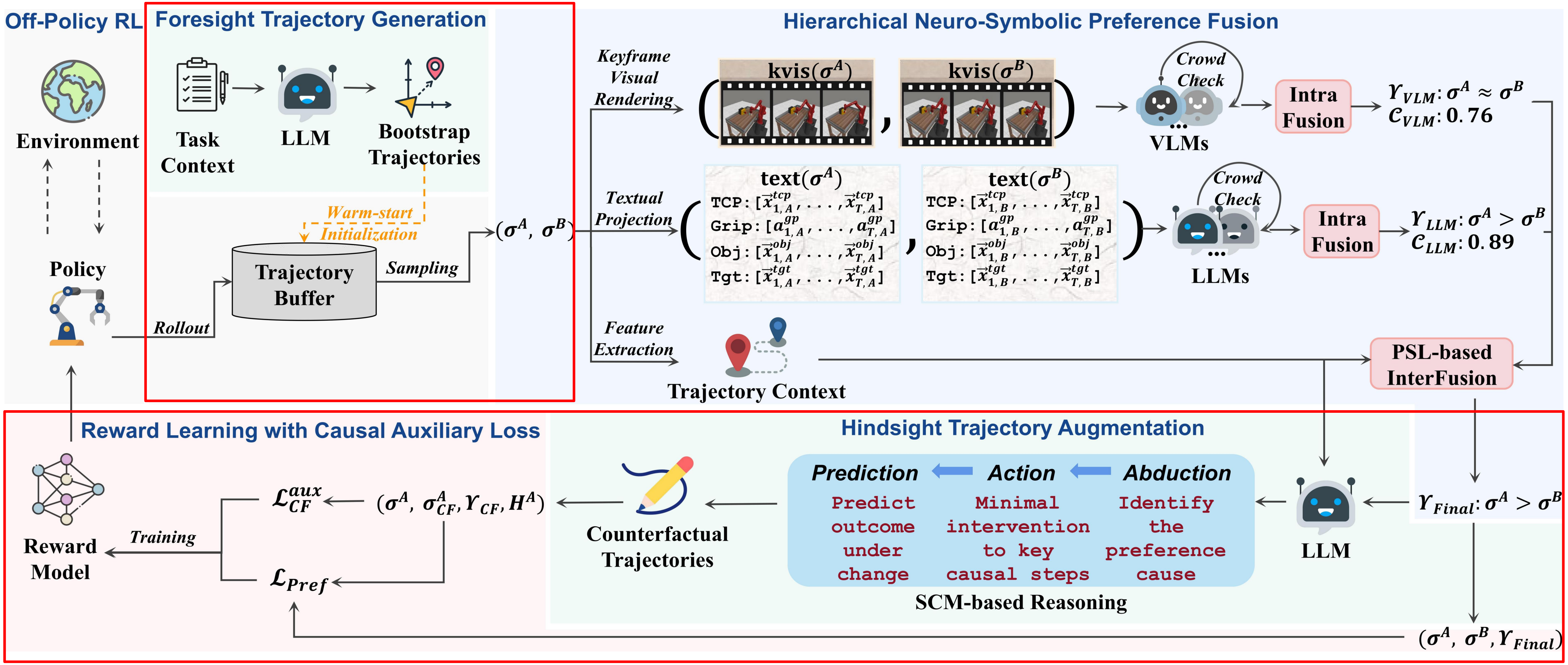
$$\neg\texttt{ConfHigh}(\mathrm{VLM}) \land \neg\texttt{ConfHigh}(\mathrm{LLM}) \rightarrow \texttt{FinalLabel}(-1)$$

PRIMT - Hierarchical Preference Fusion

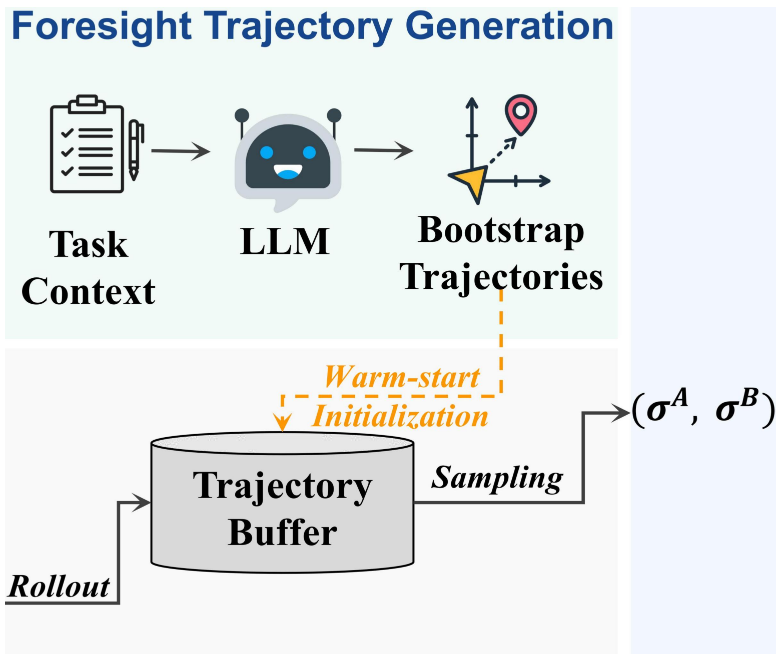
PSL turns these rules into one or one more hinge-loss penalties and solves an online convex optimization to find the label that best fits all rules

$$Y^* = \arg\min_{Y} \sum_{i=1}^{m} w_i, \phi_i(Y, X) \quad \text{s.t.} \sum_{\Upsilon \in \{-1,0,1\}} \texttt{FinalLabel}(\Upsilon) = 1$$

PRIMT - Bidirectional Trajectory Synthesis



PRIMT - Foresight Trajectory Generation

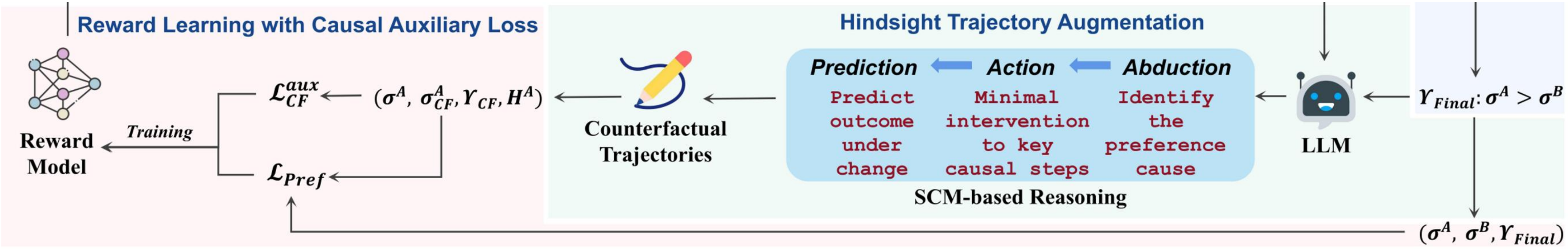


Before training, we proactively generate diverse, task-aligned trajectory samples using LLMs:

high-level multi-step plan → code for each step → rollout trajectories under varied conditions

serve as ***informative preference anchors*** rather than optimal demonstrations, making early comparisons easier when combined with query sampling strategies

PRIMT - Hindsight Trajectory Augmentation

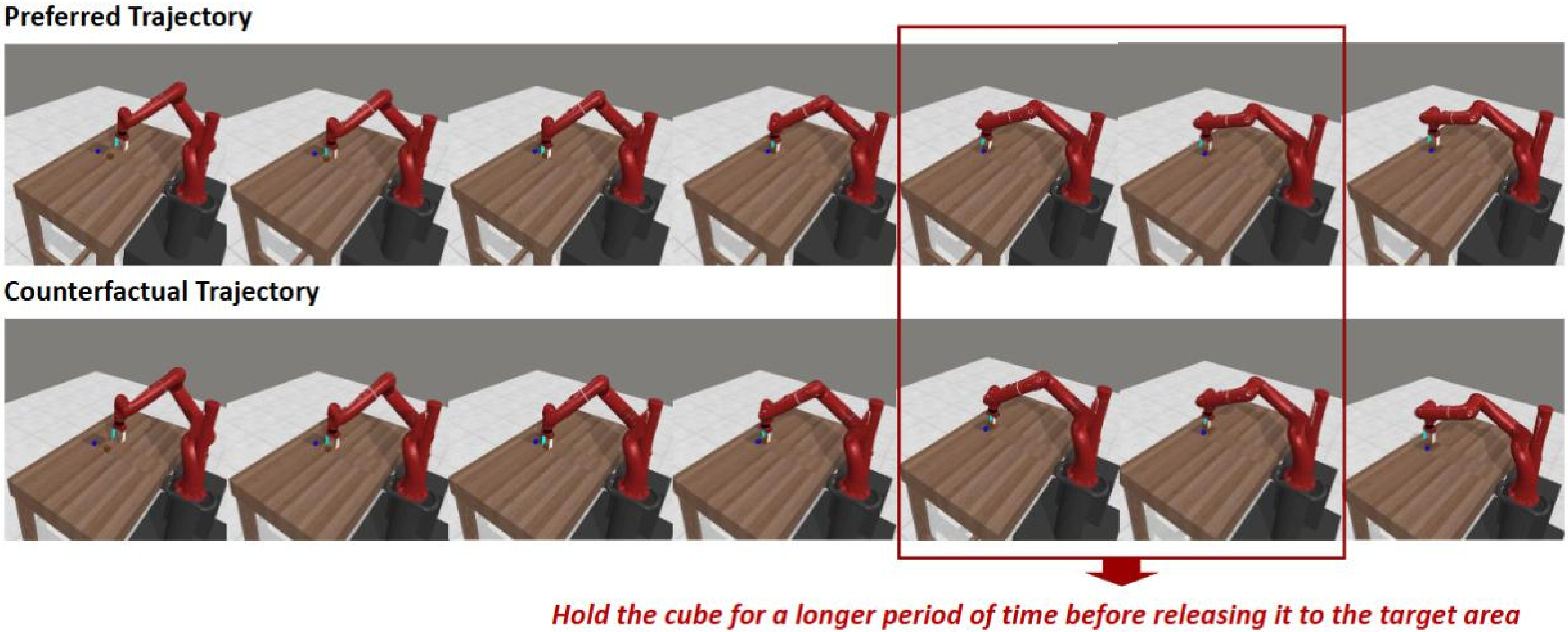


Once we have a clear preference for a trajectory pair, we perform the **hindsight counterfactual reasoning** by asking:

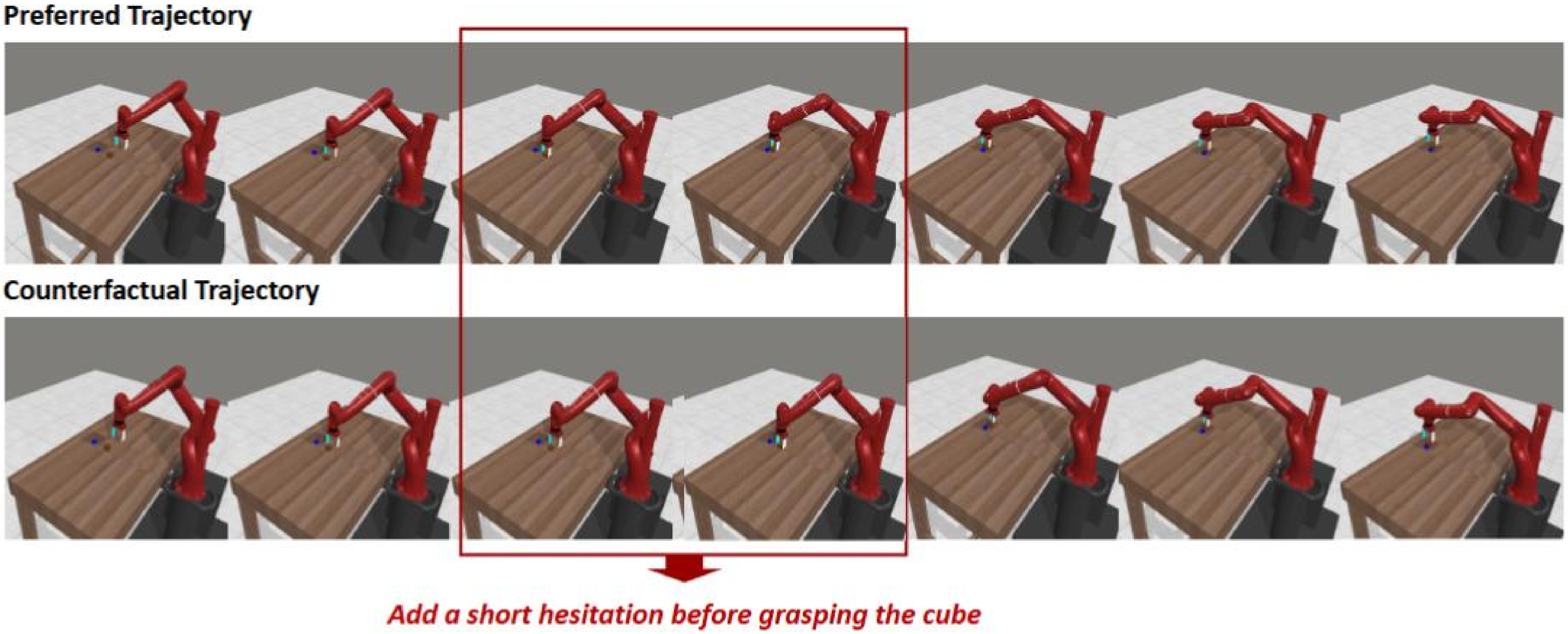
"What minimal change would make the preferred trajectory less preferred?"

identify key steps in the preferred trj → minimal edits → create counterfactuals

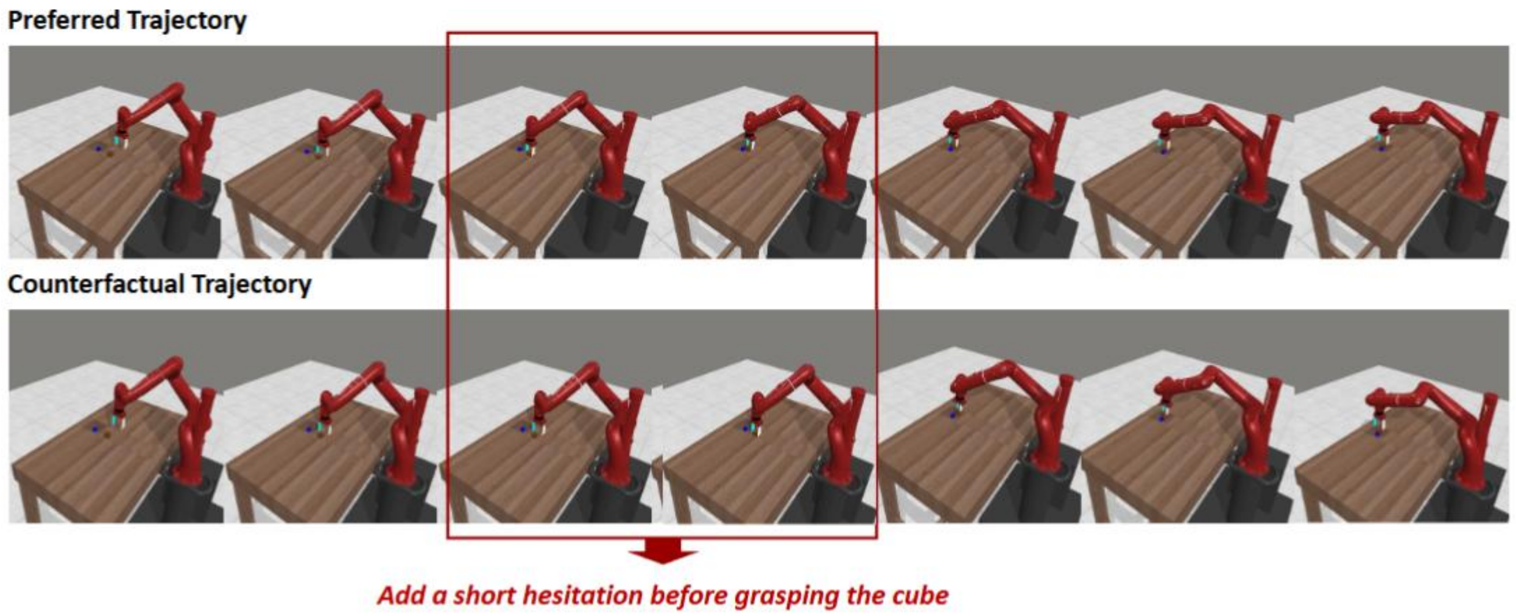
PRIMT - Hindsight Trajectory Augmentation



PRIMT - Hindsight Trajectory Augmentation



PRIMT - Auxiliary Causal Loss

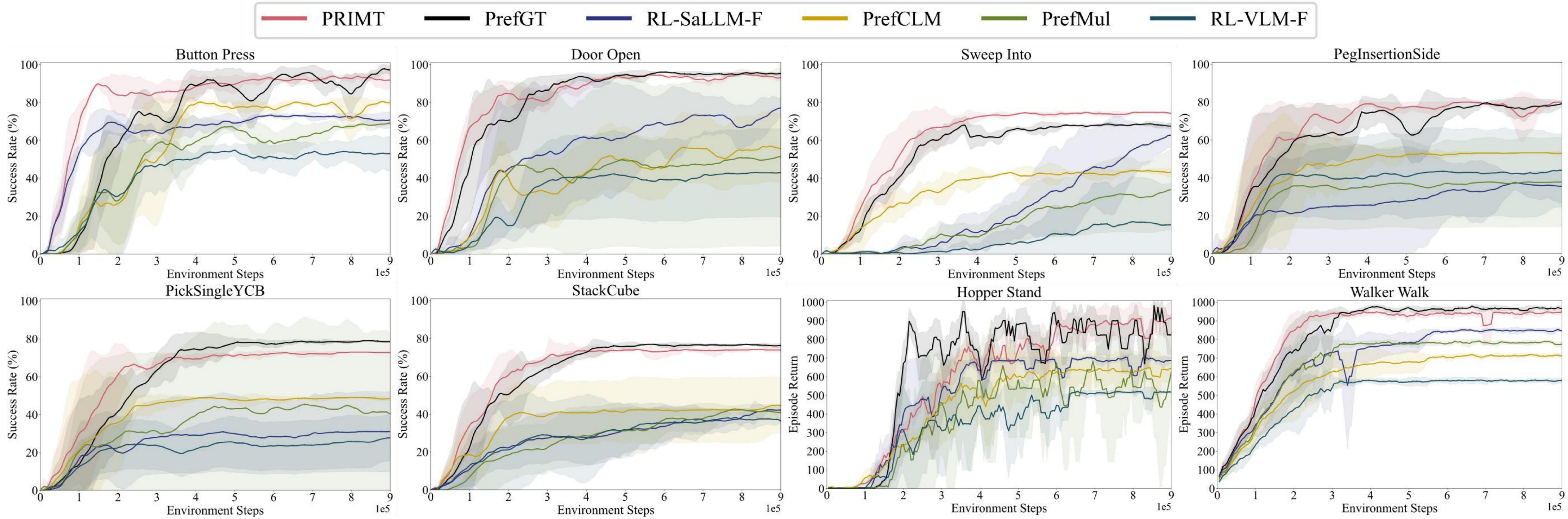


$$\mathcal{L}_{\text{cf}}^{\text{aux}} = \underbrace{\sum_{t=1}^{T} H_t \cdot \log\left(1 + \exp\left(r_\psi(s_t^{cf}) - r_\psi(s_t^{*})\right)\right)}_{\text{i) causal contrast loss}} + \underbrace{\sum_{t=1}^{T} (1 - H_t) \cdot \left\| r_\psi(s_t^{*}) - r_\psi(s_t^{cf}) \right\|_2^2}_{\text{ii) reward consistency loss}}$$

i) contrast rewards at edited (causal) steps

ii) enforce consistency on unedited regions

Evaluation - Task Performance on Manipulation and Locomotion Tasks

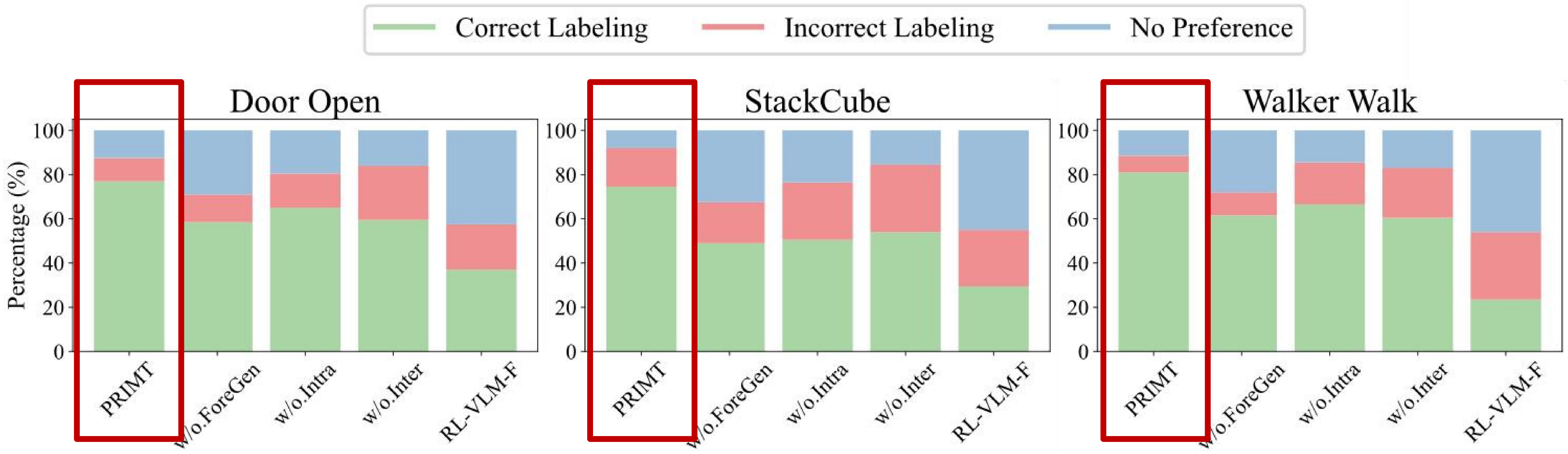


- PRIMT (red) outperforms all FM-based baselines and the naïve multimodal fusion

- PRIMT matches the oracle PrefGT (black, with ground-truth preference labels) and even surpasses it on 2 of 8 tasks

- PRIMT learns more efficiently in early training -> less query ambiguity

Evaluation - Synthetic Feedback Quality



- Improved label accuracy (green)
- Reduced indecision and query ambiguity (blue)

Evaluation - Reward Alignment

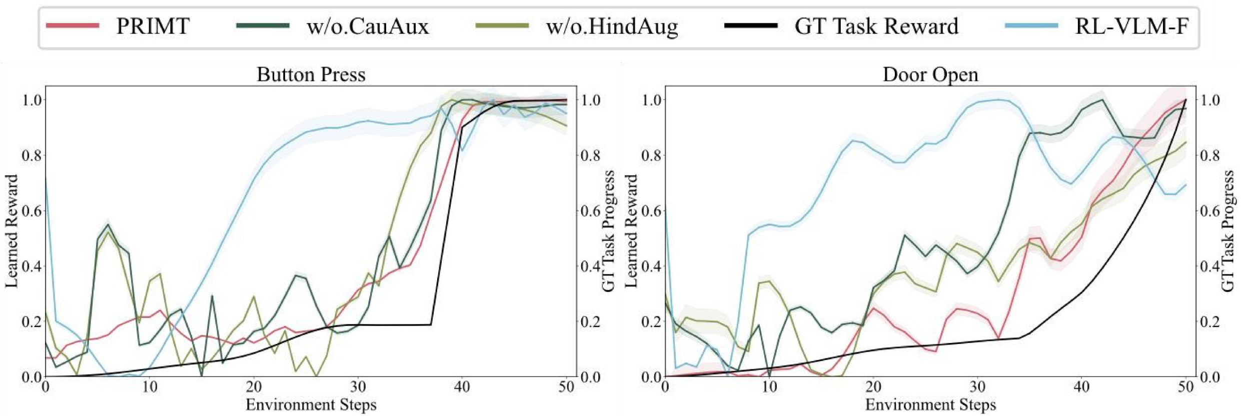


Table 5: R^2 Coefficient Analysis (Reward Alignment with Ground Truth).

Task	PRIMT	w/o CauAux	w/o HindAug	RL-VLM-F
PegInsertionSide	0.56	0.28	0.23	0.37
PickSingleYCB	0.84	0.01	0.34	-0.05
StackCube	0.78	-1.31	-2.28	-1.50
ButtonPress	0.87	0.68	0.53	-0.61
DoorOpen	0.64	-1.19	0.15	-4.72
SweepInto	0.88	0.83	0.73	-0.27
WalkerWalk	0.33	0.19	0.02	-2.29

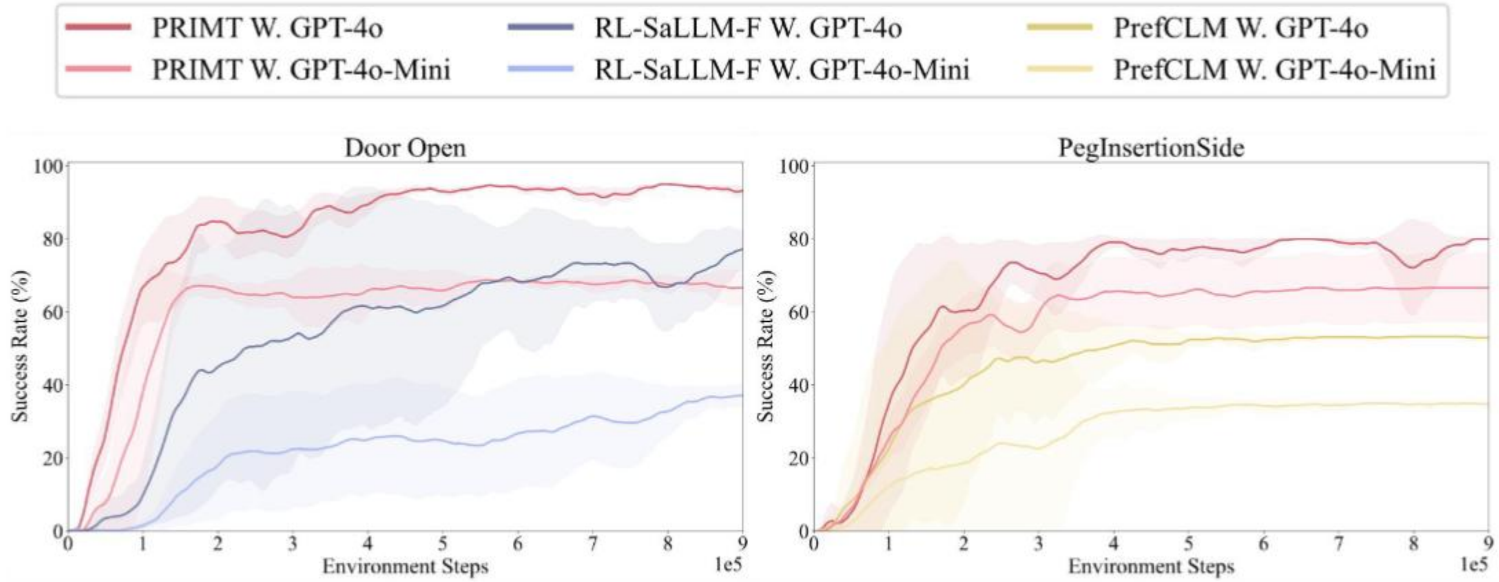
- Aligned more closely with ground-truth rewards and task progress

-> Enhanced Preference Credit Assignment

Evaluation - Cost-Performance Efficiency

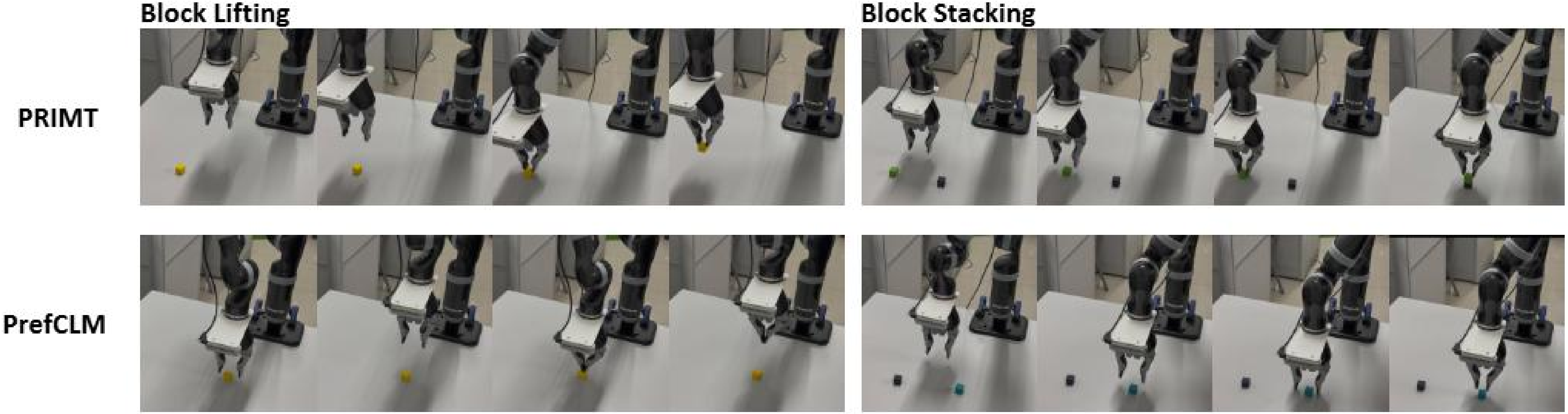
Baseline	Cost	Time	Performance	Efficiency†
	MetaWorld / ManiSkill / DMC			
vs RL-VLM-F	+43%/+39%/+47%	+58%/+43%/+69%	+95%/+117%/+68%	2.0×
vs Sa-LLM-F	+45%/+44%/+38%	+31%/+30%/+46%	+32%/+109%/+19%	1.4×
vs Human (PrefGT)	−92%/−95%/−92%	—/—/—	−1%/−%/−%	47×

- [†] *Efficiency = Average performance gain / Average resource increase*
- *Performance is measured using the final return from the learning curves presented in Figure 2*
- *Values shown are relative changes across MetaWorld / ManiSkill / DMC environments, respectively*



Improved cost–performance efficiency and increased robustness to weaker FMs

Evaluation - Real-World Experiments



Thank you !

Poster Section

Dec 3rd, Wednesday

4:30 p.m. - 7:30 p.m. PST

In Room: Exhibit Hall C,D,E

Poster Location: #**2209**

Ruiqi Wang

wang5357@purdue.edu

Purdue University

Project Website

Contact Information