NEURAL INFORMATION
PROCESSING SYSTEMS

RLLAB
http://rllab.snu.ac.kr

Sequential Preference Ranking for Efficient Reinforcement Learning from Human Feedback

NeurIPS 2023 Accepted

Poster Session: Wed 13 Dec 8:45 am -10:45 a.m. PST
Great Hall & Hall B1+B2 #1300

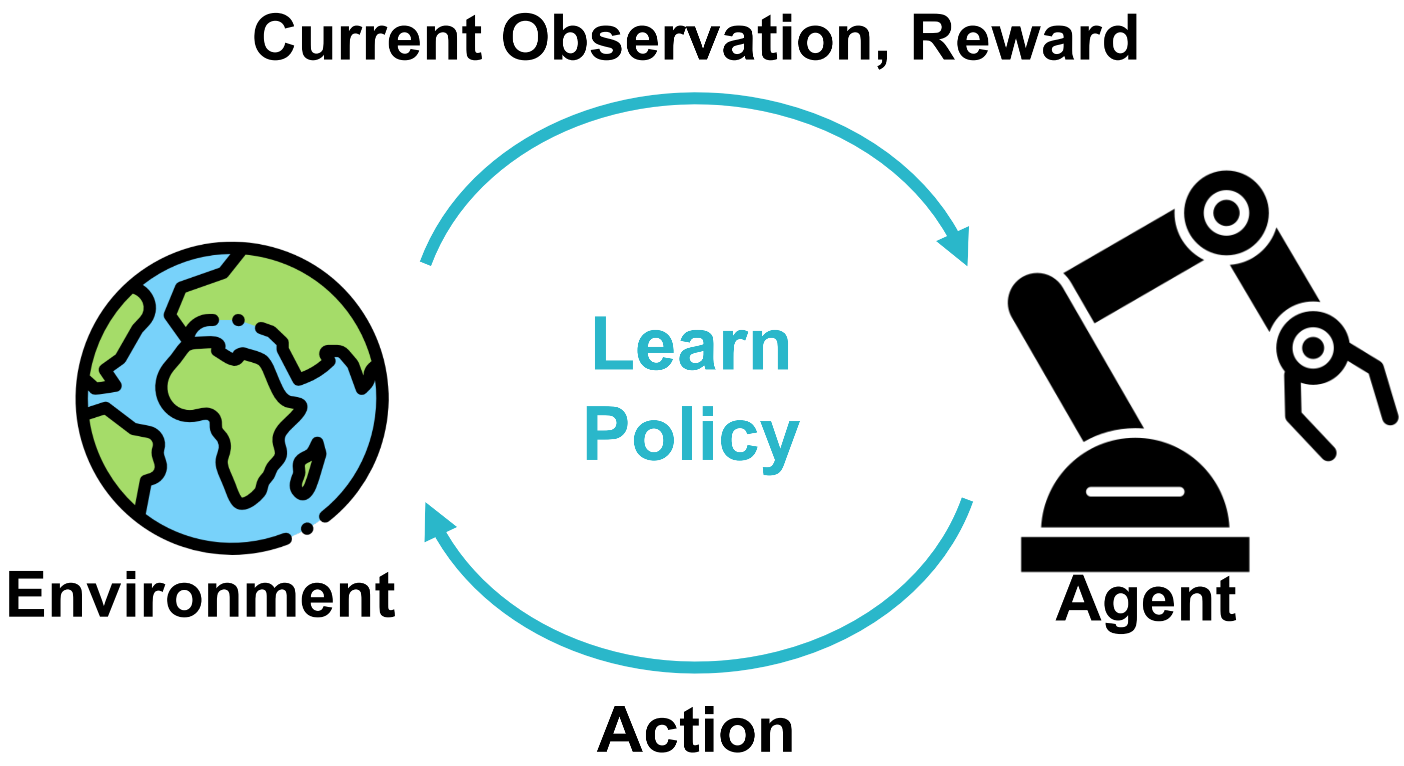
Minyoung Hwang[1], Gunmin Lee[1], Hogun Kee[1], Chan Woo Kim[2], Kyungjae Lee[2*], Songhwai Oh[1*]

[1]Electrical and Computer Engineering and ASRI, Seoul National University
[2]Department of Artificial Intelligence, Chung-Ang University

Contents

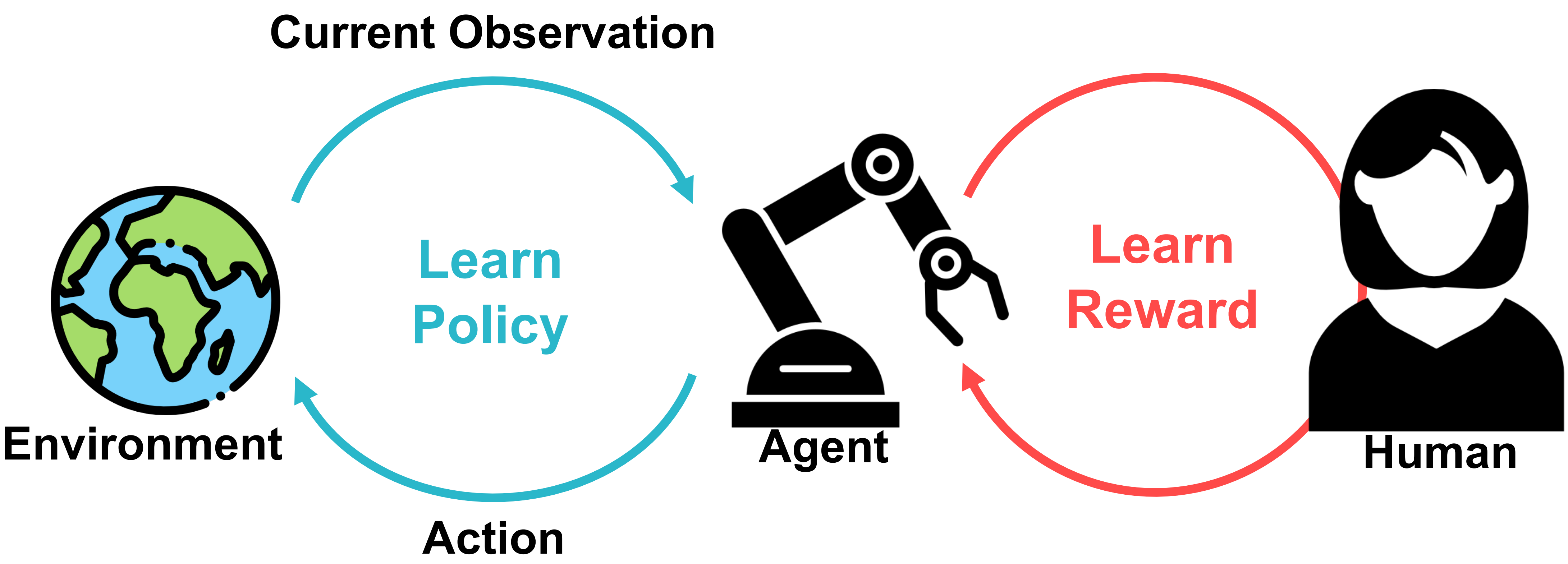
1) Reinforcement Learning from Human Feedback (RLHF)
2) SeqRank
3) Theoretical Analyses
4) Simulation Experiments: DMControl
5) Simulation Experiments: Meta-World
6) Experiments: Real Human Feedback
7) Real Robot Experiments

Problems in Reinforcement Learning



Designing a suitable reward function in RL often requires **task-specific prior knowledge**.
Additionally, we need **sufficient time to design** the reward function to capture the true task objective.

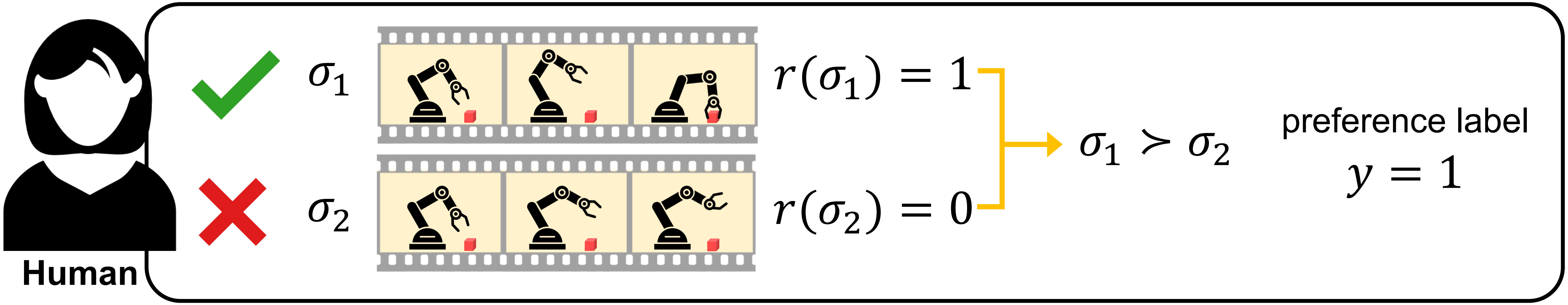
Reinforcement Learning from Human Feedback



Reinforcement learning from human feedback (RLHF) directly learns from **human's preferences** without the need for a hand-crafted reward function.

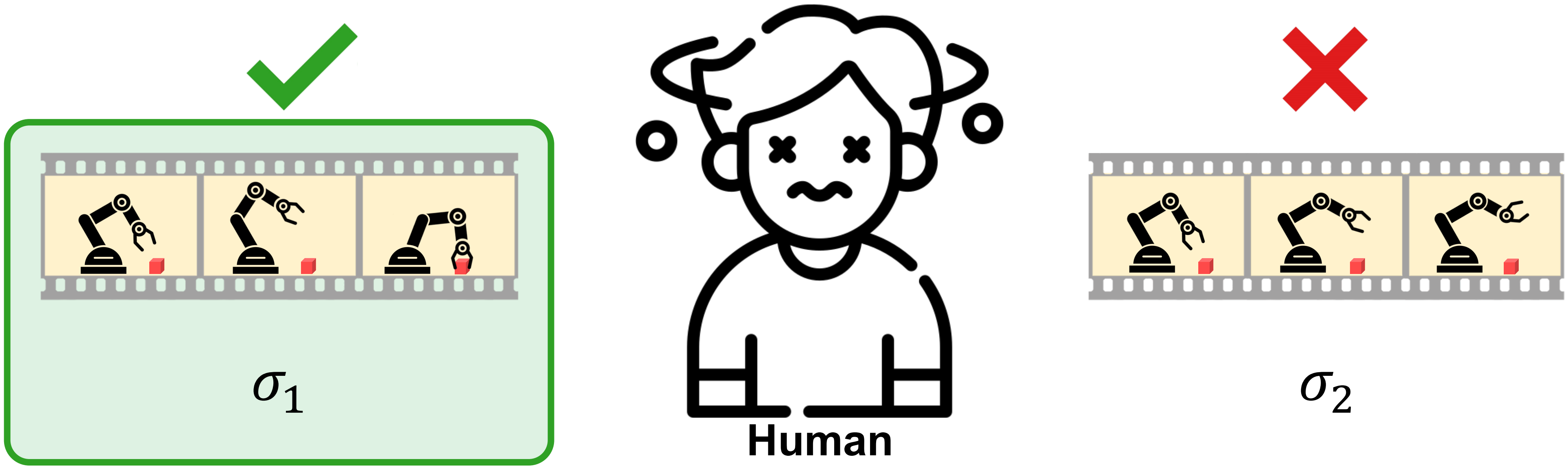
Reinforcement Learning from Human Feedback

A conventional way to learn a reward function in RLHF is **pairwise comparison**.



Reinforcement Learning from Human Feedback

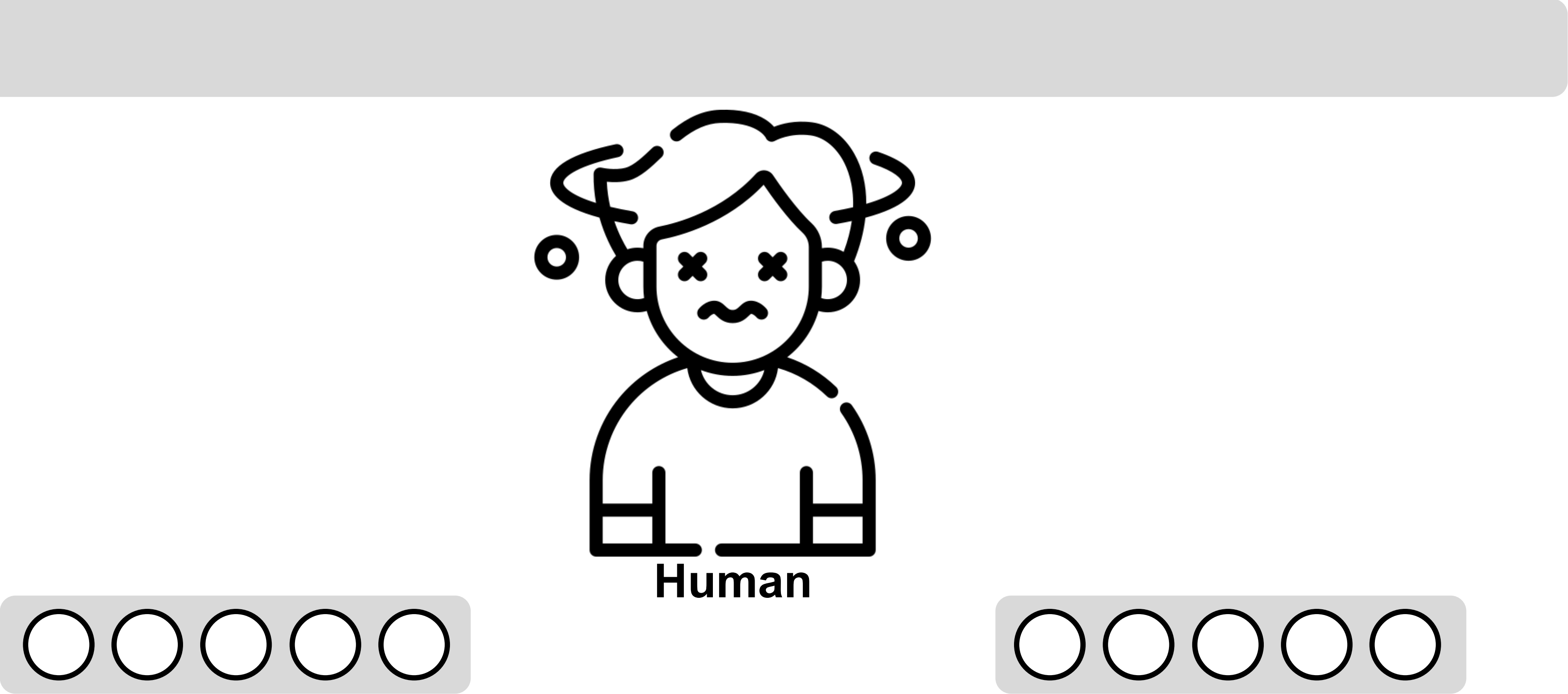
Using **pairwise comparison**, the agent queries a human to compare two different trajectories.



A critical limitation of pairwise comparison is **high cognitive load**.
A human must remember $2N$ different trajectories to determine N preferences.

Reinforcement Learning from Human Feedback

Using **pairwise comparison**, the agent queries a human to compare two different trajectories.



The feedback efficiency is also fixed as a standardized level, 1.

$$feedback\ efficiency := \frac{\#\ trajectory\ pairs}{\#\ number\ of\ feedbacks}$$

preferred
not preferred

SeqRank

We propose a novel RLHF framework called **SeqRank**.
Our method uses **sequential preference ranking** to enhance the feedback efficiency and reduce human's labeling effort.

Human

If A is preferred over B and B is preferred over C, A is preferred over C.

$$A \succ B \land B \succ C \rightarrow A \succ C$$

The key idea of our approach is to utilize the **preference relationships** of the previous trajectory pairs.
Bringing the nature of **transitivity in human preferences**, we can **augment** preference data.

SeqRank

Our method **samples trajectories in a sequential manner** by iteratively selecting
a **defender** from the set of previously chosen trajectories $\mathcal{K}$ and
a **challenger** from the set of unchosen trajectories $\mathcal{U} \setminus \mathcal{K}$.

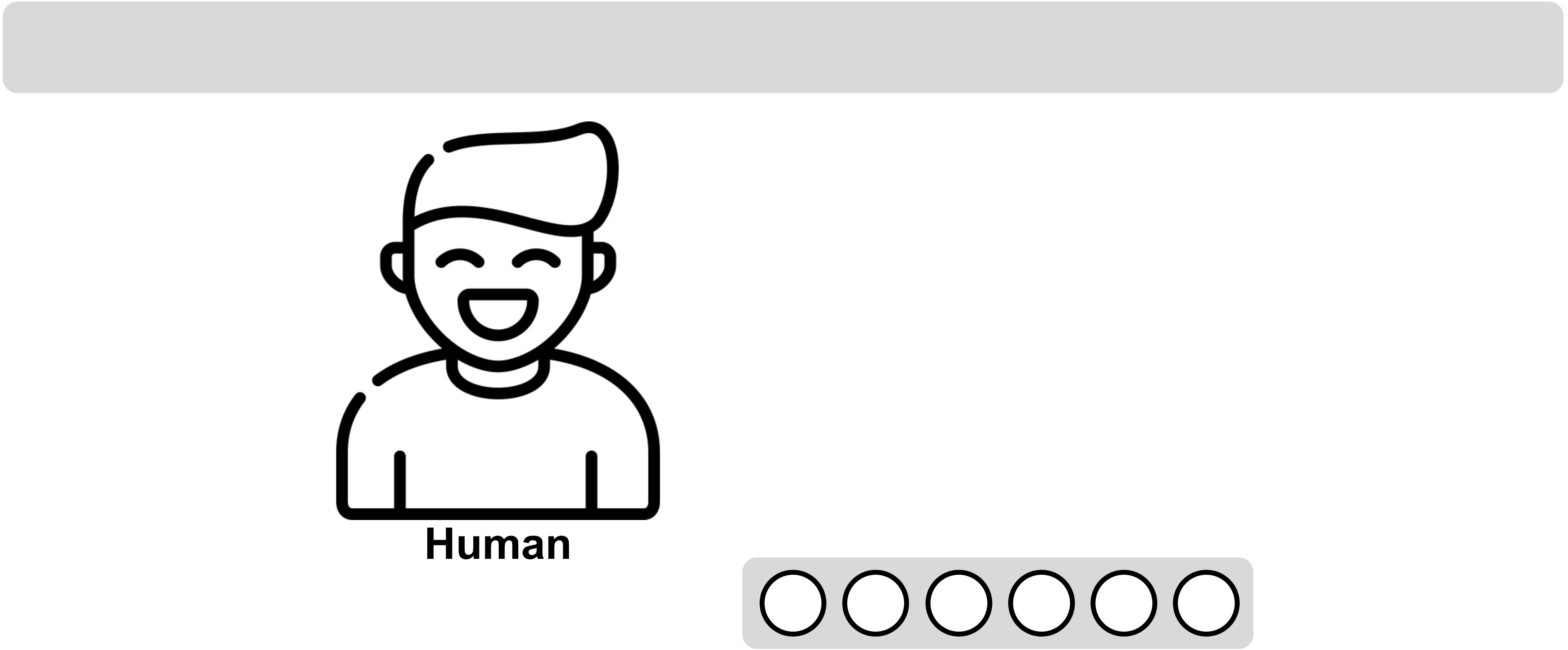
(1) Sequential Pairwise Comparison
defender = most recently sampled trajectory

(2) Root Pairwise Comparison
defender = previously most preferred trajectory

Specifically, we propose two trajectory comparison methods with different defender sampling strategies.

SeqRank

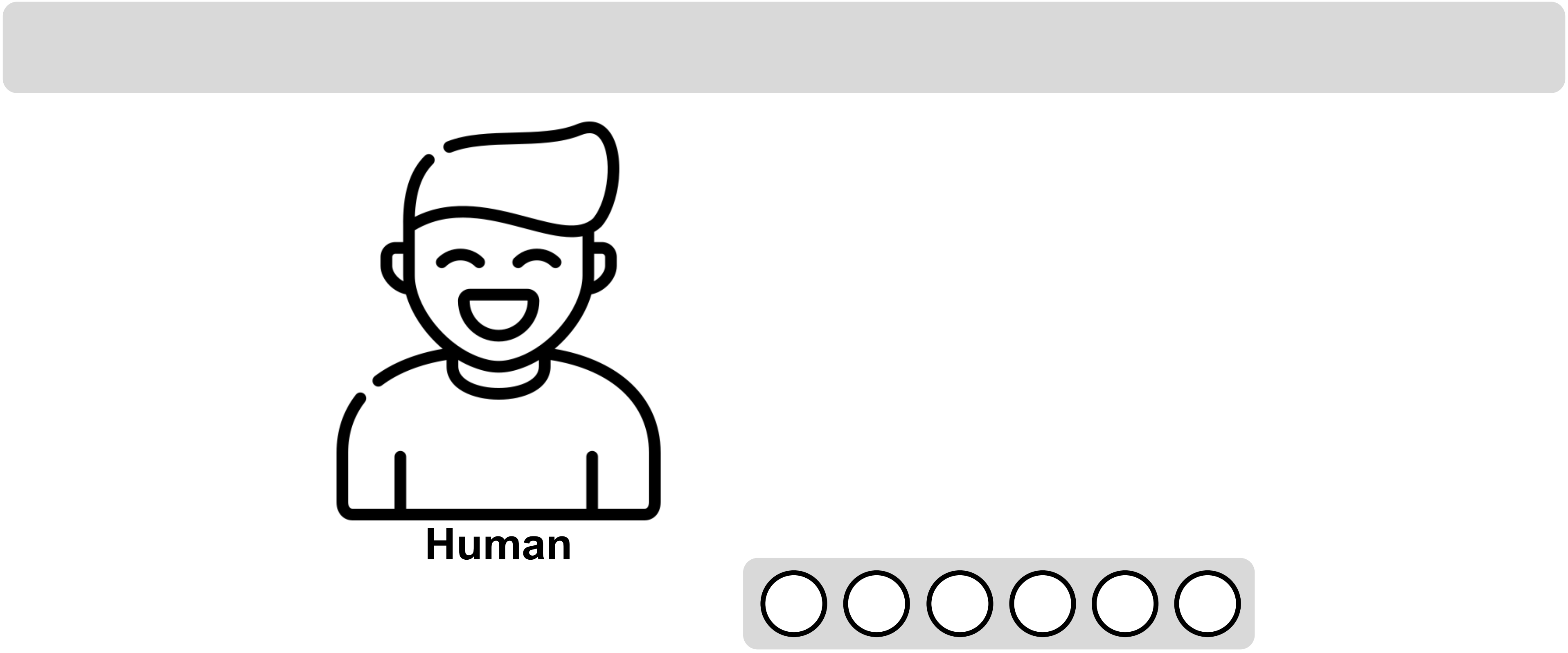
Sequential pairwise comparison selects the **most recently sampled trajectory** as the defender.



preferred
not preferred

SeqRank

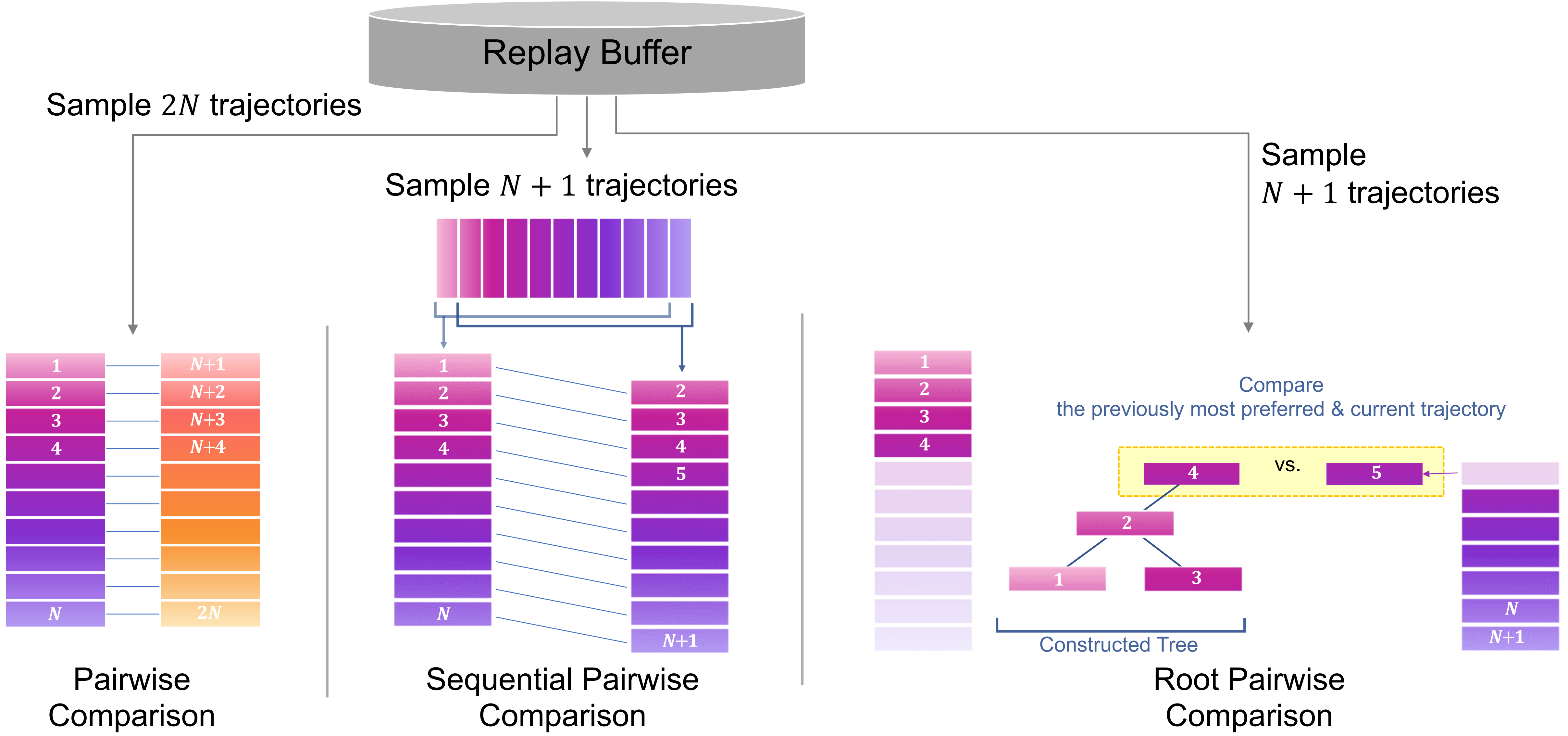
Root pairwise comparison selects the **previously most preferred trajectory** as the defender.



preferred
not preferred

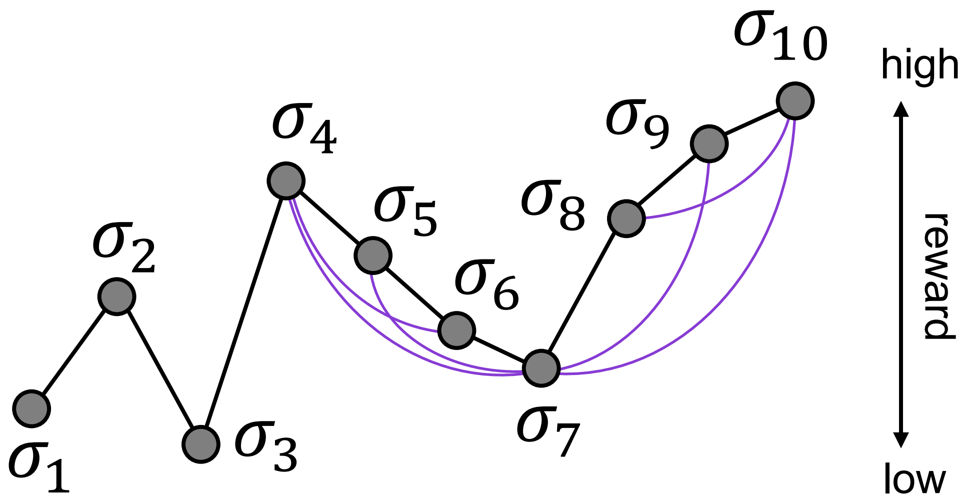
SeqRank

Both sequential and root pairwise comparison can augment additional preference data due to transitivity.

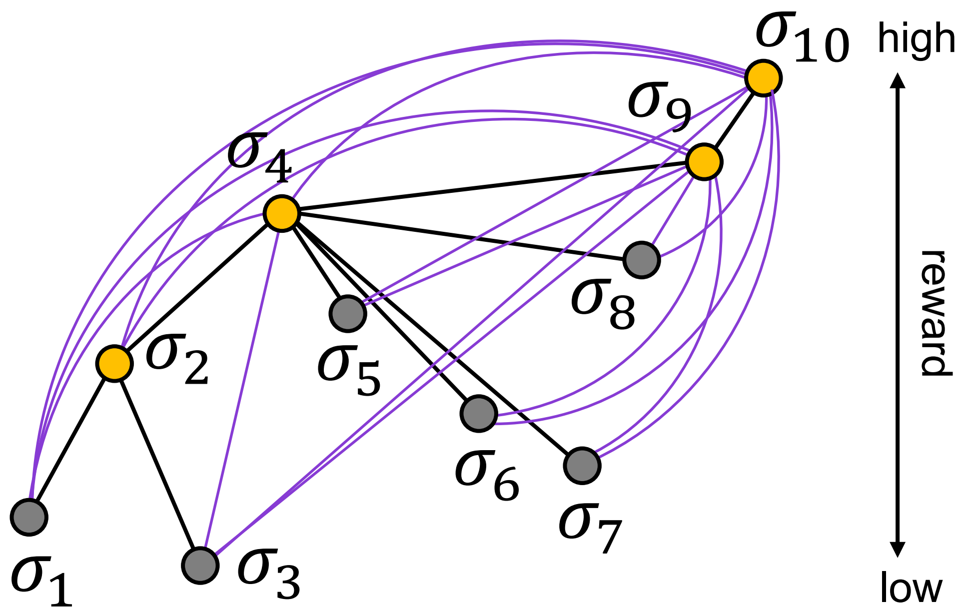


Toy Example

Suppose the reward values for segments $\sigma_1, \sigma_2, \ldots, \sigma_{10}$ are $2, 5, 1, 8, 6, 4, 3, 7, 9, 10$, respectively.
Then, we can construct a **graph** for each trajectory comparison method.



Sequential Pairwise Comparison



Root Pairwise Comparison

Black lines indicate actual pairs that receive **true preference labels** from human feedback.
Purple lines describe **augmented labels** for non-adjacent pairs.

Theoretical Analyses

NEURAL INFORMATION PROCESSING SYSTEMS

Method	M	Best		Average		Worst	
		p_N	η	p_N	η	p_N	η
Pairwise	$2N$	N	1	N	1	N	1
Sequential Pairwise	$N+1$	$N(N+1)/2$	$(N+1)/2$	$1.392(N-0.324)$	1.392	N	1
Root Pairwise	$N+1$	$N(N+1)/2$	$(N+1)/2$	$2(N+1-\sum_{n=1}^{N+1}\frac{1}{n})$	2	N	1

We prove that sequential and root pairwise comparison show 39.2% and 100% **higher average feedback efficiency** compared to conventional pairwise comparison.

Theoretical Analyses

We show that **the convergence rate of the empirical risk** is $\mathcal{O}(2n_B/(\beta TM))$.

Based on our analysis, the reward model is likely to **converge faster** in the **order of root pairwise, sequential pairwise, and pairwise comparison** in terms of the global iteration T.

We show that **the convergence rate of the generalization bounds as follows:**

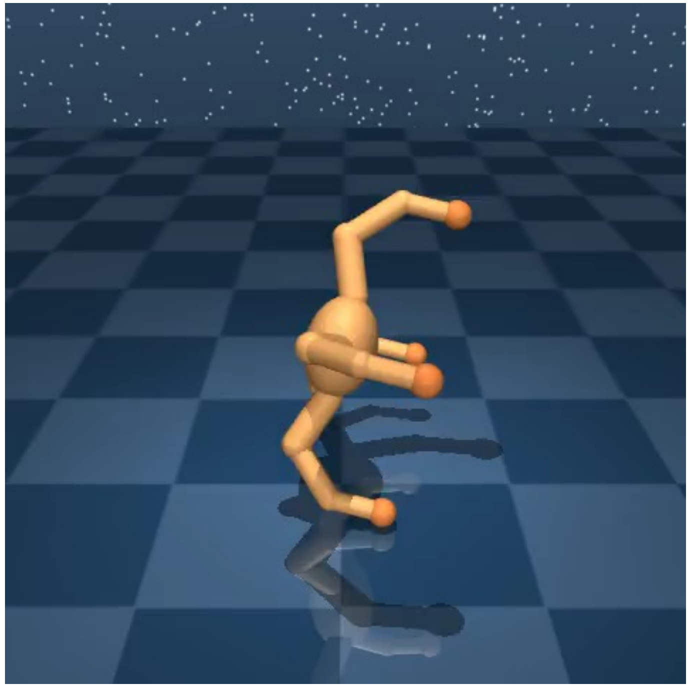
Corollary 3.1. *For a fixed $M \geq 2$, the generalization bounds of the reward model for pairwise, sequential pairwise, and root pairwise comparison converge at rates of $\mathcal{O}(\sqrt{\ln(T)/T})$, $\mathcal{O}(\sqrt{(\ln(T))^2/T})$, and $\mathcal{O}(\sqrt{\ln(T)/T})$, respectively, with probability at least $1 - 1/T$.*

Root pairwise comparison demonstrates a **faster convergence rate** than sequential pairwise comparison, and has the **same convergence rate** as pairwise comparison.

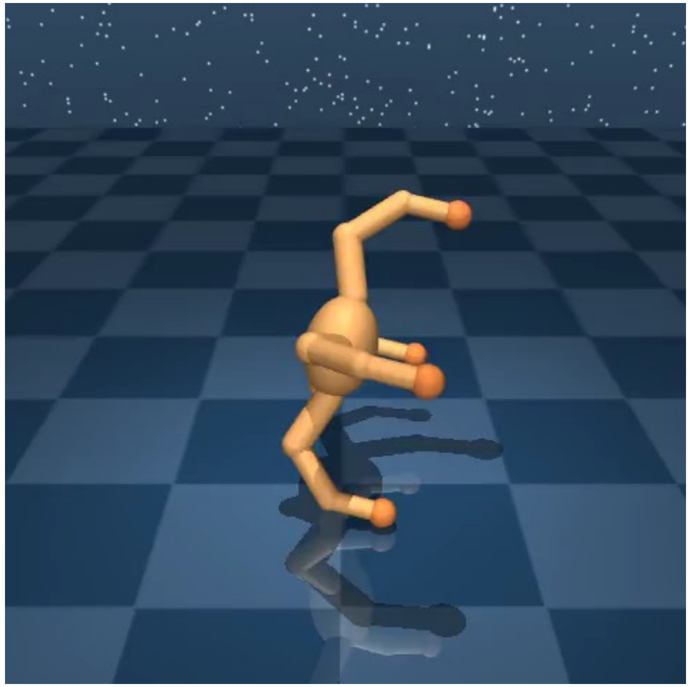
Simulation Experiments: DMControl

We show that the overall performance in DMControl is in the order of root, sequential, and pairwise comparison.

In the example trajectories in the **quadruped walk** task,
the agent trained using **pairwise** comparison **fails** to turn its body upside down.



Pairwise
Comparison
(Reward = 112.2)



Sequential Pairwise
Comparison
(Reward = 457.7)

Root Pairwise
Comparison
(Reward = 934.0)

Simulation Experiments: DMControl

We show that the overall performance in DMControl is in the order of root, sequential, and pairwise comparison.

In the example trajectories in the **walker walk** task,
the agent trained using **root** pairwise comparison shows **the fastest and most stable gait.**

Pairwise Comparison (Reward = 730.2)

Sequential Pairwise Comparison (Reward = 820.9)

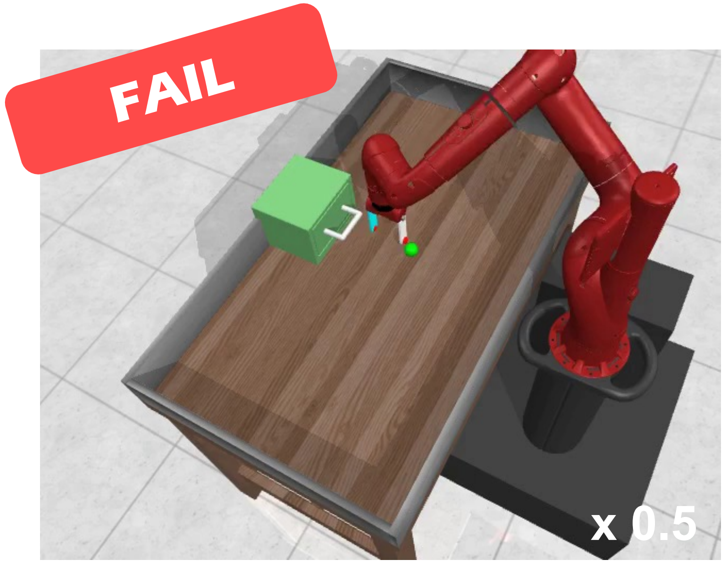
Root Pairwise Comparison (Reward = 985.1)

Simulation Experiments: Meta-World

NEURAL INFORMATION PROCESSING SYSTEMS

We show that the overall performance in Meta-World is in the order of root, sequential, and pairwise comparison.

In the first scenario in the **drawer open** task,
the agent trained using **pairwise** comparison **fails to open the drawer**.



Pairwise Comparison (Reward = 2578.1)

Sequential Pairwise Comparison (Reward = 4192.8)

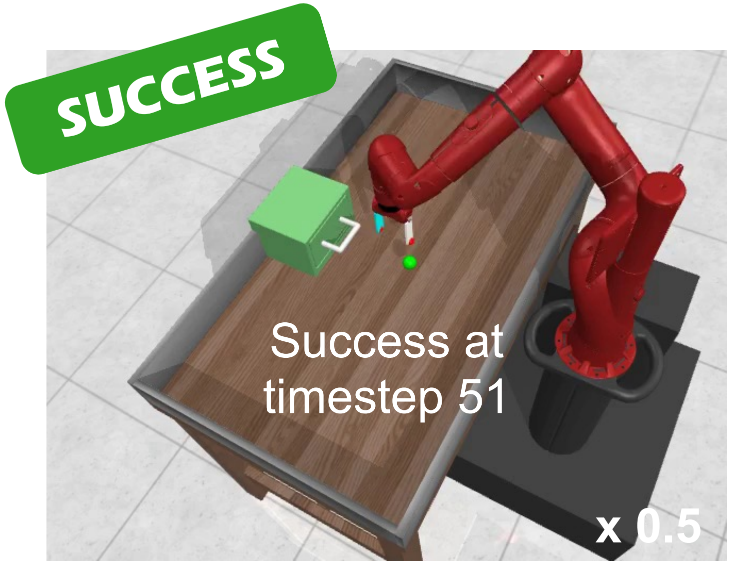
Root Pairwise Comparison (Reward = 4718.0)

Simulation Experiments: Meta-World

NEURAL INFORMATION PROCESSING SYSTEMS

We show that the overall performance in Meta-World is in the order of root, sequential, and pairwise comparison.

In the second scenario in the **drawer open** task,
all agents open the drawer, but the agents trained using **pairwise** and **sequential** pairwise comparison are **unstable** because their **end effectors oscillate** with a **large** and **small amplitude**, respectively.



Pairwise Comparison (Reward = 3880.0)

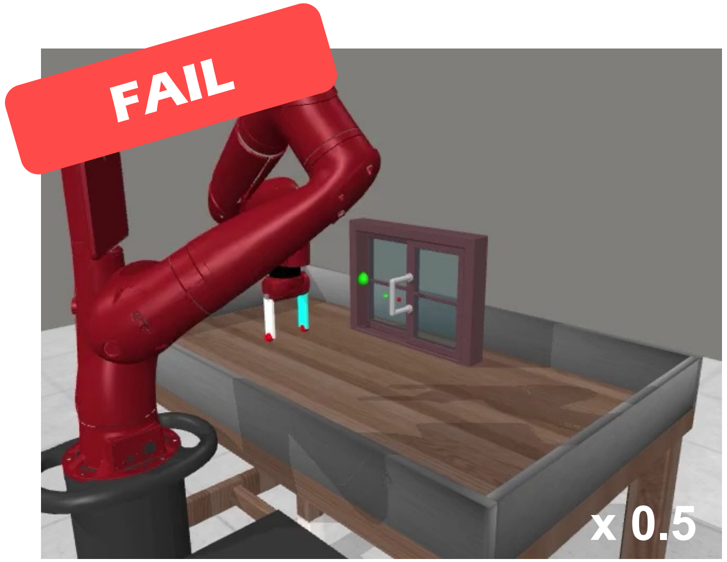
Sequential Pairwise Comparison (Reward = 4196.6)

Root Pairwise Comparison (Reward = 4766.9)

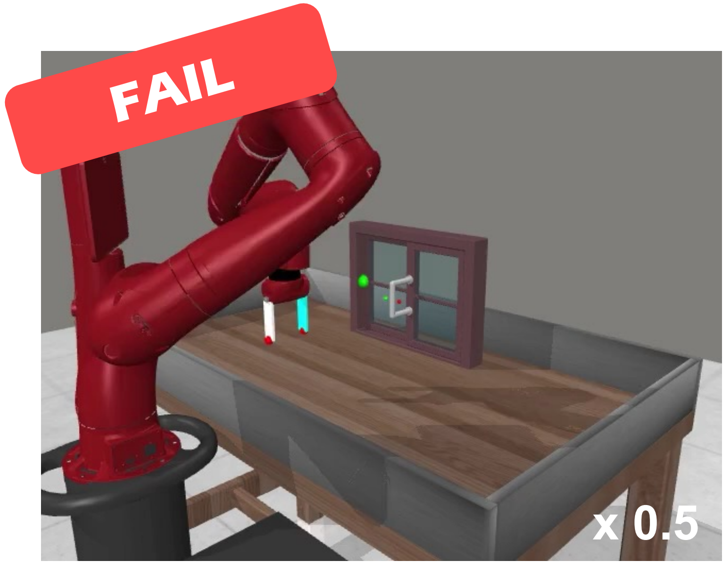
Simulation Experiments: Meta-World

We show that the overall performance in Meta-World is in the order of root, sequential, and pairwise comparison.

In the example trajectories in the **window open** task,
only the agent trained using **root** pairwise comparison **succeeds in opening** the window.



Pairwise Comparison
(Reward = 288.3)



Sequential Pairwise Comparison
(Reward = 762.3)

Root Pairwise Comparison
(Reward = 853.6)

Simulation Experiments: Meta-World

NEURAL INFORMATION PROCESSING SYSTEMS

We show that the overall performance in Meta-World is in the order of root, sequential, and pairwise comparison.

In the example trajectories in the **hammer** task,
agents trained using **sequential** and **root** pairwise comparison **succeed in driving a nail** into the wooden box.

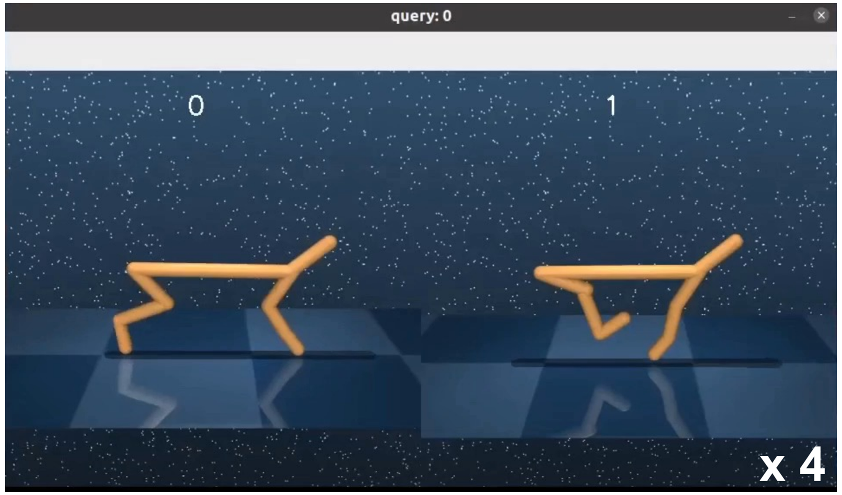
Pairwise Comparison (Reward = 1974.9)

Sequential Pairwise Comparison (Reward = 4093.7)

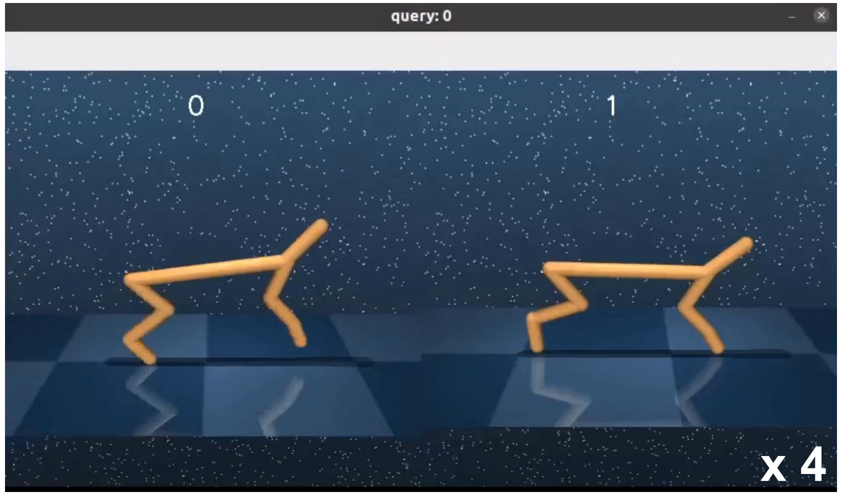
Root Pairwise Comparison (Reward = 4142.0)

Experiments: Real Human Feedback

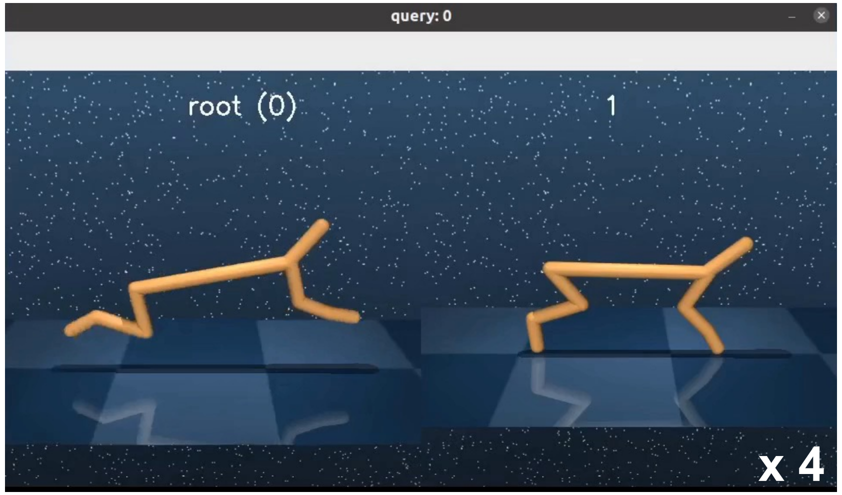
We conduct experiments with **real human feedback** to compare the user stress level for each method.
Each participant trained a **cheetah to run** as fast as it can.



Pairwise
Comparison



Sequential Pairwise
Comparison



Root Pairwise
Comparison

Experiments: Real Human Feedback

NEURAL INFORMATION PROCESSING SYSTEMS

After the experiments end, the participants took a survey.

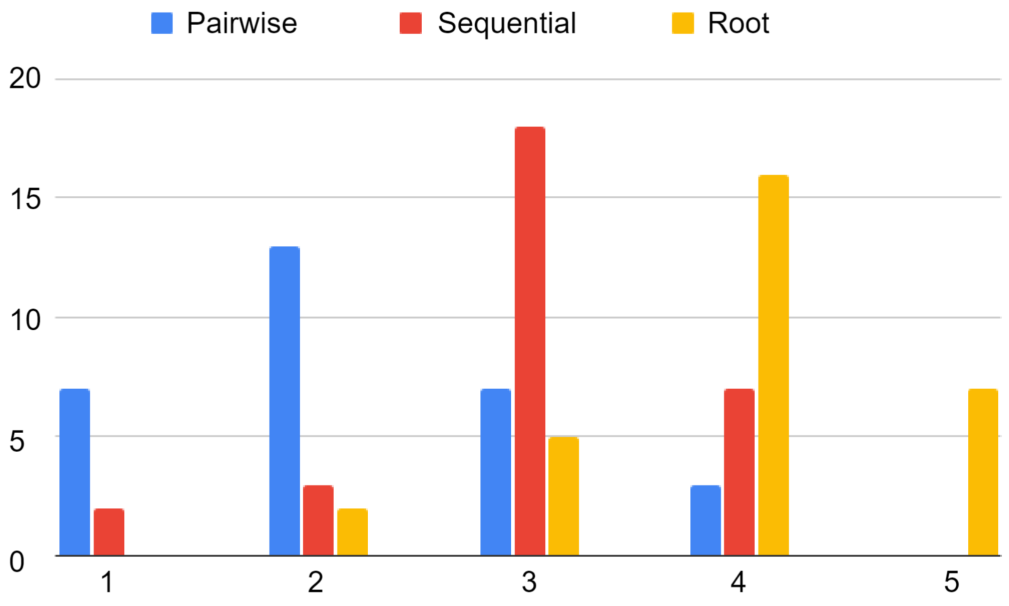
Survey Questions:

Q1) Express the user satisfaction that you have experienced from each trajectory comparison method in levels from 1 to 5. A higher score indicates more satisfaction and less stress, while a lower score indicates less satisfaction and more stress. (1: strong stress, 2: weak stress, 3: no stress or satisfaction, 4: weak satisfaction, 5: strong satisfaction)

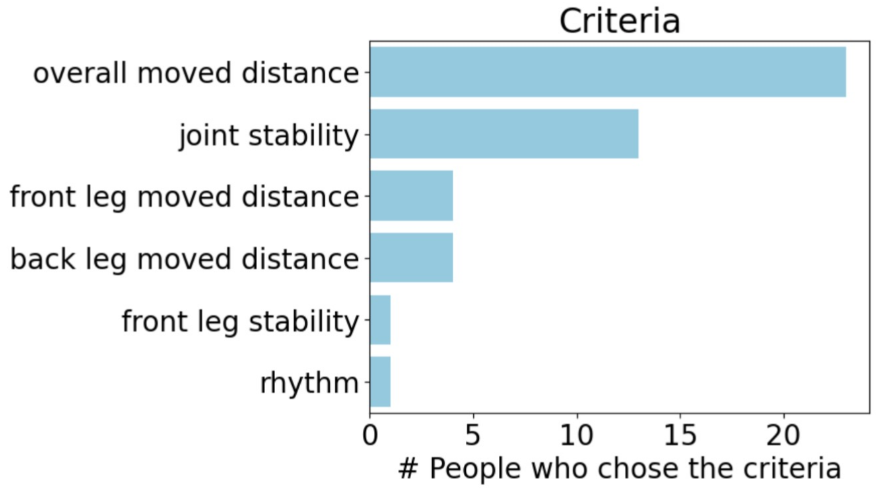
Q2) What were your own criteria for selecting one trajectory over the other? If you have multiple criteria, please write them down in order of priority.

Experiments: Real Human Feedback

Participants responed that the **user satisfaction scores** are 2.20 (pairwise), 3.00 (sequential), and 3.93 (root). The **most significant preference criterion** was the **overall moved distance** of the agent.



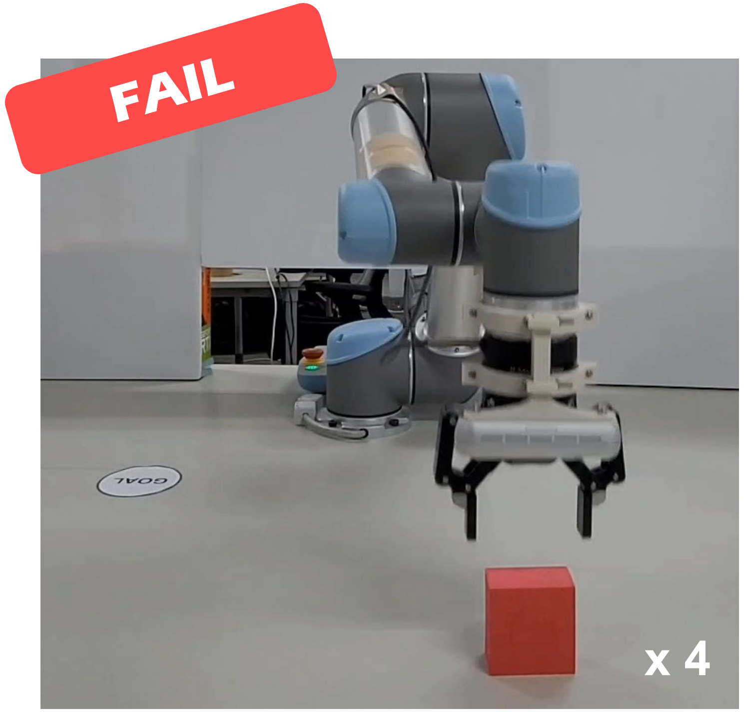
(a) User Satisfaction



(b) Preference Criteria

Real Robot Experiments

To demonstrate our method in **real-world environments**, we conduct a **block placing** task using a **real UR-5 robot**.



Pairwise
Comparison

Sequential Pairwise
Comparison

Root Pairwise
Comparison

Contributions

NEURAL INFORMATION PROCESSING SYSTEMS

1) We propose a **novel RLHF framework** that utilizes **sequential preference ranking** to **enhance human feedback efficiency**. We prove the proposed sequential and root pairwise comparison substantially improve the average feedback efficiency and speeds up the estimation of the reward function.
2) We derive the convergence rates of the empirical risk and the generalization bound of the reward model using the proposed sequential and root pairwise comparison. We address **the trade-off between feedback efficiency and data dependency** required for successful reward learning.
3) We empirically show that **prioritizing the feedback efficiency** is significantly important by evaluating in simulation and real-world environments. Both sequential and root pairwise comparison outperform conventional pairwise comparison on average. **Root pairwise comparison** shows the **most substantial improvement** against the baseline by 29.0% and 25.0% in DMControl locomotion and Meta-World manipulation tasks, respectively.

Thank you for your attention