S-EPOA: Overcoming the Indistinguishability of Segments with Skill-Driven Preference-Based Reinforcement Learning

Ni Mu*, Yao Luan*, Yiqin Yang, Bo Xu, Qing-Shan Jia

2025.08.31

{mn23, luany23}@mails.tsinghua.edu.cn

RL is Powerful, but *Reward Design* is Challenging

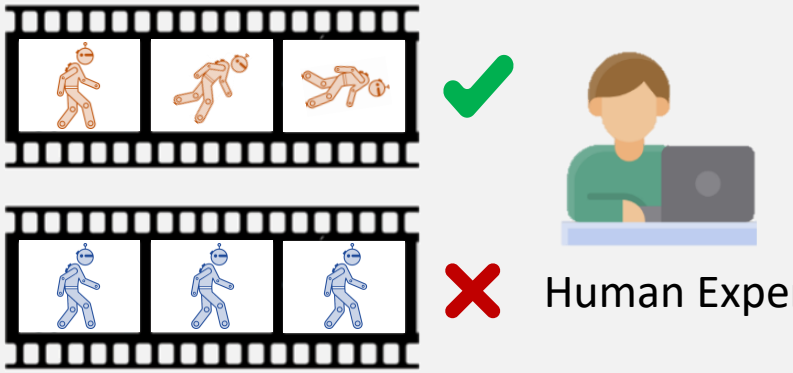
- Reinforcement Learning (RL):
  - Succeeds in gaming [1], robotics [2], and autonomous systems [3].
  - Requires ***carefully designed reward functions*** [4].
- **Example**: A cleaning robot must:
  - ① Pick up trash efficiently,
  - ② Avoid obstacles, and
  - ③ Save battery life.
- Poor rewards lead to *unintended behaviors*.

Reward design becomes a bottleneck of RL application.

[1] David Silver, et al. Mastering the game of go with deep neural networks and tree search. nature, 529(7587):484–489, 2016
[2] Yuanpei Chen, et al. Towards human-level bimanual dexterous manipulation with reinforcement learning. NeurIPS 2022.
[3] Marc G Bellemare, et al. Autonomous navigation of stratospheric balloons using reinforcement learning. Nature, 588(7836):77–82, 2020.
[4] Kimin Lee, Laura M Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. ICML 2021.

Preference-Based RL (PbRL): A Reward-Free Paradigm

- PbRL replaces rewards with ***human preferences***.
- Humans compare behaviors to guide learning.
- ***Advantages:***
  - No need for complex reward design.
  - Uses human intuition directly.
- ***PbRL is widely applied:***
  - Robotics [5].
  - LLM training (e.g., RLHF) [6].
  - Autonomous Driving [7].



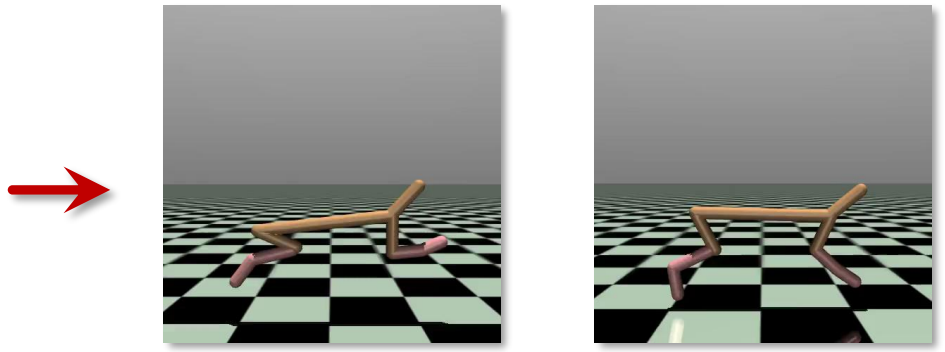
Utilize human preference

Learn a reward model $\hat{r}_\psi$,
which serves as the reward function

[5] Donald Joseph Hejna III and Dorsa Sadigh. Few-shot preference learning for humanin-the-loop RL. CoRL 2023.
[6] Ouyang, Long, et al. "Training language models to follow instructions with human feedback." NeurIPS 2022.
[7] Ni Mu, Yao Luan, and Qing-Shan Jia. "Preference-based Multi-Objective Reinforcement Learning." IEEE Transactions on Automation Science and Engineering (2025).

Challenge in PbRL: *Segment Indistinguishability*

- What is *Segment Indistinguishability*?
  - ***Humans struggle to label preferences*** when behavior segments are too similar.
  - Existing methods [4] prioritize "informative" queries, which produces similar behavior pairs, ***aggravating this issue***.



- **We validate it through human experiments:**
  - When performance differences shrink ↓, human labeling accuracy drops ↓.

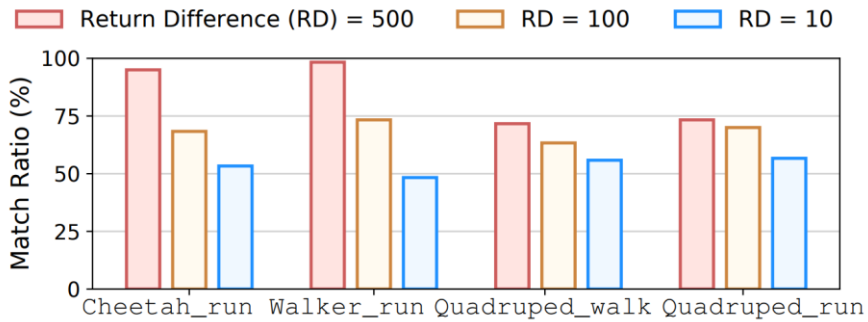
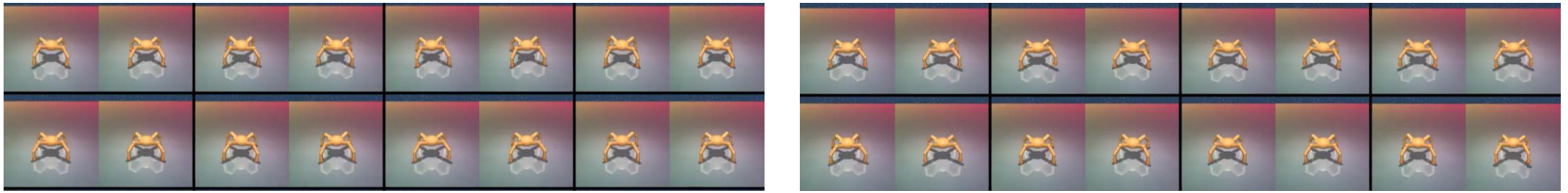


Figure 2: Human-labeled preferences match ratio with ground truth. As the return differences decrease, labeling errors increase.

Motivation: Utilize *Skills* from Unsupervised RL

- Unsupervised RL [8]:
  - Learns policies $a \sim \pi(\cdot \mid s, z)$ with different skills z
- Different skills produce **naturally distinguishable behaviors** [9].
  - Example: Stationary vs. Running, easy to tell apart.
- Motivation: ***Use different skills to generate distinguishable behaviors***.



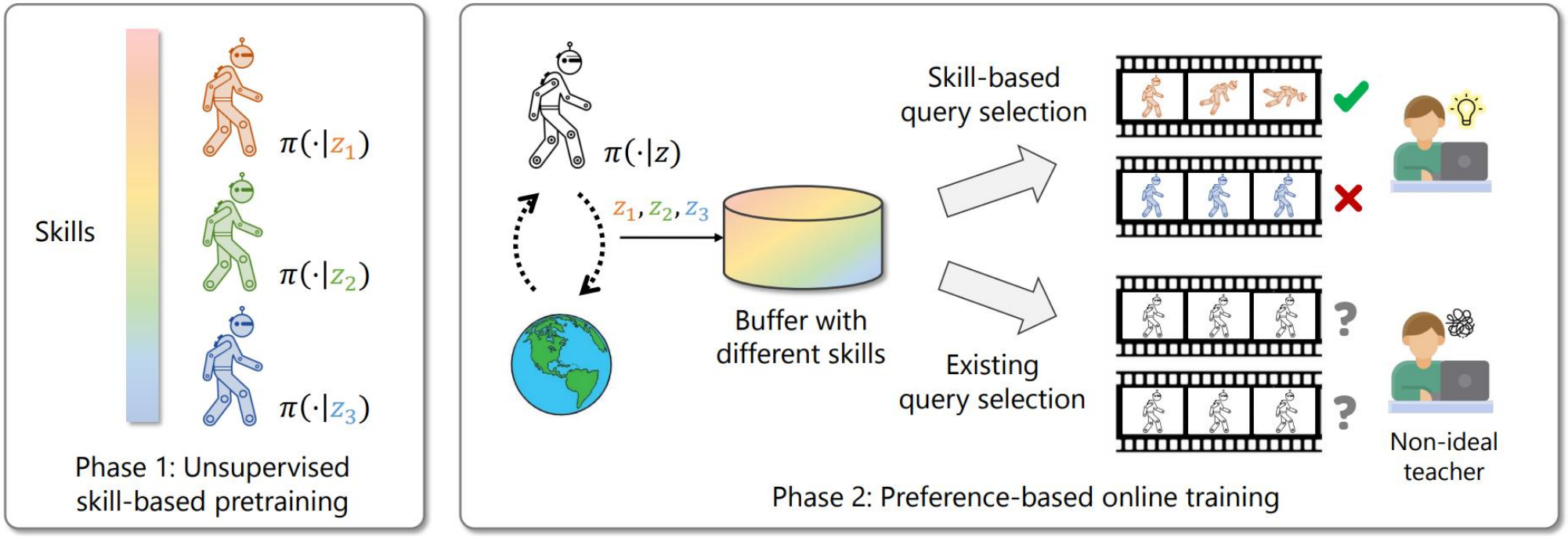
The unsupervised skills learned by METRA [9].

[8] Hao Liu and Pieter Abbeel. Aps: Active pretraining with successor features. ICML 2021.
[9] Seohong Park, Oleh Rybkin, and Sergey Levine. Metra: Scalable unsupervised RL with metric-aware abstraction. ICLR 2024.

The S-EPOA Framework

- To address *Segment Indistinguishability*,
  we propose **S**kill-**E**nhanced **P**reference **O**ptimization **A**lgorithm **(S-EPOA).**
- **Phase 1: Skill Pretraining**
  - Discover diverse skills via unsupervised RL pretraining.
  - Generate distinct behaviors (e.g., stationary, running).
- **Phase 2: Skill-Driven PbRL**
  - Estimate skill returns: Train $R(z) = \mathbb{E}_{\tau \sim \pi(\cdot|\cdot, z)} \sum_{(s,a)} \hat{r}(s, a)$
  - Select queries $(\text{behavior}_0, \text{behavior}_1)$ with largest return differences $|R(z_0) - R(z_1)|$.

The S-EPOA Framework

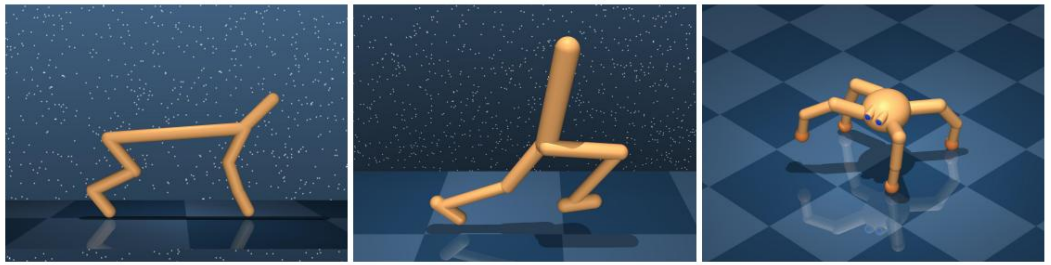


① In the pretraining phase, we learn diverse skills based on unsupervised skill discovery.

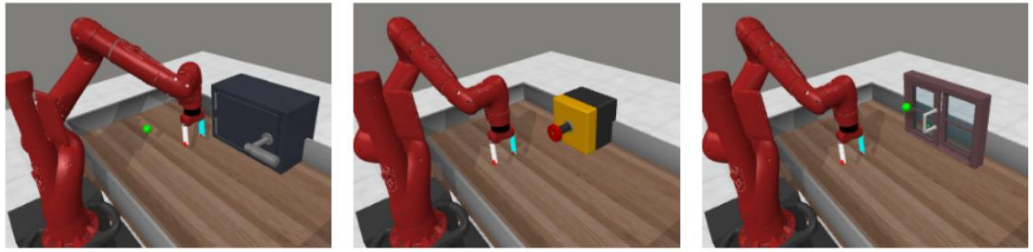
② In the online training phase, we leverage a novel skill-based query selection method to ***generate distinguishable queries*** for non-ideal teachers.

Experimental Setup

- **Tasks**:
  - 3 tasks in DeepMind Control Suite.
  - 4 tasks in Metaworld.
- **Baselines**:
  - PEBBLE, SURF, RUNE, RIME.
- **Real-world simulation**:
  - Introduced a noisy scripted teacher to mimic human feedback errors.



(a) Cheetah_run (b) Walker_run (c) Quadruped_walk and Quadruped_run



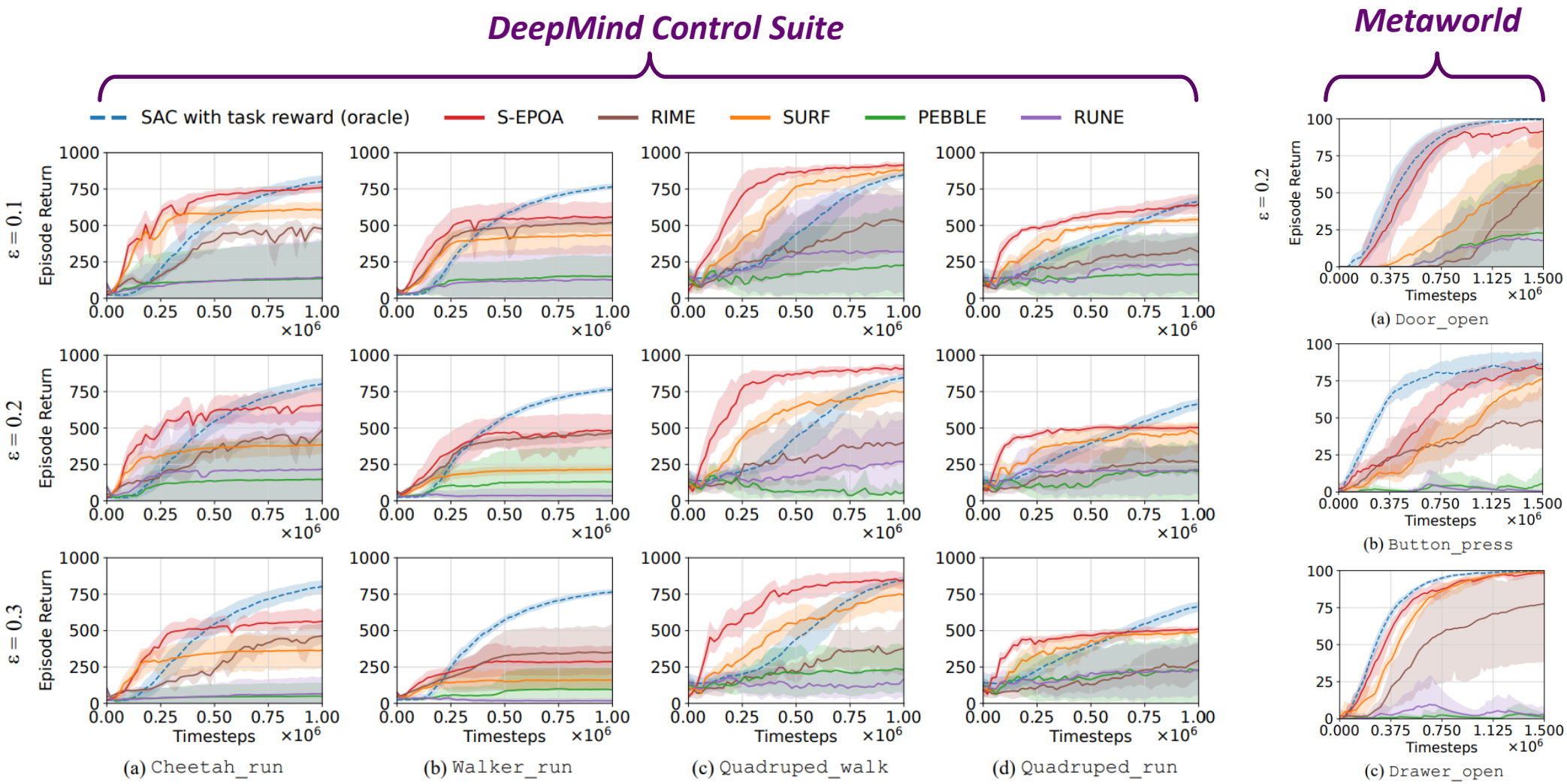
(d) Door_open (e) Button_press (f) Window_open

Specifically, for a query (σ_0, σ_1) with underlying trajectories (τ_0, τ_1) and ground truth reward function r_{gt}, if

$$\left| \sum_{(s,a)\in\tau_0} r_{\text{gt}}(s,a) - \sum_{(s,a)\in\tau_1} r_{\text{gt}}(s,a) \right| < \epsilon \cdot R_{\text{avg}}, \qquad (10)$$

a random label is assigned. Here, R_{avg} is the average return of the most recent ten trajectories. We refer to $\epsilon \in (0, 1)$ as the error rate.

Experimental Results



S-EPOA outperforms all baseline across different noisy feedback settings.

Human Experiments

- To assess the distinguishability of queries, we conduct human experiments.
  - Humans label preferences for behavior pairs selected by S-EPOA and baselines.
- S-EPOA generates more distinguishable behavior pairs for human.

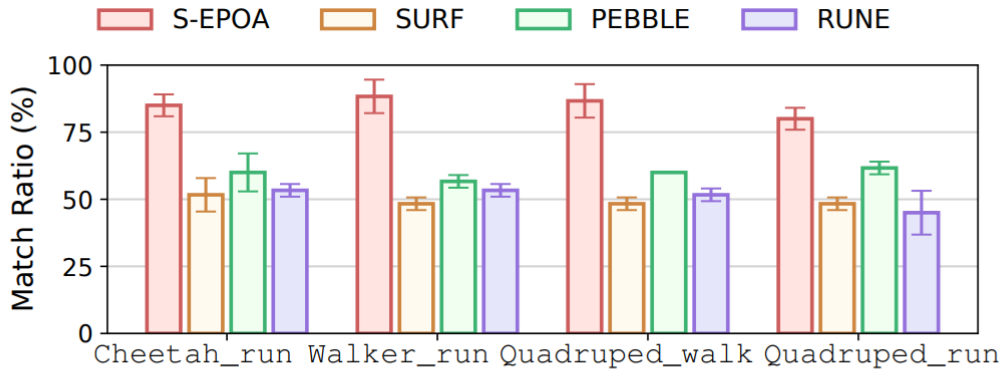


Figure 6: Human-labeled preferences match ratio with ground truth. Queries selected by S-EPOA are more distinguishable than others.

*S-EPOA successfully addresses **Segment Indistinguishability** issue, making PbRL more **applicable**.*

Takeaway

- **Challenge**:
  - PbRL is powerful, but ***Segment Indistinguishability*** limits its application.
- **Motivation**:
  - S-EPOA uses ***skill mechanisms*** in unsupervised RL to address this issue.
- **Method**:
  - ① Pretraining diverse skills z,
  - ② Comparing behavior pairs with large skill return differences $|R(z_0) - R(z_1)|$.
- **Results**:
  - S-EPOA shows **superior performance** across tasks and feedback setting.
  - S-EPOA selects behavior pairs that **human recognized as distinguishable**.
  - S-EPOA successfully addresses *Segment Indistinguishability* issue.

Thank you!

Ni Mu*, Yao Luan*, Yiqin Yang, Bo Xu, Qing-Shan Jia

2025.08.31

{mn23, luany23}@mails.tsinghua.edu.cn