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

**Ni Mu**[1*], **Yao Luan**[1*], **Yiqin Yang**[2†], **Bo Xu**[2] and **Qing-Shan Jia**[1†]
[1]Beijing Key Laboratory of Embodied Intelligence Systems,
Department of Automation, Tsinghua University
[2] The Key Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences
{mn23, luany23}@mails.tsinghua.edu.cn, yiqin.yang@ia.ac.cn, jiaqs@tsinghua.edu.cn

## Abstract

Preference-based reinforcement learning (PbRL) stands out by utilizing human preferences as a direct reward signal, eliminating the need for intricate reward engineering. However, despite its potential, traditional PbRL methods are often constrained by the indistinguishability of segments, which impedes the learning process. In this paper, we introduce Skill-Enhanced Preference Optimization Algorithm (S-EPOA), which addresses the segment indistinguishability issue by integrating skill mechanisms into the preference learning framework. Specifically, we first conduct the unsupervised pretraining to learn useful skills. Then, we propose a novel query selection mechanism to balance the information gain and distinguishability over the learned skill space. Experimental results on a range of tasks, including robotic manipulation and locomotion, demonstrate that S-EPOA significantly outperforms conventional PbRL methods in terms of both robustness and learning efficiency. The results highlight the effectiveness of skill-driven learning in overcoming the challenges posed by segment indistinguishability.

## 1 Introduction

Reinforcement Learning (RL) has made significant progress across a variety of fields, including gameplay [Mnih *et al.*, 2013; Silver *et al.*, 2016], robotics [Chen *et al.*, 2022] and autonomous systems [Bellemare *et al.*, 2020; Mu *et al.*, 2024; Luan *et al.*, 2025]. Yet, the success of RL often relies on the careful construction of reward functions, a process that can be both labor-intensive and costly. To solve this issue, Preference-based Reinforcement Learning (PbRL) emerges as a compelling alternative [Christiano *et al.*, 2017; Lee *et al.*, 2021b]. PbRL uses human-provided preferences among various agent behaviors to serve as the reward signal, thereby eliminating the need for hand-crafted reward functions.

Existing PbRL methods [Lee *et al.*, 2021b; Park *et al.*, 2022a; Shin *et al.*, 2023; Kim *et al.*, 2023] focus on enhancing feedback efficiency, by maximizing the expected return with minimal feedback queries. However, these methods rely on high-quality or even ideal expert feedback, overlooking an important issue in preference labeling: **indistinguishability of segments**. For example, asking humans to specify preferences between two similar trajectories can be challenging, as it is difficult for human observers to discern which is superior. Consequently, the resulting preference labels are often incorrect, further degrading the performance of the algorithm [Lee *et al.*, 2021a]. Therefore, the segment indistinguishability issue limits the broader applicability of PbRL.

In the field of unsupervised reinforcement learning, information-theoretic methods have been shown to discover useful and diverse skills, without the need for rewards [Eysenbach *et al.*, 2019; Hansen *et al.*, 2019]. In contrast to the unlabeled agent behaviors in PbRL, these discovered skills possess a higher degree of distinguishability, allowing humans to easily express preferences between different skills. However, the skill-driven approach has not been fully explored in PbRL, and how to apply the discovered skills to preference learning remains unclear. This naturally leads to the following question:

*How can we integrate the skill mechanism with PbRL to overcome the indistinguishability of segments?*

In this work, we aim to provide an effective solution to the important and practical issue in PbRL: indistinguishability of segments, by incorporating the skill mechanism. First, we conduct skill-based unsupervised pretraining to learn useful and diverse skills. Next, we introduce a novel query selection mechanism in the learned skill space, which effectively balances the information gain with the distinguishability of the query. We name our method as **S**kill-**E**nhanced **P**reference **O**ptimization **A**lgorithm (S-EPOA). Our experiments demonstrate the necessity of the above two techniques, showing that S-EPOA significantly outperforms conventional PbRL methods in terms of both robustness and learning efficiency.

In summary, our contributions are threefold:

- First, we highlight the critical issue of segment indistinguishability, validate its practical significance through human experiments, and theoretically analyze the limitations of current mainstream query selection methods, such as disagreement.
- Second, we propose S-EPOA, a skill-driven reward learning framework, which selects queries in a highly

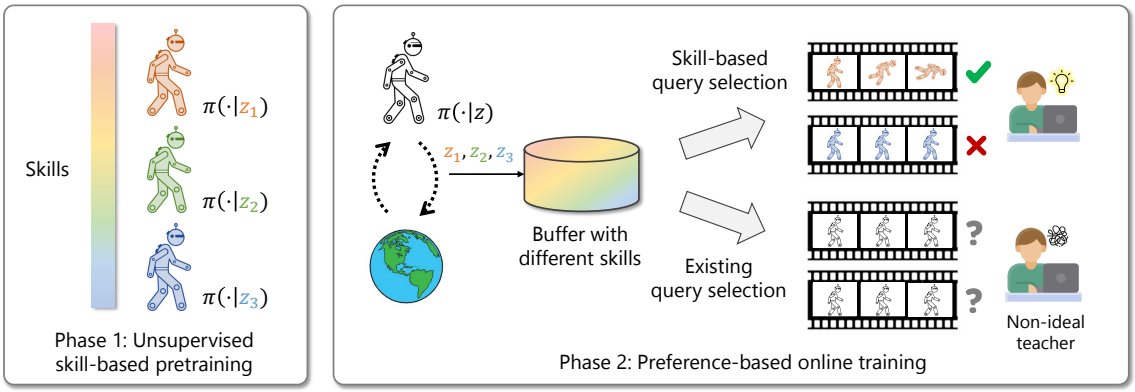



Figure 1: The framework of S-EPOA. In the pretraining phase, we learn diverse skills based on unsupervised skill discovery methods. In the online training phase, we leverage a novel skill-based query selection method to generate distinguishable queries for non-ideal teachers.

distinguishable skill space to address the segment indistinguishability issue.

- Lastly, we conduct extensive experiments to show that S-EPOA significantly outperforms conventional PbRL methods under non-ideal feedback conditions. The results indicate that by introducing the skill mechanism, we can effectively mitigate the segment indistinguishability issue, thereby broadening the application of PbRL.

## 2 Preliminaries

**Reinforcement Learning.** A Markov Decision Problem (MDP) could be characterized by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor balancing instant and future rewards. A policy $\pi$ interacts with the environment by sampling action $a$ from distribution $\pi(s, a)$ when observing state $s$. The goal of RL agent is to learn a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, which maximizes the expectation of a discounted cumulative reward: $\mathcal{L}(\pi) = \mathbb{E}_{\mu_0,\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$.

For any policy $\pi$, the corresponding state-action value function is $Q^\pi(s, a) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | S_t = s, A_t = a, \pi]$. The state value function is $V^\pi(s) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | S_t = s, \pi]$. It follows from the Bellman equation that $V^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) Q^\pi(s, a)$.

**Preference-based Reinforcement Learning.** In PbRL, the reward function $r$ is replaced by human-provided preferences over segment pairs, denoted as $\sigma_0, \sigma_1$. A segment $\sigma$ is a continuous sequence in a fixed length $H$ of states and actions, i.e. $\{s_k, a_k, \ldots, s_{k+H-1}, a_{k+H-1}\}$. Preferences are expressed as one-hot labels $y \in \{(0, 1), (1, 0)\}$, indicating the preferred segment. All preference triples $(\sigma_0, \sigma_1, y)$ are stored in the dataset $D$. The algorithm first estimates a reward function $\hat{r}_\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, parameterized by $\psi$, using provided preferences, This learned reward function $\hat{r}_\psi$ is then used to train the policy with standard RL algorithms. To construct $\hat{r}_\psi$, we employ the Bradley-Terry model [Bradley and Terry, 1952; Christiano *et al.*, 2017] as follows:

$$P_\psi[\sigma_1 \succ \sigma_0] = \frac{\exp \sum_t \hat{r}_\psi(s_t^1, a_t^1)}{\sum_{i\in\{0,1\}} \exp \sum_t \hat{r}_\psi(s_t^i, a_t^i)}. \tag{1}$$

where $\sigma_1 \succ \sigma_0$ indicates the human prefer $\sigma_1$ than $\sigma_0$. $\hat{r}_\psi$ can be trained by minimizing the cross-entropy loss:

$$\mathcal{L}_{\text{reward}}(\psi) = -\mathop{\mathbb{E}}_{(\sigma_0,\sigma_1,y)\sim\mathcal{D}}\Big[y(0) \log P_\psi[\sigma_0 \succ \sigma_1] + y(1) \log P_\psi[\sigma_1 \succ \sigma_0]\Big]. \tag{2}$$

## 3 Indistinguishability of Segments

In an ideal PbRL scenario, human labelers can give the true preference between two distinct behaviors. In practice, however, humans often have to label similar trajectories, which can lead to mistakes. These errors could reduce the precision of the trained reward function, and consequently degrade the performance. We name this labeling issue as the **Indistinguishability of Segments**. The issue significantly limits the broad application of PbRL, particularly in safety control fields [Fulton and Platzer, 2018] where the precision of the reward function is strictly required.

**Human experiments.** To validate this issue, we conduct human experiments, where human labelers provide preferences between segments with various return differences. Then, we calculate the match ratio between human labels and ground truth. Specifically, the human labelers watch a video rendering each segment and select the one that better achieves the objective based on the instruction. For example, the instruction given to human teachers in the `Cheetah_run` task is to run as fast as possible. The results in Figure 2 shows that as the return differences decrease, the match rate between human-labeled and ground truth preferences diminishes, indicating an increase in labeling errors. This confirms the practical significance of the indistinguishability issue. Please refer to Appendix C for experimental details.

Furthermore, the segment indistinguishability issue can be particularly severe when the query selection method prioritizes "informative" queries. A representative example is the

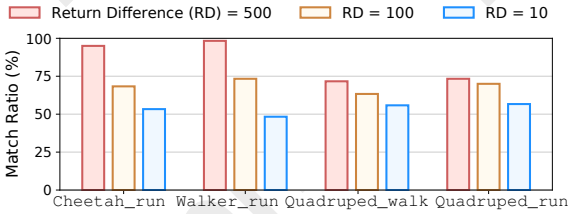



Figure 2: Human-labeled preferences match ratio with ground truth. As the return differences decrease, labeling errors increase.

**Algorithm 1** S-EPOA Framework

**Require:** frequency of feedback $K$, number of queries $M$ per feedback session, total feedback number $N_{\text{total}}$

Initialize $\pi(a|s,z) \leftarrow$ UNSUPERVISED PRETRAIN
**for** each iteration **do**
  **if** iteration % $K$ == 0 and total feedback $< N_{\text{total}}$ **then**
    Update trajectory estimator $R_\theta(z)$ based on Eq. 6
    **for** $m$ in $1 \cdots M$ **do**
      $(\sigma_0, \sigma_1) \sim$ QUERY SELECTION
      Query the teacher for preference label $y$
      Store preference $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\sigma_0, \sigma_1, y)\}$
    **end for**
    Update the reward model $\hat{r}_\psi$ using $\mathcal{D}$
    Relabel replay buffer $\mathcal{B}$ using $\hat{r}_\psi$
  **end if**
  **if** the current episode ends **then**
    Update $z_{\text{task}}$ based on Eq. 9
  **end if**
  Obtain action $a_t \sim \pi(a|s_t, z_{\text{task}})$ and next state $s_{t+1}$
  Store transitions $\mathcal{B} \leftarrow \mathcal{B} \cup \{(s_t, a_t, s_{t+1}, \hat{r}_\psi(s_t, a_t)\}$
  Sample transitions $(s, a, s', \hat{r}_\psi)$ from $\mathcal{B}$
  Minimize $\mathcal{L}_{\text{critic}}$ and $\mathcal{L}_{\text{actor}}$ with $\hat{r}_\psi$
**end for**

disagreement mechanism in PEBBLE [Lee *et al.*, 2021b], which selects queries based on the highest variance in predictions from an ensemble of reward models. To illustrate why this occurs, we conduct the following theoretical analysis. As shown in Proposition 1, the disagreement mechanism tends to select queries with similar return values, resulting in segments with similar and often indistinguishable behaviors.

**Proposition 1.** *Let $\{\hat{r}^i\}$ be an ensemble of i.i.d. reward estimators, and $(\sigma_1, \sigma_2)$ be a segment pair with ground-truth cumulative discounted reward $r_1 \geq r_2$. Suppose $\hat{r}^i$ estimates the cumulative discounted reward of $\sigma_j$ as $\hat{r}^i_j \sim N(r_j, c)$ ($c$ is a constant), and induces preference*

$$\hat{P}_i[\sigma_1 \succ \sigma_2] = \frac{\exp \hat{r}^i_1}{\exp \hat{r}^i_1 + \exp \hat{r}^i_2} = \text{sigmoid}(\hat{r}^i_1 - \hat{r}^i_2). \quad (3)$$

*Then the disagreement of induced preference across $\{\hat{r}^i\}$, i.e. $\text{Var}[\hat{P}[\sigma_1 \succ \sigma_2]]$, approximately and monotonically increases as the dissimilarity of segment pair $\Delta = r_1 - r_2$ decreases.*

# 4 Skill-Driven PbRL

To address the issue of segment indistinguishability, we hope to choose segments with different behaviors. Skill discovery methods can discover diverse skills, which align with this requirement. In this section, we propose the Skill-Enhanced Preference Optimization Algorithm (S-EPOA), which leverages skill discovery techniques to enhance PbRL's reward learning by selecting distinguishable queries. S-EPOA can be integrated with any skill discovery method, by introducing the following two key components:

- Skill-based unsupervised pretraining, where the agent explores the environment and learns useful skills without supervision (see Section 4.1).
- Skill-based query selection, which can select more distinguishable queries based on the learned skill space (see Section 4.2).

We show the overall framework of S-EPOA in Figure 1 and Algorithm 1.

## 4.1 Skill-based Unsupervised Pretraining

Previous unsupervised pretraining methods [Lee *et al.*, 2021b; Park *et al.*, 2022a] aim to help the agent explore the state space by maximizing the state entropy $H(s)$. However, these methods often fail to produce clearly distinguishable behaviors, as the learned policies primarily focus on exploration, rather than diversity. As a result, the queries may be difficult for human teachers to distinguish. Skill-based unsupervised pretraining addresses this issue by guiding the agent to discover diverse skills, resulting in more distinct behaviors, and enabling the selection of distinguishable queries in the early training process.

In skill discovery, the policy is in the form of $\pi(a|s,z)$, where $z \in \mathcal{Z}$ denotes the skill and $\mathcal{Z}$ represents the skill space. To ensure that the policies generated by each skill have distinct behaviors, we aim to maximize the mutual information between the policy's behavior and its skill. For tractability, we optimize a variational lower bound, as shown in Eq. 4, where $p$ is the underlying distribution, and $q_\phi$ is learned via maximum likelihood on data sampled from $p$.

$$\begin{aligned} I(s;z) &= \mathbb{E}_{s,z\sim p(s,z)}[\log p(s|z)] - \mathbb{E}_{s\sim p(s)}[\log p(s)] \\ &\geq \mathbb{E}_{s,z\sim p(s,z)}[\log q_\phi(s|z)] - \mathbb{E}_{s\sim p(s)}[\log p(s)] \end{aligned} \quad (4)$$

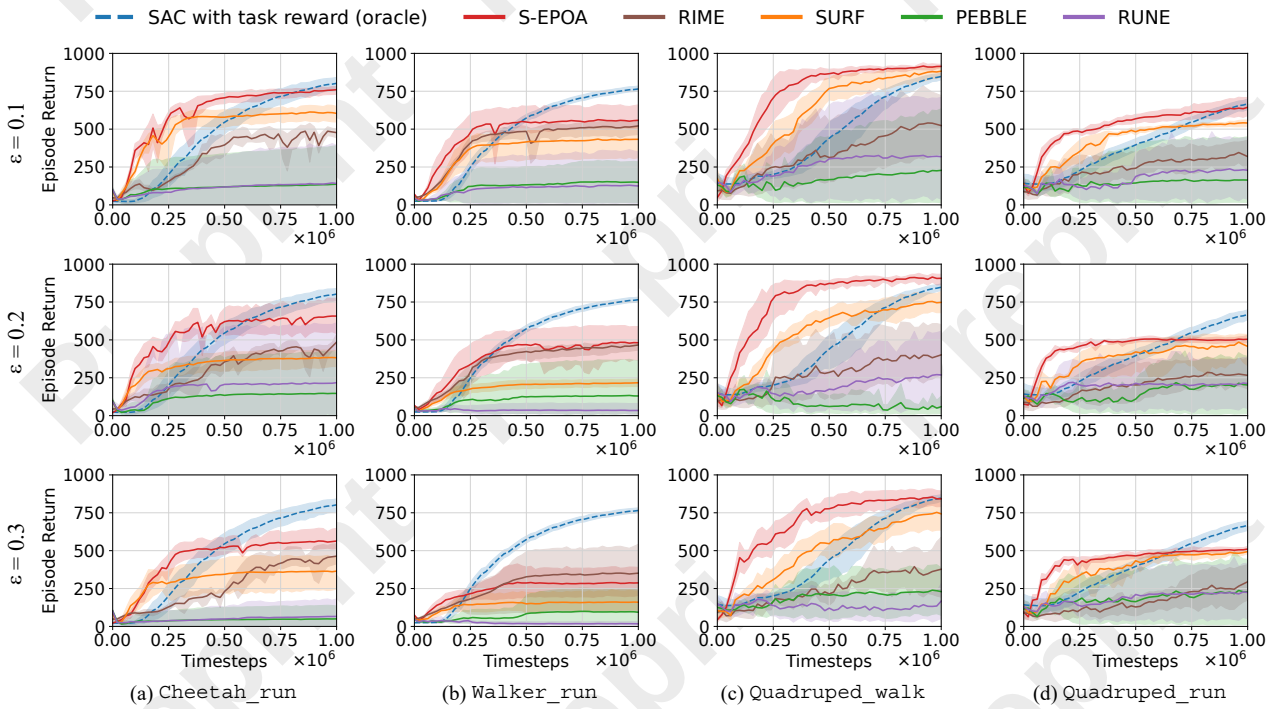
Skill discovery methods use intrinsic rewards to encourage agents to explore different behaviors. A typical form of the intrinsic reward is shown in Eq. 5:

$$r^{\text{int}} = \log q_\phi(s|z) - \log p(s). \quad (5)$$

In this way, after pre-training in this stage, we obtained a policy in the form of $\pi(a|s,z)$. By selecting different skills $z$, we can generate segments with diverse behaviors.

## 4.2 Skill-based Query Selection

In this subsection, we focus on selecting distinguishable queries to enhance reward learning. To achieve this, the two segments being compared should come from different skills with distinct behaviors and performances. We introduce a trajectory estimator $R_\theta(z)$ to estimate the expected return of




(a) `Cheetah_run` (b) `Walker_run` (c) `Quadruped_walk` (d) `Quadruped_run`

Figure 3: Learning curves on locomotion tasks from DMControl, where each row corresponds to a different error rate $\epsilon$, and each column represents a specific task. SAC serves as an oracle, using the ground-truth reward unavailable in PbRL settings. The solid line and shaded regions respectively represent the mean and standard deviation of episode return, across 5 seeds.

trajectories generated by skill $z$, which helps us identify skills with significant performance differences. The training objective of $R_\theta(z)$ is:

$$\min \mathcal{L}_{\text{est}}(\theta) = \mathbb{E}_z \left[ R_\theta(z) - \mathbb{E}_{\tau \sim \pi(\cdot|\cdot,z)} \sum_{(s,a)\in\tau} \hat{r}_\psi(s,a) \right]^2, \tag{6}$$

where $\tau$ is the trajectory generated by $z$, and $\hat{r}_\psi$ is the learned reward model. In practice, we normalize the targets of $R_\theta(z)$ to the range of $[0,1]$ for training stability. Based on $R_\theta(z)$, we define the skill-based selection criteria $I(\sigma_0,\sigma_1)$ for query $(\sigma_0,\sigma_1)$ with underlying skills $(z_0,z_1)$:

$$I(\sigma_0,\sigma_1) = (1 + |R_\theta(z_0) - R_\theta(z_1)|) \cdot (1 + \text{Var}(P_\psi[\sigma_1 \succ \sigma_0])), \tag{7}$$

where $P_\psi$ is the probability that reward model prefers $\sigma_1$ than $\sigma_0$, as defined in Eq. 1. The first term assesses the difference between the skills, while the second term measures the reward model uncertainty, which is commonly used in prior works [Lee *et al.*, 2021b; Liang *et al.*, 2022]. For training stability, we normalize both terms to the $[0,1]$ range and add 1 to balance the two values.

Thus, we propose the skill-based query selection method: For each query $(\sigma_0,\sigma_1)$, we calculate the skill-based selection criteria $I(\sigma_0,\sigma_1)$, and select queries with the highest value. Based on Eq. 7, this approach not only considers the query's uncertainty to maximize the information gain, but also considers the differences between skills, ensuring the segments have clearly distinguishable skill explanations. The specific method is illustrated in Algorithm 2.

**Algorithm 2** QUERY SELECTION

1: Randomly sample $N$ queries $(\sigma_0,\sigma_1)$, where $\sigma_0,\sigma_1$ are segments generated by skills $z_0,z_1$
2: Calculate the difference $|R_\theta(z_0) - R_\theta(z_1)|$
3: Calculate the uncertainty $\text{Var}(P_\psi[\sigma_0 \succ \sigma_1])$
4: Normalize $|R_\theta(z_0) - R_\theta(z_1)|$ and $\text{Var}(P_\psi[\sigma_0 \succ \sigma_1])$
5: Calculate $I(\sigma_0,\sigma_1)$ for each query
6: Select the query with maximum $I(\sigma_0,\sigma_1)$

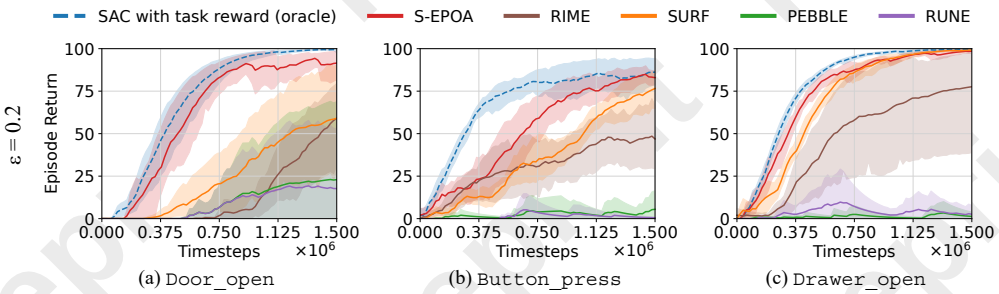
### 4.3 Implementation Details

To convert the unsupervised pretraining policy $\pi(a|s,z)$ in Section 4.1 to PbRL's policy $\pi(a|s)$ in Section 4.2, we attempt to obtain the skill nearest to the current task, i.e., performs the best under the current estimated reward function. This skill is denoted as $z_{\text{task}}$. Intuitively, $z_{\text{task}}$ can be derived by solving the optimization problem in Eq. 8:

$$z_{\text{task}} = \arg\max_z \mathbb{E}_{\mu_0,\pi(a_t|s_t,z)} \left[ \sum_{t=0}^{\infty} \gamma^t \hat{r}(s_t,a_t) \right] \tag{8}$$

Using the learned trajectory estimator $R_\theta(z)$, we sample $N_z$ skills $z_i$ from $\mathcal{Z}$ uniformly, and approximate $z_{\text{task}}$ with

$$\hat{z}_{\text{task}} = \arg\max_{z_i} \{R_\theta(z_i)\}_{i=1}^{N_z}. \tag{9}$$




(a) `Door_open` (b) `Button_press` (c) `Drawer_open`

Figure 4: Learning curves on locomotion tasks from Metaworld, with error rate $\epsilon = 0.2$, across 5 seeds.

Based on the above discussion, we obtain the overall process of S-EPOA, as shown in Algorithm 1. Firstly, we initialize the policy with skill-based unsupervised pretraining. Then, for each feedback session, we update the trajectory estimator as in Eq. 6, and select queries based on the skill-based selection criteria in Eq. 7. The PbRL reward model $\hat{r}_\psi$ is trained using the selected queries based on Eq. 2. Finally, we train the critic and actor using the learned reward function $\hat{r}_\psi$.

Besides the two key components in Section 4.1 and 4.2, we also adopt the semi-supervised data augmentation technique for reward learning [Park *et al.*, 2022a]. To elaborate, we randomly sub-sample several shorter pairs of $(\hat{\sigma}_0, \hat{\sigma}_1)$ from the queried segments $(\sigma_0, \sigma_1, y)$, and use these $(\hat{\sigma}_0, \hat{\sigma}_1, y)$ to optimize the cross-entropy loss in Eq. 2. Moreover, we sample a batch of unlabeled segments $(\sigma_0, \sigma_1)$, generate the artificial labels $\hat{y}$, if $P_\psi[\sigma_0 \succ \sigma_1]$ or $P_\psi[\sigma_1 \succ \sigma_0]$ reaches a predefined confidence threshold.

## 5 Experiments

We design our experiments to answer the following questions: *Q1:* How does S-EPOA compare to other state-of-the-art methods under non-ideal teachers? *Q2:* Can S-EPOA select queries with higher distinguishability? *Q3:* Can S-EPOA be integrated with various skill discovery methods? *Q4:* What is the contribution of each of the proposed techniques in S-EPOA?

### 5.1 Setup

**Domains.** We evaluate S-EPOA on several complex robotic manipulation and locomotion tasks from DMControl [Tassa *et al.*, 2018] and Metaworld [Yu *et al.*, 2020]. Specifically, We choose 4 complex tasks in DMControl: `Cheetah_run`, `Walker_run`, `Quadruped_walk`, `Quadruped_run`, and 3 complex tasks in Metaworld: `Door_open`, `Button_press`, `Window_open`. The details of experimental tasks are shown in Appendix B.1.

**Baselines and Implementation.** We compare S-EPOA with several state-of-the-art methods, including PEBBLE [Lee *et al.*, 2021b], SURF [Park *et al.*, 2022a], RUNE [Liang *et al.*, 2022], and RIME [Cheng *et al.*, 2024], a robust PbRL method. We also train SAC with ground truth rewards as a performance upper bound. For PEBBLE, SURF, and RUNE, we employ the disagreement query selection, which performs the best among all the query selection methods. In our experiment, we use APS [Liu and Abbeel, 2021a] for unsupervised skill discovery. The impact of skill discovery methods is discussed in the ablation study. More implementation details are provided in Appendix B.2 and D.

**Noisy scripted teacher imitating humans.** Following prior works [Lee *et al.*, 2021b; Liang *et al.*, 2022], we use a scripted teacher for systematic evaluation, which provides preferences between segments based on the sum of ground-truth rewards. To better mimic human decision-making uncertainty, we introduce a noisy scripted teacher. When the performance difference between two policies is marginal, humans often struggle to distinguish between them, as Section 3 shows. To imitate this, we implement an error mechanism: if the ground truth returns of two trajectories are nearly identical, we randomly assign a preference label of 0 or 1. The core idea is to evaluate policy performance by comparing the overall returns of entire trajectories, which aligns with how humans assess policies based on their overall effectiveness. Specifically, for a query $(\sigma_0, \sigma_1)$ with underlying trajectories $(\tau_0, \tau_1)$ and ground truth reward function $r_{\text{gt}}$, if

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

a random label is assigned. Here, $R_{\text{avg}}$ is the average return of the most recent ten trajectories. We refer to $\epsilon \in (0, 1)$ as the error rate. For fairness, we constrain that the two segments in a query come from different trajectories. It is important to note that our noisy teacher differs from the "mistake" teacher in B-Pref [Lee *et al.*, 2021a], which randomly flips correct labels. Our teacher only introduces errors in too-close queries.

### 5.2 Results on Benchmark Tasks

**Locomotion tasks in DMControl.** Figure 3 shows the learning curves of S-EPOA and baselines on the four DMControl tasks with error rates $\epsilon \in \{0.1, 0.2, 0.3\}$. S-EPOA outperforms baselines in most environments and remains robust under non-ideal conditions, while other PbRL methods are unstable and even fail.

**Robotic manipulation tasks in Metaworld.** Figure 4 shows the learning curves for S-EPOA and baselines on the three Metaworld tasks with $\epsilon = 0.2$. These results further demonstrate that S-EPOA improves robustness against non-ideal feedback across diverse tasks.

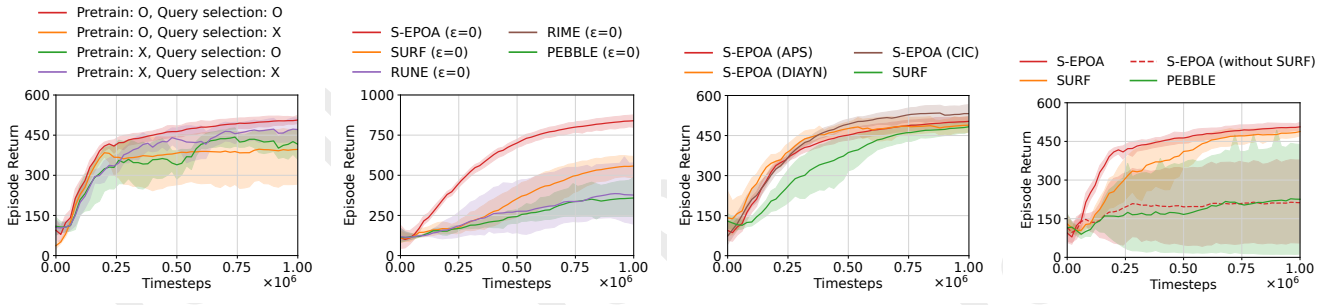



Figure 5: Ablation studies on the `Quadruped_run` task. (a) Contribution of each technique in S-EPOA, under $\epsilon = 0.3$. (b) Demonstration of enhanced learning efficiency of S-EPOA under the ideal scripted teacher with error rate $\epsilon = 0$. (c) The learning curve of S-EPOA and baselines, with and without data augmentation under $\epsilon = 0.3$. (d) Integrating S-EPOA with other skill discovery methods, under $\epsilon = 0.3$.

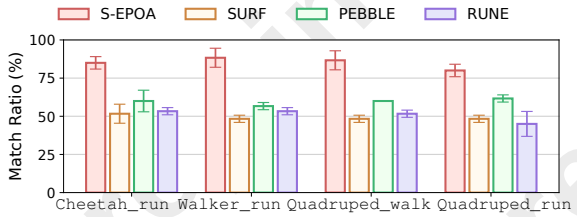



Figure 6: Human-labeled preferences match ratio with ground truth. Queries selected by S-EPOA are more distinguishable than others.

| | Disagreement | Skill-based |
|---|---|---|
| `Cheetah_run` | 0.3270 | **0.4839** |
| `Walker_run` | 0.2448 | **0.4648** |
| `Quadruped_walk` | 0.3570 | **0.3800** |
| `Quadruped_run` | 0.2545 | **0.2856** |

Table 1: Ratios of queries that can be distinguished by the noisy scripted teacher for both skill-based and disagreement query selection methods, with $\epsilon = 0.3$.

| # of Queries | S-EPOA | SURF |
|---|---|---|
| 500 | 444.25 ± 38.94 | 403.85 ± 116.18 |
| 1000 | 464.36 ± 37.04 | 438.29 ± 95.56 |
| 2000 | 506.57 ± 19.24 | 488.33 ± 21.78 |
| 10000 | 568.86 ± 197.01 | 468.49 ± 22.92 |

Table 2: Performance of S-EPOA and SURF using different numbers of queries, under $\epsilon = 0.3$.

**Enhanced query distinguishability.** To evaluate that S-EPOA selects more distinguishable queries, we compare the ratios of queries that can be distinguished by the noisy teacher, for both S-EPOA and disagreement query selection methods. As shown in Table 1, S-EPOA achieves higher distinguishability ratios across all environments, demonstrating its effectiveness in selecting easily differentiable queries. To further assess the distinguishability of queries, we conduct human experiments. We compare S-EPOA with three state-of-the-art PbRL methods in DMControl tasks, where labelers provide 20 preference labels for each run, across 3 random seeds. As shown in Figure 6, humans find the queries selected by S-EPOA easier to distinguish. Please refer to Appendix C for experimental details.

**Query visualizations.** We visualize the segment pairs selected by both S-EPOA and the disagreement mechanism. As shown in Figure 7, the pair chosen by the disagreement mechanism has similar behaviors, while the pair selected by S-EPOA clearly differs, with $\sigma_1$ being preferred. These results confirm that S-EPOA can select more distinguishable queries, enhancing both the labeling accuracy of noisy teachers and the quality of human preference labeling.

## 5.3 Ablation Study

**Component analysis.** To evaluate the effect of each technique in S-EPOA individually, we incrementally apply skill-based unsupervised pretraining and skill-based query selection to our backbone algorithm. Figure 5(a) shows the learning curves on the `Quadruped_run` task with error rate $\epsilon = 0.3$. First, skill-based unsupervised pretraining significantly boosts performance, for both skill-based (red vs. green) and disagreement-based query selection (orange vs. blue). This improvement is because the pretrained policy generates diverse behaviors, which leads to a better reward function. Additionally, skill-based query selection enhances results (red vs. orange), as it selects more distinguishable queries, enabling the agent to obtain more accurate labels. In summary, both components of S-EPOA are effective, and their combination is crucial to the method's success.

**Enhanced learning efficiency under the ideal teacher.** Under the ideal scripted teacher ($\epsilon = 0$), S-EPOA can also significantly enhance learning efficiency. As shown in Figure 5(b), S-EPOA converges faster and achieves better final performance. This advantage results from selecting skills with high distinguishability, highlighting the robustness and effectiveness of our approach.

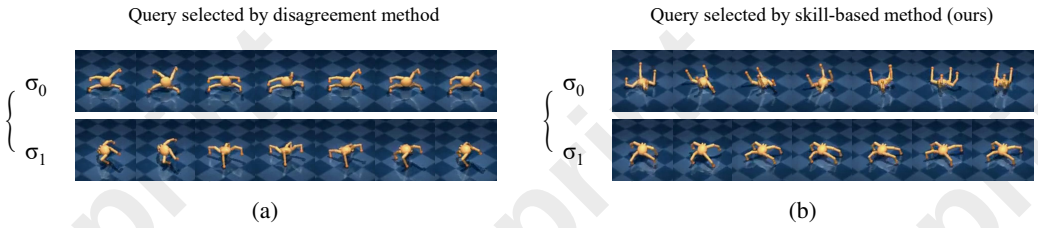



Figure 7: Visualization of segment pairs selected by (a) disagreement mechanism and (b) skill-based mechanism, under the `Quadruped_run` task, with error rate $\epsilon = 0.3$.

**Integrate S-EPOA with other skill discovery methods.** To show that S-EPOA is integratable with various skill learning methods, we replace APS with DIAYN [Eysenbach *et al.*, 2019] and CIC [Laskin *et al.*, 2022], which are typical and commonly used skill discovery methods. The details of these methods are shown in Appendix D. As is shown in Figure 5(c), the performance is consistent and similar across all skill discovery methods, which demonstrates that S-EPOA is compatible with various skill discovery methods.

**Enhanced query efficiency.** We compare the performance of S-EPOA and SURF using different numbers of queries. As is shown in Table 2, S-EPOA outperforms SURF consistently, even if only 500 queries are provided, which demonstrates the ability of S-EPOA to make better use of the limited queries.

**Effect of data augmentation in S-EPOA.** We conduct an ablation study to evaluate the impact of data augmentation in S-EPOA. Figure 5(d) shows that without data augmentation, both S-EPOA and the baselines (PEBBLE as the backbone) show similar performance deficiencies. This highlights the importance of data augmentation for achieving optimal performance, while it is not the sole factor in our method's success and does not undermine the innovation of our approach.

## 6 Related Work

**Preference-based reinforcement learning.** PbRL enables humans (or supervisors in other forms, like script teachers) to guide the RL agent toward desired behaviors by providing preferences on segment pairs, where the feedback efficiency is a primary concern [Lee *et al.*, 2021b; Park *et al.*, 2022a]. Prior works improve the feedback efficiency from various perspectives. Some works focus on the query selection scheme, trying to improve the information quality of queries [Ibarz *et al.*, 2018; Biyik *et al.*, 2020; Hejna III and Sadigh, 2023]. Some works integrate unsupervised pretraining to avoid the waste on initial nonsense queries [Lee *et al.*, 2021b]. Some works augment queries from humans to better utilize limited human feedback [Park *et al.*, 2022a; Liu *et al.*, 2023]. These methods depend on reliable feedback. However, humans could make mistakes, especially when the segment pair for comparison is slightly different, which restricts and even harms the performance in practice [Lee *et al.*, 2021a; Cheng *et al.*, 2024].

**Unsupervised pretraining for RL.** Unsupervised pretraining has been well studied in RL [Xie *et al.*, 2022], where unlabeled data (i.e., transitions without task-specific rewards) are used to learn a policy or set of policies that effectively explore the state space through intrinsic rewards. The method to calculate the intrinsic reward varies in different unsupervised pretraining works, including uncertainty measures like prediction errors [Pathak *et al.*, 2017; Pathak *et al.*, 2019; Burda *et al.*, 2019], state entropy [Hazan *et al.*, 2019; Liu and Abbeel, 2021b], pseudo-counts [Bellemare *et al.*, 2016; Ostrovski *et al.*, 2017] and empowerment measures like mutual information [Eysenbach *et al.*, 2019; Sharma *et al.*, 2020; Liu and Abbeel, 2021a; Park *et al.*, 2022b; Park *et al.*, 2023]. The learned policy could serve as a strong initialization for downstream tasks, enhancing the sample efficiency in multi-task and few-shot RL.

**Unsupervised skill discovery methods.** Unsupervised skill discovery methods are a subset of unsupervised pretraining methods, which use empowerment measures as the intrinsic reward, trying to discover a set of distinguishable primitives. Mutual information $I(s, z)$ is a common choice for the empowerment measure, where $s$ is a state, and $z$ is a latent variable indicating the skill. Some studies consider the reverse form $I(s, z) = H(z) - H(z|s)$ [Eysenbach *et al.*, 2019; Park *et al.*, 2022b], which train a parameterized skill discriminator $q(z|s)$ together with the policy. On the other hand, the forward form $I(s, z) = H(s) - H(s|z)$ [Sharma *et al.*, 2020; Liu and Abbeel, 2021a; Laskin *et al.*, 2022] can be integrated with model-based RL and state entropy-based unsupervised pretraining algorithms. Additionally, some studies design the skill latent space for unique properties by parameterizing the distribution $q(z|s)$ or $q(s|z)$. VISR [Hansen *et al.*, 2019] and APS [Liu and Abbeel, 2021a] let the latent $z$ be the successor feature to enable fast task inference. LSD [Park *et al.*, 2022b] and METRA [Park *et al.*, 2023] bind the distance in state space and latent space to force a significant travel distance in a trajectory, thereby capturing dynamic skills.

## 7 Conclusion

This paper presents S-EPOA, a robust and efficient PbRL algorithm designed to address the segment indistinguishability issue. By leveraging skill mechanisms, S-EPOA learns diverse behaviors through unsupervised learning and generates distinguishable queries through skill-based query selection. Experiments show that S-EPOA outperforms PbRL baselines in robustness and efficiency, with ablation studies confirming the effectiveness of skill-based query selection. In future work, we aim to extend S-EPOA to broader applications.

## Acknowledgments

This work is supported by the NSFC (No. 62125304, 62192751, and 62073182), the Beijing Natural Science Foundation (L233005), BNRist project (BNR2024TD03003), the 111 International Collaboration Project (B25027) and the Strategic Priority Research Program of the Chinese Academy of Sciences (No.XDA27040200).

## Contribution Statement

*N. Mu and Y. Luan contributed equally.
†Y. Yang and Q-S. Jia are corresponding authors.

## References

[Barreto *et al.*, 2017] André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[Bellemare *et al.*, 2016] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

[Bellemare *et al.*, 2020] Marc G Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C Machado, Subhodeep Moitra, Sameera S Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

[Biyik *et al.*, 2020] Erdem Biyik, Nicolas Huynh, Mykel Kochenderfer, and Dorsa Sadigh. Active preference-based gaussian process regression for reward learning. In *Robotics: Science and Systems*, 2020.

[Bradley and Terry, 1952] 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.

[Burda *et al.*, 2019] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *Seventh International Conference on Learning Representations*, pages 1–17, 2019.

[Chen *et al.*, 2022] Yuanpei Chen, Tianhao Wu, Shengjie Wang, Xidong Feng, Jiechuan Jiang, Zongqing Lu, Stephen McAleer, Hao Dong, Song-Chun Zhu, and Yaodong Yang. Towards human-level bimanual dexterous manipulation with reinforcement learning. *Advances in Neural Information Processing Systems*, 35:5150–5163, 2022.

[Cheng *et al.*, 2024] Jie Cheng, Gang Xiong, Xingyuan Dai, Qinghai Miao, Yisheng Lv, and Fei-Yue Wang. Rime: Robust preference-based reinforcement learning with noisy preferences. *arXiv preprint arXiv:2402.17257*, 2024.

[Christiano *et al.*, 2017] 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.

[Eysenbach *et al.*, 2019] Benjamin Eysenbach, Julian Ibarz, Abhishek Gupta, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[Fulton and Platzer, 2018] Nathan Fulton and André Platzer. Safe reinforcement learning via formal methods: Toward safe control through proof and learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Hansen *et al.*, 2019] Steven Hansen, Will Dabney, Andre Barreto, Tom Van de Wiele, David Warde-Farley, and Volodymyr Mnih. Fast task inference with variational intrinsic successor features. *arXiv preprint arXiv:1906.05030*, 2019.

[Hazan *et al.*, 2019] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR, 2019.

[Hejna III and Sadigh, 2023] Donald Joseph Hejna III and Dorsa Sadigh. Few-shot preference learning for human-in-the-loop rl. In *Conference on Robot Learning*, pages 2014–2025. PMLR, 2023.

[Huber, 2020] Marco F Huber. Bayesian perceptron: Towards fully bayesian neural networks. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3179–3186. IEEE, 2020.

[Ibarz *et al.*, 2018] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.

[Kim *et al.*, 2023] Changyeon Kim, Jongjin Park, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. Preference transformer: Modeling human preferences using transformers for rl. *arXiv preprint arXiv:2303.00957*, 2023.

[Kristiadi *et al.*, 2020] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pages 5436–5446. PMLR, 2020.

[Laskin *et al.*, 2021] Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, and Pieter Abbeel. Urlb: Unsupervised reinforcement learning benchmark. *arXiv preprint arXiv:2110.15191*, 2021.

[Laskin *et al.*, 2022] Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, 35:34478–34491, 2022.

[Lee *et al.*, 2021a] Kimin Lee, Laura Smith, Anca Dragan, and Pieter Abbeel. B-pref: Benchmarking preference-based reinforcement learning. *arXiv preprint arXiv:2111.03026*, 2021.

[Lee *et al.*, 2021b] Kimin Lee, Laura M Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In *International Conference on Machine Learning*, pages 6152–6163. PMLR, 2021.

[Liang *et al.*, 2022] Xinran Liang, Katherine Shu, Kimin Lee, and Pieter Abbeel. Reward uncertainty for exploration in preference-based reinforcement learning. *arXiv preprint arXiv:2205.12401*, 2022.

[Liu and Abbeel, 2021a] Hao Liu and Pieter Abbeel. Aps: Active pretraining with successor features. In *International Conference on Machine Learning*, pages 6736–6747. PMLR, 2021.

[Liu and Abbeel, 2021b] Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. *Advances in Neural Information Processing Systems*, 34:18459–18473, 2021.

[Liu *et al.*, 2023] Yi Liu, Gaurav Datta, Ellen Novoseller, and Daniel S Brown. Efficient preference-based reinforcement learning using learned dynamics models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2921–2928. IEEE, 2023.

[Luan *et al.*, 2025] Yao Luan, Qing-Shan Jia, Yi Xing, Zhiyu Li, and Tengfei Wang. An efficient real-time railway container yard management method based on partial decoupling. *IEEE Transactions on Automation Science and Engineering*, 22:14183–14200, 2025.

[Mnih *et al.*, 2013] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[Mu *et al.*, 2024] Ni Mu, Xiao Hu, Qing-Shan Jia, Xu Zhu, and Xiao He. Large-scale data center cooling control via sample-efficient reinforcement learning. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pages 2780–2785. IEEE, 2024.

[Ostrovski *et al.*, 2017] Georg Ostrovski, Marc G Bellemare, Aäron Oord, and Rémi Munos. Count-based exploration with neural density models. In *International conference on machine learning*, pages 2721–2730. PMLR, 2017.

[Park *et al.*, 2022a] Jongjin Park, Younggyo Seo, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. Surf: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning. *arXiv preprint arXiv:2203.10050*, 2022.

[Park *et al.*, 2022b] Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. Lipschitz-constrained unsupervised skill discovery. In *International Conference on Learning Representations*, 2022.

[Park *et al.*, 2023] Seohong Park, Oleh Rybkin, and Sergey Levine. Metra: Scalable unsupervised rl with metric-aware abstraction. *arXiv preprint arXiv:2310.08887*, 2023.

[Pathak *et al.*, 2017] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

[Pathak *et al.*, 2019] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International conference on machine learning*, pages 5062–5071. PMLR, 2019.

[Sharma *et al.*, 2020] Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. In *International Conference on Learning Representations*, 2020.

[Shin *et al.*, 2023] Daniel Shin, Anca D Dragan, and Daniel S Brown. Benchmarks and algorithms for offline preference-based reward learning. *arXiv preprint arXiv:2301.01392*, 2023.

[Silver *et al.*, 2016] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[Singh *et al.*, 2003] Harshinder Singh, Neeraj Misra, Vladimir Hnizdo, Adam Fedorowicz, and Eugene Demchuk. Nearest neighbor estimates of entropy. *American journal of mathematical and management sciences*, 23(3-4):301–321, 2003.

[Tassa *et al.*, 2018] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[Xie *et al.*, 2022] Zhihui Xie, Zichuan Lin, Junyou Li, Shuai Li, and Deheng Ye. Pretraining in deep reinforcement learning: A survey. *arXiv preprint arXiv:2211.03959*, 2022.

[Yu *et al.*, 2020] 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*, pages 1094–1100. PMLR, 2020.