# Reinforcement Learning with Human Feedback

## Preference-based Reinforcement Learning 4

**2025. 09. 26**

**발표자: 허종국**

# 발표자 소개

❖ 이름 : 허종국 (Jong Kook, Heo)

- Data Mining & Quality Analytics Lab
- Ph.D. Student (2021.03~)
- 지도 교수 : 김성범 교수님

❖ 관심 연구 분야

- Reinforcement Learning with Human Feedback
- Self-Supervised Learning

❖ 연락망


- E-mail : hjkso1406@korea.ac.kr

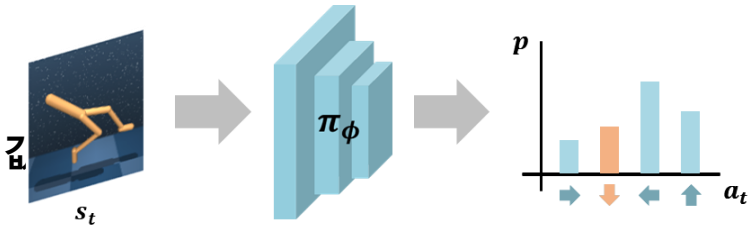
# 목 차

# Introduction

## Challenges with applying RL in the real-world

### ❖ Reinforcement Learning Framework

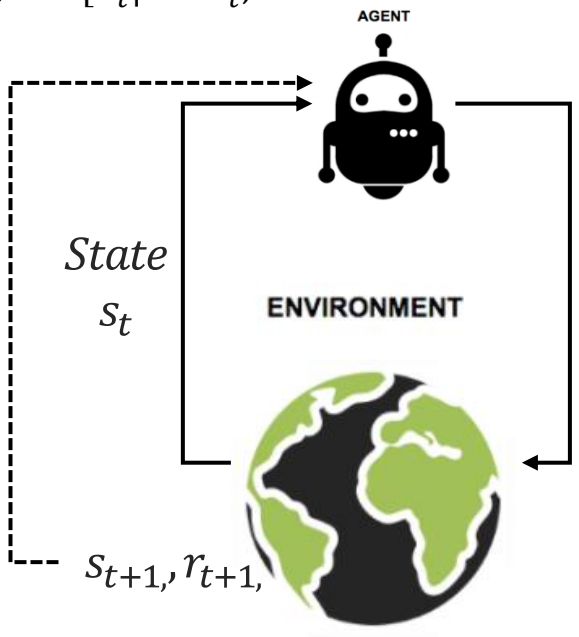
- 경험(Experience) : $(s_t, a_t, r_{t+1}, s_{t+1})$
- $G_t$ : 현재 시점 $t$ 이후부터 에피소드 끝까지 받을 수 있는 누적 보상**(확률 변수)**
  - ✓ $G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} \dots$
- 정책 함수(Policy) $\boldsymbol{\pi(a|s)}$ : 확률 변수 $a$ 에 대한 **조건부 확률 함수**
- 행동 가치 함수(Action-Value Function) $\boldsymbol{Q(s_t, a_t)}$: 확률 변수 $G_t$에 대한 **조건부 기댓값**
  - ✓ $Q(s_t, a_t) = E[G_t | S = s_t, A = a_t]$




**Policy Function**

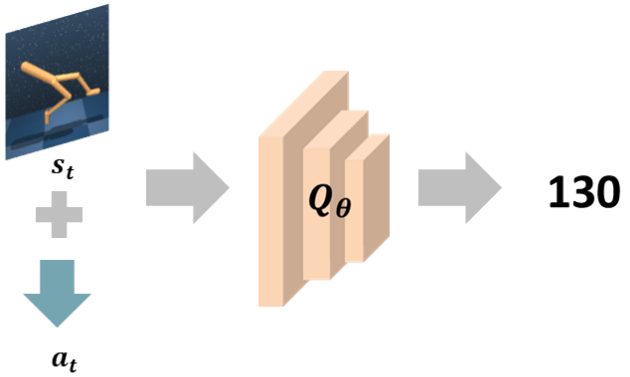
Transition Probability

$P(s', r|s, a)$




Action

$a_t$



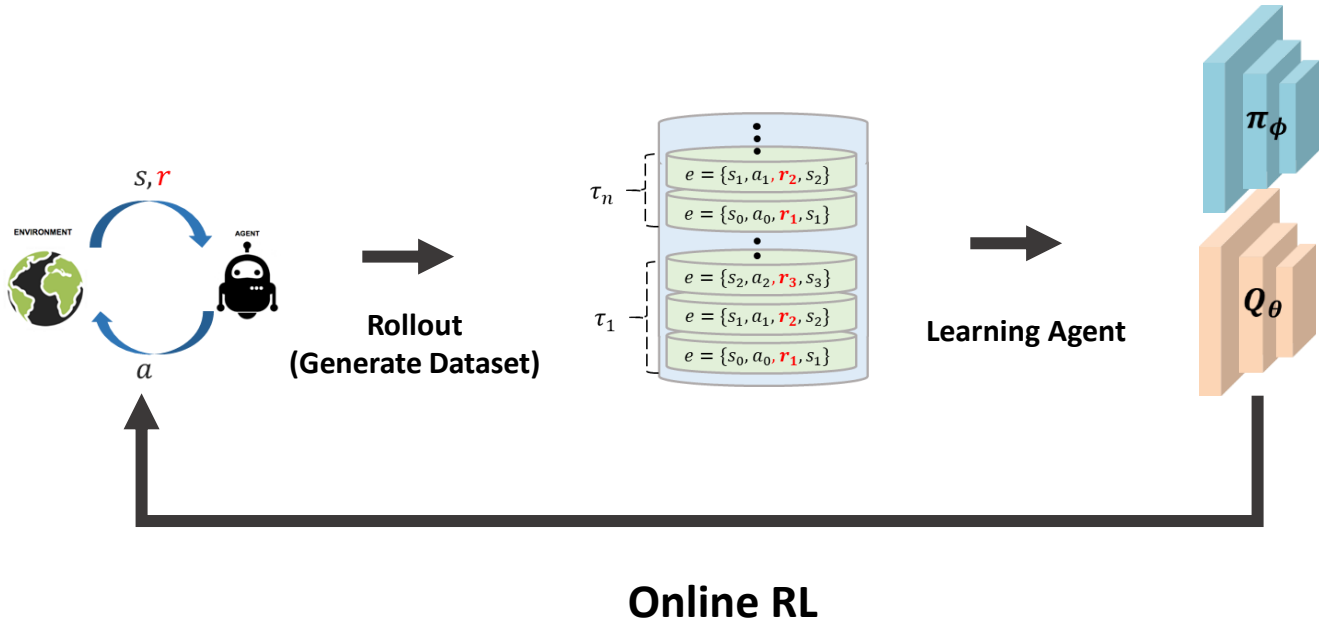

**Action-value Function**

# Introduction

## Challenges with applying RL in the real-world

❖ Reinforcement Learning Basics – Actor Critic

- 정책 함수 (Policy Function) $\pi_\phi$: 상태가 주어졌을 때 행동을 선택하는 함수
- 행동 가치함수 (Action-value Function) $Q_\theta$: 상태에 대한 행동이 얼마나 좋은지 판단하는 함수
- Actor Critic Methods : 정책 함수와 가치 함수를 함께 학습하는 강화학습 알고리즘들 (i.e., SAC, TD3, DrQ-v2, IQL)




$$L_\phi = -E_{s\sim D}[Q_\theta(s,a) log \pi_\phi(a|s)]$$

$$L_{\theta=} E_{s,a,r,s\prime\sim D}[(Q_\theta(s,a) - Q'(s,a)^2]$$
$$where\ Q'(s,a) = \boldsymbol{r} + \gamma E_{a\prime}\,[Q_\theta(s',a')]$$

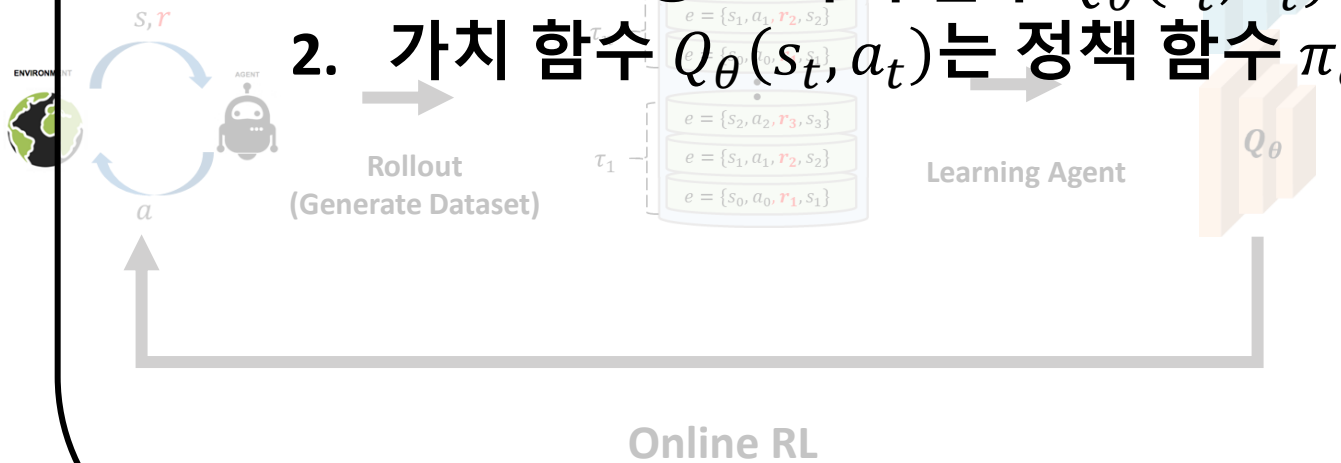
# Introduction

Challenges with applying RL in the real-world

- Reinforcement Learning Basics – Actor Critic
  - 정책 함수 (Policy Function) $\pi_\phi$: 상태가 주어졌을 때 행동을 선택하는 함수
  - 행동 가치함수 (Action-value Function) $Q_\theta$: 상태에 대한 행동이 얼마나 좋은지 판단하는 함수
  - Actor Critic Methods : 정책 함수와 가치 함수를 함께 학습하는 강화학습 알고리즘들 (i.e., SAC, TD3, DrQ-v2, IQL)

**강화학습에서...**

1. **보상은 가치 함수 $Q_\theta(s_t, a_t)$를 학습하는데 사용된다.**
2. **가치 함수 $Q_\theta(s_t, a_t)$는 정책 함수 $\pi_\phi(a|s)$를 학습하는데 사용된다.**



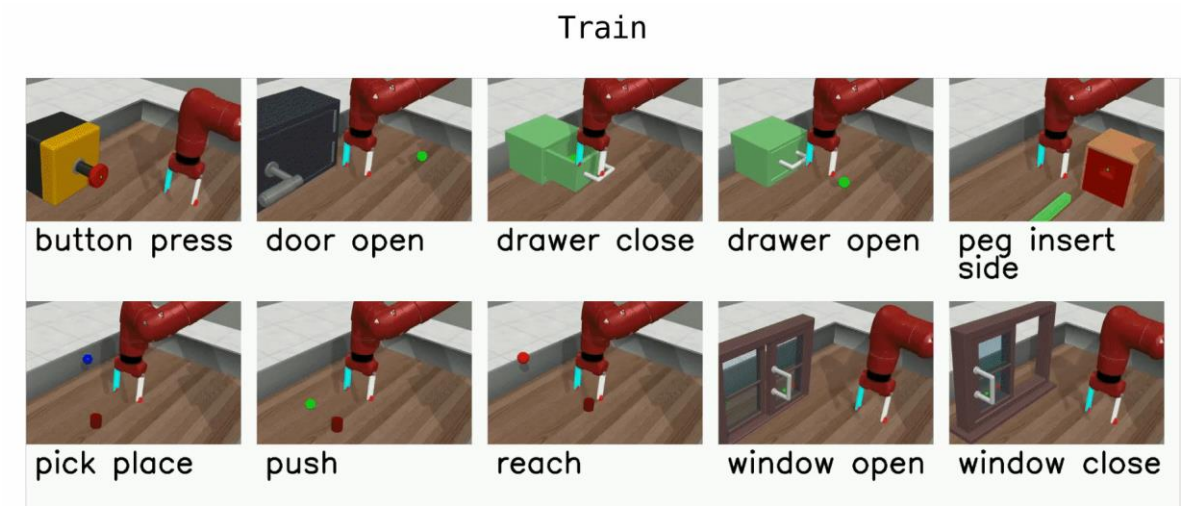

$$L_\theta = E_{s,a\sim D}[-Q_\theta(s,a) log \pi_\phi(a|s)]$$

$$L_\theta = E_{s,a,r,s' \sim D}[(Q_\theta(s,a) - Q'(s,a)^2]$$
$$where\ Q'(s,a) = r + \gamma E_{a'}[Q_\theta(s',a')]$$

# Introduction

## Challenges with applying RL in the real-world

- ❖ Meticulous Reward Design
  - How to formulate reward in robotic manipulation task?
    - ✓ How much reward for pressing the button?
    - ✓ How much reward for opening the door?




**Metaworld Environment**

**E.1.11 Door Unlock**

$$R = \begin{cases} 2L(\|\langle 1,4,2\rangle \cdot (o - h + \langle 0, 0.055, 0.07\rangle)\|, \\ 0, \\ 0.02, \\ \|\langle 1,4,2\rangle \cdot (o_i - h_i + \langle 0, 0.055, 0.07\rangle)\|)) + 8L(|t_{(x)} - o_{i,(x)}|, 0, 0.005, 0.1) \end{cases}$$

**E.1.12 Door Open**

$$alt = \mathbb{I}_{\|h_{(xy)} - o_{(xy)}\| > 0.12} \cdot \left(0.4 + 0.04 \log\left(\|h_{(xy)} - o_{(xy)}\| - 0.12\right)\right)$$

$$ready = \begin{cases} T_{H_0}\left(L(\|h - o - \langle 0.05, 0.03, -0.01\rangle\|, 0, 0.06, 0.5), L(alt - h_{(z)}, 0, 0.01, \frac{alt}{2}),\right) & h_{(z)} < alt \\ L(\|h - o - \langle 0.05, 0.03, -0.01\rangle\|, 0, 0.06, 0.5) & otherwise \end{cases}$$

$$R = \begin{cases} 2T_{H_0}(g, ready) + 8\left(0.2\mathbb{I}_{o_{(\theta)} < 0.03} + 0.8L(o_{(\theta)} + \frac{2\pi}{3}, 0, 0.5, \frac{\pi}{3})\right) & |t_{(x)} - o_{(x)}| > 0.08 \\ 10 & otherwise \end{cases}$$

**E.1.13 Box Close**

$$alt = \mathbb{I}_{\|h_{(xy)} - o_{(xy)}\| > 0.02} \cdot \left(0.4 + 0.04 \log\left(\|h_{(xy)} - o_{(xy)}\| - 0.02\right)\right)$$

$$ready = \begin{cases} T_{H_0}\left(L(\|h - o\|, 0, 0.02, 0.5), L(alt - h_{(z)}, 0, 0.01, \frac{alt}{2}),\right) & h_{(z)} < alt \\ L(\|h - o\|, 0, 0.02, 0.5) & otherwise \end{cases}$$

$$R = \begin{cases} 2T_{H_0}\left(\frac{g+1}{2}, ready\right) + 8\left(0.2\mathbb{I}_{o_{(x)} > 0.04} + 0.8L(\langle 1,1,3\rangle \|t - o\|, 0, 0.05, 0.25)\right) & |t - o| \geq 0.08 \\ 10 & otherwise \end{cases}$$

**E.1.14 Drawer Open**

$$R = 5\left(L\left(\|t - o\|, 0, 0.02, 0.2\right) + L\left(\|(o - h) \cdot \langle 3,3,1\rangle\|, 0, 0.01, \|(o_i - h_i) \cdot \langle 3,3,1\rangle\|\right)\right)$$

**E.1.15 Drawer Close**

$$R = \begin{cases} T_{H_0}\left(L(\|t - o\|, 0, 0.05, \|t - o_i\| - 0.05), T_{H_0}\left(g, L(\|o - h\|, 0, 0.005, \|o_i - h_i\| - 0.005)\right)\right) & \|t - o\| > 0.065 \\ 10 & otherwise \end{cases}$$

**E.1.16 Faucet Close**

$$R = \begin{cases} 4L(\|o - h\|, 0, 0.01, \|o_i - h_i\| - 0.01) + 6L(\|t - o\|, 0, 0.07, \|t - o_i\| - 0.07) & \|t - o\| > 0.07 \\ 10 & otherwise \end{cases}$$

**E.1.17 Faucet Open**

$$R = \begin{cases} (4L(\|o - h + \langle -.04, 0, .03\rangle\|, 0, 0.01, \|o_i - h_i\| - 0.01) \\ \quad + 6L(\|t - o + \langle -.04, 0, .03\rangle\|, 0, 0.07, \|t - o_i\| - 0.07)) & \|t - o + \langle -.04, 0, .03\rangle\| > 0.07 \\ 10 & otherwise \end{cases}$$

**Too many physics...**

https://meta-world.github.io/

보상 함수 (Reward Function)를 사람이 디자인하지 않고 강화학습 에이전트를 학습시킬 수는 없을까?

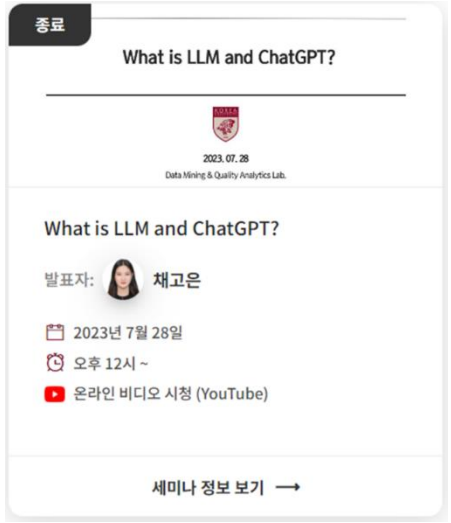
# →Preference-based RL

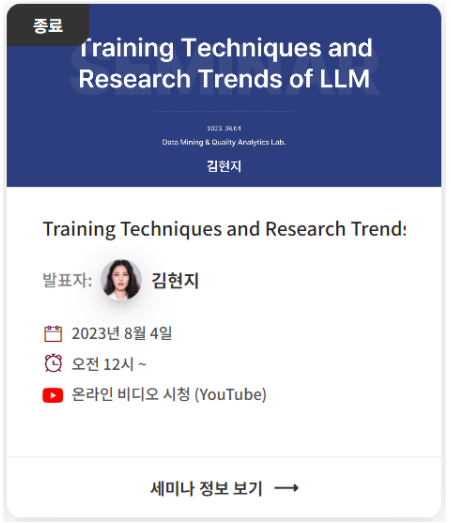
# Introduction

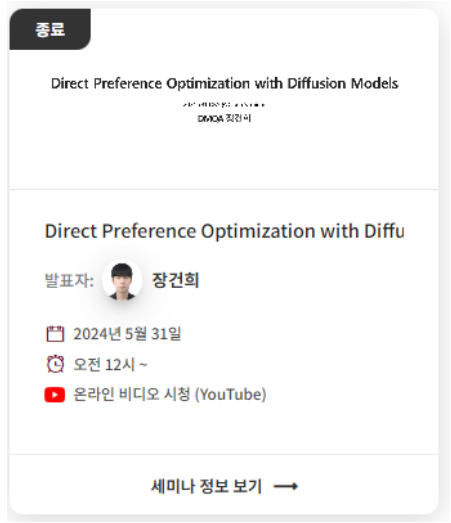
## RLHF in Large Language Model

### ❖ Details

- What is LLM and ChatGPT?
  - ✓ Seq2Seq, Transformer, GPT~InstructGPT
- Training Techniques and Research Trends of LLM
  - ✓ RLHF(Alignment Tuning), LLaMA, Alpaca, Vicuna, Falcon, etc.
- Direct Preference Optimization with Diffusion Models
  - ✓ RLHF, DPO, Diffusion DPO, DCO



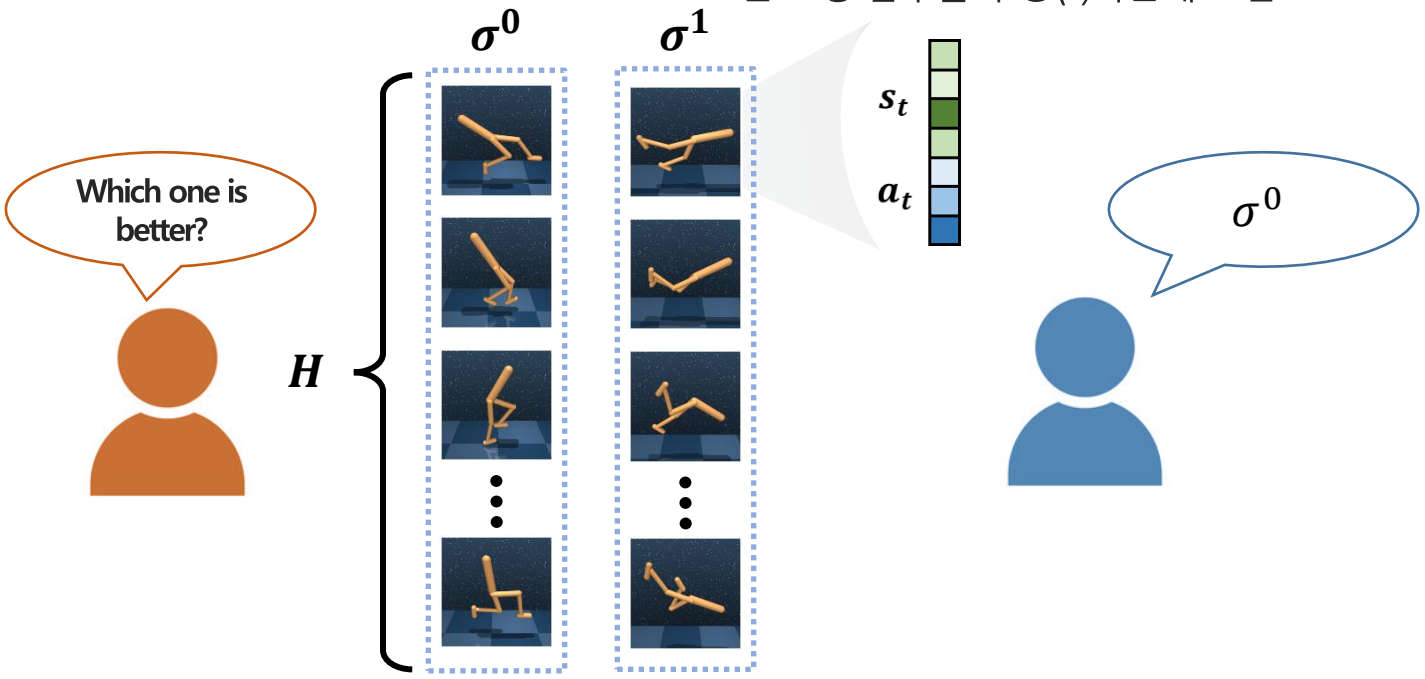

http://dmqa.korea.ac.kr/activity/seminar/416
http://dmqa.korea.ac.kr/activity/seminar/417
http://dmqa.korea.ac.kr/activity/seminar/452

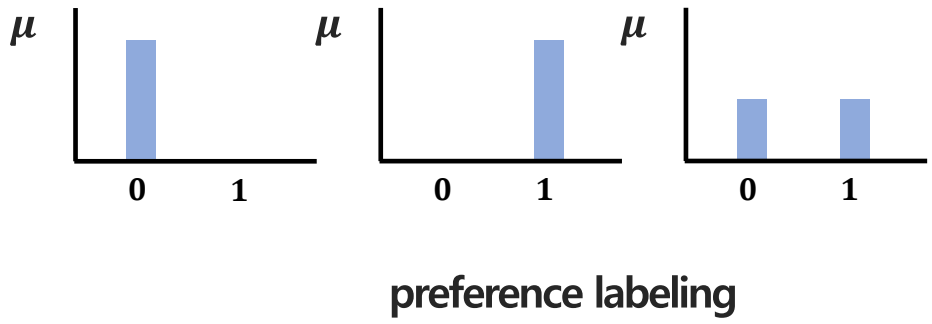
# Preliminaries

## REMIND : PbRL Basics

❖ How to define preference?

- **Trajectory Segment $\sigma^i$** : Sequence of state-action pairs $(s_t, a_t)$
- **Query** : 두 Trajectory Segment 사이의 선호도를 질문하는 것
- **Preference Annotation** : 수집된 경로들 중 두 Trajectory Segment를 추출하여 비교하고, 선호도를 레이블링($\mu$) 하는 것
  - ✓ $(\sigma^0, \sigma^1, \mu)$ 로 이루어진 Preference Dataset 구성
  - ✓ Preference Dataset은 보상 함수를 추정($\hat{r}$)하는데 쓰임



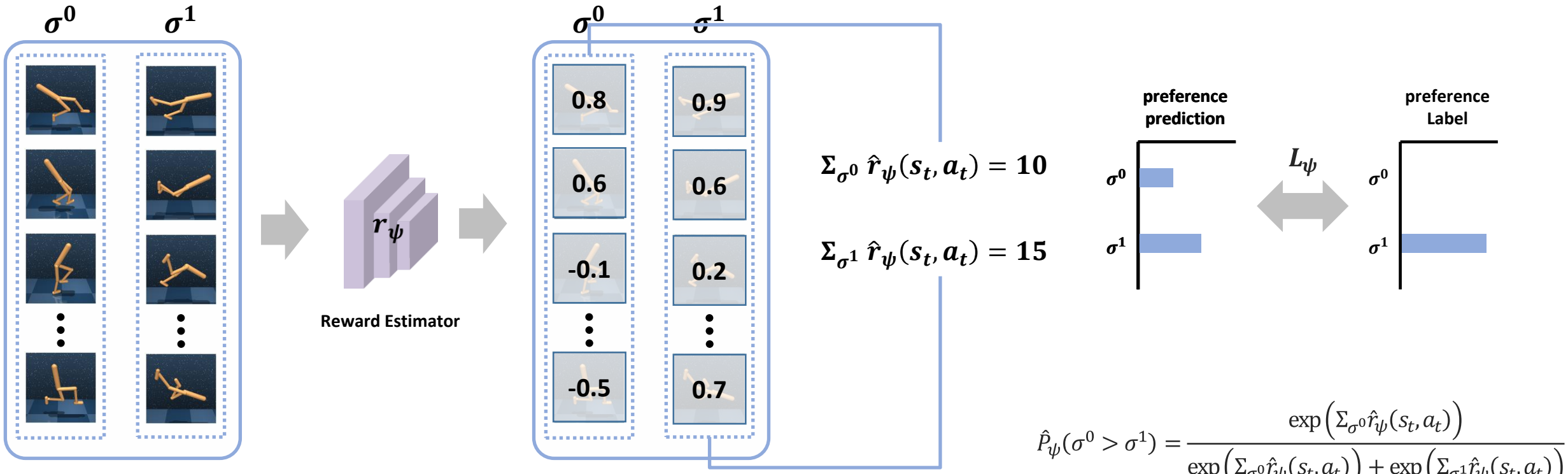




Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30.

# Preliminaries

## REMIND : PbRL Basics

❖ Fitting Reward Function with Human Preferences – Bradley Terry Model




$$\hat{P}_\psi(\sigma^0 > \sigma^1) = \frac{\exp\left(\Sigma_{\sigma^0}\hat{r}_\psi(s_t, a_t)\right)}{\exp\left(\Sigma_{\sigma^0}\hat{r}_\psi(s_t, a_t)\right) + \exp\left(\Sigma_{\sigma^1}\hat{r}_\psi(s_t, a_t)\right)}$$

$$L_\psi = -\Sigma_{(\sigma^0,\sigma^1,y)\in D}(y(0)log\hat{P}_\psi(\sigma^0 > \sigma^1) + y(1)log\hat{P}_\psi(\sigma^0 < \sigma^1))$$

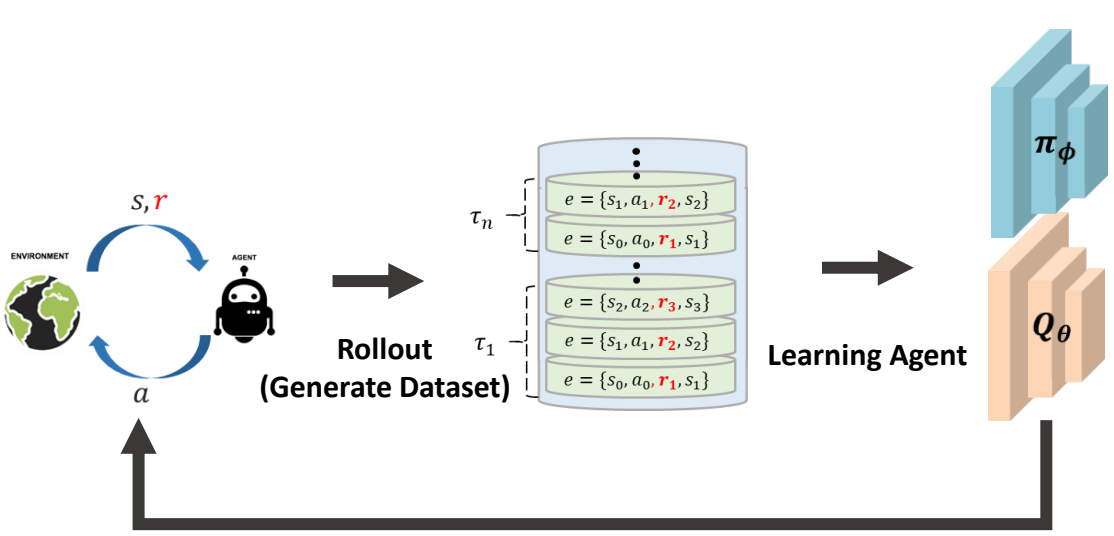

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30.

# Preliminaries

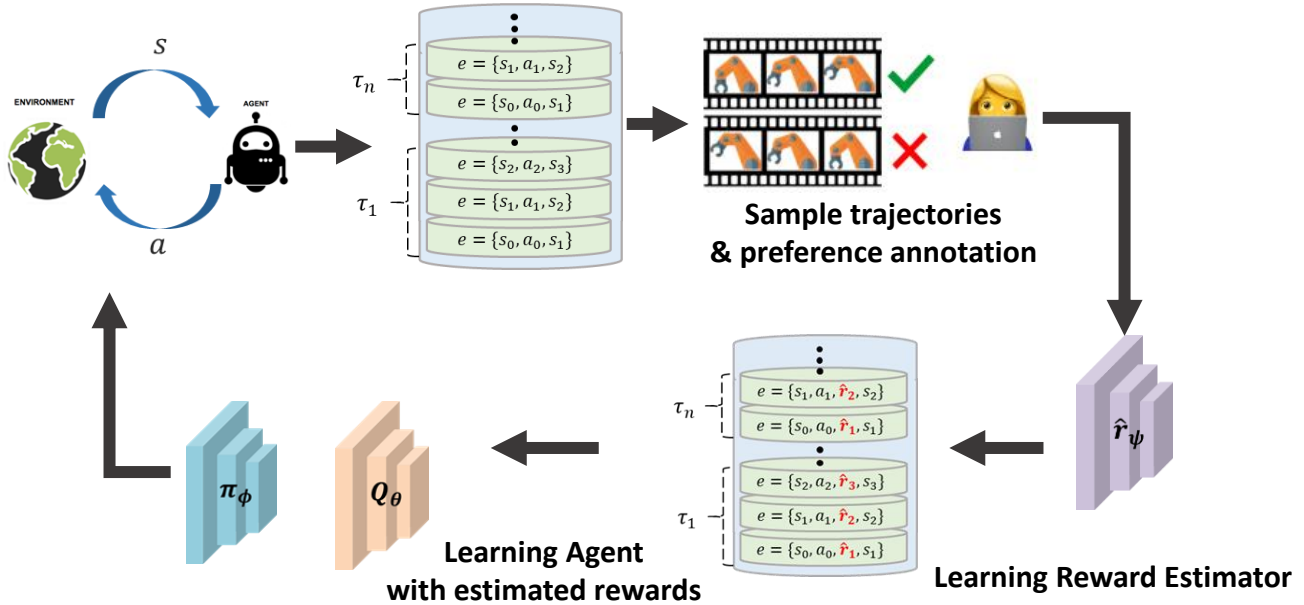
## REMIND : PbRL Basics

### ❖ Online Setting

- Online RL : 에이전트가 **환경과 상호 작용하며 데이터를 수집**하며 가치함수 $Q_\theta$ 와 정책함수 $\pi_\phi$를 학습 (시뮬레이터 필요o)
  - ✓ PPO, A3C, DrQ-v2, SAC, TD3…
- Online PbRL : 에이전트가 **환경과 상호 작용하며 보상이 없는 데이터를 수집, 선호도 레이블링 및 보상 함수 학습**과 에이전트 학습을 반복




**Online RL**



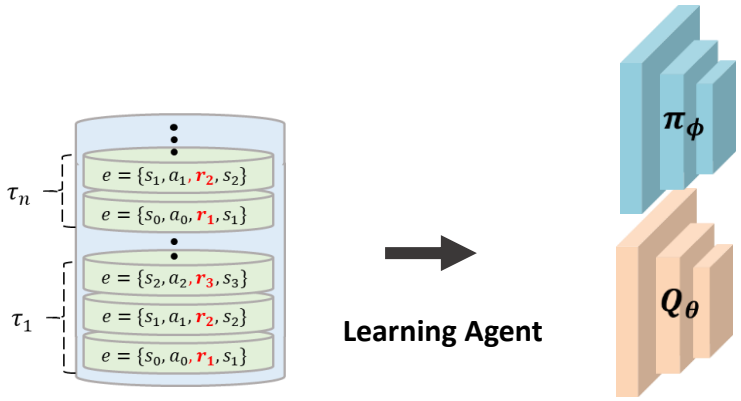

**Online PbRL**

# Preliminaries

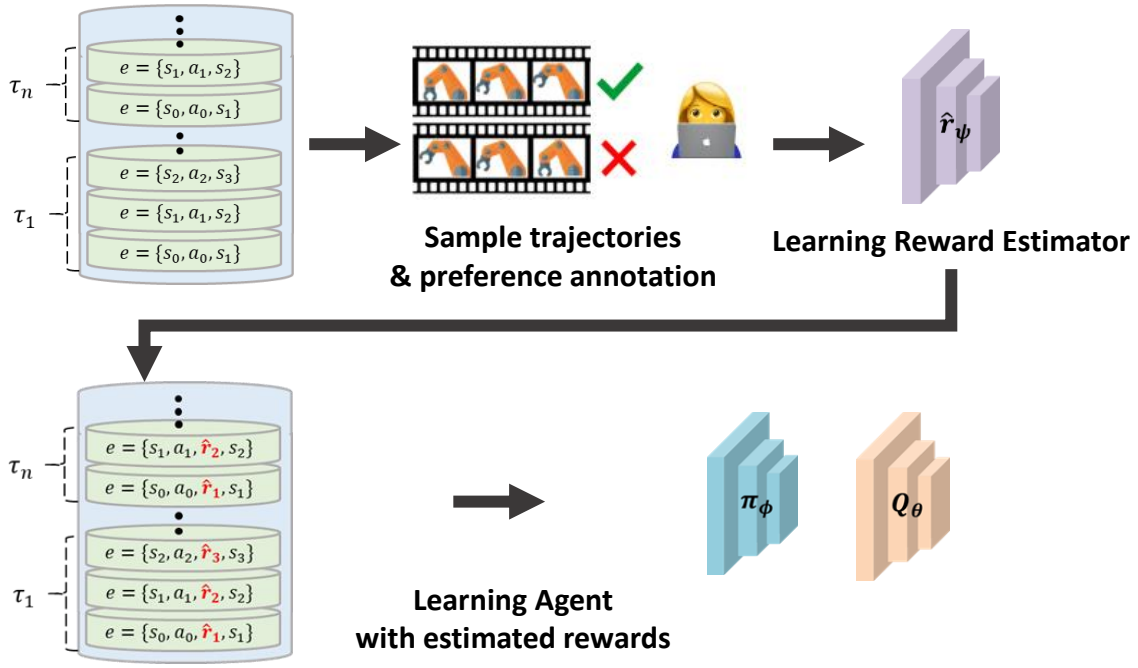
## REMIND : PbRL Basics

### ❖ Offline Setting

- Offline RL : **사전에 수집된 데이터셋** (보상 o)을 통해 에이전트를 학습 (시뮬레이터 필요 x)
  - ✓ CQL, IQL, XQL, AWR, AWAC....
- Offline PbRL : **사전에 수집된 데이터셋** (보상 x)에 사람이 레이블링을 하여 보상 함수 $\hat{r}$를 먼저 학습 후, 에이전트를 학습




**Offline RL**




**Offline PbRL**

# Advanced Methods

PrefPPO/PrefA3C (2017 NeurIPS)

Few-shot Preference Learning (2022 CoRL)

Language Instructed RL for Human AI Coordination (2023 ICML)

DPO (2023 NeurIPS)

RIME (2024 ICML)

PrefMMT (2025 IROS)

Reward Learning with Demonstrations (2018 NeurIPS)

Meta-Reward Net (2022 NeurIPS)

OPPO (2023 ICML)

SeqRank (2023 NeurIPS)

FuRL (2024 ICML)

APPO (2025 ICLR)

PEBBLE (2021 ICML)

MIL NRM (2022 NeurIPS)

OPRL (2023 TMLR)

RoboCLIP (2023 NeurIPS)

LiRE (2024 ICML)

Sim OPRL (2025 ICLR)

Multimodal Rewards from Rankings (2021 CoRL)

InsturctGPT (2022 NeurIPS)

QDP-HRL (2023 IEEE TNNLS)

Diverse Human Preferences (2024 IJCAI)

RL VLM F (2024 ICML)

Beyond Bradley-Terry (2025 ICML)

SkiP (2021 CoRL)

Preference Transformer (2023 ICLR)

REED (2023 CoRL)

CPL (2024 ICLR)

Adaptive Preference Scaling (2024 NeurIPS)

S-EPOA (2025 IJCAI)

SURF (2022 ICLR)

Causal Confusion and Reward Misidentification (2023 ICLR)

DPPO (2023 NeurIPS)

QPA (2024 ICLR)

RL saLLM F (2025 AAMAS)

CLARIFY (2025 ICML)

RUNE (2022 ICLR)

Reward Design with Language Models (2023 ICLR)

IPL (2023 NeurIPS)

Eureka (2024 ICLR)

PrefCLM (2025 IEEE Robotics & Automation Letters)

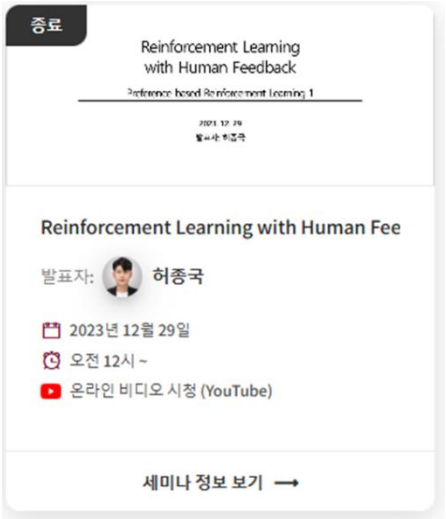
TREND (2025 ICRA)

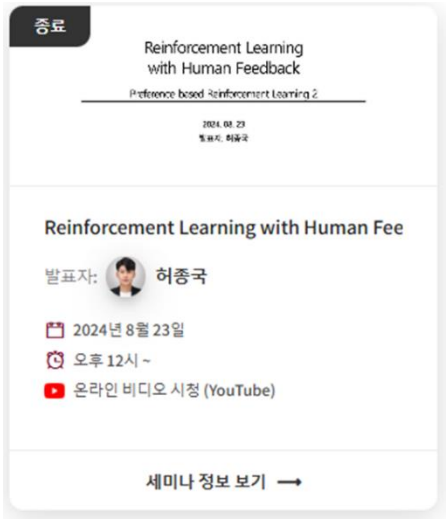
# Advanced Methods

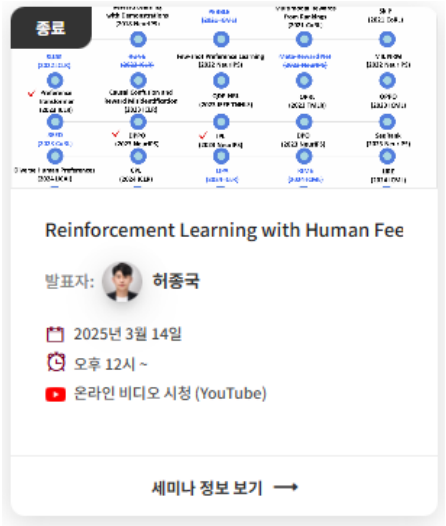
## Review

### ❖ Details

- RLHF - reference-based Reinforcement Learning 1
  - ✓ PrefPPO/A3C, PEBBLE, SURF, RUNE (Online PbRL Algorithms)
- RLHF : Preference-based Reinforcement Learning 2
  - ✓ Meta-Reward Net, REED, QPA, RIME (Online PbRL Algorithms)
- RLHF : Preference-based Reinforcement Learning 3
  - ✓ Preference Transformer, DPPO, IPL (Offline PbRL Algorithms)




http://dmqa.korea.ac.kr/activity/seminar/435
http://dmqa.korea.ac.kr/activity/seminar/463

# Advanced Methods

## SeqRank

❖ Sequential Preference Ranking for Efficient Reinforcement Learning from Human Feedback (Hwang et al., NeurIPS 2023)

- 기존 Online PbRL 방법론들이 **독립적으로 쿼리를 추출하기 때문에 하나의 피드백 당 하나의 정보밖에 얻을 수 없음을 지적**
- 독립적인 쿼리 추출 방식이 아니라 **순차적인 쿼리 추출 방식**을 제안
- N 개의 선호 데이터를 얻기 위해 필요한 원본 trajectory의 개수를 **피드백 효율성**으로 정의

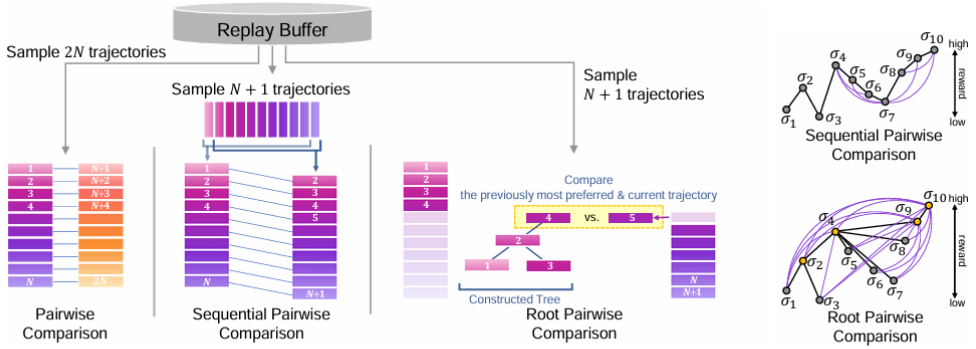


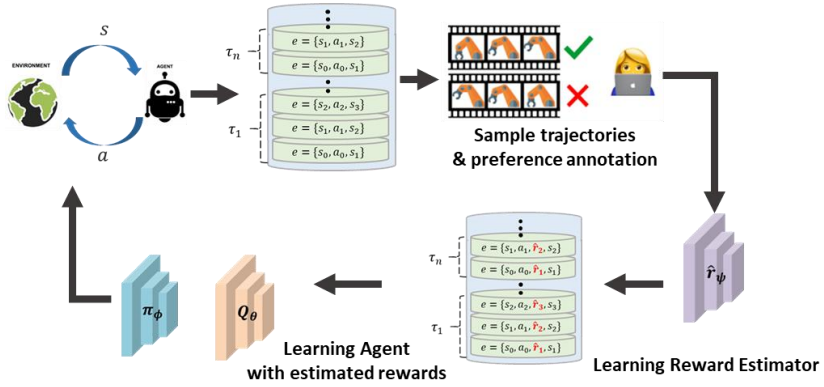

**Figure 2: Trajectory comparison methods.** (a) All three methods query a human to provide $N$ feedback. Pairwise comparison samples $2N$ trajectories to obtain $N$ feedback, while sequential or root pairwise comparison samples $N + 1$ trajectories. Despite sampling fewer trajectories, sequential or root pairwise comparison shows higher feedback efficiency than pairwise comparison by adopting sequential preference ranking. (b) Gray nodes illustrate fixed-length trajectory segments sampled from the replay buffer. Suppose the reward values for segments $\sigma_1, ..., \sigma_{10}$ are $2, 5, 1, 8, 6, 4, 3, 7, 9, 10$, respectively. Black lines indicate actual pairs that receive true preference labels from human feedback. The upper node for each black line represents the preferred trajectory. For root pairwise comparison, orange nodes ● describe non-leaf nodes in the tree. Using sequential or root pairwise comparison, the agent can obtain augmented labels for non-adjacent pairs illustrated with purple lines.


Hwang, M., Lee, G., Kee, H., Kim, C. W., Lee, K., & Oh, S. (2023). Sequential preference ranking for efficient reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, *36*, 49088-49099.

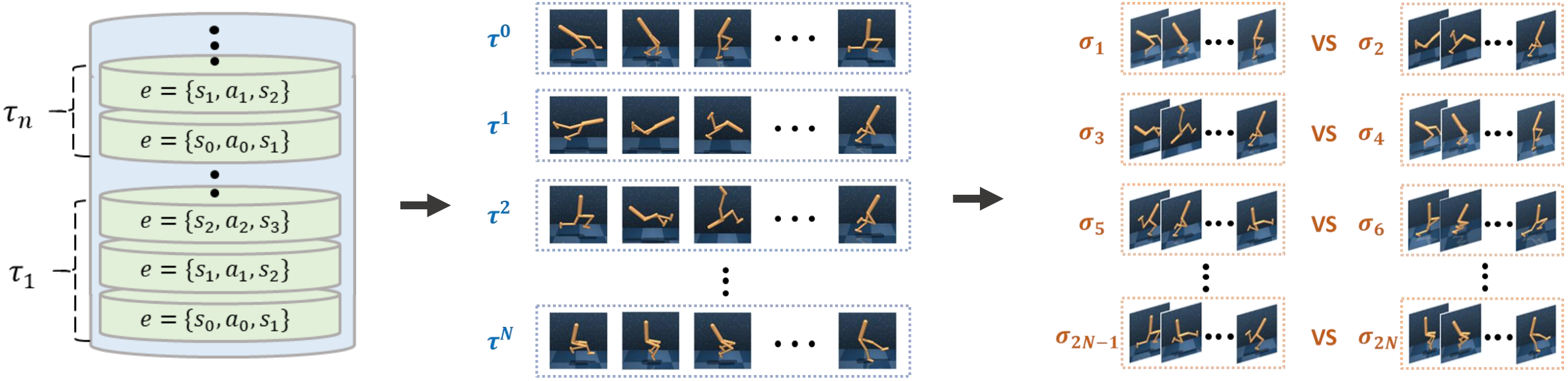
# Advanced Methods

## SeqRank

- Previous sampling schemes (Uniform Sampling)
  - Replay buffer에 존재하는 여러 개의 에피소드 중 **무작위로 N개의 쿼리**를 추출
  - N 개의 쿼리는 **독립적인 2N개의 trajectory**로 구성되어 있음




**Online PbRL**



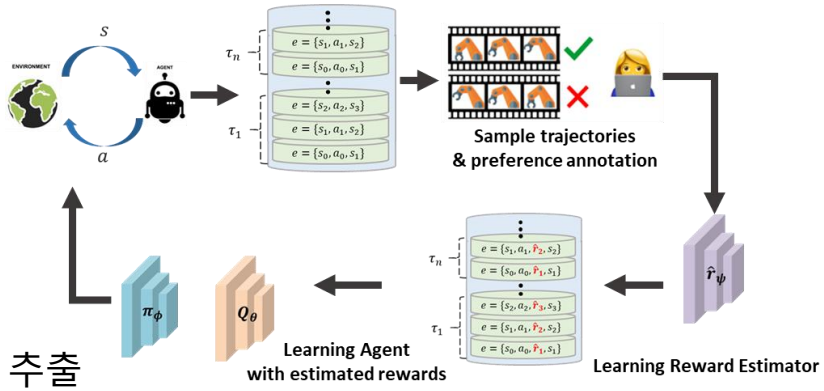

**Replay Buffer** **Collected Episodes** **Sample Queries for Annotation**


Hwang, M., Lee, G., Kee, H., Kim, C. W., Lee, K., & Oh, S. (2023). Sequential preference ranking for efficient reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, *36*, 49088-49099.

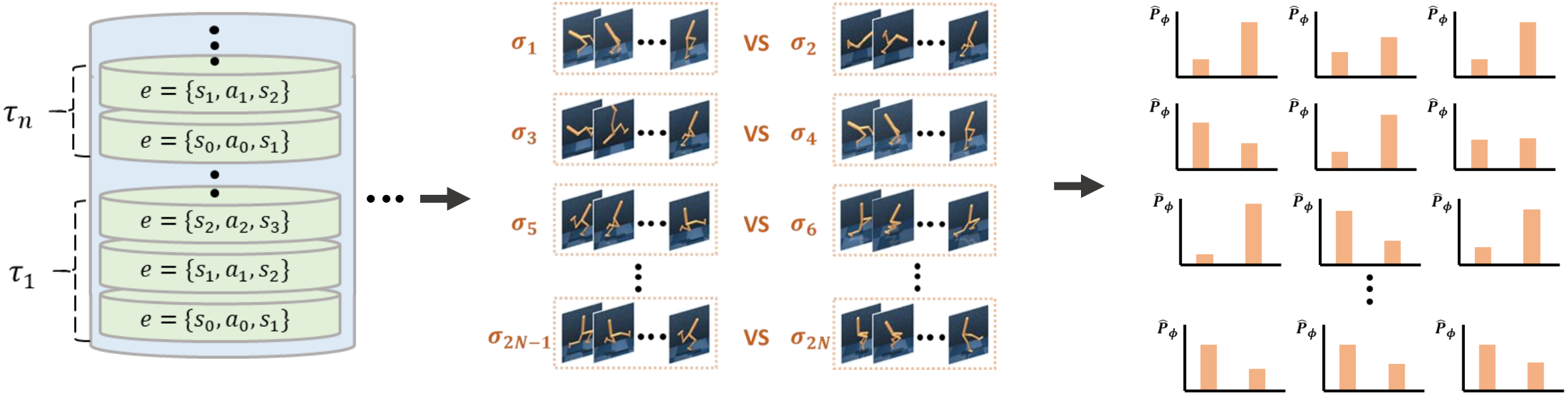
# Advanced Methods

## SeqRank

- Previous sampling schemes (Disagreement Sampling)
  - Replay buffer에 존재하는 여러 개의 에피소드 중 **불확실성이 높은 N개의 쿼리**를 추출
  - N 개의 쿼리는 **독립적인 2N개의 trajectory**로 구성되어 있음




**Online PbRL**




**Replay Buffer** **Sample Initial Batch of Queries** **Ensemble Uncertainty**

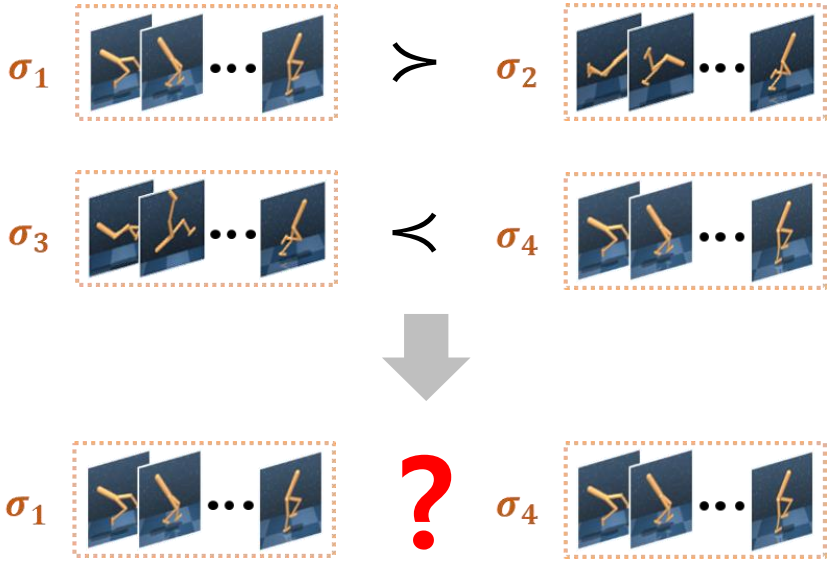

Hwang, M., Lee, G., Kee, H., Kim, C. W., Lee, K., & Oh, S. (2023). Sequential preference ranking for efficient reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, *36*, 49088-49099.

# Advanced Methods

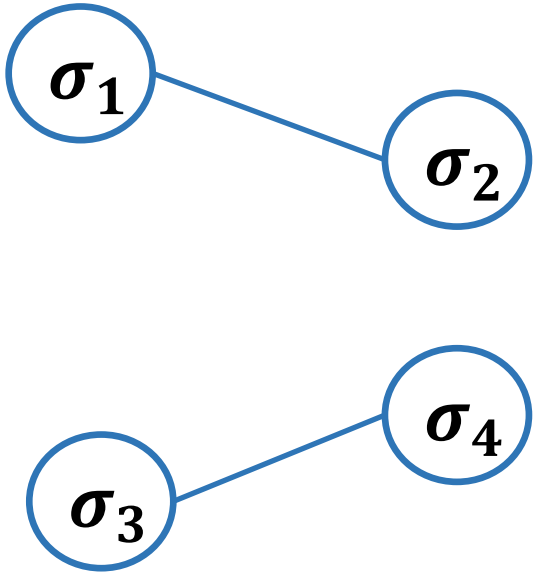
## SeqRank

❖ Drawback of previous sampling schemes

- N 개의 쿼리는 독립적인 2N개의 trajectory로 구성되어 있음
- 다시 말해, **2N개의 trajectory 에서 얻을 수 있는 정보는 N개**




**Independent Queries**





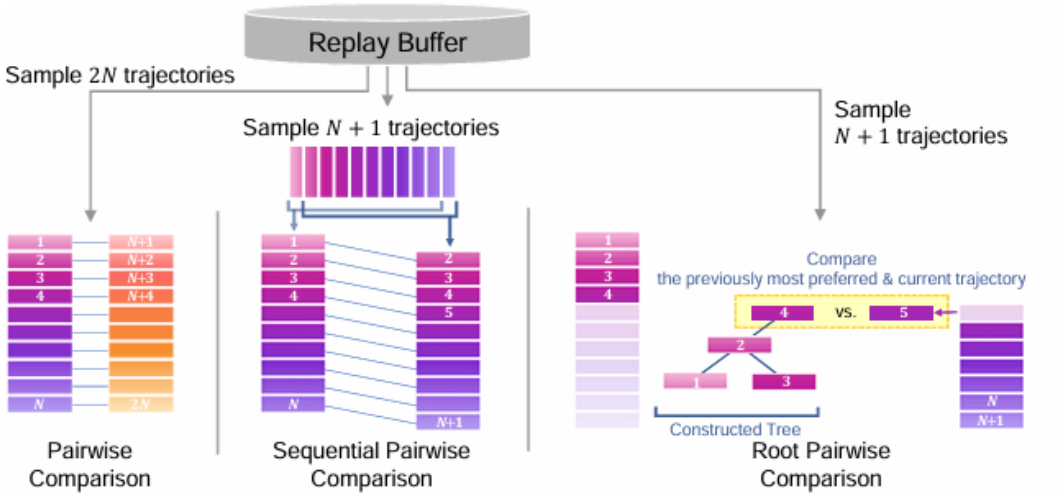
Hwang, M., Lee, G., Kee, H., Kim, C. W., Lee, K., & Oh, S. (2023). Sequential preference ranking for efficient reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, *36*, 49088-49099.

# Advanced Methods

## SeqRank

### ❖ SeqRank

- Pairwise Comparison : 기존 샘플링 방식으로, N개의 피드백 데이터를 얻기 위해 2N개의 trajectory가 필요
- **Sequential Pairwise Comparison** : 순차적으로 피드백 비교, N개의 피드백 데이터를 얻기 위해 N+1개의 trajectory가 필요
- **Root Pairwise Comparison** : 순차적 + 가장 좋았던 것과 비교, N개의 피드백 데이터를 얻기 위해 N+1개의 trajectory가 필요
- 제안하는 2가지 비교 방식은 N개의 피드백 뿐만 아니라 연관성으로 인한 추가 정보 획득 가능




(a) Three trajectory comparison methods.

Hwang, M., Lee, G., Kee, H., Kim, C. W., Lee, K., & Oh, S. (2023). Sequential preference ranking for efficient reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, *36*, 49088-49099.

# Advanced Methods

## SeqRank

### ❖ Feedback Efficiency of SeqRank

- $\boldsymbol{M}$ : $\boldsymbol{N}$개의 인간 피드백 레이블을 얻기 위해 필요한 trajectory의 개수
- $\boldsymbol{p_N}$ : $\boldsymbol{N}$개의 인간 피드백 레이블로부터 얻을 수 있는 정보(# of labeled queries, preferences)의 개수
- $\boldsymbol{\eta}$ : 피드백 효율성 $(\boldsymbol{p_N}/\boldsymbol{N})$

| Method | $M$ | Best | | Average | | Worst | |
|---|---|---|---|---|---|---|---|
| | | $p_N$ | $\eta$ | $p_N$ | $\eta$ | $p_N$ | $\eta$ |
| 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 |

**Table 1: Trajectory comparison methods.** We compare three trajectory comparison methods: pairwise, sequential pairwise, and root pairwise. $\eta$ denotes the feedback efficiency.

Hwang, M., Lee, G., Kee, H., Kim, C. W., Lee, K., & Oh, S. (2023). Sequential preference ranking for efficient reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, *36*, 49088-49099.

# Advanced Methods

## SeqRank

### ❖ Experimental Results

- DMControl 과 Meta-world 환경에서 실험
- Upper Bound : SAC trained with true reward (Oracle)
- Lower Bound : MRN trained with pairwise preference queries (Pairwise)

| Task | # feedback | Oracle | Pairwise | Sequential Pairwise | Root Pairwise |
|---|---|---|---|---|---|
| Walker Walk | 0.4K | 957.3 ± 2.3 | 862.5 ± 85.3 | 883.7 ± 71.9 | **907.5** ± 66.2 |
| Cheetah Run | 0.2K | 886.9 ± 43.2 | 690.2 ± 93.2 | 715.8 ± 117.4 | **739.4** ± 100.8 |
| Quadruped Walk | 1K | 843.3 ± 247.3 | 353.5 ± 183.9 | 479.5 ± 234.7 | **728.8** ± 274.1 |
| Humanoid Walk | 40K | 272.5 ± 117.0 | 114.3 ± 80.5 | 141.3 ± 34.4 | **163.8** ± 71.6 |
| Hopper Hop | 4K | 273.5 ± 47.0 | 26.2 ± 37.0 | 37.3 ± 59.8 | **100.1** ± 70.7 |

**Table 2: Rewards after convergence in DMControl locomotion tasks.** Two elements in each cell denote the average value and standard deviation of rewards across runs with 10 random seeds.

| Task | # feedback | Oracle | Pairwise | Sequential Pairwise | Root Pairwise |
|---|---|---|---|---|---|
| Button Press | 10K | 99.3 ± 0.9 | 95.6 ± 7.6 | 97.0 ± 4.3 | **97.6** ± 5.6 |
| Door Open | 10K | 100.0 ± 0.0 | 77.9 ± 41.3 | 77.4 ± 40.8 | **97.8** ± 5.0 |
| Drawer Open | 20K | 99.9 ± 0.3 | 65.3 ± 41.6 | 72.0 ± 45.4 | **89.9** ± 31.6 |
| Sweep Into | 10K | 88.8 ± 29.9 | 68.4 ± 35.2 | 55.7 ± 48.1 | **88.0** ± 31.0 |
| Window Open | 1K | 99.9 ± 0.3 | 55.9 ± 45.1 | 66.1 ± 36.5 | **70.7** ± 40.8 |
| Hammer | 10K | 91.5 ± 26.5 | 31.0 ± 24.7 | 30.8 ± 35.4 | **48.8** ± 41.2 |

**Table 3: Success rates in Meta-World manipulation tasks.** Two elements in each cell denote the average value and standard deviation of success rates across runs with 10 random seeds.

Hwang, M., Lee, G., Kee, H., Kim, C. W., Lee, K., & Oh, S. (2023). Sequential preference ranking for efficient reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, *36*, 49088-49099.

# Advanced Methods

## SeqRank

### ❖ Experimental Results

- UR-5 로봇 팔을 이용한 블록 옮기기 실험
- Discrete action space에서 SAC 대신 DQN을 활용

| Method | # feedback | Success Rate↑ | True Reward↑ | Episode Length↓ | Reward Accuracy↑ |
|---|---|---|---|---|---|
| Pairwise | 0.8K | 34.4 | 17.1 | 29.6 | 75.8 |
| Sequential Pairwise | 0.8K | 41.0 | 20.1 | 28.4 | 80.9 |
| Root Pairwise | 0.8K | **50.5** | **25.9** | **27.9** | **84.9** |

**Table 4: Average performance in the real robot manipulation task.** For each method, the agent is fine-tuned in the real world for 3,000 steps.

**Figure 5: Block placing using a UR-5 robot.**

Hwang, M., Lee, G., Kee, H., Kim, C. W., Lee, K., & Oh, S. (2023). Sequential preference ranking for efficient reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, *36*, 49088-49099.

# Advanced Methods

## LiRE

❖ Listwise Reward Estimation for Offline Preference-based Reinforcement Learning (Choi et al., ICML 2024)

- Offline PbRL에서 개별적인 선호 데이터가 독립적이기 때문에, 선호도 간의 상대적 강도를 고려한 second-order preference를 고려할 수 없음을 지적
- SeqRank와 마찬가지로 독립적인 선호 데이터 구축이 아닌, 다른 데이터와의 관계를 구축하여 ranked list of trajectories (RLT)를 구축

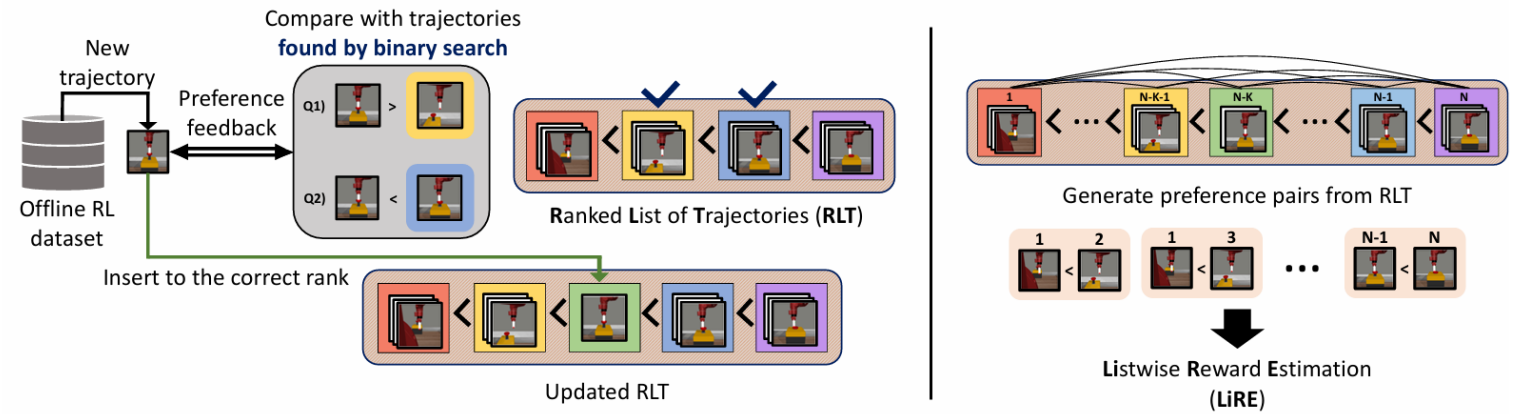



Figure 1: **An overview of LiRE.** The figure shows an example of a *button-press-topdown* task. We sample a trajectory segment and sequentially obtain the preference feedback for existing trajectories in RLT. We use binary search to find the correct rank (left) efficiently. Multiple preference pairs are generated from RLT to learn the reward model (right).

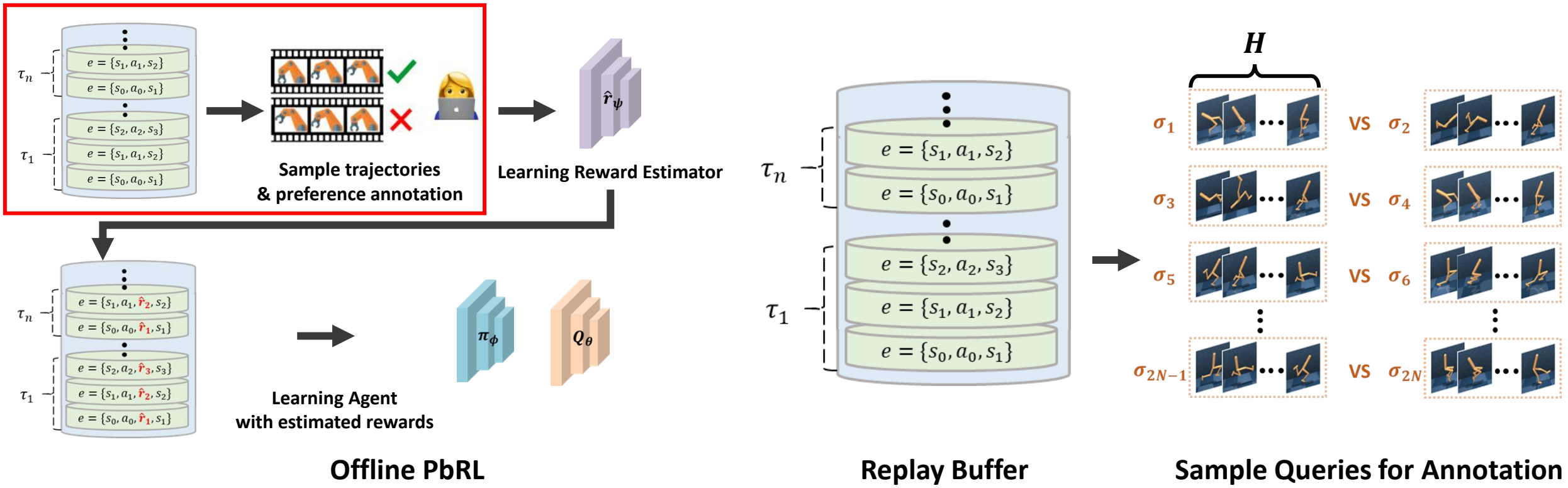

Choi, H., Jung, S., Ahn, H., & Moon, T. (2024, July). Listwise Reward Estimation for Offline Preference-based Reinforcement Learning. In *International Conference on Machine Learning* (pp. 8651-8671). PMLR.

# Advanced Methods

## LiRE

❖ Listwise Reward Estimation for Offline Preference-based Reinforcement Learning (Choi et al., ICML 2024)

- Offline PbRL은 Online PbRL과 달리 사전에 수집된 선호 데이터를 활용하여 정책 함수를 최적화하는 분야





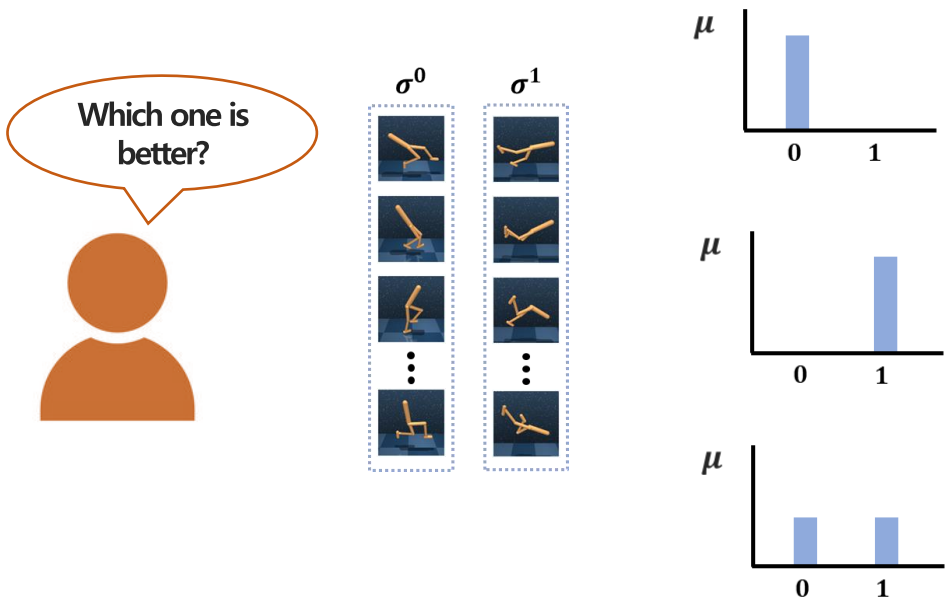
Choi, H., Jung, S., Ahn, H., & Moon, T. (2024, July). Listwise Reward Estimation for Offline Preference-based Reinforcement Learning. In *International Conference on Machine Learning* (pp. 8651-8671). PMLR.

# Advanced Methods

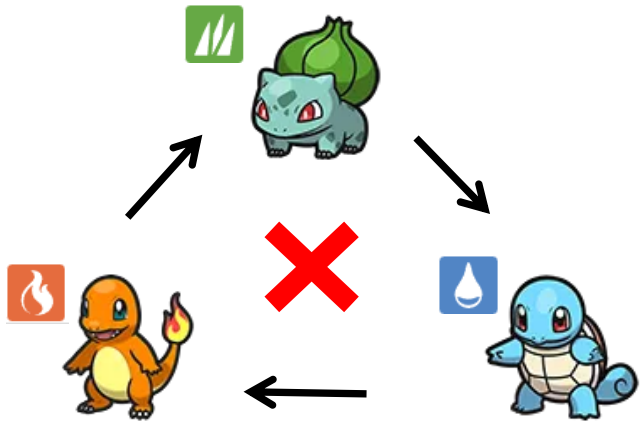
## LiRE

### ❖ Assumption

- **Completeness Assumption** : 인간 피드백은 (0, 1), (1, 0), **(0.5, 0.5)** 셋 중 하나만 가능
- **Transitivity Assumption** : $A < B \,\&\, B < C \rightarrow A < C$를 만족하거나 $\boldsymbol{A = B \,\&\, B = C \rightarrow A = C}$를 만족해야함




**Completeness Assumption**



**Transitivity Assumption**

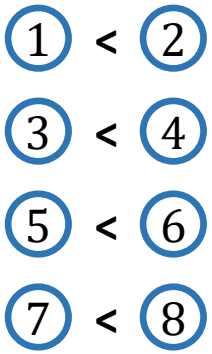
Choi, H., Jung, S., Ahn, H., & Moon, T. (2024, July). Listwise Reward Estimation for Offline Preference-based Reinforcement Learning. In *International Conference on Machine Learning* (pp. 8651-8671). PMLR.

# Advanced Methods

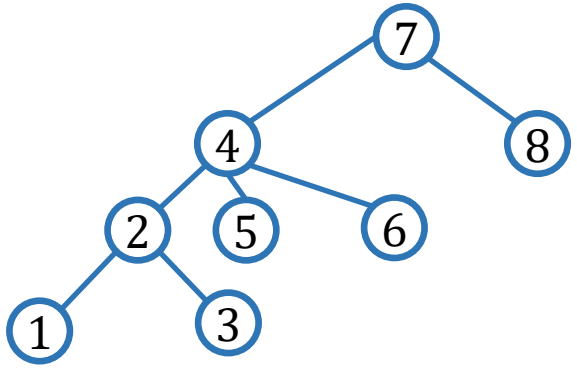
## LiRE

❖ Construction of Ranked List of Trajectories (RLT)

- SeqRank : 순차적 샘플링을 활용하여 **다른 데이터와의 선호도를 일부 추출 가능**
- LiRE : 순차적 샘플링을 활용하여 **다른 데이터와의 선호도를 모두 추출 가능**




**Independent**
**(4 preferences/8 data)**




**SeqRank(Root Pairwise)**
**(14 preferences/8 data)**


**LiRE**
**(28 preferences/8 data)**

Choi, H., Jung, S., Ahn, H., & Moon, T. (2024, July). Listwise Reward Estimation for Offline Preference-based Reinforcement Learning. In *International Conference on Machine Learning* (pp. 8651-8671). PMLR.

# Advanced Methods

## LiRE

### ❖ Construction of Ranked List of Trajectories (RLT)

- Process of RLT

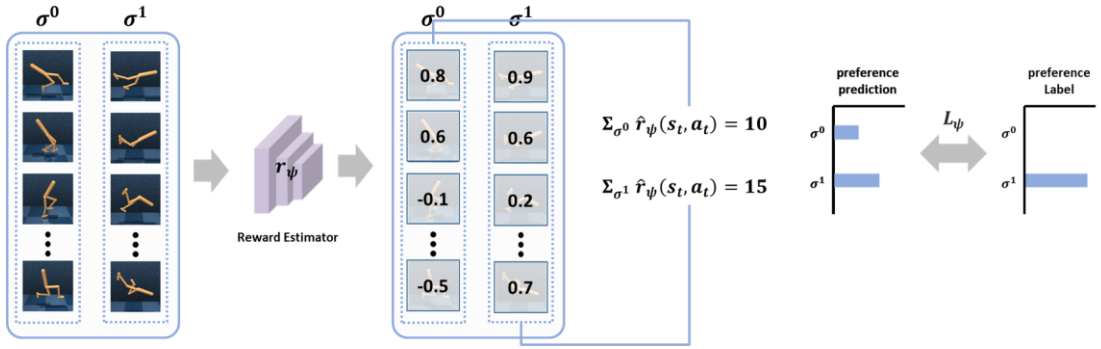
**Algorithm 1** RLT Construction

**function** BINARYSEARCH $(\sigma, \texttt{low}, \texttt{high}, L)$ :
  **if** $\texttt{low} = \texttt{high}$ **then**
    insert a new group $\{\sigma\}$ to $L$ right behind to $g_{\texttt{low}+1}$
    (i.e., $g_{\texttt{low}} \prec \{\sigma\} \prec g_{\texttt{low}+1}$)
  **else**
    /* Human Feedback */
    compare $\sigma$ to $\sigma_s \in g_{\texttt{mid}}$ where $\texttt{mid} = \lceil \frac{\texttt{low}+\texttt{high}}{2} \rceil$
    **if** $\sigma_s \prec \sigma$ **then**
      BINARYSEARCH$(\sigma, \texttt{mid}, \texttt{high}, L)$
    **else if** $\sigma \prec \sigma_s$ **then**
      BINARYSEARCH$(\sigma, \texttt{low}, \texttt{mid} - 1, L)$
    **else**
      $g_{\texttt{mid}} \leftarrow g_{\texttt{mid}} \cup \{\sigma\}$
**Init:** List $L = [\,]$
**repeat**
  sample $\sigma_1, \sigma_2, \cdots \in D_s$
  **if** $L$ is empty **then**
    $L \leftarrow [\{\sigma_i\}]$
  **else**
    BINARYSEARCH$(\sigma_i, 0, l, L)$
**until** end of feedback
**Output:** $L$

Choi, H., Jung, S., Ahn, H., & Moon, T. (2024, July). Listwise Reward Estimation for Offline Preference-based Reinforcement Learning. In *International Conference on Machine Learning* (pp. 8651-8671). PMLR.

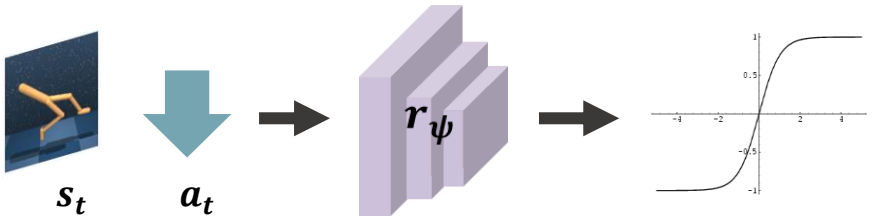
# Advanced Methods

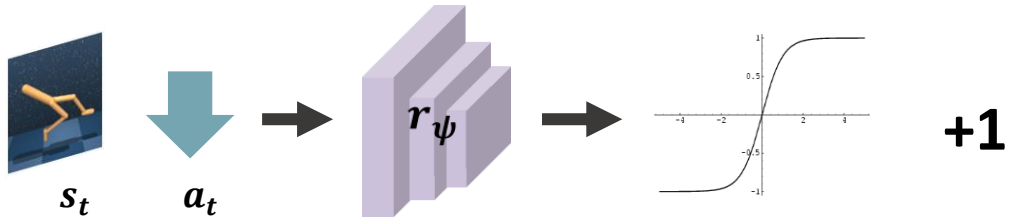
## LiRE




❖ (Additional) Bounding Reward Model

- **Reward Model with Exponential Logit :** 보상의 범위가 -1~1 사이기 때문에 positive logit을 위해 exponential 사용
- **Reward Model with Linear Logit** : 보상에 1을 더해 positive logit으로 만든 후 그대로 사용
- Linear Logit이 Exponential Logit 보다 선호도로 인한 보상의 차이를 더욱 극대화할 수 있어서 학습에 용이




$$\hat{P}_\psi(\sigma^0 > \sigma^1) = \frac{\exp\left(\Sigma_{\sigma^0}\hat{r}_\psi(s_t, a_t)\right)}{\exp\left(\Sigma_{\sigma^0}\hat{r}_\psi(s_t, a_t)\right) + \exp\left(\Sigma_{\sigma^1}\hat{r}_\psi(s_t, a_t)\right)}$$

**Reward Estimator with Exponential Logit**

$$\hat{P}_\psi(\sigma^0 > \sigma^1) = \frac{\Sigma_{\sigma^0}(\hat{r}_\psi(s_t, a_t) + 1)}{\Sigma_{\sigma^0}(\hat{r}_\psi(s_t, a_t) + 1) + \Sigma_{\sigma^1}(\hat{r}_\psi(s_t, a_t) + 1)}$$

**Reward Estimator with Linear Logit**


Choi, H., Jung, S., Ahn, H., & Moon, T. (2024, July). Listwise Reward Estimation for Offline Preference-based Reinforcement Learning. In *International Conference on Machine Learning* (pp. 8651-8671). PMLR.

# Advanced Methods

## LiRE

❖ Experimental Results

- Metaworld 환경에서 실험
- Upper Bound : IQL trained with true Reward
- Comparison with : Preference Transformer (PT), DPPO, IPL, SeqRank, OPRL

| # of feedbacks | Algorithm | button-press -topdown | box-close | dial-turn | sweep | button-press -topdown-wall | sweep-into | drawer-open | lever-pull |
|---|---|---|---|---|---|---|---|---|---|
| - | IQL with GT rewards | 88.33 ± 4.76 | 93.40 ± 3.10 | 75.40 ± 5.47 | 98.33 ± 1.87 | 56.27 ± 6.32 | 78.80 ± 7.96 | 100.00 ± 0.00 | 98.47 ± 1.77 |
| 500 | MR | 9.60 ± 5.74 | 10.33 ± 8.23 | 50.20 ± 8.51 | 79.80 ± 13.36 | 0.13 ± 0.50 | 24.80 ± 5.28 | 98.07 ± 3.20 | 50.53 ± 8.55 |
| | PT (Kim et al., 2022) | 22.87 ± 9.06 | 0.33 ± 1.16 | **68.67** ± 12.39 | 43.07 ± 24.57 | 0.87 ± 1.43 | 20.53 ± 8.26 | 88.73 ± 11.64 | 82.40 ± 22.69 |
| | OPRL (Shin et al., 2022) | 12.13 ± 5.75 | 4.73 ± 3.24 | 54.33 ± 11.47 | **94.13** ± 5.95 | 0.20 ± 0.60 | 25.87 ± 8.58 | 94.13 ± 6.41 | 54.67 ± 12.79 |
| | DPPO (An et al., 2023) | 3.93 ± 4.34 | 10.20 ± 11.47 | 26.67 ± 22.23 | 10.47 ± 15.84 | 0.80 ± 1.51 | 23.07 ± 7.02 | 35.93 ± 11.18 | 10.13 ± 12.19 |
| | IPL (Hejna & Sadigh, 2024) | **34.73** ± 13.92 | 5.93 ± 5.81 | 31.53 ± 12.50 | 27.20 ± 23.81 | **8.93** ± 9.84 | **32.20** ± 7.35 | 19.00 ± 13.63 | 31.20 ± 15.76 |
| | SeqRank (Hwang et al., 2023) | 17.6 ± 11.94 | **13.2** ± 12.72 | 65.6 ± 12.84 | **83.4** ± 9.76 | 1.73 ± 1.98 | 25.67 ± 11.02 | **99.53** ± 0.36 | **95.67** ± 4.04 |
| | LiRE (ours) | **67.20** ± 18.97 | **51.53** ± 18.48 | **79.07** ± 10.96 | 77.53 ± 10.50 | **79.13** ± 15.19 | **49.13** ± 15.85 | **99.40** ± 1.65 | **95.67** ± 6.26 |
| 1000 | MR | 9.27 ± 5.30 | 17.07 ± 9.56 | 59.07 ± 7.57 | **90.80** ± 9.74 | 0.60 ± 1.87 | 26.07 ± 8.57 | 96.47 ± 4.02 | 50.87 ± 10.89 |
| | PT (Kim et al., 2022) | 18.27 ± 10.62 | 2.27 ± 2.86 | 68.80 ± 5.50 | 29.13 ± 14.55 | 2.13 ± 2.96 | 20.27 ± 7.84 | 95.40 ± 7.27 | 72.93 ± 10.16 |
| | OPRL (Shin et al., 2022) | 11.00 ± 7.84 | 15.07 ± 11.19 | 51.33 ± 10.08 | **85.53** ± 5.43 | 0.33 ± 0.75 | 28.27 ± 6.40 | **99.20** ± 1.42 | 53.20 ± 6.67 |
| | DPPO (An et al., 2023) | 3.20 ± 3.04 | 9.33 ± 9.60 | 36.40 ± 21.95 | 8.73 ± 16.37 | 0.27 ± 0.85 | 23.33 ± 7.80 | 36.47 ± 7.30 | 8.53 ± 9.96 |
| | IPL (Hejna & Sadigh, 2024) | **36.67** ± 17.40 | 6.73 ± 8.41 | 43.93 ± 13.37 | 38.33 ± 24.87 | **14.07** ± 11.47 | **30.40** ± 7.74 | 28.53 ± 18.37 | 40.40 ± 17.38 |
| | SeqRank (Hwang et al., 2023) | 13.93 ± 8.11 | **46.60** ±12.53 | **70.67** ± 7.58 | 74.93 ±22.35 | 2.47 ±2.67 | 29.33 ± 11.59 | 98.6 ± 3.92 | **95.47** ± 3.86 |
| | LiRE (ours) | **83.07** ± 6.38 | **89.13** ± 6.02 | **76.93** ± 7.55 | 75.87 ± 6.81 | **81.47** ± 10.04 | **57.73** ± 13.11 | **99.73** ± 0.85 | **99.47** ± 1.15 |

# Advanced Methods

## LiRE

❖ Experimental Results

- Metaworld 환경에서 실험
- Upper Bound : IQL trained with true Reward
- Comparison with : Preference Transformer (PT), DPPO, IPL, SeqRank, OPRL

### 5.1. Settings

**Dataset** Previous offline PbRL papers are evaluated mainly on D4RL (Fu et al., 2020), but D4RL has the problem that RL performance can be high even when wrong rewards are used (Li et al., 2023; Shin et al., 2022). To that end, we newly collect the offline PbRL dataset with Meta-World (Yu et al., 2020) and DeepMind Control Suite (DMControl) (Tassa et al., 2018) following the protocols of previous work: `medium-replay` dataset, *e.g.,* (Yu et al., 2021a; Mazoure et al., 2023; Gulcehre et al., 2020) and `medium-expert` dataset, *e.g.,* (Yu et al., 2021b; Sinha et al., 2022; Hejna & Sadigh, 2024; Li et al., 2023).

# Advanced Methods

## LiRE

❖ Experimental Results

- 개별 순위 정렬 리스트 당 budget의 개수를 조절해가며 실험

Table 4: Average success rates of LiRE when adjusting the $Q$ budget. We use a total of 500 preference feedbacks.

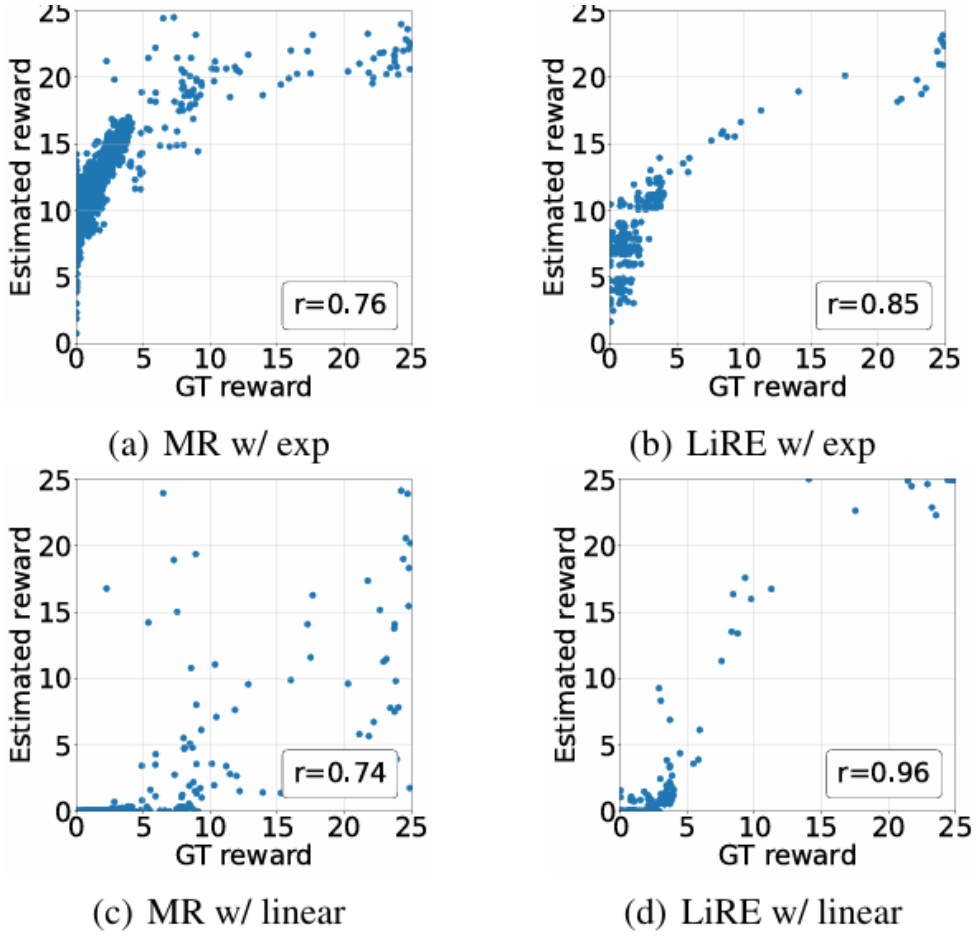
| Dataset | Q=1 (MR w/ linear) | Q=2 | Q=10 | Q=20 | Q=50 | Q=100 | Q=500 | SeqRank w/ linear |
|---|---|---|---|---|---|---|---|---|
| button-press-topdown | 36.87 ± 13.75 | 59.47 ± 2.18 | 53.60 ± 10.82 | 65.13 ± 14.24 | **71.26** ± 12.95 | 67.20 ± 18.97 | **77.67** ± 18.13 | 54.87 ± 9.89 |
| lever-pull | 70.20 ±18.03 | 69.80 ± 3.79 | 70.47 ±18.19 | 92.7 ±7.16 | 93.4 ±7.90 | **95.67** ± 6.26 | **99.33** ± 1.18 | 74.33 ± 18.50 |

# Advanced Methods

## LiRE

❖ Experimental Results

- 쿼리 추출 방식, 선호도 함수 모델링 방식에 따른 보상 함수 scatter plot



# Advanced Methods

## CPL

❖ Contrastive Preference Learning: Learning from Human Feedback without RL (Hejna et al., ICLR 2024)

- 기존 PbRL은 1. 인간 피드백을 활용하여 보상 함수를 학습하고, 2. 이를 통해 강화 학습 에이전트를 최적화하는 방식
- Offline 세팅에서 2단계 학습 방식과 강화학습 학습 과정에서의 high variance를 줄이고자, advantage를 기반으로 보상함수 없이 직접 에이전트를 학습
- 누적 보상 기반 선호도 모델링이 아닌 regret 기반 선호도 모델링 제안

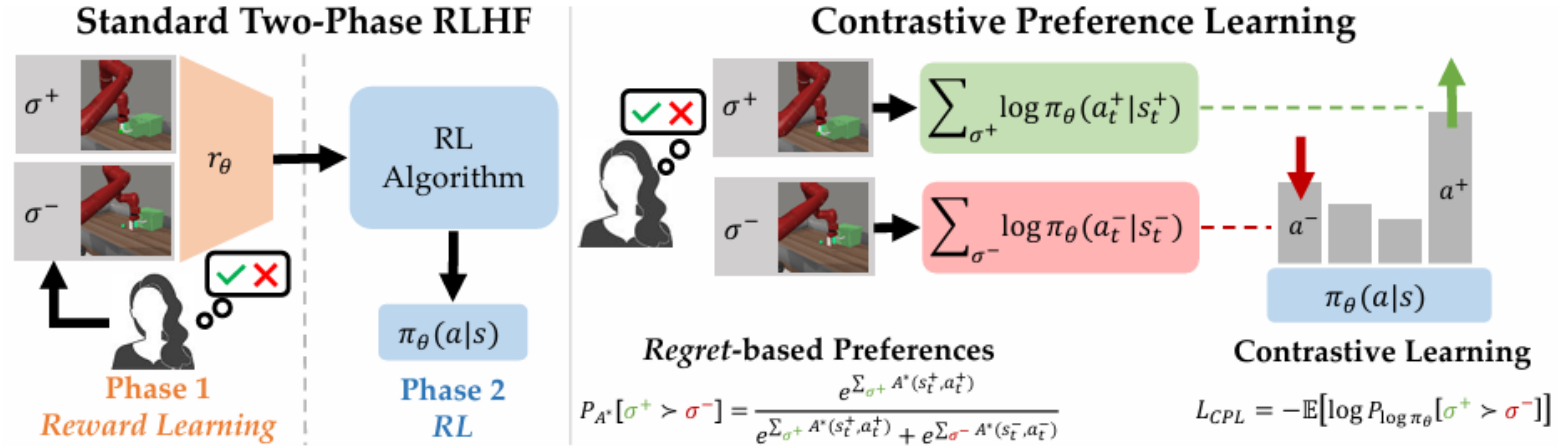



Figure 1: While most RLHF algorithms use a two-phase reward learning, then RL approach, CPL directly learns a policy using a contrastive objective. This is enabled by the regret preference model.

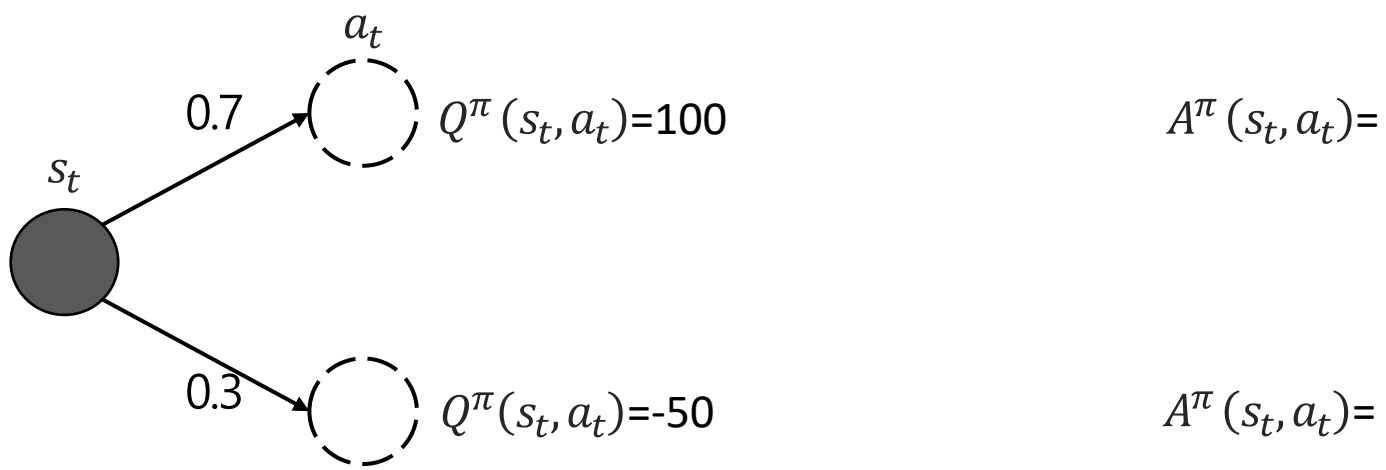

Hejna, J., Rafailov, R., Sikchi, H., Finn, C., Niekum, S., Knox, W. B., & Sadigh, D. Contrastive Preference Learning: Learning from Human Feedback without Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

# Advanced Methods

## CPL

❖ Advantage Function (Bellman Expectation)

- 행동 가치 함수 $(Q^{\pi}(s,a))$: 현재 상태에서 선택한 행동이 얼마나 좋은가?
- 상태 가치 함수 $(V^{\pi}(s))$ : 현재 상태가 얼마나 좋은가? → **행동 가치 함수의 평균**
- $V^{\pi}(s) = E_{a\sim\pi}[Q^{\pi}(s,a)]$
- $A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$

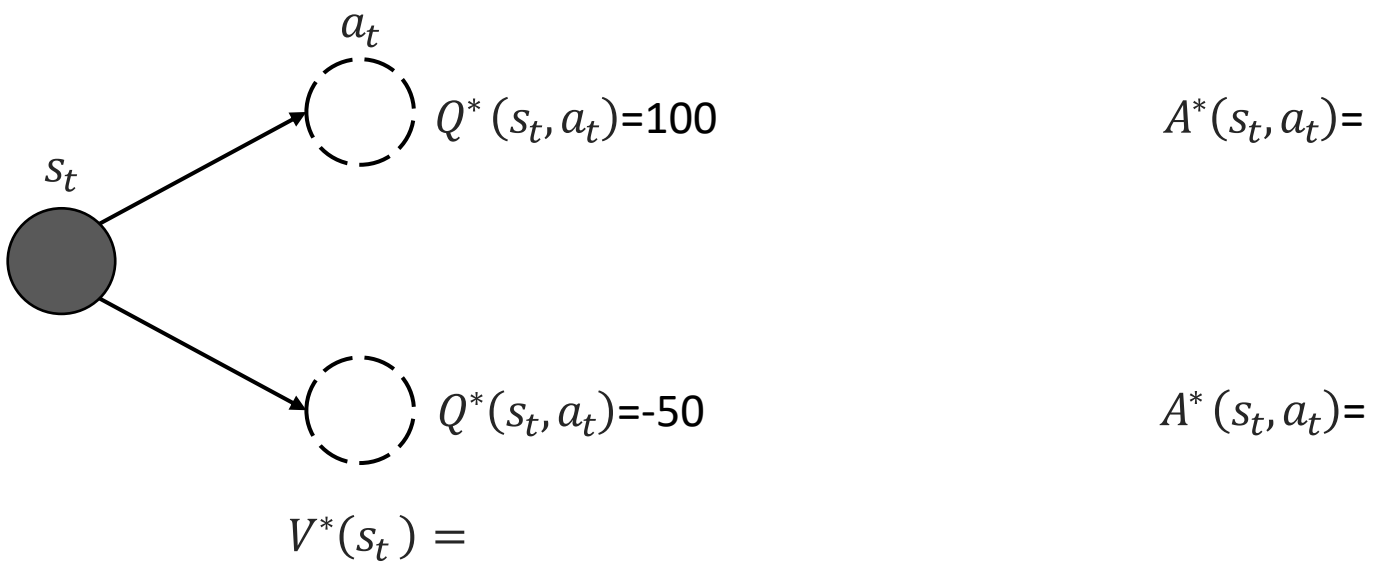


# Advanced Methods

## CPL

❖ Advantage Function (Bellman Optimality)

- 상태 가치 함수 $(V^{\pi}(s))$ : 현재 상태가 얼마나 좋은가?
- 행동 가치 함수 $(Q^{\pi}(s,a))$: 현재 상태에서 선택한 행동이 얼마나 좋은가?
- $V^{*}(s) = \max_{a}[Q^{*}(s,a)]$
- $A^{*}(s,a) = Q^{*}(s,a) - V^{*}(s)$



# Advanced Methods

## CPL

❖ Regret-based Preference Modeling

- $A^*(s_t, a_t) = Q^*(s_t, a_t) - V^*(s_t)$

Preference model based on cumulative rewards

$$\hat{P}_\psi(\sigma^0 > \sigma^1) = \frac{\exp\left(\Sigma_{\sigma^0}\hat{r}_\psi(s_t, a_t)\right)}{\exp\left(\Sigma_{\sigma^0}\hat{r}_\psi(s_t, a_t)\right) + \exp\left(\Sigma_{\sigma^1}\hat{r}_\psi(s_t, a_t)\right)}$$

Preference model based on optimal advantage (regret)

$$\hat{P}_{A^*}(\sigma^0 > \sigma^1) = \frac{\exp\left(\Sigma_{\sigma^0}\boldsymbol{A^*}(s_t, a_t)\right)}{\exp\left(\Sigma_{\sigma^0}\boldsymbol{A^*}(s_t, a_t)\right) + \exp\left(\Sigma_{\sigma^1}\boldsymbol{A^*}(s_t, a_t)\right)}$$

Considering time penalty through discount factor

$$\hat{P}_{A^*}(\sigma^0 > \sigma^1) = \frac{\exp\left(\Sigma_{\sigma^0}\boldsymbol{\gamma^t A^*}(s_t, a_t)\right)}{\exp\left(\Sigma_{\sigma^0}\boldsymbol{\gamma^t A^*}(s_t, a_t)\right) + \exp\left(\Sigma_{\sigma^1}\boldsymbol{\gamma^t A^*}(s_t, a_t)\right)}$$

Hejna, J., Rafailov, R., Sikchi, H., Finn, C., Niekum, S., Knox, W. B., & Sadigh, D. Contrastive Preference Learning: Learning from Human Feedback without Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

# Advanced Methods

## CPL

❖ Relationship between $A^*$ and $\pi^*$ in maximum entropy RL

- Maximum Entropy RL : 최적 행동 뿐만 아니라 엔트로피를 고려하여 다양한 행동에 대한 탐험을 할 수 있도록 하는 강화학습 정책

$$\pi^*(a|s) = argmax_a A^*(s,a)$$

$$\pi^*(a|s) = \frac{e^{A^*(s,a)/\alpha}}{\int e^{A^*(s,a)/\alpha}\, da}$$

$s_t$

**RL Policy**

**Maximum Entropy RL Policy**

Hejna, J., Rafailov, R., Sikchi, H., Finn, C., Niekum, S., Knox, W. B., & Sadigh, D. Contrastive Preference Learning: Learning from Human Feedback without Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

# Advanced Methods

## CPL

❖ Relationship between $A^*$ and $\pi^*$ in maximum entropy RL

- Maximum Entropy RL : 최적 행동 뿐만 아니라 엔트로피를 고려하여 다양한 행동에 대한 탐험을 할 수 있도록 하는 강화학습 정책

$$\pi^*(a|s) = \frac{e^{A^*(s,a)/\alpha}}{\int e^{A^*(s,a)/\alpha}\, da}$$

Hejna, J., Rafailov, R., Sikchi, H., Finn, C., Niekum, S., Knox, W. B., & Sadigh, D. Contrastive Preference Learning: Learning from Human Feedback without Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

# Advanced Methods

## CPL

❖ CPL Loss Function with Preference Dataset

- Advantage 함수를 정책 함수에 관련한 식으로 대입

$$\hat{P}_{A^*}(\sigma^0 > \sigma^1) = \frac{\exp\left(\Sigma_{\sigma^0}\gamma^t A^*(s_t, a_t)\right)}{\exp\left(\Sigma_{\sigma^0}\gamma^t A^*(s_t, a_t)\right) + \exp\left(\Sigma_{\sigma^1}\gamma^t A^*(s_t, a_t)\right)}$$

$$A^* = \alpha log \pi^*(a|s)$$

$$\hat{P}_{A^*}(\sigma^0 > \sigma^1) = \frac{\exp\left(\Sigma_{\sigma^0}\gamma^t \alpha \log \pi^*(a_t|s_t)\right)}{\exp\left(\Sigma_{\sigma^0}\gamma^t \alpha \log \pi^*(a_t|s_t)\right) + \exp\left(\Sigma_{\sigma^1}\gamma^t \alpha \log \pi^*(a_t|s_t)\right)}$$

$$L_{CPL,\theta} = E_{(\sigma^+,\sigma^-)\sim D_{pref}}\left[-\log \frac{\exp\left(\Sigma_{\sigma^+}\gamma^t \alpha \log \pi_\theta(a_t^+|s_t^+)\right)}{\exp\left(\Sigma_{\sigma^+}\gamma^t \alpha \log \pi_\theta(a_t^+|s_t^+)\right) + \exp\left(\Sigma_{\sigma^-}\gamma^t \alpha \log \pi_\theta(a_t^-|s_t^-)\right)}\right]$$

Hejna, J., Rafailov, R., Sikchi, H., Finn, C., Niekum, S., Knox, W. B., & Sadigh, D. Contrastive Preference Learning: Learning from Human Feedback without Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

# Advanced Methods

## CPL

### ❖ CPL Loss Function with Preference Dataset

- 해당 score function을 단순히 사용할 경우, 절대적인 scale에 invariant하게 됨
- $\lambda \in (0,1)$ 을 도입하여 절대적 거리가 늘어날 수록 score가 감소하도록 수정 (DPPO에서 착안)

$$L_{CPL,\theta} = E_{(\sigma^+,\sigma^-)\sim D_{pref}}[-\log \frac{exp\left(\Sigma_{\sigma^+}\gamma^t \alpha \log \pi_\theta(a_t^+|s_t^+)\right)}{exp\left(\Sigma_{\sigma^+}\gamma^t \alpha \log \pi_\theta(a_t^+|s_t^+)\right) + exp\left(\lambda \Sigma_{\sigma^-}\gamma^t \alpha \log \pi_\theta(a_t^-|s_t^-)\right)}]$$

$$\log \frac{\exp(-3)}{\exp(-3)+\exp(-5)} = 0.88 \qquad \log \frac{\exp(-13)}{\exp(-13)+\exp(-15)} = 0.88$$

$$\lambda = 0.5$$

$$\log \frac{\exp(-3)}{\exp(-3)+\exp(-5\lambda)} = 0.37 \qquad \log \frac{\exp(-13)}{\exp(-13)+\exp(-15\lambda)} = 0.04$$

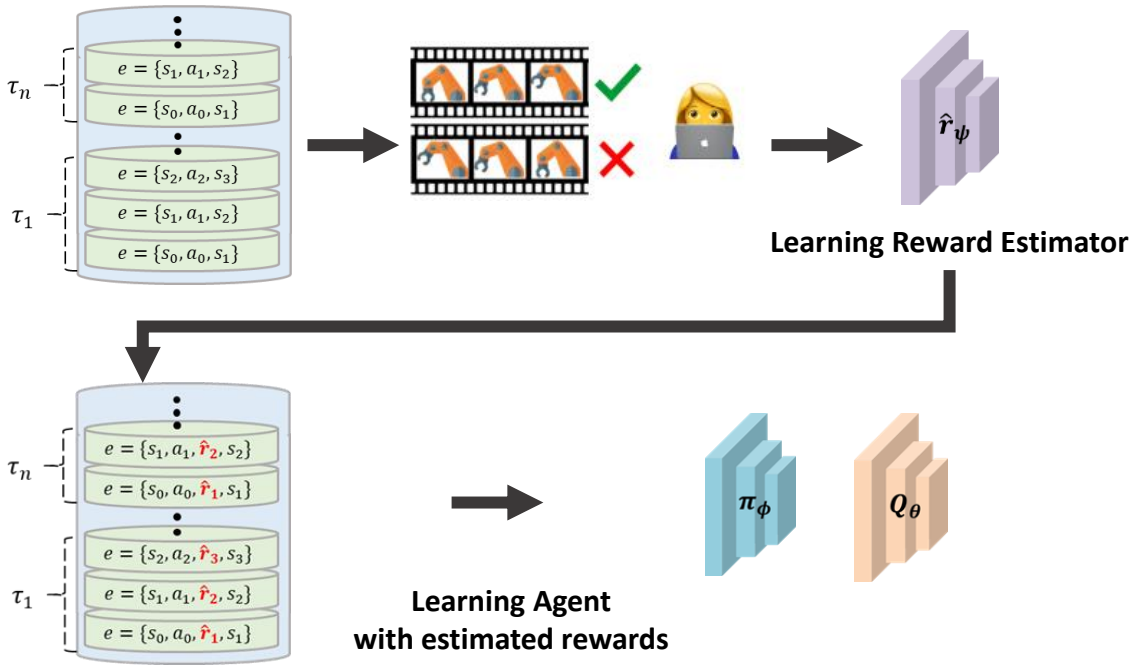
Hejna, J., Rafailov, R., Sikchi, H., Finn, C., Niekum, S., Knox, W. B., & Sadigh, D. Contrastive Preference Learning: Learning from Human Feedback without Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

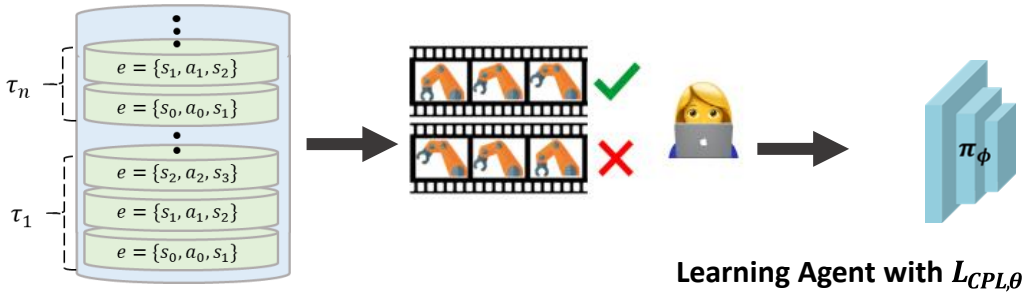
# Advanced Methods

## CPL

❖ CPL Loss Function with Preference Dataset

- DPPO, IPL과 마찬가지로 보상 함수 설계 없이 선호 데이터만으로 정책함수를 직접 최적화




$$L_{CPL,\theta} = E_{(\sigma^+,\sigma^-)\sim D_{pref}}\left[-\log \frac{exp\left(\Sigma_{\sigma^+}\gamma^t \alpha \log \pi_\theta(a_t^+|s_t^+)\right)}{exp\left(\Sigma_{\sigma^+}\gamma^t \alpha \log \pi_\theta(a_t^+|s_t^+)\right) + exp\left(\lambda \Sigma_{\sigma^-}\gamma^t \alpha \log \pi_\theta(a_t^-|s_t^-)\right)}\right]$$

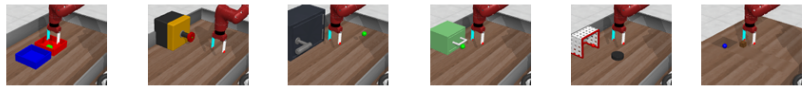

Hejna, J., Rafailov, R., Sikchi, H., Finn, C., Niekum, S., Knox, W. B., & Sadigh, D. Contrastive Preference Learning: Learning from Human Feedback without Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

# Advanced Methods

## CPL

### ❖ Experimental Results

- Metaworld 환경에서 실험
- SFT : 선호된 데이터만을 활용하여 behavior cloning
- P-IQL : 선호 데이터를 통해 MLP 기반 보상 함수 학습 후, IQL 강화학습 알고리즘을 최적화
- PPO : LLM의 RLHF 방식과 동일하게 KL Divergence 기반으로 학습
- Vector/Image 등 다양한 형태의 인풋의 State를 사용



| | | Bin Picking | Button Press | Door Open | Drawer Open | Plate Slide | Sweep Into |
|---|---|---|---|---|---|---|---|
| State 2.5k Dense | PPO | 83.7 ± 3.7 | 22.7 ± 1.9 | 79.3 ± 1.2 | 66.7 ± 8.2 | 51.5 ± 3.9 | 55.3 ± 6.0 |
| | SFT | 66.9 ± 2.1 | 21.6 ± 1.6 | 63.3 ± 1.9 | 62.6 ± 2.4 | 41.6 ± 3.5 | 51.9 ± 2.1 |
| | P-IQL | 70.6 ± 4.1 | 16.2 ± 5.4 | 69.0 ± 6.2 | 71.1 ± 2.3 | 49.6 ± 3.4 | 60.6 ± 3.6 |
| | CPL | **80.0 ± 2.5** | **24.5 ± 2.1** | **80.0 ± 6.8** | **83.6 ± 1.6** | **61.1 ± 3.0** | **70.4 ± 3.0** |
| Image 2.5k Dense | SFT | 74.7 ± 4.8 | 20.8 ± 2.4 | 62.9 ± 2.3 | 64.5 ± 7.6 | 44.5 ± 3.2 | 52.5 ± 2.5 |
| | P-IQL | **83.7 ± 0.4** | 22.1 ± 0.8 | 68.0 ± 4.6 | 76.0 ± 4.6 | 51.2 ± 2.4 | **67.7 ± 4.4** |
| | CPL | 80.0 ± 4.9 | **27.5 ± 4.2** | **73.6 ± 6.9** | **80.3 ± 1.4** | **57.3 ± 5.9** | **68.3 ± 4.8** |
| State 20k Sparse | PPO | 68.0 ± 4.3 | 24.5 ± 0.8 | 82.8 ± 1.6 | 63.2 ± 6.6 | 60.7 ± 4.2 | 58.2 ± 0.6 |
| | SFT | 67.0 ± 4.9 | 21.4 ± 2.7 | 63.6 ± 2.4 | 63.5 ± 0.9 | 41.9 ± 3.1 | 50.9 ± 3.2 |
| | P-IQL | 75.0 ± 3.3 | 19.5 ± 1.8 | **79.0 ± 6.6** | 76.2 ± 2.8 | **55.5 ± 4.2** | 73.4 ± 4.2 |
| | CPL | **83.2 ± 3.5** | **29.8 ± 1.8** | 77.9 ± 9.3 | **79.1 ± 5.0** | **56.4 ± 3.9** | **81.2 ± 1.6** |
| Image 20k Sparse | SFT | 71.5 ± 1.9 | 22.3 ± 2.9 | 65.2 ± 2.2 | 67.5 ± 1.1 | 41.3 ± 2.8 | 55.8 ± 2.9 |
| | P-IQL | **80.0 ± 2.3** | 27.2 ± 4.1 | **74.8 ± 5.8** | **80.3 ± 1.2** | 54.8 ± 5.8 | **72.5 ± 2.0** |
| | CPL | 78.5 ± 3.1 | **31.3 ± 1.6** | 70.2 ± 2.1 | **79.5 ± 1.4** | **61.0 ± 4.2** | **72.0 ± 1.8** |
| State % BC | 10% | 62.6 ± 2.6 | 18.9 ± 1.7 | 57.5 ± 3.0 | 61.5 ± 3.7 | 39.1 ± 2.5 | 49.3 ± 2.1 |
| | 5% | 64.6 ± 4.1 | 18.2 ± 0.6 | 59.8 ± 1.6 | 61.3 ± 1.8 | 38.6 ± 2.5 | 49.2 ± 1.9 |

Hejna, J., Rafailov, R., Sikchi, H., Finn, C., Niekum, S., Knox, W. B., & Sadigh, D. Contrastive Preference Learning: Learning from Human Feedback without Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

# Conclusion

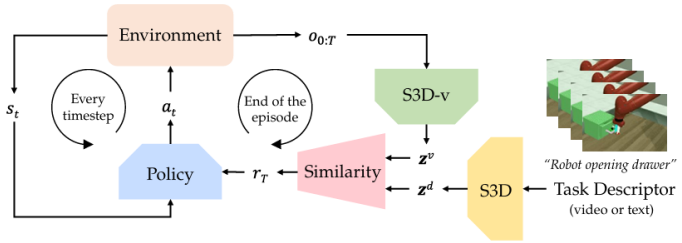
## Summary

- SeqRank (Hwang et al., NeurIPS 2024)
  - 순차적 쿼리 샘플링을 활용해 피드백 효율성을 높인 Online PbRL 방법론
- LiRE (Choi et al., ICML 2024)
  - 순차적 쿼리 샘플링을 활용해 모든 쿼리 데이터의 선호 순위를 활용한 Offline PbRL 방법론
- CPL (Hejna et al., ICLR 2024)
  - 대조학습과 Advantage 함수를 활용해 보상 함수 설계 없이 정책함수를 최적화하는 Offline PbRL 방법론

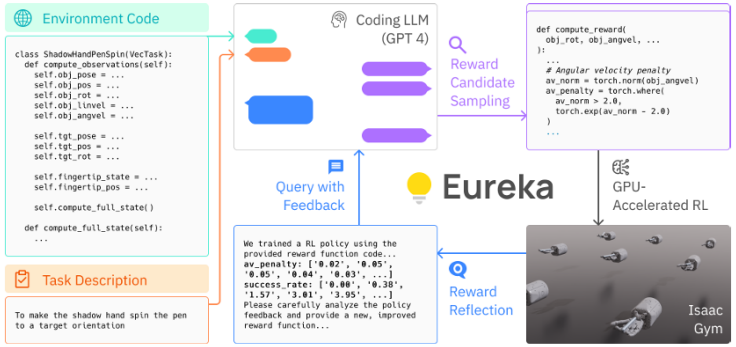
# Conclusion

## Trailer (Reinforcement Learning with Human Feedback: VLM-based Reward Design 1)

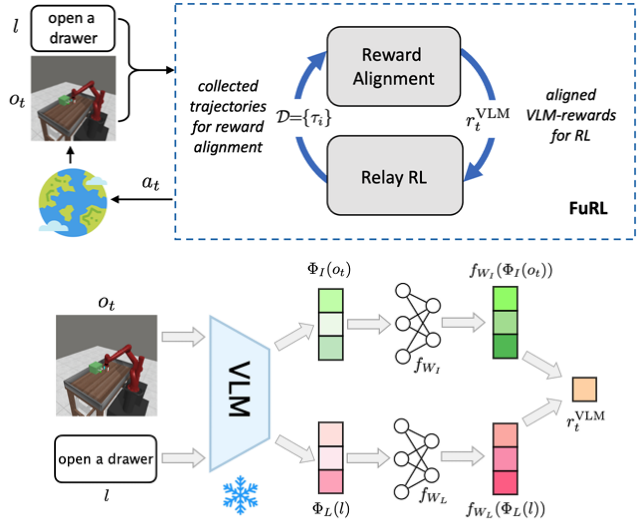
- ❖ RoboCLIP (Sontakke et al., NeurIPS 2023)
  - Video Demonstration, Text Description, 그리고 VLM을 이용해 보상 함수 구축
- ❖ Eureka (Ma et al., ICLR 2024)
  - GPT 4를 활용해 자동으로 보상 함수를 생성하고 개선
- ❖ FuRL (Fu et al., ICML 2024)
  - 단순히 VLM을 통해 보상 함수를 설계할 때 문제점을 지적하고, 단순 인간 피드백 뿐만 아니라 환경 신호도 함께 활용




**RoboCLIP**




**Eureka**




**FuRL**