Pretraining A Shared Q-Network for Data Efficient Offline Reinforcement Learning

Jongchan Park*
HYUNDAI MOTOR GROUP

Mingyu Park*
KAIST

Donghwan Lee
KAIST

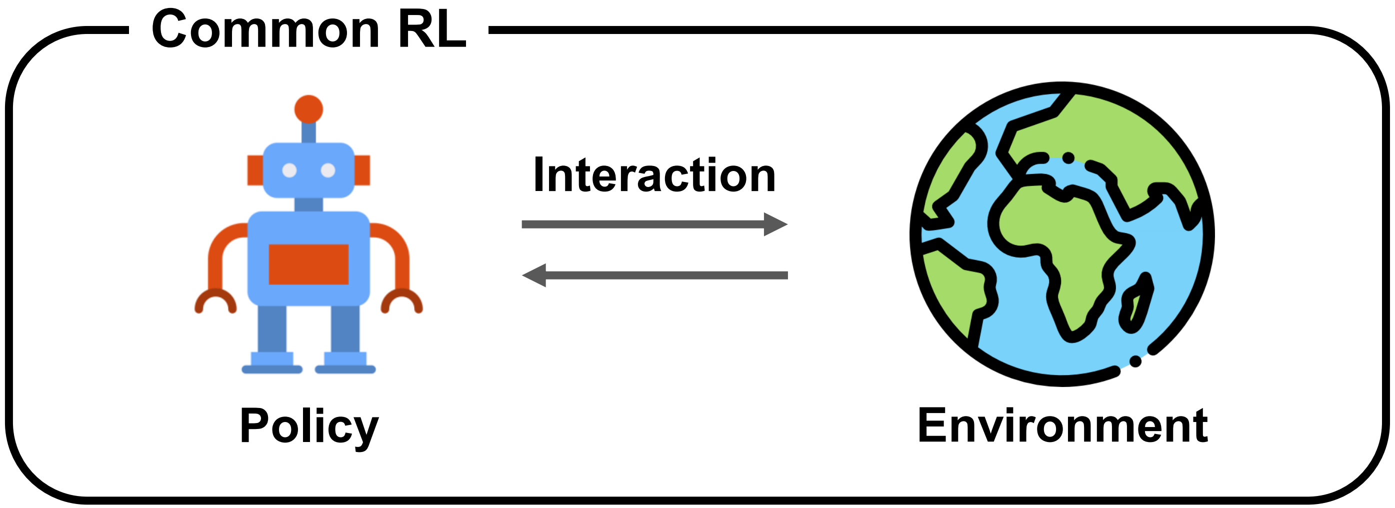
2025.12.

*Equal Contribution. Correspondence to Donghwan Lee.

Introduction

- **Sample-efficiency** means how fast a reinforcement learning (RL) learns a policy
- Sample-efficiency is the **crucial issue** in the common RL research field since RL agents are learned from **interaction** samples (Figure 1)
- Interaction with environments arises at a cost and is difficult in some cases; *accident, surgery, etc.*

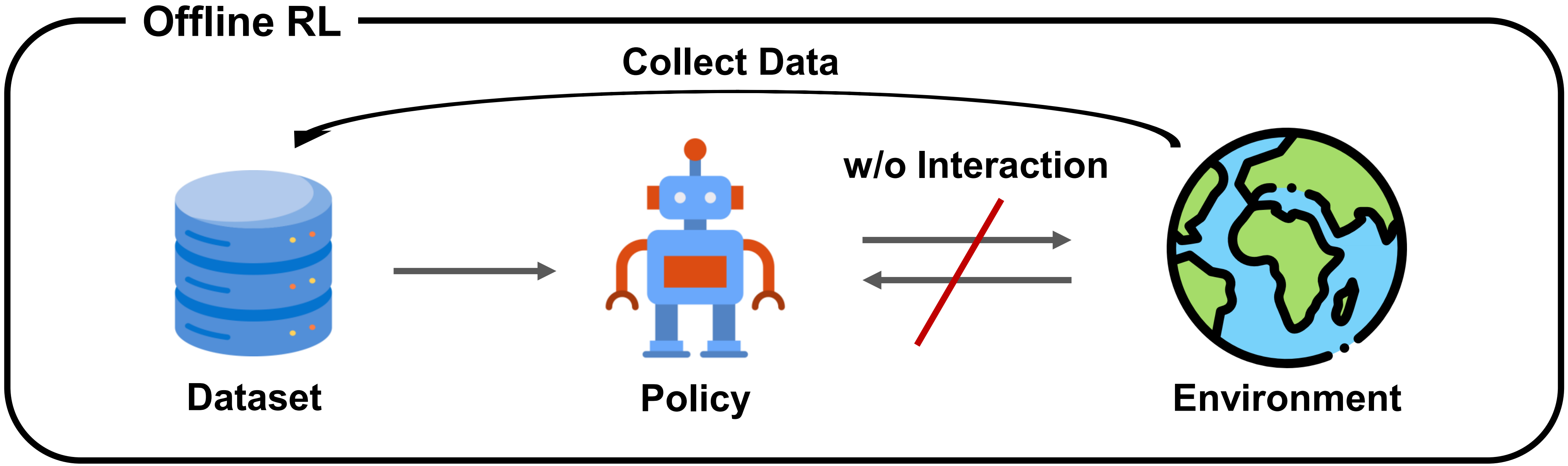
Figure 1.



Introduction

- To overcome this **sample-efficiency** issue, offline RL (batch RL) is suggested
- **Offline RL** aims to learn a policy with a static dataset **without any interaction** with an environment (Figure 2)
- While offline RL is actively studied in various problems, training with small amounts of data has not been considered enough

Figure 2.



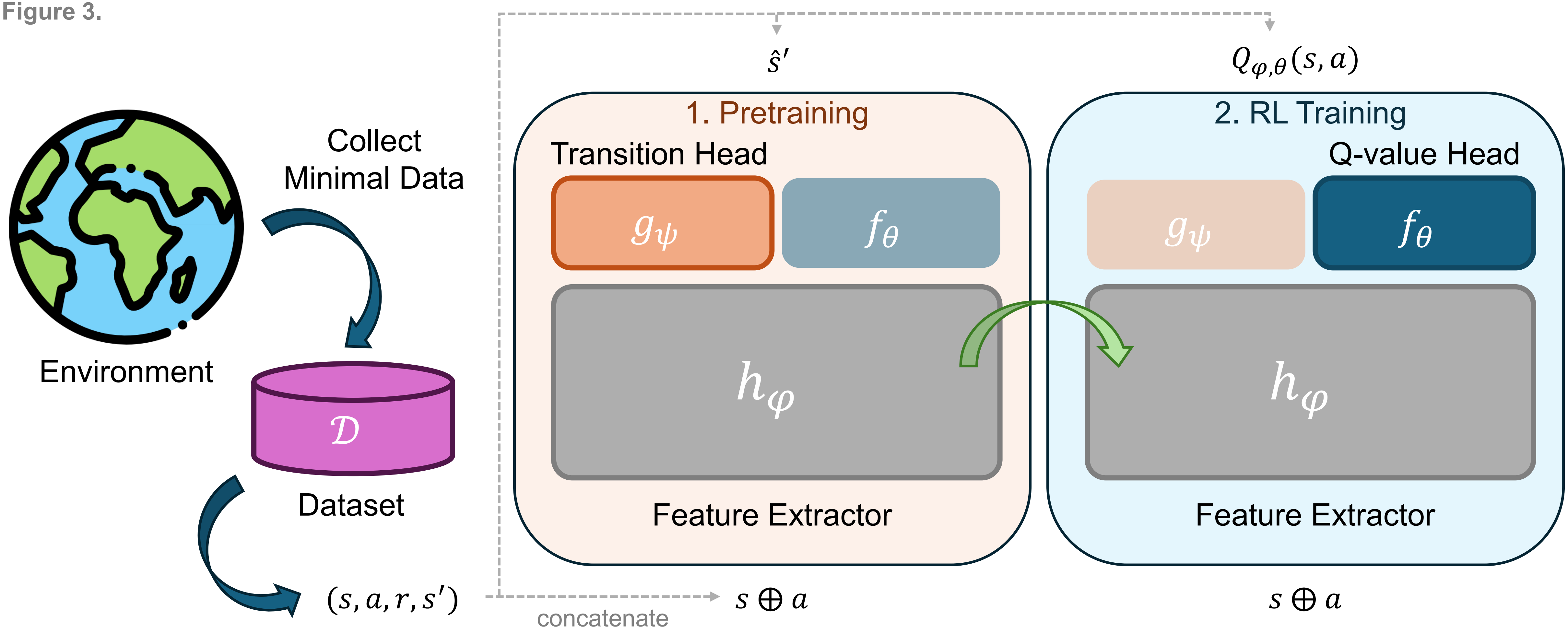
Introduction

- In this work, we aim to learn an offline RL policy with **less data** and name this problem as **data-efficiency**
- Following sections, we propose a simple yet effective **data-efficient offline RL method** that pretrains a Q-network with transition model estimation
- We demonstrate that the proposed method is indeed data-efficient by empirical experiment results

Method

- The proposed method consists of two stages on shared Q-Network (Figure 3)

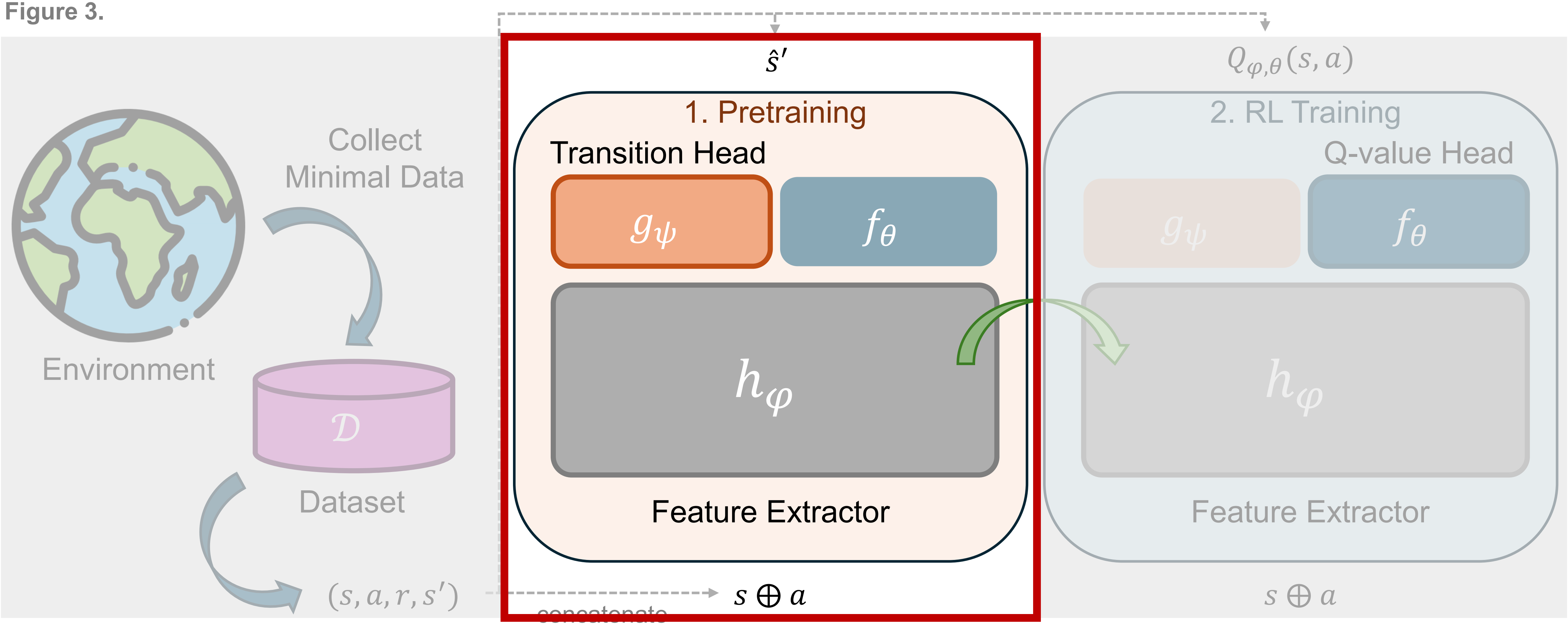
Figure 3.



Method

- First, pretrain a feature extractor h_φ of the shared Q-network with transition head

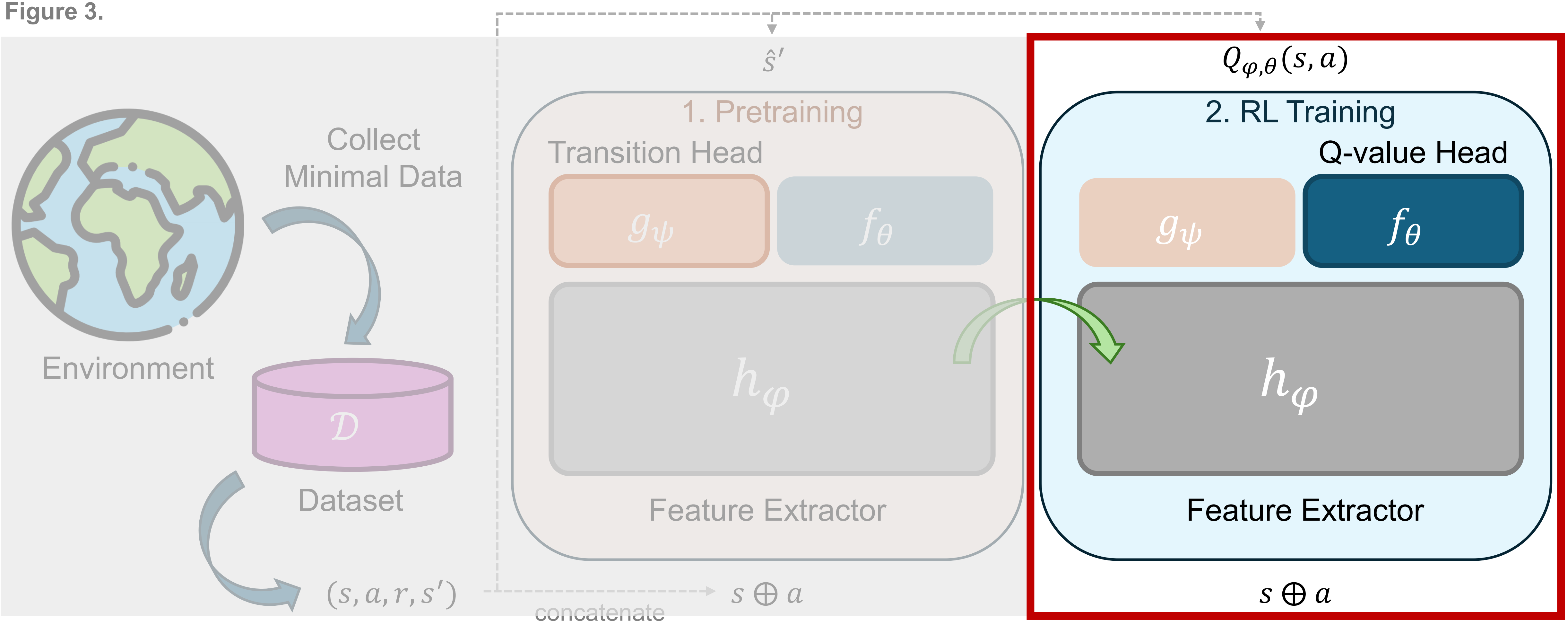
Figure 3.



Method

- Second, train an existing offline RL algorithm with the pretrained feature extractor

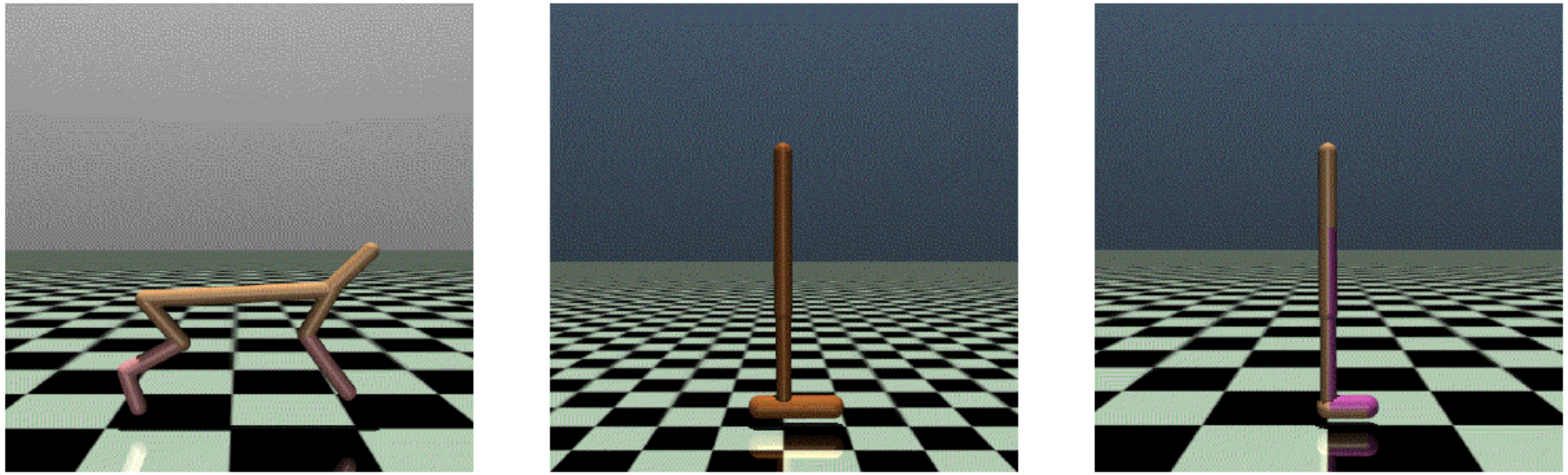
Figure 3.



Experiments - Dataset

- **D4RL** (Fu et al., 2020) Open AI Gym locomotion tasks
  - Three different embodiments: HalfCheetah, Hopper and Walker2d (Figure 4)
  - Five different reward qualities: random, medium, medium replay, medium expert, expert

Figure 4.



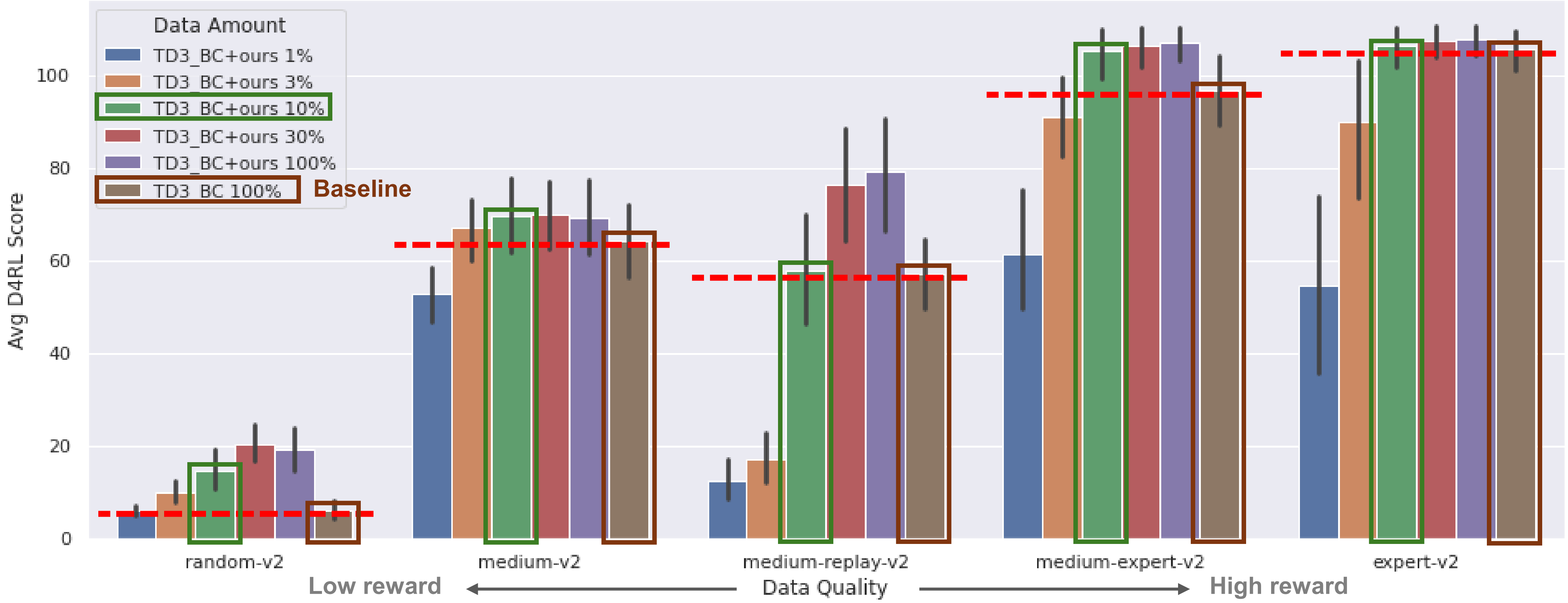
(a) HalfCheetah (b) Hopper (c) Walker2d

Fu, Justin, et al. "D4rl: Datasets for deep data-driven reinforcement learning." *arXiv preprint arXiv:2004.07219* (2020).

Experiments - Results

- We evaluate our method with TD3+BC over various sizes of D4RL datasets across **data qualities** of reward
- Figure 5 shows that **only with 10% of dataset**, our method **outperforms** the baseline

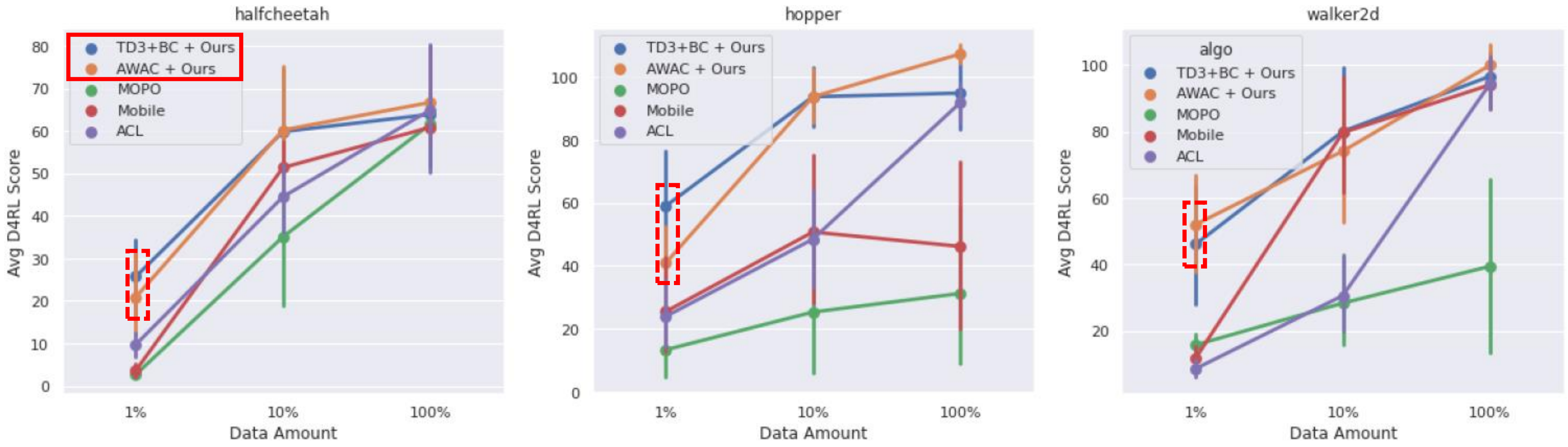
Figure 5.



Experiments - Results

- We also compare our method with offline model-based RL and representation learning approaches on D4RL
- Figure 6 shows overall results of medium, medium replay and medium expert datasets and **our method outperforms in reduced datasets**, especially in 1%

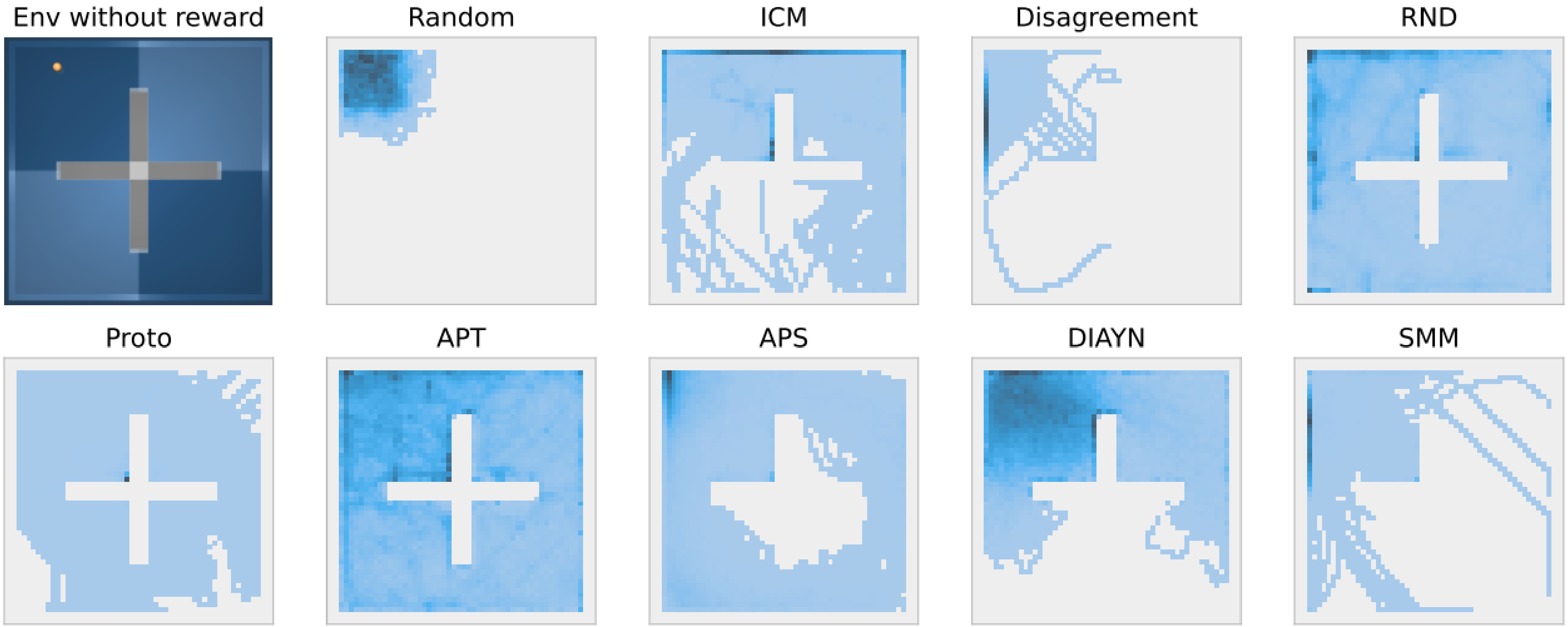
Figure 6.



Experiments - Dataset

- **ExoRL** (Yarats et al., 2022) collects the datasets by utilizing **various exploration strategies** (Figure 7)

Figure 7.



Yarats, Denis, et al. "Don't change the algorithm, change the data: Exploratory data for offline reinforcement learning." arXiv preprint arXiv:2201.13425 (2022).

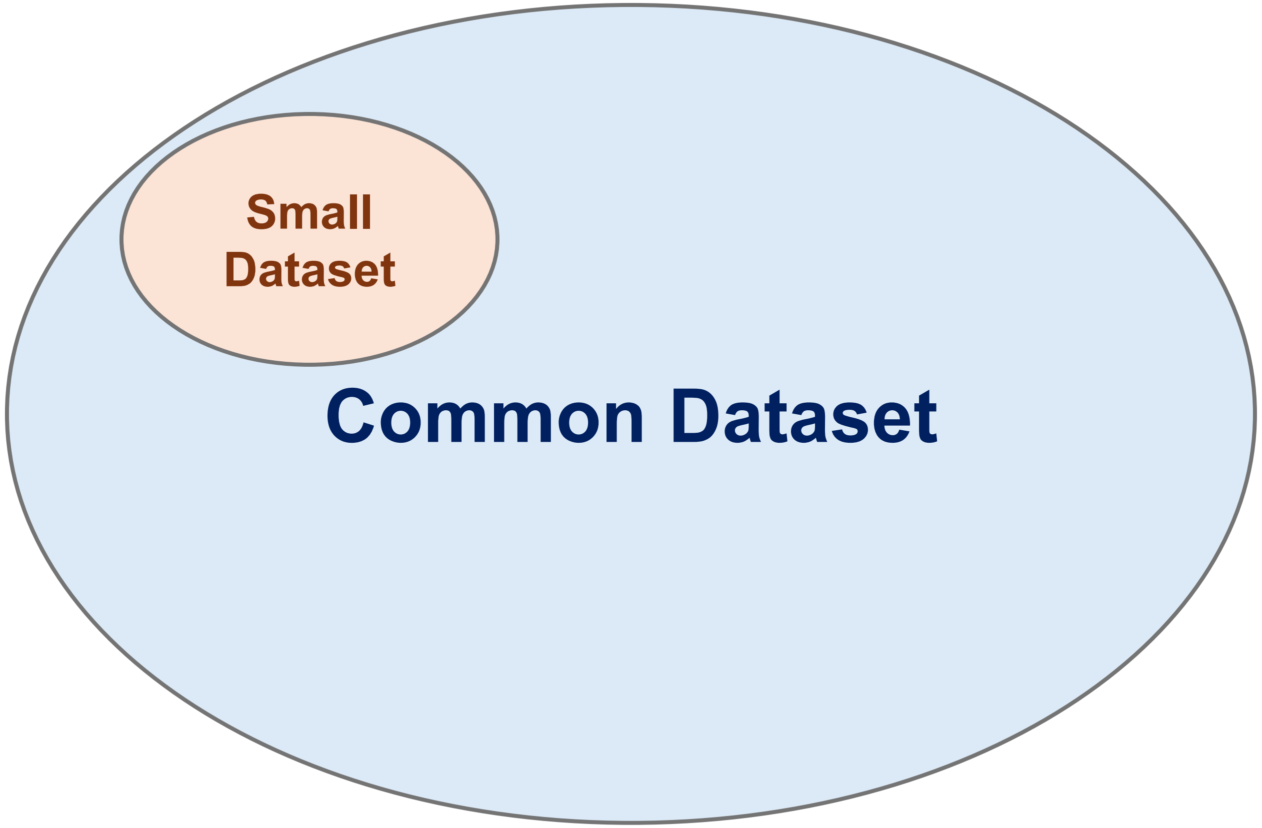
HYUNDAI MOTOR GROUP

KAIST

Experiments

- **Assumption**: small datasets have **a distinctive data distribution** compared to common, large datasets

Figure 8.



Experiments - Results

- We evaluate our method on *walker walk* task in ExoRL from three different **collection strategies** (*i.e., SMM, RND, ICM*)
- Figure 9 shows that TD3 with our method outperforms the baseline overall
- With **10% of datasets**, our method outperforms the baseline with full datasets

Figure 9.

Conclusion

- We propose **pretraining a shared Q-network** method to deal with **data-efficiency**
- The **shared network** structure for Q-function leads a simple yet effective framework
- We demonstrate that our method is indeed **data-efficient regardless** of the dataset **qualities** and **distribution**