

We show the need for large and diverse dataset to learn a goal-oriented value function in offline setting

Introduction

- We explore the application of offline Reinforcement Learning (RL), specifically focusing on learning a goal-oriented knowledge representation framework called World Value Function (WVF).
- We benchmark the performance of selected offline RL algorithms, including offline Deep Q-Network (DQN) and Batch-Constrained deep Q-learning (BCQ), under varying data buffer sizes.
- The rationale for investigating these algorithms for learning WVFs lies in the potential benefits of leveraging historical data to accelerate learning and improve sample efficiency.
 - This can be particularly advantageous in scenarios where online data collection is limited.

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right]$$

Methodology

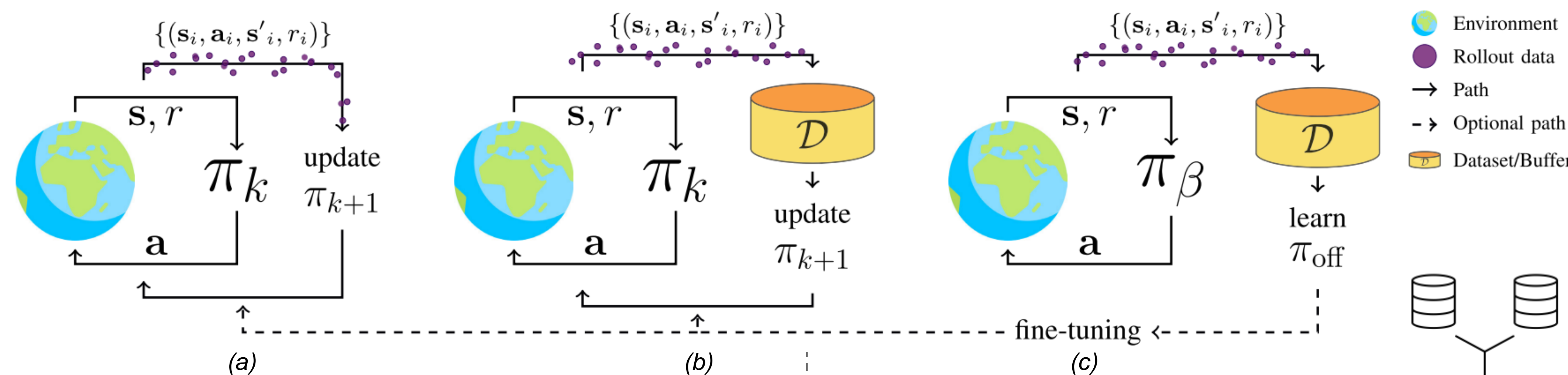


Figure 2. Offline RL (c), learn a control policy without interacting with the environment (Prudencio et al., 2023).

- Notably, both the DQN and BCQ algorithms were modified to learn goal-oriented value functions:

Algorithm 1 Offline DQN for WVF

```

1: Initialize: WVF  $\bar{Q}$  with random weights  $\theta$ , target network  $\bar{Q}'$  with weights  $\theta' = \theta$ , goal buffer  $\mathcal{G}$  replay buffer  $\mathcal{D}$ , learning rate  $\alpha$ , batch size  $N$ 
2: for each timestep do
3:   Sample minibatch of transitions  $(s, g, a, r, s')$  of size  $N$  from  $\mathcal{D}$ 
4:   for each  $(s, g, a, r, s')$  in minibatch do
5:     if  $s'$  is absorbing then
6:        $\mathcal{G} \leftarrow \mathcal{G} \cup \{s'\}$ 
7:     end if
8:     for each  $g' \in \mathcal{G}$  do
9:        $\bar{r} \leftarrow R_{\text{MIN}}$  if  $g' \neq s'$  and  $s' \in \mathcal{G}$  else  $r$ 
10:       $\delta \leftarrow [\bar{r} + \max_{a'} \bar{Q}(s', g', a'; \theta')] - \bar{Q}(s, g', a; \theta)$ 
11:      Perform a gradient descent step on  $(\delta)^2$  with respect to the network parameters  $\theta$ 
12:    end for
13:    Every  $C$  steps update  $\bar{Q}' = \bar{Q}$ 
14:  end for
15: end for
    
```

- We used a similar approach for BCQ, but the policy is constrained to improve upon a behaviour policy that is close to the data collection policy, as in (Fujimoto et al., 2019b)

Experiments: Boxman

- We trained an offline DQN and discrete BCQ agent using various dataset sizes to assess their performance under different data sizes.

- In the context of discrete BCQ, setting the threshold parameter, τ to 0 results in Q-learning, while setting τ to 1 yields an imitation of the actions present in the dataset (Fujimoto et al., 2019a)

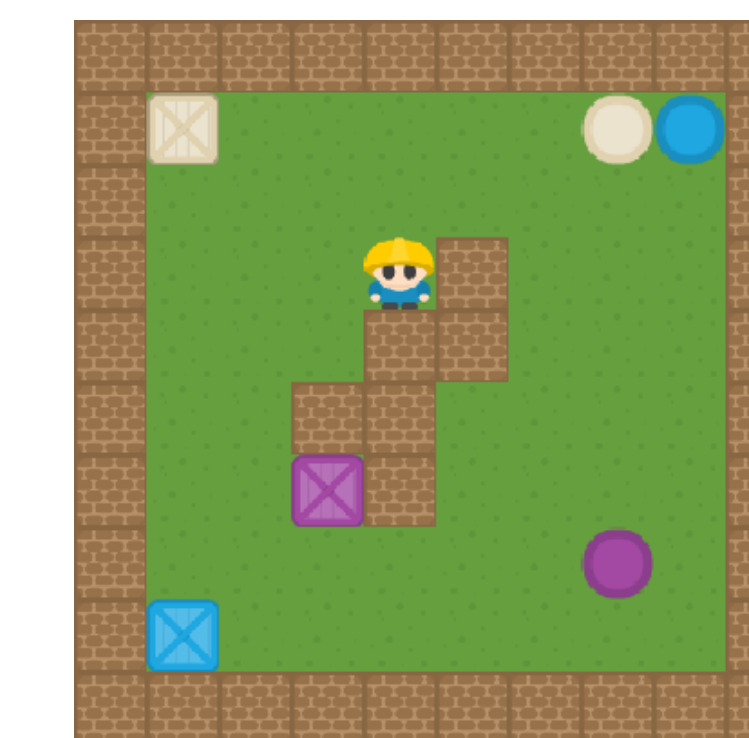
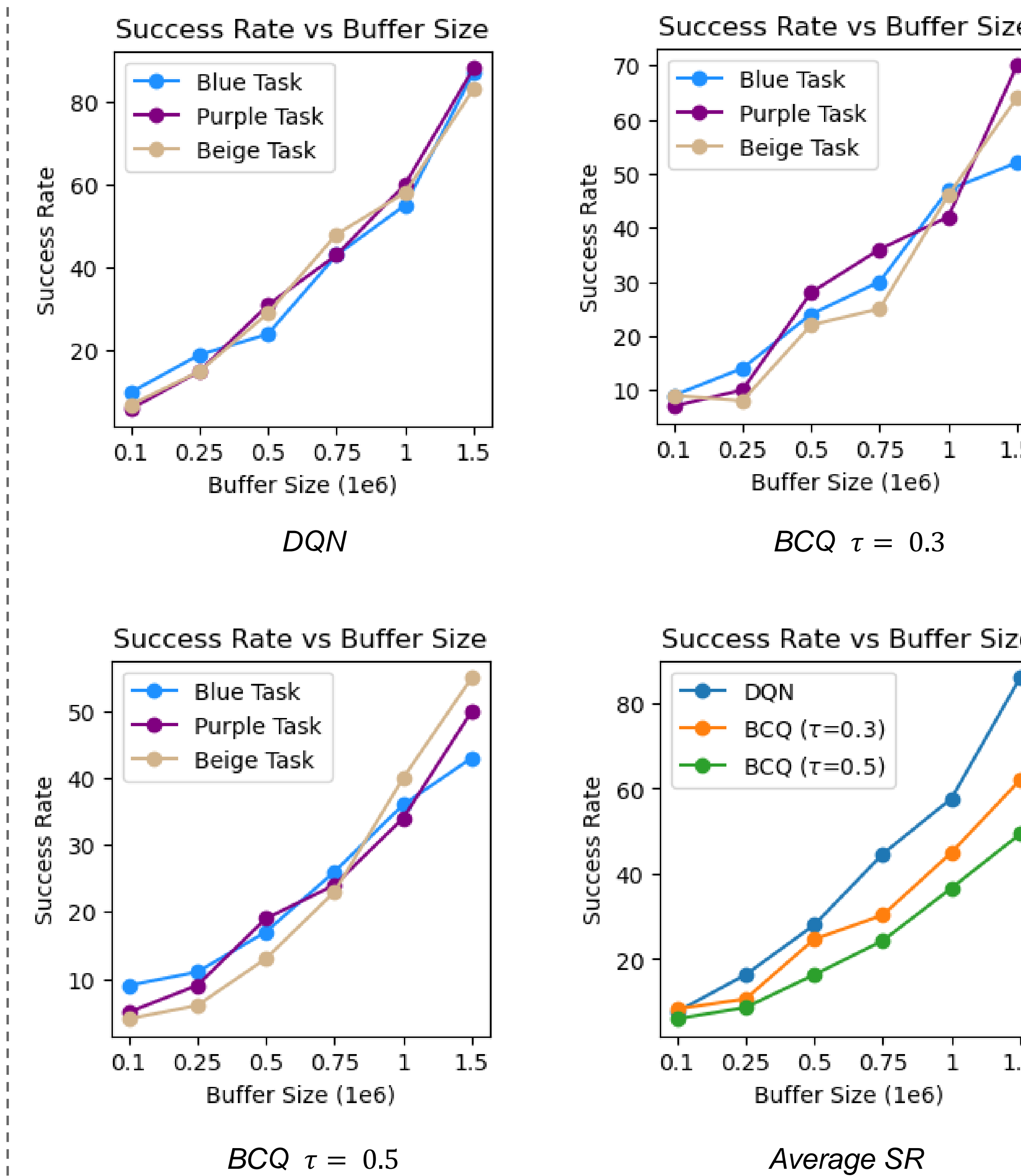


Figure 2. Layout of Boxman



- We believe that because the offline DQN does not have this constraint, it is able to learn better than BCQ given the combined replay buffer.
- We conclude that offline DQN is more suitable for learning WVFs in a discrete domain.

Experiments: Panda-Gym

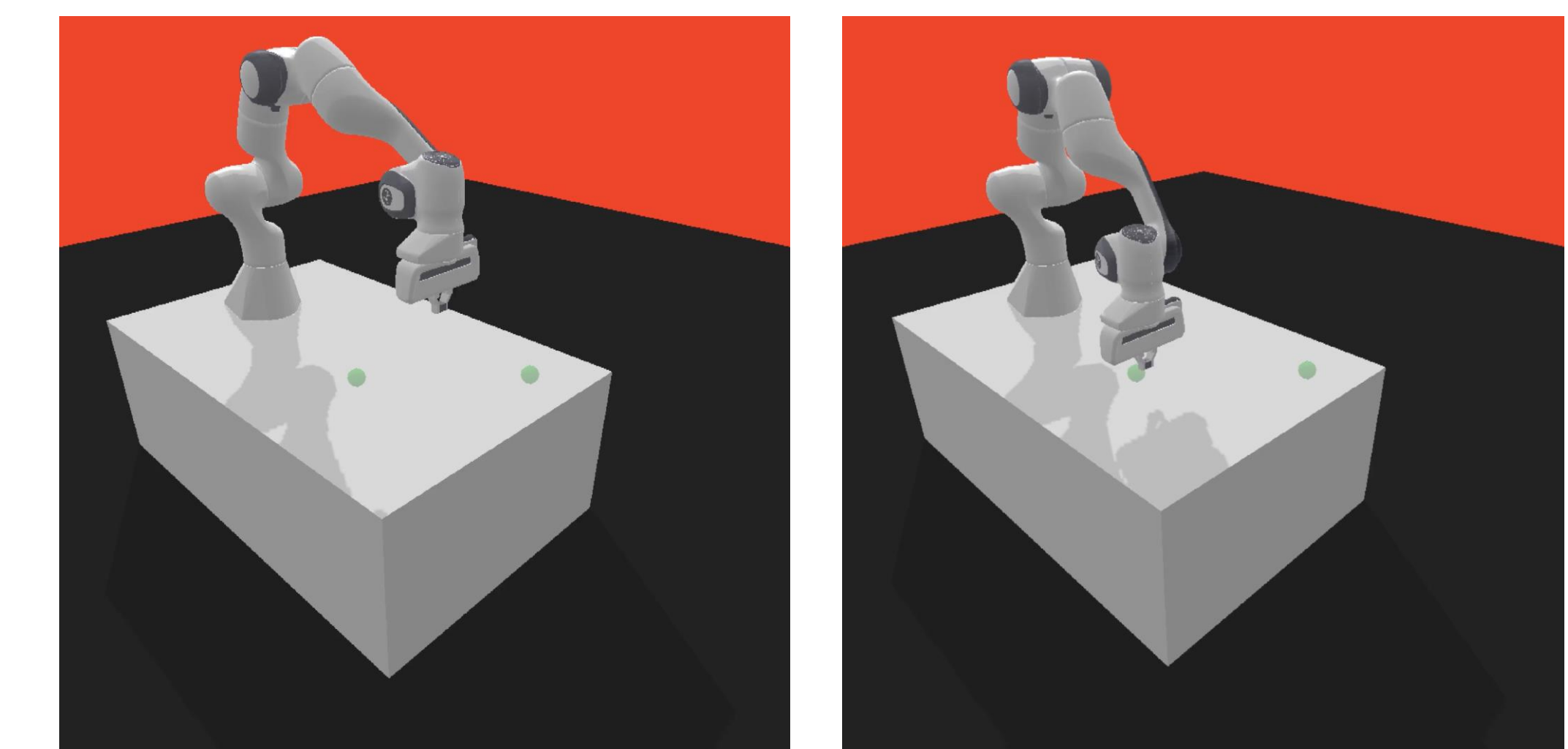
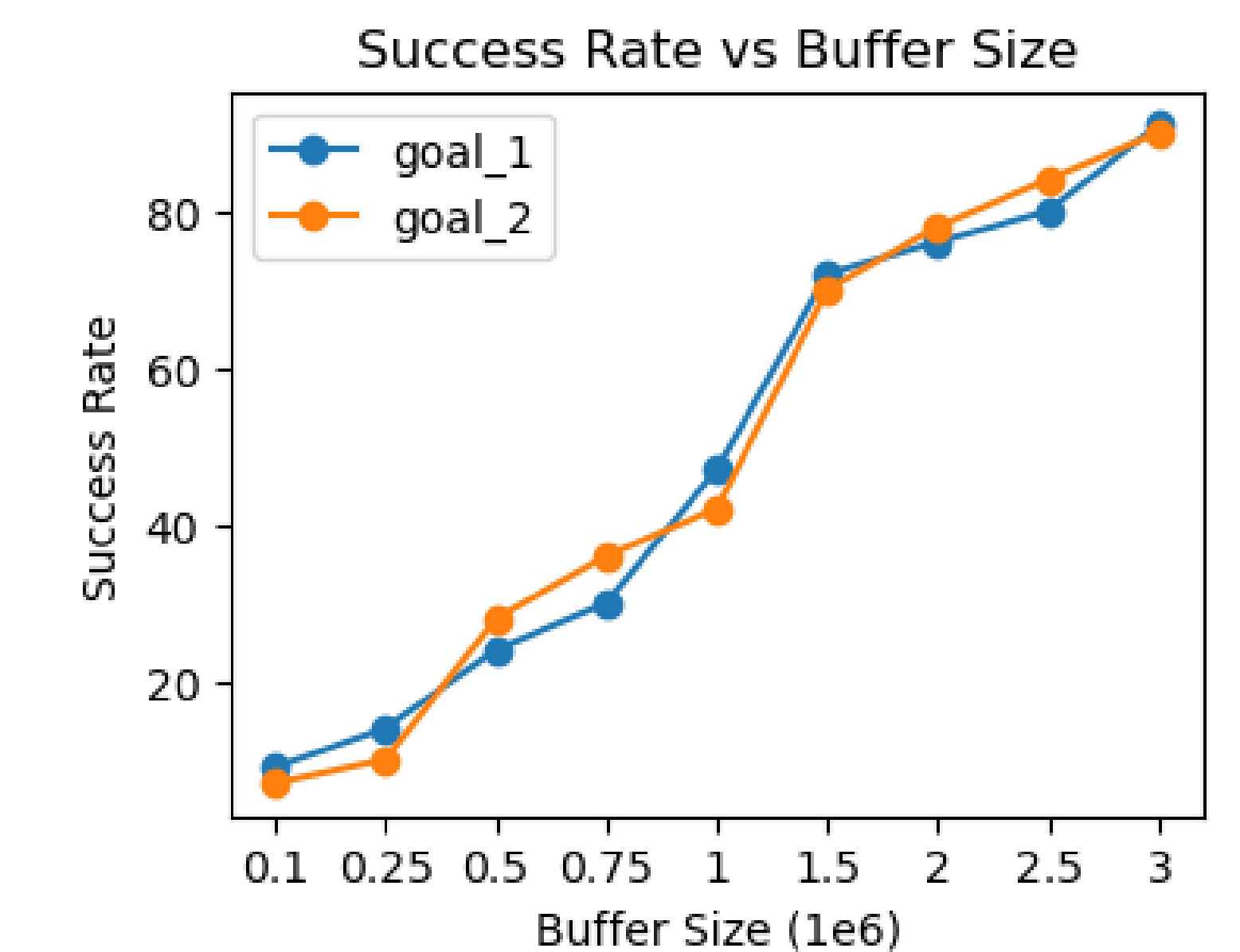


Figure 3. The figure presented illustrates the application of learned WVF. It's important to note that the WVF encapsulates all tasks related to reaching the agent's internal goals (Tasse et al., 2022b).



- While continuous BCQ has mechanisms to handle issues like extrapolation error and overestimation bias, the effectiveness of these mechanisms are greatly enhanced with more data.

References

1. Tasse, G. N., Rosman, B., and James, S. *World value functions: Knowledge representation for learning and planning*, 2022b.
2. Fujimoto, S., Meger, D., and Precup, D. *Off-policy deep reinforcement learning without exploration*, 2019b.
3. Fujimoto, S., Conti, E., Ghavamzadeh, M., and Pineau, J. *Benchmarking batch deep reinforcement learning algorithms*, 2019a.