

---

# Learning Human-Like RL Agents through Trajectory Optimization with Action Quantization

---

Jian-Ting Guo<sup>1\*</sup>, Yu-Cheng Chen<sup>1\*</sup>, Ping-Chun Hsieh<sup>1†</sup>, Kuo-Hao Ho<sup>1</sup>  
 Po-Wei Huang<sup>1</sup>, Ti-Rong Wu<sup>2†</sup>, I-Chen Wu<sup>1,3</sup>

<sup>1</sup>Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan

<sup>2</sup>Institute of Information Science, Academia Sinica, Taiwan

<sup>3</sup>Research Center for Information Technology Innovation, Academia Sinica, Taiwan

{gjt.cs13, yucheng.cs11, pinghsieh, lukewayne123.cs05}@nycu.edu.tw  
 a311551048.cs11@nycu.edu.tw, tirongwu@iis.sinica.edu.tw, icwu@cs.nycu.edu.tw

## Abstract

Human-like agents have long been one of the goals in pursuing artificial intelligence. Although reinforcement learning (RL) has achieved superhuman performance in many domains, relatively little attention has been focused on designing human-like RL agents. As a result, many reward-driven RL agents often exhibit unnatural behaviors compared to humans, raising concerns for both interpretability and trustworthiness. To achieve human-like behavior in RL, this paper first formulates human-likeness as trajectory optimization, where the objective is to find an action sequence that closely aligns with human behavior while also maximizing rewards, and adapts the classic receding-horizon control to human-like learning as a tractable and efficient implementation. To achieve this, we introduce Macro Action Quantization (MAQ), a human-like RL framework that distills human demonstrations into macro actions via Vector-Quantized VAE. Experiments on D4RL Adroit benchmarks show that MAQ significantly improves human-likeness, increasing trajectory similarity scores, and achieving the highest human-likeness rankings among all RL agents in the human evaluation study. Our results also demonstrate that MAQ can be easily integrated into various off-the-shelf RL algorithms, opening a promising direction for learning human-like RL agents. Our code is available at <https://rlg.iis.sinica.edu.tw/papers/MAQ>.

## 1 Introduction

Designing human-like agents has been an important goal on the path toward artificial intelligence. One of the most well-known benchmarks is the *Turing Test*, which evaluates whether an agent can perform intelligent behavior that is indistinguishable from a human. To pass the Turing Test, agents are designed not only to achieve the goal of the given task but also to behave in a human-like manner.

While the pursuit of human-like agents has been investigated in natural language processing (NLP), such as in large language models (LLMs) that aim to generate responses aligned with human preferences [1], it has not yet been widely explored in the field of deep reinforcement learning (DRL). In recent years, DRL has made significant progress in various domains, such as gaming [2, 3, 4] or robotics control [5, 6]. However, researchers have noticed RL agents often exhibit unnatural behaviors,

---

\*These authors contributed equally.

†Corresponding authors.

making they easily distinguishable from those of human players, e.g., smoothness [7, 8, 9, 10], navigation issues [11, 12, 13], and (unnatural) shaking and spinning actions [14]. Although RL agents can accomplish these tasks, they fail to achieve the goal of designing human-like intelligence. This disparity raises an important and underexplored challenge in human-like RL research: *How to design agents that not only succeed in tasks but also behave like humans?*

To address this challenge, this paper proposes to formulate human-likeness in reinforcement learning as a trajectory optimization problem, where the objective is to produce trajectories that closely align with human behavior. Following the principle of receding-horizon control, we optimize over short action segments rather than entire trajectories to capture human behavior more effectively. To achieve this, we introduce *Macro Action Quantization* (MAQ), a human-likeness aware RL framework for developing competitive agents with human-like behavior. MAQ first distills human behavior into *macro actions* – sequences of actions – from offline human demonstrations using a conditional Vector-Quantized Variational Autoencoder (VQVAE) [15]. The VQVAE generates a discrete set of macro actions, stored in a learned *codebook*, where each entry represents possible human behavior segments for a given state. By leveraging these macro actions, MAQ transforms the action space from low-level primitive actions to high-level, human-like macro actions, effectively constraining the agent to operate within a human-like behavioral space.

We evaluate MAQ on one of the standard RL benchmarks, the Adroit tasks in D4RL [16], which has readily available human demonstrations. We apply MAQ to three RL algorithms, including IQL, SAC, and RLPD. Experimental results show that MAQ not only achieves higher success rates in task completion but also substantially improves human-likeness. When measuring trajectory similarity using Dynamic Time Warping (DTW) and Wasserstein distance metrics, MAQ significantly increases the similarity scores across all tasks. In addition, we conduct a human evaluation study – similar to the Turing Test – in which participants are asked to distinguish between human demonstrations and agent behaviors. The results show that while participants can easily identify traditional RL agents as non-human, they struggle to distinguish MAQ agents from humans. In conclusion, our findings demonstrate that MAQ effectively captures human-like behavior, offering a promising direction for future research in human-like RL studies.

## 2 Background

### 2.1 Human-like Reinforcement Learning

While most RL research focuses on designing reward-driven RL agents, there are a few works exploring human-like RL in two directions. The first involves adopting *behavioral constraints* to enforce human-likeness. For example, Fujii et al. [17] proposed to penalize actions identified as non-human, while Ho et al. [14] introduced the adaptive behavior cost to discourage non-human-like behavior, such as spinning and shaking. Although effective, these methods rely on pre-defined behavior constraints or rule-based penalties, requiring substantial effort for handcrafted design. An alternative approach to avoid the handcrafted design of behavior costs is to directly learn from *offline human datasets*. Notably, human demonstrations are available and widely used for RL training in various domains, such as Atari games [18], self-driving car [19], and robot arm manipulation [16]. These pre-collected demonstrations can be used to capture the characteristics of human behavior through imitation learning [20], such as behavior cloning or inverse RL [21]. However, while imitation learning improves human likeness, its performance is often limited by the quality of human demonstrations and fails to compete with the non-human-like RL benchmarks. In summary, how to effectively train human-like RL agents remains an open and underexplored challenge.

### 2.2 Semi-Markov Decision Process and Macro Action

Semi-Markov Decision Process (SMDP) extends MDP by incorporating *macro actions*, which are sequences of consecutive actions executed over multiple timesteps. Given a state  $s_t$ , a macro action with a length  $L$  is defined as  $m_t = (a_t, a_{t+1}, \dots, a_{t+L-1})$ . Compared to MDP, SMDP is denoted by  $(\mathcal{S}, \mathcal{M}, \mathcal{R}, \mathcal{P}, \gamma)$ , where  $\mathcal{M}$  is the set of macro actions and  $\mathcal{R}$  is the cumulative reward obtained over the macro actions:  $R(s_t, m_t) = \sum_{k=t}^{t+L-1} r(s_k, a_k)$  [22]. Similarly, the objective is to learn a policy that selects macro actions to maximize the cumulative rewards:  $J(\pi_\theta) = \mathbb{E}_{(s_t, m_t) \sim \pi_\theta} [\sum_{t=0}^{\infty} \gamma^t R(s_t, m_t)]$ , where  $m_t \in M$ , and  $t$  represents the timestep for selecting macro

actions. SMDP simplifies complex tasks by enabling agents to focus on long-term planning, as macro actions reduce decision-making frequency while capturing sequential dependencies.

### 2.3 VQVAE

*Vector quantized variational autoencoder* (VQVAE) [15], an extension of the *variational autoencoder* (VAE) [23], is a generative model that incorporates vector quantization to learn discrete latent embeddings. VQVAE consists of an encoder, a codebook of discrete latent embeddings, and a decoder. The encoder transforms input data into a latent vector, which is then quantized by replacing it with the nearest vectors from the codebook. The decoder aims to reconstruct the same input data using this discrete latent code. By utilizing discrete latent representations, VQVAE allows learning more structured representations.

Recent studies have explored the application of VQVAE in reinforcement learning. For example, Ozair et al. [24] and Antonoglou et al. [25] propose using conditional-VQVAE [26] as a state transition model in stochastic environments, using the codebook’s latent vectors to represent possible chance events. Given a state and a latent embedding (chance event) from the codebook, the decoder aims to generate the corresponding next states. In addition, Luo et al. [27] proposes utilizing conditional-VQVAE to discretize continuous action spaces in robotic control tasks. Specifically, each latent embedding in the codebook represents a possible action for a given state. The decoder reconstructs the action from the corresponding latent embedding. This approach transforms continuous action spaces into discrete action spaces, improving learning efficiency in robotic control tasks.

## 3 Human-Like Trajectory Optimization

**Trajectory optimization.** The canonical trajectory optimization problem is formulated based on an MDP  $(\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma, T, p_0)$  with continuous state space  $\mathcal{S}$  and action space  $\mathcal{A}$ , transition function  $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})^3$ , reward function  $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ , discount factor  $\gamma \in [0, 1)$ , episode length  $T$ , and initial state distribution  $p_0$ . Let  $\tau = (s_0, a_0, \dots, s_{T-1}, a_{T-1}, s_T)$  denote a trajectory generated under an action sequence  $a_{0:T-1} := (a_0, a_1, \dots, a_{T-1})$ . Our goal is to find an action sequence that can maximize the total discounted return.

$$a_{0:T-1}^* := \arg \max_{a_{0:T-1} \in \mathcal{A}^T} \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma^t R(s_t, a_t) \middle| s_0 \sim p_0; a_{0:T-1} \right], \quad (1)$$

where the expectation is taken over the randomness of the initial state, reward realizations, and state transitions.

**Human-like receding-horizon control (HRC).** As exact full-horizon optimization in equation (1) can be intractable even for moderately large  $T$ , we resort to the technique of receding-horizon control, which is a generic control method that involves repeatedly solving a optimization problem on a short moving time horizon to choose action sequences. Specifically, at the decision step  $t$ , receding-horizon control would find an action sequence of length  $H$  (usually much smaller than  $T$ ) by solving

$$\arg \max_{a_{t:t+H-1}} \mathbb{E} \left[ \sum_{i=t}^{t+H-1} \gamma^{i-t} R(s_i, a_i) \middle| s_t; a_{t:t+H-1} \right]. \quad (2)$$

Notably, under receding-horizon control, such short-term replanning is done every  $k$  steps (with  $1 \leq k \leq H$ ), i.e., the first  $j$  actions in  $a_{t:t+H-1}^*$  ( $1 \leq j \leq H$ ) are executed and the environment returns  $s_{t+j}$  and equation (2) is re-optimized. This *receding-horizon* loop reproduces the closed-loop robustness observed in [28] while bounding per-step computation.

To enforce human-likeness, we introduce the concept of *human-like sequence constraint* as follows: Let  $\mathcal{D} = \{\tau^{(i)}\}_{i=1}^N$  be a set of human-generated action sequences and define *human manifold*  $\mathcal{H} = \{\tau \mid \tau \in \mathcal{D}\}$ . Accordingly, human-likeness can be enforced by constraining the search space of action sequences to  $\mathcal{H}$ . Hence, human-like receding-horizon control can be formally described as finding an action sequence as

$$a_{t:t+H-1}^* = \arg \max_{a_{t:t+H-1} \in \mathcal{H}} \mathbb{E} \left[ \sum_{i=t}^{t+H-1} \gamma^{i-t} R(s_i, a_i) \middle| s_t; a_{t:t+H-1} \right]. \quad (3)$$

---

<sup>3</sup>Throughout this paper, we use  $\Delta(\mathcal{X})$  to denote the set of all probability distributions over a set  $\mathcal{X}$ .

Under HRC, executing a larger prefix  $j$  of the  $H$ -step action sequence generates longer, uninterrupted human-style motion, whereas  $j = 1$  maximizes reactivity.

**Action-segment quantization for efficient HRC.** Implementing equation (3) presents two obstacles: (1) Searching for all  $|\mathcal{H}|$  demonstration segments at every step is prohibitively slow. (2) Although the dataset records each visited state, the starting state of a stored segment still rarely matches the current state  $s_t$  exactly, so naively re-using segments requires expensive alignment or local optimization. To alleviate these issues, we propose to perform *segment-level planning*: HRC first proposes a complete  $H$ -step sequence  $(a_t, a_{t+1}, \dots, a_{t+H-1})$ , executes all  $H$  actions, observes the resulting state  $s_{t+H}$ , and then replans a new  $H$ -step action sequence from that state. This commit-and-replan loop amortizes the optimization overhead over fixed-length segments while fully re-using each trajectory rollout.

Moreover, searching for this segment in continuous action space is still computationally very costly, so we propose to adopt the *macro action quantization* strategy. Inspired by [27], we replace each segment of continuous actions with a compact sequence of discrete codebook indices, effectively reducing the search over the full space  $\mathcal{A}^H$  to a lookup on a finite codebook. Since all demonstrations can be rolled out and scored offline, the HRC controller needs only a constant-time table lookup at runtime, slashing planning latency while still retaining the multi-step look-ahead and human-style fidelity provided by  $H$ -step segments.

## 4 Macro Action Quantization

To concretize the trajectory optimization introduced in Section 3, we propose a human-likeness aware framework called *Macro Action Quantization* (MAQ). MAQ addresses the challenges of enforcing human-likeness in receding-horizon control by learning a quantized set of action segments – macro actions – from offline human demonstrations. These macro actions serve as  $H$ -step segments in the trajectory optimization process. The primary goal of MAQ is to achieve high performance while preserving the key characteristics of human-like behavior. We present the details of the MAQ framework as follows.

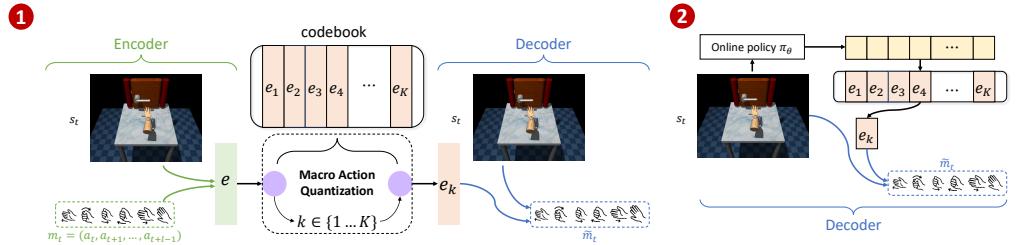


Figure 1: The overall training process of MAQ. (1) **Human behavior distillation:** A Conditional-VQVAE is trained on the state ( $s_t$ ) and the macro action ( $m_t = (a_t, a_{t+1}, \dots, a_{t+H-1})$ ) to learn a discrete codebook; the macro actions ( $m_t$ ) are extracted from human demonstrations via a sliding window over action trajectories. (2) **Reinforcement learning with MAQ:** An online policy ( $\pi_\theta$ ) acts in the learned discrete code space by selecting codebook indices, which the VQVAE decodes into macro action  $\tilde{m}_t$ , which aims to replicate the original macro action  $m_t$ .

**Human behavior distillation.** To capture the characteristics of human behavior, specifically the *action sequence* patterns, we propose training a Conditional-VQVAE to distill macro actions from human demonstrations, as illustrated on the left in Figure 1. Specifically, given a state  $s_t$  and a macro action  $m_t$ , where  $m_t = (a_t, a_{t+1}, \dots, a_{t+H-1})$  represents a sequence of  $H$  consecutive actions executed by a human at  $s_t$ , the encoder generates a latent vector  $e$ . This vector  $e$  is then compared with the codebook to identify the nearest latent vector  $e_k$ . Finally,  $e_k$  along with the state  $s_t$  is passed to the decoder to produce a reconstructed macro action  $\tilde{m}_t$ , which aims to replicate the original macro action  $m_t$ . During training, the loss function is updated as follows:

$$\mathcal{L} = \|m - \tilde{m}\|^2 + \|sg[e] - e_k\|^2 + \beta \|e - sg[e_k]\|^2, \quad (4)$$

where  $sg$  denotes the stop-gradient operation and  $\beta$  is the coefficient of commitment loss. MAQ not only captures the human behavior of a specific state and macro action pair  $(s_t, m_t)$  but also

generalizes the macro actions across similar states, with the codebook serving as a collection of human-like behaviors mapped to different states.

MAQ significantly reduces the macro action space. Since macro actions consist of sequences of actions, the number of possible macro actions grows exponentially with the sequence length. For example, given a sequence length  $H$  and assuming each state has  $n$  potential primitive actions, there are  $n^H$  possible macro actions. Fortunately, as the goal of MAQ is to capture human-like behavior, a vast number of these macro actions can be disregarded in practice since they do not appear in human demonstration. Therefore, we limit the codebook size to  $K$ , focusing only on macro actions that align with human behavior, effectively reducing the action space and distilling the key macro actions.

**Reinforcement learning with MAQ.** MAQ can be seamlessly integrated into any reinforcement learning algorithm. Conceptually, by quantizing macro actions, the action space is transformed from primitive actions (MDP) to macro actions (SMDP), while all other aspects of the RL training remain unchanged. Since these macro actions are distilled from human demonstration, RL agents aim to achieve high performance while operating within the constraints of human-like behaviors.

Specifically, we train a policy  $\pi_\theta$  to select the index  $k$  corresponding to the codebook entry. The policy  $\pi_\theta$  takes a given state  $s$  as input and produces a distribution with  $K$  policy logits, each corresponding to one macro action embedding in the codebook. At each state  $s_t$ ,  $\pi_\theta$  samples an index according to the policy distribution. Suppose the index  $k$  is selected, the corresponding latent vector  $e_k$  is then retrieved from the codebook. With this latent vector  $e_k$  and the state  $s_t$ , the decoder reconstructs a macro action  $\tilde{m}_t$  to interact with the environment, as illustrated on the right in Figure 1.

To optimize  $\pi_\theta$ , we follow equation (3) to find a macro action that maximizes the expected reward

$$m_t^* = \arg \max_{m_t \in \mathcal{H}} \mathbb{E} \left[ R(s_t, m_t) \mid s_t; m_t \right], \quad (5)$$

where  $m_t = a_{t:t+H-1}$  is the macro action and  $R(s_t, m_t) = \sum_{i=t}^{t+H-1} \gamma^{i-t} R(s_i, a_i)$  is the cumulative reward over the segment. Through this process, MAQ allows agents to plan and act with human-like behavior while optimizing for long-term return.

## 5 Experiments

This section empirically evaluates the proposed MAQ framework. Our experiments are designed to answer the two main questions: *Q1) Can MAQ accomplish tasks while optimizing trajectories that align with human demonstrations, and Q2) Can MAQ exhibit human-like behavior and convince human evaluators into believing it is human in a Turing Test evaluation?*

### 5.1 Experiment Setup

Our experiments are conducted on four Adroit tasks from D4RL [16], including *Door* (opening a door), *Hammer* (hammering a nail), *Pen* (twirling a pen), and *Relocate* (lifting and moving a ball). We first train three off-the-shelf RL algorithms, including IQL [29], SAC [30], and the state-of-the-art offline algorithm RLPD [6]. Next, we apply MAQ to each of these algorithms, resulting in three human-like RL agents – MAQ+IQL, MAQ+SAC, and MAQ+RLPD – to evaluate their generality across different RL algorithms. When incorporating MAQ, each RL algorithm learns a policy over a discrete codebook, replacing the original primitive action space with macro actions distilled from human demonstrations. This flexibility makes MAQ easy to apply to a wide range of RL algorithms. Note that to support the discrete action space introduced by MAQ, we adopt Discrete SAC (DSAC) [31] when applying it to SAC. In addition, we train a Behavioral Cloning (BC) [32] agent as a baseline for evaluating human-likeness.

Each task includes 25 human demonstration trajectories, which are split into training and testing datasets with a 9:1 ratio. Agents that require human demonstrations, including BC, IQL, and all MAQ-based agents, are trained on the same training datasets. All agents are evaluated based on trajectory similarity on the testing dataset. To ensure a fair comparison, all agents (including baseline and MAQ-based agents) are trained for a total of  $10^6$  steps. For IQL and MAQ+IQL, which include both offline and online training stages, we train  $10^6$  steps to each stage. For MAQ-based agents, we first train a VQVAE on human demonstrations to obtain the macro action codebook. The VQVAE is

trained with  $\beta = 0.25$  in Eq. (4), a hidden size of 256, a batch size of 32, and 100 training episodes. The same trained VQVAE is shared across all MAQ-based agents to ensure consistency. Other detailed hyperparameters are provided in Appendix A.

## 5.2 Trajectory Similarity Evaluation

### 5.2.1 Evaluation Metrics

To evaluate how closely agent behaviors align with human demonstrations, we use two trajectory similarity metrics: Dynamic Time Warping (DTW) and the Wasserstein Distance (WD). DTW measures the alignment between trajectories, while WD captures distributional similarity. Both metrics are computed between agent and human trajectories from the same testing dataset.

**Dynamic time warping.** DTW is a technique used to measure the similarity between two sequences, even if they differ in length or timing. It finds the optimal alignment between two sequences by stretching or compressing segments to minimize the overall distance. Generally, a lower distance indicates higher similarity. Specifically, let  $\mathcal{T}^H = \{\tau_1^H, \tau_2^H, \dots, \tau_n^H\}$  and  $\mathcal{T}^A = \{\tau_1^A, \tau_2^A, \dots, \tau_m^A\}$  denote the sets of trajectories played by the human and the agent, consisting of  $n$  and  $m$  trajectories, respectively. The DTW distance between  $\mathcal{T}^H$  and  $\mathcal{T}^A$  is calculated by  $\text{DTW} = \frac{1}{n} \sum_{i=1}^n (\frac{1}{m} \sum_{j=1}^m \text{DTW}(\tau_i^H, \tau_j^A))$ . Moreover, we compare two types of DTW distances: state distance  $\text{DTW}_s$  and action distance  $\text{DTW}_a$ . The  $\text{DTW}_s$  and  $\text{DTW}_a$  are calculated by applying DTW to the state and action sequences in  $\text{DTW}(\tau^H, \tau^A)$ , respectively. A lower  $\text{DTW}_s$  indicates that the agent frequently encounters states in a similar order to those in human demonstrations, while a lower  $\text{DTW}_a$  suggests that the agent performs actions in an order similar to human demonstrations.

**Wasserstein distance.** Compared to DTW, WD evaluates distributional similarity with the Wasserstein distance between the agent and human demonstrations:  $W(\rho_{\text{agent}}, \rho_{\text{human}})$ . The calculation is performed on the same normalized feature space used for DTW and uses the POT library’s emd2 solver to compute the Wasserstein distance. Similar to DTW, we evaluate two Wasserstein variants: state-based Wasserstein  $\text{WD}_s$  and action-based Wasserstein  $\text{WD}_a$ . The  $\text{WD}_s$  and  $\text{WD}_a$  are calculated by applying  $W(\rho_{\text{agent}}, \rho_{\text{human}})$  to the state and action.

### 5.2.2 Performance of MAQ-based Agents

Table 1 summarizes results for the four Adroit tasks, reporting state and action-based Dynamic Time Warping ( $\text{DTW}_s$ ,  $\text{DTW}_a$ ) and Wasserstein distance ( $\text{WD}_s$ ,  $\text{WD}_a$ ) scores together with task success rates. For better comparability across tasks, we normalize the similarity scores for each task using  $1 - \frac{\text{agent score} - \text{human score}}{\text{random score} - \text{human score}}$ , where the agent score is the similarity between the agent and human trajectories, the human score is the similarity between human trajectories, and the random score is the similarity between a random agent and human trajectories. After transformation, higher values indicate more human-like behavior.

Table 1 shows that incorporating MAQ significantly increases trajectory similarity across all tasks, according to both DTW and WD scores. For example, in the *Door* task,  $\text{DTW}_s$  improves from 0.43 with IQL to 0.84 with MAQ+IQL, from -0.39 with SAC to 0.8 with MAQ+SAC, and from -0.06 with RLPD to 0.76 with MAQ+RLPD. Similarly, in the *Hammer* task,  $\text{WD}_a$  improves from -0.19 with SAC to 0.78 with MAQ+SAC, and from 0.2 with RLPD to 0.85 with MAQ+RLPD. Most importantly, these improvements in trajectory similarity are achieved without significantly sacrificing task success rate. In addition, although the BC agent learns directly from human demonstrations and achieves moderate similarity scores, it performs the worst in terms of task success rate among all agents. This is because the BC agent learns to imitate state-action pairs from the demonstrations, without optimizing for successful outcomes. Overall, the largest improvement is observed between SAC and MAQ+SAC, where the average  $\text{DTW}_s$  increases from -0.49 to 0.56. This is because SAC learns solely from environment interaction without using any human demonstrations, unlike IQL and RLPD. As a result, its trajectories deviate more significantly from those of humans. Our findings also suggest that MAQ has significant potential when applied to more complex problems, such as tasks requiring intricate scenarios with numerous macro actions (e.g., continuous action spaces).

Table 1: The results for Adroit benchmark from D4RL. The baseline of each algorithm incorporate with MAQ generally improves all of the original algorithm similarity score in each control task.

Tasks		BC	IQL	MAQ+IQL	SAC	MAQ+SAC	RLPD	MAQ+RLPD
Door	DTW <sub>s</sub> (↑)	0.18 ± 0.09	0.43 ± 0.06	<b>0.84 ± 0.06</b>	-0.39 ± 0.10	<b>0.80 ± 0.08</b>	-0.06 ± 0.04	<b>0.76 ± 0.04</b>
	DTW <sub>a</sub> (↑)	0.42 ± 0.13	0.61 ± 0.04	<b>0.95 ± 0.01</b>	-0.25 ± 0.04	<b>0.91 ± 0.03</b>	0.28 ± 0.08	<b>0.91 ± 0.05</b>
	WD <sub>s</sub> (↑)	0.32 ± 0.05	0.48 ± 0.05	<b>0.75 ± 0.05</b>	-0.28 ± 0.02	<b>0.71 ± 0.08</b>	-0.14 ± 0.04	<b>0.71 ± 0.03</b>
	WD <sub>a</sub> (↑)	0.41 ± 0.08	0.50 ± 0.02	<b>0.81 ± 0.03</b>	-0.15 ± 0.02	<b>0.77 ± 0.07</b>	0.10 ± 0.02	<b>0.76 ± 0.03</b>
	Success(↑)	0.02 ± 0.01	0.16 ± 0.06	<b>0.93 ± 0.04</b>	0.43 ± 0.23	<b>0.56 ± 0.50</b>	<b>0.96 ± 0.07</b>	0.93 ± 0.05
Hammer	DTW <sub>s</sub> (↑)	-0.16 ± 0.07	-0.14 ± 0.34	<b>0.64 ± 0.17</b>	-1.10 ± 0.35	<b>0.61 ± 0.21</b>	-0.03 ± 0.14	<b>0.68 ± 0.17</b>
	DTW <sub>a</sub> (↑)	0.47 ± 0.02	0.45 ± 0.20	<b>0.92 ± 0.06</b>	-0.33 ± 0.11	<b>0.91 ± 0.10</b>	0.37 ± 0.08	<b>0.94 ± 0.07</b>
	WD <sub>s</sub> (↑)	0.11 ± 0.06	0.12 ± 0.13	<b>0.75 ± 0.03</b>	-0.44 ± 0.09	<b>0.64 ± 0.12</b>	-0.03 ± 0.08	<b>0.76 ± 0.04</b>
	WD <sub>a</sub> (↑)	0.30 ± 0.03	0.30 ± 0.10	<b>0.84 ± 0.02</b>	-0.19 ± 0.04	<b>0.78 ± 0.11</b>	0.20 ± 0.03	<b>0.85 ± 0.03</b>
	Success(↑)	0.00 ± 0.00	<b>0.01 ± 0.01</b>	0.00 ± 0.00	<b>0.01 ± 0.01</b>	0.00 ± 0.00	<b>1.00 ± 0.00</b>	0.56 ± 0.37
Pen	DTW <sub>s</sub> (↑)	0.53 ± 0.13	0.34 ± 0.09	<b>0.55 ± 0.17</b>	0.06 ± 0.20	<b>0.58 ± 0.17</b>	0.48 ± 0.24	<b>0.54 ± 0.18</b>
	DTW <sub>a</sub> (↑)	0.58 ± 0.05	0.51 ± 0.05	<b>0.58 ± 0.09</b>	-0.34 ± 0.16	<b>0.58 ± 0.11</b>	0.40 ± 0.17	<b>0.59 ± 0.13</b>
	WD <sub>s</sub> (↑)	0.59 ± 0.12	0.54 ± 0.11	<b>0.59 ± 0.13</b>	0.29 ± 0.10	<b>0.61 ± 0.12</b>	0.49 ± 0.08	<b>0.59 ± 0.12</b>
	WD <sub>a</sub> (↑)	0.65 ± 0.14	0.63 ± 0.12	<b>0.66 ± 0.15</b>	0.22 ± 0.15	<b>0.67 ± 0.14</b>	0.44 ± 0.12	<b>0.66 ± 0.14</b>
	Success(↑)	0.40 ± 0.03	0.40 ± 0.05	<b>0.42 ± 0.07</b>	0.32 ± 0.09	<b>0.41 ± 0.01</b>	<b>0.62 ± 0.09</b>	0.42 ± 0.05
Relocate	DTW <sub>s</sub> (↑)	0.09 ± 0.14	0.20 ± 0.20	<b>0.52 ± 0.06</b>	-0.55 ± 0.20	<b>0.25 ± 0.19</b>	0.03 ± 0.13	<b>0.27 ± 0.14</b>
	DTW <sub>a</sub> (↑)	0.47 ± 0.15	0.51 ± 0.11	<b>0.82 ± 0.01</b>	-0.10 ± 0.16	<b>0.66 ± 0.09</b>	0.32 ± 0.13	<b>0.69 ± 0.10</b>
	WD <sub>s</sub> (↑)	0.27 ± 0.15	0.36 ± 0.06	<b>0.47 ± 0.07</b>	-0.22 ± 0.06	<b>0.40 ± 0.07</b>	0.02 ± 0.07	<b>0.38 ± 0.09</b>
	WD <sub>a</sub> (↑)	0.45 ± 0.12	0.50 ± 0.04	<b>0.65 ± 0.03</b>	-0.05 ± 0.03	<b>0.61 ± 0.03</b>	0.20 ± 0.03	<b>0.55 ± 0.08</b>
	Success(↑)	0.01 ± 0.02	0.00 ± 0.00	<b>0.20 ± 0.10</b>	0.00 ± 0.00	<b>0.14 ± 0.07</b>	0.14 ± 0.03	<b>0.17 ± 0.10</b>
Average	DTW <sub>s</sub> (↑)	0.16 ± 0.29	0.21 ± 0.25	<b>0.63 ± 0.14</b>	-0.49 ± 0.48	<b>0.56 ± 0.23</b>	0.10 ± 0.25	<b>0.56 ± 0.21</b>
	DTW <sub>a</sub> (↑)	0.49 ± 0.07	0.52 ± 0.06	<b>0.82 ± 0.17</b>	-0.26 ± 0.11	<b>0.76 ± 0.17</b>	0.34 ± 0.05	<b>0.78 ± 0.17</b>
	WD <sub>s</sub> (↑)	0.32 ± 0.20	0.38 ± 0.18	<b>0.64 ± 0.14</b>	-0.17 ± 0.32	<b>0.59 ± 0.14</b>	0.08 ± 0.28	<b>0.61 ± 0.17</b>
	WD <sub>a</sub> (↑)	0.45 ± 0.15	0.48 ± 0.14	<b>0.74 ± 0.10</b>	-0.04 ± 0.18	<b>0.71 ± 0.08</b>	0.24 ± 0.15	<b>0.70 ± 0.13</b>
	Success(↑)	0.11 ± 0.19	0.14 ± 0.19	<b>0.39 ± 0.40</b>	0.19 ± 0.22	<b>0.28 ± 0.25</b>	<b>0.68 ± 0.40</b>	0.52 ± 0.32

### 5.2.3 Human-Like Trajectory Optimization with Different Macro Action Lengths

We examine how the macro action length  $H$  in the MAQ codebook affects the human-likeness similarity score. We evaluate the MAQ+RLPD across different sequence lengths ( $H = 1$  to  $9$ ), using the same five metrics illustrated in Table 1, averaged over the four *Adroit* tasks. Figure 2 shows the results, where each cell is color-coded with darker shades indicating higher values. The color scale is normalized independently for each metric. The results show that  $H = 9$  yields the highest similarity scores as well as the best task success rate. Notably,  $WD_a$  remains relatively stable across all values of  $H$ , likely due to the nature of the Adroit action space, which consists solely of Shadow Hand joint positions. This suggests that even shorter sequences can appear human-like in action space without being effective at completing the task. As  $H$  increases, both trajectory similarity and success rate improve, indicating that longer macro actions not only enhance human-likeness similarity score but also contribute to more effective task execution, highlighting the benefits of temporally extended planning in human-like trajectory optimization. We have also provided additional analysis of MAQ in Appendix B.

## 5.3 Human Evaluation Study

While the trajectory similarity scores demonstrate that MAQ-based agents align more closely with human behavior compared to other methods, we also conduct a human evaluation study with 19 human evaluators to verify whether their behavior is recognized as human-like by human evaluators. Specifically, each evaluator study is asked to complete two sets of questions, including a Turing Test and a human-likeness ranking test. Detailed settings for human evaluation stduy are provided in Appendix C. The following subsections provide the results on each evaluation test.

### 5.3.1 Turing Test

In the Turing Test, we conduct several two-alternative forced-choice (2AFC) questions, where evaluators are shown two videos – one from human demonstrations and the other from one of the seven trained agents listed in Table 1 – and asked to choose which one was performed by humans. To ensure fairness, the evaluation consists of seven questions for each Adroit task, each corresponding to a different agent. Namely, each agent appears exactly once, and the order of appearance is randomized for each evaluator.



Figure 2: Trajectory similarity scores and success rates with different lengths in MAQ+RLPD.

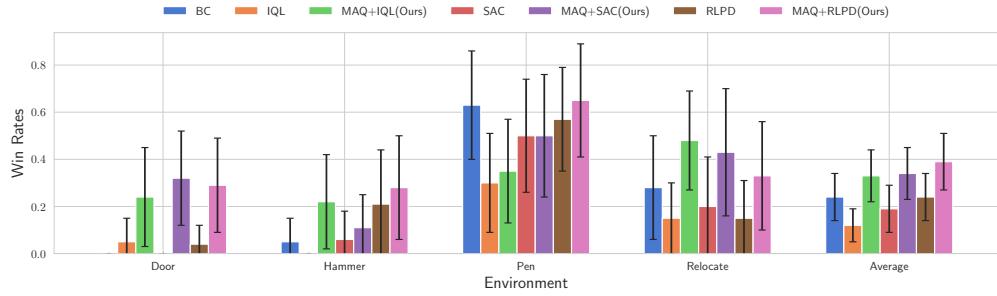


Figure 3: Turing Test. Error bars represent 96% confidence intervals.

Figure 3 shows the results of the Turing Test. Bars indicate the win rate of each agent, i.e., the percentage of questions in which evaluators were fooled by the agent and believed its behavior was more human-like than the human demonstration. The results are consistent with the findings in Subsection 5.2, where MAQ-based agents achieve higher win rates compared to RL algorithms without MAQ. This further corroborates that optimizing trajectories with longer macro action lengths is strongly correlated with human-likeness. Overall, when comparing the average win rate across four tasks, the win rate evaluated by the evaluators is: MAQ+RLPD (39%) > MAQ+SAC (34%) > MAQ+IQL (32%) > BC (24%) = RLPD (24%) > SAC (19%) > IQL (13%). Notably, MAQ+RLPD improves a win rate of 15% higher than non-MAQ agents, demonstrating that our method significantly enhances the perceived human-likeness of agent behavior.

The figure also shows some interesting results. The *Pen* task exhibits the highest win rate for all agents, indicating that evaluators have the greatest difficulty distinguishing agent behavior from human demonstrations in this setting. This suggests that pen manipulating is relatively easy for agents to master, and that the visual differences between human and agent behavior are subtle, making deception more likely. In addition, the SAC agent shows nearly 0% win rates in the *Door* task. Despite achieving a 43% success rate, its behavior remains noticeably different from human demonstrations. This discrepancy clearly demonstrates that reward-driven RL agents are not necessarily human-like.

### 5.3.2 Human-Likeness Ranking Test

The questions in the human-likeness ranking test are similar to those in the Turing Test, but with a key difference, as described below. Since the goal is to compare human-likeness among agents, each question presents evaluators with two videos; however, both videos may be from trained agents.

Evaluators are asked to choose the video that appears more human-like. For fairness, each agent pair appears only once per evaluator, and not all possible agent pairs are shown for evaluators to reduce fatigue. Each evaluator is asked to complete approximately 17 questions during this test. After completing the test, each evaluator is shown two videos – one selected from the most human-like and the other from the least human-like agents or humans based on their responses – and is invited to leave comments explaining their choices.

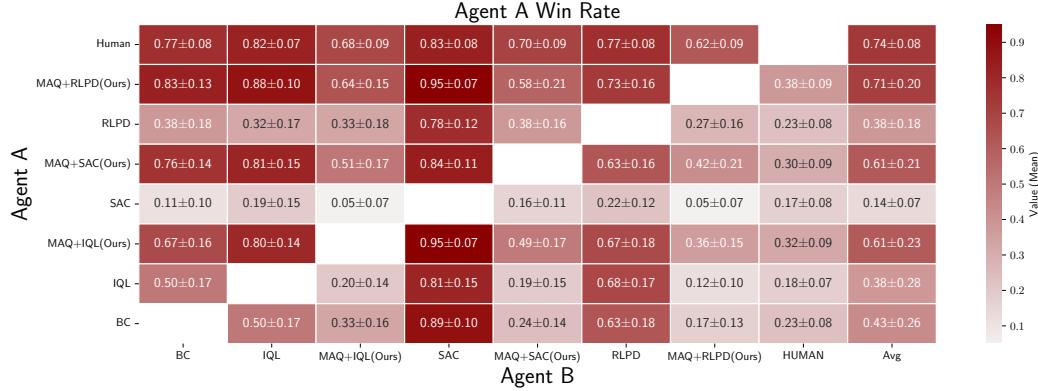


Figure 4: Human-likeness ranking test.

Figure 4 summarizes the results across the four tasks. The heatmap shows the win rate, where each cell indicates how often evaluators judged agents on the y-axis to behave more human-like than agents on the x-axis when comparing their videos directly. Overall, the results are consistent with the Turning Test, where the overall ranking among all agents and humans as follows: Human (74%) > MAQ+RLPD (71%) > MAQ+SAC (61%) = MAQ+IQL (61%) > BC (43%) > RLPD (38%) = IQL (38%) > SAC (14%). Interestingly, the ranking order remains the same as in the Turing Test, except that the positions of IQL and SAC are reversed. This indicates that human evaluators are consistent in their judgments and can distinguish between human behavior and that of RL agents. Moreover, MAQ+RLPD achieves a 71% win rate across all agent pairs, achieving performance comparable to the human demonstration's 74% win rate. This suggests that MAQ+RLPD can convincingly fool human evaluators into believing its behavior is human-like.

### 5.3.3 Human Feedback on Behavior Analysis

We further investigate the feedback provided by human evaluators. Figure 5 shows video clips in *Door*. The MAQ+RLPD agent demonstrates precise control by firmly holding, rotating the door handle, and pulling to open the door, closely mimicking human behavior. In contrast, while the RLPD agent achieves a high success rate in this task, it employs an unnatural method to open the door – using its backhand to press the handle while simultaneously pulling the door without properly holding the handle. Although this approach successfully opens the door, the behavior appears less human-like. Therefore, several evaluators judged the MAQ+RLPD agent to be more human-like. As one evaluator noted, “*Because it shows how a human casually opens a door — the grip on the handle is stable, without forcefully grabbing it all the way*”. Conversely, many evaluators believe RLPD agent does not exhibit human-like behavior, remarking, “*Instead of grabbing, it is glitching its hand through the door handle*” and “*Its wrist performed a weird gesture and opened the door rudely*.”

## 6 Discussions

In this paper, we discuss a critical and yet underexplored challenge in reinforcement learning for designing human-like agents. We first formulate human-likeness as a trajectory optimization problem, and realize this formulation by proposing MAQ, a human-likeness aware framework that distills human behaviors into macro actions. Our experiments show that MAQ not only completes tasks but also aligns with human behaviors by improving the trajectory similarity scores across four Adroit control tasks, and achieves the highest human-likeness rankings in human evaluation. These results demonstrate that MAQ agents are both effective and convincingly human-like.



(a) MAQ+RLPD



(b) RLPD

Figure 5: Agent behavior in *Door*.

Although MAQ exhibits more human-like behavior, one limitation is its reliance on human demonstrations to distill macro actions. The quality of these demonstrations can affect the effectiveness of MAQ. However, since our primary goal is to pursue human-likeness, obtaining human demonstrations is both necessary and appropriate within the scope of this work. Moreover, our experiments demonstrate that MAQ can be easily integrated into various RL algorithms. We also show that MAQ has the potential to generalize to more complex domains, such as real-time strategy games, where agents often act at unnaturally high frequencies to maximize performance. Rather than relying on manually designed constraints (e.g., delaying action execution [3]), MAQ naturally regulates decision frequency through learned human-like macro actions, allowing agents to behave more realistically. In conclusion, MAQ offers a promising direction for learning human-like RL agents.

## Broader Impact

This work investigates human-like reinforcement learning by introducing a human-likeness aware framework, called MAQ, that allows AI agents to behave in a more natural and human-like manner. Human-like behavior can offer practical benefits in real-world applications, particularly in scenarios involving human-robot collaboration. For example, a robot that assists with human-like movements can improve safety and trust, reducing the risk of accidents. On the other hand, human-like agents may be misused in deceptive or manipulative ways, such as cheating in competitive games. In summary, the proposed method and human-likeness metrics provide a standardized approach for evaluating and quantifying human-like behavior, benefiting the broader RL community.

## Acknowledgement

This research is partially supported by the National Science and Technology Council (NSTC) of the Republic of China (Taiwan) under Grant Numbers 113-2221-E-001-009-MY3, 113-2634-F-A49-004, 114-2221-E-A49-005, 114-2221-E-A49-006, and 114-2628-E-A49-002. We also thank the National Center for High-performance Computing (NCHC) for providing computational and storage resources. The authors would also like to thank the anonymous reviewers for their valuable comments and the anonymous participants for the Human Evaluation Study.

## References

- [1] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, December 2022.
- [2] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy

- Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, December 2020.
- [3] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, November 2019.
- [4] OpenAI, Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with Large Scale Deep Reinforcement Learning, December 2019.
- [5] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pages 1–7, April 2025.
- [6] Philip J. Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient Online Reinforcement Learning with Offline Data. In *Proceedings of the 40th International Conference on Machine Learning*, pages 1577–1594. PMLR, July 2023.
- [7] Siddharth Mysore, Bassel Mabsout, Renato Mancuso, and Kate Saenko. Regularizing Action Policies for Smooth Control with Reinforcement Learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1810–1816. IEEE Press, May 2021.
- [8] I Lee, Hoang-Giang Cao, Cong-Tinh Dao, Yu-Cheng Chen, and I-Chen Wu. Gradient-based Regularization for Action Smoothness in Robotic Control with Reinforcement Learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 603–610, October 2024.
- [9] Qianli Shen, Yan Li, Haoming Jiang, Zhaoran Wang, and Tuo Zhao. Deep Reinforcement Learning with Robust and Smooth Policy. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8707–8718. PMLR, November 2020.
- [10] William Koch. Flight Controller Synthesis Via Deep Reinforcement Learning, September 2019.
- [11] Stephanie Milani, Arthur Juliani, Ida Momennejad, Raluca Georgescu, Jaroslaw Rzepecki, Alison Shaw, Gavin Costello, Fei Fang, Sam Devlin, and Katja Hofmann. Navigates Like Me: Understanding How People Evaluate Human-Like AI in Video Games. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI ’23*, pages 1–18. Association for Computing Machinery, April 2023.
- [12] Sam Devlin, Raluca Georgescu, Ida Momennejad, Jaroslaw Rzepecki, Evelyn Zuniga, Gavin Costello, Guy Leroy, Ali Shaw, and Katja Hofmann. Navigation Turing Test (NTT): Learning to Evaluate Human-Like Navigation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2644–2653. PMLR, July 2021.
- [13] Evelyn Zuniga, Stephanie Milani, Guy Leroy, Jaroslaw Rzepecki, Raluca Georgescu, Ida Momennejad, Dave Bignell, Mingfei Sun, Alison Shaw, Gavin Costello, Mikhail Jacob, Sam Devlin, and Katja Hofmann. How Humans Perceive Human-like Behavior in Video Game Navigation. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, CHI EA ’22*, pages 1–11. Association for Computing Machinery, April 2022.
- [14] Kuo-Hao Ho, Ping-Chun Hsieh, Chiu-Chou Lin, You-Ren Lou, Feng-Jian Wang, and I.-Chen Wu. Towards Human-Like RL: Taming Non-Naturalistic Behavior in Deep RL via Adaptive Behavioral Costs in 3D Games. In *Proceedings of the 15th Asian Conference on Machine Learning*, pages 438–453. PMLR, February 2024.

- [15] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [16] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for Deep Data-Driven Reinforcement Learning, February 2021.
- [17] Nobuto Fujii, Yuichi Sato, Hironori Wakama, Koji Kazai, and Haruhiro Katayose. Evaluating Human-like Behaviors of Video-Game Agents Autonomously Acquired with Biological Constraints. In *Proceedings of the 10th International Conference on Advances in Computer Entertainment - Volume 8253*, ACE 2013, pages 61–76. Springer-Verlag, November 2013.
- [18] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, Gabriel Dulac-Arnold, John Agapiou, Joel Leibo, and Audrunas Gruslys. Deep Q-learning From Demonstrations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018.
- [19] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to End Learning for Self-Driving Cars, April 2016.
- [20] Maryam Zare, Parham M. Kebria, Abbas Khosravi, and Saeid Nahavandi. A Survey of Imitation Learning: Algorithms, Recent Developments, and Challenges. *IEEE Transactions on Cybernetics*, 54(12):7173–7186, December 2024.
- [21] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.
- [22] Ishan P. Durugkar, Clemens Rosenbaum, Stefan Dernbach, and Sridhar Mahadevan. Deep Reinforcement Learning With Macro-Actions, June 2016.
- [23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL <https://arxiv.org/abs/1312.6114>.
- [24] Sherjil Ozair, Yazhe Li, Ali Razavi, Ioannis Antonoglou, Aaron Van Den Oord, and Oriol Vinyals. Vector Quantized Models for Planning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8302–8313. PMLR, July 2021.
- [25] Ioannis Antonoglou, Julian Schrittwieser, Sherjil Ozair, Thomas K. Hubert, and David Silver. Planning in Stochastic Environments with a Learned Model. In *International Conference on Learning Representations*, October 2021.
- [26] Andros Tjandra, Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura. VQVAE Unsupervised Unit Discovery and Multi-Scale Code2Spec Inverter for Zerospeech Challenge 2019. In *Proc. Interspeech 2019*, pages 1118–1122, 2019.
- [27] Jianlan Luo, Perry Dong, Jeffrey Wu, Aviral Kumar, Xinyang Geng, and Sergey Levine. Action-Quantized Offline Reinforcement Learning for Robotic Skill Learning. In *Proceedings of The 7th Conference on Robot Learning*, pages 1348–1361. PMLR, December 2023.
- [28] Nicklas A. Hansen, Hao Su, and Xiaolong Wang. Temporal Difference Learning for Model Predictive Control. In *Proceedings of the 39th International Conference on Machine Learning*, pages 8387–8406. PMLR, June 2022.
- [29] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline Reinforcement Learning with Implicit Q-Learning. In *International Conference on Learning Representations*, October 2021.
- [30] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1861–1870. PMLR, July 2018.
- [31] Haibin Zhou, Tong Wei, Zichuan Lin, junyou li, Junliang Xing, Yuanchun Shi, Li Shen, Chao Yu, and Deheng Ye. Revisiting Discrete Soft Actor-Critic, November 2024.

- [32] Christopher G. Atkeson and Stefan Schaal. Robot Learning From Demonstration. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 12–20. Morgan Kaufmann Publishers Inc., July 1997.
- [33] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.
- [34] gwthomas. Iql-pytorch. <https://github.com/gwthomas/IQL-PyTorch>, 2021.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Our contributions are summarized in both the abstract and introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discussed the limitations in Section 6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We describe our assumption in Section 3.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all the information for reproducing the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We have provided all necessary information for reproducing our main results. Once the paper is accepted, we will release all codes and trained models.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The experimental details are discussed in both Section 5 and Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide the error bars for all tables and figures.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The experiment details are provided in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The paper conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We have discussed broader impacts in the Discussion.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We don't release models or datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Yes, the dataset in this paper is properly credited and mentioned.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We don't release models or datasets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: We have provided detailed instructions for all participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper has no such issues.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: LLM is used only for editing to polishing the writing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

## A Experimental settings

**Experimental Setup** We conducted our experiments on a machine equipped with an E5-2678 CPU and four NVIDIA GeForce GTX 1080 Ti GPUs. Training was performed on various games, each requiring a different number of GPU hours depending on the specific task. Our method begins by training a VQVAE with varying sequence lengths  $H$  and codebook sizes  $K$ . For each game, we used three different random seeds: 1, 10, and 100. On average, VQVAE training required less than 5 minutes. Each MAQ-based method is built upon a pre-trained VQVAE. All MAQ variants share the same VQVAE model when trained with the same random seed. We consistently selected the model from the final training iteration as the VQVAE used in our online reinforcement learning (RL) experiments. For online RL training, we evaluated the best performing checkpoint using two similarity metrics: Dynamic Time Warping (DTW) and Wasserstein distance. Among all variants, MAQ+IQL was the most computationally intensive, requiring 9 GPU hours per training run. In contrast, MAQ+SAC and MAQ+RLPD required only 1 GPU hour each. Regarding the vanilla RL, IQL, and RLPD, the training required approximately 8 GPU hours per run, and SAC training required 1 GPU hour per run. This difference stems from the experimental setup: MAQ+IQL was trained in a single-environment setting, while MAQ+SAC and MAQ+RLPD were trained using 8 parallel environments. As a result, the training time for MAQ+IQL was approximately 8 times longer, and for MAQ+SAC and MAQ+RLPD, updates were performed 8 times per iteration.

**Hyperparameters** All hyperparameters used in the online RL methods are detailed in Table 2, and those for Conditional VQVAE are listed in Table 3. For SAC, we used the default hyperparameters from Stable-Baselines3 [33]. For IQL [29], we adopted the PyTorch implementation by gwthomas’s repository released on Github [34], which most closely follows the original paper. For RLPD [6], we used the official implementation released on GitHub.

In our MAQ-based methods, we made the following modifications. In MAQ+IQL, we built upon the IQL codebase and integrated the VQVAE decoder after the policy, enabling the policy to make decisions conditioned on the codebook size  $K$ . In MAQ+SAC, we extended the implementation of Discrete SAC [31] by incorporating the VQVAE decoder. For MAQ+RLPD, we built on MAQ+SAC and additionally implemented the symmetric sampling scheme from RLPD, using a symmetric ratio of 0.5 (equivalent to the offline ratio shown in Table 2).

**Dataset Segmentation** For data partitioning, we used the human dataset provided by D4RL. Each task consists of 25 successful human demonstrations, which we split into training and testing sets in a 9:1 ratio, resulting in 22 trajectories for training and 3 for testing. We used the testing dataset to evaluate similarity between agents. The normalization procedure mentioned in the paper involves generating a random agent and using its performance on the testing dataset to normalize the similarity scores. Table 4 presents all the detailed results, including the raw data, training data, and testing data.

**Training Curves for All Environments** Figure 6 presents the training curves of the experiments described in Subsection 5.2. The results are based on the normalized rewards across different environments. For Offline-to-Online RL (O2ORL) methods such as IQL and MAQ+IQL, training curves are plotted in environment steps and the x-axis is scaled by 0.5, so that 2M steps are compressed to the same span as 1M steps in other settings. Shaded regions indicate the standard deviation across three different random seeds.

**Environment Reward and Success Rate** In D4RL, the success rate indicates whether the agent completes the task within the episode limit (e.g., opening the Door within 200 steps). In contrast, the normalized reward is a heuristic score defined by the environment that reflects how close the agent is to completing the task. For example, in the *Door* task, the reward increases as the door approaches a fully open state, but the episode is marked as successful only when the door is completely opened.

As shown in Figure 6 and summarized in Table 4, RLPD and MAQ+RLPD achieve similar success rates. However, their normalized rewards differ substantially: RLPD tends to maximize reward by opening the door quickly, whereas MAQ+RLPD opens the door more slowly and naturally while maintaining the same success rate.

**Computational Cost Analysis of MAQ** Since there are 25 trajectories per task, the VQVAE can be trained in a very short time. Specifically, training VQVAE takes only 3 to 4 minutes, whereas RL

training requires several hours. While MAQ-based RL introduces some additional computational overhead, Table 5 presents a comparison of inference time for each RL algorithm, with and without MAQ.

It is also worth noting that MAQ can offer better inference efficiency, as it executes an entire macro action (e.g., 9 consecutive actions in the paper) with a single forward pass, whereas vanilla RL must do a forward pass at every timestep. In this case, MAQ-based RL can even achieve lower overall inference time than vanilla RL, especially when longer action sequences are used.

Table 2: Hyperparameters for training Adroit tasks.

Parameter	BC	IQL	MAQ+IQL	SAC	MAQ+SAC	RLPD	MAQ+RLPD
Batch size	256	256	256	128	128	256	128
Learning rate	3e-4	3e-4	3e-4	3e-4	3e-4	-	-
Actor learning rate	-	-	-	-	-	3e-4	3e-4
Critic learning rate	-	-	-	-	-	3e-4	1e-3
Temperature learning rate	-	-	-	-	-	3e-4	3e-4
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam	Adam
Offline training steps	1M	1M	1M	-	-	-	-
Online training steps	-	1M	1M	1M	1M	1M	1M
Discount factor $\gamma$	-	0.99	0.99	0.99	0.99	0.99	0.99
Warm-up steps	-	-	-	1e2	8e3	1e4	8e3
Update epoch	-	-	-	1	8	1	8
Value coeff	-	-	-	-	-	-	-
Entropy coeff	-	-	-	auto	auto	auto	auto
Offline ratio	-	-	-	-	-	0.5	0.5
Temperature alpha $\alpha$	-	-	-	-	-	0.2	1.0
Replay buffer size	-	2M	2M	1M	1M	1M	1M
Target network update rate $\tau$	-	0.005	0.005	0.005	0.005	0.005	0.005
Advantage coeff $\lambda$	-	-	-	-	-	-	-
Asymmetric loss coeff $\tau$	-	0.7	0.7	-	-	-	-
Inverse temperature $\beta$	-	3.0	3.0	-	-	-	-
Std of Gaussian exploration noise	-	0.03	0.03	-	-	-	-
Range to clip noise	-	0.5	0.5	-	-	-	-
MAQ Codebook size $K$	-	-	16	-	8	-	16
MAQ Macro action length $H$	-	-	9	-	8	-	9

Table 3: Hyperparameters of training MAQ’s Conditional-VQVAE.

Parameter	Adroit
Latent size	256
Learning rate	3e-4
Codebook size $K$	[8, 16, 32]
Batch size	32
Commitment loss coeff $\beta$	0.25
Macro length $H$	[1 ... 9]
Optimizer	Adam

Table 4: Detailed numerical results corresponding to the main table, including raw scores, training and testing data, and normalization baselines used for similarity evaluation.

Tasks	BC	IQL	MAQ+IQL	SAC	MAQ+SAC	RLPD	MAQ+RLPD	Training Dataset	Testing Dataset	Random
Door	DTW <sub>s</sub> ( $\downarrow$ )	560.738 ± 39.892	451.199 ± 27.652	<b>266.994 ± 26.499</b>	820.125 ± 47.292	<b>283.977 ± 34.359</b>	67.954 ± 19.864	<b>302.186 ± 19.757</b>	285.085 ± 21.432	193.165 ± 30.694
	DTW <sub>a</sub> ( $\downarrow$ )	700.093 ± 110.944	536.414 ± 33.763	<b>237.304 ± 6.737</b>	1285.045 ± 37.101	<b>274.658 ± 24.079</b>	823.852 ± 70.334	<b>275.209 ± 41.988</b>	299.011 ± 11.922	192.035 ± 12.539
	WD <sub>s</sub> ( $\downarrow$ )	6.721 ± 0.261	5.864 ± 0.263	<b>4.388 ± 0.258</b>	9.988 ± 0.127	<b>4.561 ± 0.455</b>	9.230 ± 0.213	<b>4.587 ± 0.189</b>	4.307 ± 0.134	3.015 ± 0.322
	WD <sub>a</sub> ( $\downarrow$ )	6.017 ± 0.531	5.429 ± 0.109	<b>3.446 ± 0.212</b>	9.697 ± 0.162	<b>3.701 ± 0.437</b>	8.076 ± 0.116	<b>3.734 ± 0.201</b>	3.304 ± 0.118	2.176 ± 0.046
Hammer	Success( $\uparrow$ )	0.020 ± 0.010	0.163 ± 0.055	<b>0.930 ± 0.040</b>	0.433 ± 0.232	<b>0.563 ± 0.497</b>	<b>0.957 ± 0.067</b>	0.933 ± 0.049	1.000 ± 0.000	1.000 ± 0.000
	DTW <sub>s</sub> ( $\downarrow$ )	894.540 ± 24.821	887.642 ± 129.362	<b>595.441 ± 63.226</b>	1246.982 ± 131.710	<b>607.701 ± 79.326</b>	845.877 ± 52.838	<b>578.816 ± 62.884</b>	642.887 ± 48.775	459.715 ± 96.337
	DTW <sub>a</sub> ( $\downarrow$ )	1058.832 ± 19.802	1083.427 ± 223.472	<b>551.378 ± 72.545</b>	1959.679 ± 118.135	<b>570.395 ± 111.086</b>	1178.736 ± 91.087	<b>528.520 ± 81.925</b>	654.830 ± 59.219	465.388 ± 104.728
	WD <sub>s</sub> ( $\downarrow$ )	8.792 ± 0.335	8.718 ± 0.691	<b>5.243 ± 0.176</b>	11.836 ± 0.509	<b>5.893 ± 0.649</b>	9.571 ± 0.443	<b>5.202 ± 0.207</b>	5.457 ± 0.166	3.892 ± 0.292
Pen	WD <sub>a</sub> ( $\downarrow$ )	6.670 ± 0.206	6.667 ± 0.586	<b>3.371 ± 0.093</b>	9.601 ± 0.234	<b>3.751 ± 0.686</b>	7.252 ± 0.185	<b>3.320 ± 0.209</b>	3.390 ± 0.077	2.394 ± 0.254
	Success( $\uparrow$ )	0.000 ± 0.000	<b>0.010 ± 0.000</b>	0.000 ± 0.000	<b>0.007 ± 0.012</b>	0.000 ± 0.000	<b>1.000 ± 0.000</b>	0.557 ± 0.368	1.000 ± 0.000	0.000 ± 0.000
	DTW <sub>s</sub> ( $\downarrow$ )	745.007 ± 52.333	810.542 ± 37.199	<b>737.831 ± 70.104</b>	932.027 ± 80.278	<b>726.706 ± 68.144</b>	<b>740.454 ± 73.628</b>	765.898 ± 61.689	556.756 ± 104.488	957.856 ± 64.304
	DTW <sub>a</sub> ( $\downarrow$ )	666.036 ± 16.540	688.794 ± 14.848	<b>666.253 ± 29.119</b>	950.579 ± 50.469	<b>665.579 ± 34.040</b>	722.633 ± 53.414	<b>664.746 ± 38.493</b>	727.791 ± 27.489	537.667 ± 10.401
Relocate	WD <sub>s</sub> ( $\downarrow$ )	8.365 ± 0.282	8.365 ± 0.289	<b>7.009 ± 0.249</b>	9.701 ± 0.246	<b>7.009 ± 0.247</b>	8.121 ± 0.246	<b>7.009 ± 0.247</b>	8.121 ± 0.246	5.085 ± 0.285
	WD <sub>a</sub> ( $\downarrow$ )	6.066 ± 0.695	6.166 ± 0.568	<b>6.049 ± 0.277</b>	8.212 ± 0.715	<b>5.970 ± 0.687</b>	7.401 ± 0.572	<b>6.051 ± 0.672</b>	5.868 ± 0.298	4.367 ± 0.336
	Success( $\uparrow$ )	0.397 ± 0.025	0.400 ± 0.053	<b>0.423 ± 0.065</b>	0.320 ± 0.085	<b>0.407 ± 0.085</b>	<b>0.617 ± 0.085</b>	0.417 ± 0.045	1.000 ± 0.000	1.000 ± 0.000
	DTW <sub>s</sub> ( $\downarrow$ )	654.866 ± 70.151	599.541 ± 102.762	<b>443.821 ± 31.627</b>	978.131 ± 102.574	<b>579.314 ± 95.599</b>	<b>686.462 ± 65.869</b>	<b>564.726 ± 69.751</b>	348.860 ± 25.274	201.109 ± 51.050
Relocate	DTW <sub>a</sub> ( $\downarrow$ )	999.567 ± 214.562	931.336 ± 152.154	<b>503.071 ± 18.131</b>	1786.058 ± 216.580	<b>731.311 ± 126.301</b>	1205.731 ± 182.452	<b>691.590 ± 132.459</b>	410.201 ± 29.469	257.894 ± 49.510
	WD <sub>s</sub> ( $\downarrow$ )	8.812 ± 1.036	8.158 ± 0.445	<b>7.389 ± 0.483</b>	12.312 ± 0.445	<b>7.903 ± 0.467</b>	10.574 ± 0.525	<b>8.048 ± 0.640</b>	5.960 ± 0.360	3.627 ± 0.150
	WD <sub>a</sub> ( $\downarrow$ )	6.937 ± 0.938	6.574 ± 0.319	<b>5.345 ± 0.274</b>	10.897 ± 0.254	<b>5.692 ± 0.245</b>	8.930 ± 0.237	<b>6.127 ± 0.615</b>	4.113 ± 0.212	2.559 ± 0.047
	Success( $\uparrow$ )	0.013 ± 0.015	0.000 ± 0.000	<b>0.203 ± 0.095</b>	0.000 ± 0.000	<b>0.143 ± 0.065</b>	0.137 ± 0.025	<b>0.173 ± 0.098</b>	1.000 ± 0.000	0.000 ± 0.000

Table 5: Comparison of computational costs with and without MAQ

	<b>Vanilla RL (ms)</b>	<b>MAQ (ms)</b>	<b>Ratio (MAQ/Vanilla RL)</b>
SAC/MAQ+SAC	1.31	1.35	1.03
IQL/MAQ+IQL	0.81	1.12	1.39
RLPD/MAQ+RLPD	0.28	0.81	2.93

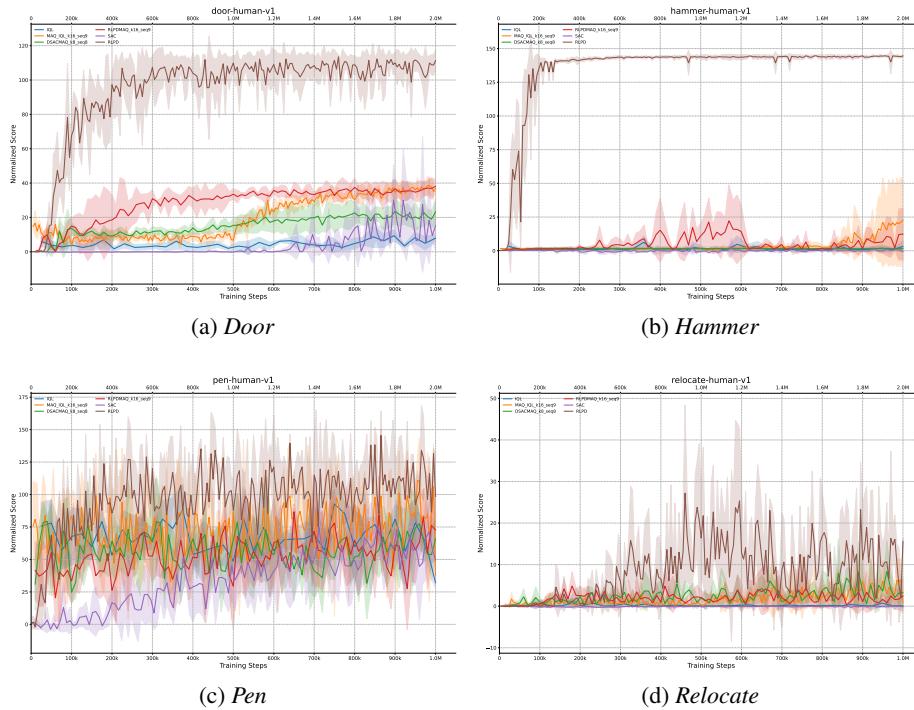


Figure 6: Training curves for Adroit tasks.

## B Further Analysis

In this section, we analyze MAQ from several aspects: (i) the effect of codebook size in Conditional-VQVAE (Appendix B.1); (ii) robustness to suboptimal demonstrations, including randomly shuffled and sorted datasets (Appendix B.2); (iii) the impact of extending the environment horizon on success rates in the *Hammer* task (Appendix B.3).

### B.1 Impact of Codebook Size on Similarity Score in MAQ

As discussed in Section 5.2.3, we previously observed that longer sequence lengths in VQVAE tend to lead to higher similarity scores when compared against human demonstrations. In this subsection, we further investigate whether varying the codebook size  $K$  exhibits a similar trend. We detail the models used in this experiment and analyze how different codebook sizes influence the similarity scores across MAQ variants, including MAQ+IQL, MAQ+SAC, and MAQ+RLPD. It is important to note that the similarity score is a metric we define to approximate the behavioral similarity between agents and human demonstrations. While it provides a quantitative means of comparison, it does not capture the full semantics or intent of human-like behavior. As also mentioned in Section 5.2.3, in D4RL control tasks, the action space only includes the positions of the Shadow Hand, while the state space contains additional information about the target object. Accordingly, we evaluate how different codebook sizes affect the similarity score for different MAQ agents. Due to the computational constraints, we limited our experiments to VQVAEs trained with codebook sizes of 8, 16, and 32.

As shown in Figures 7, 8, and 9, we observe a consistent trend across all MAQ variants: as the macro action length increases, the similarity score also improves, regardless of the codebook size. Moreover, the similarity scores remain within a narrow variance range, indicating stable performance across different configurations.

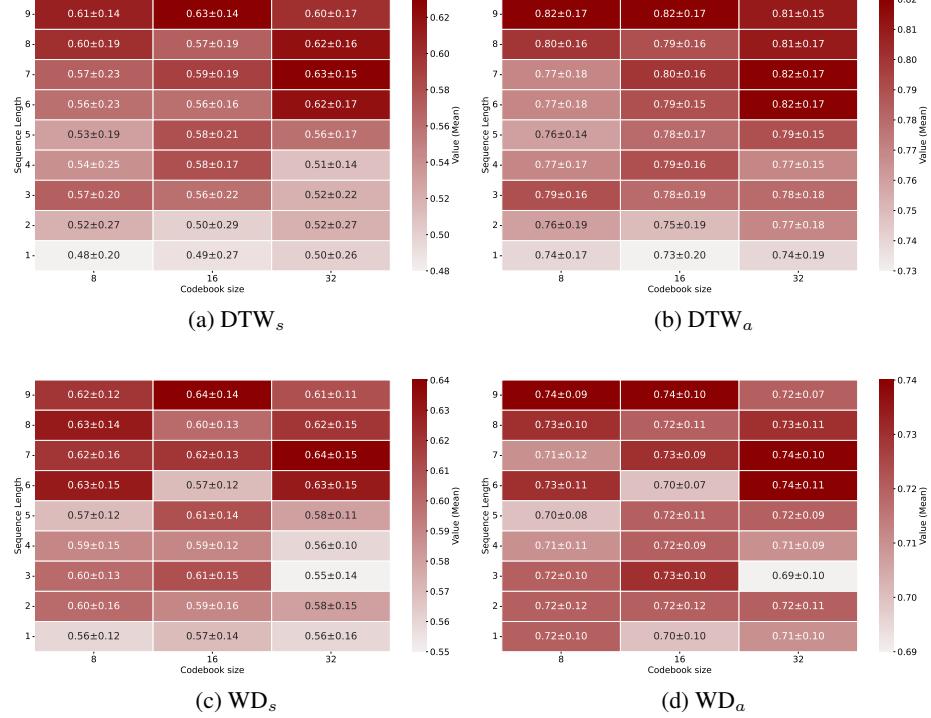


Figure 7: Similarity heatmaps of MAQ+IQL under different codebook sizes, measured by DTW and WD on state and action.

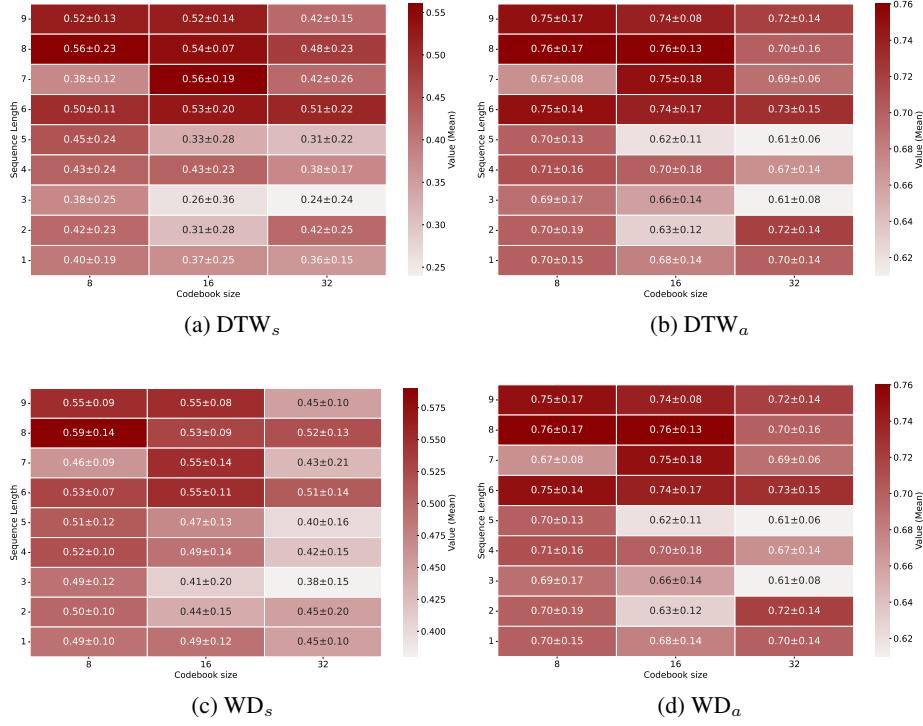


Figure 8: Similarity heatmaps of MAQ+SAC under different codebook sizes, measured by DTW and WD on state and action sequences.

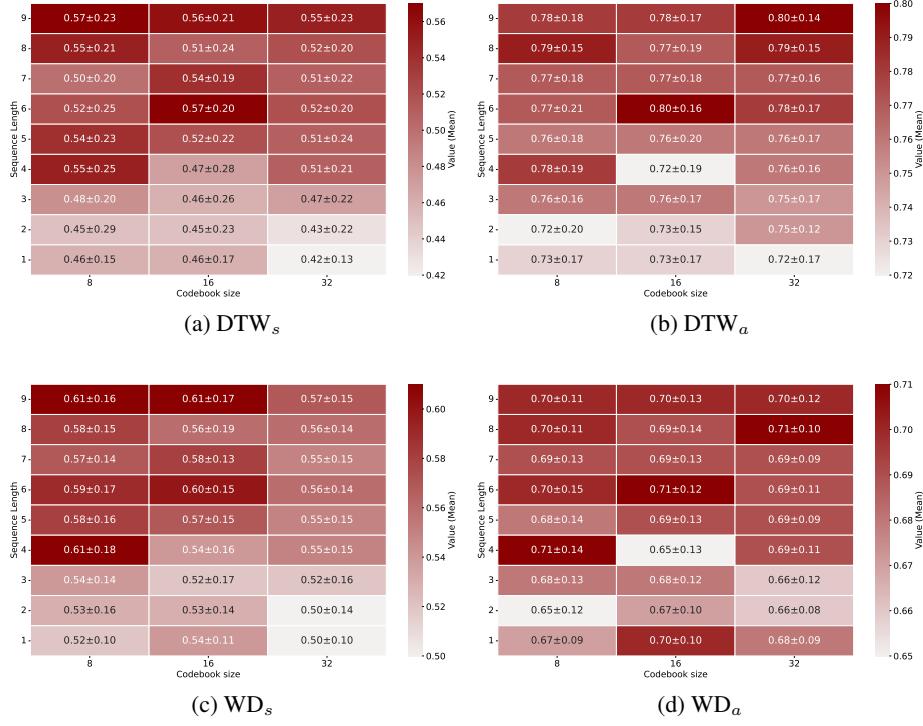


Figure 9: Similarity heatmaps of MAQ+RLPD under different codebook sizes, measured by DTW and WD on state and action.

In contrast, Figure 10 highlights how the effect of varying codebook sizes interacts differently with each underlying RL algorithm. For MAQ+IQL, which benefits from offline pretraining, success rates remain relatively consistent across different codebook sizes. MAQ+SAC, on the other hand, struggles under sparse reward conditions and fails to achieve a success rate above 30% regardless of codebook size or macro action length. Interestingly, MAQ+RLPD, which incorporates symmetric sampling, achieves a significantly higher success rate of up to 52%, demonstrating its advantage in such challenging environments.

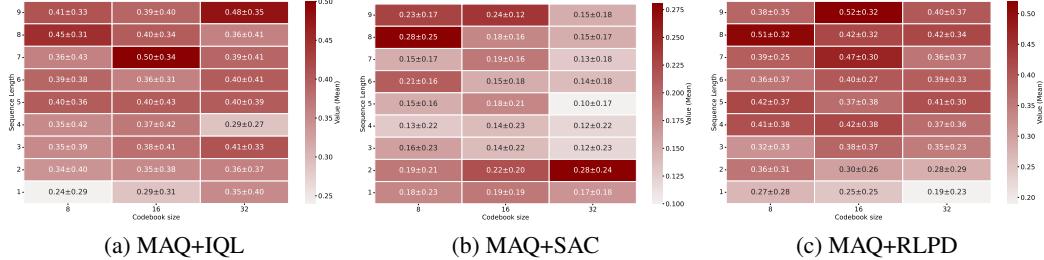


Figure 10: Similarity heatmaps of MAQ agents under different codebook sizes, measured by success rates.

## B.2 Impact of Using Suboptimal Demonstration

In this subsection, we evaluate MAQ under different demonstration qualities, including suboptimal ones. We conduct an ablation study using subsets of trajectories with the lowest task rewards as well as random subsets. The results are shown in Table 6.

In this experiment, using the Pen task as an example, the success rate remains stable across demonstrations of varying quality. Moreover, MAQ trained on small subsets of demonstrations (e.g., 25% of the dataset under random sampling or the bottom 25% in terms of trajectory-wise reward) can still achieve higher similarity scores (DTW and WD values in the table are not normalized; lower is better). Even when using different random subsets, which may include higher-return trajectories, MAQ maintains a competitive success rate.

Table 6: Ablation of MAQ trained on suboptimal demonstrations. Similarity scores remain stable across demonstration ratios in **random** and **lowest**, and MAQ+RLPD achieves the best success rate and similarity.

Tasks	MAQ+RLPD	75%	75%	50%	50%	25%	25%
		random	lowest	random	lowest	random	lowest
Door	DTW <sub>s</sub> (↓)	302.19	261.29	<b>239.77</b>	277.42	298.26	350.23
	DTW <sub>a</sub> (↓)	275.20	257.20	<b>242.36</b>	263.66	294.37	335.20
	WD <sub>s</sub> (↓)	4.59	<b>4.38</b>	4.70	4.42	4.77	5.19
	WD <sub>a</sub> (↓)	3.73	3.37	3.85	<b>3.27</b>	3.77	4.10
	Success(↑)	0.93	<b>0.98</b>	0.79	0.85	0.55	0.48
Hammer	DTW <sub>s</sub> (↓)	578.82	<b>593.07</b>	736.09	605.79	643.57	623.78
	DTW <sub>a</sub> (↓)	528.52	519.13	641.82	<b>515.70</b>	550.62	545.67
	WD <sub>s</sub> (↓)	5.20	5.37	5.54	<b>5.15</b>	5.26	5.48
	WD <sub>a</sub> (↓)	3.32	3.37	3.53	<b>3.21</b>	3.40	3.40
	Success(↑)	<b>0.56</b>	0.33	0.31	0.30	0.28	0.23
Pen	DTW <sub>s</sub> (↓)	740.45	679.96	<b>590.02</b>	669.14	649.56	686.26
	DTW <sub>a</sub> (↓)	664.75	629.19	<b>560.92</b>	624.25	607.43	654.41
	WD <sub>s</sub> (↓)	8.55	8.02	9.70	<b>7.79</b>	8.47	8.24
	WD <sub>a</sub> (↓)	6.05	5.95	8.42	<b>5.53</b>	6.01	5.85
	Success(↑)	<b>0.42</b>	0.33	0.32	0.33	0.23	0.20
Relocate	DTW <sub>s</sub> (↓)	<b>564.73</b>	582.06	596.79	641.00	625.72	652.59
	DTW <sub>a</sub> (↓)	<b>691.59</b>	739.38	741.85	857.08	772.72	862.77
	WD <sub>s</sub> (↓)	8.05	<b>7.84</b>	8.08	8.34	8.83	8.87
	WD <sub>a</sub> (↓)	6.13	6.11	<b>5.89</b>	6.54	6.47	6.60
	Success(↑)	<b>0.17</b>	0.08	0.04	0.04	0.02	0.00

### B.3 Impact of Different Environment Horizon

In Appendix A, we discussed the distinction between environment reward and success rate, as well as the score gap between RLPD and MAQ+RLPD in the training curves. In the *Hammer* task, the success rates of RLPD and MAQ+RLPD still differ markedly. We conducted a deeper analysis of this task. We found that in the 25 human demonstration trajectories, humans take an average of 451.4 steps to successfully complete the task. However, in the D4RL environment, the agent is restricted to a maximum of 200 steps per trajectory, after which the episode is terminated and directly considered a failure.

Since MAQ learns from human demonstrations, it tends to favor smooth but slow movements, while non-human-like RL agents tend to complete the task more quickly but less naturally. To validate this hypothesis, we increased the maximum episode length from 200 to 450 steps. As shown in Table 7, MAQ+RLPD’s success rate improves significantly from 0.56 to 0.72 while maintaining human-likeness scores.

Table 7: Ablation of MAQ+RLPD on *Hammer* under different training horizons  $h$ .

Tasks	RLPD ( $h = 200$ )	MAQ+RLPD ( $h = 200$ )	MAQ+RLPD ( $h = 450$ )
Hammer	DTW <sub>s</sub> (↓)	845.88	<b>578.82</b>
	DTW <sub>a</sub> (↓)	1178.74	<b>528.52</b>
	WD <sub>s</sub> (↓)	9.57	<b>5.20</b>
	WD <sub>a</sub> (↓)	7.25	<b>3.32</b>
	Success(↑)	1.00	<b>0.56</b>

## C Details of Human-likeness Survey

### C.1 Survey Setup and Evaluation Protocol

This subsection describes the interface shown to the evaluators. Each questionnaire contains four “tasks,” where each task is split into two stages, and each stage consists of a series of two-alternative forced-choice (2AFC) trials.

**Stage 1: Human-Detection Phase (Figure 11a).** The assessors are told that *at least one* of the two clips was produced by a human and must identify which one. They also mark their confidence on a five-point scale, but our main analysis considers only the binary choice and ignores the confidence scores.

**Stage 2: Ranking Phase (Figure 11b).** Here, we test whether the behavior generated by MAQ-based agents appears more human-like than that of baseline agents or even the human demonstrations. Model clips (optionally mixed with human reference clips) are presented in 2AFC pairs that are scheduled with a shuffled single-elimination and round-robin *mini-tournament*. Up to eight clips yield no more than 17 head-to-head comparisons; each win earns one point. Clips are ranked by win-rate  $w_i := \frac{\text{wins}_i}{\text{appearances}_i}$ , with mean reported confidence used only to break ties, producing an ordering from the least to the most human-like one.

**Post-ranking feedback.** After the Ranking Phase, each evaluator is shown the clips judged the *most* and the *least* human-like and may optionally explain those judgments in free text (Figure 11c). We analyze these comments in Section C.2.

### C.2 Additional Results of Human Feedback on Behavior Analysis

In this subsection, we discuss the additional behavioral results and the qualitative feedback received for each game. Figure 12 displays the distribution of the Top-1 and Top-2 rankings obtained in the Ranking Phase. Across all games, human demonstrations achieve the highest Top-1 probability, confirming that evaluators can reliably identify genuine human behavior. Among the learned agents, MAQ+IQL attains the second-highest Top-1 rate, followed by MAQ+RLPD, showing that our methods frequently convince evaluators that their behavior is indeed human-like.

In the *Hammer* task, five evaluators placed the human demonstration at Rank 1. The next most human-like agents were the MAQ variants—MAQ+RLPD and MAQ+IQL, each with four first-place votes, followed by MAQ+SAC with one first-place vote.

**What evaluators liked about MAQ policies.** Comments converge on three strengths: deliberate grip preparation, a realistic multi-strike rhythm, and a smooth follow-through.

“Is able to hammer the nail.”

“First takes hold of the hammer; because you need to aim, the first hit is lighter and the second harder.”

“Humans can aim accurately when hammering a nail and probably won’t drive the nail completely in at once.”

“It performs smoothly and hits the nail several times.”

“There is a back-and-forth hammer-swinging motion.”

“The feeling of driving it in on the last strike is very human-like.”

**Why baseline agents were judged less human-like.** Non-MAQ policies drew noticeably harsher remarks:

“Can’t even lift the hammer.”

“It threw the hammer.”

“It fails the task.”

“It cannot even lift the hammer.”

“Hits too far from the nail and releases the hammer in an unnatural manner.”

MAQ does more than replicating the gross kinematics of hammering: several evaluators remarked that the policy “re-aims and strikes again” after an initial impact, a behavior they naturally expect from humans. Figure 13 contrasts MAQ+RLPD with the plain RLPD agent. MAQ+RLPD delivers a series of well-timed blows that gradually seat the nail, whereas RLPD drives the nail in a single, overly forceful hit. Although MAQ+RLPD does not chase the maximum task score, it acts purposefully the way a person would, delivering several well-timed blows, while vanilla RLPD, trained only to maximize reward, drives the nail in a single, overly forceful hit that evaluators consistently judged “unlike a human.”

In the *Pen* task, six evaluators placed **MAQ+IQL** at Rank 1 more than any other policy, while vanilla RLPD received five first-place votes and the human demonstration received only two.

**What evaluators liked about MAQ policies.** Written feedback centers on a natural, well-coordinated grip and fine adjustments that keep the pen aligned with the target:

- “It is using all the fingers.”
- “It feels like all fingers are used.”
- “Humans would likely adjust the pen to be as close as possible to the target model.”
- “The middle and ring fingers move with the rest of the hand, which aligns with ergonomic principles.”
- “Holding the pen like this is very steady and human-like.”
- “The finger angles are not too strange and the task is completed quickly.”

**Why baseline agents were judged less human-like.** Agents trained without MAQ conditioning were criticized for awkward finger placement and lack of re-aiming:

- “It’s unnatural for a human to leave one finger open when trying to grasp.”
- “I feel like fingers are out of control.”
- “After holding the pen, I wouldn’t deliberately stick out a single finger.”
- “It does not adjust the pen.”
- “He basically shows no intention of aiming or aligning.”
- “People don’t hold a pen vertically.”
- “The angles of the fingers appear twisted.”

These remarks echo the ranking statistics: MAQ conditioning encourages full-hand coordination and incremental alignment, which are features that evaluators intuitively associate with human pen manipulation, whereas reward-only policies often adopt grasp patterns that look distinctly non-human.

In the *Relocate* task, six evaluators placed the human demonstration at Rank 1, but **MAQ+IQL** was a close second with five first-place votes, followed by MAQ+RLPD (three) and MAQ+SAC (one).

**What evaluators liked about MAQ policies.** Positive remarks emphasize three recurring traits: a direct, purposeful approach to the ball, continuous motion once the object is secured, and smooth point-to-point transfer.

- “It moves towards the ball.”
- “After the ball is grasped, the motion continues without any pause.”
- “Humans should smoothly move an object from the source to the destination point the hand is gently and smoothly picking, moving, and putting.”
- “At least it puts the ball near the target position.”
- “The hand shows slight grasping motions, and the movement trajectory feels quite natural.”
- “People normally go straight to grab the ball.”

**Why baseline agents were judged less human-like.** When an agent broke the direct-and-smooth pattern, the evaluators reacted strongly:

*“It moves away from the ball.”*

*“The hand stopped moving before it actually grasped the ball.”*

*“It looks like it can’t even pick up the ball.”*

*“It does not move the object at all.”*

*“The arm and finger movements are both very chaotic.”*

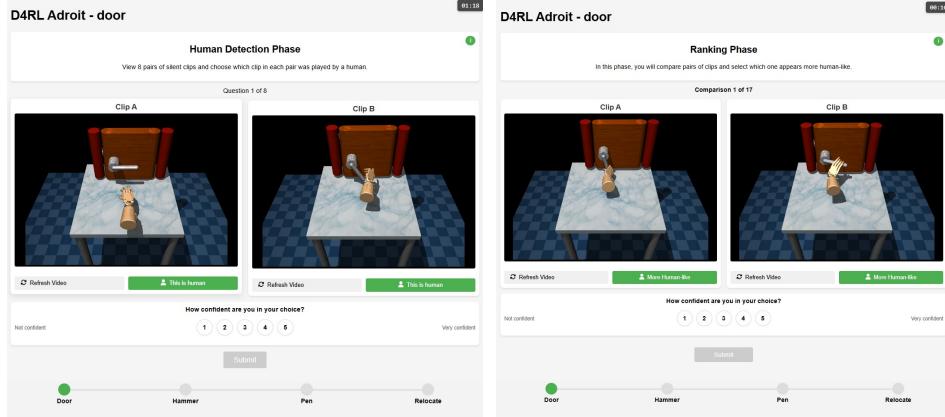
*“When I grab the ball, I would align my palm to the ball, not my wrist.”*

Taken together, these comments show that MAQ conditioning steers agents toward the straight-line reach, uninterrupted grasp-and-place sequence that evaluators intuitively associate with human relocation behavior, whereas reward-only policies often hesitate, wander, or execute awkward wrist-first contacts that look distinctly non-human.

**Summary of MAQ Advantages.** Across all appendix tasks, including *Hammer*, *Pen*, and *Relocate*, the MAQ-conditioned agents consistently receive the highest human-likeness rankings among the learned policies. Evaluators repeatedly highlight three qualities that MAQ alone imparts:

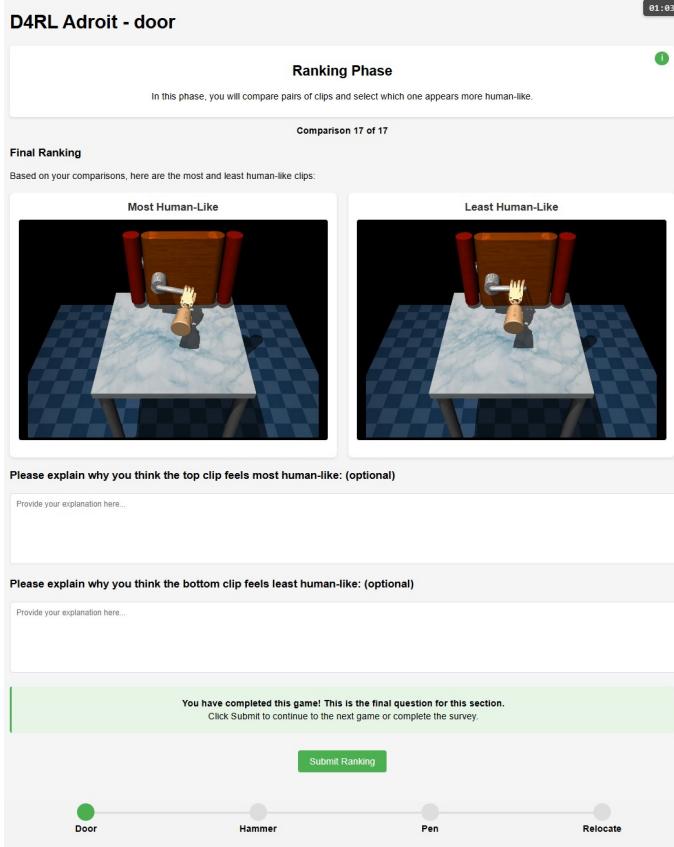
1. **Preparatory alignment.** MAQ policies re-aim, re-grip, or re-strike in ways that mirror how humans make minor adjustments before the decisive action.
2. **Appropriate force modulation.** MAQ agents hammer with several well-timed taps, guide the pen with gentle finger coordination, and transport the ball with a continuous grasp-and-place sequence—behaviors evaluators intuitively regard as natural.

These observations reinforce our central claim: MAQ does not simply optimize task scores; it systematically shapes behavior toward the timing, grip strategy, and motion fluidity that humans recognize as their own.



(a) Human Detection Phase.

(b) Ranking Phase.



(c) Post-ranking feedback.

Figure 11: Evaluation protocol.

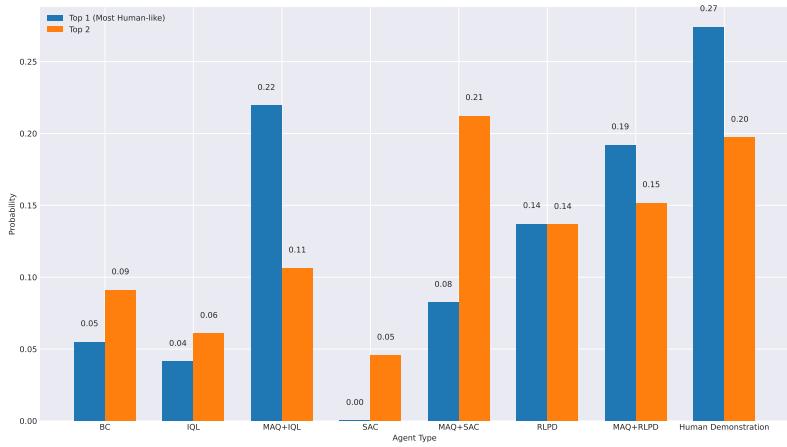


Figure 12: Probability distribution of agents ranked top 1 and top 2 in Ranking Phase.

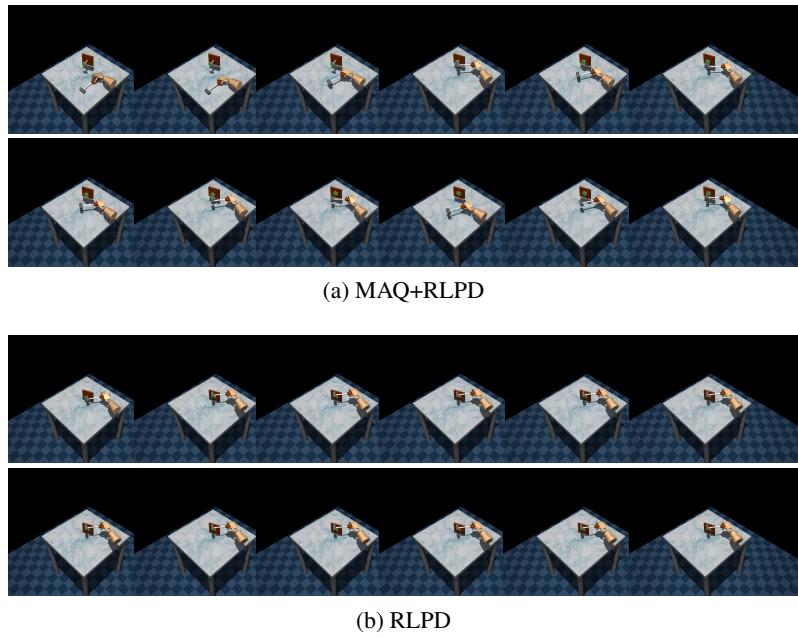


Figure 13: Agent behavior in *Hammer*.