

## A Experimental Settings and Detailed Results

**Experimental Settings** 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 approximately 1 GPU hour per run. Each MAQ method is built upon a pre-trained VQVAE. Therefore, all MAQ variants share the same VQVAE model when trained with the same random seed. We consistently selected the model from the final training generation as the VQVAE used in our online reinforcement learning (RL) experiments. 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. 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 4, and those for the 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, 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 released Discrete SAC [31] implementation 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 4).

**Dataset Segmentation** For data partitioning, we used the human dataset provided by D4RL. Each task consists of 25 human demonstrations, which we split into training and testing sets at a 9:1 ratio, resulting in 22 trajectories for training and 3 for testing. We used the testing set to evaluate similarity between agents. The normalization procedure mentioned in the paper involves generating a random agent and using its performance on the testing set to normalize the similarity scores. Table 2 presents all detailed results, including raw data, training data, and testing data.

**Training Curves for All Environments** Figure 6 presents the training curves for the experiments described in Subsection 5.2. The results are based on normalized rewards across different environments. For O2ORL methods such as IQL and MAQ+IQL, we represent training progress using a scale of 2M steps (with overall time scaled by 0.5 for comparability). Shaded regions indicate the standard deviation across three different random seeds.

Table 2: Detailed numerical results corresponding to the main table, including raw scores, training and testing data, and normalization baselines used for similarity evaluation.

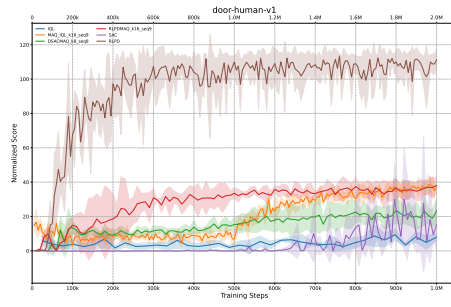
Tasks		BC	IQL	MAQ+IQL	SAC	MAQ+SAC	RLPD	MAQ+RLPD	Training Dataset	Testing Dataset	Random
Door	DTW <sub>a</sub>	564.738 ± 39.892	451.199 ± 27.652	266.994 ± 26.499	820.125 ± 47.292	283.977 ± 34.359	672.954 ± 19.864	302.186 ± 19.757	285.085 ± 21.432	193.165 ± 30.694	643.789 ± 23.380
	DTW <sub>s</sub>	700.093 ± 110.944	536.414 ± 33.763	237.304 ± 6.737	1285.045 ± 37.101	274.658 ± 24.079	823.852 ± 70.334	275.200 ± 41.988	299.011 ± 11.922	192.035 ± 12.539	1068.101 ± 26.872
	WD <sub>a</sub>	6.721 ± 0.261	5.864 ± 0.263	4.388 ± 0.258	9.988 ± 0.127	4.561 ± 0.455	9.230 ± 0.213	4.587 ± 0.189	4.307 ± 0.134	3.013 ± 0.322	8.446 ± 0.188
	WD <sub>s</sub>	6.017 ± 0.531	5.429 ± 0.109	3.446 ± 0.212	9.697 ± 0.162	3.701 ± 0.437	8.076 ± 0.116	3.734 ± 0.201	3.304 ± 0.118	2.176 ± 0.046	8.738 ± 0.093
	Success	0.020 ± 0.010	0.163 ± 0.055	0.930 ± 0.040	0.433 ± 0.232	0.563 ± 0.497	0.957 ± 0.067	0.933 ± 0.049	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000
Hammer	DTW <sub>a</sub>	894.540 ± 24.821	887.642 ± 129.362	595.441 ± 63.226	1246.982 ± 131.710	607.021 ± 79.326	845.877 ± 52.838	578.816 ± 62.884	642.887 ± 48.775	459.715 ± 96.337	834.893 ± 45.341
	DTW <sub>s</sub>	1056.832 ± 19.802	1083.427 ± 223.472	551.378 ± 72.545	1959.679 ± 118.135	570.395 ± 111.008	1178.736 ± 91.087	528.520 ± 81.925	654.830 ± 59.219	465.388 ± 104.728	1588.783 ± 81.516
	WD <sub>a</sub>	8.792 ± 0.335	8.718 ± 0.696	5.243 ± 0.176	11.836 ± 0.509	5.893 ± 0.649	9.571 ± 0.443	5.202 ± 0.207	5.457 ± 0.166	3.892 ± 0.292	9.393 ± 0.251
	WD <sub>s</sub>	6.670 ± 0.206	6.667 ± 0.586	3.371 ± 0.093	9.661 ± 0.234	3.751 ± 0.686	7.252 ± 0.185	3.320 ± 0.209	3.390 ± 0.077	2.394 ± 0.254	8.469 ± 0.105
	Success	0.000 ± 0.000	0.010 ± 0.010	0.000 ± 0.000	0.007 ± 0.012	0.000 ± 0.000	1.000 ± 0.000	0.557 ± 0.368	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000
Pen	DTW <sub>a</sub>	745.007 ± 52.333	819.542 ± 37.199	737.831 ± 70.104	932.027 ± 80.278	726.706 ± 68.144	767.026 ± 94.645	740.454 ± 73.628	765.896 ± 61.689	556.756 ± 104.488	957.856 ± 64.804
	DTW <sub>s</sub>	666.036 ± 16.540	688.794 ± 14.848	666.253 ± 29.119	950.579 ± 50.449	665.579 ± 34.040	722.633 ± 53.414	664.746 ± 38.493	727.791 ± 27.489	537.667 ± 62.392	845.175 ± 10.401
	WD <sub>a</sub>	8.555 ± 0.782	8.865 ± 0.689	8.559 ± 0.793	10.498 ± 0.616	8.461 ± 0.737	9.211 ± 0.491	8.550 ± 0.736	8.112 ± 0.360	5.965 ± 0.285	12.305 ± 1.054
	WD <sub>s</sub>	6.066 ± 0.695	6.166 ± 0.568	6.040 ± 0.727	8.212 ± 0.715	5.970 ± 0.687	7.101 ± 0.572	6.051 ± 0.672	5.868 ± 0.298	4.367 ± 0.336	9.275 ± 1.150
	Success	0.397 ± 0.025	0.400 ± 0.053	0.423 ± 0.065	0.320 ± 0.085	0.407 ± 0.006	0.617 ± 0.085	0.417 ± 0.045	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000
Relocate	DTW <sub>a</sub>	654.866 ± 70.151	599.541 ± 102.262	443.821 ± 31.627	978.131 ± 102.574	579.314 ± 93.509	686.462 ± 65.869	564.726 ± 69.751	348.860 ± 25.274	201.109 ± 51.059	702.083 ± 91.180
	DTW <sub>s</sub>	999.567 ± 214.562	931.336 ± 152.154	503.071 ± 18.131	1786.058 ± 216.580	731.311 ± 128.301	1205.731 ± 182.452	691.590 ± 132.459	410.301 ± 29.469	257.894 ± 49.510	1646.299 ± 229.892
	WD <sub>a</sub>	8.812 ± 1.036	8.158 ± 0.445	7.389 ± 0.483	12.312 ± 0.445	7.903 ± 0.467	10.574 ± 0.525	8.048 ± 0.640	5.960 ± 0.360	3.627 ± 0.150	10.730 ± 0.308
	WD <sub>s</sub>	6.937 ± 0.938	6.574 ± 0.319	5.345 ± 0.274	10.897 ± 0.254	5.692 ± 0.245	8.930 ± 0.237	6.127 ± 0.615	4.113 ± 0.212	2.559 ± 0.047	10.516 ± 0.329
	Success	0.013 ± 0.015	0.000 ± 0.000	0.203 ± 0.095	0.000 ± 0.000	0.143 ± 0.065	0.137 ± 0.025	0.173 ± 0.098	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000

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

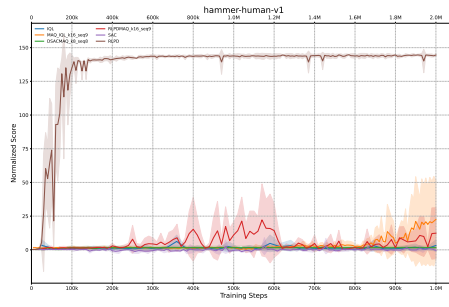
Parameter	Adroit
Latent size	256
Learning rate	3e-4
Codebook size $K$	[8, 16, 32]
Batch size	32
Commitment loss coeff $\beta$	0.25
Macro length $H$	[1 ... 9]
Optimizer	Adam

Table 4: Hyperparameters for training Adroit tasks.

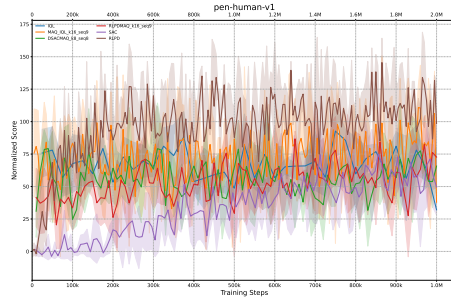
Parameter	BC	IQL	MAQ+IQL	SAC	MAQ+SAC	RLPD	MAQ+RLPD
Batch size	256	256	256	128	128	256	128
Learning rate	3e-4	3e-4	3e-4	3e-4	3e-4	-	-
Actor learning rate	-	-	-	-	-	3e-4	3e-4
Critic learning rate	-	-	-	-	-	3e-4	1e-3
Temperature learning rate	-	-	-	-	-	3e-4	3e-4
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam	Adam
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



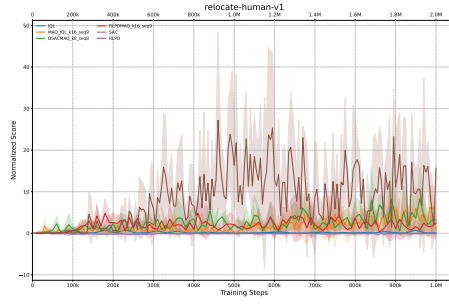
(a) *Door*



(b) *Hammer*



(c) *Pen*

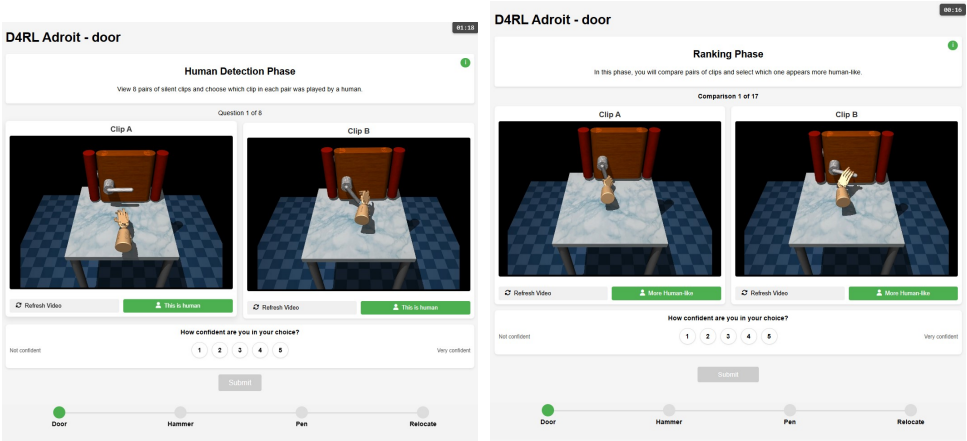


(d) *Relocate*

Figure 6: Training curves for Adroit tasks.

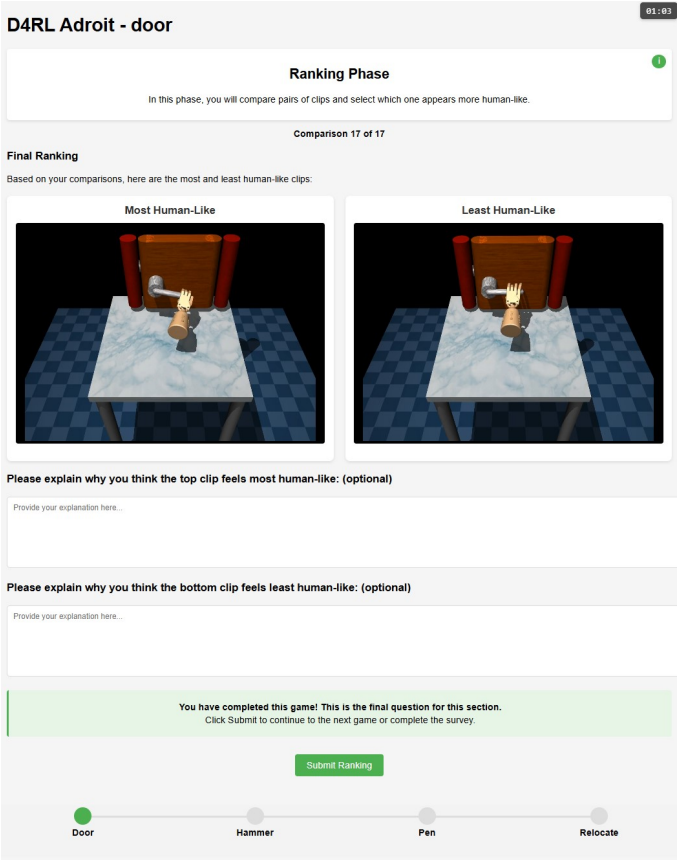
831 **B Human-likeness Survey in RL**

832 **B.1 Survey Setup and Evaluation Protocol**



(a) Human Detection Phase

(b) Ranking Phase



(c) Post-ranking feedback

Figure 7: Evaluation Protocol.

833 This section describes the interface shown to evaluators. Each questionnaire contains four “games,”  
834 every game split into two stages, and both stages consist of a series of two-alternative forced-choice  
835 (2AFC) trials.

836 **Stage 1: Human-Detection Phase (Figure 7a).** Evaluators are told that *at least one* of the two  
837 clips was produced by a human and must identify which one. They also mark their confidence on a  
838 five-point scale, but our main analysis considers only the binary choice and ignores the confidence  
839 scores.

840 **Stage 2: Ranking Phase (Figure 7b).** Here we test whether behavior generated by MAQ-based  
841 agents appears more human-like than that of baseline agents or even the human demonstrations.  
842 Model clips (optionally mixed with human reference clips) are presented in 2AFC pairs that are  
843 scheduled with a shuffled single-elimination and round-robin *mini-tournament*. Up to eight clips yield  
844 no more than 17 head-to-head comparisons; each win earns one point. Clips are ranked by win-rate  
845  $w_i = \frac{\text{wins}_i}{\text{appearances}_i}$ , with mean reported confidence used only to break ties, producing an ordering from  
846 least to most human-like.

847 **Post-ranking feedback.** After the Ranking Phase, each evaluator is shown the clips judged *most*  
848 and *least* human-like and may optionally explain those judgments in free text (Figure 7c). We analyze  
849 these comments in Section B.2.

## 850 B.2 More Results For Human Feedback on Behavior Analysis

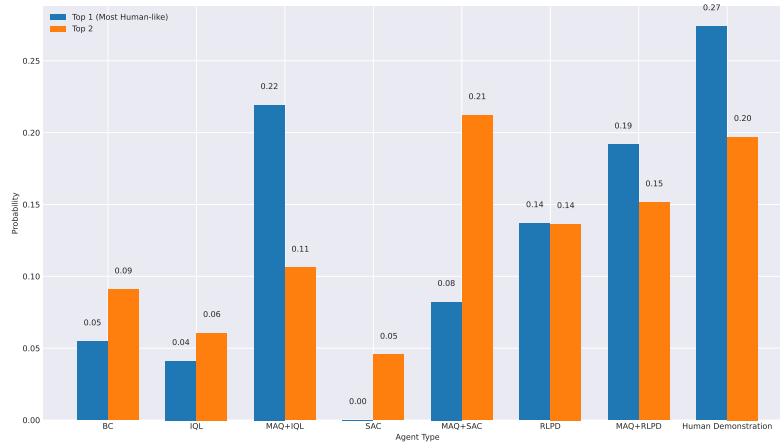


Figure 8: Probability Distribution of Agents Ranked Top 1 and Top 2 in Ranking Phase

851 In this section we discuss additional behavioral results and the qualitative feedback received for  
852 each game. Figure 8 displays the distribution of Top-1 and Top-2 rankings obtained in the Ranking  
853 Phase. Across all games, human demonstrations achieve the highest Top-1 probability, confirming  
854 that evaluators can reliably identify genuine human behavior. Among the learned agents, MAQ+IQL  
855 attains the second-highest Top-1 rate, followed by MAQ + RLPD, showing that our methods frequently  
856 convince evaluators that their behavior is human-like.

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

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

862 “Is able to hammer the nail.”

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

865 *“Humans can aim accurately when hammering a nail and probably won’t drive the*  
 866 *nail completely in at once.”*  
 867 *“It performs smoothly and hits the nail several times.”*  
 868 *“There is a back-and-forth hammer-swinging motion.”*  
 869 *“The feeling of driving it in on the last strike is very human-like.”*

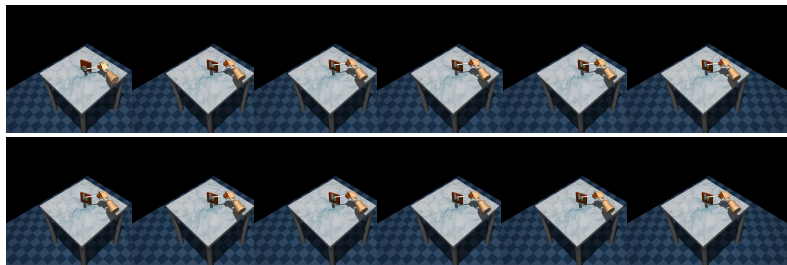
870 **Why baseline agents were judged less human-like.** Non-MAQ policies drew noticeably harsher  
 871 remarks:

872 *“Can’t even lift the hammer.”*  
 873 *“It threw the hammer.”*  
 874 *“It fails the task.”*  
 875 *“It cannot even lift the hammer.”*  
 876 *“Hits too far from the nail and releases the hammer in an unnatural manner.”*

877 MAQ does more than replicate the gross kinematics of hammering: several evaluators remarked that  
 878 the policy “re-aims and strikes again” after an initial impact, a behavior they naturally expect from  
 879 humans. Figure 9 contrasts MAQ+RLPD with the plain RLPD agent. MAQ+RLPD delivers a series  
 880 of well-timed blows that gradually seat the nail, whereas RLPD drives the nail in a single, overly  
 881 forceful hit. Although MAQ+RLPD does not chase the maximum task score, it purposefully acts the  
 882 way a person would, delivering several well-timed blows—whereas vanilla RLPD, trained only to  
 883 maximize reward, drives the nail in a single overly forceful hit that evaluators consistently judged  
 884 “unlike a human.”



(a) MAQ+RLPD



(b) RLPD

Figure 9: Agent behavior in *Hammer*.

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

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

889 *“It is using all the fingers.”*  
 890 *“It feels like all fingers are used.”*

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

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

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

905 These remarks echo the ranking statistics: MAQ conditioning encourages full-hand coordination  
 906 and incremental alignment—features that evaluators intuitively associate with human pen manipula-  
 907 tion—whereas reward-only policies often adopt grasp patterns that look distinctly non-human.

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

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

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

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

923 *“It moves away from the ball.”*  
 924 *“The hand stopped moving before it actually grasped the ball.”*  
 925 *“It looks like it can’t even pick up the ball.”*  
 926 *“It does not move the object at all.”*  
 927 *“The arm and finger movements are both very chaotic.”*  
 928 *“When I grab the ball, I would align my palm to the ball, not my wrist.”*

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

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

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

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

## 944 C Impact of Codebook Size on Similarity Score in MAQ

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

957 As shown in Figures 10, 11, and 12, we observe a consistent trend across all MAQ variants: as the  
958 macro action length increases, the similarity score also improves, regardless of the codebook size.  
959 Moreover, the similarity scores remain within a narrow variance range, indicating stable performance  
960 across different configurations.

961 In contrast, Figure 13 highlights how the effect of varying codebook sizes interacts differently with  
962 each underlying RL algorithm. For MAQ+IQL, which benefits from offline pretraining, success  
963 rates remain relatively consistent across different codebook sizes. MAQ+SAC, on the other hand,  
964 struggles under sparse reward conditions and fails to achieve a success rate above 30% regardless of  
965 the codebook size or macro action length. Interestingly, MAQ+RLPD, which incorporates symmetric  
966 sampling, achieves a significantly higher success rate of up to 52%, demonstrating its advantage in  
967 such challenging environments.



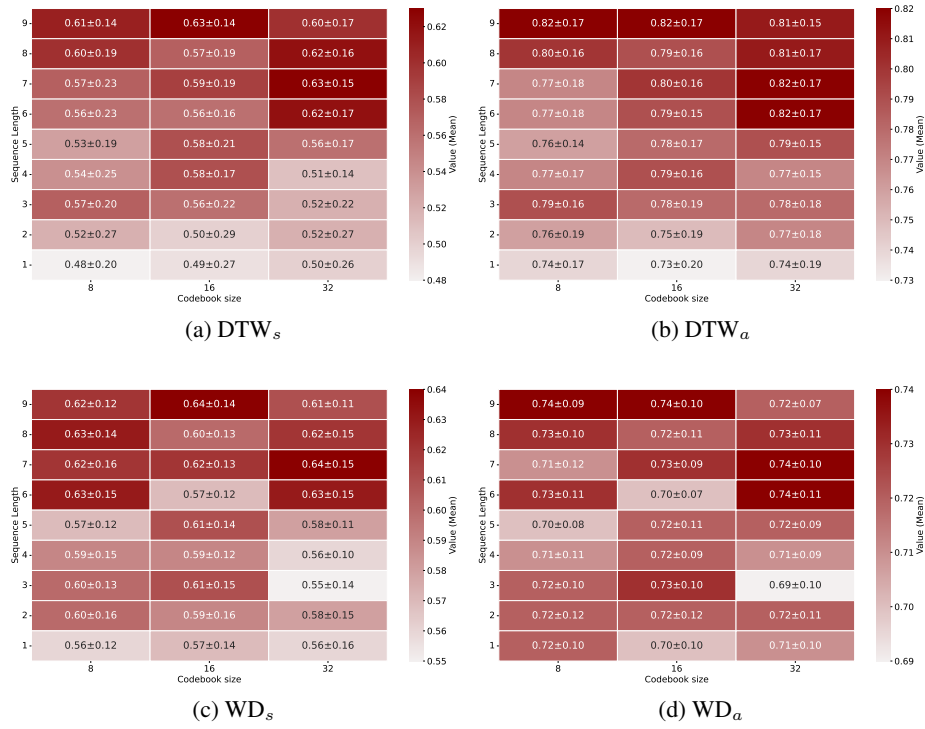


Figure 10: Similarity heatmaps of MAQ+IQL under different codebook sizes, measured by DTW and WD on state and action.

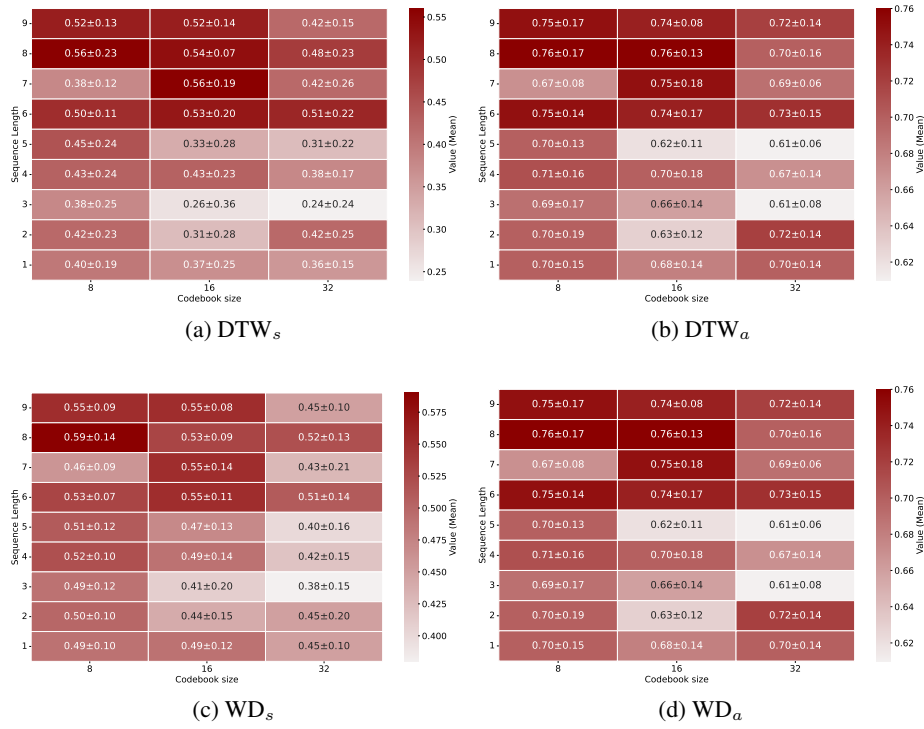


Figure 11: Similarity heatmaps of MAQ+SAC under different codebook sizes, measured by DTW and WD on state and action sequences.

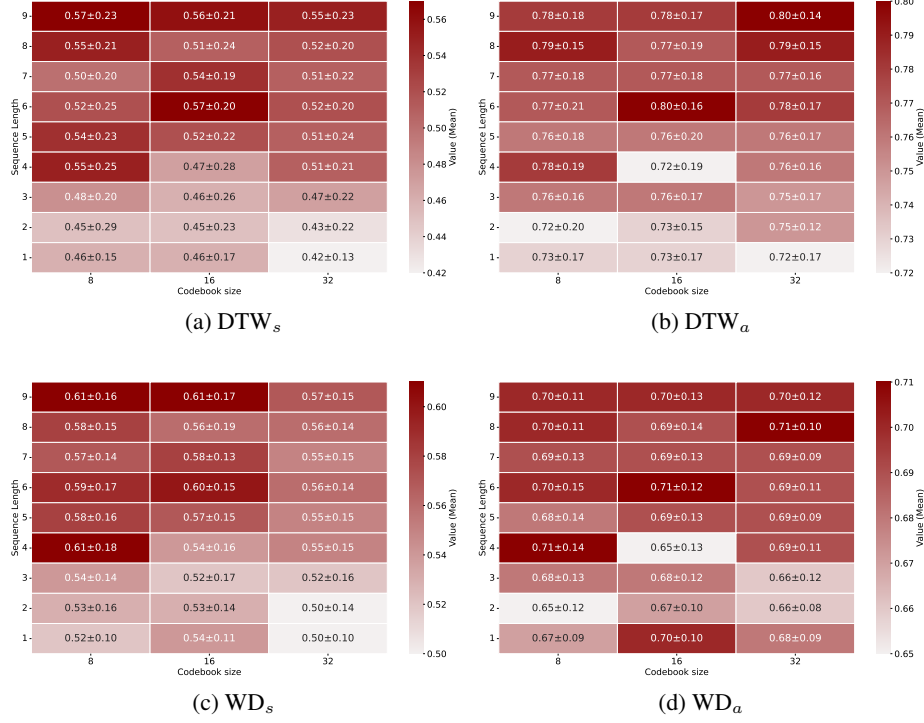


Figure 12: Similarity heatmaps of MAQ+RLPD under different codebook sizes, measured by DTW and WD on state and action.

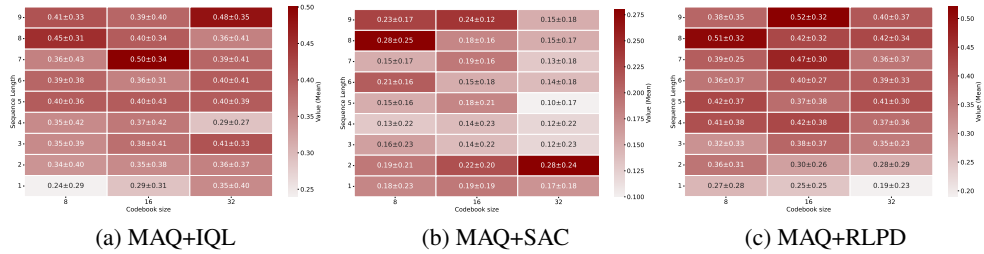


Figure 13: Similarity heatmaps of MAQ agents under different codebook sizes, measured by success rates.