NEURAL INFORMATION PROCESSING SYSTEMS

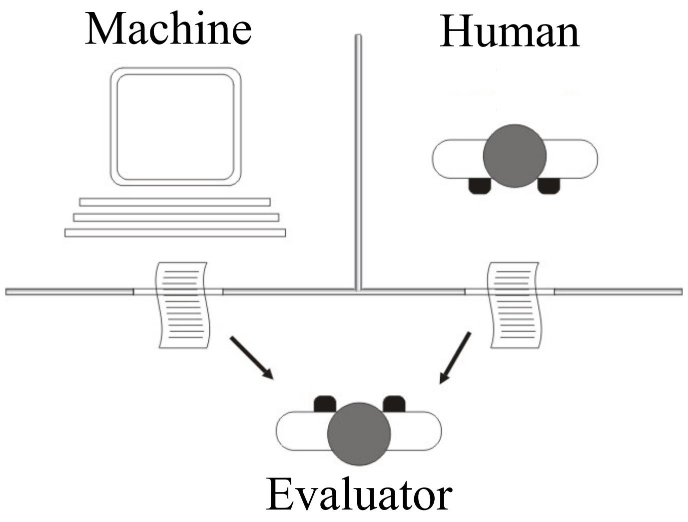
Learning Human-Like RL Agents Through Trajectory Optimization With Action Quantization

Jian-Ting Guo[1], Yu-Cheng Chen[1], Ping-Chun Hsieh[1], Kuo-Hao Ho[1]
Po-Wei Huang[1], Ti-Rong Wu[2], I-Chen Wu[1,2]

[1] National Yang Ming Chiao Tung University
[2] Academia Sinica

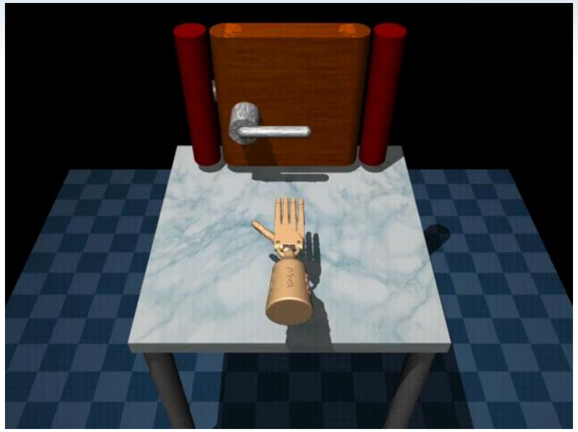
Turing test

- The *Turing Test* is one of the most well-known benchmarks to evaluate 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
- However, most of these benchmarks are in text, not in RL



Human-Like Reinforcement Learning

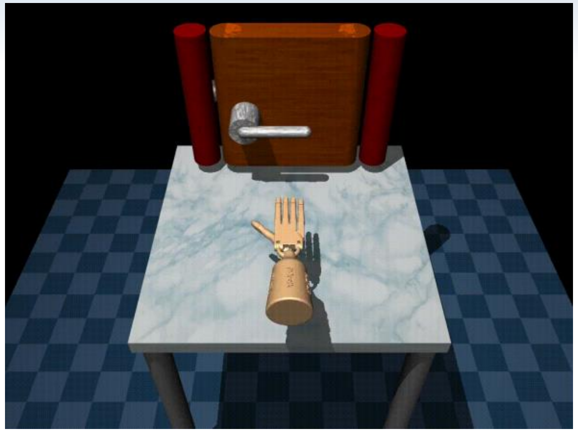
- Most RL research focuses on designing reward-driven agents
  - The behavior is unnatural, not human-like



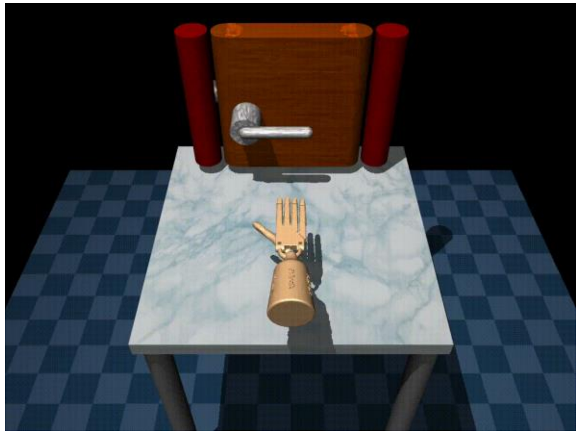
Reward-driven RL agents

Human-Like Reinforcement Learning

- Most RL research focuses on designing reward-driven agents
  - The behavior is unnatural, not human-like
- *Human-Like Reinforcement Learning* seeks both human-like behavior and optimal performance
  - Remains underexplored in the RL community
  - Most of the previous methods rely on pre-defined behavior constraints or rule-based penalties



Reward-driven RL agents



Human-like RL agents (Ours)

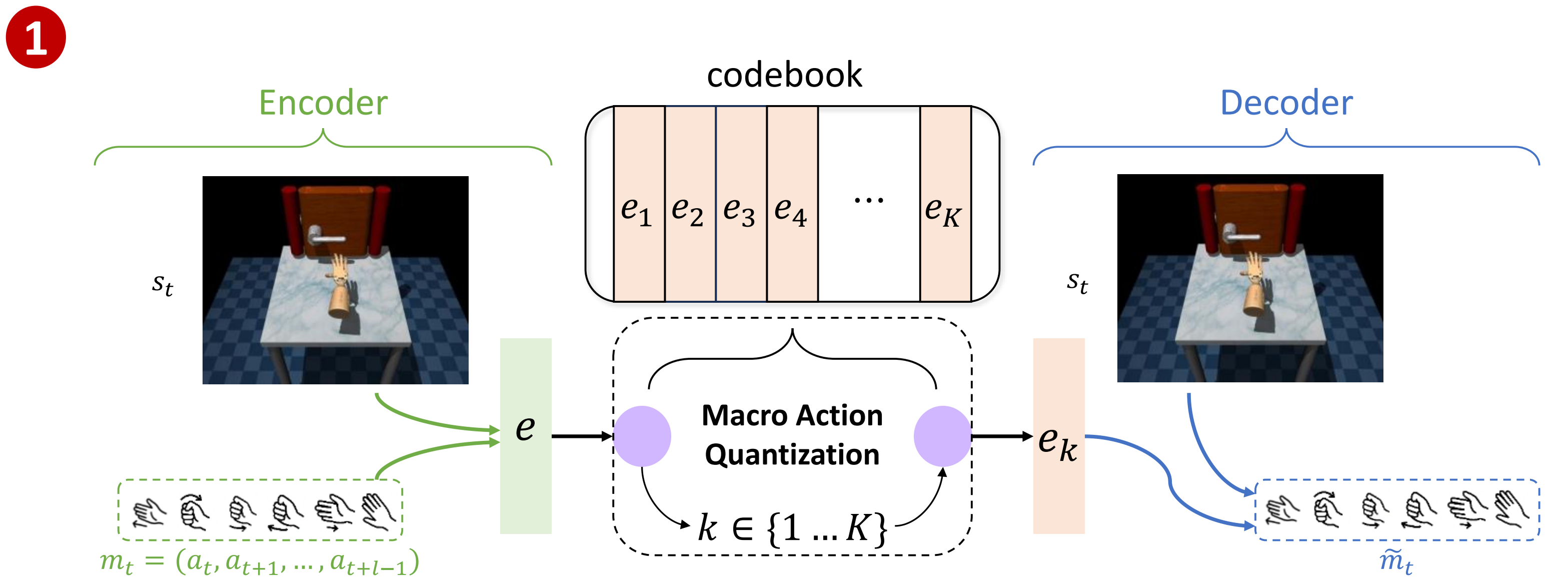
Human-likeness

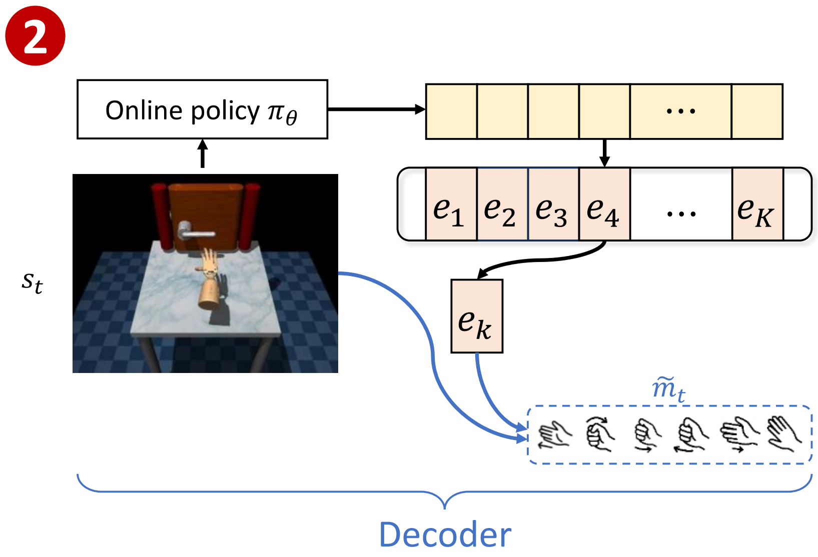
- We introduce the concept of *human-like sequence*, which is a set of human-generated action sequences
  - Human-likeness can be enforced by constraining the search space of action sequences
  - The longer the action lengths are, the more human-like it is

- To achieve the human-like sequence:
  - Distill human behavior into ***macro action*** – sequences of actions – from human demonstrations
  - Then, select macro actions to interact with the environment at each timestep

Macro Action Quantization (MAQ)

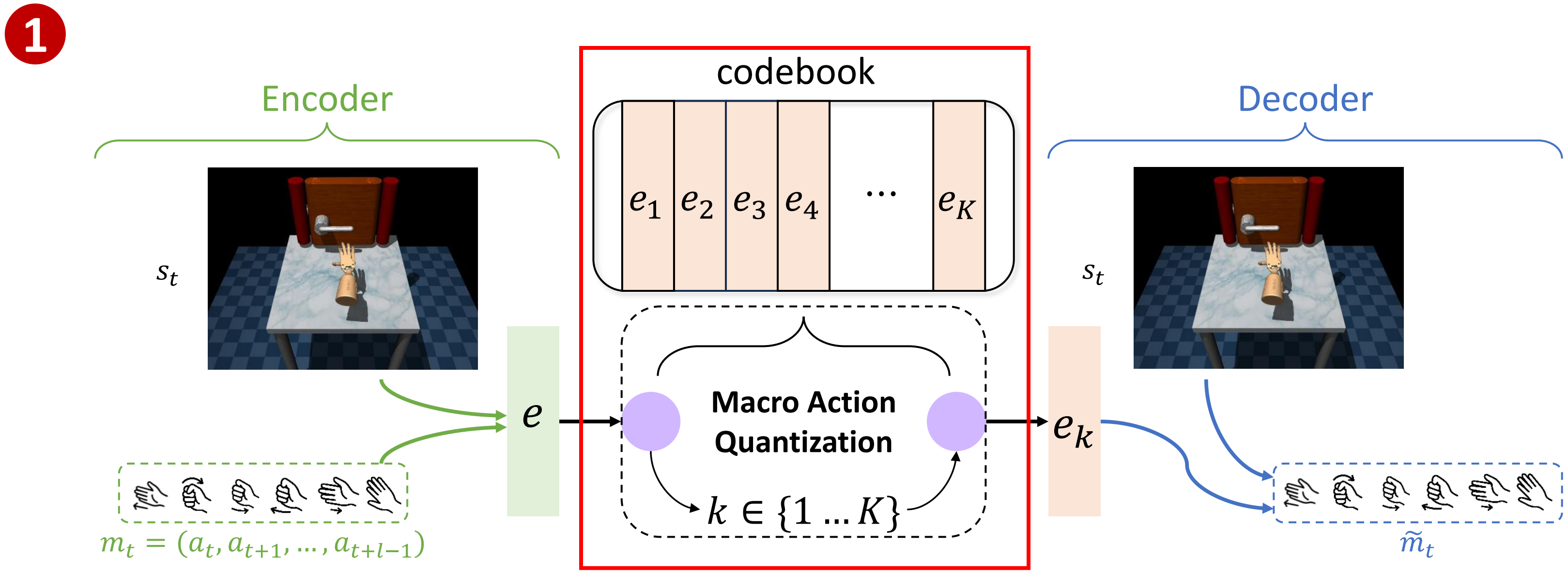
- MAQ consists of two stages
  - ❶ Human behavior distillation
  - ❷ Reinforcement learning with macro actions





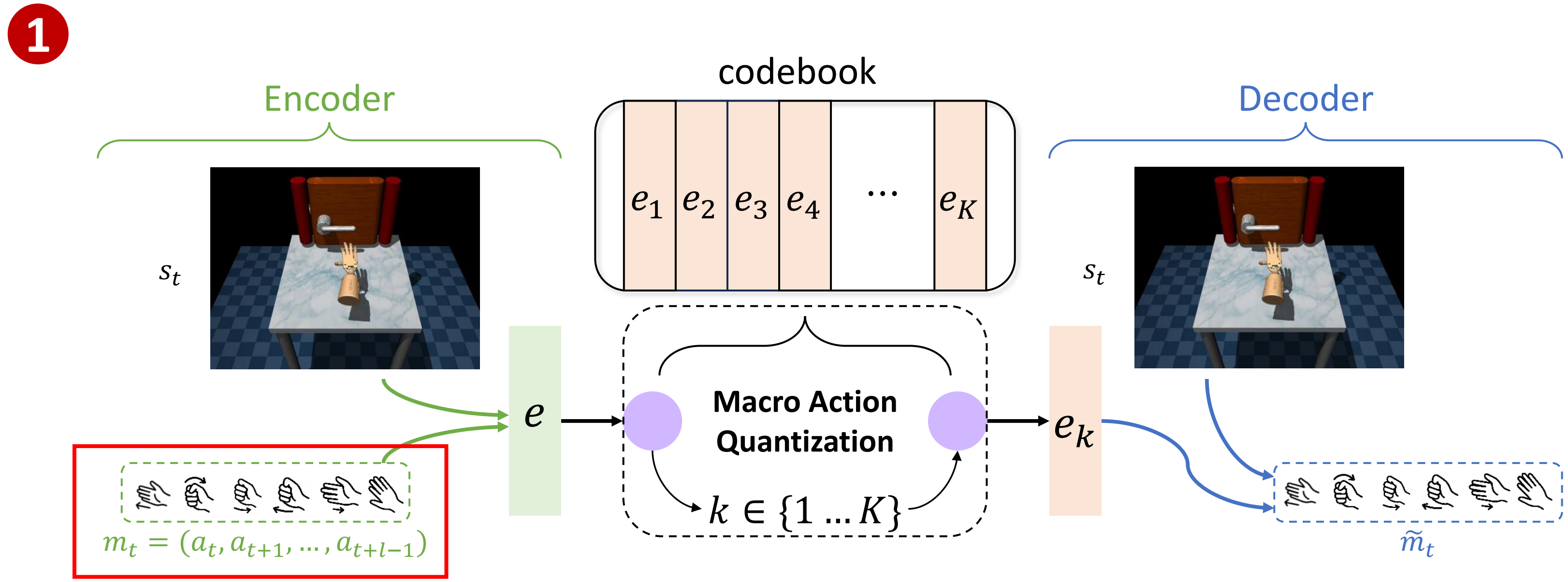
Macro Action Quantization (MAQ)

- Human behavior distillation
  - We train a Conditional-VQVAE to distill **macro action** from human demonstration to learn a discrete codebook



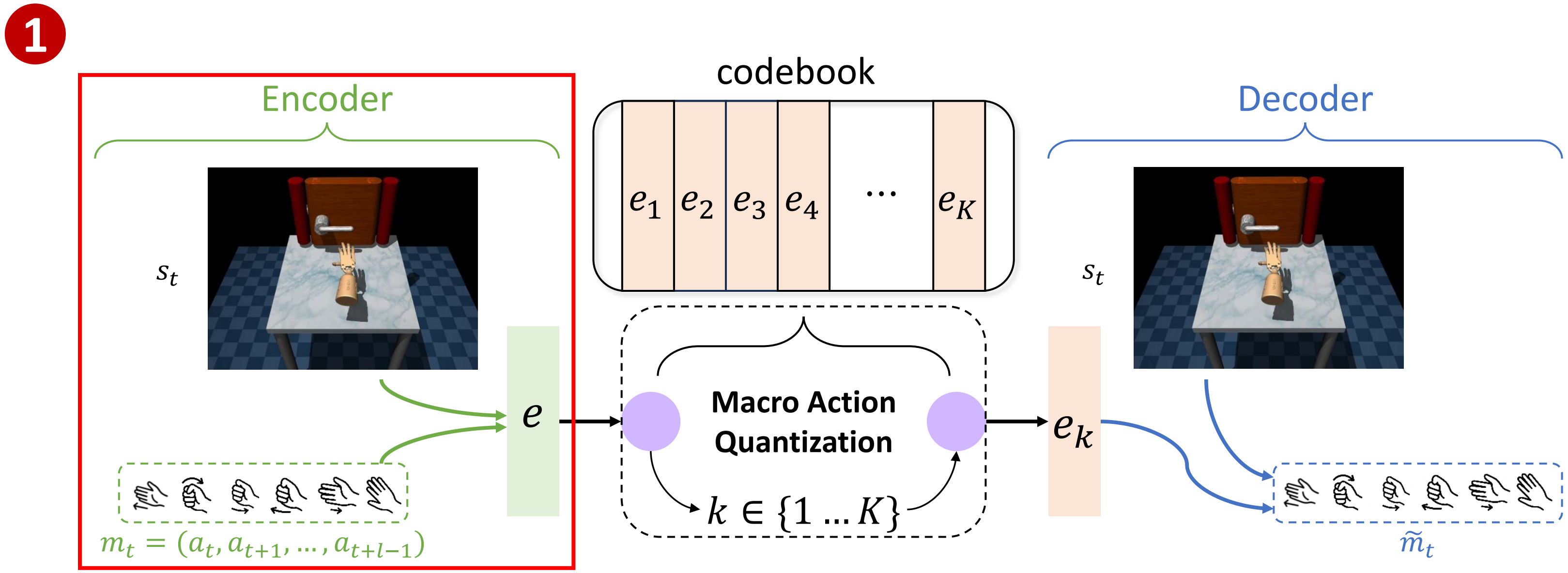
Macro Action Quantization (MAQ)

- Human behavior distillation
  - We train a Conditional-VQVAE to distill **macro action** from human demonstration to learn a discrete codebook



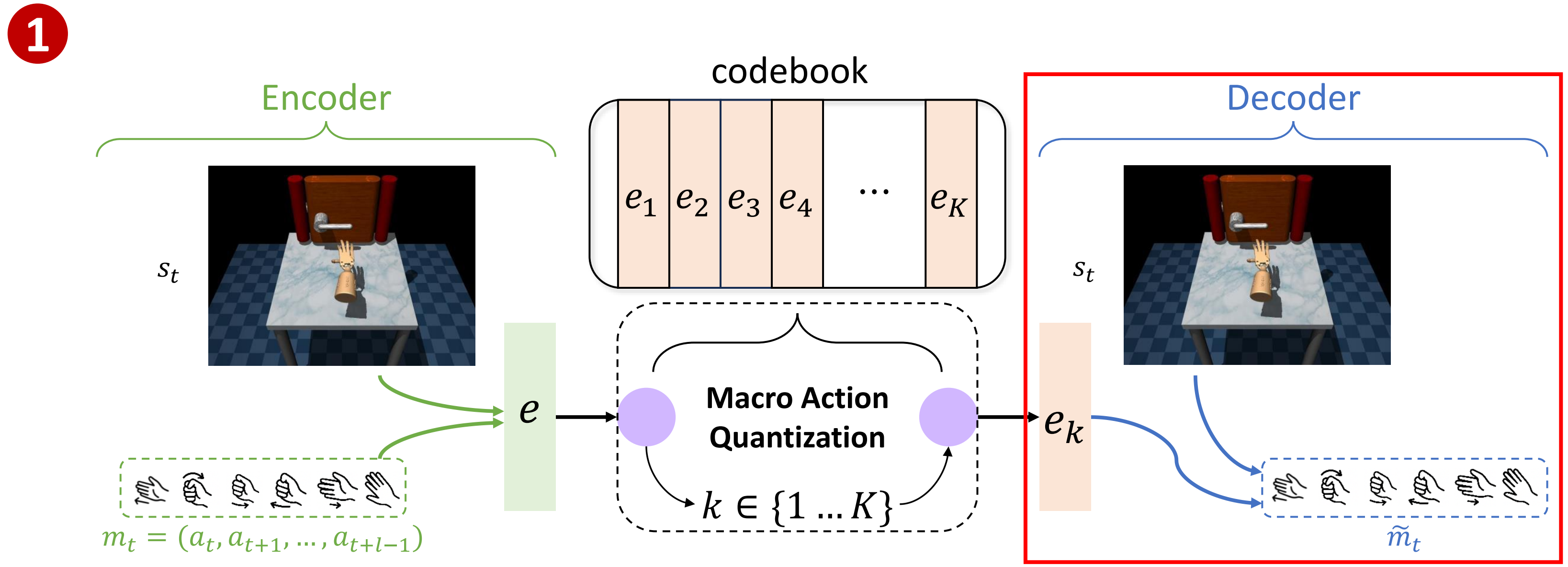
Macro Action Quantization (MAQ)

- Human behavior distillation
  - We train a Conditional-VQVAE to distill **macro action** from human demonstration to learn a discrete codebook



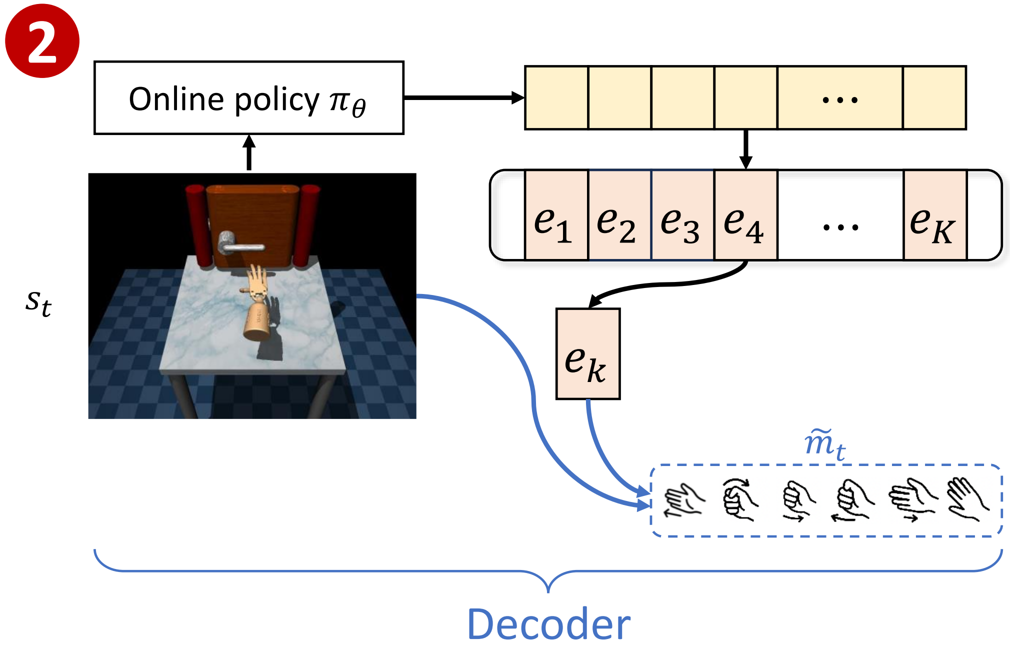
Macro Action Quantization (MAQ)

- Human behavior distillation
  - We train a Conditional-VQVAE to distill **macro action** from human demonstration to learn a discrete codebook



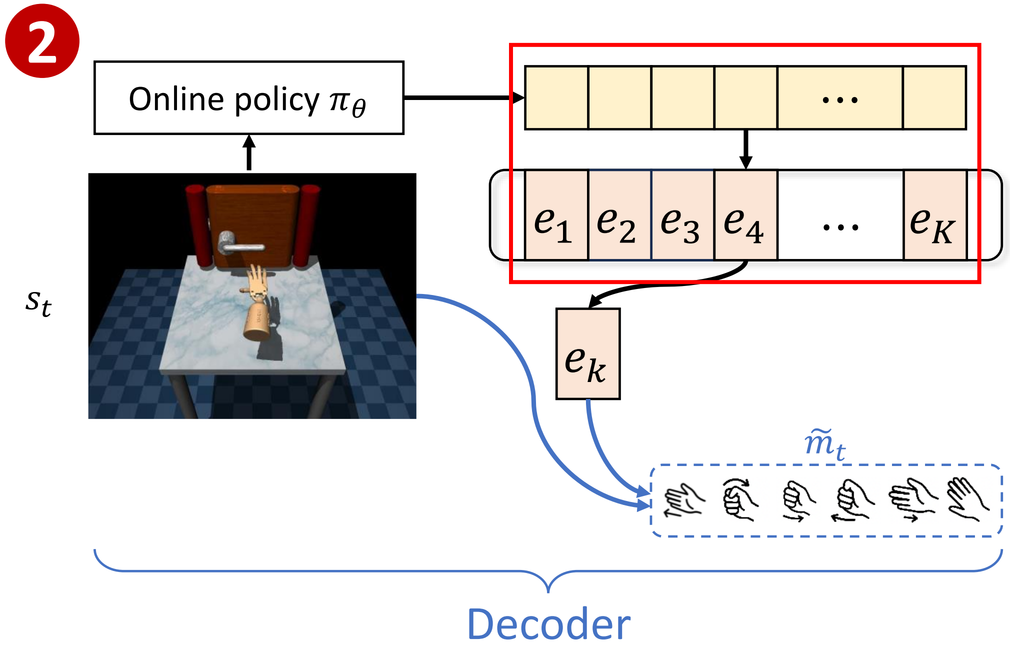
Macro Action Quantization (MAQ)

- Reinforcement learning with Macro Actions
  - We train an online policy (π_θ) that acts in the **learned discrete code space** by selecting **codebook indices**



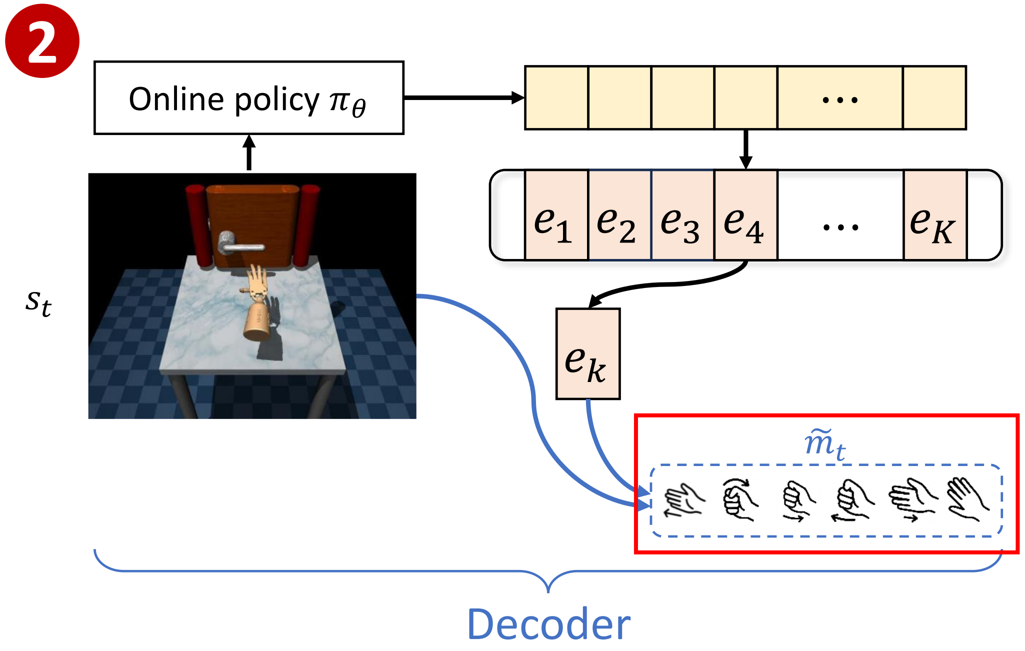
Macro Action Quantization (MAQ)

- Reinforcement learning with Macro Actions
  - We train an online policy (π_θ) that acts in the **learned discrete code space** by selecting **codebook indices**



Macro Action Quantization (MAQ)

- Reinforcement learning with Macro Actions
  - We train an online policy (π_θ) that acts in the **learned discrete code space** by selecting **codebook indices**



Experiments

- We train three off-the-shelf RL algorithms (w/ and w/o MAQ) on four Adroit tasks and evaluate their human-likeness and performance
  - Human-likeness: MAQ-based agents significantly outperform original RL
  - Performance: MAQ-based agents also achieve comparable success rates

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

DTW: Dynamic Time Warping
WD: Wasserstein Distance

$_s$: state distance between agent and humans
$_a$: action distance between agent and humans

Experiments

- We train three off-the-shelf RL algorithms (w/ and w/o MAQ) on four Adroit tasks and evaluate their human-likeness and performance
  - Human-likeness: MAQ-based agents significantly outperform original RL
  - Performance: MAQ-based agents also achieve comparable success rates

Tasks		BC	IQL	MAQ+IQL	SAC	MAQ+SAC	RLPD	MAQ+RLPD
Door	$DTW_s(\uparrow)$	0.18 ± 0.09	0.43 ± 0.06	$\mathbf{0.84 \pm 0.06}$	-0.39 ± 0.10	$\mathbf{0.80 \pm 0.08}$	-0.06 ± 0.04	$\mathbf{0.76 \pm 0.04}$
	$DTW_a(\uparrow)$	0.42 ± 0.13	0.61 ± 0.04	$\mathbf{0.95 \pm 0.01}$	-0.25 ± 0.04	$\mathbf{0.91 \pm 0.03}$	0.28 ± 0.08	$\mathbf{0.91 \pm 0.05}$
	$WD_s(\uparrow)$	0.32 ± 0.05	0.48 ± 0.05	$\mathbf{0.75 \pm 0.05}$	-0.28 ± 0.02	$\mathbf{0.71 \pm 0.08}$	-0.14 ± 0.04	$\mathbf{0.71 \pm 0.03}$
	$WD_a(\uparrow)$	0.41 ± 0.08	0.50 ± 0.02	$\mathbf{0.81 \pm 0.03}$	-0.15 ± 0.02	$\mathbf{0.77 \pm 0.07}$	0.10 ± 0.02	$\mathbf{0.76 \pm 0.03}$
	Success$(\uparrow)$	0.02 ± 0.01	0.16 ± 0.06	$\mathbf{0.93 \pm 0.04}$	0.43 ± 0.23	$\mathbf{0.56 \pm 0.50}$	$\mathbf{0.96 \pm 0.07}$	0.93 ± 0.05
Hammer	$DTW_s(\uparrow)$	-0.16 ± 0.07	-0.14 ± 0.34	$\mathbf{0.64 \pm 0.17}$	-1.10 ± 0.35	$\mathbf{0.61 \pm 0.21}$	-0.03 ± 0.14	$\mathbf{0.68 \pm 0.17}$
	$DTW_a(\uparrow)$	0.47 ± 0.02	0.45 ± 0.20	$\mathbf{0.92 \pm 0.06}$	-0.33 ± 0.11	$\mathbf{0.91 \pm 0.10}$	0.37 ± 0.08	$\mathbf{0.94 \pm 0.07}$
	$WD_s(\uparrow)$	0.11 ± 0.06	0.12 ± 0.13	$\mathbf{0.75 \pm 0.03}$	-0.44 ± 0.09	$\mathbf{0.64 \pm 0.12}$	-0.03 ± 0.08	$\mathbf{0.76 \pm 0.04}$
	$WD_a(\uparrow)$	0.30 ± 0.03	0.30 ± 0.10	$\mathbf{0.84 \pm 0.02}$	-0.19 ± 0.04	$\mathbf{0.78 \pm 0.11}$	0.20 ± 0.03	$\mathbf{0.85 \pm 0.03}$
	Success$(\uparrow)$	0.00 ± 0.00	$\mathbf{0.01 \pm 0.01}$	0.00 ± 0.00	$\mathbf{0.01 \pm 0.01}$	0.00 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.56 ± 0.37
Pen	$DTW_s(\uparrow)$	0.53 ± 0.13	0.34 ± 0.09	$\mathbf{0.55 \pm 0.17}$	0.06 ± 0.20	$\mathbf{0.58 \pm 0.17}$	0.48 ± 0.24	$\mathbf{0.54 \pm 0.18}$
	$DTW_a(\uparrow)$	0.58 ± 0.05	0.51 ± 0.05	$\mathbf{0.58 \pm 0.09}$	-0.34 ± 0.16	$\mathbf{0.58 \pm 0.11}$	0.40 ± 0.17	$\mathbf{0.59 \pm 0.13}$
	$WD_s(\uparrow)$	0.59 ± 0.12	0.54 ± 0.11	$\mathbf{0.59 \pm 0.13}$	0.29 ± 0.10	$\mathbf{0.61 \pm 0.12}$	0.49 ± 0.08	$\mathbf{0.59 \pm 0.12}$
	$WD_a(\uparrow)$	0.65 ± 0.14	0.63 ± 0.12	$\mathbf{0.66 \pm 0.15}$	0.22 ± 0.15	$\mathbf{0.67 \pm 0.14}$	0.44 ± 0.12	$\mathbf{0.66 \pm 0.14}$
	Success$(\uparrow)$	0.40 ± 0.03	0.40 ± 0.05	$\mathbf{0.42 \pm 0.07}$	0.32 ± 0.09	$\mathbf{0.41 \pm 0.01}$	$\mathbf{0.62 \pm 0.09}$	0.42 ± 0.05
Relocate	$DTW_s(\uparrow)$	0.09 ± 0.14	0.20 ± 0.20	$\mathbf{0.52 \pm 0.06}$	-0.55 ± 0.20	$\mathbf{0.25 \pm 0.19}$	0.03 ± 0.13	$\mathbf{0.27 \pm 0.14}$
	$DTW_a(\uparrow)$	0.47 ± 0.15	0.51 ± 0.11	$\mathbf{0.82 \pm 0.01}$	-0.10 ± 0.16	$\mathbf{0.66 \pm 0.09}$	0.32 ± 0.13	$\mathbf{0.69 \pm 0.10}$
	$WD_s(\uparrow)$	0.27 ± 0.15	0.36 ± 0.06	$\mathbf{0.47 \pm 0.07}$	-0.22 ± 0.06	$\mathbf{0.40 \pm 0.07}$	0.02 ± 0.07	$\mathbf{0.38 \pm 0.09}$
	$WD_a(\uparrow)$	0.45 ± 0.12	0.50 ± 0.04	$\mathbf{0.65 \pm 0.03}$	-0.05 ± 0.03	$\mathbf{0.61 \pm 0.03}$	0.20 ± 0.03	$\mathbf{0.55 \pm 0.08}$
	Success$(\uparrow)$	0.01 ± 0.02	0.00 ± 0.00	$\mathbf{0.20 \pm 0.10}$	0.00 ± 0.00	$\mathbf{0.14 \pm 0.07}$	0.14 ± 0.03	$\mathbf{0.17 \pm 0.10}$

DTW: Dynamic Time Warping
WD: Wasserstein Distance

$_s$: state distance between agent and humans
$_a$: action distance between agent and humans

Human Study

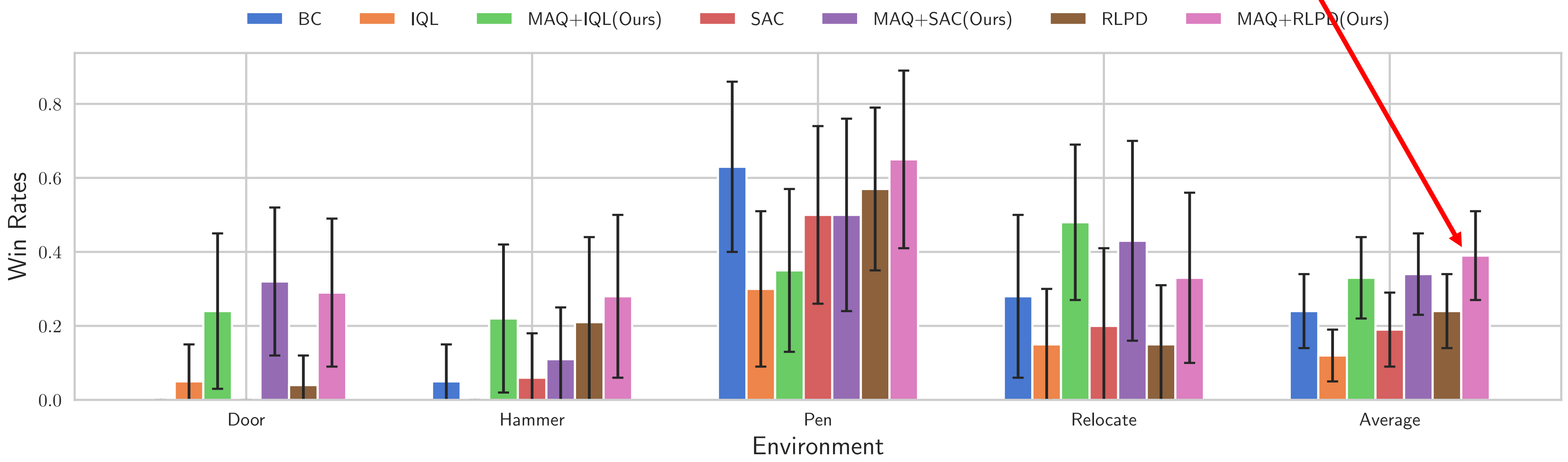
- We further evaluate the human-likeness of the agents through a human evaluation study

- We conduct a two-stage 2AFC questionnaire
  - **Turing Test**: evaluators are shown two videos, one from a *human demonstration* and the other from *the trained agent*

Experiments (Turing Test)

- MAQ-based agents achieve higher win rates compared to non-MAQ-based agents

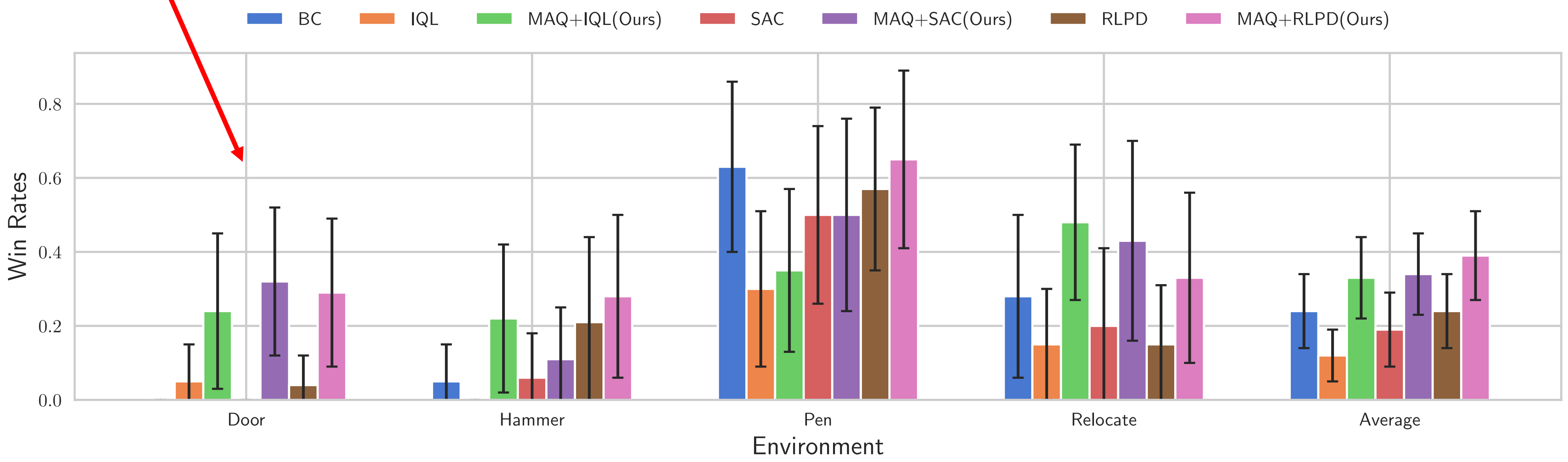
MAQ+RLPD gets the highest win rate overall



Experiments (Turing Test)

- MAQ-based agents achieve higher win rates compared to non-MAQ-based agents

RLPD and SAC, reward-driven RL agents achieve nearly 0% win rates



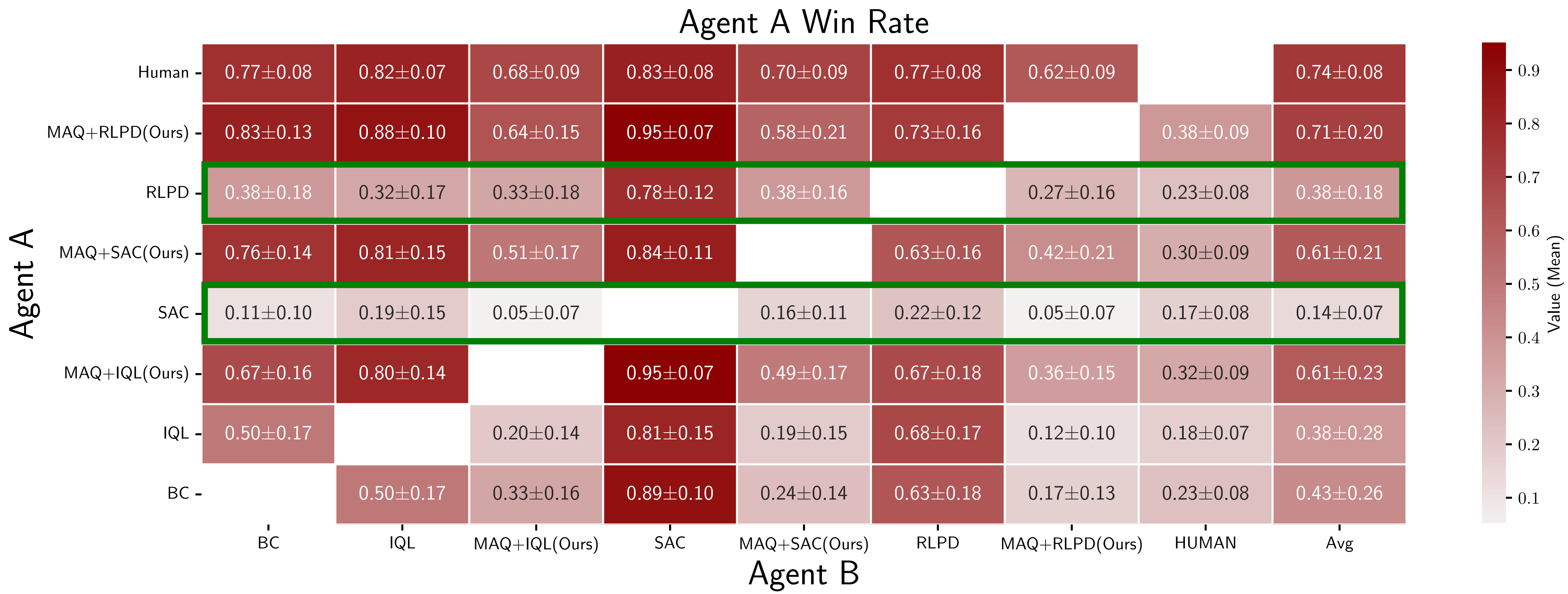
Human Study

- We further evaluate the human-likeness of the agents through a human evaluation study

- We conduct a two-stage 2AFC questionnaire
  - **Turing Test**: evaluators are shown two videos, one from a *human demonstration* and the other from *the trained agent*
  - **Human-likeness ranking test**: Similar to the Turing Test, but both may be trained agents

Experiments (Human-likeness Ranking Test)

- Reward-driven agents are easy to distinguish from their play by a computer



Experiments (Human-likeness Ranking Test)

- MAQ+RLPD achieves a 71% win rate across all agent pairs, achieving performance comparable to the human demonstration's 74% win rate

Agent A Win Rate

Agent A \ Agent B	BC	IQL	MAQ+IQL(Ours)	SAC	MAQ+SAC(Ours)	RLPD	MAQ+RLPD(Ours)	HUMAN	Avg
Human	0.77±0.08	0.82±0.07	0.68±0.09	0.83±0.08	0.70±0.09	0.77±0.08	0.62±0.09		0.74±0.08
MAQ+RLPD(Ours)	0.83±0.13	0.88±0.10	0.64±0.15	0.95±0.07	0.58±0.21	0.73±0.16		0.38±0.09	0.71±0.20
RLPD	0.38±0.18	0.32±0.17	0.33±0.18	0.78±0.12	0.38±0.16		0.27±0.16	0.23±0.08	0.38±0.18
MAQ+SAC(Ours)	0.76±0.14	0.81±0.15	0.51±0.17	0.84±0.11		0.63±0.16	0.42±0.21	0.30±0.09	0.61±0.21
SAC	0.11±0.10	0.19±0.15	0.05±0.07		0.16±0.11	0.22±0.12	0.05±0.07	0.17±0.08	0.14±0.07
MAQ+IQL(Ours)	0.67±0.16	0.80±0.14		0.95±0.07	0.49±0.17	0.67±0.18	0.36±0.15	0.32±0.09	0.61±0.23
IQL	0.50±0.17		0.20±0.14	0.81±0.15	0.19±0.15	0.68±0.17	0.12±0.10	0.18±0.07	0.38±0.28
BC		0.50±0.17	0.33±0.16	0.89±0.10	0.24±0.14	0.63±0.18	0.17±0.13	0.23±0.08	0.43±0.26

Value (Mean): 0.1 – 0.9

Summary

- We investigate Human-Like Reinforcement Learning, which seeks both human-like behavior and optimal performance
  - An important direction for the future
  - Benefit to human-AI collaboration/interaction and trust in the future
- We propose MAQ, which learns macro actions from human demonstration
  - Directly use human actions to interact with the environment
  - Without the need to predefine behavior constraints or rule-based penalties
- Our results show that MAQ achieves comparable performance and strong human-likeness
  - In the human evaluation study, MAQ-based agent is nearly indistinguishable from humans

Thank You for Your Attention

Our code and data are available at
https://rlg.iis.sinica.edu.tw/papers/MAQ