Uni-RLHF: Universal Platform and Benchmark Suite for Reinforcement Learning with Diverse Human Feedback

Yifu Yuan, Jianye Hao, Yi Ma, Zibin Dong, Hebin Liang, Jinyi Liu, Zhixin Feng, Kai Zhao, Yan Zheng

ICLR

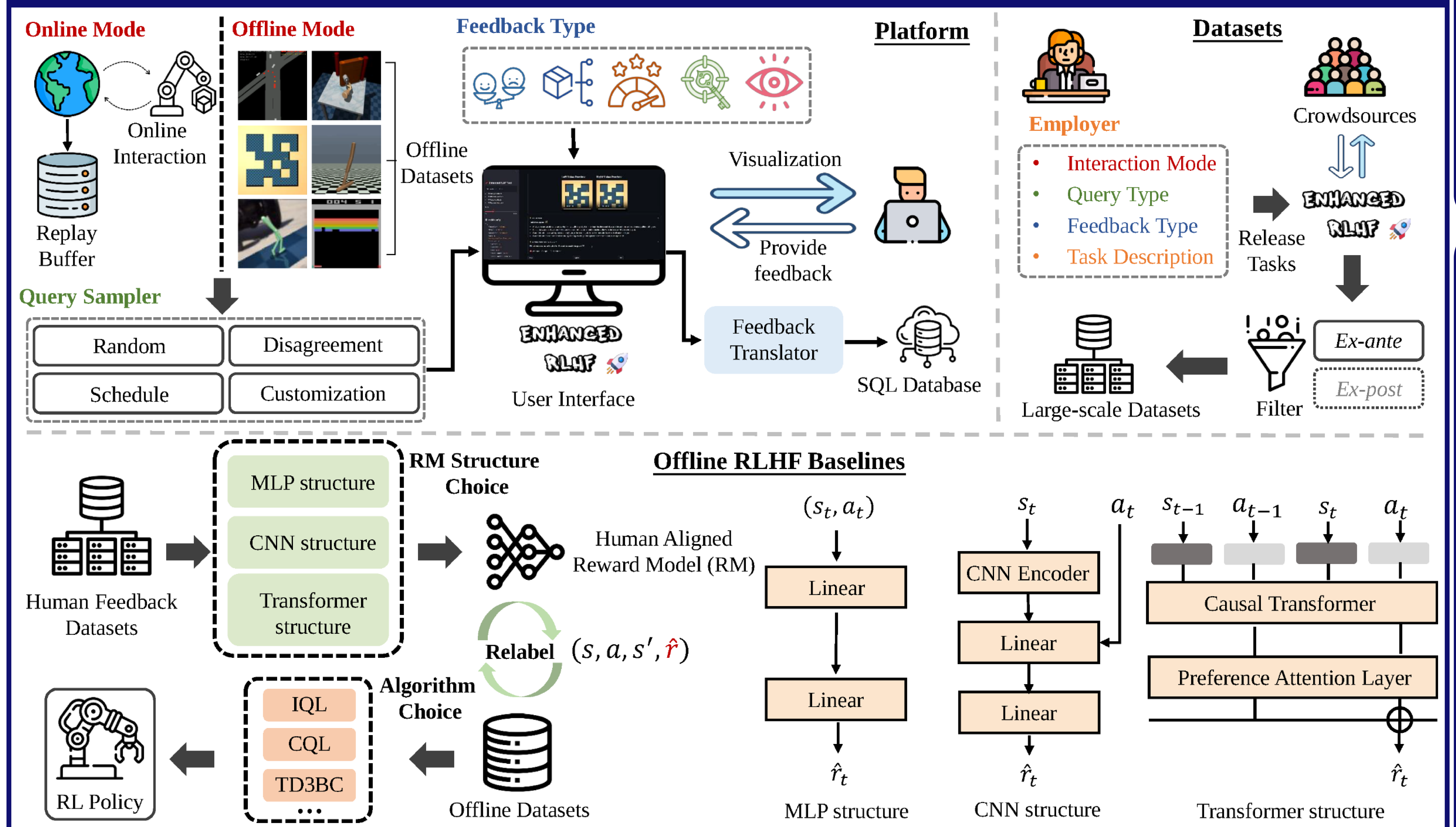
☞: https://uni-rlhf.github.io/ Tianjin University ✉: yuanyf@tju.edu.cn

Overview

We introduce **Uni-RLHF**, a comprehensive system implementation tailored for RLHF. It aims to provide a complete workflow from ***real human feedback***. Uni-RLHF contains:

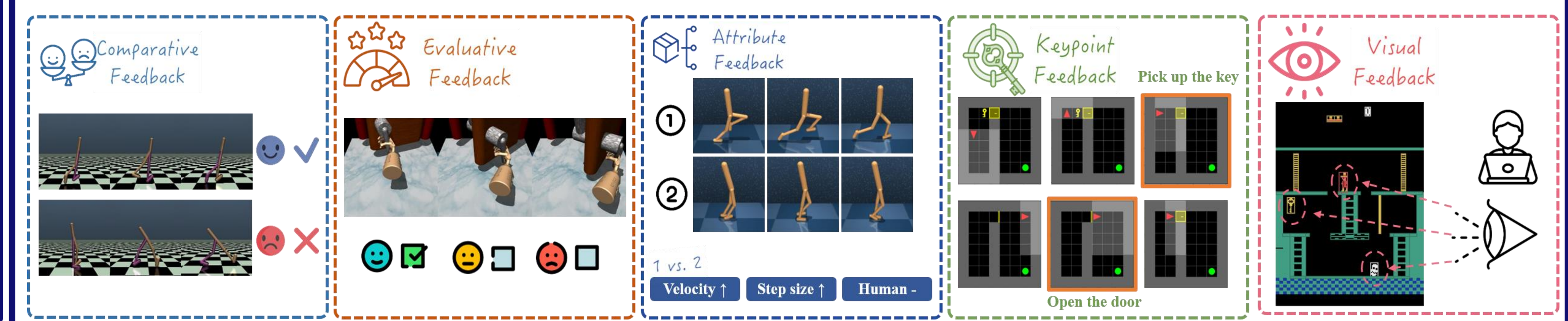
- **A universal multi-feedback annotation platform** → **32 tasks**
- **Large-scale crowdsourced feedback datasets** → **15 million annotation**
- **Modular offline RLHF baseline implementations** → **3 RM structure**

① Multi-feedback Annotation Platform



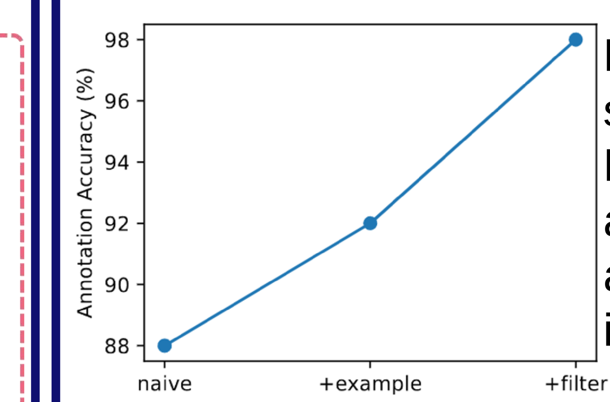
- The interface supports **Offline and Online Mode**, and can be extended to access new environments through simple interface extensions
- The **Query Sampler** determines sampling strategies and what data needs to be labelled
- The **User Interface** allows crowdsourcing to view available track clips and provide feedback responses, offering a range of video clip and image annotation methods
- **Feedback Translator** convert different feedback labels into a standard format

② Standardized Feedback Encoding Format



- **Comparative Feedback:** Gives relative binary feedback comparison between two trajectories
- **Attribute Feedback:** Gives a relative feedback comparison between two trajectories based on multiple attributes
- **Evaluative Feedback:** Gives multiple levels of evaluation options for a trajectory
- **Visual Feedback:** Selects and labels the visual highlights of a track
- **Keypoint Feedback:** Capture and mark keyframes in a trajectory

③ Large-scale Annotation Pipeline



Experts first annotate 300 segments as Oracle. Next, the researchers asked 5 crowdsource to annotate 100 trajectories in 3 settings.

- **Naïve:** only the task description
- **Example:** five annotated samples and detailed analyses
- **Filter:** added filters

Each component significantly improves the reliability of the annotation, **ultimately achieving a 98% agreement rate with expert annotations.**

④ Evaluating Offline RL With Comparative Feedback

Oracle: ***Ground Truth*** **ST**: ***Scripted Teacher*** **CS**: ***Crowd Sourced***

Dataset	IQL					CQL					TD3BC				
	Oracle	ST-MLP	ST-TFM	CS-MLP	CS-TFM	Oracle	ST-MLP	ST-TFM	CS-MLP	CS-TFM	Oracle	ST-MLP	ST-TFM	CS-MLP	CS-TFM
walker2d-m	80.91*	73.7	75.39	78.4	79.36	80.75*	76.9	75.62	76.0	77.22*	80.91	86.0	80.26	26.3	84.11*
walker2d-m-r	82.15*	68.6	60.33	67.3	56.52	73.09*	-0.3	33.18	20.6	1.82	82.15*	82.8*	24.3	47.2	61.94
walker2d-m-e	111.72*	109.8	109.16	109.4	109.12	109.56*	108.9	108.83	92.8	98.96	111.72*	110.4	110.13	74.5	110.75*
hopper-m	67.53	51.8	37.47	50.8	67.81*	59.08*	57.1	44.04	54.7	63.47*	60.37	58.6	62.89	48.0	99.42*
hopper-m-r	97.43*	70.1	64.42	87.1	22.65	95.11*	2.1	2.08	1.8	52.97	64.42*	44.4	24.35	25.8	41.44
hopper-m-e	107.42	107.7	109.16	94.3	111.43*	99.26*	57.5	57.27	57.4	57.05	101.17	103.7*	104.14*	97.4	91.18
halfcheetah-m	48.31*	47.0	45.10	43.3	43.24	47.04*	43.9	43.26	43.4	43.5	48.10	50.3*	48.06	34.8	46.62
halfcheetah-m-r	44.46*	43.0	40.63	38.0	39.49	45.04*	42.8	40.73	41.9	40.97	44.84*	44.2*	36.87	38.9	29.58
halfcheetah-m-e	94.74*	92.2	92.91	91.0	92.20	95.63*	69.0	63.84	62.7	64.86	90.78	94.1*	78.99	73.8	80.83
mujoco average	81.63	73.7	69.9	73.29	69.09	78.28	50.9	52.09	50.14	55.65	76.45	74.8	63.33	51.86	71.76
antmaze-u	77.00*	71.59	74.67*	74.22*	68.44	92.75*	93.71*	91.71*	63.95	91.34*	70.75	93.51*	92.90*	90.25	92.30*
antmaze-u-d	54.25	51.66	59.67	54.60	63.82*	37.25*	34.05	25.11	6.77	22.75	44.75	73.19*	36.45	51.88	59.58
antmaze-m-p	65.75	74.24*	71.67*	72.31*	65.25	65.75*	7.98	62.39*	60.26*	64.67*	0.25*	0.21*	0.00*	0.25*	0.39*
antmaze-m-d	73.75*	65.74	66.00	62.69	64.91	67.25	17.50	63.27	46.95	69.74*	0.25	3.33*	0.39	0.10	0.32
antmaze-l-p	42.00	40.79	43.33*	49.86*	44.63*	20.75	1.70	18.45	44.45*	19.33	0.00	0.07*	0.00	0.00	0.00
antmaze-l-d	30.25	49.24*	29.67	21.97	29.67	20.50	20.88	12.39	0.00	33.00*	0.00	0.00	0.00	0.00	0.00
antmaze average	57.17	58.91	57.67	55.94	56.12	50.71	29.3	45.55	37.06	50.14	19.33	28.38	21.62	23.75	25.43
pen-human	78.49*	50.15	63.66	57.26	66.07	13.71	9.80	20.31*	6.53	23.77*	-3.88*	-3.94*	-3.94*	-3.71*	-2.81*
pen-cloned	83.42*	59.92	64.65	62.94	62.26	1.04*	3.82*	3.721*	2.88*	3.18*	5.13	10.84*	14.52*	6.71	19.13*
pen-expert	128.05	132.85*	127.29	120.15	122.42	-1.41	138.34*	119.60	121.14	122.41	122.53*	14.41	34.62	11.45	30.28
door-human	3.26	3.46	6.8*	5.05*	3.22	5.53*	4.68	4.92	10.31*	8.81*	-0.13*	-0.32	-0.32	-0.33	-0.32
door-cloned	3.07*	-0.08	-0.06	-0.10	-0.02	-0.33*	-0.34*	-0.34*	-0.34*	-0.34*	0.29*	-0.34	-0.34	-0.34	-0.34
door-expert	106.65*	105.35	105.05	105.72	105.00	-0.32	103.90*	103.32*	102.63	103.15*	-0.33*	-0.34*	-0.33*	-0.34*	-0.34*
hammer-human	1.79*	1.43*	1.85*	1.03*	0.54*	0.14	0.85*	1.41*	0.70*	1.13*	1.02*	1.00*	0.96*	0.46	1.02*
hammer-cloned	1.50*	0.70*	1.87*	0.67*	0.73*	0.30*	0.28*	0.29*	0.28*	0.29*	0.25	0.25	0.27	0.45*	0.35*
hammer-expert	128.68*	127.4	127.36	91.22	126.50	0.26	120.16*	120.66*	117.65*	117.60*	3.11*	2.22	3.13*	2.13	3.03*
adroit average	59.43	53.46	55.39	49.33	54.08	2.1	42.39	41.57	40.2	42.22	14.22	2.64	5.4	1.83	5.62

- The **IQL-based baseline is the most stable**, and IQL-CS's perform as well as IQL-Oracle
- The TFM structure outperforms the MLP structure, especially in the environment of sparse reward
- Compared to Scripted Teacher (ST), **Crowd Sourced (CS) can achieve comparable or even superior results** in most environments

⑤ SMARTS Experiments

Can the RLHF method successfully replace hand-designed reward functions on real complex tasks?

Item	Value	Definition
Distance-Traveled Reward	Single-step travelling distance	Encouraging vehicles to travel towards the goal
Reach Goal Reward	+20	If ego vehicle reach the goal
Near Goal Reward	min(dis_to_goal/5,10)	If ego vehicle close the goal
Collision Penalty	-40	Collision
Mistake Penalty	-6	Ego vehicle triggers off road, off lane or wrong way
Lane Selection Reward	+0.3	The ego vehicle is in the right lane
Lane Change Penalty	-1	The ego vehicle changes lane
Time Penalty	-0.3	Each time step gives a fixed penalty term

- Multi-Metric Evaluation for SMARTS

	IQL-Oracle			IQL-CrowdSource		
	success rate↑	speed↑	comfort↓	success rate↑	speed↑	comfort↓
left-c	0.53 ± 0.03	9.75 ± 0.32	7.10 ± 0.06	**0.70 ± 0.03**	**10.04 ± 0.16**	**6.98 ± 0.26**
cruise	**0.71 ± 0.03**	**13.65 ± 0.03**	1.85 ± 0.08	0.62 ± 0.03	12.61 ± 0.44	**1.84 ± 0.49**
cutin	**0.85 ± 0.04**	13.84 ± 0.05	0.95 ± 0.23	0.80 ± 0.03	**13.90 ± 0.01**	**0.86 ± 0.05**
avg	0.70	**12.42**	3.30	**0.71**	12.19	**3.23**

⑥ Attribute Feedback

Walker is able to follow the user's preference for speed and height