CLARIFY: Contrastive Preference Reinforcement Learning for Untangling Ambiguous Queries

通过对比学习得到轨迹 embedding space，解决 PbRL 打不出偏好的问题

牟倪 2025.03.24

Preference-based Reinforcement Learning (PbRL)

- PbRL：基于偏好信息（preference）的强化学习
- 环境不提供 reward
- 使用专家对于两个轨迹段（trajectory segment）的偏好，进行策略优化
- 优势：以自然的方式利用人类专家的经验，避免了困难的 reward engineering 过程

Query：包含两个轨迹段 (σ_0, σ_1)

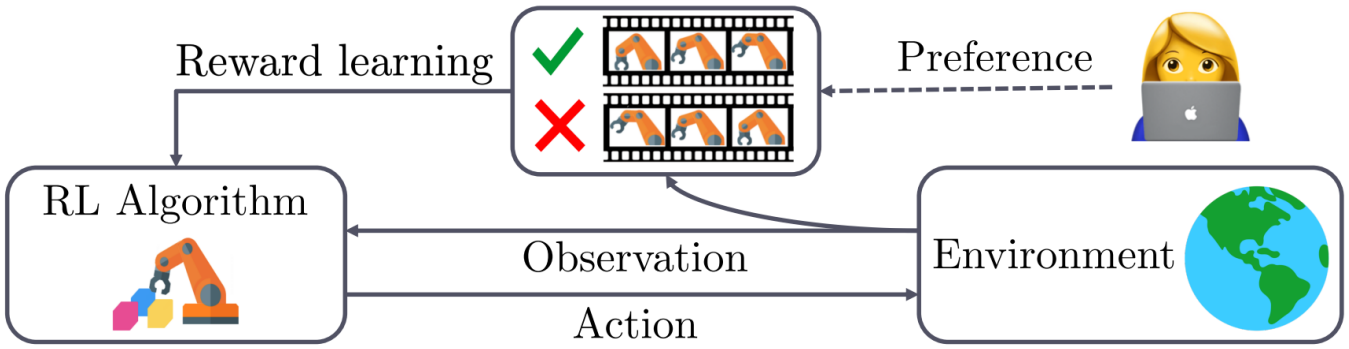
Preference $p = (1,0)$
代表 $\sigma_0 \succ \sigma_1$

经典的 PbRL 方法

- 基于 Bradley-Terry 模型，使用形式为 (σ_0, σ_1, p) 的 preference 信息，学习 reward model $\hat{r}_\psi(s,a)$，用于代替传统 RL 中的 reward function

$$P_\psi[\sigma^0 \succ \sigma^1] = \frac{\exp \sum_t \hat{r}_\psi\left(s_t^{\sigma_0}, a_t^{\sigma_0}\right)}{\exp \sum_t \hat{r}_\psi\left(s_t^{\sigma_0}, a_t^{\sigma_0}\right) + \exp \sum_t \hat{r}_\psi\left(s_t^{\sigma_1}, a_t^{\sigma_1}\right)}$$

- 即，将偏好概率建模为 segment reward 之和的 softmax

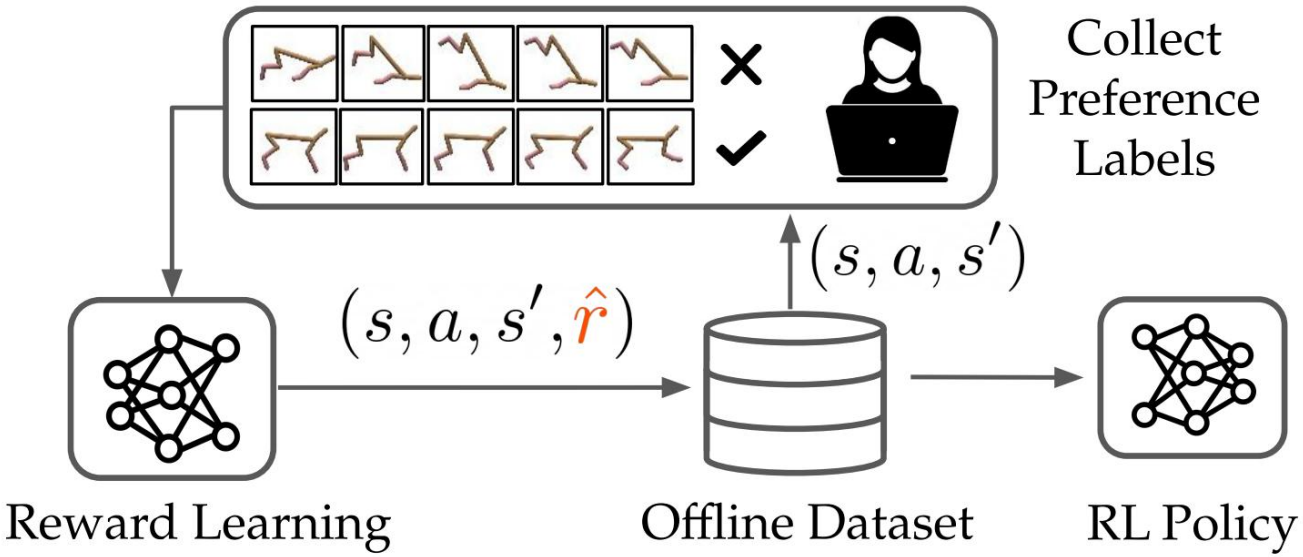


流程：

- 收集 preference 信息
- 学习 reward model
- 给 replay buffer 里的数据标注 reward
- 优化 policy
- 继续收集 preference 信息
- ……

Offline PbRL（我们工作关注的 setting）

- 提供不含 reward 信息的 offline dataset D
- 通过询问专家两段轨迹 $(\sigma_0, \sigma_1) \sim D$ 哪段更好，训练得到 reward model $\hat{r}_\psi(\mathrm{s}, \mathrm{a})$，然后执行普通 offline RL

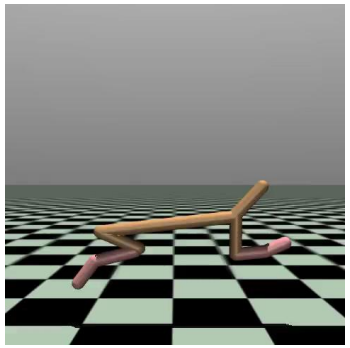


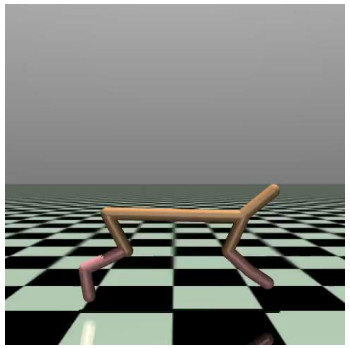
PbRL 领域主要关注的研究方向

- 关注 preference feedback efficiency 的工作：
  - 希望利用更少的 preference feedback 学出策略，节省专家成本。
  - PEBBLE (2021), SURF (2022), OPRL (2023), QPA (2023), …
- 关注 reward learning 鲁棒性的工作：
  - 考虑专家给出的 preference 有误 / 不够明确的情况，希望学到更鲁棒的 reward model。
  - B-Pref (2021), PT (2023), RIME (2023), S-EPOA (2024), …
- 使用不同于二元比较（比较两个 segment）的数据形式，来学习 reward model：
  - SeqRank (2023), LiRE (2024), …
- LLM 微调中的应用（RLHF）：
  - 对于同一 prompt 下的两个回答，让专家 / 用户评价哪个更好，从而改进 LLM 策略。

Gap & Motivation

- Gap：对于两段相似的轨迹，人类专家会难以区分、打出 preference，只能跳过。





- Motivation：
  - 如果能够训练轨迹的 embedding space，将 preference 信息融入 embedding space 里，使得相似轨迹的 embedding 相近，区分度大的轨迹的 embedding 距离较远。
  - 则可根据该 embedding space，选择人类可区分的 query，即 embedding 距离较远的两个轨迹，从而缓解人类难以打出 preference 导致的 feedback 效率低下的问题。

验证“人类难以打出 preference”现象是真实存在的

- 在机械臂操作任务上进行验证。
- 让人类给具有不同 segment return 差值大小的 query 标注 preference，统计：
  1. 人类可以打出 preference 的比率；
  2. 人类的 preference 标注与 ground truth preference 匹配的比率。
- 实验结果： segment return 差值越小，人类越容易打不出 preference 和犯错。
- （ground truth preference 通过比较 (σ_0, σ_1) 的 return 大小生成）

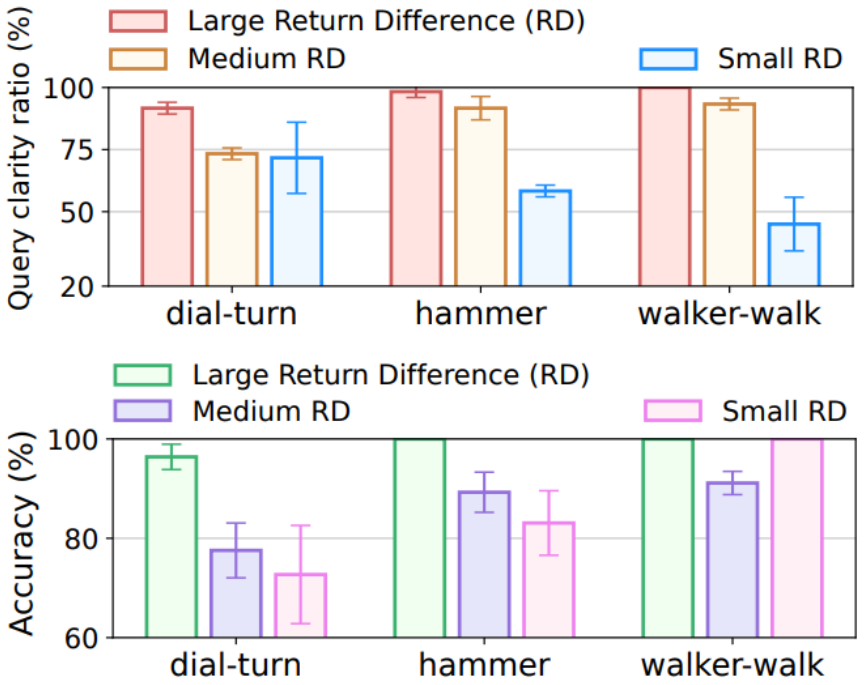


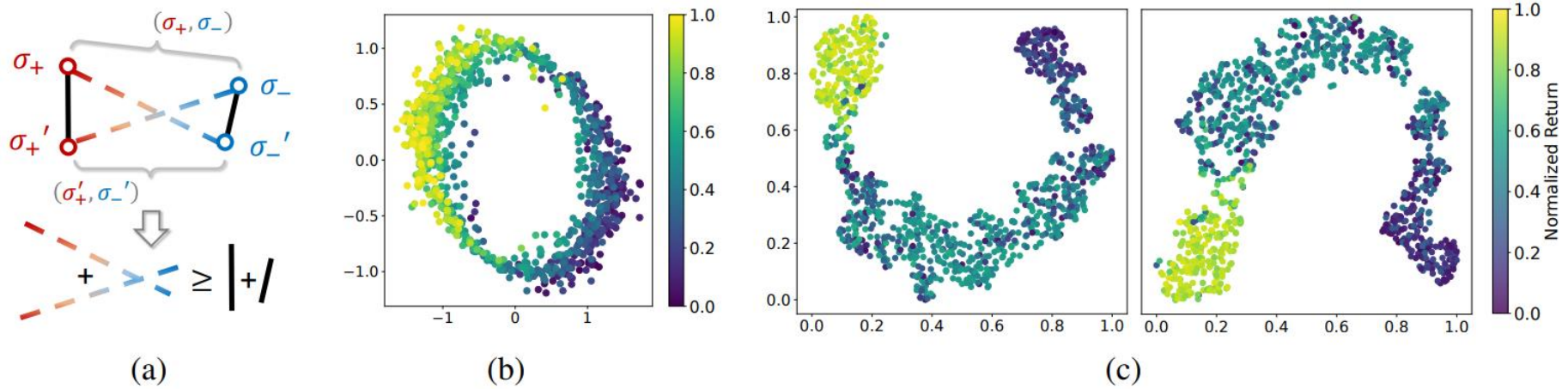
Figure 5: Query clarity ratio and accuracy of human labels for queries with varying return differences (RD). For dial-turn and hammer, RD values are 300, 100, and 10; for walker-walk, RD are 30, 10, and 1.

核心创新点：用对比学习 loss 建模 preference 信息

- 对比学习的主要思想：
  - 利用数据集中的正负样本，从数据中提取特征；提高正样本对 (x, x^+) 的特征的相似度，降低负样本对 (x, x^-) 的特征相似度
- 为把 preference 信息 (σ_0, σ_1, p) 融合进轨迹的 embedding space，我们提出两个对比学习 loss
- 对比学习 loss 1：利用“一个 query 能否被专家区分”信息
  - 可区分的 query：将 preference 标签为 (1, 0) (0, 1) 的 embedding 距离拉远
  - 不可区分的 query：将 preference 标签为 (0.5, 0.5) 的 embedding 距离拉近
  - $\mathcal{L}_{\text{amb}} = -\left[\mathbb{E}_{p\in\{(0,1)(1,0)\}} d(z_0, z_1) - \mathbb{E}_{p=(0.5,0.5)} d(z_0, z_1)\right]$

核心创新点：用对比学习 loss 建模 preference 信息

- 对比学习 loss 2：利用“可区分的 query 的具体偏好”信息，构建四边形 loss
  - 假设给定可区分的两个 query $(\sigma^+, \sigma^-), (\sigma^{+\prime}, \sigma^{-\prime})$，应该满足“正样本之间距离 + 负样本之间距离 ≤ 一正一负样本之间的距离”
  - $\mathcal{L}_{\text{quad}} = -\mathbb{E}_{p,p' \in \{(0,1)(1,0)\}}\left[d(z^+, z^{-\prime}) + d(z^-, z^{+\prime}) - d(z^+, z^{+\prime}) - d(z^-, z^{-\prime})\right]$



(a) (b) (c)

(a) 四边形 loss 示意图；(b) demo 实验中的 embedding 可视化；(c) 实验中训练的 embedding 可视化。

具体技术路线

- 总体框架：Reward learning + Policy learning（直接使用 IQL）
- Reward learning：

  收集 preference 数据 →

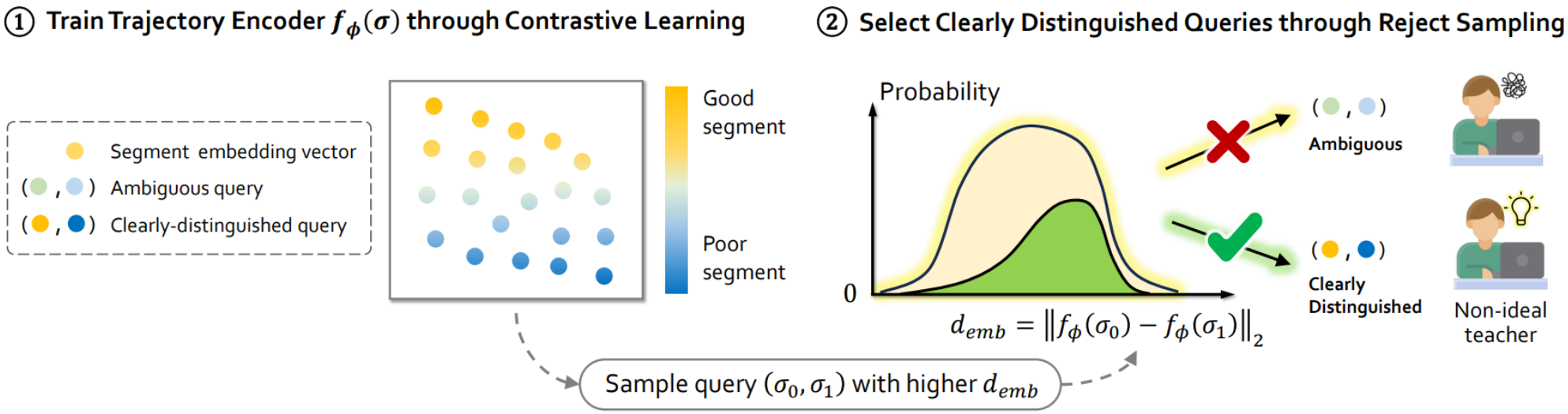
  训练 embedding 和 reward model → 基于 embedding 选择 query →

  得到新 preference 数据 →

  更新 embedding 和 reward model → 基于 embedding 选择 query……
- 轨迹编码器 $f_\phi(\sigma)$：使用 Transformer 进行序列编码，具体实现基于 Hindsight Information Matching（HIM）的 Bidirectional Decision Transformer（BDT）

具体技术路线

- 如何基于 embedding 选择 query：
  - 使用拒绝采样方法。基于已收集 query 的 embedding 距离 $d_{emb} = \left\| f_\phi(\sigma_0) - f_\phi(\sigma_1) \right\|_2$ 的分布，构建拒绝采样后的目标分布 $q(d_{emb})$。
  - 从目标分布中 $q(d_{emb})$ 采样 query，其可区分的概率更大，不可区分的概率更小。



具体技术路线

- 如何基于 embedding 选择 query：
  - 使用拒绝采样方法。基于已收集 query 的 embedding 距离 $d_{emb} = \left\| f_\phi(\sigma_0) - f_\phi(\sigma_1) \right\|_2$ 的分布，构建拒绝采样后的目标分布 $q(d_{emb})$。
  - 从目标分布中 $q(d_{emb})$ 采样 query，其可区分的概率更大，不可区分的概率更小。

$$\rho_1(d_{\text{emb}}) = \frac{\max\left(0, \rho_{\text{clr}}(d_{\text{emb}}) - \rho_{\text{amb}}(d_{\text{emb}})\right)}{\int \max\left(0, \rho_{\text{clr}}(d') - \rho_{\text{amb}}(d')\right) dd'},$$

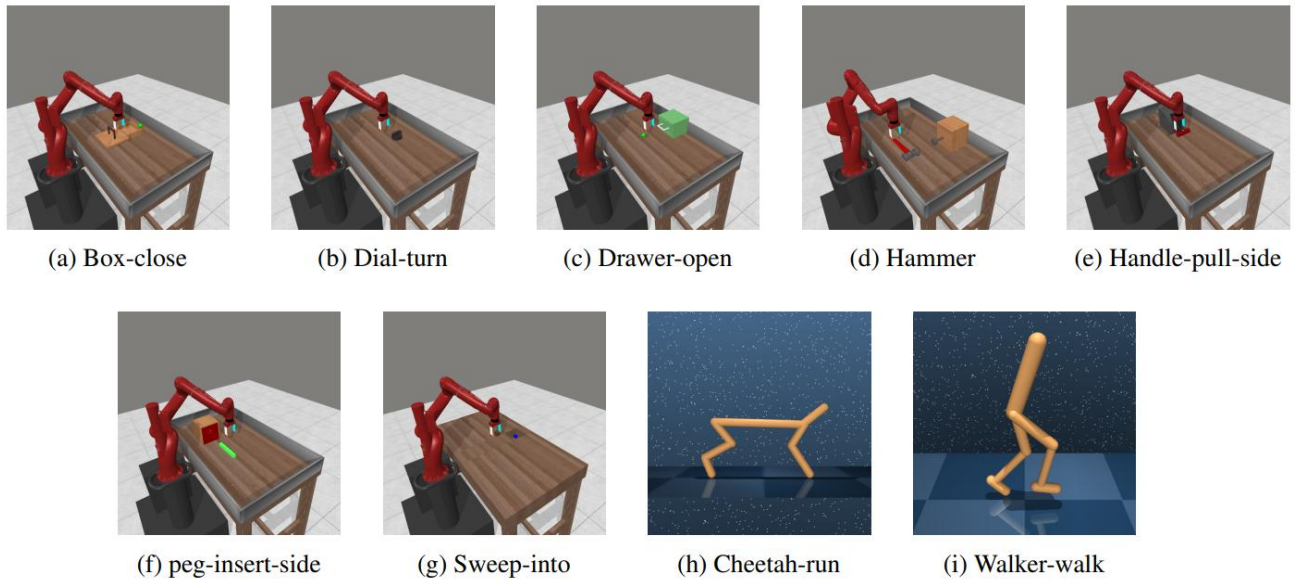
$$\rho_2(d_{\text{emb}}) = \frac{\rho_{\text{clr}}(d_{\text{emb}})/\rho_{\text{amb}}(d_{\text{emb}})}{\int (\rho_{\text{clr}}(d')/\rho_{\text{amb}}(d'))dd'}, \tag{7}$$

$$\rho(d_{\text{emb}}) = 0.5(\rho_1(d_{\text{emb}}) + \rho_2(d_{\text{emb}})).$$

$$q(d_{\text{emb}}) = p(d_{\text{emb}}) \cdot \rho(d_{\text{emb}}), \tag{8}$$

实验设置

- 实验场景： Metaworld、DMControl
- 如何生成 preference 信息：
  - 基于 ground truth reward 构造“scripted teacher”，跳过 segment return 相似的 query（与第 7 页验证实验相对应）。使用 skip rate ϵ 控制跳过 query 的数量，ϵ 越大，跳过越多。
  - 在人类实验中，人类直接标注 preference。



(a) Box-close (b) Dial-turn (c) Drawer-open (d) Hammer (e) Handle-pull-side

(f) peg-insert-side (g) Sweep-into (h) Cheetah-run (i) Walker-walk

实验结果

- 在不同的 skip rate ϵ 设置下，我们的方法（CLARIFY）性能都超过了 baselines

Table 1: Success rates on Metaworld tasks (the first 7 tasks) and episodic returns on DMControl tasks (the last 2 tasks), over 6 random seeds. We use skip rate $\epsilon = 0.5, 0.7$ and report the average performance and standard deviation of the last 5 trained policies. The yellow and gray shading represent the best and second-best performances, respectively.

Skip Rate	Algorithm	box-close	dial-turn	drawer-open	handle -pull-side	hammer	peg-insert -side	sweep-into	cheetah-run	walker-walk
-	IQL	94.90 ± 1.42	76.50 ± 1.73	98.60 ± 0.45	99.30 ± 0.59	72.00 ± 2.35	88.40 ± 1.30	79.40 ± 1.91	607.46 ± 8.13	830.15 ± 20.08
0.5	MR	0.26 ± 0.02	14.46 ± 5.27	50.47 ± 6.24	79.73 ± 12.19	0.14 ± 0.08	9.23 ± 2.03	23.25 ± 9.67	205.04 ± 50.53	322.93 ± 161.07
	OPRL	**8.25** ± 3.16	**57.33** ± 25.02	**72.67** ± 2.87	83.92 ± 7.98	14.80 ± 5.27	**22.00** ± 4.64	**61.00** ± 7.52	531.96 ± 48.75	646.40 ± 51.35
	PT	0.15 ± 0.20	30.54 ± 7.24	61.64 ± 10.07	**89.75** ± 6.07	0.11 ± 0.10	10.20 ± 2.94	46.35 ± 3.35	384.59 ± 99.26	599.43 ± 39.42
	OPPO	0.56 ± 0.79	13.06 ± 12.74	11.67 ± 6.24	0.56 ± 0.79	2.78 ± 4.78	0.00 ± 0.00	15.56 ± 11.57	346.01 ± 127.78	311.27 ± 42.73
	LiRE	3.60 ± 0.69	46.20 ± 6.75	63.47 ± 10.38	52.07 ± 33.58	**16.20** ± 13.37	21.60 ± 4.00	57.47 ± 2.74	**553.61** ± 43.16	**789.18** ± 28.77
	CLARIFY	**29.40** ± 16.27	**77.50** ± 7.37	**83.50** ± 7.40	**95.00** ± 1.22	**26.75** ± 11.26	**24.25** ± 6.65	**68.00** ± 7.13	**617.31** ± 14.43	**796.34** ± 12.87
0.7	MR	0.20 ± 0.11	16.57 ± 8.22	45.51 ± 10.25	74.82 ± 17.10	0.06 ± 0.09	5.21 ± 2.63	17.22 ± 6.05	234.77 ± 81.40	306.39 ± 134.72
	OPRL	7.40 ± 6.97	**63.40** ± 9.46	46.50 ± 18.83	**84.00** ± 7.11	5.00 ± 2.28	**23.75** ± 5.36	52.00 ± 9.33	513.94 ± 55.45	664.16 ± 89.47
	PT	0.18 ± 0.18	25.64 ± 7.40	**64.72** ± 18.44	74.94 ± 12.56	0.09 ± 0.05	8.21 ± 4.07	28.52 ± 6.67	429.19 ± 44.92	647.68 ± 38.53
	OPPO	0.56 ± 0.79	2.78 ± 2.83	11.67 ± 6.24	0.56 ± 0.79	4.44 ± 6.29	0.00 ± 0.00	15.56 ± 11.57	355.69 ± 59.35	304.19 ± 16.25
	LiRE	**7.67** ± 16.79	38.40 ± 8.52	39.10 ± 6.78	50.60 ± 31.59	**14.25** ± 7.30	**23.40** ± 3.40	**56.00** ± 6.40	**514.75** ± 10.02	**795.02** ± 22.80
	CLARIFY	**17.50** ± 6.87	**79.40** ± 3.83	**78.60** ± 10.52	**95.00** ± 1.10	**28.75** ± 15.58	23.00 ± 1.22	**59.67** ± 10.09	**593.24** ± 22.75	**816.54** ± 11.08

人类提供 preference 的实验

- ① 让人类给我们的方法和 baselines 选出来的 query 标注 preference，发现我们的方法拥有更高的
  1. 人类可以打出 preference 的比率；
  2. 人类 preference 与 ground truth preference 匹配的比率。
- ② 使用人类标注的 preference 训练策略，发现我们的方法能训出性能更好的策略。

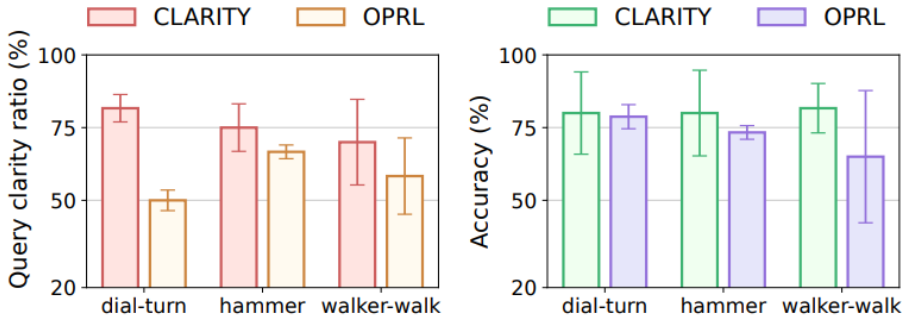


Figure 4: Query clarity ratio and accuracy of human labels for CLARIFY and OPRL, with CLARIFY-selected queries are more clearly distinguished for humans.

Table 5: Performance of CLARIFY and OPRL on walker-walk task, under real human labelers.

	CLARIFY	OPRL
Episodic Returns	**420.75** ± 52.02	265.91 ± 33.57
Query Clarity Ratio (%)	63.33 ± 8.50	53.33 ± 6.24
Accuracy (%)	**87.08** ± 9.15	66.67 ± 4.71

感谢倾听！

牟倪　自动化系　mn23@mails.tsinghua.edu.cn