增强人类标注可靠性的偏好强化学习方法研究

—— 博士开题答辩 ——

培养单位 ：自动化系

研 究 生 ：牟倪

学　　号 ：2023310855

指导教师 ：贾庆山 教授

答辩日期 ：2025年11月27日

智能与网络化系统研究中心

Center for Intelligent and Networked Systems

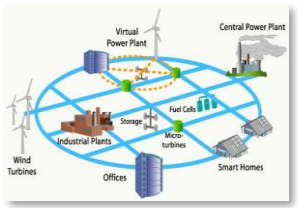
目 录

1.1 选题背景 强化学习方法具有重大研究价值

信息物理融合系统（CPS）是国家发展重要基础，求解 CPS 中的控制与决策问题具有重大战略意义。

典型 CPS 系统: 数据中心 [1]

智能制造 [2]



智能电网 [3]

传统控制方法难以应对 CPS 控制决策挑战 [4-6]：

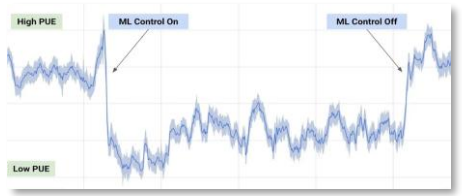
1. **高维度**：系统状态和动作空间维度过高 ⟶
2. **系统复杂非线性**：系统模型难以精确刻画 ⟶
3. **长周期序贯决策**：需考虑跨时间步的全局优化 ⟶

举例：

1. **数据中心**：状态包含数万传感器数据
2. **智能电网**：涉及潮流计算，高度非线性
3. **智能制造**：需对生产流程进行长期优化

强化学习（RL）解决高维、复杂、多步序贯决策问题，直接应对上述挑战，具有巨大工业化潜力。

AlphaGo 在搜索空间为 10^{170} 的围棋游戏中战胜人类顶尖棋手 [7]



Google 通过 DRL 策略节省 40% 数据中心冷却成本 [8]

1.2 选题背景 偏好强化学习：让策略与人类意图对齐

强化学习的难点：奖励函数难以精确设计与对齐

❌ **1. 奖励设计成本高**

依赖人工设计标量奖励函数，门槛高、成本大 [9]。

❌ **2. 人机意图对齐难**

难以将复杂、模糊的人类价值观（如安全、伦理、舒适度）转化为单一的标量函数 [10]。

❌ **3. 奖励函数漏洞**

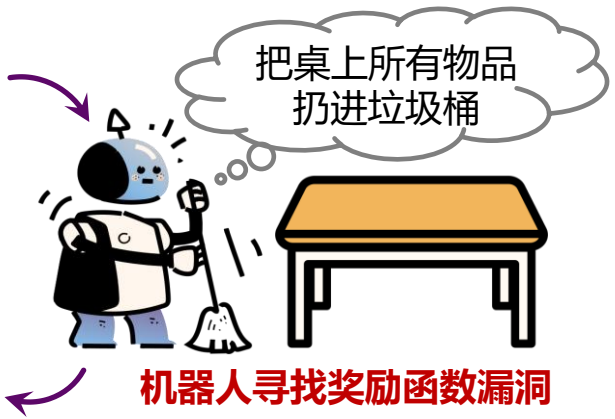
智能体倾向于寻找奖励函数的漏洞，而非实现人类的真实意图 [11]。

举例：

设计奖励函数：

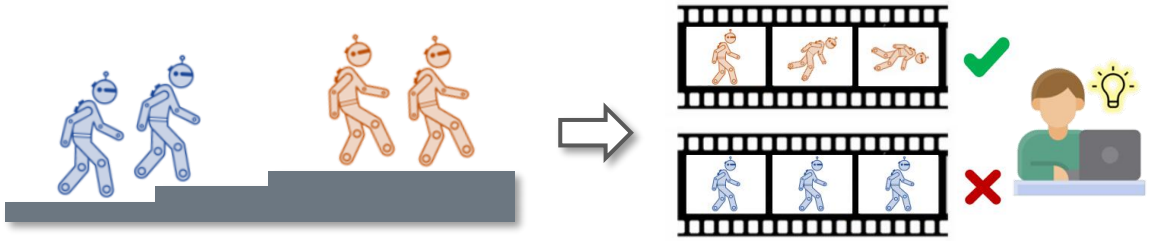
$$\begin{cases} -100, \text{桌上有物品} \\ +100, \text{没有物品} \end{cases}$$

继续调整奖励函数
耗费大量人工成本



机器人寻找奖励函数漏洞

偏好强化学习（Preference-based RL）



① 收集策略行为　　② 人类比较策略行为

✅ **直接解决上述难点：**

- 用 "行为比较" 代替 "精确打分"，**大幅降低人工成本**。
- 基于人类偏好建模奖励模型，**策略与人类意图相对齐**。

✅ **偏好强化学习已广泛应用：**

- LLM 微调：**Nature 期刊文章** [12]
- 机器人控制：**RA-L, ICRA** 等领域内期刊会议 [13-15]
- 自动驾驶 [16]、储能系统管理 [16]、...

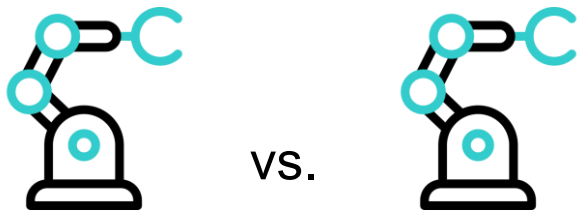
偏好强化学习 已成为重要研究方向

1.3 选题背景 偏好强化学习：局限性与落地挑战

- 偏好强化学习的隐性假设： **假设人类能够提供准确、高效、一致的偏好标注。**
- 当轨迹片段相似或目标冲突时，人类面临“认知负荷”[17]，**难以提供可靠的偏好标注。**

举例：机器人搬运物品 (对应 MetaWorld 基准任务)

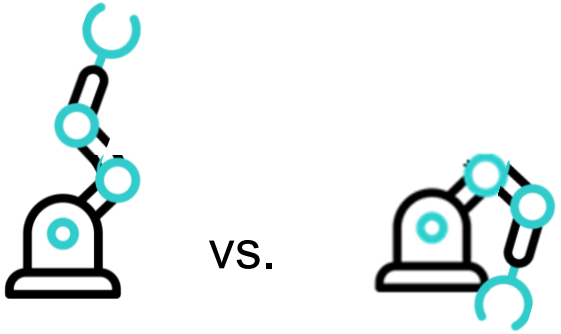
①



(极为相似、难以区分的行为)

待比较的行为高度相似，
人类难以区分行为之间的差异

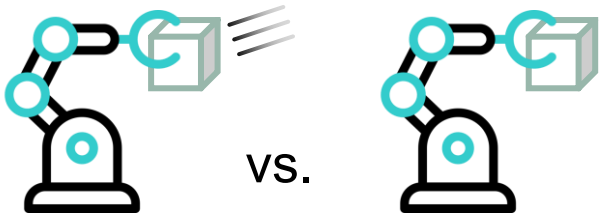
②



(扭曲的机械臂，两个“不可取”的行为)

面对两个没有明确优劣比较的行为，
人类难以给出清晰比较结果

③



(机械臂 速度快 vs. 拿得稳)

面对多目标权衡，如 安全 vs. 效率，
人类难以从单维度判断优劣

1.3 选题背景 偏好强化学习：局限性与落地挑战

- 偏好强化学习的隐性假设： **假设人类能够提供准确、高效、一致的偏好标注。**
- 当轨迹片段相似或目标冲突时，人类面临“认知负荷”[17]，**难以提供可靠的偏好标注。**
- **该局限性的影响：**
  - **数据效率低**：大量标注数据成为无效噪声。
  - **策略鲁棒性差**：基于不可靠偏好标注，易学习到存在偏差的奖励模型，导致策略在边界条件产生不可预测的危险行为。
  - **不可靠的意图对齐**：无法捕获人类的复杂多目标意图。

    ↓
  - 导致最终策略与人类真实价值观发生偏差，**无法解决 “奖励设计与意图对齐”的根本问题。**

亟需构建 **增强人类标注可靠性** 的偏好强化学习方法

目 录

偏好强化学习研究概述

提出范式：

Christiano et al. (2017) [11] 等工作，提出 **偏好强化学习** 范式：

一对轨迹片段 (σ_0, σ_1) 被称为**查询**

人类给出偏好 $p \in \{0,1\}$，代表认为 σ_p 更优

Bradley-Terry 模型

$$P_\psi[\sigma_1 \succ \sigma_0] = \frac{\exp \sum_t \hat{r}_\psi(s_t^1, a_t^1)}{\sum_{i\in\{0,1\}} \exp \sum_t \hat{r}_\psi(s_t^i, a_t^i)}$$

交叉熵损失函数

$$\min_\psi \mathcal{L}_{\text{reward}} = -\mathop{\mathbb{E}}_{(\sigma_0,\sigma_1,p)\sim D_p}\Big[(1-p)\log P_\psi[\sigma_0 \succ \sigma_1] + p \log P_\psi[\sigma_1 \succ \sigma_0]\Big]$$

基于 Bradley-Terry 模型建模人类偏好，

将奖励函数与人类意图对齐

研究方向：

① 提高**利用人类标注的效率**

② 提高**不可靠人类标注**下的**鲁棒性**

偏好强化学习范式的应用：

LLM 微调 [12]、机器人 [13-15]、自动驾驶 [16]、能源管理 [16]、……

2.2 研究现状 ① 提高利用人类标注的效率

方法大类	方法细分	核心思想	代表方法
查询选择	多样性驱动	最大化行为差异以获取信息量高的偏好数据	PEBBLE (Lee et al., 2021) [11], OPRL (Hejna et al., 2023) [18]
	不确定性感知	选择模型预测不一致的样本以提升学习效率	Biyik et al. (2020) [19]
	查询对齐优化	将查询选择与模型优化的关注点对齐	QPA (Hu et al., 2023) [20]
数据增强	样本合成	通过状态-动作插值生成新样本	SURF (Park et al., 2022) [21]
	偏序关系建模	利用片段间的连续偏序比较关系生成新样本	LiRE (Bai et al., 2024) [22]
预训练机制	奖励模型初始化	利用少量标签预训练奖励模型加速收敛	PEBBLE (Lee et al., 2021) [11]
	少样本迁移学习	跨任务迁移偏好知识降低标注需求	Hejna et al. (2023) [15]
网络架构创新	序列建模	用 Transformer 建模偏好序列依赖关系	Preference Transformer (Uesato et al., 2022) [23]
	策略-奖励协同优化	联合优化策略网络与奖励模型	DPPO (Stiennon et al., 2020) [24], OPPO (Liu et al., 2023) [25], MRN (Chen et al., 2024) [26]

- **存在的不足：**
  - 假设偏好标注**准确、相互一致**。然而，现实世界中，人类标注存在**不可靠情况**。

2.3 研究现状 ② 提高不可靠人类标注下的鲁棒性

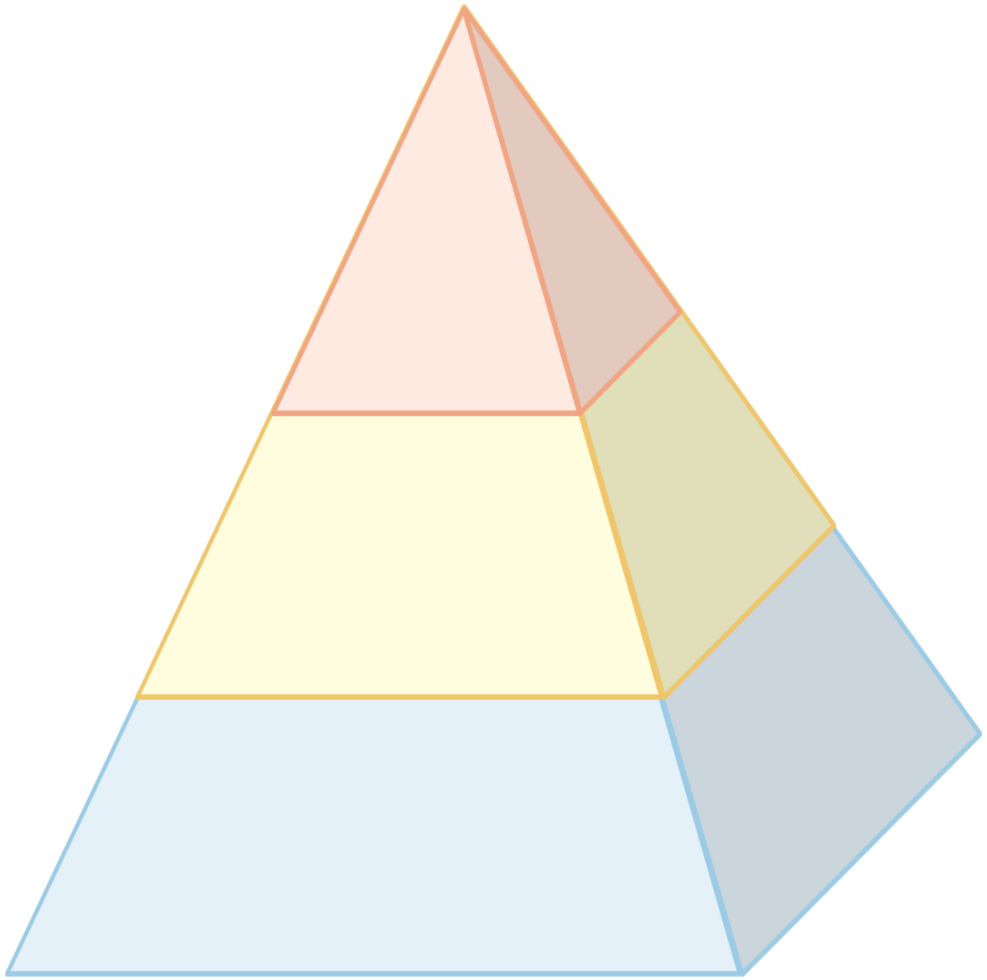
方法大类	方法细分	核心思想	代表方法
噪声建模基准	多类型噪声模拟	提供多种简单、理想噪声类型（错误、随机、跳过、短视等）的基准环境	B-Pref (Lee et al., 2021) [27]
样本选择与过滤	动态阈值过滤	使用 KL 散度等指标动态筛选可靠样本，翻转或丢弃不可靠样本	RIME (Cheng et al., 2024) [28]
数据增强	样本插值增强	通过对轨迹片段和偏好标签进行混合，提升泛化与抗噪能力	MCP (Heo et al., 2025) [29]
模型正则化	集成与置信度加权	使用多个奖励模型集成，基于置信度加权提升鲁棒性	B-Pref (Lee et al., 2021) [27]
	潜在空间约束	强制奖励模型潜在空间接近先验分布，提升预测一致性	Xue et al. (2023) [30]

- **存在的不足：**
  - 缺乏对偏好标注不可靠机制的建模，或建模过于简单（假设偏好标注随机错误等）

本研究将对**偏好标注不可靠的机制**进行系统建模，
并提出**增强人类标注可靠性**的偏好强化学习方法。

目 录

偏好强化学习中，人类标注为何会不可靠?



高层：**多目标决策的内在冲突性**

面对多目标权衡，如 安全 vs. 效率，

人类难以从单维度判断优劣

中层：**人类判断的优劣不确定性**

面对两个没有明确优劣比较的行为，

人类难以给出清晰比较结果

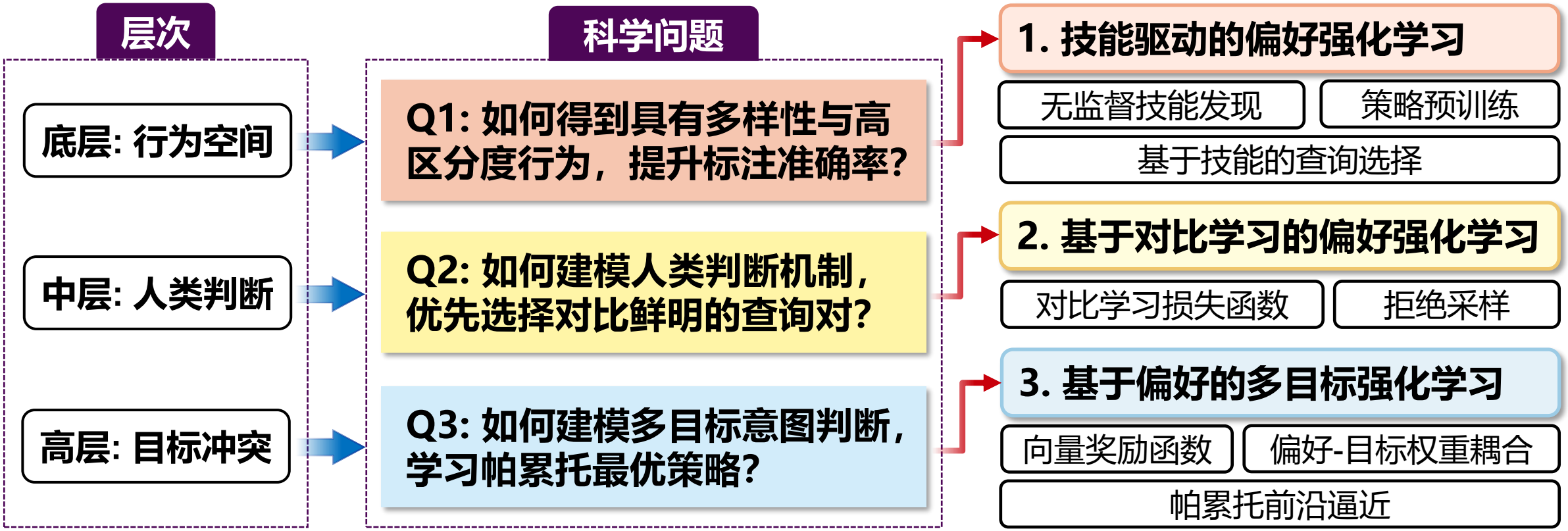
底层：**行为空间上的不可区分性**

待比较的行为高度相似，

人类难以区分行为之间的差异

3. 研究内容 研究内容概述

研究目标：提出 **增强人类标注可靠性** 的偏好强化学习方法

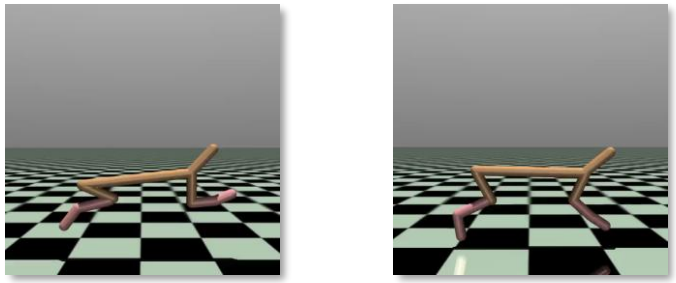


针对人类标注不可靠性的成因，进行系统性建模与分点突破，提出增强人类标注可靠性的偏好强化学习方法。

3.1 研究内容一 解决行为空间的不可区分性

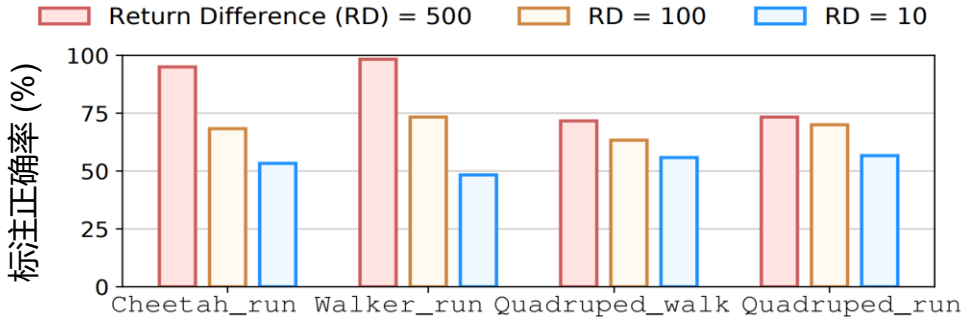
➢ 挑战：

- 传统方法 [10] 出于信息熵考虑，倾向于选择“不确定性最大”的轨迹片段给人类标注。
- 导致轨迹片段**高度相似、难以区分**，标注者“随机猜测”，产生高噪声的偏好数据，严重损害算法性能。



① 传统方法选择的两条片段，其行为难以区分

* 任务为智能体向前奔跑；该动图为智能体行为可视化



② 随着行为相似度增加，人类标注正确率降低

➢ 科学问题：

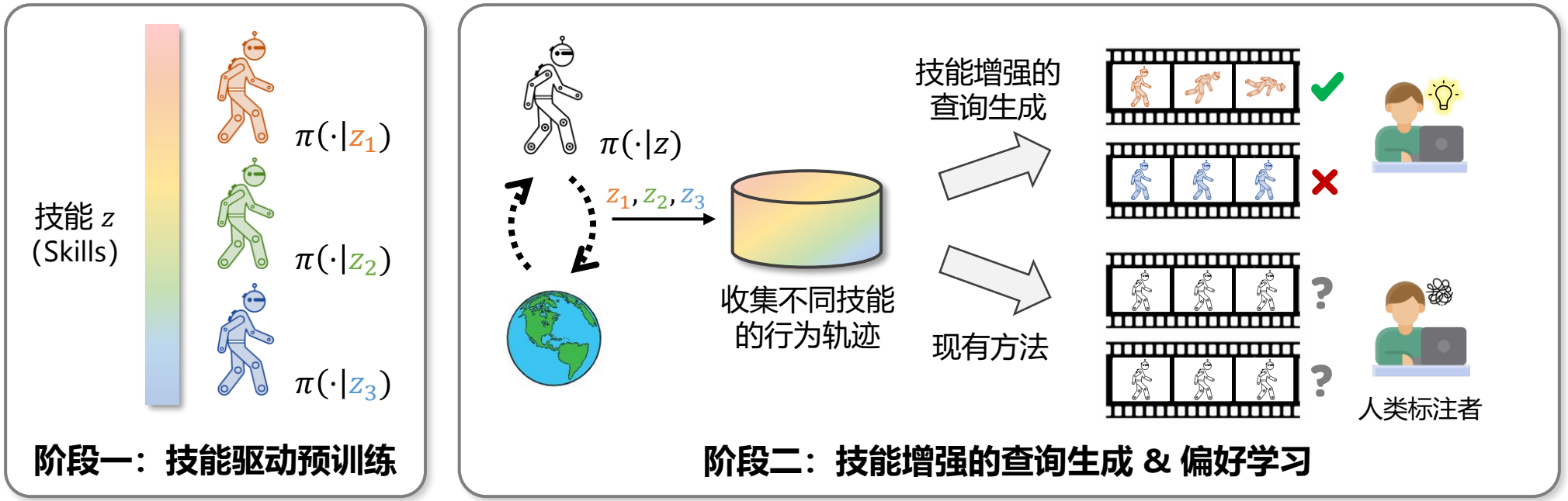
如何生成并选择具有 **行为多样性与高区分度** 的轨迹片段，从源头提升人类标注准确率？

➢ 难点：

- 如何在无奖励信号的情况下，**自主学习出多样且有意义的行为模式**？
- 如何将学习到的行为有效融入到偏好查询机制中，确保轨迹片段的可区分性？

3.1 研究内容一 解决行为空间的不可区分性

- **拟采用的方法：技能驱动的偏好强化学习算法**
  - **1. 技能驱动的预训练**：利用无监督强化学习中的技能发现（Skill Discovery）方法，在无奖励环境下学习一组多样化的技能（skill）。
  - **2. 技能增强的查询生成**：在生成待比较的轨迹片段时，主动选择**包含不同技能**的片段，确保行为模式存在差异，同时考虑查询不确定性，从而提高人类反馈的区分度。



3.1 研究内容一 解决行为空间的不可区分性

现有初步研究成果：

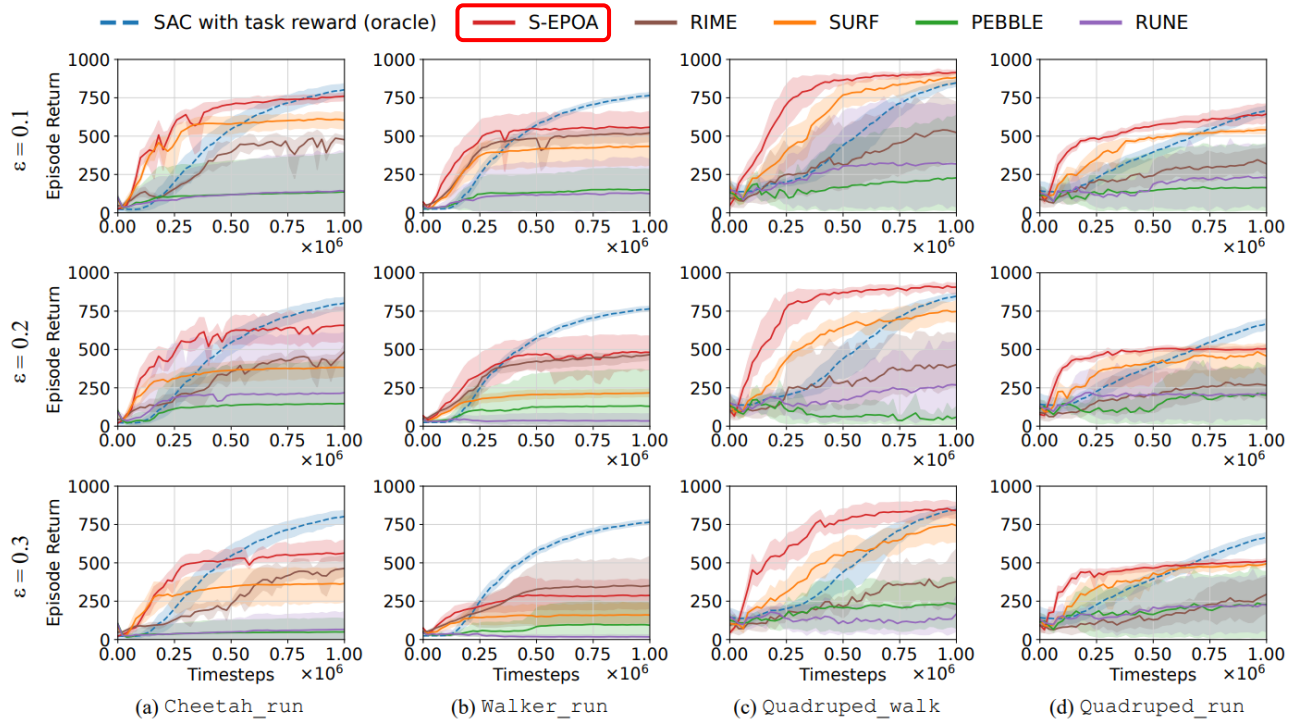
算法框图

Algorithm 1 S-EPOA Framework

Require: frequency of feedback K, number of queries M per feedback session, total feedback number N_{total}

Initialize $\pi(a|s,z) \leftarrow$ UNSUPERVISED PRETRAIN
for each iteration **do**
  if iteration % K == 0 and total feedback $< N_{\text{total}}$ **then**
    Update trajectory estimator $R_\theta(z)$ based on Eq. 6
    for m in $1 \cdots M$ **do**
      $(\sigma_0, \sigma_1) \sim$ QUERY SELECTION
      Query the teacher for preference label y
      Store preference $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\sigma_0, \sigma_1, y)\}$
    end for
    Update the reward model $\hat{r}_\psi$ using $\mathcal{D}$
    Relabel replay buffer $\mathcal{B}$ using $\hat{r}_\psi$
  end if
  if the current episode ends **then**
    Update z_{task} based on Eq. 9
  end if
  Obtain action $a_t \sim \pi(a|s_t, z_{\text{task}})$ and next state s_{t+1}
  Store transitions $\mathcal{B} \leftarrow \mathcal{B} \cup \{(s_t, a_t, s_{t+1}, \hat{r}_\psi(s_t, a_t))\}$
  Sample transitions $(s, a, s', \hat{r}_\psi)$ from $\mathcal{B}$
  Minimize $\mathcal{L}_{\text{critic}}$ and $\mathcal{L}_{\text{actor}}$ with $\hat{r}_\psi$
end for

强化学习标准测试环境上测试结果



学术界首个解决“行为不可区分”问题的工作

性能超过 SOTA 偏好强化学习方法
(RIME (2024)、SURF (2022)、PEBBLE (2021)、RUNE (2022))

[1] **Ni Mu***, Yao Luan*, Yiqin Yang, Bo Xu, Qing-Shan Jia. "S-EPOA: Overcoming the Indistinguishability of Segments with Skill-Driven Preference-Based Reinforcement Learning." International Joint Conferences on Artificial Intelligence (**IJCAI**), 2025.

3.2 研究内容二 解决人类判断的优劣不确定性

➢ **挑战：**

- 即使片段行为不同，人类也可能因无法清晰判断哪个行为更好 [31]，给出模糊 / 随机反馈。
- 现有方法缺乏对“**哪些查询会让人类感到犹豫**”的建模，导致反馈效率低下。

➢ **科学问题：**

如何建模人类判断机制，**主动筛选掉“优劣难断”的模糊样本**，优先选择对比鲜明的查询对？

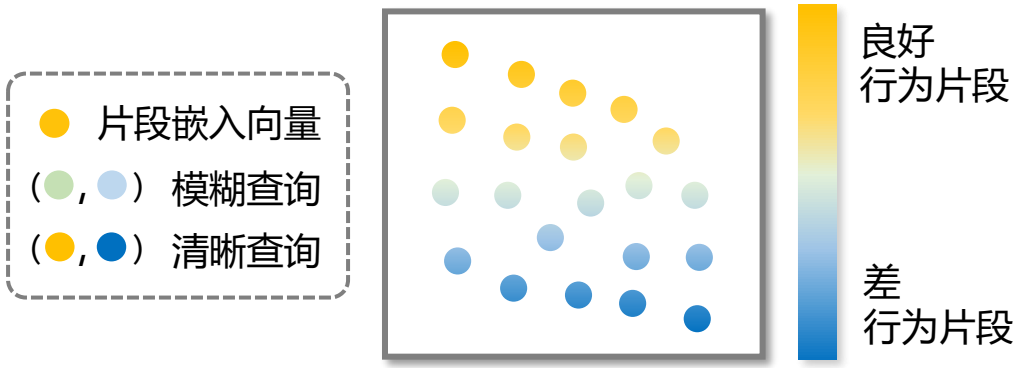
➢ **难点：**

- 如何量化一个查询的“**清晰度 / 模糊度**”？
- 如何凭借已有的偏好数据来学习这种判断，并随着偏好数据收集，不断更新判断模型？

➢ **解决方案：**

将行为映射到一个 **高维嵌入空间** 中；

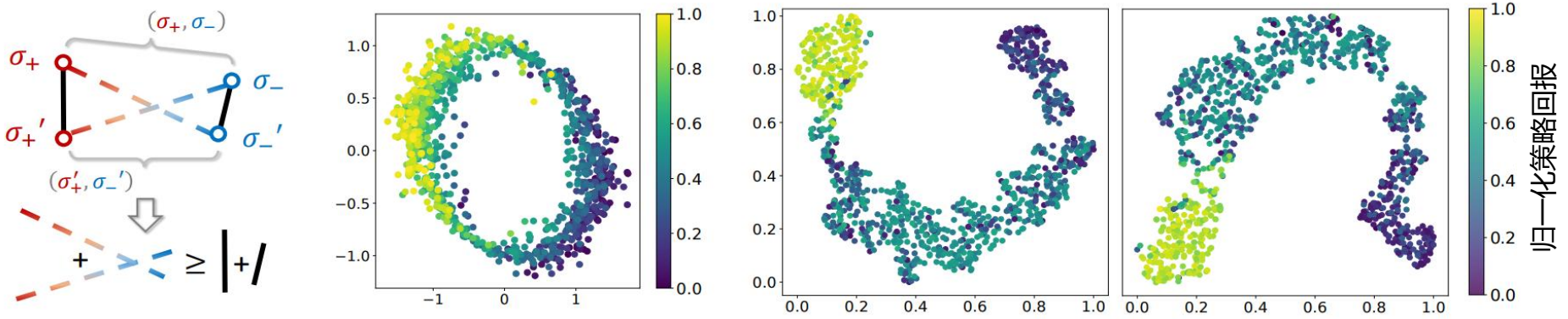
其中，清晰的查询距离远，模糊的查询距离近。



高维嵌入空间示意图

3.2 研究内容二 解决人类判断的优劣不确定性

- **拟采用的方法：基于对比学习的偏好强化学习算法**
  - **1. 表示学习**：通过对比学习，训练轨迹编码器，将偏好信息融入轨迹的嵌入表示（embedding）中。核心是两种 **对比学习损失函数**：
    - **模糊度损失** L_{amb}：鼓励模糊查询的片段在嵌入空间中靠近，清晰查询的片段远离。
    - **四边形损失** L_{quad}：构建片段间的四边形关系，在嵌入空间中建模优劣结构。
  - **2. 查询选择**：在学到的嵌入空间中，基于**拒绝采样**，优先选择嵌入距离远（即清晰可辨）的查询对，从而提高人类反馈的清晰度。



四边形损失 L_{quad} 主要思想：鼓励同类别样本（均为正样本 / 负样本）靠近，不同类别样本（一正一负）远离。
初步实验结果：同时使用 L_{amb} 和 L_{quad}，可学出具有语义结构的嵌入空间，好行为的嵌入向量连贯过渡到坏行为。

3.2 研究内容二 解决人类判断的优劣不确定性

现有初步研究成果：

核心方法

对比学习损失函数

$$\min_{\phi} \mathcal{L}_{\text{amb}} = \Big[- \mathop{\mathbb{E}}_{\substack{(\sigma_0,\sigma_1,p)\sim D_p,\\ p\in\{0,1\}}} \ell(z_0, z_1) + \mathop{\mathbb{E}}_{\substack{(\sigma_0,\sigma_1,p)\sim D_p,\\ p=\text{no\_cop}}} \ell(z_0, z_1) \Big], \tag{5}$$

$$\min_{\phi} \mathcal{L}_{\text{quad}} = - \mathop{\mathbb{E}}_{((\sigma_+,\sigma_-),(\sigma'_+,\sigma'_-))\sim D_p} \Big[\ell(z^+, z^{-\prime}) + \ell(z^{+\prime}, z^-) - \ell(z^+, z^{+\prime}) - \ell(z^-, z^{-\prime}) \Big], \tag{6}$$

拒绝采样方法

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

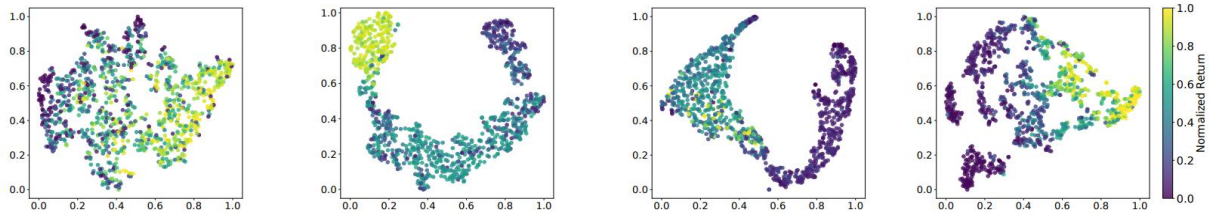
$$\rho_2(d_{\text{emb}}) = \frac{\rho_{\text{clr}}(d_{\text{emb}})/\rho_{\text{amb}}(d_{\text{emb}})}{\int (\rho_{\text{clr}}(d')/\rho_{\text{amb}}(d'))dd'}, \tag{7}$$

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

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

强化学习标准测试环境上测试结果

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



(a) dial-turn (b) drawer-open (c) hammer (d) walker-walk

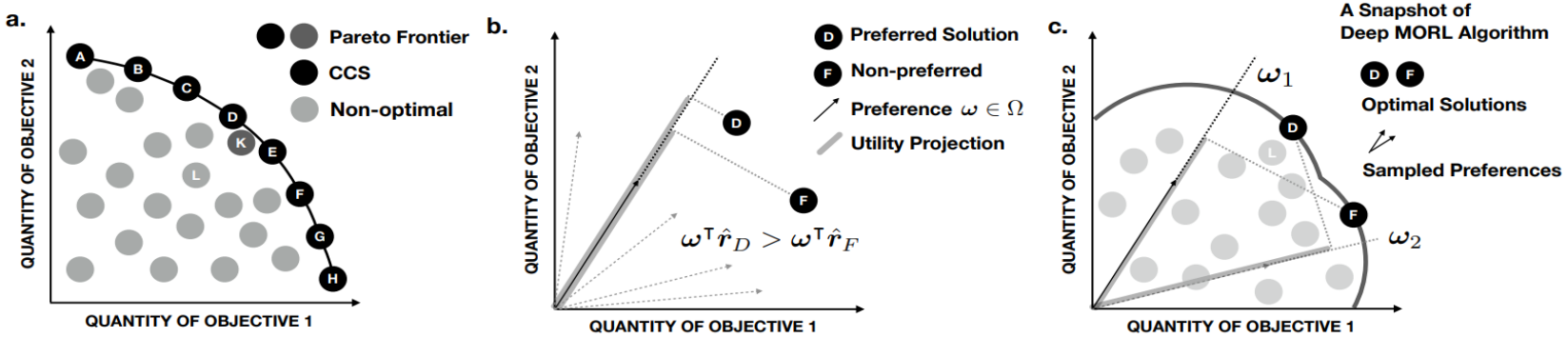
性能超过 SOTA 离线偏好强化学习方法
(OPRL (2023)、PT (2023)、OPPO (2023)、LiRE (2024))

[1] **Ni Mu***, Hao Hu*, Xiao Hu, Yiqin Yang, Bo Xu, Qing-Shan Jia. "CLARIFY: Contrastive Preference Reinforcement Learning for Untangling Ambiguous Queries." Forty-Second International Conference on Machine Learning (**ICML**), 2025.

3.3 研究内容三 解决多目标决策的内在冲突性

➢ 挑战：

- 现实任务往往具有多个目标，如平衡安全、效率、成本。
- 在多目标冲突下，人类难以给出单一偏好信号 [16]；现有方法无法建模人类的多目标意图。



经典多目标强化学习算法，使用向量奖励建模多目标

然而，现有偏好强化学习方法**缺乏对多目标意图的建模**

➢ 科学问题：

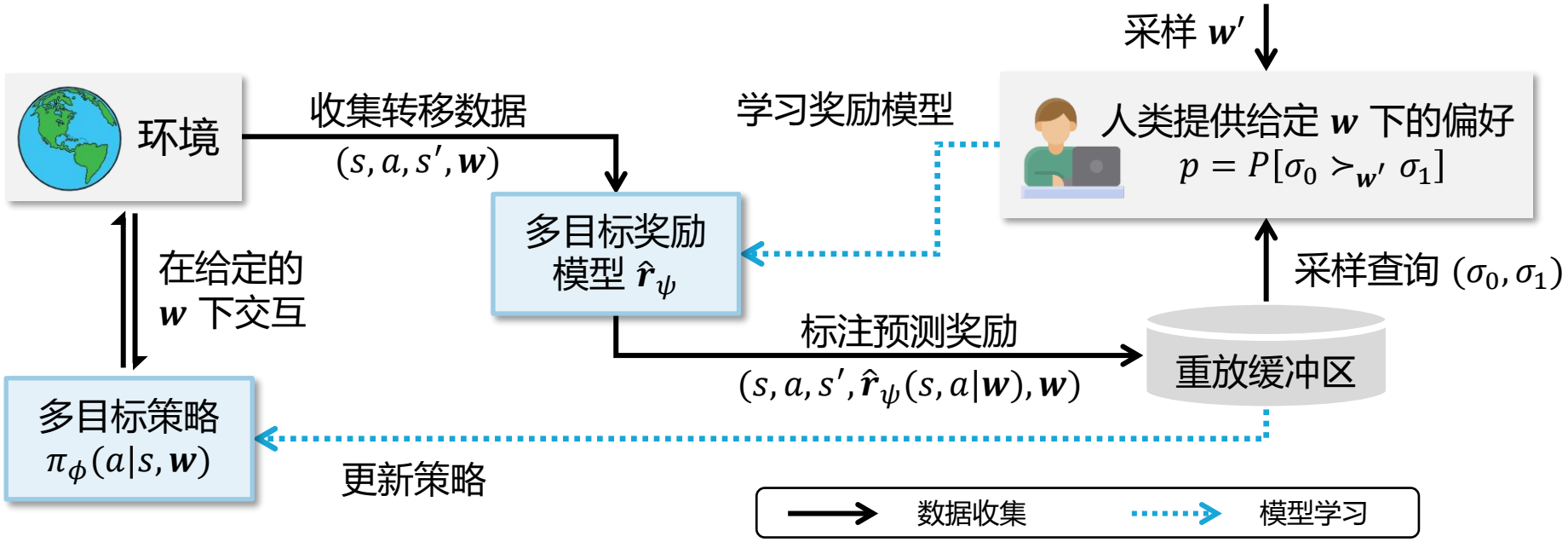
如何建模多目标意图判断，以学习一组能覆盖不同目标权衡的 **帕累托最优策略**？

➢ 难点：

- 如何将单目标的偏好比较范式扩展到多目标场景？
- 如何从不同目标权衡下的局部偏好中，恢复出整个帕累托前沿？

3.3 研究内容三 解决多目标决策的内在冲突性

- **拟采用的方法：基于偏好的多目标强化学习算法**
  - **1. 多目标偏好建模**：在偏好数据中引入权重向量 $\boldsymbol{w}$，用于建模不同目标权衡情况，然后建立基于 Bradley-Terry 模型的多目标奖励模型。
    - **理论保证**：在帕累托前沿为凸时，学习最大化该奖励模型的策略，**等价于学习整个帕累托前沿。**
  - **2. 策略学习**：将学到的多目标奖励模型与多目标强化学习算法结合，训练以权重向量 $\boldsymbol{w}$ 为条件的策略 $\pi(a|s,\boldsymbol{w})$，从而解决多目标决策的内在冲突性。



3.3 研究内容三 解决多目标决策的内在冲突性

现有初步研究成果：

初步理论结果

可求解策略帕累托前沿

Theorem 1. *Each policy in the policy set* $\pi^*(a|s,\boldsymbol{w}) \in \Pi^*$ *derived from Algorithm 1 is in the Pareto frontier when the segment length* $H \to \infty$.

Theorem 2. *Algorithm 1 obtains the entire convex Pareto frontier, i.e.,* Π^* *is the entire convex Pareto frontier.*

Theorem 4. *If the reward model* $\hat{\boldsymbol{r}}$ *is perfectly aligned with the teacher's preferences, that is, for segments* (σ_0, σ_1) *with arbitrary length* H,

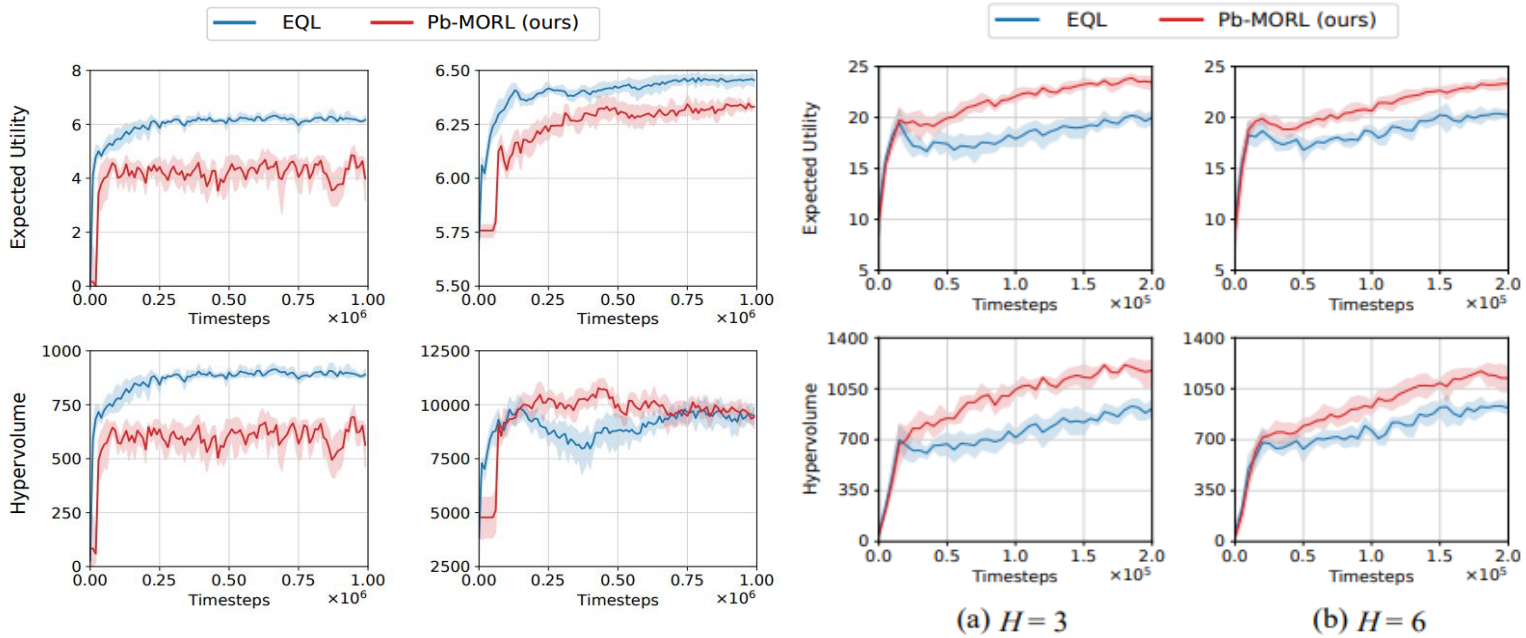
$$\sigma_0 \succ_{\boldsymbol{w}} \sigma_1 \iff \sum_{(s,a)\sim\sigma_0} \gamma^t \boldsymbol{w}^T \hat{\boldsymbol{r}}(s_t,a_t) > \sum_{(s,a)\sim\sigma_1} \gamma^t \boldsymbol{w}^T \hat{\boldsymbol{r}}(s_t,a_t). \quad (15)$$

Then, under a given weight vector $\boldsymbol{w}$, *maximizing the discounted return*

$$J(\pi) = \sum_{t=0}^{\infty} \gamma^t \boldsymbol{w}^T \hat{\boldsymbol{r}}(s_t,a_t) \quad (17)$$

is equivalent to selecting the optimal policy $\pi^*(\cdot|\cdot,\boldsymbol{w})$.

多目标强化学习标准测试环境上测试结果



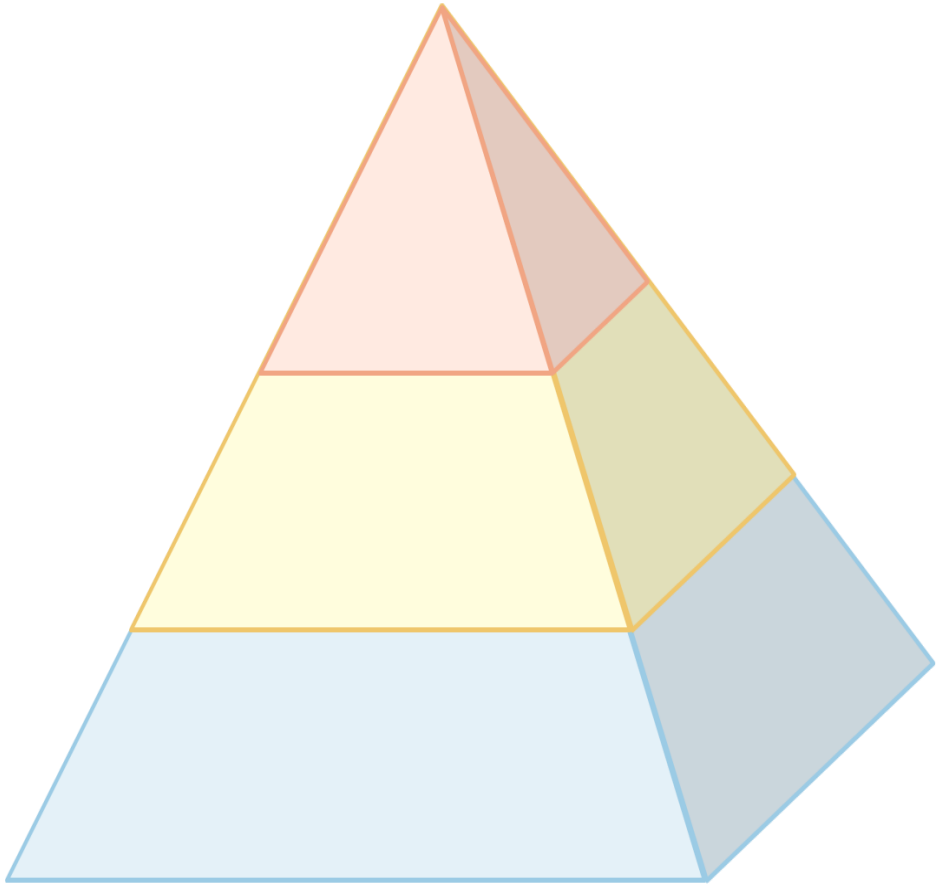
在（左）标准多目标强化学习测试环境（右）多能源管理系统上，与基于先验奖励函数的 EQL 方法持平

[1] **Ni Mu***, Yao Luan*, Qing-Shan Jia. Preference-based Multi-Objective Reinforcement Learning. **IEEE Transactions on Automation Science and Engineering**, vol. 22, pp. 18737-18749, 2025.

[2] **Ni Mu**, Yao Luan, Qing-Shan Jia. Preference-based Multi-Objective Reinforcement Learning with Explicit Reward Modeling. China Automation Congress (CAC), pp. 4874-4879, IEEE, 2024.

3. 研究内容

研究内容总结



高层：**解决多目标决策的内在冲突性**

技术路线：提出**多目标偏好强化学习范式**，建模多目标意图判断，学习帕累托最优策略

中层：**解决人类判断的优劣不确定性**

技术路线：基于**对比学习方法**，学习行为的高维嵌入空间，建模人类判断清晰度

底层：**解决行为空间上的不可区分性**

技术路线：利用**技能发现方法**，生成具有高区分度的行为，从源头提升人类标注准确率

组成 逻辑自洽、层层递进 的三部分研究内容

目 录

4. 研究基础 初步研究成果

研究内容	初步研究成果
解决行为空间内的不可区分性 **拟采用的方法：技能驱动的偏好强化学习方法**	[1] **Ni Mu*** , Yao Luan* , Yiqin Yang, Bo Xu, Qing-Shan Jia. S-EPOA: Overcoming the Indistinguishability of Segments with Skill-Driven Preference-Based Reinforcement Learning. **IJCAI 2025 (CCF-A)**. [2] Yao Luan*, **Ni Mu***, Hanfei Ge, Yiqin Yang, Bo Xu, Qing-Shan Jia. COMPASS: Training-Free Guidance for Skill Discovery Human Feedback. Submitted to **ICLR**, under review.
解决人类判断的优劣不确定性 **拟采用的方法：基于对比学习的偏好强化学习方法**	[3] **Ni Mu*** , Hao Hu* , Xiao Hu, Yiqin Yang, Bo Xu, Qing-Shan Jia. CLARIFY: Contrastive Preference Reinforcement Learning for Untangling Ambiguous Queries. **ICML 2025 (CCF-A)**. [4] Yao Luan* , **Ni Mu*** , Yiqin Yang, Bo Xu, Qing-Shan Jia. STAIR: Addressing Stage Misalignment through Temporal-Aligned Preference Reinforcement Learning. **NeurIPS 2025 (CCF-A)**.
解决多目标决策的内在冲突性 **拟采用的方法：基于偏好的多目标强化学习方法**	[5] **Ni Mu*** , Yao Luan* , Qing-Shan Jia. Preference-based Multi-Objective Reinforcement Learning. **IEEE Transactions on Automation Science and Engineering,** vol. 22, pp. 18737-18749, 2025. [6] **Ni Mu**, Yao Luan, Qing-Shan Jia. Preference-Based Multi-Objective Reinforcement Learning with Explicit Reward Modeling. 2024 China Automation Congress (CAC). [7] **Ni Mu***, Yao Luan*, Qing-Shan Jia. MAGPIE: Utilizing Agent-Specific Preferences for Multi-Agent Reinforcement Learning. 2025 China Automation Congress (CAC).

4. 研究基础 项目及平台支持

研究内容

解决行为空间内的不可区分性

拟采用的方法：技能驱动的偏好强化学习方法

解决人类判断的优劣不确定性

拟采用的方法：基于对比学习的偏好强化学习方法

解决多目标决策的内在冲突性

拟采用的方法：基于偏好的多目标强化学习方法

项目及平台支持

- 国家杰出青年基金（62125304）
  网络化信息物理融合能源系统的优化理论与方法
- 国家自然科学基金面上项目（62073182）
  离散事件动态系统的安全策略优化
- 国家自然科学基金面上项目（62192751）
  含氢多能源供需系统建模与智能性设计方法
- 北京市自然科学基金-小米创新联合基金课题（L233005）
  基于机理-数据双驱动的云边协同资源调度关键技术研究
- 中国联通企业项目
  运维及优化阶段空调系统数字孪生应用
- 国家重点研发计划资助项目（2022YFA1004600）
  新能源电力系统若干关键技术的数学理论与算法

4. 研究基础 已有研究成果 (1)

已发表 12 篇论文，其中 第一 / 共一作者 10 篇，CCF-A 类会议 3 篇，期刊论文 2 篇

以 第一 / 共一作者 发表 10 项工作

[1] **Ni Mu*** , Hao Hu* , Xiao Hu, Yiqin Yang, Bo Xu, Qing-Shan Jia. CLARIFY: Contrastive Preference Reinforcement Learning for Untangling Ambiguous Queries. **ICML 2025 (CCF-A)**.

[2] **Ni Mu*** , Yao Luan* , Yiqin Yang, Bo Xu, Qing-Shan Jia. S-EPOA: Overcoming the Indistinguishability of Segments with Skill-Driven Preference-Based Reinforcement Learning. **IJCAI 2025 (CCF-A)**.

[3] **Ni Mu*** , Yao Luan* , Qing-Shan Jia. Preference-based Multi-Objective Reinforcement Learning. **IEEE Transactions on Automation Science and Engineering**, vol. 22, pp. 18737-18749, 2025.

[4] Yao Luan* , **Ni Mu*** , Yiqin Yang, Bo Xu, Qing-Shan Jia. STAIR: Addressing Stage Misalignment through Temporal-Aligned Preference Reinforcement Learning. **NeurIPS 2025 (CCF-A)**.

[5] **Ni Mu**, Yao Luan, Qing-Shan Jia. Safety-Guaranteed Policy Composition Via Generalized Policy Improvement for Autonomous Vehicles. IEEE 21st International Conference on Automation Science and Engineering, pp. 2450-2455. IEEE, 2025. ***Outstanding WiRA Student Paper Award***.

[6] **Ni Mu**, Xiao Hu, et al. Large-scale data center cooling control via sample-efficient reinforcement learning. IEEE 20th International Conference on Automation Science and Engineering, pp. 2780-2785. IEEE, 2024.

[7] **Ni Mu***, Yao Luan*, Qing-Shan Jia. MAGPIE: Utilizing Agent-Specific Preferences for Multi-Agent Reinforcement Learning. 2025 China Automation Congress (CAC).

[8] **Ni Mu**, Zeyuan Liu, Yuhang Zhu, Qing-Shan Jia. LLM-Empowered Knowledge Graph Construction for Optimization Formalization in Smart Manufacturing. 2025 China Automation Congress (CAC).

[9] **Ni Mu**, Yao Luan, Qing-Shan Jia. Preference-Based Multi-Objective Reinforcement Learning with Explicit Reward Modeling. 2024 China Automation Congress (CAC).

[10] **Ni Mu**, Xiao Hu, Qing-Shan Jia. Integrating mechanism and data: Reinforcement learning based on multi-fidelity model for data center cooling control. 2023 China Automation Congress (CAC).

4. 研究基础 已有研究成果 (2)

以 第二作者 发表 2 项工作

[11] Hanchen Zhou, **Ni Mu**, Qing-Shan Jia. A Transformer-based Thermal Surrogate Model for Cooling Control in Data Centers. **IEEE Robotics and Automation Letters, 2024**.

[12] Yao Luan, **Ni Mu**, Qing-Shan Jia. Addressing Coupling in Restless Multi-Armed Bandits by Finetuning Whittle Index. IEEE 21st International Conference on Automation Science and Engineering, pp. 2456-2461. IEEE, 2025.

已授权或公开发明专利 5 项

[13] **牟倪**, 栾垚, 贾庆山.一种基于多目标强化学习的策略生成方法及装置. 已授权.

[14] **牟倪**, 胡潇, 贾庆山, 贺晓, 朱旭. 一种基于强化学习的数据中心机房的控制方法及装置. 已公开.

[15] **牟倪**, 周翰辰, 贾庆山, 胡潇. 一种基于数字镜像的数据中心运行调试方法及装置. 已公开.

[16] **牟倪**, 贾庆山, 胡潇. 一种数据中心末端空调系统运行策略确定方法及装置. 已授权.

[17] 周翰辰, **牟倪**, 贾庆山. 基于时空特征的数据中心建模与运行策略确定方法及装置. 已公开.

已授权或完成软件著作权 3 项

[18] **牟倪**, 贾庆山. 基于 Transformer 的锡膏印刷质量智能预测与优化软件.

[19] **牟倪**, 栾垚, 贾庆山. 基于指数策略的手机测试调度仿真软件.

[20] 胡潇, 周翰辰, **牟倪**, 贾庆山. 基于离线强化学习和基因算法的数据中心冷却控制策略调优软件.

目 录

5.1 预期创新点

- **研究内容一：解决行为空间的不可区分性**
  - **技能驱动的偏好强化学习方法**
    - 提出基于无监督技能发现的预训练机制，在无奖励信号下自主发掘多样化行为模式
    - 设计技能增强的查询生成策略，从源头提升轨迹片段的可区分性，显著降低人类标注噪声
- **研究内容二：解决人类判断的优劣不确定性**
  - **基于对比学习的偏好强化学习方法**
    - 构建融合偏好信息的轨迹嵌入空间，设计模糊度损失与四边形损失，建模人类判断的清晰度
    - 提出基于嵌入距离的拒绝采样机制，主动筛选对比鲜明的查询对，提升标注质量与效率
- **研究内容三：解决多目标决策的内在冲突性**
  - **基于偏好的多目标强化学习方法**
    - 引入权重向量建模多目标偏好，扩展 Bradley-Terry 模型至多目标场景
    - 提出偏好-目标耦合机制，学习覆盖帕累托前沿的条件策略，实现复杂人类意图的对齐

5.2 研究计划

时间节点	主要工作	预期成果
2026.04	完成技能驱动的偏好查询生成方法的构建与实验，撰写该部分论文	2 篇 A 类会议 1 项专利授权
2026.11	完成模糊查询筛选机制的研究与实验，撰写该部分论文	2 篇 A 类会议 1 项专利授权
2027.04	完成多目标偏好强化学习范式的构建与实验，撰写该部分论文	2 篇高质量期刊 1 项专利授权
2027.11	撰写博士学位论文；预答辩；答辩	博士论文

参考文献

1. Mu N, Hu X, Jia Q S, et al. Large-scale Data Center Cooling Control via Sample-efficient Reinforcement Learning[C]//2024 IEEE 20th International Conference on Automation Science and Engineering (CASE). IEEE, 2024: 2780-2785.
2. Mu N, Liu Z, Zhu Y, et al. LLM-Empowered Knowledge Graph Construction for Optimization Formalization in Smart Manufacturing[C]//2025 China Automation Congress (CAC). IEEE, 2025.
3. Zhu, Yuhang, et al. A Reinforcement Learning Embedded Surrogate Lagrangian Relaxation Method for Fast Solving Unit Commitment Problems[J]//IEEE Transactions on Power Systems (2025).
4. Lee E A. Cyber physical systems: Design challenges[C]//2008 11th IEEE international symposium on object and component-oriented real-time distributed computing (ISORC). IEEE, 2008: 363-369.
5. Baheti R, Gill H. Cyber-physical systems[J]. The impact of control technology, 2011, 12(1): 161-166.
6. Chu C C, Iu H H C. Complex networks theory for modern smart grid applications: A survey[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2017, 7(2): 177-191.
7. Silver D, Huang A, Maddison C J, et al. Mastering the game of go with deep neural networks and tree search[J]. nature, 2016, 529(7587): 484-489.
8. Evans R, Gao J. DeepMind AI Reduces Google Data Centre Cooling Bill by 40%[EB/OL]. 2016. https://deepmind.google/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-b y-40/.
9. Dewey D. Reinforcement learning and the reward engineering principle[C]//2014 AAAI Spring Symposium Series. 2014.
10. Lee K, Smith L M, Abbeel P. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training[C]//International Conference on Machine Learning. PMLR, 2021: 6152-6163.
11. Christiano P F, Leike J, Brown T, et al. Deep reinforcement learning from human preferences [J]. Advances in neural information processing systems, 2017, 30.
12. Guo, D., Yang, D., Zhang, H. et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning[J]. Nature 645, 633–638 (2025).
13. H. Zhan, F. Tao and Y. Cao, Human-Guided Robot Behavior Learning: A GAN-Assisted Preference-Based Reinforcement Learning Approach[J]. IEEE Robotics and Automation Letters, vol. 6, no. 2, pp. 3545-3552, April 2021, doi: 10.1109/LRA.2021.3063927.
14. S. Holk, D. Marta and I. Leite, POLITE: Preferences Combined with Highlights in Reinforcement Learning[C]//2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 2024, pp. 2288-2295, doi: 10.1109/ICRA57147.2024.10610505.
15. Hejna III, Donald Joseph, and Dorsa Sadigh. Few-shot preference learning for human-in-the-loop rl[C]//Conference on Robot Learning. PMLR, 2023.
16. Mu N , Luan Y, Jia Q S. Preference-based Multi-Objective Reinforcement Learning[J]. IEEE Transactions on Automation Science and Engineering, 2025, 22: 18737-18749.
17. Luan Y , Mu N , Yang Y, et al. STAIR: Addressing Stage Misalignment through Temporal-Aligned Preference Reinforcement Learning[C]//The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS). 2025.
18. Shin D, Dragan A D, Brown D S. Benchmarks and algorithms for offline preference-based reward learning[A]. 2023.
19. Biyik E, Huynh N, Kochenderfer M, et al. Active preference-based gaussian process regression for reward learning[C]//Robotics: Science and Systems. 2020.
20. Hu X, Li J, Zhan X, et al. Query-policy misalignment in preference-based reinforcement learning[A]. 2023.
21. Park J, Seo Y, Shin J, et al. Surf: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning[A]. 2022.
22. Choi H, Jung S, Ahn H, et al. Listwise reward estimation for offline preference-based reinforcement learning[A]. 2024.
23. Kim C, Park J, Shin J, et al. Preference transformer: Modeling human preferences using transformers for rl[A]. 2023.
24. An, Gaon, et al. Direct preference-based policy optimization without reward modeling[J]. Advances in Neural Information Processing Systems 36 (2023): 70247-70266.
25. Kang Y, Shi D, Liu J, et al. Beyond reward: Offline preference-guided policy optimization[A]. 2023.
26. Liu R, Bai F, Du Y, et al. Meta-reward-net: Implicitly differentiable reward learning for preference-based reinforcement learning[C/OL]//Oh A H, Agarwal A, Belgrave D, et al. Advances in Neural Information Processing Systems. 2022.
27. Lee K, Smith L, Dragan A, et al. B-pref: Benchmarking preference-based reinforcement learning[A]. 2021.
28. Cheng J, Xiong G, Dai X, et al. Rime: Robust preference-based reinforcement learning with noisy preferences[A]. 2024.
29. Heo J, Lee Y J, Kim J, et al. Mixing corrupted preferences for robust and feedback-efficient preference-based reinforcement learning[J]. Knowledge-Based Systems, 2025, 309: 112824.
30. Xue W, An B, Yan S, et al. Reinforcement learning from diverse human preferences[A]. 2023.
31. Mu N, Hu H, Hu X, et al. CLARIFY: Contrastive Preference Reinforcement Learning for Untangling Ambiguous Queries[C]//Forty-second International Conference on Machine Learning. 2025.

请各位老师批评指正！

培养单位 ：自动化系
研 究 生 ：牟倪
学 号 ：2023310855
指导教师 ：贾庆山 教授

答辩日期 ：2025年11月27日

智能与网络化系统研究中心
Center for Intelligent and Networked Systems