项目九：多智能体逆强化学习与对比实验最终报告

摘要

本项目围绕多智能体逆强化学习（reward recovery）开展系统复现与对比实验，完成了三个任务链路：T1（GridWorld，MA-SPI 数据）、T2（MPE simple_spread，PPO 数据）与 T3（MultiWalker，PPO 数据）。

项目的核心问题有两条：

第一，在 T2 的 PPO 学习者数据上，使用基于 MA-SPI 假设的 MA-LfL 作为"错配基线"时，常用的 1b KL 指标是否仍然可信；

第二，I-LOGEL（Stage A）与 I-LOLA（Stage B）在 KL 指标上是否呈现可解释差异。

所有实验在 seeds=0..4 上运行，并完成了统一汇总与可视化。

主要发现是：

- T1 在匹配设定下 MA-LfL 的 KL 误差稳定很小；T2 中 mismatch 的 1a 误差很大，但 1b 误差反而接近 0，导致 $\mathrm{ratio} = \mathrm{err}_{1b}/\mathrm{err}_{1a}$ 极小，表明 1b 指标在该错配条件下会失真；
- 同时在当前数据生成强度下，T2 的任务层 induced 回报没有优于 random，PPO 基线本身也弱于 random。
- T3 中 I-LOLA 的 KL 误差与 induced 回报整体稳定，但 induced 与 baseline 差距很小。

---

1. 问题定义与动机

多智能体环境中的 reward recovery 面临两类困难：

- 其一，学习者之间相互影响，使得"从轨迹推回奖励"的问题更不适定；
- 其二，评估指标往往依赖对学习者更新规则的假设，一旦假设与真实学习过程不一致，指标可能给出误导性结论。

本项目选择对比三条链路来回答一个具体问题：当我们把"学习者更新规则"从匹配（T1 的 MA-SPI，或 T2/T3 的 I-LOLA 假设）切换到错配（T2 上用 MA-SPI 假设解释 PPO 学习者）时，KL 类指标与任务层证据是否仍然一致。

课程项目要求报告必须覆盖问题定义、数据/环境、方法、相关工作与评估（含定性与定量），并允许出现负结果，只要给出清晰分析。本项目的目标是把"能跑通"升级为"能给出可复现、可解释的对比结论"。

---

2. 环境、数据与实验配置

2.1 三个任务与数据来源

T1 使用 GridWorld 环境生成 MA-SPI 轨迹数据，适合做离散策略的 KL 评估与 reward heatmap 的直观检查。

T2 使用 MPE 的 simple_spread 环境生成 PPO 学习者数据，并在此基础上做 StageA/StageB/mismatch 的对比。

T3 使用 MultiWalker 的 PPO 数据，用于检验连续控制环境中的 KL 评估与 induced 训练是否能稳定运行。

所有数据按 seed 存为 `outputs/data/t{1,2,3}/t*_seed{seed}_*.pkl`。

在下游评估中，每个 metrics/induced 结果文件都会记录 `data_path` 与 seed 相关字段，用于追溯和一致性检查。

2.2 统一的运行流程与可复现性

本项目采用"每个 seed 一条完整流水线"的方式运行：先生成数据，再运行对应算法评估与 induced 训练，最后汇总统计并生成图表。完整的顺序与命令已经被记录在 `log.txt` 中，可作为复现实验的执行证据。

2.3 指标定义与统计口径

1a/1b 指标都使用 KL：
$\mathrm{KL}(p|q) = \sum_i p_i \log \frac{p_i}{q_i}$。
在本项目里，(p) 对应观测到的"真实下一步策略"π_{real}，q 对应由某种 lookahead/预测规则得到的 π_{pred}。

- err_{1a}：使用参考奖励（或固定的 proxy 基准）做 lookahead 得到的 KL，作为该模型假设下的"上界/天花板"误差。
- err_{1b}：使用恢复奖励 $\hat{W}$ 做 lookahead 得到的 KL，用于衡量恢复奖励是否能解释策略变化。

任务层评估使用 induced 回报：`R_random`、`R_ppo_best`、`R_induced_mean`、`R_induced_last`，评估 episodes 统一为不少于 20，以降低方差。

---

3. 方法概述

3.1 MA-LfL（T1 匹配；T2 mismatch 基线）

在 T1 中，MA-LfL 在匹配设定（与数据生成假设一致）下用于 sanity check：应当得到较小且稳定的 KL。
在 T2 中，MA-LfL 被用作"错配基线"：数据来自 PPO 学习者，但 lookahead 仍按 MA-SPI 的假设构造。其目的不是追求 MA-LfL 的最强实现，而是构造一个固定、可复现的错配对照，从而观察评估口径在错配时会发生什么。

T2 的 mismatch 还提供两类补强评估：cross-phase 与 holdout（train/test states 切分），用于避免"同一批 states 上一步拟合导致 1b 假小"的问题。

3.2 I-LOGEL（Stage A）与 I-LOLA（Stage B）

在 T2 中，Stage A（I-LOGEL）与 Stage B（I-LOLA）形成对比：Stage A 给出一个更基础的恢复/预测结果，Stage B 在此基础上进行进一步优化（本项目实现中用 CMA-ES 等方式搜索权重），从而观察两阶段在 KL 指标上的差异。T3 中同样运行 I-LOLA，用于连续控制环境下的稳定性检查。

3.3 Induced（任务层证据）

对 T2/T3，使用恢复的奖励 $\hat{W}$ 训练 induced policy，并与 random 与 `ppo_best` 对比。该评估用于回答：即使 KL 指标看起来很好，恢复奖励是否真的能带来任务层回报改善。

---

4. 相关工作（简述）

本项目所实现的方法属于多智能体 reward learning/逆强化学习的典型路线：通过对学习者策略更新的建模，反推出能够解释行为的奖励或偏好。

MA-LfL 与 MA-SPI 路线强调在给定学习者更新规则下，对“下一步策略变化”进行可解释建模；

I-LOGEL/I-LOLA 则以两阶段或带优化的方式进一步提高对策略变化的解释能力。

本项目与上述工作的关系是：在课程提供的框架与任务建议基础上完成端到端实现与复现实验，并在一个明确的错配条件（MA-SPI 假设对 PPO 数据）下系统检验 KL 指标与任务层证据是否一致。

---

5. 实验结果与分析

5.1 T1：GridWorld（匹配 sanity check）

在 5 个 seeds 上，T1 的 MA-LfL 得到的 KL 很小且稳定：
$\mathrm{err}_{1b} = (9.35 \times 10^{-6} \pm (6.54 \times 10^{-6})$。
对应的可视化包括 `t1_kl_seed0.png` 与 `t1_reward_heatmap.png`，用于从定性上检查 KL 与奖励形状是否符合预期。
该结果支持一个基本结论：在匹配设定下，MA-LfL 链路是可用的，评估口径不会自动失真。

5.2 T2：MPE simple_spread（核心对比）

5.2.1 1b（StageA vs StageB vs mismatch）的直接对比

在 5 个 seeds 上，T2 的 err_{1b} 统计为：

- Stage A（I-LOGEL）：$1.73 \times 10^{-7} \pm 7.13 \times 10^{-8}$
- Stage B（I-LOLA）：$5.86 \times 10^{-7} \pm 4.73 \times 10^{-7}$
- mismatch（MA-LfL）：$1.13 \times 10^{-7} \pm 1.47 \times 10^{-7}$

这组数字的关键现象是：仅看 1b，mismatch 并没有显著“更差”，反而与 Stage A 同量级，甚至更小。相关图表为 `summary_err1b_t2.png`，以及跨任务总览的 `summary_err1b.png`（log 轴）。

5.2.2 1a 与 ratio 揭示：mismatch 下 1b 指标失真

当引入 1a 后，现象发生根本变化。T2 的 err_{1a} 统计为：

- Stage A（I-LOGEL）：$1.73 \times 10^{-7} \pm 7.13 \times 10^{-8}$
- Stage B（I-LOLA）：$1.73 \times 10^{-7} \pm 7.13 \times 10^{-8}$
- mismatch（MA-LfL）：1.42 ± 1.32

mismatch 的 1a 在数量级上远大于 Stage A/B，这意味着：在错配假设下，基准口径已经显示“预测下一步策略”很困难。但与此同时，mismatch 的 1b 仍然停留在 10^{-7} 量级。将二者做比值 $\mathrm{ratio} = \mathrm{err}_{1b}/\mathrm{err}_{1a}$，得到：

- Stage A：ratio ≈ 1（同量级）
- Stage B：ratio $\approx 3.33 \pm 2.33$（同量级或略大）
- mismatch：ratio $\approx 4.80 \times 10^{-8} \pm 4.89 \times 10^{-8}$

这说明一个清晰结论：在 T2 的错配条件下，1b KL 会“贴地”，从而不能反映真实难度与误差差异；此时必须联合 1a 与 ratio 才能给出可信解释。对应图表为 `t2_err1a_summary.png` 与 `t2_ratio_summary.png`，其中 ratio 图包含 ratio=1 的参考线，便于读者判断“同量级”还是“塌缩”。

5.2.3 任务层 induced：当前设置下没有正向提升

T2 的 induced 回报（episodes≥20）在 5 个 seeds 上为：

- $R_{\text{random}} = -52.54 \pm 3.41$
- $R_{\text{ppo\_best}} = -95.51 \pm 29.72$
- $R_{\text{induced\_mean}} = -89.87 \pm 20.93$
- $R_{\text{induced\_last}} = -86.31 \pm 21.05$

图表 `summary_induced_t2.png` 与 `summary_induced_delta_t2.png` 显示：相对 random 的提升 Δ 为负，induced 并未改善任务回报。换句话说，在当前数据生成强度下，PPO 基线本身弱于 random（`ppo_best` 更差），因此 reward recovery 与 induced 很难在任务层体现正向收益。这是一条可解释的负结果：任务层证据并不支持"恢复奖励带来更好控制"，其原因更可能是学习者轨迹质量不足，而不是单纯的实现错误。

5.3 T3：MultiWalker（连续控制稳定性）

T3 的 I-LOLA 在 5 个 seeds 上得到：
$\text{err}_{1b} = 0.01276 \pm 0.00293$，且 $\text{err}_{1a} = 0$（按该任务的定义口径）。对应图表为 `summary_err1b_t3.png`。

任务层 induced 回报为：

- $R_{\text{random}} = 33.77 \pm 0.12$
- $R_{\text{ppo\_best}} = 33.76 \pm 0.19$
- $R_{\text{induced\_mean}} = 33.79 \pm 0.16$
- $R_{\text{induced\_last}} = 33.74 \pm 0.19$

`summary_induced_t3.png` 与 `summary_induced_delta_t3.png` 显示 induced 与 baseline 的差异很小，整体稳定，不存在明显退化，但也没有显著提升。这一部分更偏"工程可行性与稳定性展示"：在连续控制环境下，KL 评估与 induced 流水线能够稳定运行并给出一致统计。

---

6. 讨论、局限与可信度说明

第一，T2 的核心结论不是"谁的 1b 更小"，而是"mismatch 条件下 1b 会塌缩"。这是一条和评估口径直接相关的发现：如果只报告 1b，会得出"错配也不差"的误导性判断；而引入 1a 与 ratio 后，错配的困难度与退化会被清晰揭示。

第二，任务层证据在 T2 上没有给出正向提升，且 `ppo_best` 劣于 random。最合理的解释是：学习者轨迹质量不足，使得"恢复奖励并用其诱导训练"无法体现优势。该负结果并不影响项目价值，因为课程项目明确允许不成功的结果，只要分析清楚原因并给出讨论。

第三，本项目的统计结论只覆盖 5 个 seeds，并且 induced 的方差在 T2 上仍然较大。更强的统计需要更多 seeds 或更稳定的学习者基线，但在课程项目预算下，当前结果已经足以支撑"评估口径失真"的主要论点。

---

7. 结论

本项目完成了三任务（T1/T2/T3）的数据生成、算法评估、induced 训练与多 seed 汇总闭环，并在 T2 上给出一条清晰且可复现的发现：在 MA-SPI 假设拟合 PPO 学习者的错配条件下，1b KL 会塌缩，从而不能反映真实误差；必须结合 1a 与 ratio 才能得到可信结论。任务层 induced 在当前设置下没有优于 random，其主要原因更可能是 PPO 学习者轨迹质量不足。整体而言，项目达到"可复现、可解释、可对比"的交付目标，具备作为正式课程报告提交的完整证据链。

---

8. 后续改进建议（可选扩展，不作为当前交付硬要求）

如果要把 T2 的任务层证据进一步做强，最直接的改进是增强 PPO 数据生成强度，使得 $R_{\text{ppo\_best}} > R_{\text{random}}$。在此基础上重复 reward recovery 与 induced，对比 $\Delta) 是否由负转正，从而把"评估口径失真"的结论进一步延伸到"任务层是否能恢复"。

---

9. 个人贡献说明

本项目为团队完成。(这里写团队分工) 完成问题选择、实验设计、代码实现、debug 与修正、多 seed 批量运行、结果汇总可视化以及最终报告撰写。项目实现基于课程提供的工程框架与任务建议，在此基础上新增了 T2 mismatch 基线、Stage A vs Stage B 对比脚本、多 seed 聚合与统计作图等模块，并完成端到端可复现的输出组织与文档化。

附录 A：主要输出文件与图表索引

- 多 seed 汇总表：`outputs/summary/metrics_summary.csv`、`outputs/summary/induced_summary.csv`（以及对应的 stats/markdown 版本）
- 关键指标图：
  `outputs/plots/summary_err1b.png`（log 总览）、`outputs/plots/summary_err1b_t1.png`、
  `outputs/plots/summary_err1b_t2.png`、`outputs/plots/summary_err1b_t3.png`
  `outputs/plots/t2_err1a_summary.png`、`outputs/plots/t2_ratio_summary.png`
- 任务层 induced 图：
  `outputs/plots/summary_induced_t2.png`、`outputs/plots/summary_induced_delta_t2.png`
  `outputs/plots/summary_induced_t3.png`、`outputs/plots/summary_induced_delta_t3.png`
- 一键运行与证据：`log.txt` 记录了 seeds=0..4 的统一执行顺序与输出生成。
- 运行指令顺序

```
# 多轮seed循环生成结果
python scripts/run_all_seeds.py > log.txt 2>&1
# 聚合结果
python scripts/aggregate_results.py
# 生成报告与图表
python scripts/run_phase3_plots_reports.py
```