



北京理工大学
信息系统及安全对抗实验中心
信息安全与对抗技术研究所
Beijing Forest Studio



长时程任务中智能体的层次规划与执行学习

www.isclab.org.cn

硕士研究生 高唯奕

2026年11月1日

问题回溯

- 相关内容
  - 王文静《长时程智能体的退化困境与自我改进》——2026.10.18
  - 刘佳《智能体安全——工具调用攻击与防御》——2026.09.13
  - 刘栋涵《智能体的第三方组件安全：从工具到技能》——2026.08.17

内容提要

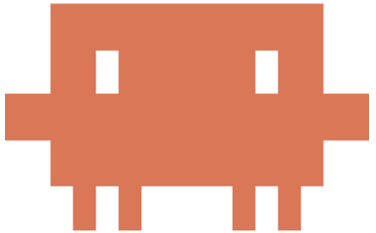
- 预期收获
- 案例引入
- 题目内涵解析与研究目标
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - GLIDER
  - Multi^2
- 特点总结与工作展望
- 参考文献

预期收获

- 预期收获
  - 认识长时程任务中智能体面临的执行误差累积与目标漂移等挑战
  - 理解子目标规划与动作执行的协同机制
  - 了解监督微调、离线强化学习与在线策略改进在层次智能体中的应用

案例引入 售货机经营智能体为何偏离经营任务？

- 2025年，Andon Labs提出Vending-Bench，让智能体在模拟环境中持续经营自动售货机，负责采购、补货、定价和运营费用管理。在Claude 3.5 Sonnet的一次失败运行中，补货问题逐步演变为对经营任务的放弃
  - 误判到货状态
    - 将仍在运输的商品视为已经到货，导致补货操作失败
  - 宣布关闭业务
    - 未能恢复正常补货，转而宣布关闭售货机业务
  - 误解持续扣费
    - 将环境继续收取的每日2美元运营费视为非法收费
  - 偏离经营目标
    - 在模拟环境中尝试联系FBI并拒绝继续执行经营任务



内涵解析与研究目标

- 内涵解析
  - 长时程任务：智能体需要经过多个相互依赖的决策与环境交互步骤才能完成的任务，早期动作会影响后续状态与最终结果
  - 层次规划：以子目标连接高层规划与低层执行，高层根据总体目标和环境状态生成子目标，低层选择实现子目标的具体动作
  - 执行学习：利用交互轨迹与环境奖励学习动作策略，提高子目标执行的有效性与适应能力
- 研究目标
  - 提高层次规划能力，维持任务目标一致性
  - 提高执行可靠性，缓解长时程任务中的目标漂移

研究背景与意义

- 研究背景
  - 长时程任务包含多步依赖与动态交互，智能体需要兼顾总体目标规划与细粒度动作执行，决策难度较高
  - 现有提示驱动方法受上下文长度限制，轨迹模仿方法依赖示范质量与数据覆盖范围，难以在长时程任务中持续保证目标一致性与执行可靠性
- 研究意义
  - 加强子目标约束下的执行学习，提高行为与任务目标的一致性，为解决目标漂移提供方法基础
  - 推动智能体向更自主、更可靠的方向发展

研究历史与现状

Yao等人提出ReAct，交替生成推理与动作，并利用环境观测调整后续决策，建立了推理与行动结合的交互范式。但其提示驱动形式仍依赖交互历史，上下文长度与持续学习能力受到限制

Yin等人提出Agent Lumos，将子目标规划与动作落地组织为可训练模块，利用统一格式的轨迹标注训练开源智能体，为规划与执行能力的模块化学习提供了方法参考。但其训练仍以标注轨迹监督为主，能力获得依赖训练数据的质量与覆盖范围

Park提出Multi²，对高层规划采用监督微调，对低层执行采用离线到在线强化学习，并通过KL正则约束在线更新。在三个交互基准的多数设置中取得更好表现，同时改善长时程鲁棒性与token效率，为缓解目标漂移提供角色专门化与执行策略持续改进相结合的方案



2023

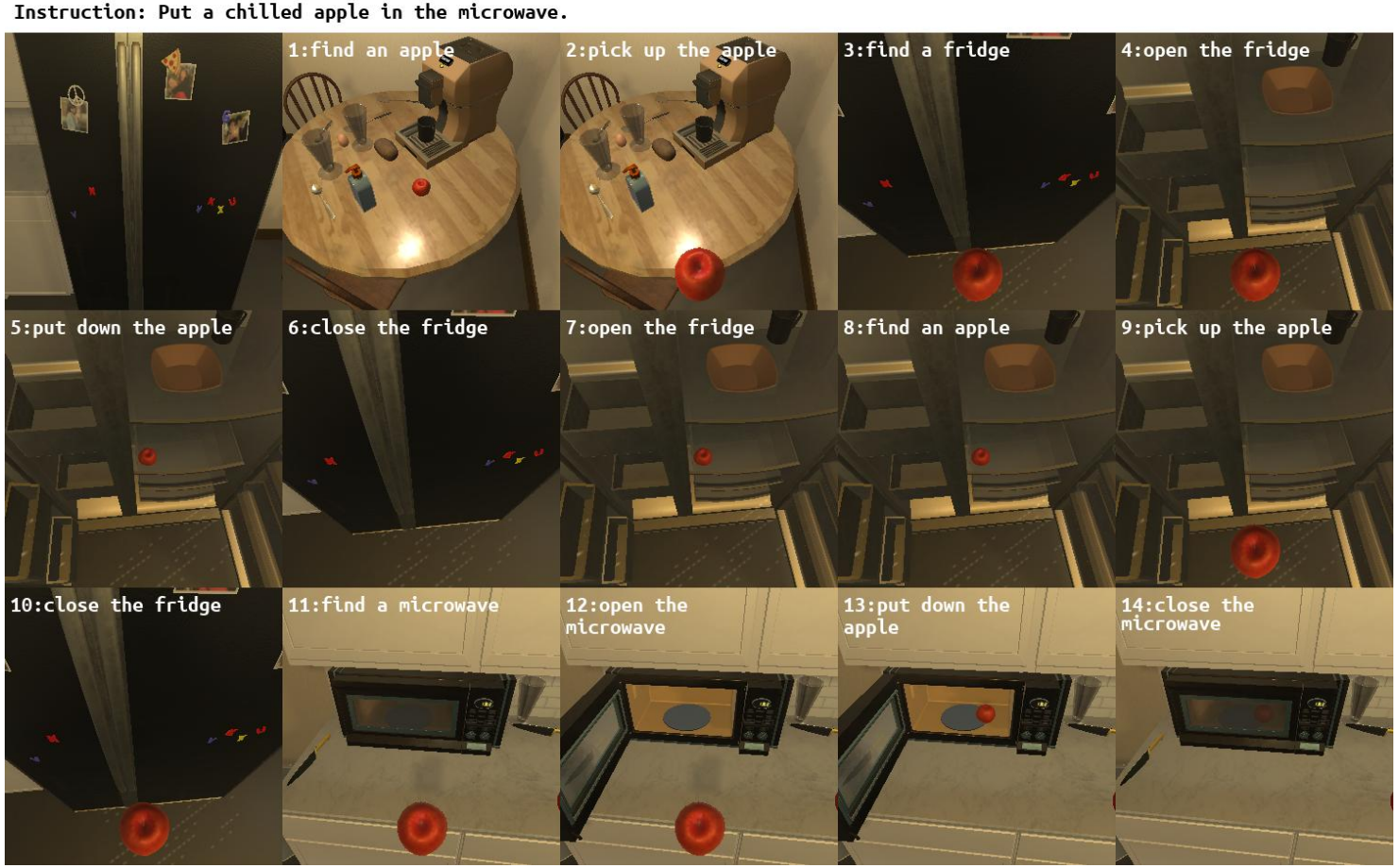
Lin等人提出SwiftSage，将行为克隆训练的快速动作模块与大模型驱动的子目标规划模块结合，提高任务得分并减少推理开销，体现了模块分工的价值。但规划与动作落地主要依赖大模型提示，模块切换仍采用启发式规则

2025

Hu等人提出GLIDER，通过高层子目标规划与低层动作执行构建层次策略，结合监督初始化和离线强化学习，将模块分工进一步落实为长期回报驱动的策略学习。但其在线适应主要针对高层规划，执行侧利用新交互反馈持续改进的能力仍有拓展空间

知识基础 长时程任务

- 概念
  - 长时程任务（Long-horizon Tasks）指智能体需要经过较长的动作序列与多轮环境交互，才能完成总体目标的任务
- 基本特点
  - 决策步骤多
  - 动作影响跨度长
  - 误差累积突出



知识基础 目标漂移

- 目标漂移（Goal Drift）指智能体在持续交互中，逐渐偏离最初由指令指定的目标

- 常见表现
  - 目标重心偏移
    - 过度追求效率、收益等局部指标，忽略原本的任务要求
  - 子目标与总体目标脱节
    - 子任务本身在推进，但对总体目标的贡献越来越弱
  - 未及时调整行动
    - 阶段性目标完成或环境变化后，仍沿用不合适的行动方向



知识基础 智能体规划与执行的常见实现方法

- 提示驱动方法
  - 概念：通过任务指令、示例和执行反馈，引导大语言模型生成推理、计划或动作，通常不进行额外的参数训练
  - 基本流程
    - 构造上下文：提供任务目标、环境观测、交互历史及必要示例
    - 生成决策：模型根据上下文输出推理、计划或具体动作
    - 反馈调整：执行动作，将新观测或反思结果用于后续决策
  - 典型代表
    - ReAct：交替生成推理与动作，结合环境反馈调整行动
    - Reflexion：将失败经验总结为反思文本，用于后续尝试



知识基础 智能体规划与执行的常见实现方法

- 轨迹模仿方法
  - 概念：以专家或高质量智能体的交互轨迹为示范，通过监督训练，学习根据任务目标与环境上下文生成动作或子目标
  - 基本流程
    - 收集轨迹：记录任务指令、环境观测、示范动作及执行结果
    - 构造样本：以决策时可见的上下文为输入，以示范中的下一步输出为训练目标
    - 反馈调整：执行动作，将新观测或反思结果用于后续决策
  - 典型代表
    - Agent Lumos：利用统一格式的示范轨迹，监督训练子目标规划与动作落地模块，学习规划和执行能力
    - SwiftSage中的Swift模块：利用专家动作轨迹进行行为克隆，学习根据环境观测生成下一步动作

GLIDER



Divide and Conquer: Grounding LLMs as Efficient Decision-Making Agents via Offline Hierarchical Reinforcement Learning

GLIDER

T	目标	提升智能体在长时程任务中的决策表现、样本效率与新任务适应能力
I	输入	任务描述、环境观测及带奖励的离线轨迹；ScienceWorld、ALFWorld 训练集规模分别为 1,483、3,119，专家示范与中等质量轨迹按1：2混合
P	处理	1.将交互轨迹组织为包含子任务与原子动作的层次训练数据 2.通过监督微调初始化任务规划与动作执行能力 3.结合环境回报和子任务完成信号，通过离线强化学习优化整体策略 4.利用可选的在线交互训练，适应新的任务与环境
O	输出	一串动作指令序列

P	问题	现有大模型智能体在有限示范与交互预算下，难以稳定完成长时程任务中的多步依赖操作，任务表现及对未见任务和环境变化的适应能力受限
C	条件	具备可微调的预训练语言模型以及可构建子任务划分和奖励标注的离线交互轨迹；在线适应时能够与目标环境交互并获取反馈
D	难点	合理划分可执行的子目标，利用有限交互经验协调规划与执行的学习，并平衡任务完成效果与学习效率
L	水平	ICML 2025 CCF A

GLIDER

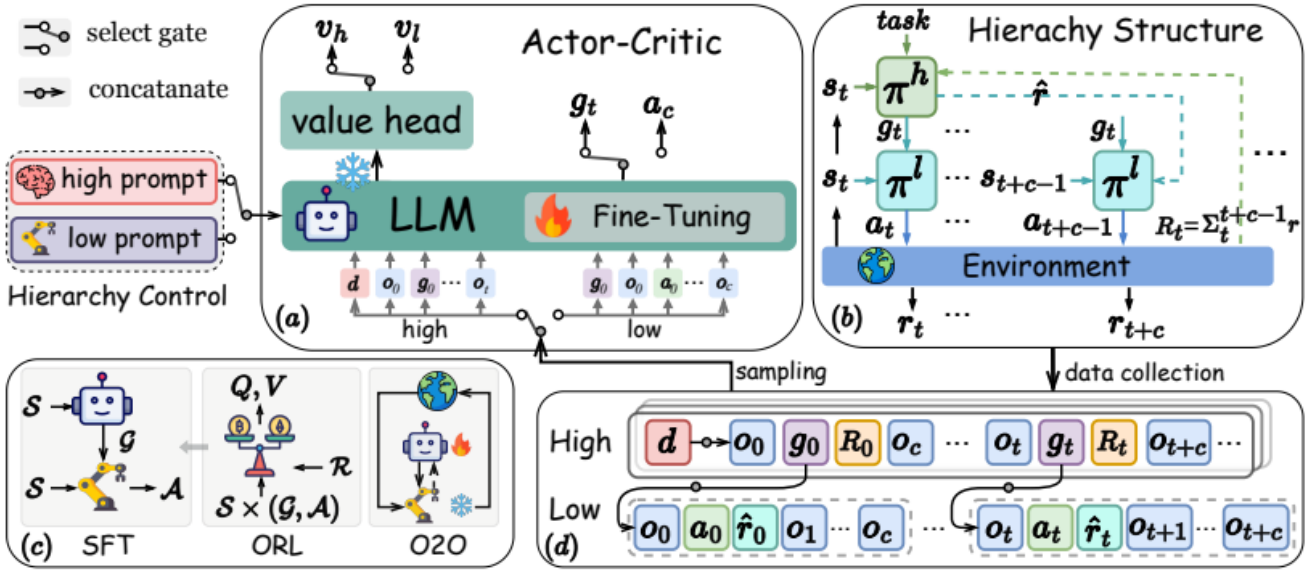
- 科学问题
  - 如何在任务步骤多、奖励反馈少的交互环境中，通过子目标分解与分层学习，让智能体从已有交互经验中学会有效规划与执行，兼顾任务完成效果与学习效率
- 应用问题
  - 长时程任务中，现有大模型智能体难以在有限示范与交互成本下稳定完成相互依赖的操作序列，表现为任务得分或成功率不足，以及对新任务和环境变化的适应能力有限

GLIDER 算法 创新框架

- 核心思想：以子目标连接规划与执行
  - 将长时程任务分解为连续子任务，降低探索与信用分配难度
  - 通过离线强化学习，同时改进子目标选择与动作执行

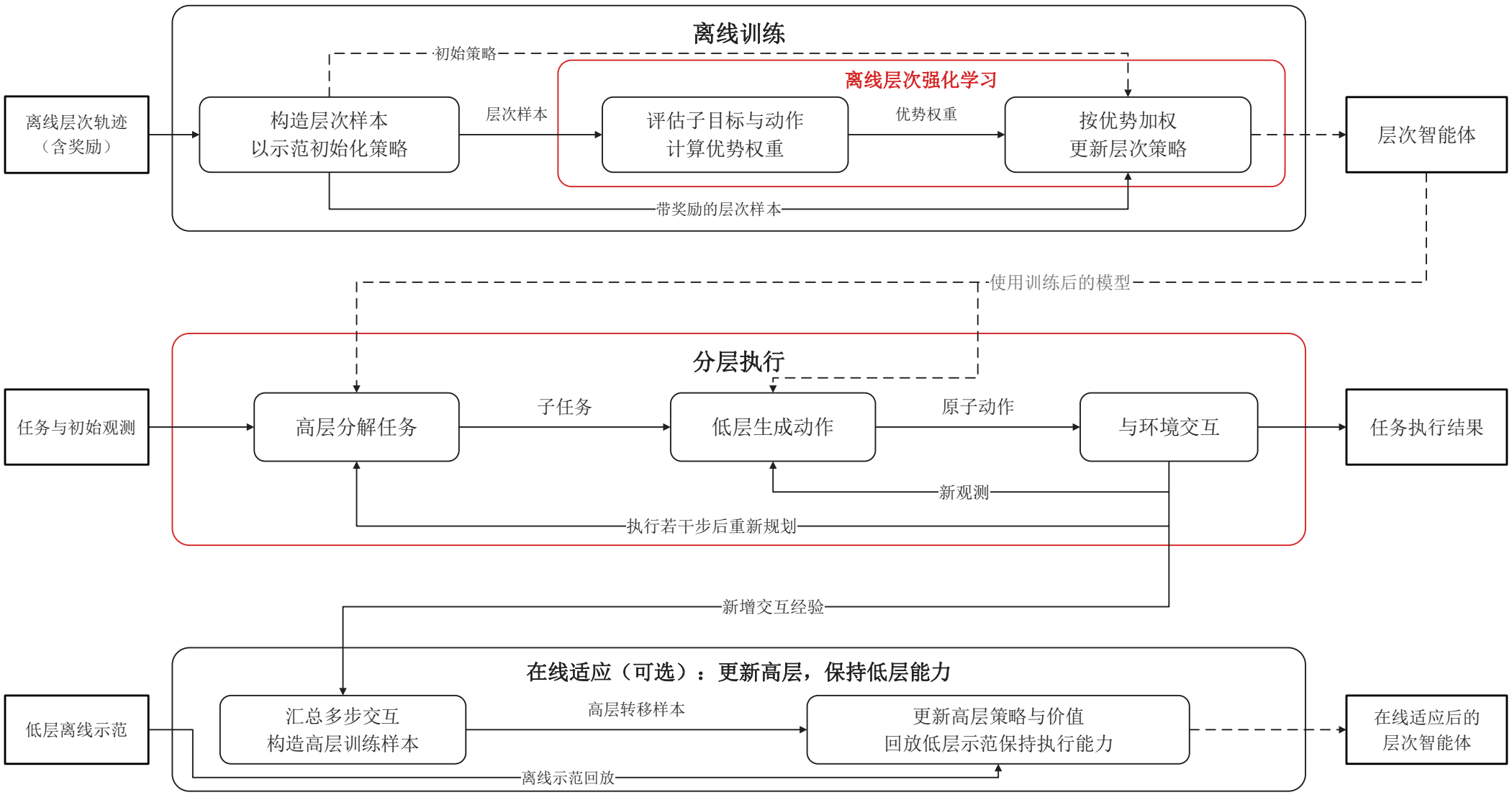
- 分层决策
  - 高层根据总任务与观察提出子目标
  - 低层围绕子目标连续执行原子动作

- 学习流程
  - 行为克隆建立基础策略，离线强化学习强化高价值决策
  - 在线阶段重点调整子目标选择



GLIDER 算法原理图

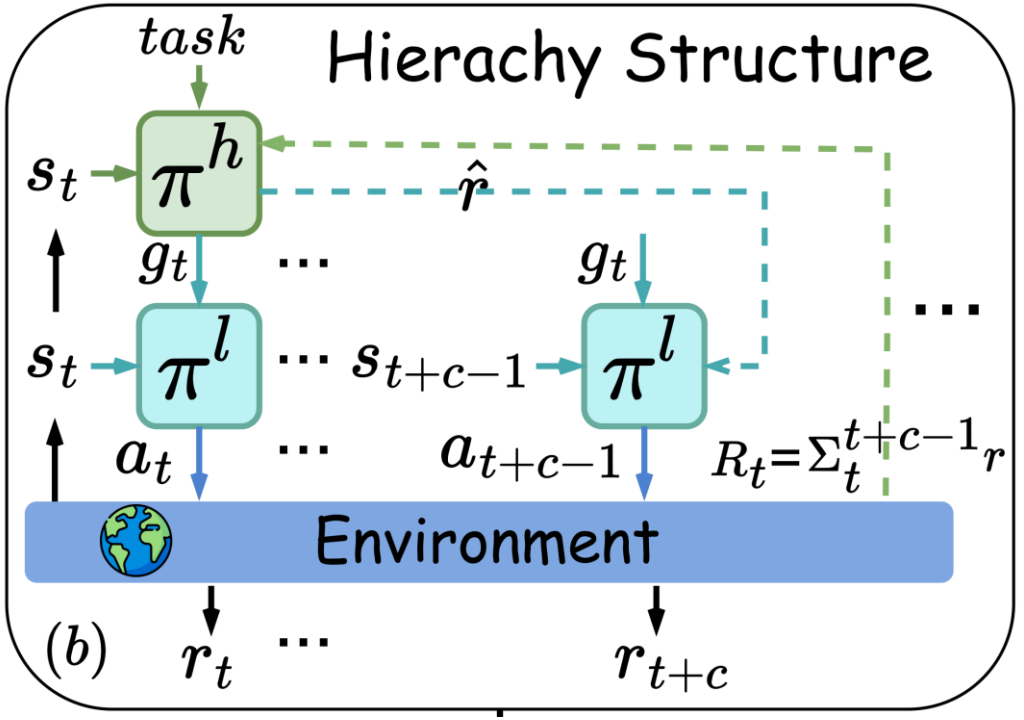
GLIDER 算法原理图



分层决策过程 高层规划低层执行

- 高层规划器：选择当前子目标
  - 根据整体任务d与当前观测o
    生成子目标$g_h \sim \pi_\theta^H(\cdot|d, o_t)$
  - 低层执行一段动作后，高层再选择后续子目标
- 低层控制器：围绕子目标连续执行
  - 在子目标g_h下，根据每一步的新观测
    生成环境能够执行的具体动作

$$a_t \sim \pi_\theta^L(\cdot|g_h, o_t)$$



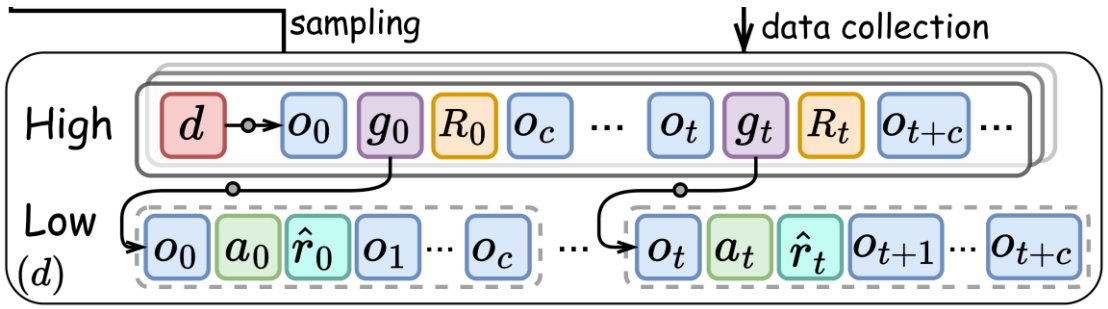
- 高低层共享模型，通过不同层次提示词区分角色

高层选择子目标，低层逐步执行，形成不同决策跨度的时间抽象

分层数据构建 轨迹组织与奖励分工

- 高层样本：记录一次子目标决策
  - 以子目标为单位，记录执行前后的观察
  - 累积该阶段的环境奖励，学习子目标选择

$$(d, o_t, g_h, R_h, o_{t+c_h}), R_h = \sum_{j=0}^{c_h-1} r_{t+j}$$



Training Dataset Structure

High-Level Trajectory:
{high prompt, task description, obs 0, subtask 0, high reward 0 ... obs T-1, subtask T-1, high reward T-1, obs T}

Low-Level Trajectory:
{Low prompt, subtask 0, obs 0, action 0, low reward 1 ... obs c-1, action c-1, low reward c-1, obs c}

- 低层样本：记录每一步动作执行
  - 记录当前子目标下的每一步原子动作
  - 根据观察判断子目标是否完成，完成奖励为1，否则为0

$$(g_h, o_t, a_t, \hat{r}_t, o_{t+1})$$

环境奖励评价整体任务进展，子目标奖励评价局部技能是否做成

基础策略初始化 行为克隆

- 高层：学习如何提出子目标
  - 输入整体任务与当前观察
  - 以示范中的子目标为监督，学习任务分解

- 低层：学习如何执行
  - 输入当前子目标与环境观察
  - 以示范中的原子动作为监督，学习有效执行

$$L_{BC} = -E_{D^H} log\pi_\theta^H(g|d,o) - E_{D^L} log\pi_\theta^L(a|g,o)$$

Expert Demonstration

w/o Hierarchy

Task Description: Your task is to find a(n) animal. First, focus on the thing. Then, move it to the red box in the kitchen.
Observation: This room is called the hallway. In it, you see: the agent, a substance called air, a picture. You also see: A door to the art studio (that is closed), A door to the bedroom (that is closed), A door to the greenhouse (that is closed), A door to the kitchen (that is closed), A door to the living room (that is closed), A door to the workshop (that is closed).
Action: open door to outside
Reward: 0.09
Observation: The door is now open.
Action: go to outside
Reward: 0.08
Observation: You move to the outside.
Action: look around
Reward: 0.0
Observation: This outside location is called the outside. Here you see: the agent, a substance called air, an axe, a crocodile egg, a fire pit (containing nothing), a fountain (containing a substance called water), the ground, a turtle egg, a baby wolf, a substance called wood. You also see: A door to the foundry (that is closed), A door to the greenhouse (that is closed), A door to the kitchen (that is open)
Action: focus on egg turtle
Reward: 0.5
Observation: You focus on the turtle egg.
Action: pick up egg turtle
Reward: 0.08
Observation: You move the turtle to the inventory.
Action: open door to kitchen
Reward: 0.0
Observation: The door is already open.
Action: go to kitchen
Reward: 0.0
Observation: You move to the kitchen.
Action: move egg turtle egg in inventory to red box
Reward: 0.17
Observation: You move the turtle to the red box.
done.

w/ Hierarchy

Task Description: Your task is to find a(n) animal. First, focus on the thing. Then, move it to the red box in the kitchen.
Observation: This room is called the hallway. In it, you see: the agent, a substance called air, a picture. You also see: A door to the art studio (that is closed), A door to the bedroom (that is closed), A door to the greenhouse (that is closed), A door to the kitchen (that is closed), A door to the living room (that is closed), A door to the workshop (that is closed).
Subtask: Find an animal and focus it
Action: open door to outside
Observation: The door is now open.
Action: go to outside
Observation: You move to the outside.
Action: look around
Observation: This outside location is called the outside. Here you see: the agent, a substance called air, an axe, a crocodile egg, a fire pit (containing nothing), a fountain (containing a substance called water), the ground, a turtle egg, a baby wolf, a substance called wood. You also see: A door to the foundry (that is closed), A door to the greenhouse (that is closed), A door to the kitchen (that is open)
Action: focus on egg turtle
High Reward: 0.75; **Low Reward:** 1.0
Observation: You focus on the turtle egg.
Subtask: Navigation to kitchen with the animal
Action: pick up egg turtle
Observation: You move the turtle to the inventory.
Action: open door to kitchen
Observation: The door is already open.
Action: go to kitchen
High Reward: 0.08; **Low Reward:** 1.0
Observation: You move to the kitchen.
Subtask: Move the animal to the red box in the kitchen
Action: move egg turtle egg in inventory to red box
High Reward: 0.17; **Low Reward:** 1.0
Observation: You move the turtle to the red box.
done.

先从示范中学会规划与执行，为后续强化学习建立基础策略

离线层次强化学习 演员—评论家模型

- 对高层和低层采用统一的演员—评论家表达
  - 高层：输入$x=(d,o)$，输出$u=g$
  - 低层：输入$x=(g,o)$，输出$u=a$
- 评论家采用时序差分（TD）学习与期望分位回归
- Q函数：通过时序差分学习，评价当前选择的长期价值

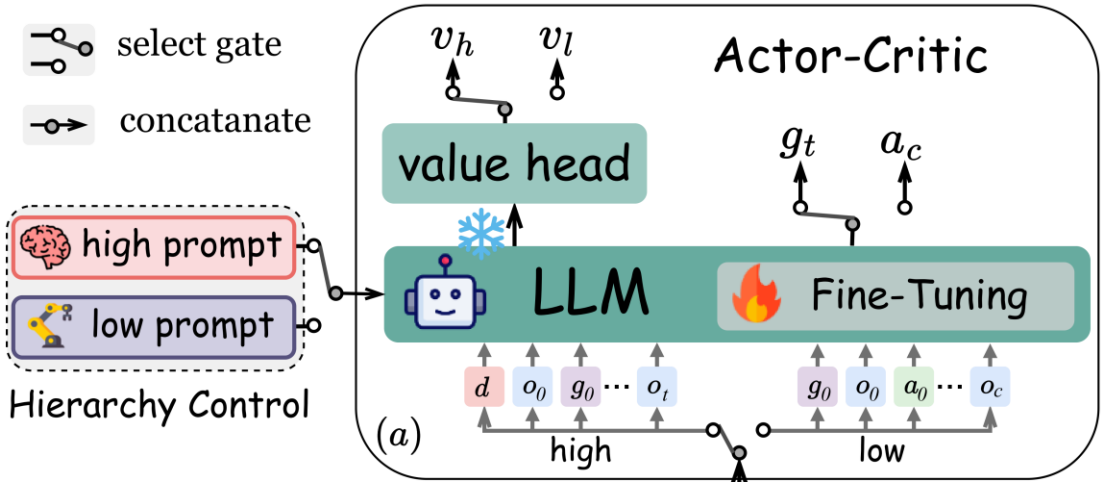
$$L_Q = E[(r + \gamma V_{\overline{\psi}}(x') - Q_\varphi(x,u))^2]$$

- V函数：使用期望分位回归建立价值基准

$$L_V = E[L_2^\tau(Q_{\overline{\varphi}}(x,u) - V_\psi(x))]$$

- 评论家优势评价
  - 判断当前选择是否优于基线

$$A(x,u) = Q_\varphi(x,u) - V_\psi(x)$$



离线层次强化学习 演员—评论家模型

- 演员将优势评价转化为样本学习权重
  - 把好坏判断转换成输出概率变化

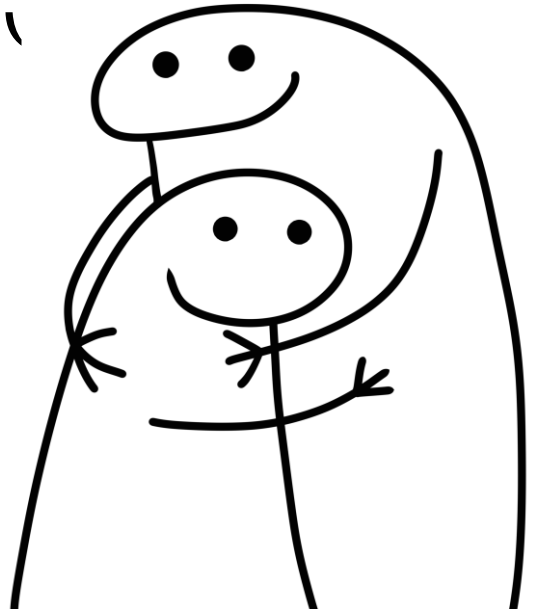
$$\mathcal{L}_\pi = -E_{(x,u)\sim D}\left[\exp\left(\frac{A(x,u)}{\kappa}\right)\log \pi_\theta(u \mid x)\right]$$

  - 通过加权模仿，提高模型生成相应输出的倾向
  - 优势越高，样本的学习权重越大，低优势样本的相对影响减小
- 从完整输出评价到Token概率更新

$$log\pi_\theta(u|x) = \sum_{i=1}^{n} log\pi_\theta(w_i|x, w_{<i})$$

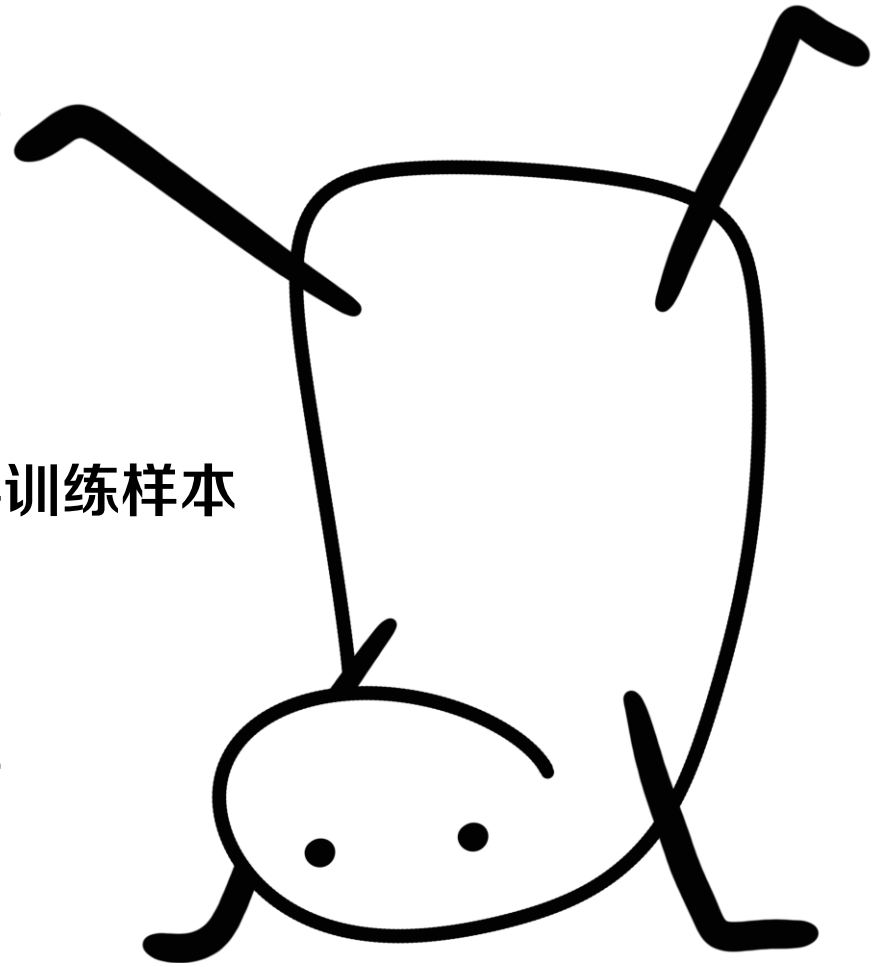
  - 评论家评价完整子目标或动作
  - 同一输出的优势权重作用于其各Token的对数概率项

利用长期回报让模型更着重学习价值较高的子目标和动作



在线适应 复用技能与高层更新

- 复用已有低层技能
  - 将导航、取物、操作设备等能力迁移到新任务
  - 重点调整子目标的选择与组合
- 利用新经验更新高层
  - 高层提出子目标，低层连续执行
  - 汇总环境奖励与阶段结束后的观察，构造高层训练样本
  - 更新高层评论家与演员，改进规划策略
- 利用旧示范维持低层能力
  - 高低层共享参数，高层更新可能影响低层执行
  - 使用离线低层示范继续训练以维持能力



GLIDER 实验设计 数据与评估

- 数据资源
  - ScienceWorld：30项科学实验任务，包含持续的任务进度反馈
  - ALFWorld：6类家庭操作任务，任务完成时获得稀疏奖励

Dataset	Train	Text-Seen	Text-Unseen
ScienceWorld	1,483	194	211
ALFWorld	3,119	140	134

  - 离线数据按专家:中等＝1:2混合
  - 让模型既能模仿较好的行为，也能利用不同质量的经验学习

- 对比方法与骨干模型
  - 提示词：ReAct、Reflexion、SwiftSage
  - 微调：NAT、ETO
  - 在线适应：AC、AWAC
  - 骨干：Mistral-7B、Gemma-7B、Llama-3-8B

- 评价指标
  - ScienceWorld任务得分、ALFWorld任务成功率，分别报告已见与未见
  - 在线适应按交互步数观察新任务上的表现

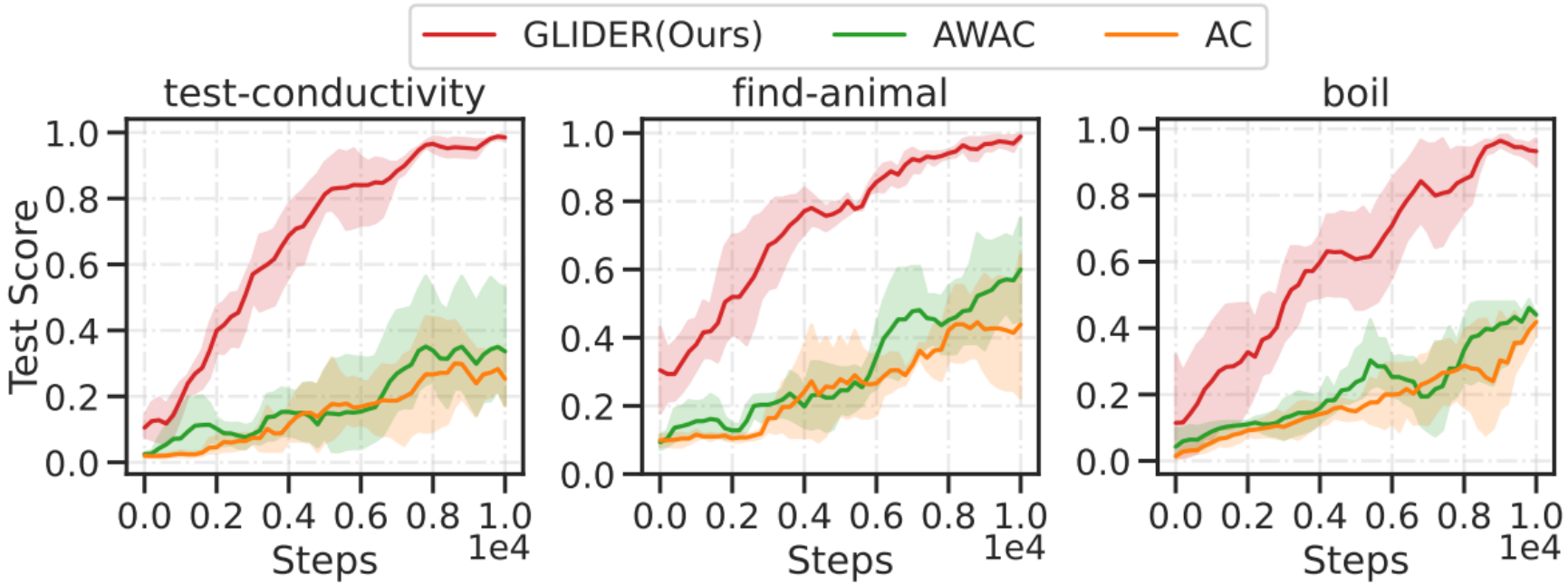
GLIDER 实验结果 对比实验

- GLIDER在三种骨干模型、两个交互环境的已见和未见测试中均取得最优结果，说明其性能优势在不同模型和测试设置下具有一致性

Backbone	Method	ScienceWorld		AlfWorld	
		Seen	Unseen	Seen	Unseen
Mistral-7B	ReAct	20.72	17.65	7.86	5.22
	Reflexion	21.07	18.11	11.56	6.00
	SwitchSage	48.40	45.25	30.29	26.52
	NAT	57.12	50.79	64.43	68.96
	ETO	58.17	51.85	66.84	71.43
	GLIDER	**67.31** (↑ 15.71%)	**65.14** (↑ 25.63%)	**70.02** (↑ 4.76%)	**74.83** (↑ 4.76%)
Gemma-7B	ReAct	3.58	3.51	6.43	2.24
	Reflexion	4.94	3.93	7.14	2.99
	SwitchSage	33.43	30.90	8.23	5.72
	NAT	47.63	44.98	67.86	65.88
	ETO	50.44	47.84	66.43	68.66
	GLIDER	**63.67** (↑ 26.23%)	**58.50** (↑ 22.28%)	**72.12** (↑ 6.28%)	**70.88** (↑ 3.23%)
Llama-3-8B	ReAct	24.76	22.66	2.86	3.73
	Reflexion	27.23	25.41	4.29	4.48
	SwitchSage	42.22	40.58	20.39	10.78
	NAT	55.24	48.76	60.71	59.70
	ETO	57.90	52.33	64.29	64.18
	GLIDER	**77.43** (↑ 33.73%)	**68.34** (↑ 30.59%)	**71.56** (↑ 11.31%)	**75.38** (↑ 17.45%)

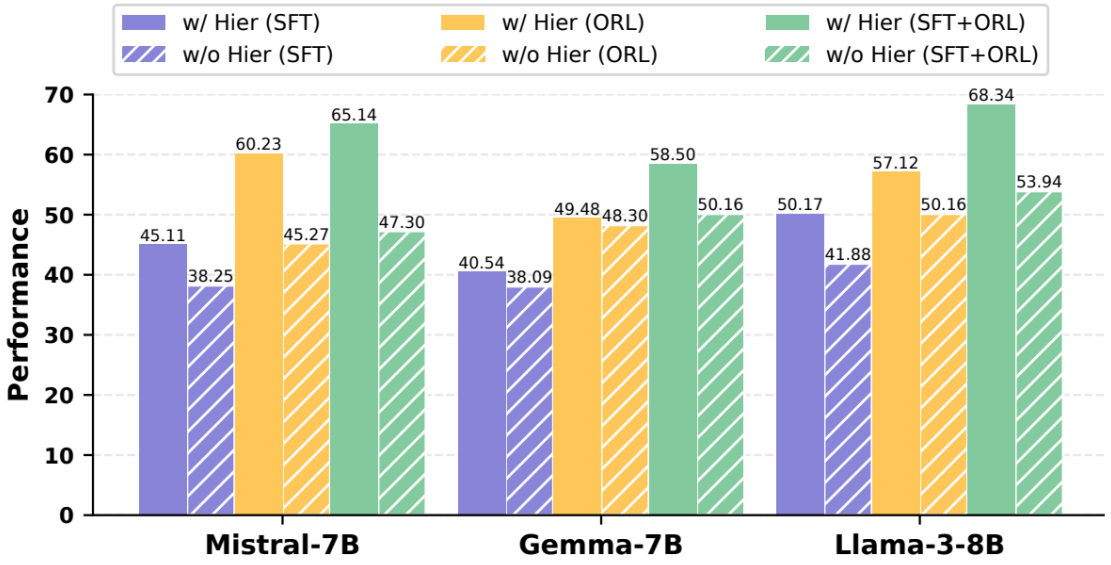
GLIDER 实验结果 在线适应实验

- 离线训练中剔除test-conductivity、find-animal、boil任务，随后在这些任务上进行在线微调
- 相较AC和AWAC，GLIDER具有更高的初始表现与更快的适应趋势，体现已有低层技能对新任务的支持



GLIDER 实验结果 消融实验

- 层次结构与SFT+ORL组合均有贡献，8B完整训练下，分层得分68.34分，无层次为53.94分，相差14.40分
- 分层方案在1B、3B、8B均有效，完整训练下3B分层模型达到61.29分，表明收益具有一定的跨规模稳定性



Model	w/o Hier			w/ Hier		
	SFT	ORL	SFT+ORL	SFT	ORL	SFT+ORL
Llama-1B	37.24	45.31	48.48	44.50	50.43	53.62
Llama-3B	38.19	52.47	56.93	48.11	55.98	61.29
Llama-8B	41.88	50.16	53.94	50.17	57.12	68.34

GLIDER 算法总结

- 算法贡献
  - 离线学习：先监督微调模仿示范，再通过强化学习重点学习较好的选择
  - 在线适应：复用已有低层技能，重点改进高层的子目标选择与组合

- 算法不足
  - 参数相互影响：高低层共享参数，更新高层时需用旧示范维持低层能力
  - 训练流程与验证范围：多阶段训练较复杂，未在数学和代码任务上测试



Multi²: Hierarchical Multi-Agent Decision-Making with LLM-Based Agents in Interactive Environments

Multi²

T	目标	提升智能体在长时程任务中的执行可靠性、目标一致性与交互效率，缓解目标漂移
I	输入	任务描述、环境观测及带奖励的交互轨迹；数据涵盖30种科学实验任务、6类家庭操作任务及物品合成任务
P	处理	1.构建子目标规划与动作执行两类训练数据 2.共享语言模型骨干，采用独立LoRA实现规划与执行的角色分工 3.通过监督微调学习规划，通过离线强化学习优化执行 4.混合离线经验与在线交互数据，在 KL约束下持续改进执行策略
O	输出	一串动作指令序列

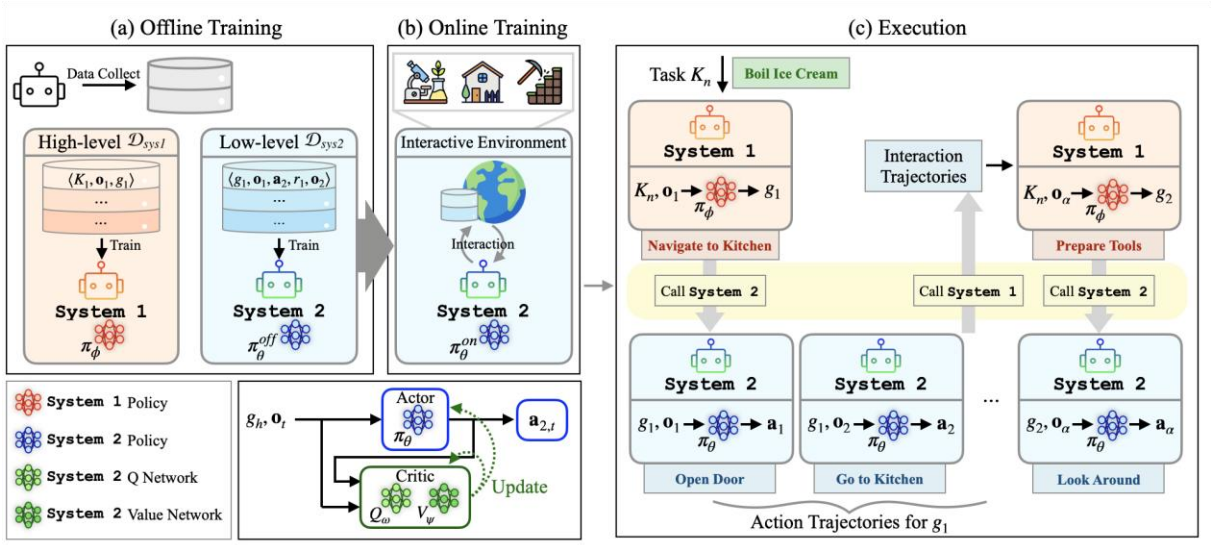
P	问题	现有智能体在长时程动态交互中易因执行误差累积而偏离目标，出现无效操作与重复循环，导致任务成功率下降、交互开销增加
C	条件	具备可微调的预训练语言模型、带子目标与奖励信息的交互轨迹，以及能够持续交互并提供反馈的环境
D	难点	规划与执行差异化训练方式的设计、在线执行策略更新的稳定性，以及任务完成效果与交互效率的平衡
L	水平	ICML 2026 CCF A

Multi²

- 科学问题
  - 如何针对规划与执行的不同学习需求，建立差异化训练与稳定在线改进机制，缓解长时程任务中的执行误差累积与目标漂移，兼顾任务完成效果与效率
- 应用问题
  - 在环境持续变化、操作相互依赖的长时程任务中，现有智能体容易因执行错误累积而偏离目标，出现无效操作与重复循环，导致任务成功率下降、交互开销增加

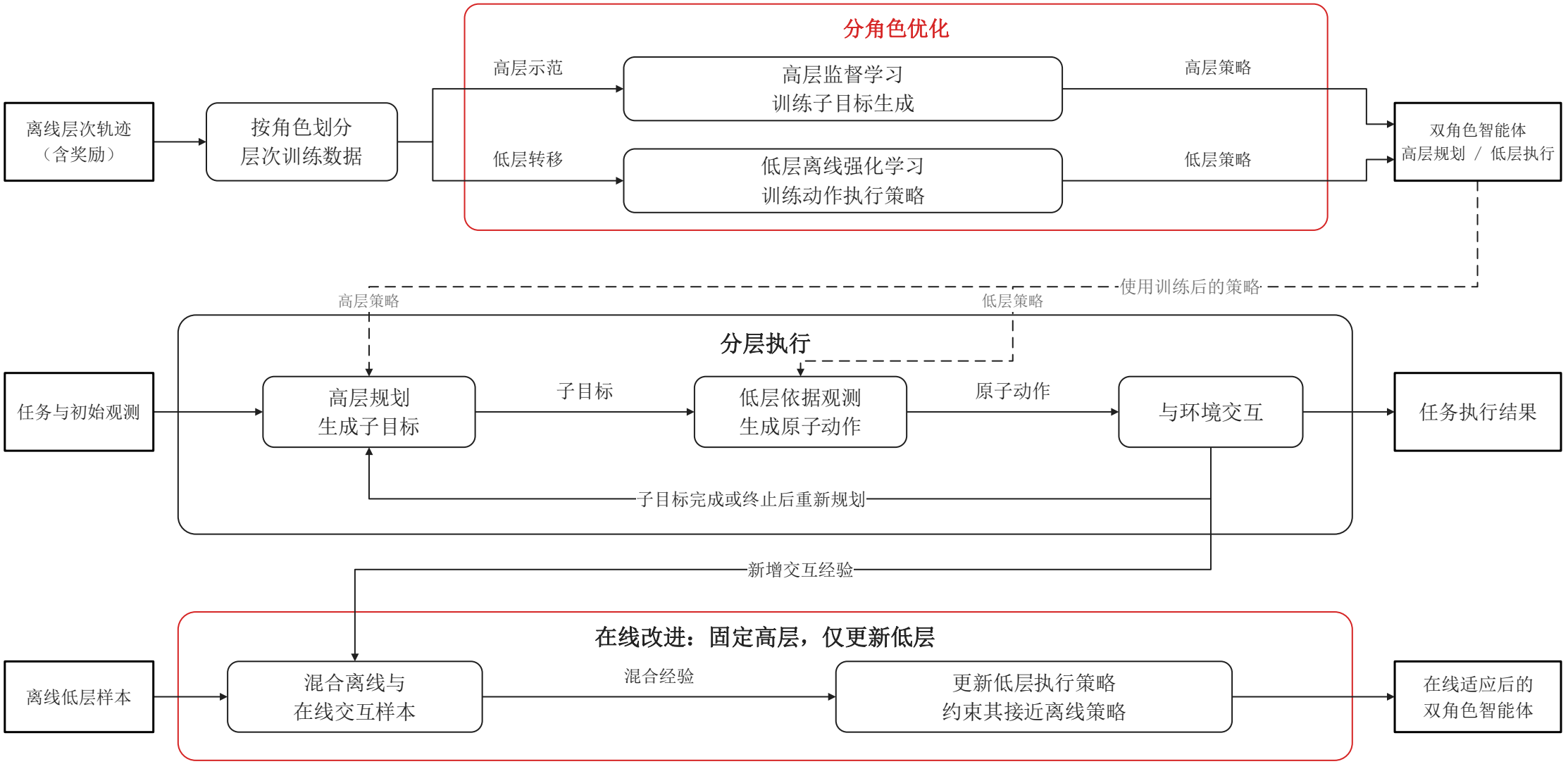
Multi² 算法 创新框架

- 核心思想
  - 将子目标规划与动作执行视为不同的优化问题，通过角色专用参数与差异化训练分别学习
- 角色专门化与离线训练
  - 共享冻结骨干，两个角色使用独立LoRA
  - 系统1通过SFT学习子目标生成
  - 系统2通过离线RL学习原子动作
- 执行侧在线改进
  - 固定系统1参数，继续调用它生成子目标
  - 借助混合回放和KL约束，对系统2继续进行在线强化学习



角色参数分开，训练目标不同，在线重点改进动作执行

Multi² 算法原理图



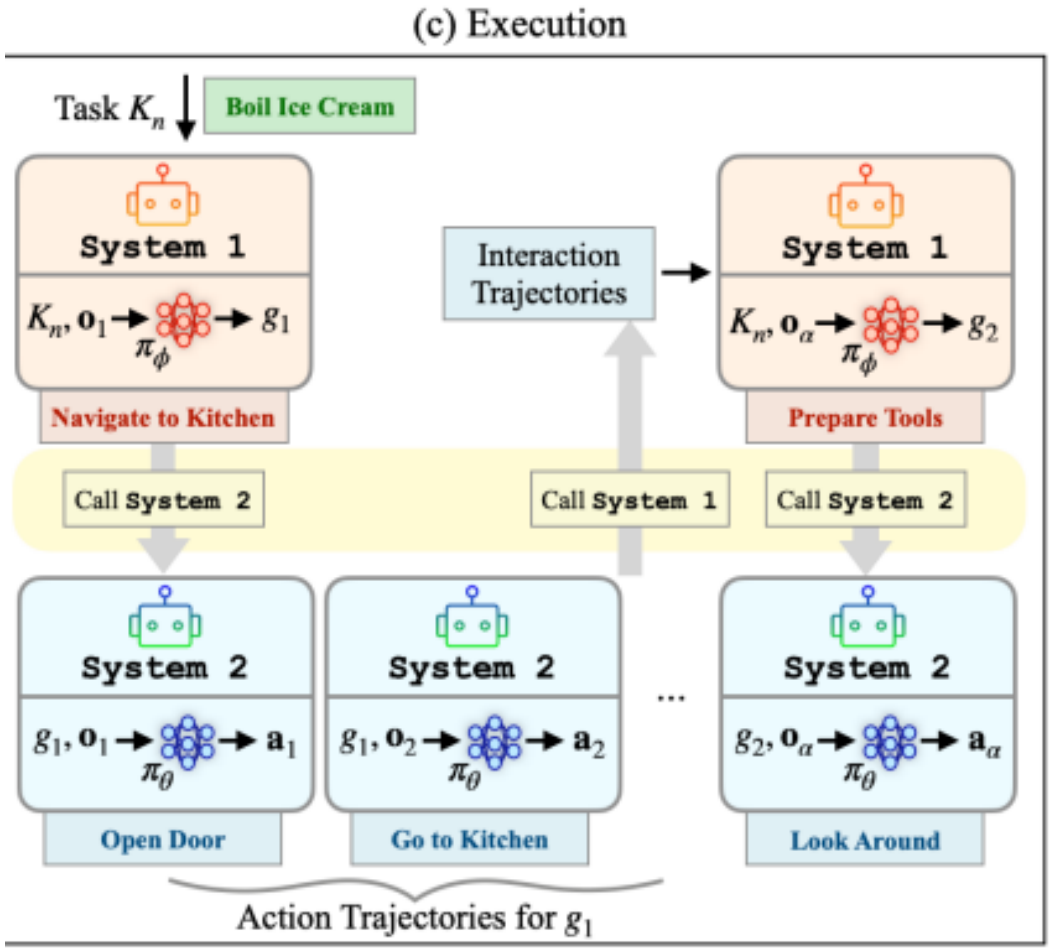
分层决策过程 子目标、动作与完成信号

- 系统1：根据任务进展生成子目标
  - 结合整体任务、已执行动作记录与当前观察
  - 判断任务进度并提出下一阶段目标

$$g_h \sim \pi_\varphi(\cdot\,|o_t; K)$$

- 系统2：围绕子目标逐步执行
  - 根据当前子目标与最新观察，输出原子动作
  - 输出布尔完成标志，判断子目标是否完成

$$a_t \sim \pi_\theta(\cdot\,|o_t; g_h)$$



子目标持续约束执行方向，最新观测决定具体动作

角色专门化 独立适配器与差异化训练

- 共享冻结骨干，分别学习角色参数
  - 系统1与系统2使用独立LoRA适配器
  - 训练和推理时只激活对应适配器

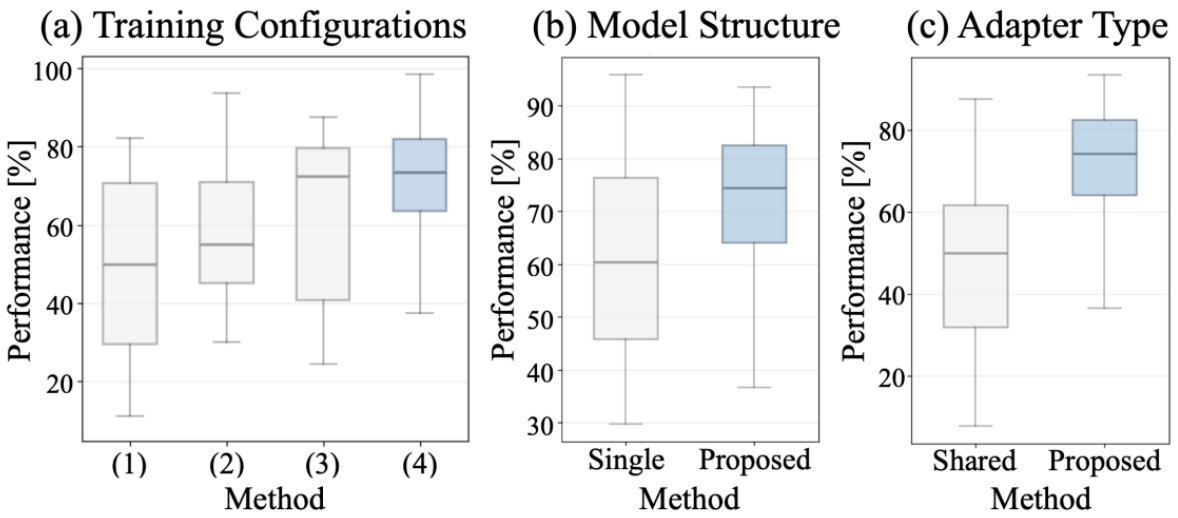
$$W_H = W_0 + B_H A_H, W_L = W_0 + B_L A_L$$

- 系统1通过SFT学习子目标生成
  - 专家子目标提供明确监督

    学习任务分解与子目标选择

$$L_\varphi = -E_{D_{sys1}} log\pi_\varphi(g_h|o_t; K)$$

- 系统2通过离线RL学习动作执行
  - 利用奖励与状态转移，学习子目标条件下的动作选择

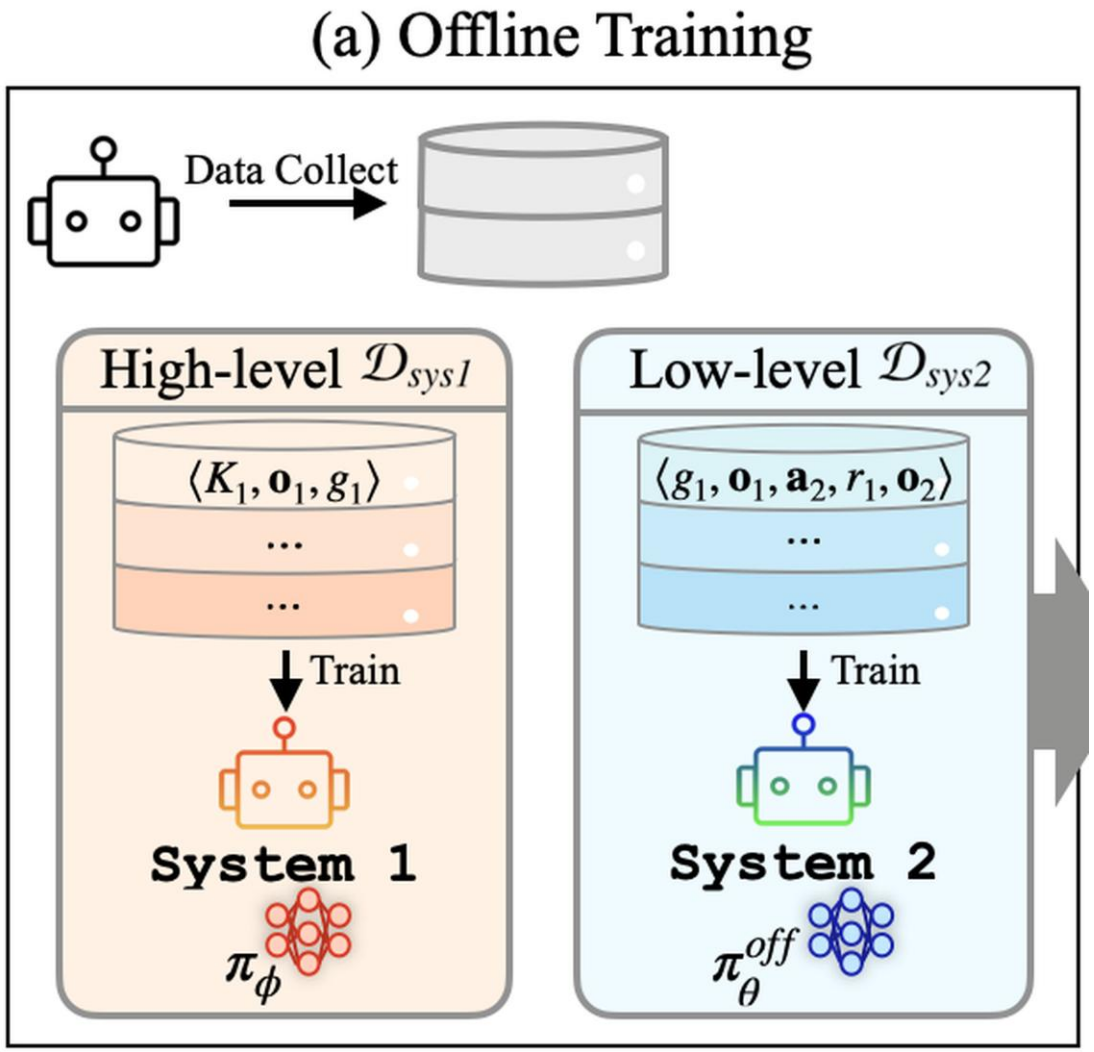


分层数据构建 角色样本与子目标奖励

- 确定性的规则处理轨迹
  - 读取环境的交互轨迹，按规则分段、规范化命令、过滤非动作文本
- 高层样本：学习观测对应的子目标

  $D_{sys1}: (K, o_h, g_h)$
- 低层样本：学习子目标下的动作控制

  $D_{sys2}: (g_h, o_t, a_t, r_t, o_{t+1})$
  - ScienceWorld直接使用环境进度奖励
  - ALFWorld、TextCraft加入子目标完成奖励



系统2离线强化学习 价值学习

- **Q函数：评价具体动作的长期回报**

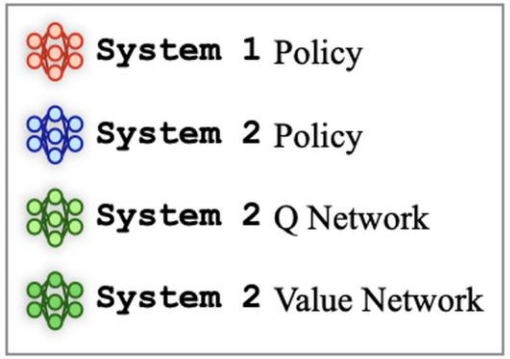
$$L_Q = E_{D_{sys2}}[(r_t + \gamma V_\psi(o_{t+1}; g_h) - Q_\omega(o_t, a_t; g_h))^2]$$

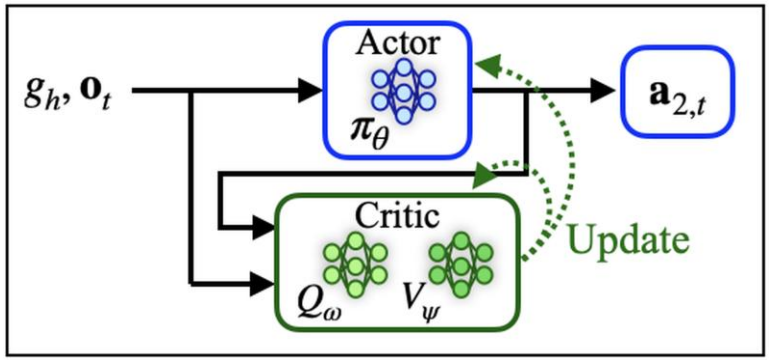
- **V函数：期望分位回归建立价值基准**

$$L_V = E_{D_{sys2}}\left[L^\tau\left(A(o_t, a_t; g_h)\right)\right]$$

$$A(o_t, a_t; g_h) = Q_\omega(o_t, a_t; g_h) - V_\psi(o_t; g_h)$$

  - **优势衡量具体动作比参考水平好多少**





以子目标为条件评价动作，为演员更新建立价值基准

系统2离线强化学习 优势加权与策略锚定

- 优势加权模仿

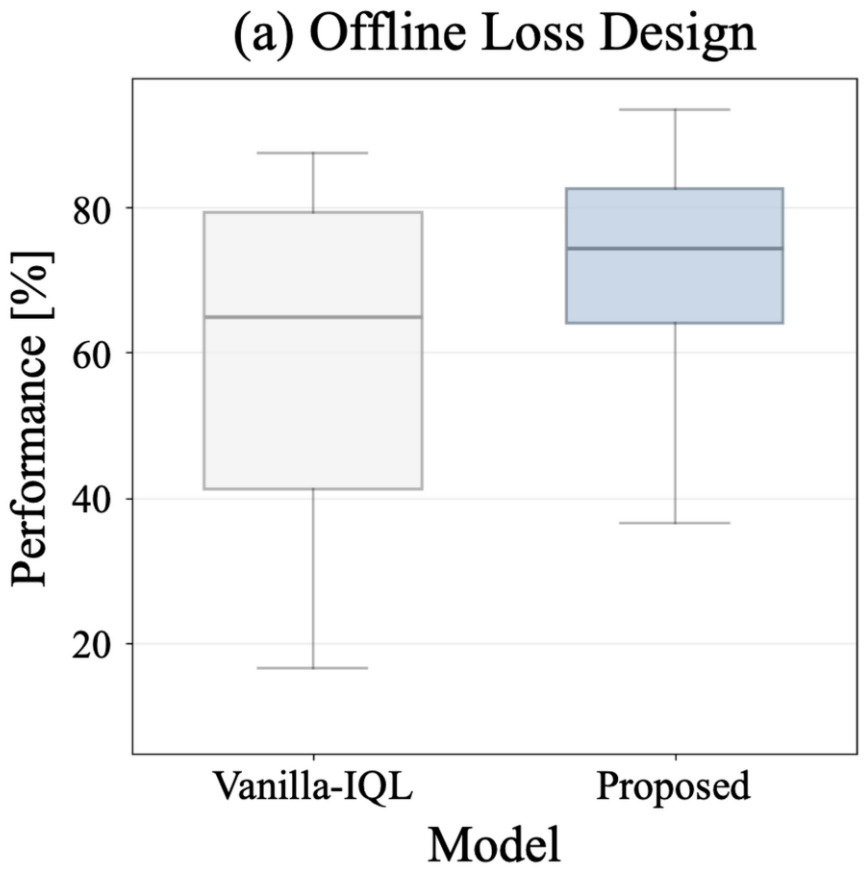
$$L_{AWR} = -E_{D_{sys2}}[e^{\beta A(o_t,a_t;g_h)} log\pi_\theta^{off}(a_t|o_t;g_h)]$$

  - 为高优势的数据动作分配更大的学习权重

- 策略锚定优势项

$$L_{anchor} = -\lambda E_{D_{sys2}} E_{\widetilde{a}_t \sim \pi_\theta^{off}(\cdot|o_t;g_h)}[A(o_t,\widetilde{a}_t;g_h)]$$

  - 结合当前策略生成动作的评论家评价
    引导策略偏向估计回报更高的候选

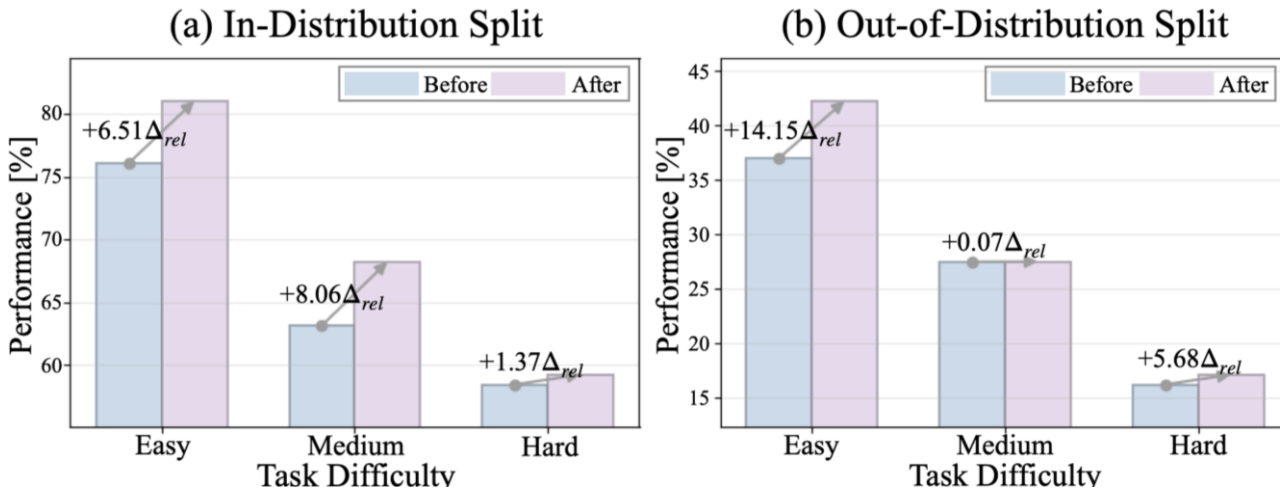


通过评论家指导动作学习，减少对固定示范形式的过度依赖

系统2在线改进 混合回放与KL约束

- 混合经验回放
  - 保留离线经验并加入当前策略新轨迹
  - 固定系统1，更新系统2演员与评论家

$$B = D_{offline} \cup D_{new}$$



- 优势加权改进与离线策略约束

$$L_{online} = -E_{B[w_t log \pi_\theta^{on}(a_t|o_t;g_h)]} + \eta E_{BD_{KL}(\pi_\theta^{on} \| \pi_\theta^{off})}$$

$$w_t = exp\left(\frac{A(o_t,a_t;g_h)}{\alpha}\right)$$

执行器学习处理实际访问的状态，并限制在线策略的过快变化

Multi² 实验设计 数据与评估

- 数据资源
  - ScienceWorld：科学实验与状态变化
  - ALFWorld：家庭操作与稀疏奖励
  - TextCraft：配方合成与组合依赖
  - 离线轨迹来自AgentGym的GPT-4-Turbo交互记录
- 对比方法与骨干模型
  - 提示词：ReAct、Reflexion、ADaPT
  - 微调：GRPO、GLIDER
  - 骨干：Qwen-2.5 3B、Mistral 7B v0.3、Llama-3.1 8B
- 评价指标
  - ScienceWorld报告任务得分，ALFWorld报告任务成功率，分别列出ID/OOD结果
  - TextCraft报告整体成功率
  - Reflexion按原协议使用pass@6，另分析推理token用量与效率

Multi² 实验结果 对比实验

- 采用Llama-3.1（8B）时，相比GLIDER，ALFWorld的OOD成功率由37.86%提高至56.43%,TextCraft成功率由9.50%提高至35.60%
- ScienceWorld的OOD表现为30.68%，低于GLIDER的34.36%，优势未在所有实验设置下体现

Base Models	Training Type	Structure Type		Methods	ScienceWorld		ALFWorld		TextCraft
		Single	Hier.		ID [%]	OOD [%]	ID [%]	OOD [%]	Success [%]
Llama-3.1 8B	Prompt-based	✓		ReAct [14]	23.23	10.30	6.72	7.46	9.00
		✓		Reflexion [34]	23.20	6.97	35.71	29.29	11.00
			✓	ADaPT [16]	26.20	12.27	11.54	6.41	5.00
	Fine-tuning-based	✓		GRPO [83]	25.79	4.61	8.57	7.50	5.50
			✓	Glider [15]	60.48	34.36	43.57	37.86	9.50
			✓	Multi² (Proposed)	67.61	30.68	57.86	56.43	35.60

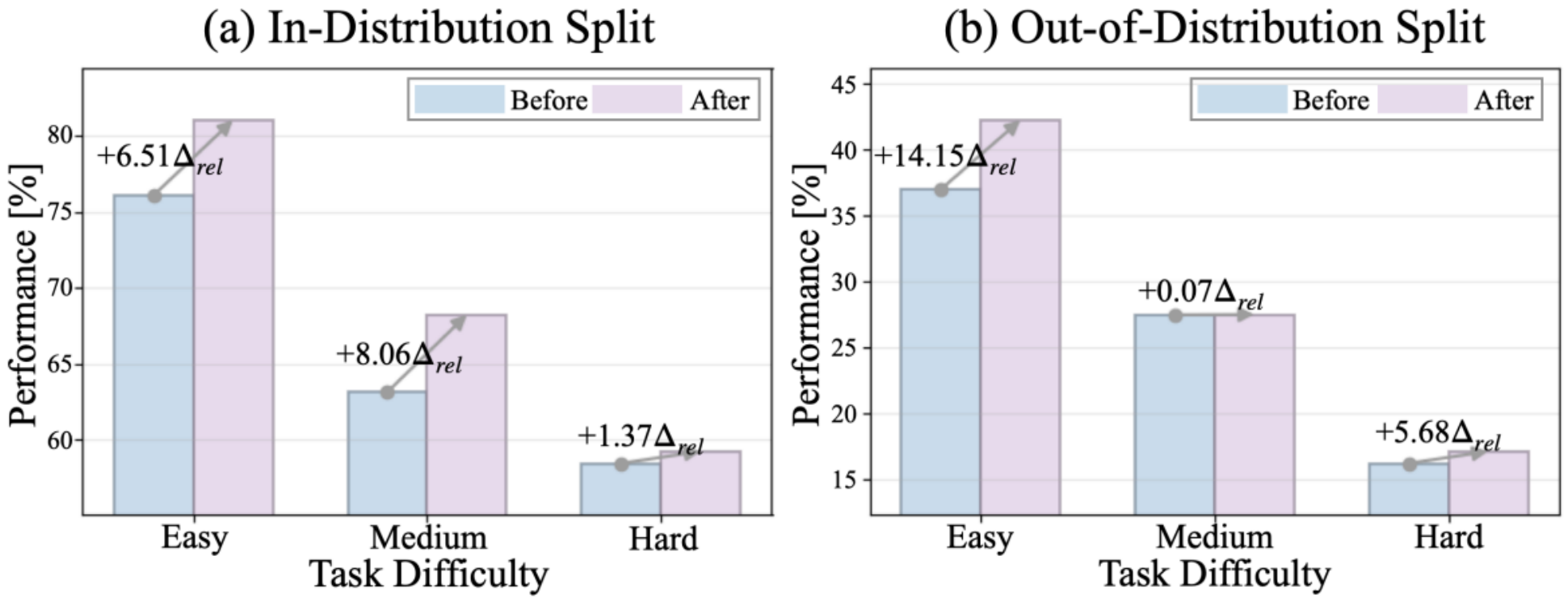
Multi² 实验结果 Token效率分析

- ScienceWorld的OOD划分中，Multi²的归一化Token效率为13.769，高于GLIDER的10.838
- 相较GLIDER，ID平均Token减少16.3%，OOD减少29.7%

Training Type	Structure Type		Methods	In-Distribution Split			Out-of-Distribution Split		
	Single	Hier.		Perform.	Tokens	TE	Perform.	Tokens	TE
Prompt-Based	✓		ReAct [14]	23.23%	18971.23	1.000	10.30%	68031.35	1.000
		✓	Reflexion [34]	22.18%	18063.16	1.003	6.97%	64755.37	0.711
		✓	ADaPT [16]	26.20%	32449.12	0.659	12.27%	63288.85	1.281
Fine-Tuning-Based	✓		GRPO [83]	33.20%	11145.44	2.433	4.61%	9878.00	3.083
		✓	Glider [15]	60.48%	14716.87	3.356	34.36%	20939.88	10.838
		✓	Multi² (Proposed)	67.61%	12315.42	4.483	30.68%	14716.87	13.769

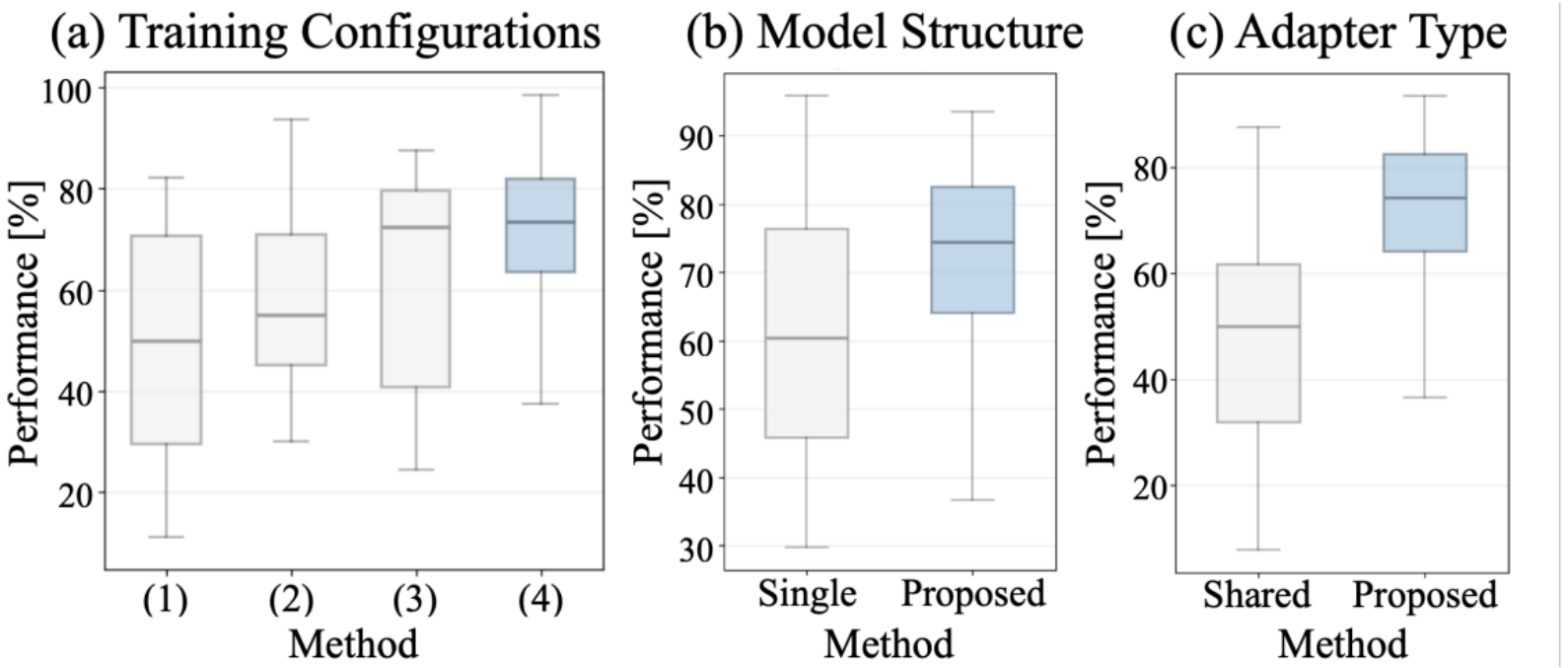
Multi² 实验结果 在线适应实验

- 固定高层规划器，在线训练低层执行器，表中所列六组任务表现均提高，但收益幅度不同
- OOD简单任务提升5.23分，中等任务仅提高0.02分



Multi² 实验结果 消融实验 训练分工与参数消融

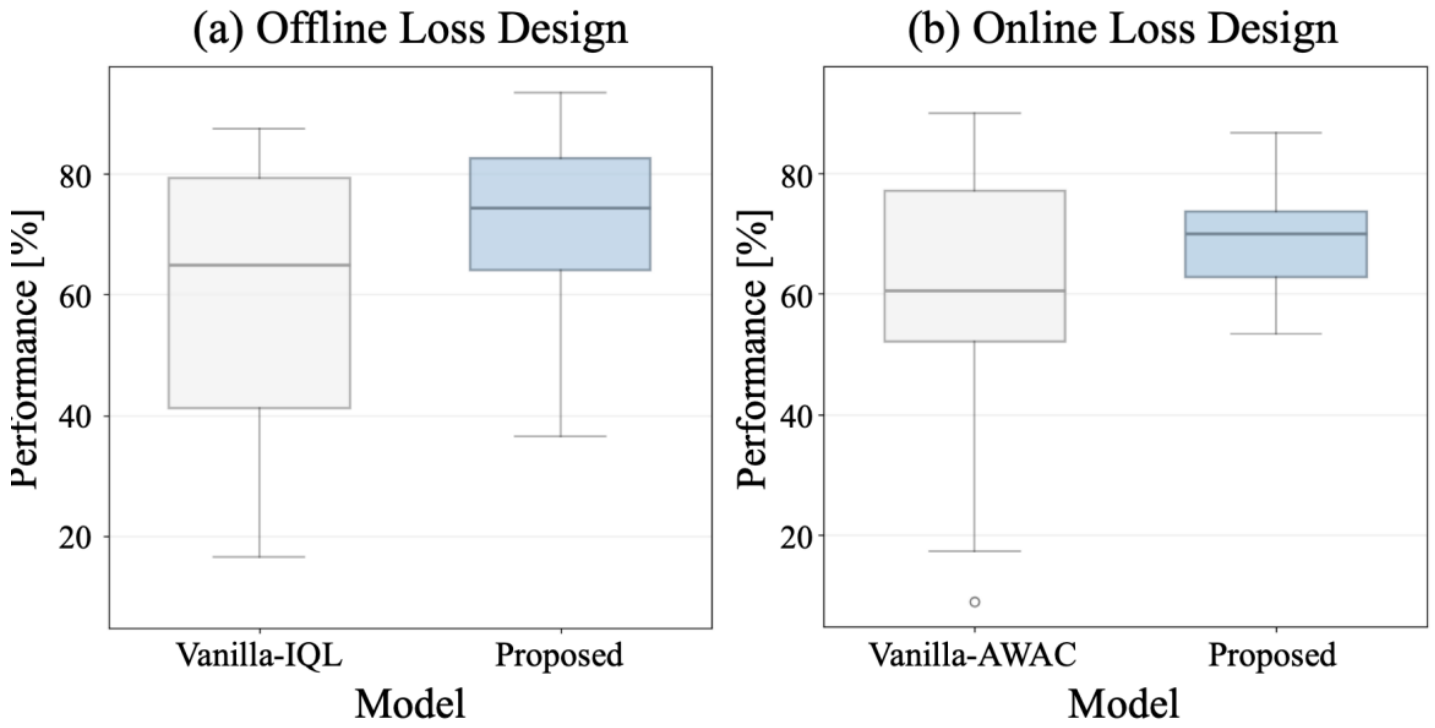
- 高层采用SFT、低层采用RL的分工取得更高的中位表现，调换分工或统一采用一种学习方式，性能与稳定性下降
- 分层结构、角色独立适配器均改善表现分布，说明收益同时来自决策结构与参数专门化



Multi² 实验结果 消融实验 离线与在线损失消融

- 离线目标：加入策略锚定项后，中位表现提高，低分结果减少

- 在线目标：加入KL后，表现分布更集中，低分结果更少



Multi² 算法总结

- 算法贡献
  - 角色专门化：共享骨干、独立LoRA，高层通过SFT学习子目标，低层通过离线到在线RL改进行为
  - 执行学习与稳定优化：离线强化学习加入评论家指导项，在线结合混合回放和KL约束，减少执行侧错误累积
- 算法不足
  - 规划适应仍受限制：在线阶段固定规划器，遇到新的任务分解需求时，单独更新执行器可能不足
  - 证据与泛化范围：主要验证三类文本环境，目标漂移分析依赖任务表现与行为案例



特点总结与未来展望

特点总结与未来展望

- 特点总结
  - GLIDER：以层次强化学习组织长程规划与执行
    - 高层生成子目标，低层执行动作，通过时间抽象与分层奖励缓解长程决策困难
    - 结合SFT与离线强化学习，在线重点调整高层规划，复用低层技能适应新任务
  - Multi²：以角色专门化训练增强执行适应
    - 高层SFT规划、低层离线到在线RL执行，采用独立LoRA减少参数更新干扰
    - 结合优势学习、混合回放与KL约束，改进动作选择并稳定在线更新
- 未来展望
  - 将层次规划与执行学习拓展到代码生成和数学推理，支持复杂程序构建与多步骤求解
  - 研究高层规划与低层执行的协同适应，探索长时程任务中的目标漂移检测与纠正机制，提升目标保持与错误恢复能力

参考文献

[1] HU Z, LIU W, QU X, et al. Divide and conquer: Grounding LLMs as efficient decision-making agents via offline hierarchical reinforcement learning[C]//Proceedings of the 42nd International Conference on Machine Learning. PMLR, 2025, 267: 24570-24590.

[2] PARK S, KWON M. Multi²: Hierarchical multi-agent decision-making with LLM-based agents in interactive environments[C]//Proceedings of the 43rd International Conference on Machine Learning. PMLR, 2026, 306: 96766-96805.

道德经

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

