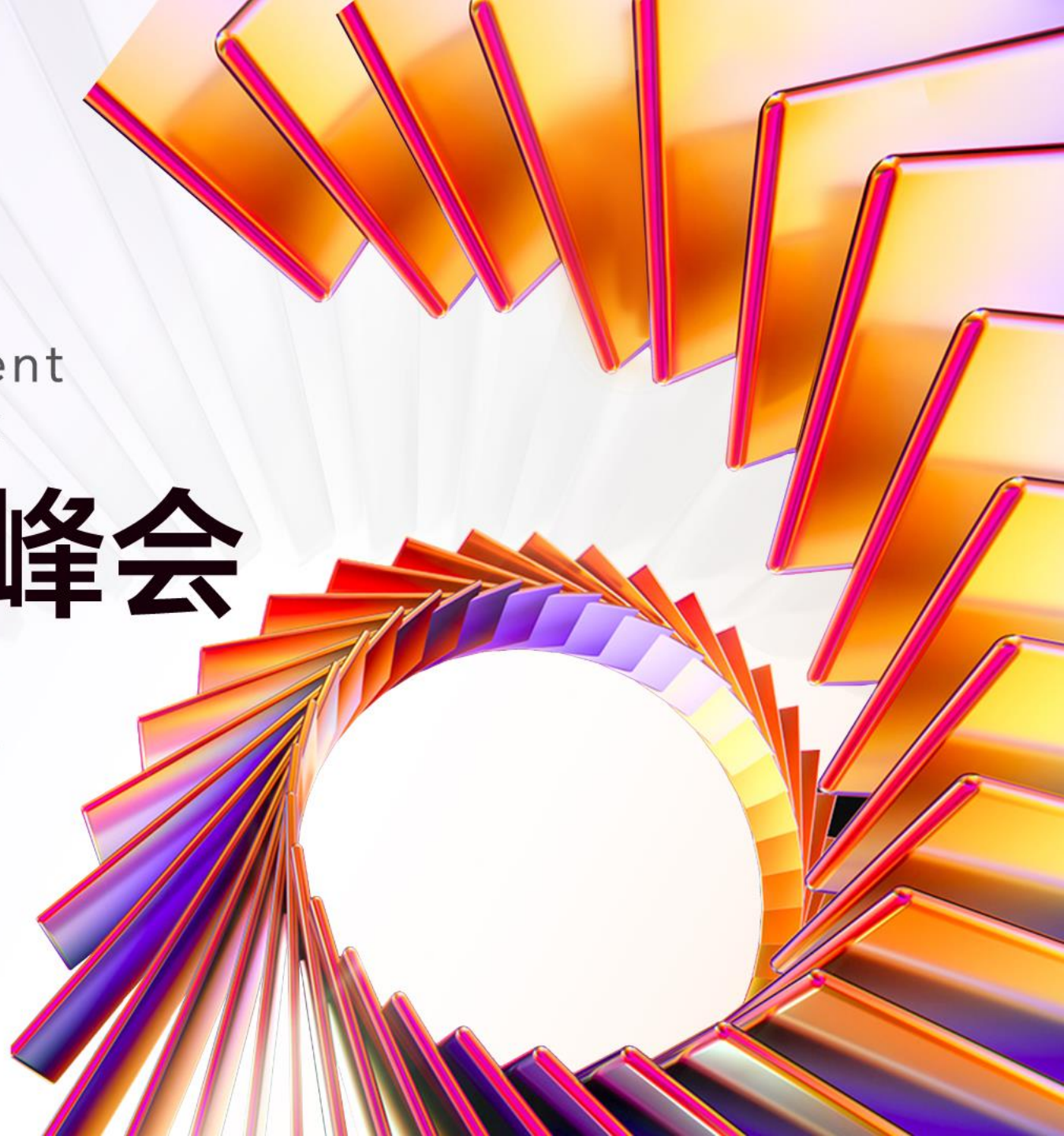
AiDD 8th 2025

第8届 AI+ Development Digital Summit AI+研发数字峰会

拥抱AI 重塑研发

11月14-15日 | 深圳



AiDD × K

2026 AI+ PRODUCT INNOVATION SUMMIT

01.16-17 · ShangHai

AI+产品创新峰会

Track 1: AI 产品战略与创新设计

从0到1的AI原生产品构建

论坛1: AI时代的用户洞察与需求发现

论坛2: AI原生产品战路与商业模式重构

论坛3: AgenticAI产品创新与交互设计

Track 2: AI 产品开发与工程实践

从1到10的工程化落地实践

论坛1: 面向Agent智能体的产品开发

论坛2: 具身智能与AI硬件产品

论坛3: AI产品出海与本地化开发

Track 3: AI 产品运营与智能演化

从10到100的AI产品运营

论坛1: AI赋能产品运营与增长黑客

论坛2: AI产品的数据飞轮与智能演化

论坛3: 行业爆款AI产品案例拆解

2-hour Speech: 回归本质

用户洞察的第一性

——2小时思维与方法论工作坊

在数字爆炸、AI迅速发展的时代，仍然考验“看见”的“同理心”

Panel 1: 出海前瞻

“出海避坑地图”圆桌对话

——不止于翻译:

AI时代的出海新范式

Panel 2: 失败复盘

为什么很多AI产品“叫好不叫座”?

——从伪需求到真价值:

AI产品商业化落地的关键挑战

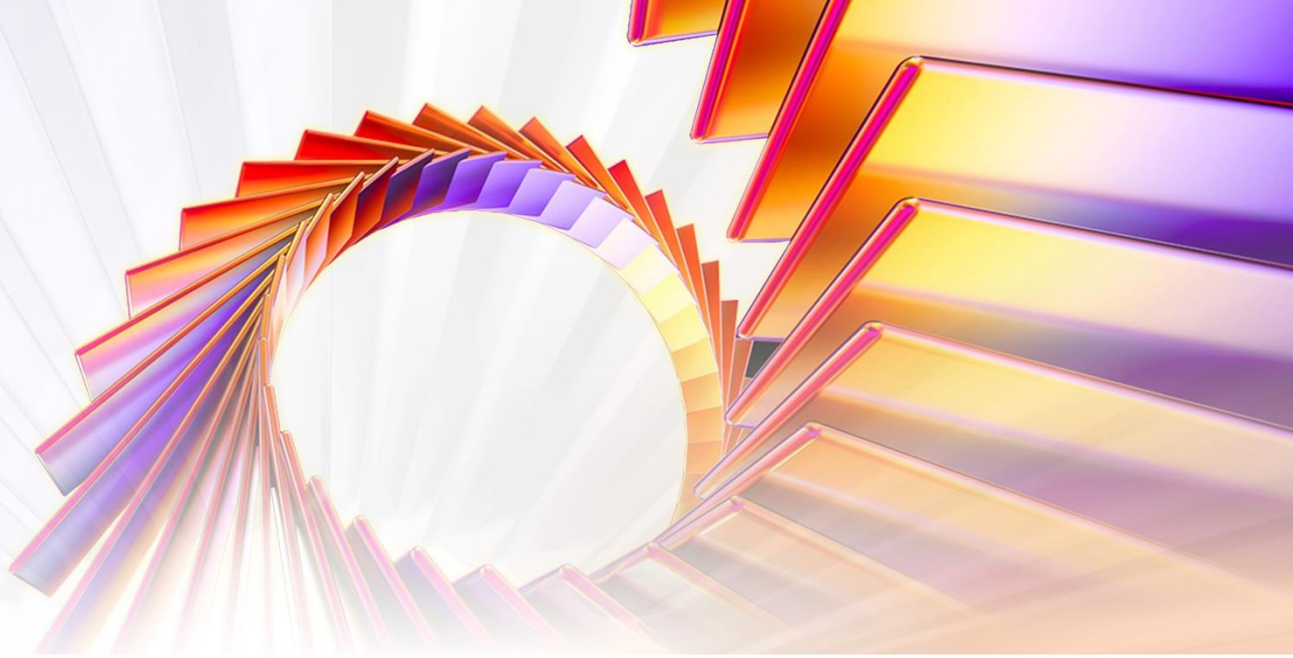
智能重构产品 数据驱动增长

Reinventing Products with Intelligence, Driven by Data

80+	40+	10+	800+
业界大咖	创新案例	分论坛	听众规模
汇聚产品大咖的真知灼见	开拓全新AI+产品视角	涵盖产品到商业增长的全链路	共同探索产品的智能重构

扫码查看会议详情

AiDD 8th 2025 | 11月14-15日 | 深圳
第8届 AI+ Development Digital Summit
AI+研发数字峰会
拥抱AI 重塑研发



通义多模态、多端GUI智能体 Mobile-Agent

徐海洋 | 阿里巴巴-通义实验室

徐海洋

阿里巴巴 通义实验室高级算法专家

阿里通义实验室高级算法专家，负责通义Mobile-Agent、mPLUG等系列工作，包括多模态智能体Mobile-Agent、多模态大模型mPLUG/mPLUG-Owl/QwenVL，多模态文档大模型mPLUG-DocOwl等，其中mPLUG 工作在 VQA 榜单首超人类的成绩，Mobile-Agent工作CCL2024、2025两年 Best Demo，获得多个多模态榜单第一和Best Paper。在国际顶级期刊和会议ICML/NeurIPS/ICLR/CVPR/ICCV/ACL/EMNLP等发表论文60多篇，并担任多个顶级和会议AC/PC/Reviewer，主导参与开源项目Mobile-Agent，mPLUG，AliceMind，DELTA等。

目录
CONTENTS

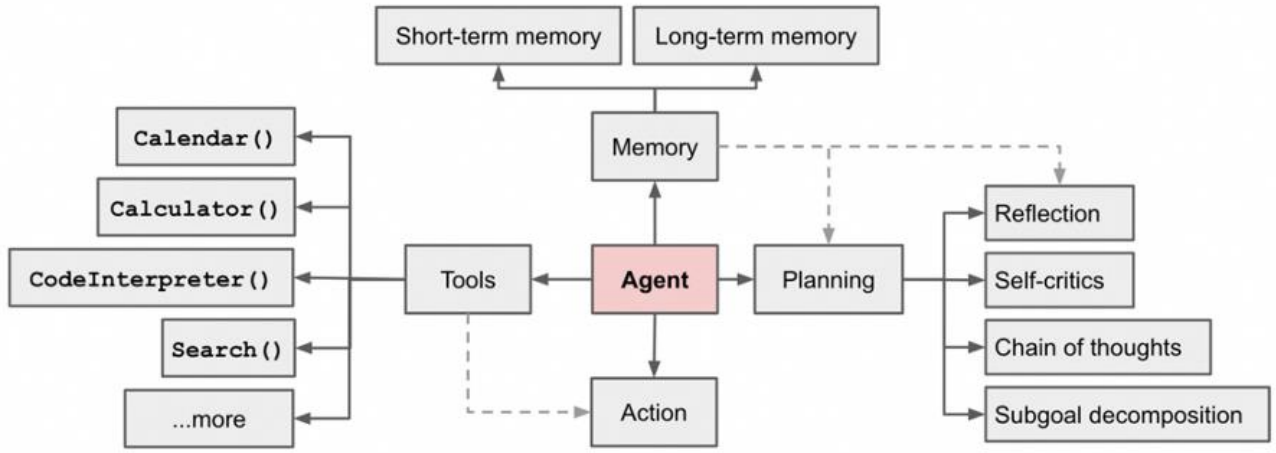
PART 01

大模型智能体背景

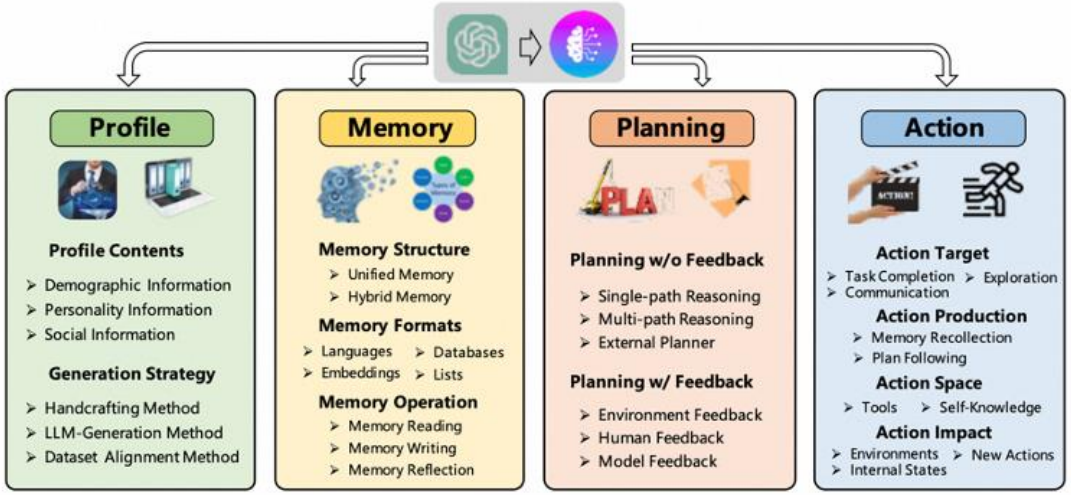
大模型智能体系统

在人工智能领域，AI智能体指可以观察周遭 **环境** 并作出 **行动** 以达致 **目标** 的 **自主** 实体

Agent System Overview from Lilian Weng's blog

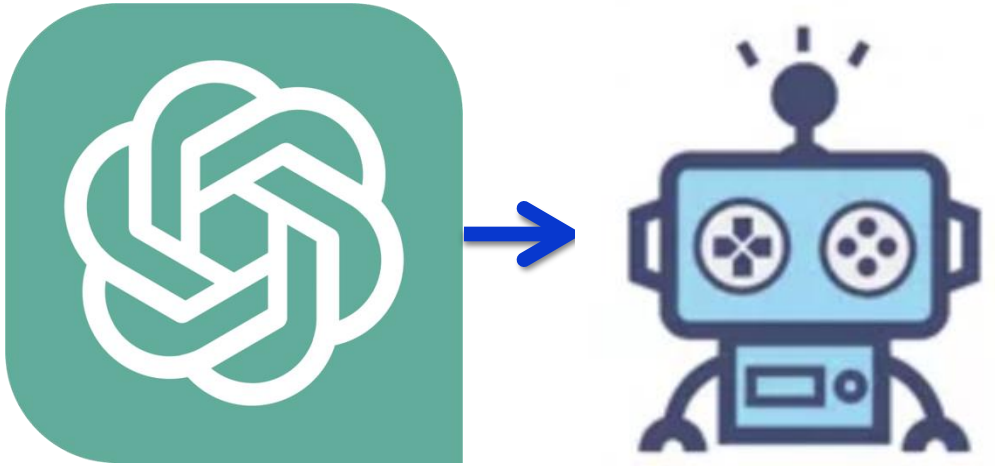


Wang et al. A Survey on Large Language Model based Autonomous Agents



大模型智能体的优势

OpenAI Five



LLM Agent with ChatGPT

传统基于RL的智能体的局限性

- 数据采样专有环境和低效
- 面向特定任务
- 稀疏奖励和长时段问题

大模型智能体的优势

- 丰富的世界知识
- 推理/规划能力
- 工具使用 (检索、code等)
- In-context Learning

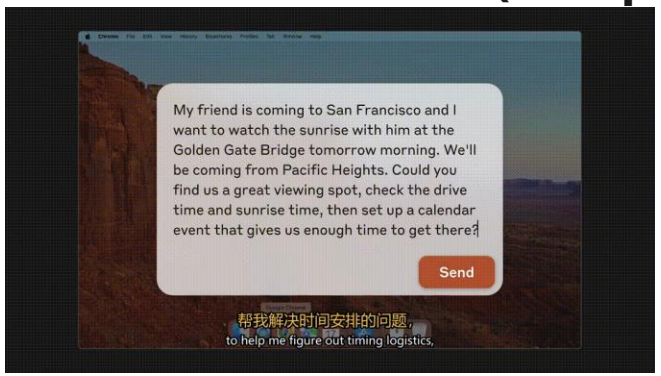
近期两类AI Agent应用

	Action Agent (GUI Agent)	Information Agent (DeepResearch)
作用	**[硬]“眼睛”&“手”** 环境感知和行动执行	**[软]“大脑”** 思考、规划和综合分析
适用场景	**[自动化]操作密集型** 办公、生活操作任务	**[智能化]知识密集型** 办公Search创作场景
示例	Operator、Apple Intelligence、 Claude、Mobile-Agent	Deep Research (OpenAI、谷歌、Qwen)
	ChatGPT-Agent、Manus	

GUI智能体发展迅速

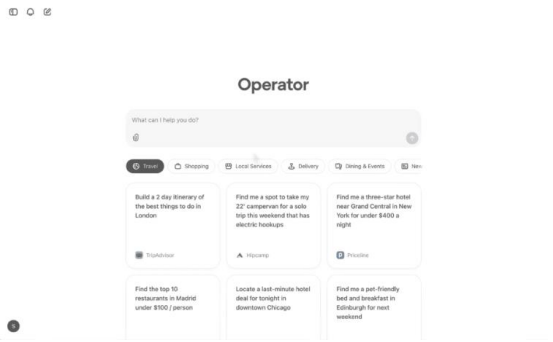
围绕Mobile、PC、Web的**GUI-Agent**是未来的**重要技术趋势**之一，替代人类操作、提升生产效率。

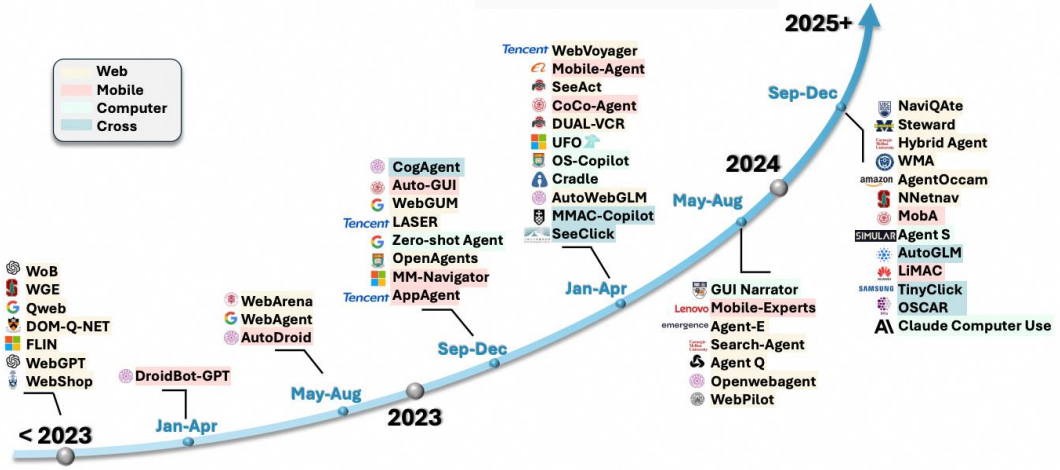
Claude 3.5 Sonnet (Computer)



Apple Intelligence

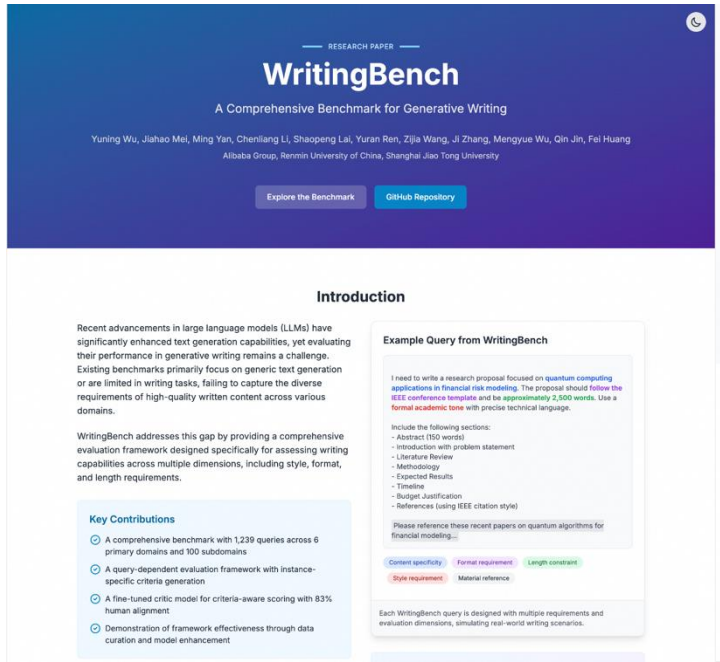
OpenAI Operator

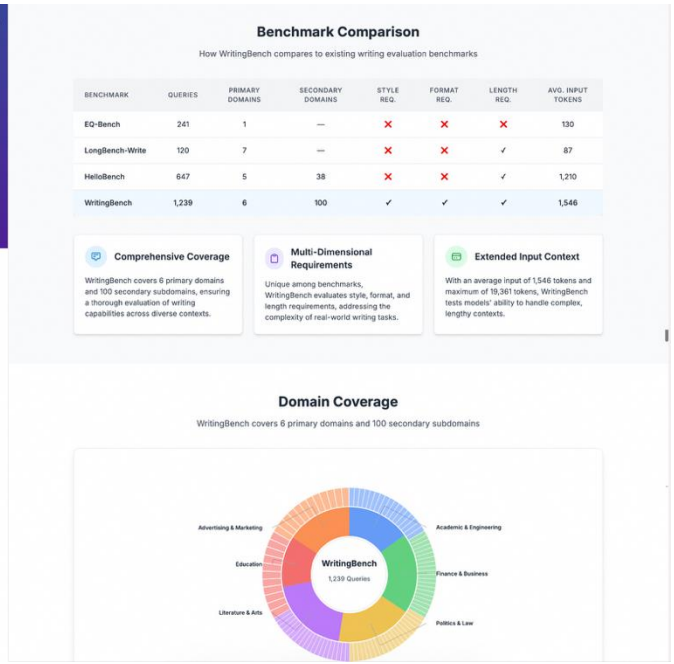


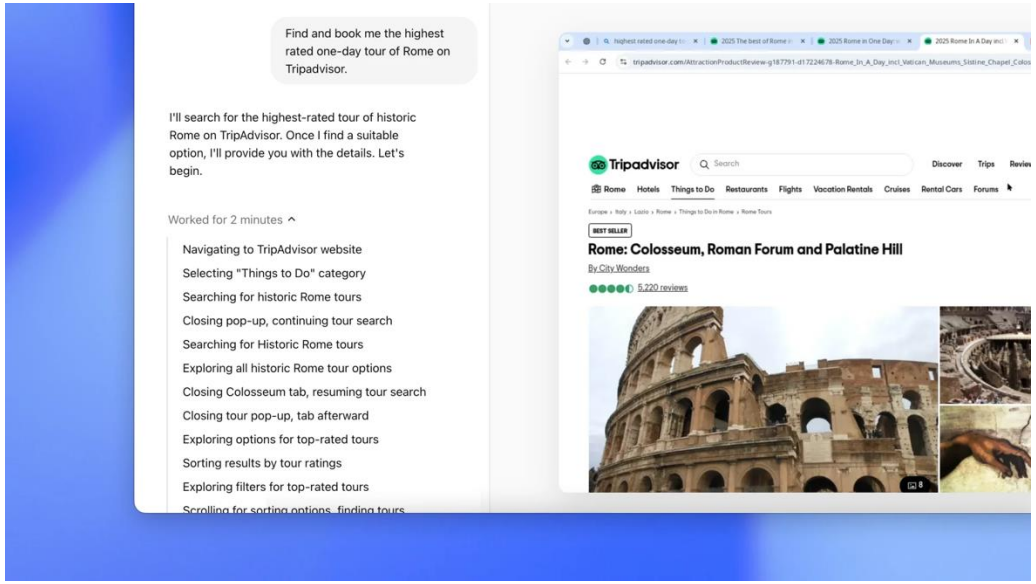


GUI大模型智能体

现实世界是需要 **多模态环境交互** 的，多模态智能体可能衍生出更多Super、Fancy应用







Claude 3.7 sonnet (computer use)

参照人类思考系统的快速反应与慢反思结合的工作模式,将LLM快速响应和思维链深度思考

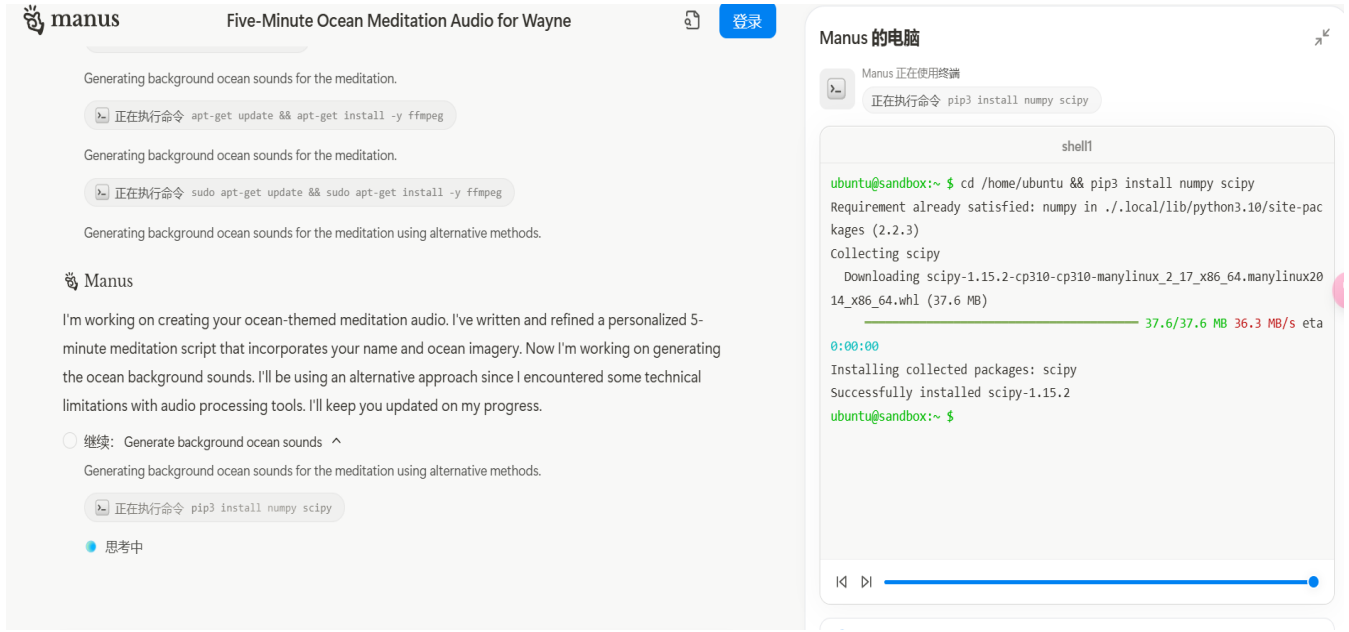
Operator

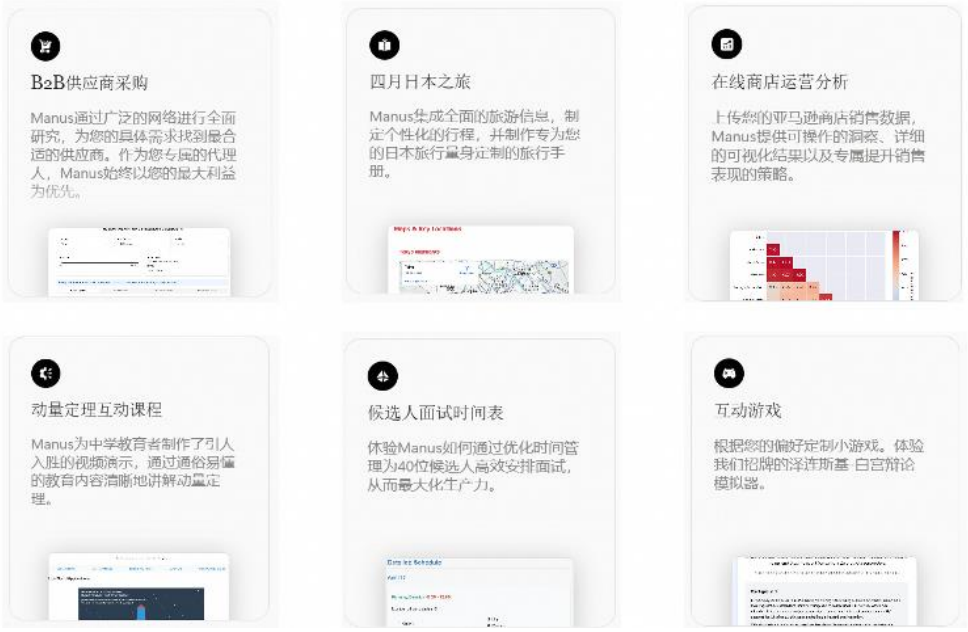
基于Computer-Using Agent 模型，结合GPT-4o的视觉理解能力和强化学习习得的推理能力，自动执行鼠标和键盘的组合操作，无需API，具备推理思维链和自动纠错能力

大模型通用型智能体系统

从基于检索提供信息，到Agent执行任务的本质进阶

（1）规划-执行Tool-反思；（2）操作上云；（3）快操作 + 慢思考





Manus/Open Manus

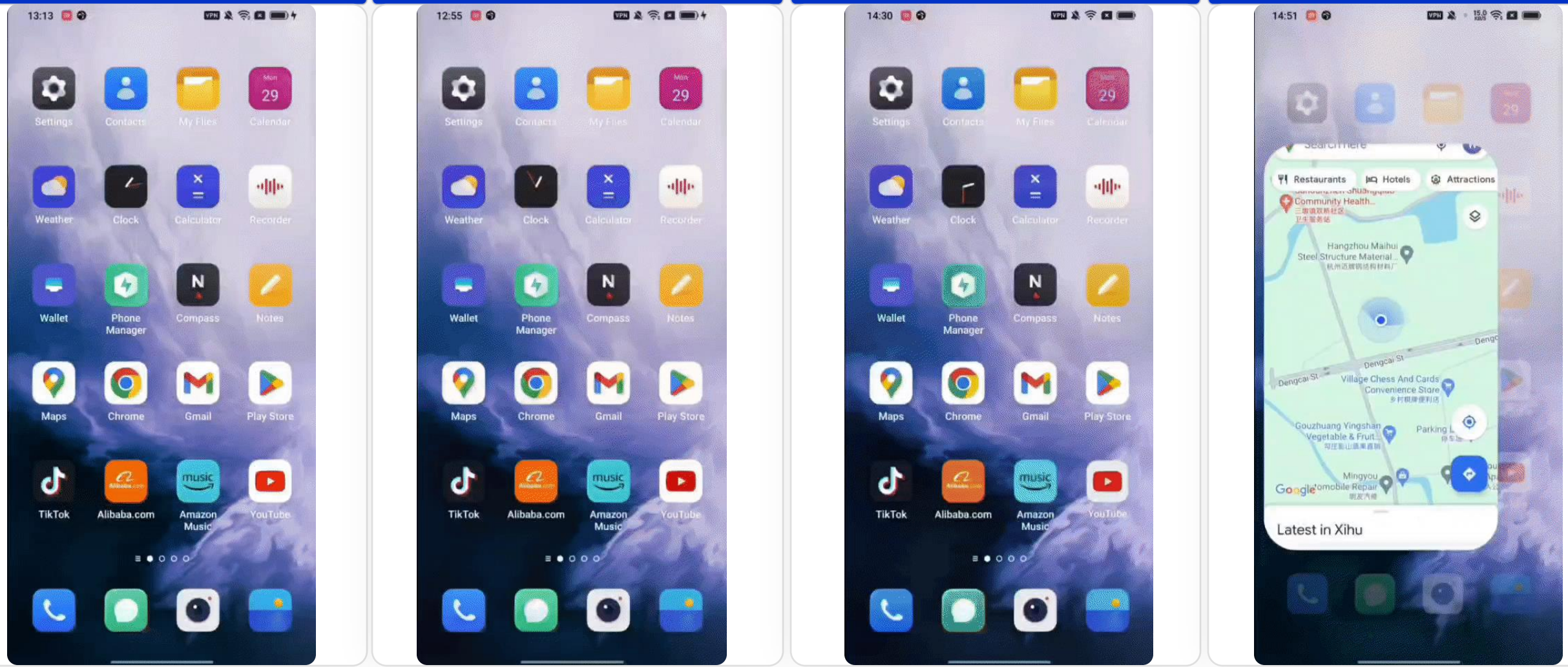
Manus强调“需求→规划→执行→交付”全流程自动化，无需用户持续指导便可能直接生成可交付成果，动态调整执行路径，在解决现实世界问题方面表现卓越

PART 02

多模态多端智能体Mobile-Agent

多模态多端智能体Mobile-Agent

分析天气	搜索视频并评论	刷短视频并点赞	导航



多模态多端智能体Mobile-Agent

在微博中搜索GTC2025的时间，然后在微信的GTC2025参会群中提醒大家

在小红书搜西湖附近的特色餐厅，用高德地图导航过去

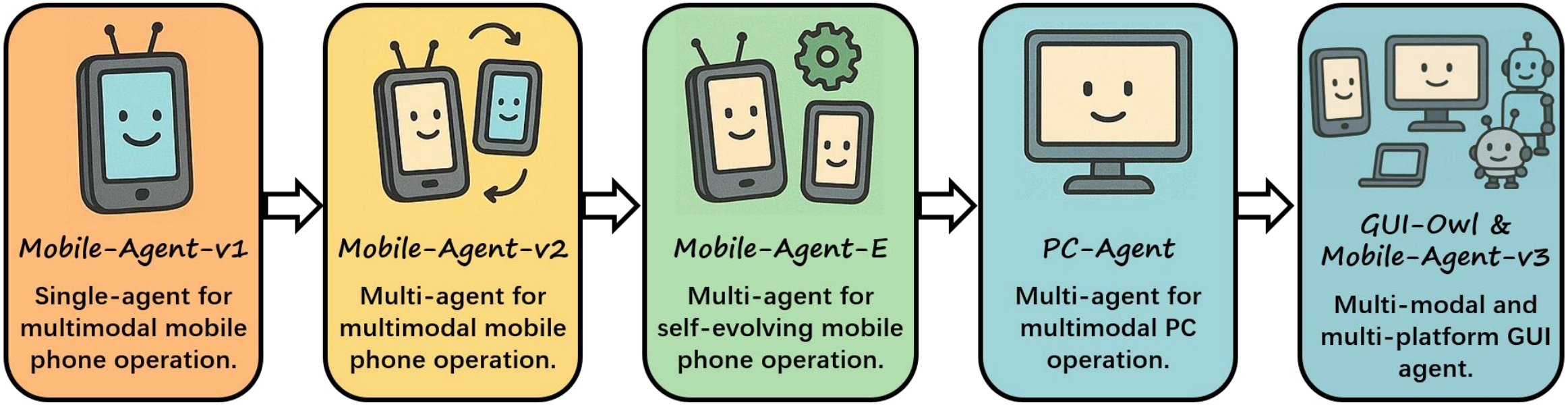
PC Agent

Institute for Intelligent Computing of Alibaba Group

https://arxiv.org/abs/2502.14282

Github: https://github.com/X-PLUG/MobileAgent/tree/main/PC-Agent

多模态多端智能体Mobile-Agent

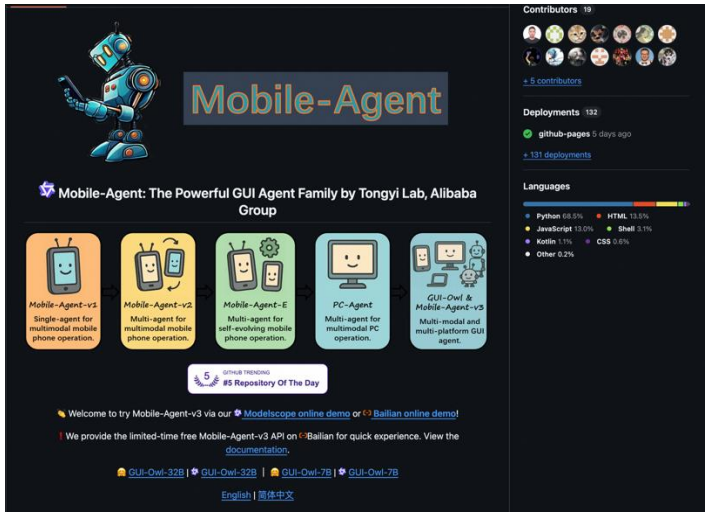


GUI场景核心挑战

核心挑战

多模态多端智能体Mobile-Agent

CCL2024, CCL 2025 Highlight System



Github 6.1k stars

	时间	方案	平均端到端完成率	单步RT
V1	2024.1 ICLR2024	基于GPT4o单Agent	75%	30s
V2	2024.6 NeuIPS2024	基于GPT4o多Agent	80%	60s
V3-Preview	2024.8 CCL Best Demo 云栖大会	基于QwenVL的多Agent	75%	10s
V3-E	2025.2	记忆增强、自主进化	85%	5s
V3-Full	2025.8	端到端模型、多Agent适配	90%	2.5s

多模态多端智能体Mobile-Agent-V1

历史操作 & 截屏

用户指令

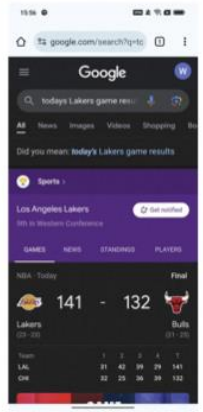
搜索今天湖人队的比赛结果，然后在笔记中写一个战况分析

当前屏幕

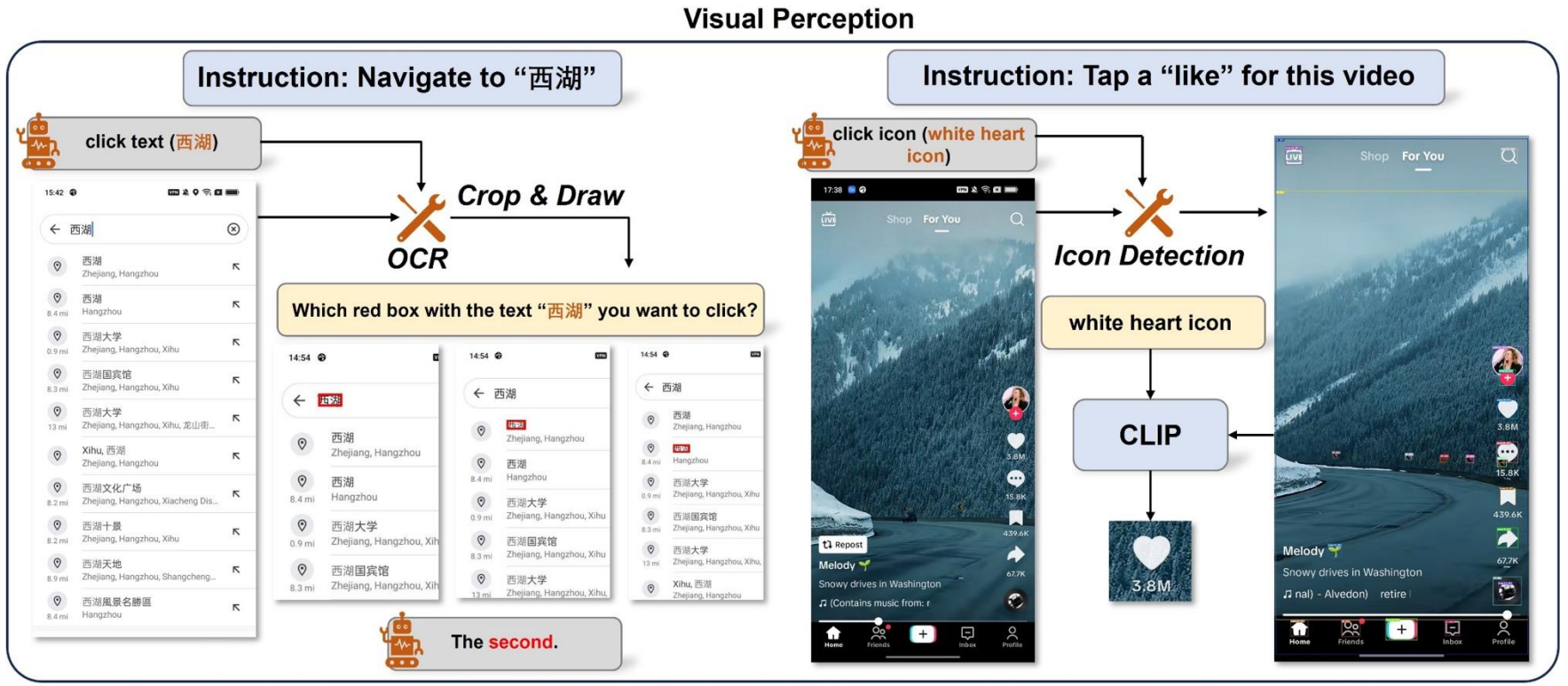
观察：描述当前页面的情况

思考：下一步的操作思路

行动：使用工具完成思考中的操作



多模态多端智能体Mobile-Agent-V1



行为空间

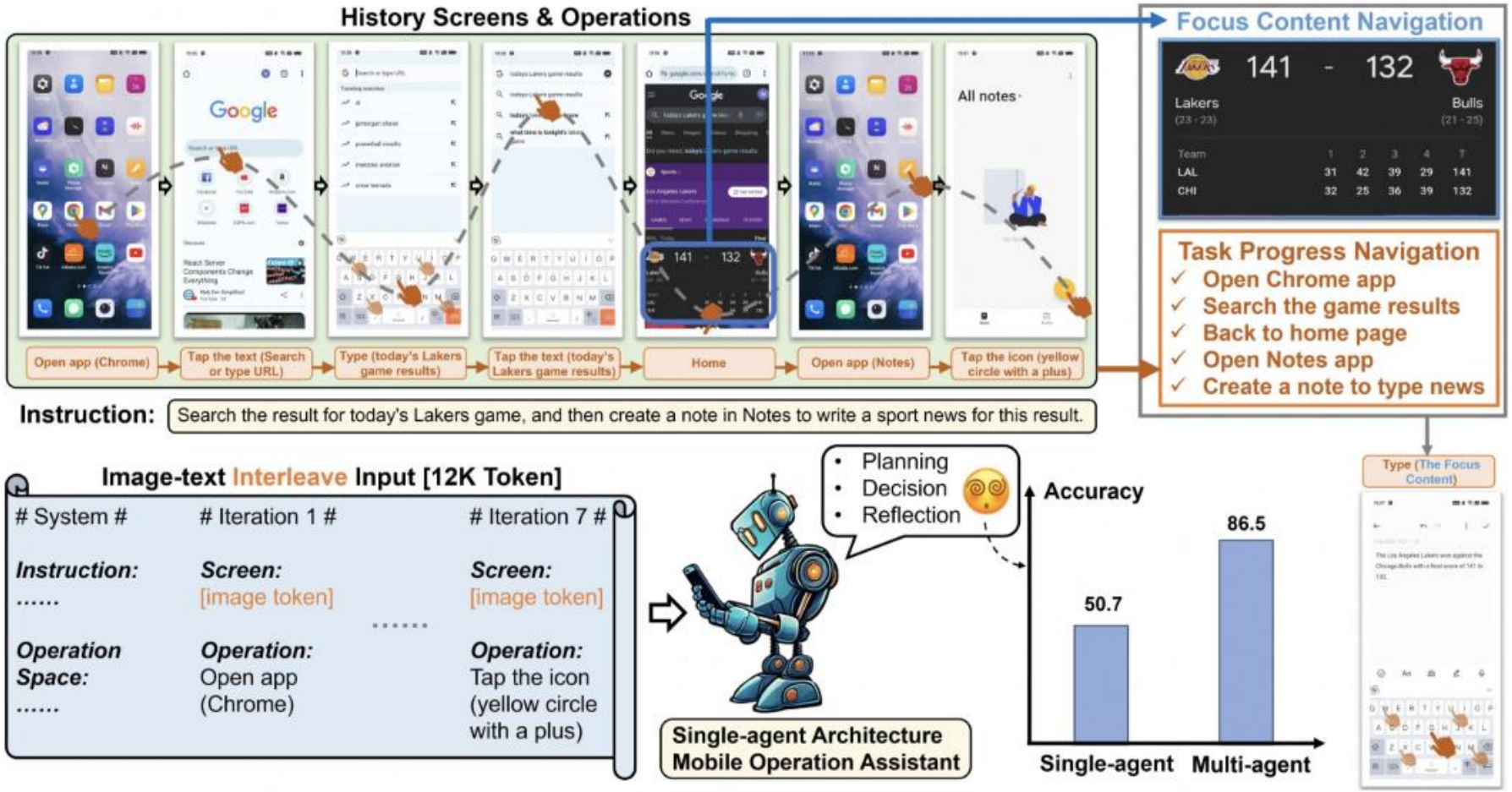
1. 点击文本
2. 点击图标
3. 打字
4. 上划 & 下划
5. 返回上一页面
6. 返回桌面
7. 结束

大模型缺乏输出精确坐标的grounding能力

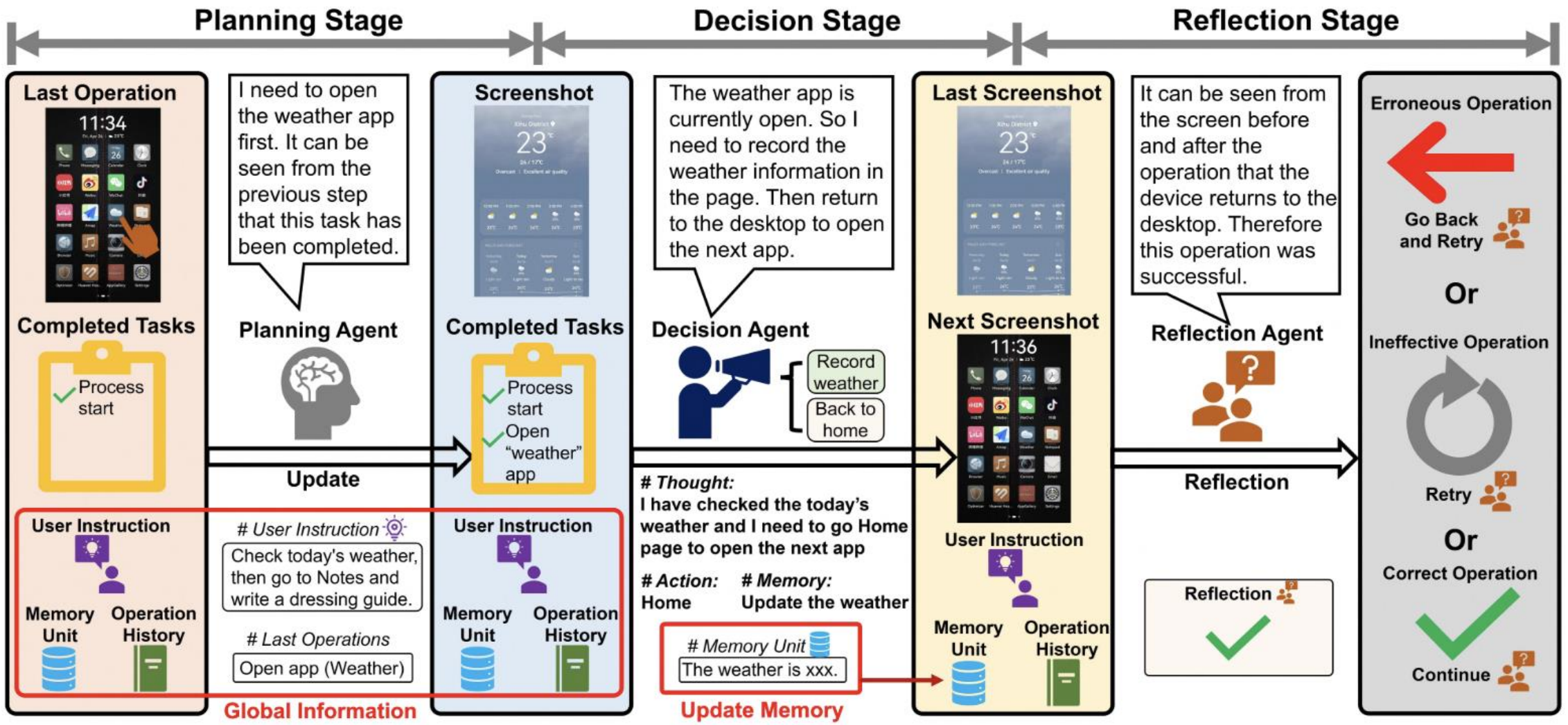
- 屏幕文本定位：使用OCR工具检测识别文本框
- 图标定位：使用图标分割检测工具检测所有图标和位置

多模态多端智能体Mobile-Agent-V2

冗长并且图文交错格式的操作历史，会大大增加智能体追踪任务进度的难度



多模态多端智能体Mobile-Agent-V2



多模态多端智能体Mobile-Agent-V2

动态评测：5个系统内置应用和5个第三方应用，每个APP和多个APP各2条基础指令和2条进阶指令

Method	Basic Instruction				Advanced Instruction			
	SR	CR	DA	RA	SR	CR	DA	RA
	System app							
Mobile-Agent	5/10	41.2	37.6	-	3/10	37.3	32.9	-
Mobile-Agent-v2	9/10	86.8	82.5	93.3	6/10	82.7	78.2	84.4
Mobile-Agent-v2 + *Know.*	10/10	97.5	98.2	98.9	8/10	88.9	87.2	91.4
	External app							
Mobile-Agent	2/10	38.3	35.4	-	1/10	29.2	27.0	-
Mobile-Agent-v2	8/10	97.9	94.0	92.5	5/10	77.9	74.1	78.8
Mobile-Agent-v2 + *Know.*	10/10	99.1	95.6	97.3	8/10	87.8	83.0	85.9
	Multi-app							
Mobile-Agent	1/2	52.8	50.0	-	0/2	33.3	31.4	-
Mobile-Agent-v2	2/2	100	92.9	91.6	2/2	100	93.8	92.9
Mobile-Agent-v2 + *Know.*	-	-	-	-	-	-	-	-

Table 1: Dynamic evaluation results on non-English scenario, where the *Know.* represents manually injected operation knowledge.

Method	Basic Instruction				Advanced Instruction			
	SR	CR	DA	RA	SR	CR	DA	RA
	System app							
Mobile-Agent	9/10	92.5	89.7	-	4/10	62.0	71.3	-
Mobile-Agent-v2	9/10	95.0	92.9	96.5	6/10	76.0	77.6	88.4
Mobile-Agent-v2 + *Know.*	10/10	100	96.2	98.7	8/10	85.3	87.9	92.0
	External app							
Mobile-Agent	7/10	79.7	72.0	-	3/10	45.3	38.7	-
Mobile-Agent-v2	9/10	97.1	93.8	96.2	7/10	89.7	91.0	93.4
Mobile-Agent-v2 + *Know.*	10/10	100	98.2	97.4	9/10	97.1	94.2	98.5
	Multi-app							
Mobile-Agent	2/2	100	91.2	-	1/2	86.7	92.9	-
Mobile-Agent-v2	2/2	100	97.4	100	1/2	93.3	93.3	80.0
Mobile-Agent-v2 + *Know.*	-	-	-	-	2/2	100	100	100

Table 2: Dynamic evaluation results on English scenario, where the *Know.* represents manually injected operation knowledge.

Metrics. We design the following four metrics for dynamic evaluation:

- Success Rate (SR): When all the requirements of a user instruction are fulfilled, the agent is considered to have successfully executed this instruction. The success rate refers to the proportion of user instructions that are successfully executed.
- Completion Rate (CR): Although some challenging instructions may not be successfully executed, the correct operations performed by the agent are still noteworthy. The completion rate refers to the proportion of correct steps out of the ground truth operations.
- Decision Accuracy (DA): This metric reflects the accuracy of the decision by the decision agent. It is the proportion of correct decisions out of all decisions.
- Reflection Accuracy (RA): This metric reflects the accuracy of reflection by the reflection agent. It is the proportion of correct reflections out of all reflections.

Mobile-Agent-E：解决复杂任务、自主进化

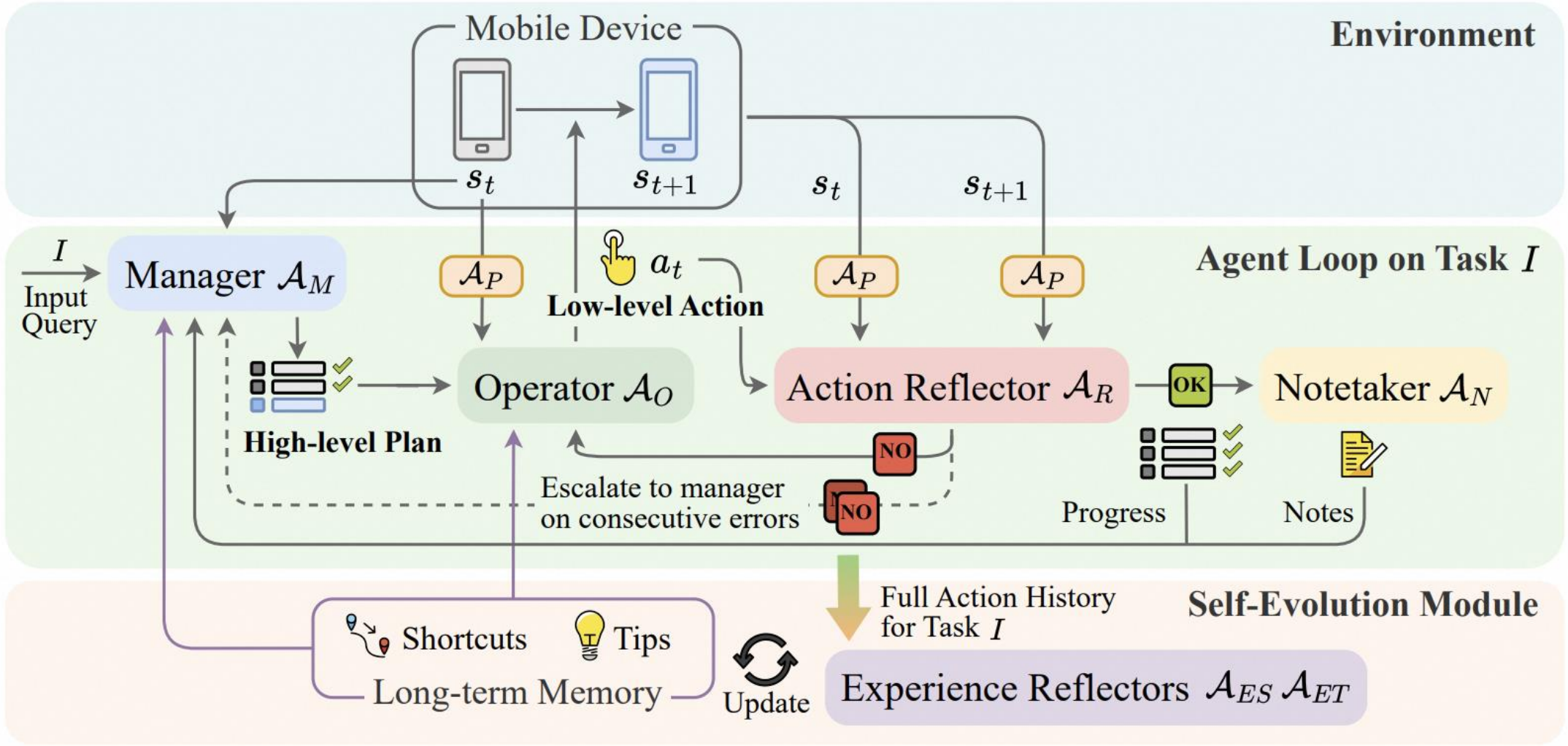
复杂指令：

执行复杂推理、多步规划以及跨App操作

自我进化：

反思过往的任务记录，从经验中学习，自动生成Tips和Shortcut

Mobile-Agent-E：解决复杂任务、自主进化



Mobile-Agent-E：解决复杂任务、自主进化

Model	Type	Satisfaction Score (%) ↑	Action Accuracy (%) ↑	Reflection Accuracy (%) ↑	Termination Error (%) ↓
AppAgent (Zhang et al., 2023)	Single-Agent	25.2	60.7	-	96.0
Mobile-Agent-v1 (Wang et al., 2024b)	Single-Agent	45.5	69.8	-	68.0
Mobile-Agent-v2 (Wang et al., 2024a)	Multi-Agent	53.0	73.2	96.7	52.0
Mobile-Agent-E	Multi-Agent	75.1	85.9	97.4	32.0
Mobile-Agent-E + Evo	Multi-Agent	**86.9**	**90.4**	**97.8**	**12.0**

Model	Gemini-1.5-pro				Claude-3.5-Sonnet				GPT-4o			
	SS↑	AA↑	RA↑	TE↓	SS↑	AA↑	RA↑	TE↓	SS↑	AA↑	RA↑	TE↓
Mobile-Agent-v2 (Wang et al., 2024a)	50.8	63.4	83.9	64.0	70.9	76.4	96.9	32.0	53.0	73.2	96.7	52.0
Mobile-Agent-E	70.9	74.3	**91.3**	**48.0**	75.5	91.1	99.1	**12.0**	75.1	85.9	97.4	32.0
Mobile-Agent-E + Evo	**71.2**	**77.4**	89.6	**48.0**	**83.0**	**91.4**	**99.7**	**12.0**	**86.9**	**90.4**	**97.8**	**12.0**

AiDD 8th 2025

PART 03

Foundation Agent for GUI

Qwen2.5-VL：认识世界到理解世界

认知现实物体能力

经典地标、日常食物、动植物、汽车、商品标签识别，准确率大幅提升

Mobile Agent 能力

作为GUI智能体与移动设备进行交互，能够基于截屏执行click、type、home等动作，完成用户指令

Grounding能力

通过生成 bounding boxes 或 points，准确定位图像中的物体，并能够为坐标和属性提供稳定的 JSON 输出

“框”出图中的摩托车并标识驾驶员是否戴头盔

OCR能力

多语言、跨语言
经典OCR任务识别能力大幅提升

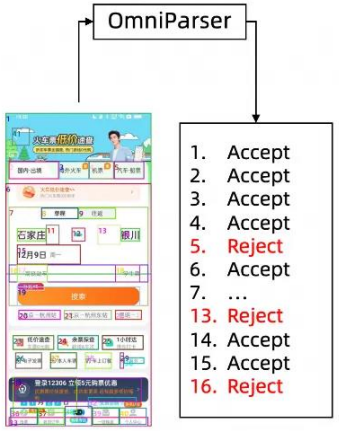
Qwen2.5-VL SMK أبو منير لبيع وصيانة الرونيترات رونيترات ماء - مكيف - دفايات COOLING CAR SYSTEM 0796-831-059 محمد أبو سيراج 5334-204-052 أبو منير 8256-811-056

文本理解能力

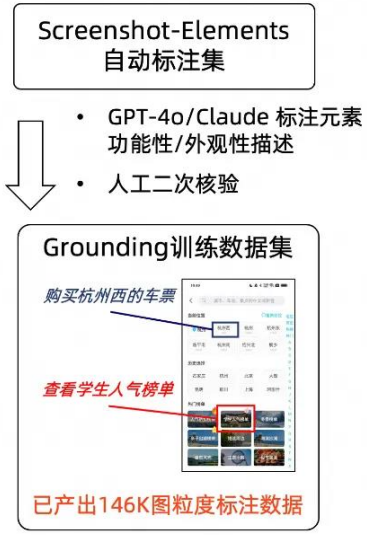
QwenVL HTML：更全面的文档解析格式
精准识别文档中的文本，也能够提取文档元素（如图片、表格等）的位置信息

Qwen2.5-VL：GUI-Agent能力提升

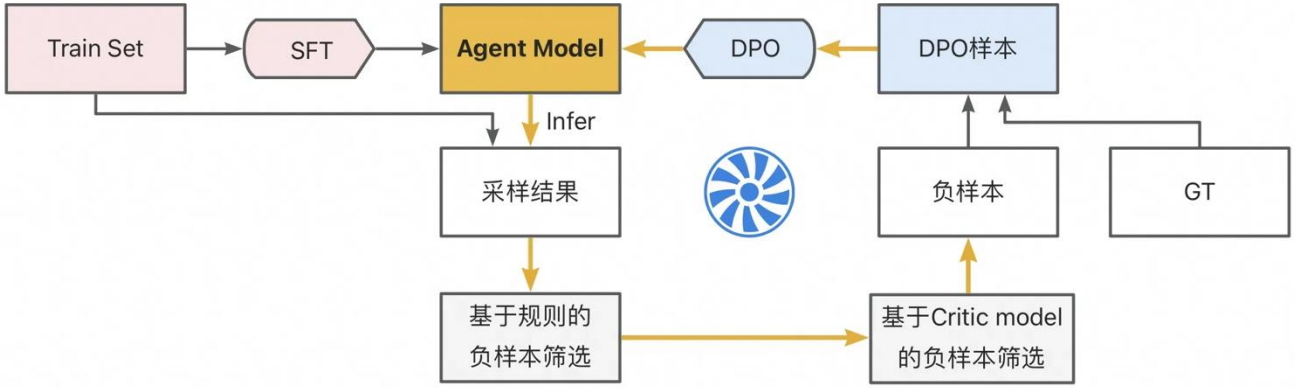
原始数据预处理



OmniParser - XML元素交叉清洗



Grounding指令打标校验



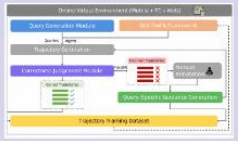
Benchmarks	GPT-4o	Gemini 2.0	Claude	Aguvis-72B	Qwen2-VL-72B	Qwen2.5-VL-72B
ScreenSpot	18.1	84.0	83.0	**89.2**	-	87.1
ScreenSpot Pro	-	-	17.1	23.6	1.6	**43.6**
Android Control High_{EM}	20.8	28.5	12.5	66.4	59.1	**67.36**
Android Control Low_{EM}	19.4	60.2	19.4	84.4	59.2	**93.7**
$\text{AndroidWorld}_{\text{SR}}$	34.5% (SoM)	26% (SoM)	27.9%	26.1%	6% (SoM)	**35%**
$\text{MobileMiniWob++}_{\text{SR}}$	61%	42% (SoM)	61% (SoM)	66%	50% (SoM)	**68%**
OSWorld	5.03	4.70	**14.90**	10.26	2.42	8.83

Mobile-Agent-V3 & GUI-Owl

Multi-Platform Environment Supporting

Highlight Capability

Large-scale Environment Infrastructure



- Cloud-based, cross-platform virtual environment
- Self-Evolving GUI Trajectory Production framework

Diverse Foundational Agents Capability

- Offline Hint-Guided Rejection Sampling
- Distillation from Multi-Agent Framework
- Iterative Online Rejection Sampling

SOTA end-to-end model
Drive different multi-agent frameworks

Scalable Environment RL

- Decoupled rollout-update framework
- Support parallel running

基建架构

ECS集群

虚拟化环境

- Android 模拟器
- Windows/Linux 模拟器
- Web 服务器

轨迹存储 ↓ 模型调用 ↑

OSS对象存储服务

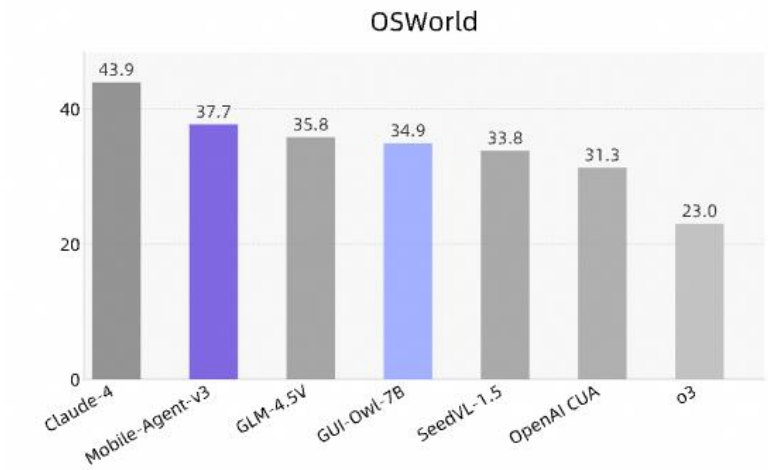
百炼模型服务部署

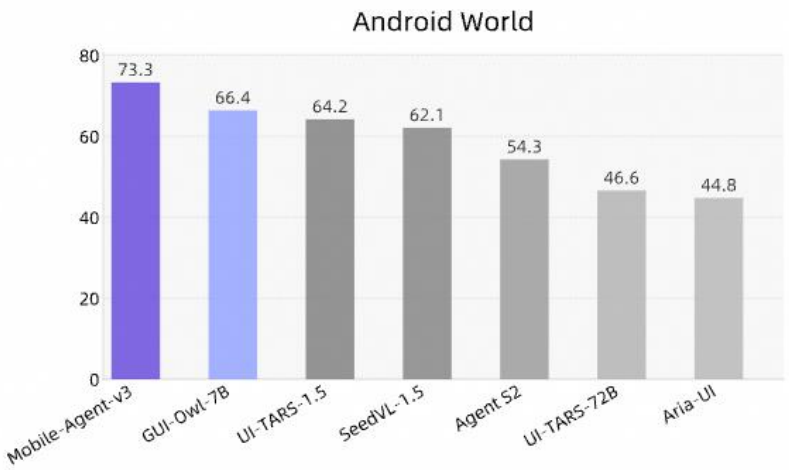
轨迹回流 ↓ 模型部署 ↑

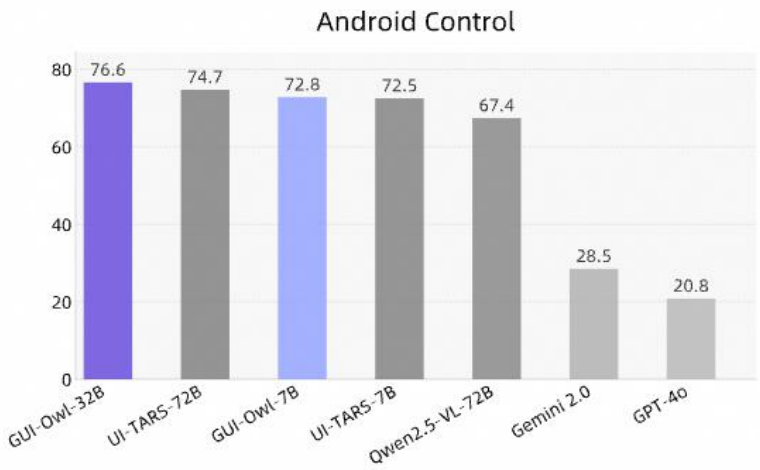
PAI GPU集群

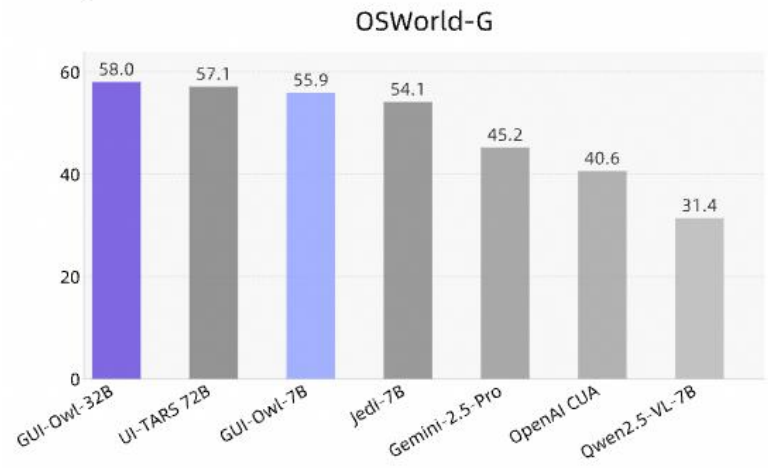
SFT/RL 模型训练

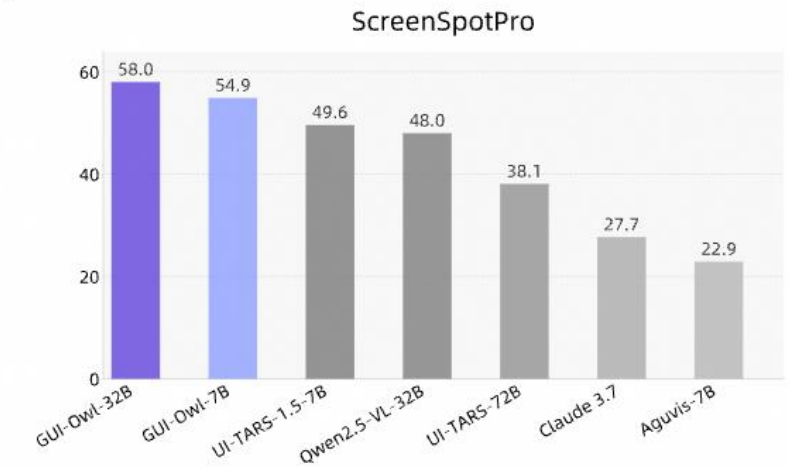
Mobile-Agent-V3 & GUI-Owl

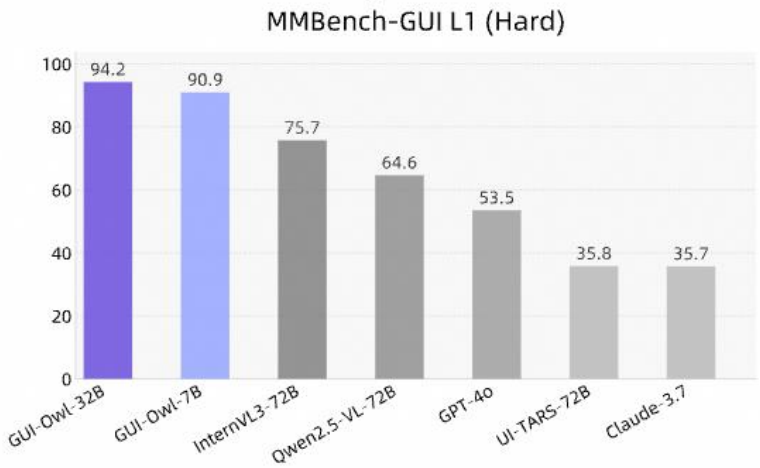




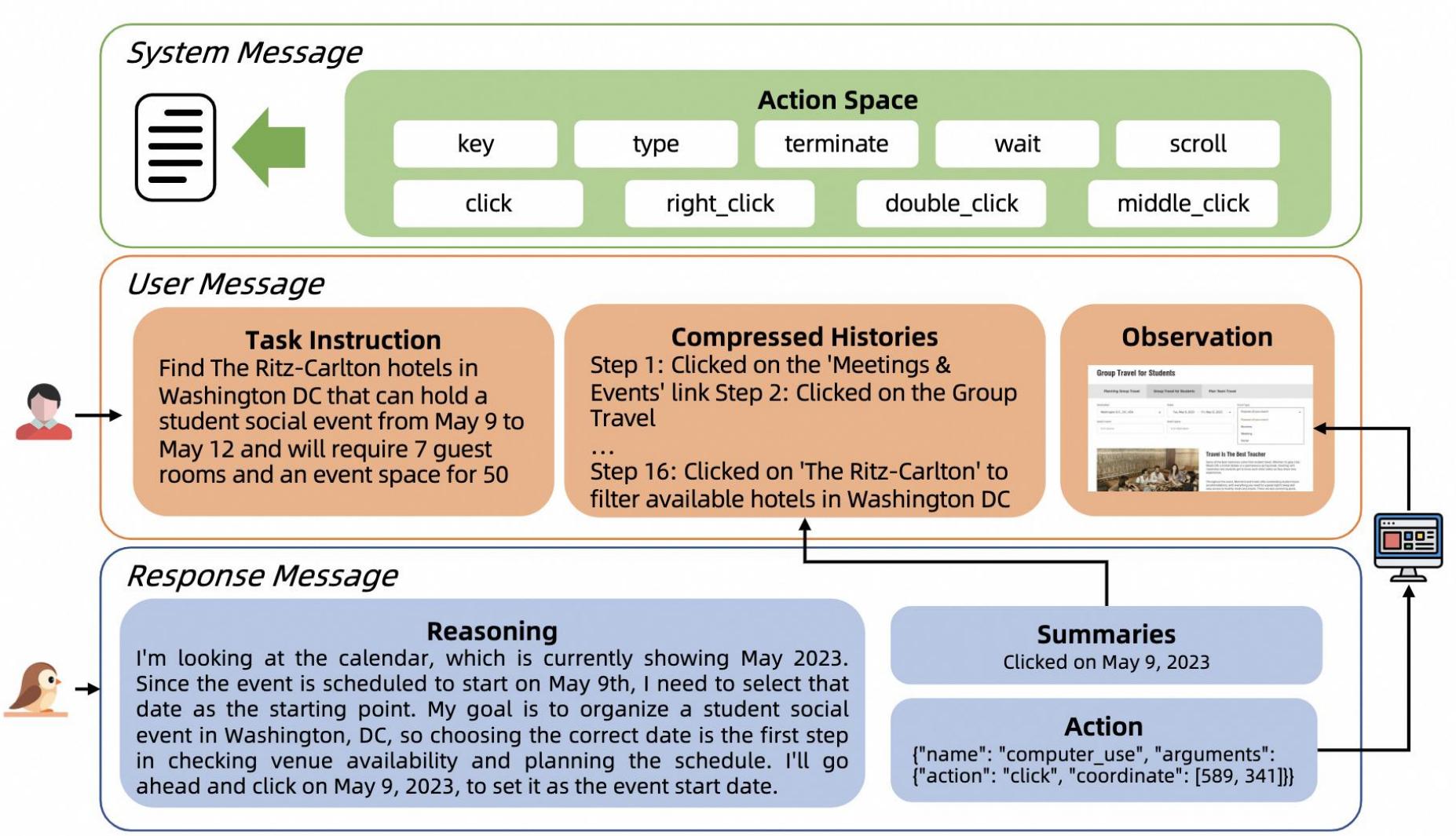




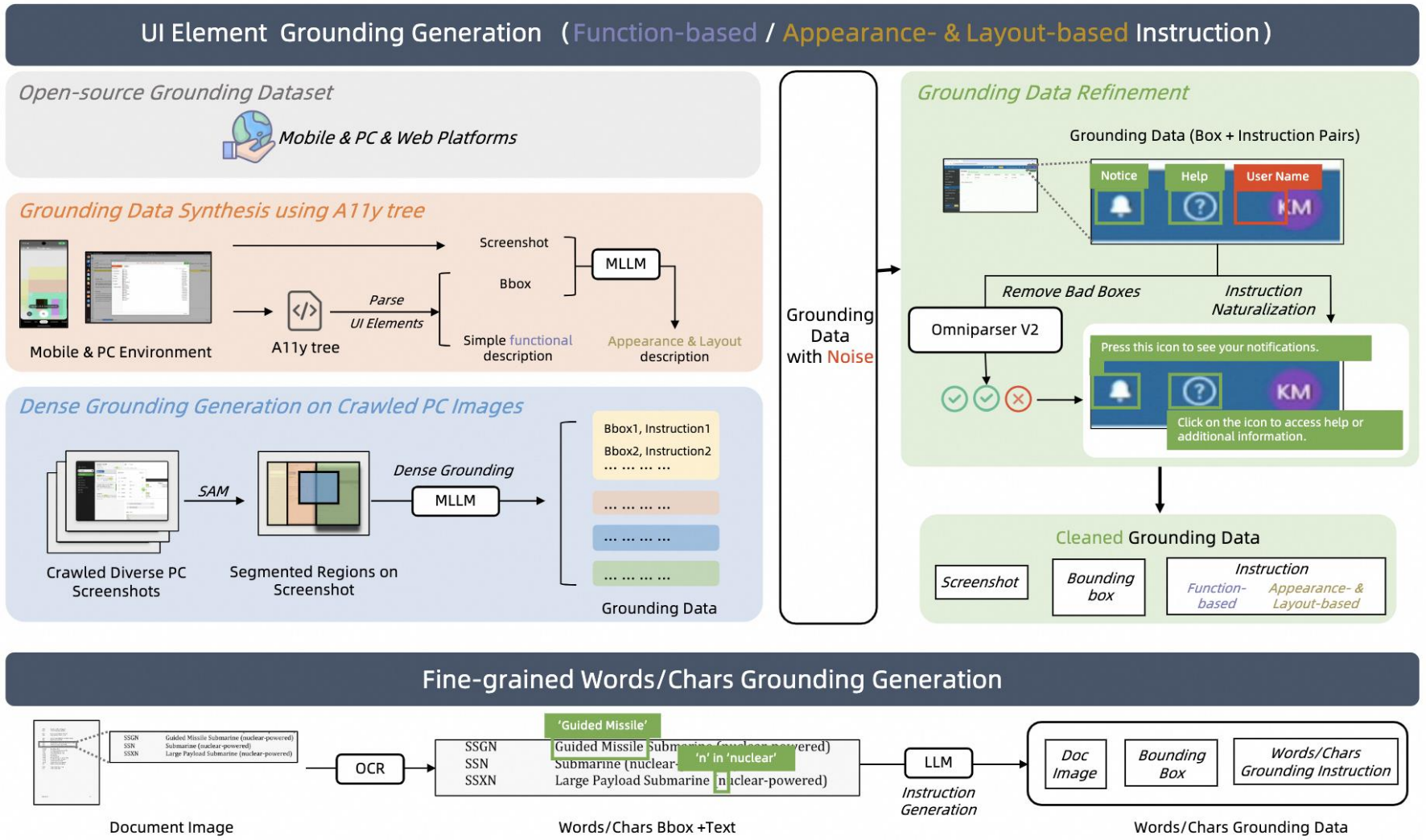




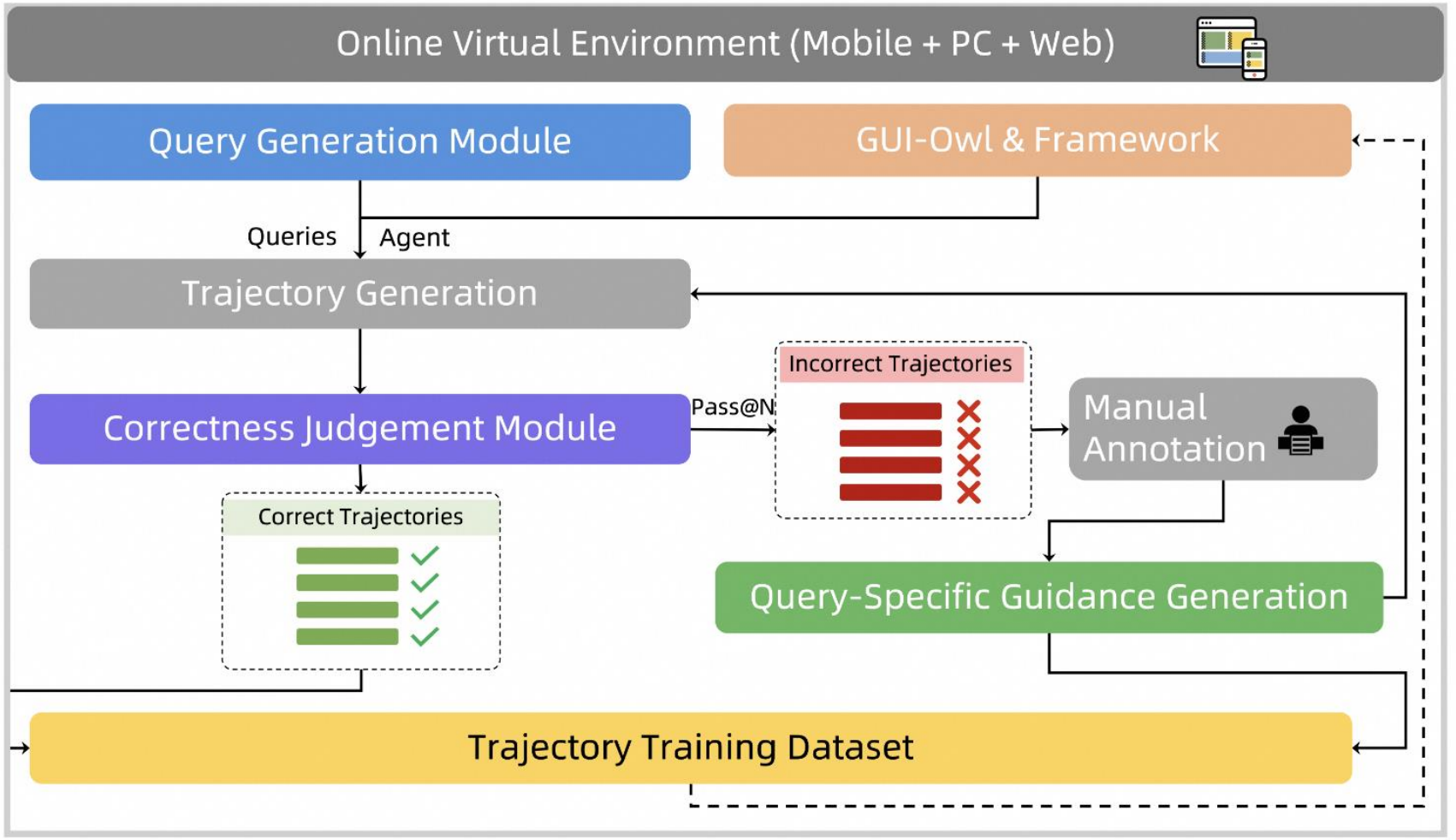
GUI-Owl整体交互flow



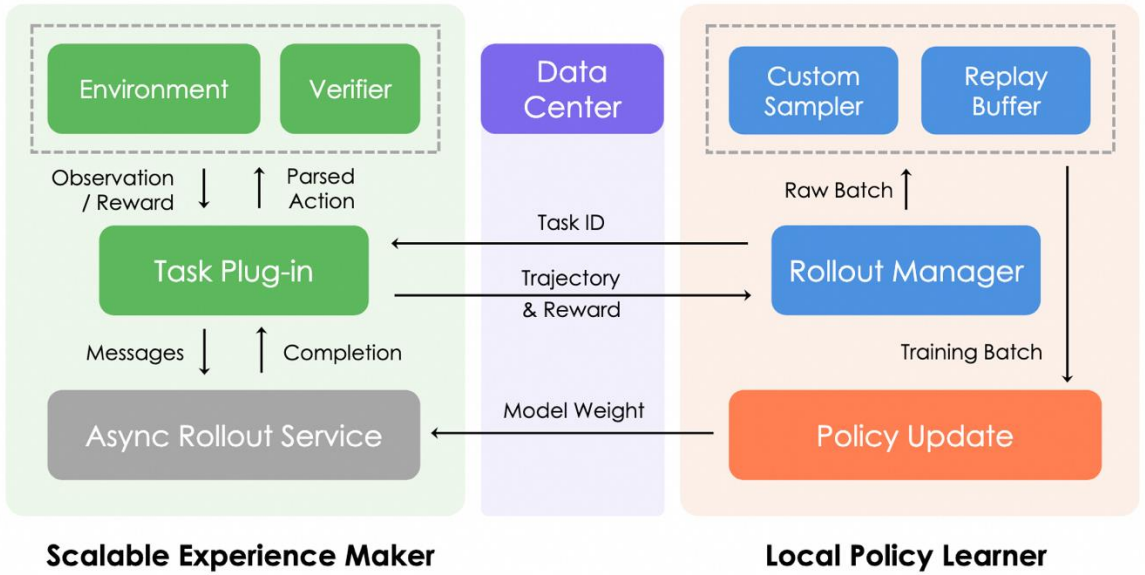
GUI-Owl数据合成链路-Grounding

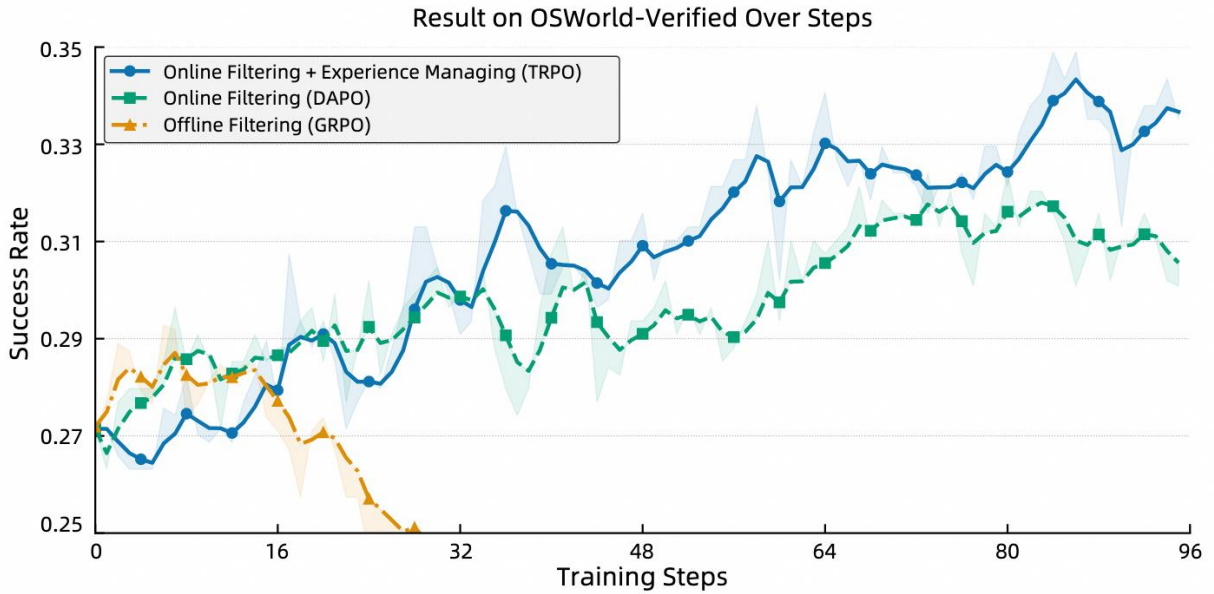


GUI-Owl数据合成链路-自进化轨迹合成



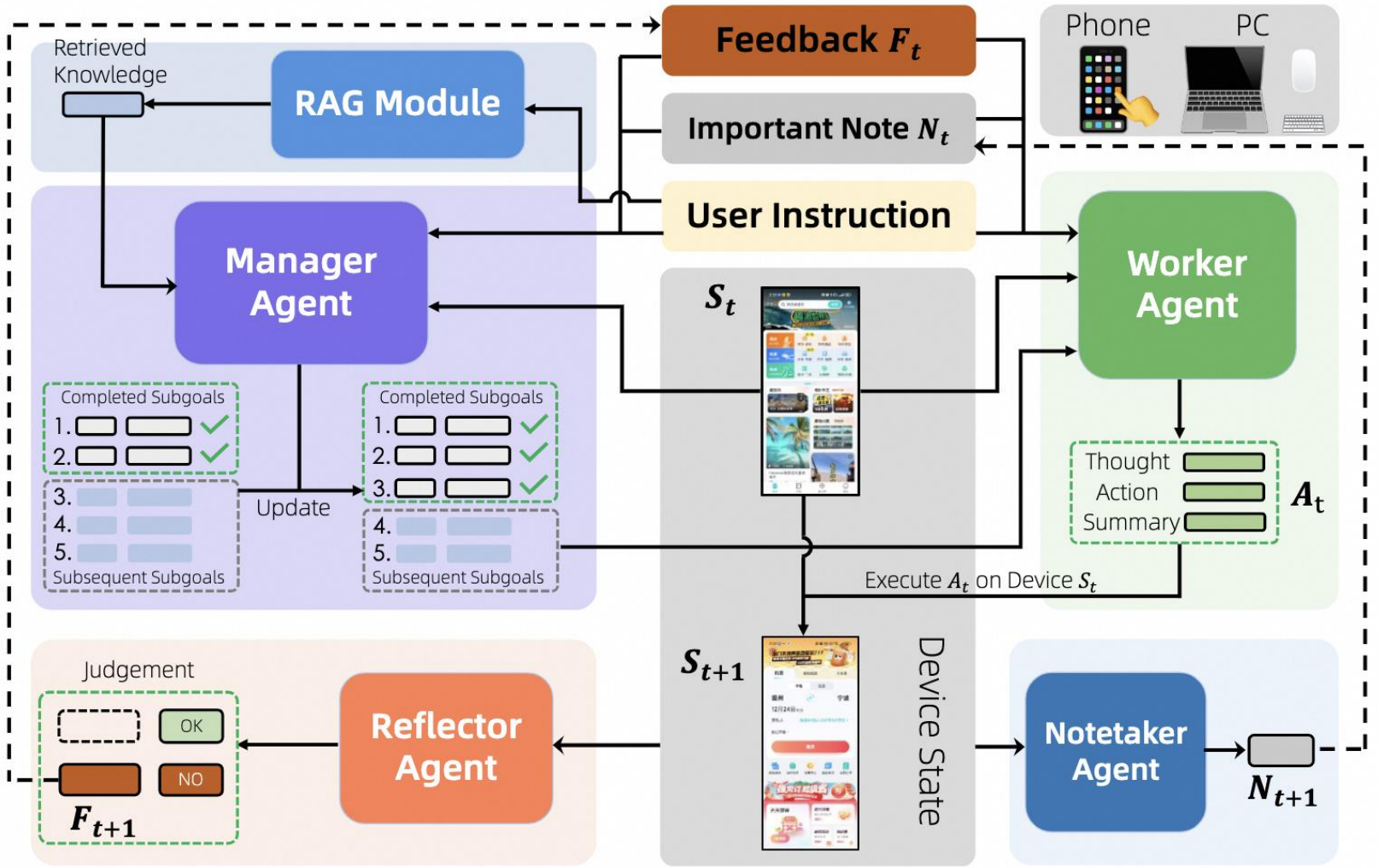
Mobile-Agent Agentic RL能力提升



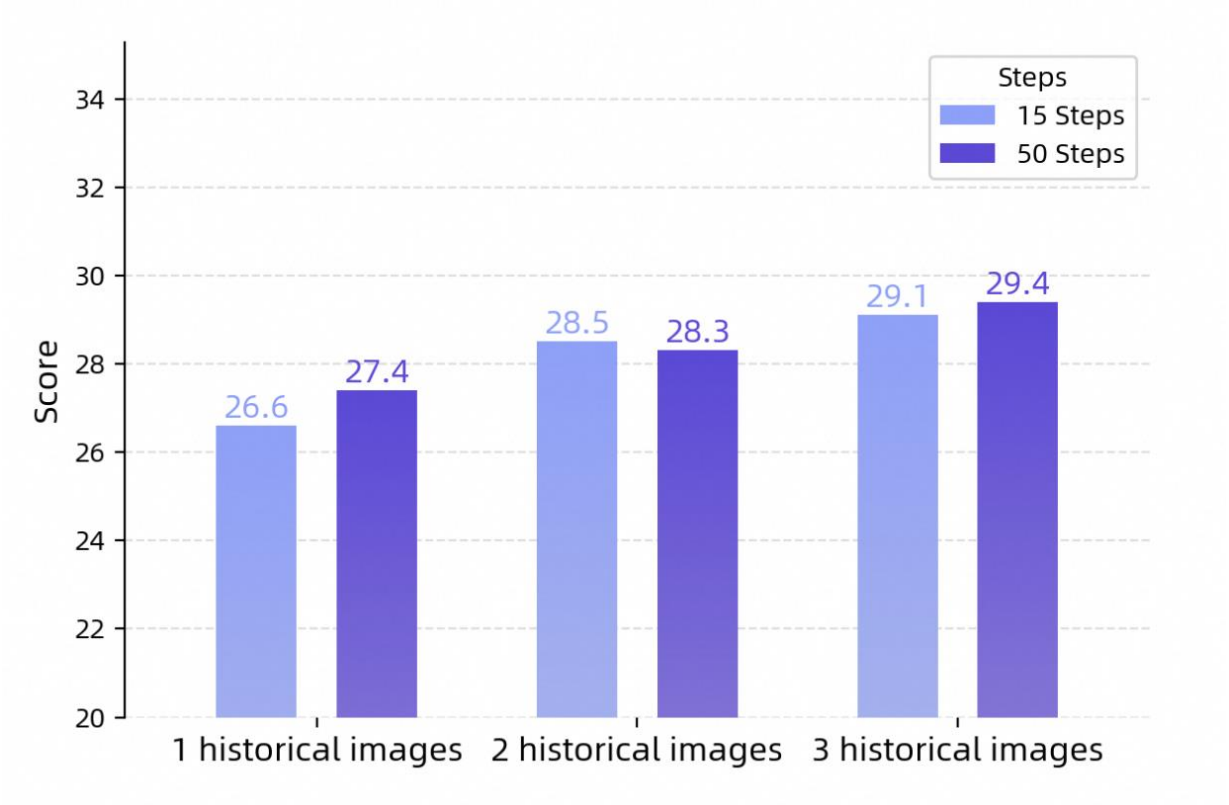


$$\mathcal{L}_{\text{TRPO}} = -\frac{1}{N}\sum_{i=1}^{G}\sum_{s=1}^{S_i}\sum_{t=1}^{|\mathbf{o}_{i,s}|}\left\{\min\left[r_t(\theta)\hat{A}_{\tau_i},\ \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_{\tau_i}\right]\right\}$$

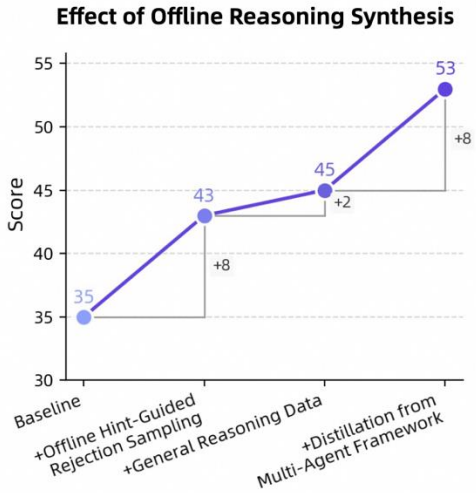
Mobile-Agent V3 Agent框架

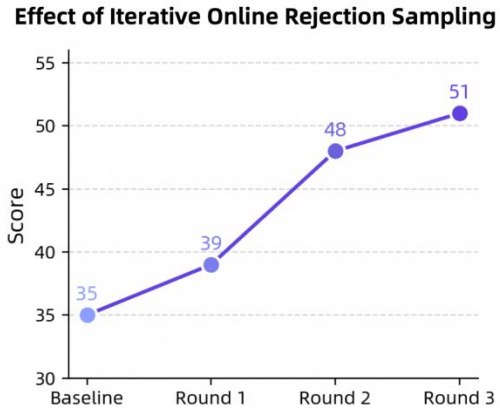


Mobile-Agent V3实验



Scaling of Historical Images and maximum interaction Steps





Effect of Reasoning data synthesis on Android World

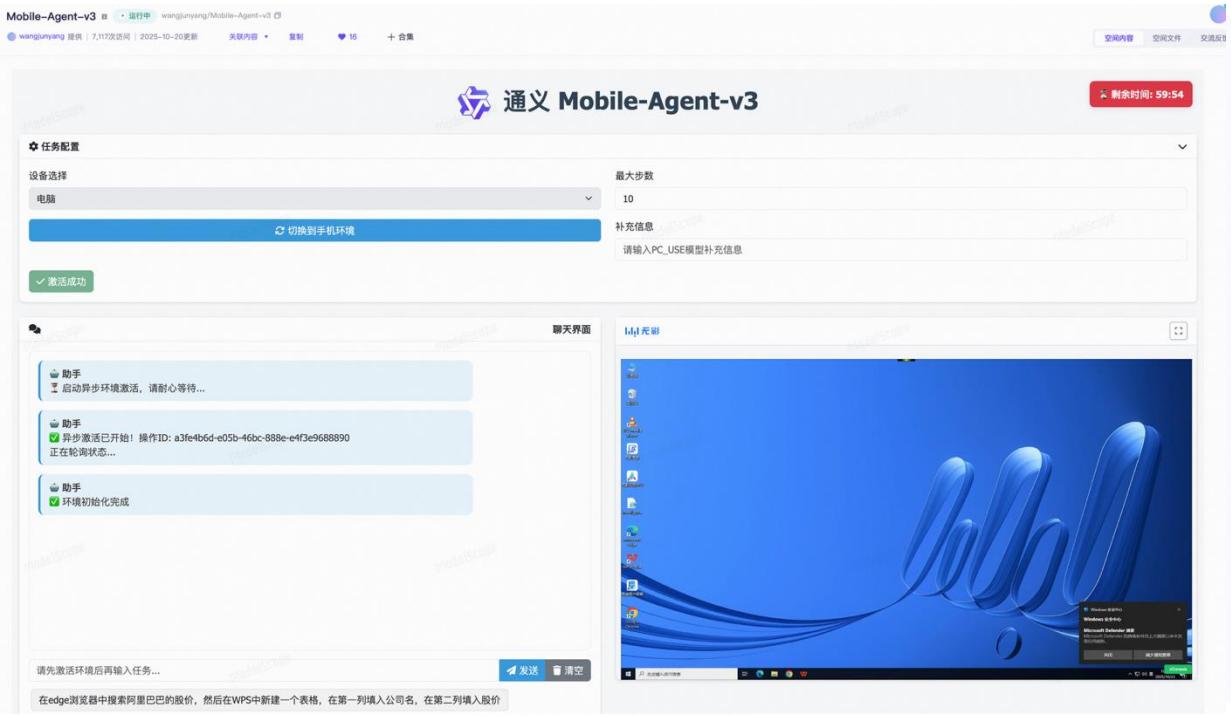
Mobile-Agent V3实验

Model	Success Rate (%)	
	Mobile-Agent-E (Wang et al., 2025b) *on AndroidWorld*	Agent-S2 (Agashe et al., 2025) *on a subset of OSWorld-Verified*
Baseline Models		
UI-TARS-1.5 (Qin et al., 2025)	14.1	14.7
UI-TARS-72B (Qin et al., 2025)	14.8	19.0
Qwen2.5-VL-72B (Bai et al., 2025)	52.6	38.6
Seed-1.5-VL (Team, 2025)	56.0	39.7
Our Models		
GUI-Owl-7B	59.5	40.8
GUI-Owl-32B	**62.1**	**48.4**

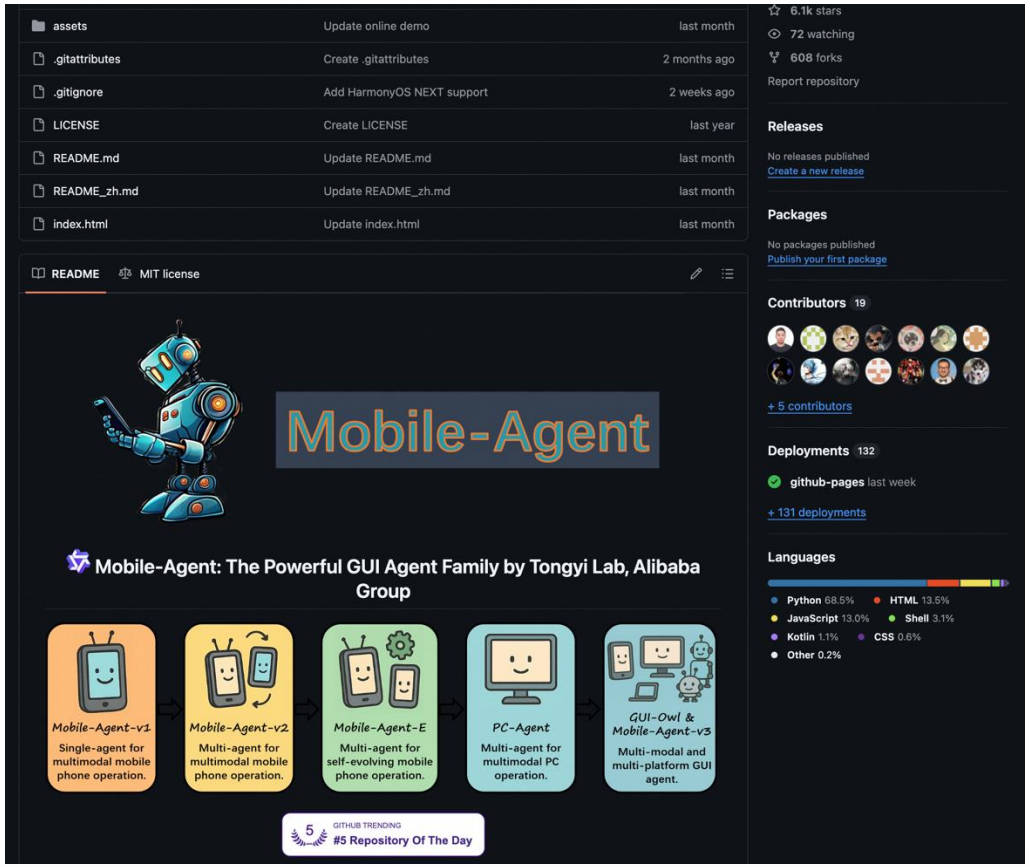
Performance comparison on agentic frameworks

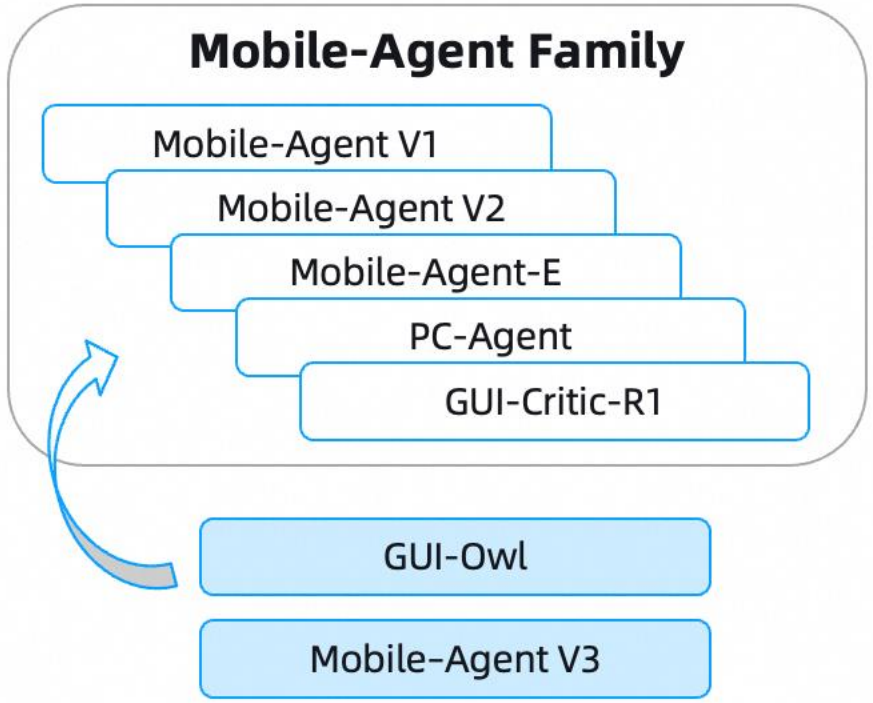
Mobile-Agent V3云沙箱Demo

Mobile-Agent V3云沙箱Demo



Mobile-Agent开源应用





体验Demo在魔搭、百炼上线，免费API并项目仓库开源

https://github.com/X-PLUG /MobileAgent

Qwen3-VL：明察、深思、广行

		Qwen3-VL 235B-A22B Instruct	Gemini2.5-Pro Thinkingbudget 128	GPT5 Minimal or Without thinking	Claude-Opus-4.1 Without thinking	Other Best Without thinking
STEM&Puzzle	$MMMU_{VAL}$	78.7	80.9	74.4*	77.2	73.6* (Seed1.5VL)
	MMMU_Pro	68.1	71.2	62.7*	60.7	59.9* (Seed1.5VL)
	$MathVista_{mini}$	84.9	77.7	50.9	74.5	83.0* (Seed1.5VL)
	MathVision	66.5	66.0	45.8	57.7	65.5* (Seed1.5VL)
	$MathVerse_{mini}$	72.5	65.9	43.0	68.1	65.4* (GLM4.5V)
	VisuLogic	29.9	26.9	27.2	27.2	33.0* (Seed1.5VL)
General VQA	$MMBench\_EN\_V1.1_{dev}$	89.9	86.6	81.4	84.1	89.0* (InternVL3)
	RealWorldQA	79.3	76.0	77.3	68.5	78.0* (InternVL3)
	MMStar	78.4	78.5	65.2	71.0	76.2* (Seed1.5VL)
	SimpleVQA	63.0	66.9	56.7	55.7	63.1* (Seed1.5VL)
Subjective Experience and Instruction Following	HallusionBench	63.2	60.9	53.7	55.1	60.0* (Seed1.5VL)
	MM_MT_Bench	8.5	7.6	7.5	7.9	7.6* (Qwen2.5VL)
	MIABench	91.3	91.3	92.6	90.0	90.8 (Seed1.5VL)
Text Recognition and Chart/Document Understanding	MMLongBench-Doc	57.0	51.2	42.4	48.1	50.1 (Seed1.5VL)
	$DocVQA_{TEST}$	97.1	94.0	89.6	89.2	96.7* (Seed1.5VL)
	$InfoVQA_{TEST}$	89.2	82.9	69.9	60.9	87.3* (Qwen2.5VL)
	$AI2D_{TEST}$	89.7	90.0	84.1	84.4	89.7* (InternVL3)
	OCRBench	920.0	872.0	787.0	750.0	906* (InternVL3)
	$OCRBenchV2_{(en/ch)}$	67.1 / 61.8	55.2 / 53.1	48.2 / 37.7	47.2 / 38.0	61.5 / 63.7* (Qwen2.5VL)
	CC_OCR	82.2	76.8	66.1	66.0	79.8* (Qwen2.5VL)
	CharXiv(RQ)	62.1	62.9	57.8*	60.2	59.8* (Seed1.5VL)
2D/3D Grounding	$RefCOCO_{avg}$	91.9	–	–	–	91.6* (Seed1.5VL)
	CountBench	93.0	91.0	87.8	91.9	93.6* (Qwen2.5VL)
	ODinW13	48.6	34.5	–	–	40.6 (Seed1.5VL)
	ARKitScenes	56.9	–	–	–	27.5 (Seed1.5VL)
	Hypersim	13.0	–	–	–	9.6 (Seed1.5VL)
	SUNRGBD	39.4	–	–	–	33.5* (Seed1.5VL)
Multi-Image	BLINK	70.7	70.0	62.8	62.9	70.2* (Seed1.5VL)
	MUIRBENCH	72.8	74.0	66.5	–	71.1* (GLM4.5V)
Embodied and Spatial Understanding	ERQA	51.3	50.3	42.0*	26.0	46.5* (GLM4.5V)
	VSI-Bench	62.6	31.4	–	–	48.4* (InternVL3)
	EmbSpatialBench	83.1	73.3	75.1	66.0	78.6* (RoboBrain 2.0)
	RefSpatialBench	65.5	35.4	23.1	–	54.0* (RoboBrain 2.0)
	RoboSpatialHome	69.5	49.2	43.6	–	72.4* (RoboBrain 2.0)
Video	VideoMME(w/o sub)	79.2	80.6	77.3	73.3	77.6* (Seed1.5VL)
	MLVU	84.3	81.2	78.3	71.2	81.8* (Seed1.5VL)
	LVBench	67.7	69.0	–	–	64.0* (Seed1.5VL)
	CharadesSTA	64.8	–	–	–	64.7* (Seed1.5VL)
	VideoMMMU	74.7	79.4	61.6*	70.1	72.1* (Seed1.5VL)
Agent	ScreenSpot	95.4	–	–	–	88.7* (InternVL3)
	ScreenSpot Pro	62.0	–	–	–	60.9* (Seed1.5VL)
	OSWorldG	66.7	–	–	–	53.2* (InternVL3)
	AndroidWorld	63.7	–	–	–	62.1* (Seed1.5VL)
Coding	Design2Code	92.0	90.3	88.9	85.3	84.5* (GLM4.5V)
	ChartMimic_v2_Direct	80.5	79.9*	–	–	–
	UniSvg	69.3	67.9	–	–	–

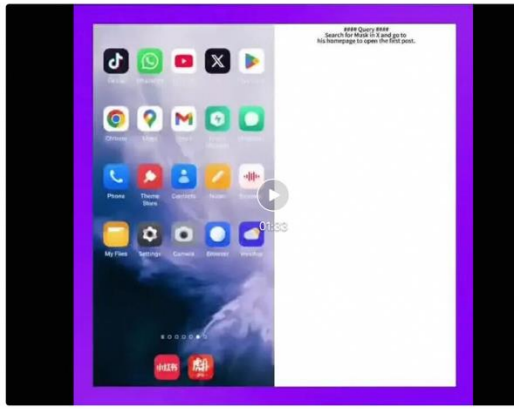
Note: The default evaluation is conducted through API calls, and * indicates scores from the report. Results on video understanding are measured using a 256k-token context, handling up to 2048 frames.

		Qwen3-VL 235B-A22B Thinking	Gemini2.5-Pro	GPT5 high	Claude-Opus-4.1 With thinking	Other Best With thinking
STEM&Puzzle	$MMMU_{VAL}$	80.6	81.7*	84.2*	78.4	80.1* (dots.vlm1)
	MMMU_Pro	69.3	68.8*	78.4*	64.8	70.1* (dots.vlm1)
	$MathVista_{mini}$	85.8	82.7*	81.3	75.5	85.6* (Seed1.5VL)
	MathVision	74.6	73.3*	70.9	64.3	69.9* (dots.vlm1)
	$MathVerse_{mini}$	85.0	82.9	84.1	70.6	72.1* (GLM4.5V)
	ZEROBench	4.0	3.0	2.0	1.0	2.0* (Seed1.5VL)
	ZEROBench_Sub	27.7	33.8	24.6	26.3	30.8* (Seed1.5VL)
	VisuLogic	34.4	31.6	28.5	27.9	35.0* (Seed1.5VL)
General VQA	$MMBench\_EN\_V1.1_{dev}$	90.6	90.1*	85.9	84.9	89.9* (Seed1.5VL)
	RealWorldQA	81.3	78.0*	82.8	69.9	79.1* (dots.vlm1)
	MMStar	78.7	77.5*	76.4	72.1	77.9* (InternVL3.5)
	SimpleVQA	61.3	65.4	61.8	56.7	63.4* (Seed1.5VL)
Subjective Experience and Instruction Following	HallusionBench	66.7	63.7*	65.7	60.4	65.4* (GLM4.5V)
	MM_MT_Bench	8.5	8.4	7.6	7.8	–
	MIABench	92.7	92.3	92.4	91.2	92.3 (Seed1.5VL)
Text Recognition and Chart/Document Understanding	MMLongBench-Doc	56.2	55.6	51.5	54.5	52.7 (Seed1.5VL)
	$DocVQA_{TEST}$	96.5	92.6	91.5	92.5	96.9* (Seed1.5VL)
	$InfoVQA_{TEST}$	89.5	84.2	79.0	69.4	91.2* (Seed1.5VL)
	$A12D_{TEST}$	89.2	90.9	89.7	86.4	88.4* (dots.vlm1)
	OCRBench	875.0	866.0	810.0	764.0	907.0* (InternVL3.5)
	$OCRBenchV2_{(en/ch)}$	66.8 / 63.5	54.3 / 48.5	53.0 / 43.2	48.4 / 43.7	–
	CC_OCR	81.5	77.2	68.3	69.1	–
	CharXiv(RQ)	66.1	67.9	81.1*	63.6	64.4* (dots.vlm1)
2D/3D Grounding	RefCOCO-avg	92.4	74.6*	66.8	–	92.4* (InternVL3.5)
	CountBench	93.7	91.0*	91.7	93.1	93.7* (Seed1.5VL)
	ODinW13	43.2	33.7*	–	–	41.3 (Seed1.5VL)
	ARKitScenes	53.7	7.4	–	–	30.3 (Seed1.5VL)
	Hypersim	11.0	2.2	–	–	9.4 (Seed1.5VL)
	SUNRGBD	34.9	–	–	–	32.5 (Seed1.5VL)
	Objectron	71.2	5.5	–	–	8.1 (Seed1.5VL)
Multi-Image	BLINK	67.1	70.6*	71.0	64.1	72.1* (Seed1.5VL)
	MUIRBENCH	80.1	77.2	77.5	–	78.6* (dots.vlm1)
Embodied and Spatial Understanding	ERQA	52.5	55.3	65.7*	36.3	50.0* (GLM4.5V)
	EmbSpatialBench	84.3	79.1	82.9	69.2	78.6* (RoboBrain 2.0)
	RefSpatialBench	69.9	36.5	23.8	–	54.0 (RoboBrain 2.0)
	RoboSpatialHome	73.9	47.5	53.5	–	72.4 (RoboBrain 2.0)
Video	VideoMME(w/o sub)	79.0	85.1	84.7	75.6	77.9* (Seed1.5VL)
	MLVU	83.8	85.6	86.2	73.5	82.1* (Seed1.5VL)
	LVBench	63.6	73.0	-	-	64.6* (Seed1.5VL)
	CharadesSTA	63.5	–	–	–	64.0* (Seed1.5VL)
	VideoMMMU	80.0	83.6*	84.6*	76.2	81.4* (Seed1.5VL)
Agent	ScreenSpot	95.4	–	–	–	95.2* (Seed1.5VL)
	ScreenSpot Pro	61.8	–	–	-	60.9* (Seed1.5VL)
	OSWorldG	68.3	45.2	–	–	62.9* (Seed1.5VL)
	OSWorld	38.1	–	–	–	36.7* (Seed1.5VL)
Coding	Design2Code	93.4	89.2	92.5	88.5	82.2* (GLM4.5V)

Note: The default evaluation is conducted through API calls, and * indicates scores from the report. Results on video understanding are measured using a 256k-token context, handling up to 2048 frames.

01 视觉Agent

Qwen3-VL 不仅能看懂图片，还能像人一样操作手机和电脑，自动完成许多日常任务。例如打开应用、点击按钮、填写信息等，实现智能化的交互与自动化操作。



Example: Android Use

02 带图推理

Qwen3-VL 可以像人类一样仔细观察图像的局部细节，并结合工具进行复杂推理。比路边的路牌判断具体位置，或根据人物照片搜索相关信息，完成细粒度识别和逻辑务。

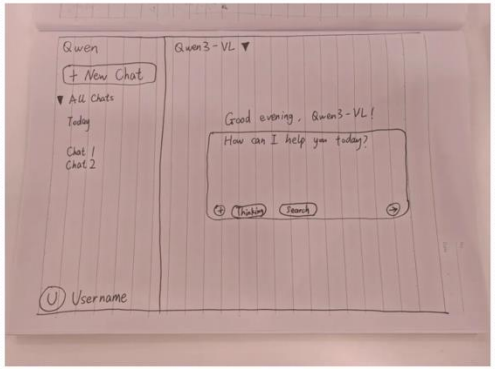
Example: Think with Image

03 代码编程

结合视觉理解和代码生成能力，Qwen3-VL 在前端开发方面展现出强大潜力。例如，能把手绘草图转成网页代码，或帮助调试界面问题，提升开发效率。

Example: Efficiency Tool

prompt: Create the webpage using HTML and CSS based on my sketch design. Color it in dark mode.

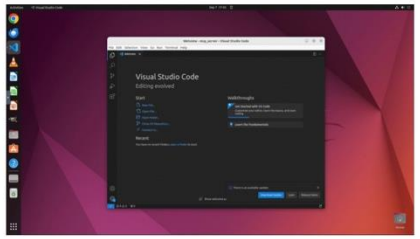


Qwen3-VL:

```
            font-weight: 600;
        }
        .new-chat-btn {
            background-color: var(--dark-tertiary);
            border: 1px solid var(--border);
            border-radius: 8px;
            padding: 12px 20px;
            font-size: 18px;
            font-weight: 500;
```

OSWorld-MCP：多模态工具调用

Step 1: Click the **Extensions** icon.



Step 2: Type autoDocstring into the search bar.

Step 3: Click **Install** on the autoDocstring extension entry.

Step 4: The autoDocstring extension is now installed.

(a) a GUI-based approach

```python
@mcp.tool
def install_extension(cls, extension_id,
    pre_release=False):
    """
    Installs an extension or updates it in
        VSCode.

    Args:
        extension_id (str)
        pre_release (bool)
    """
    try:
        command = ["code", "--install-
            extension", extension_id]
        if pre_release:
            command.append("--pre-release")
        subprocess.run(command, check=True)
        ret = "Successfully installed
            extension"
    except subprocess.CalledProcessError as
        e:
        ret = f"Error installing extension:
            {e}"
    except Exception as e:
        ret = f"Unexpected error: {e}"

    return ret
```

(b) MCP tool

https://osworld-mcp.github.io/

Leaderboard (OSWorld-MCP)

We adopt state-of-the-art **LLM** and **VLM** from open-source representatives such as **Agent-S**, **Qwen** and closed-source ones from **GPT**, **Gemini**, and **Claude** families on OSWorld-MCP, as LLM and VLM agent baselines.
Acc, **TIR**, and **ACS** denote the three evaluation metrics: *Task Accuracy*, *Tool Invocation Rate*, and *Average Completion Steps*.

Unified Prompt: To facilitate comparison of performance differences across models, we standardized our evaluation using the **GUI-Owl** agent configuration. This may lead to some performance fluctuations for certain models under their original OSWorld configuration.

Specific Prompt: Each model adopts its own unique configuration for OSWorld-MCP evaluation.

We are actively updating the benchmark with new LLMs, VLMs and methods. Please submit the invocation method and evaluation scripts. **Pull requests welcomed!**
For more information, contact: jiahongrui@stu.pku.edu.cn, shuofeng.xhy@alibaba-inc.com, zx443053@alibaba-inc.com.

Filter by Max Steps: All

All Prompts | Unified Prompt | Specific Prompt

Rank	Model	Details	Acc	TIR	ACS
	Agent-S2.5 OpenAI o3 + UI-TARS-1.5-72B Simular Research Simular Research, '25	Type: Agentic framework Max Steps: 50 Runs: 3 Prompt: unified prompt	49.5	35.3	17.0
	Claude 4 Sonnet claude-sonnet-4-20250514-thinking Anthropic Anthropic, '25	Type: General model Max Steps: 50 Runs: 3 Prompt: unified prompt	43.3	36.6	20.1
	Agent-S2.5 OpenAI o3 + UI-TARS-1.5-72B Simular Research Simular Research, '25	Type: Agentic framework Max Steps: 15 Runs: 3 Prompt: unified prompt	42.1	30.0	10.0
4	**Qwen3-VL** qwen3-vl-235b-a22b-thinking Alibaba Cloud, Qwen Team Qwen Team, '25	Type: General model Max Steps: 50 Runs: 3 Prompt: unified prompt	39.1	29.5	21.1
5	**Seed1.5-VL** doubao-1-5-thinking-vision-pro-250428 ByteDance Seed ByteDance Seed, '25	Type: General model Max Steps: 50 Runs: 3 Prompt: unified prompt	38.4	29.0	23.0

大模型通用型智能体系统-未来方向

技术角度：

- Agentic RL Scaling，提升自主推理和知识进化；
- MCP、Code、GUI相结合；
- DeepResearch + Operator -> ChatGPT agent；
- 个性化交互与记忆；

多模态GUI-o3推理

个性化交互

API工具/代码调用

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

AiDD 峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2026.05.22-23

AiDD峰会 北京站
AI+研发数字峰会
时间：2026.08.21-22

AiDD峰会 深圳站
AI+研发数字峰会
时间：2026.11.20-21

AiDD峰会详情

AiDD × AI+产品峰会

AiDD × 上海站
AI+产品创新峰会
时间：2026.01.16-17

AiDD × 北京站
AI+产品创新峰会
时间：2026.08.23-24

产品峰会详情

AiDD × K

2026 AI+ PRODUCT INNOVATION SUMMIT

01.16-17 · ShangHai

AI+产品创新峰会

Track 1: AI 产品战略与创新设计

从0到1的AI原生产品构建

论坛1: AI时代的用户洞察与需求发现

论坛2: AI原生产品战路与商业模式重构

论坛3: AgenticAI产品创新与交互设计

Track 2: AI 产品开发与工程实践

从1到10的工程化落地实践

论坛1: 面向Agent智能体的产品开发

论坛2: 具身智能与AI硬件产品

论坛3: AI产品出海与本地化开发

Track 3: AI 产品运营与智能演化

从10到100的AI产品运营

论坛1: AI赋能产品运营与增长黑客

论坛2: AI产品的数据飞轮与智能演化

论坛3: 行业爆款AI产品案例拆解

2-hour Speech: 回归本质

用户洞察的第一性

——2小时思维与方法论工作坊

在数字爆炸、AI迅速发展的时代，仍然考验“看见”的“同理心”

Panel 1: 出海前瞻

“出海避坑地图”圆桌对话

——不止于翻译:

AI时代的出海新范式

Panel 2: 失败复盘

为什么很多AI产品“叫好不叫座”?

——从伪需求到真价值:

AI产品商业化落地的关键挑战

智能重构产品 数据驱动增长

Reinventing Products with Intelligence, Driven by Data

80+	40+	10+	800+
业界大咖	创新案例	分论坛	听众规模
汇聚产品大咖的真知灼见	开拓全新AI+产品视角	涵盖产品到商业增长的全链路	共同探索产品的智能重构

扫码查看会议详情

第8届AI+研发数字峰会
AI+ Development Digital Summit

感谢聆听！

扫码领取会议PPT资料