

多模态、多端GUI智能体 Mobile-Agent

徐海洋

通义实验室算法科学家

CONTENT 目录

01 大模型智能体背景

02 多模态、多端智能体Mobile-Agent

03 Foundation Agent for GUI

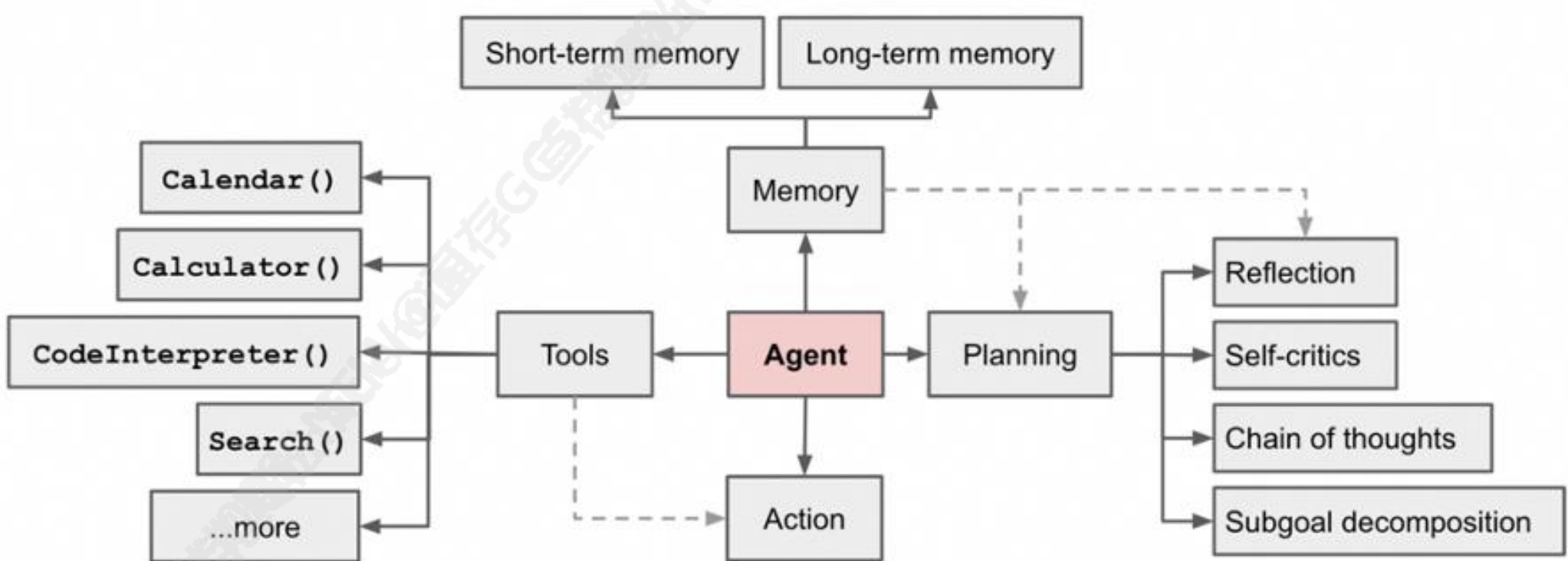
04 Mobile-Agent开源应用

01 大模型智能体背景

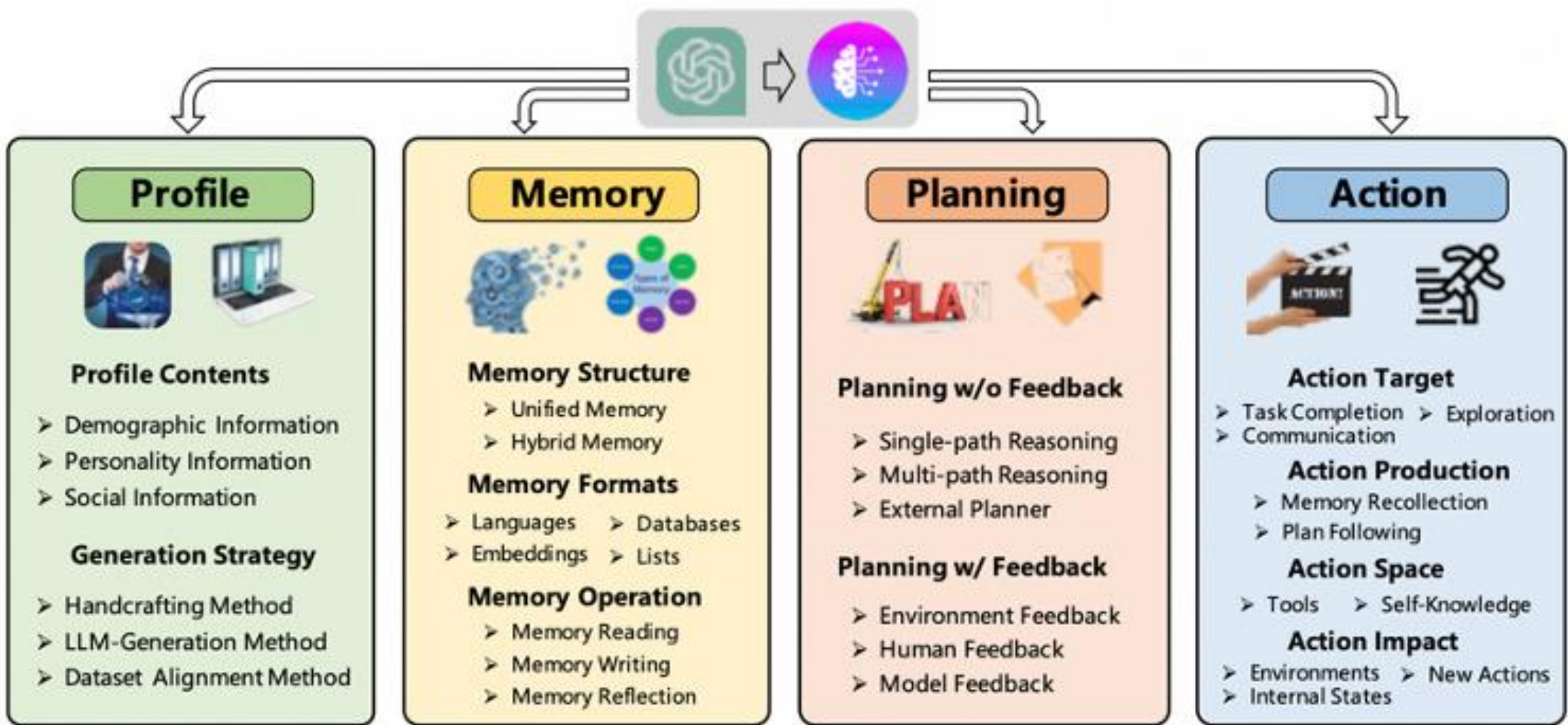
大模型智能体系统

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



大模型智能体优势



OpenAI Five



DeepMind AlphaStar



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

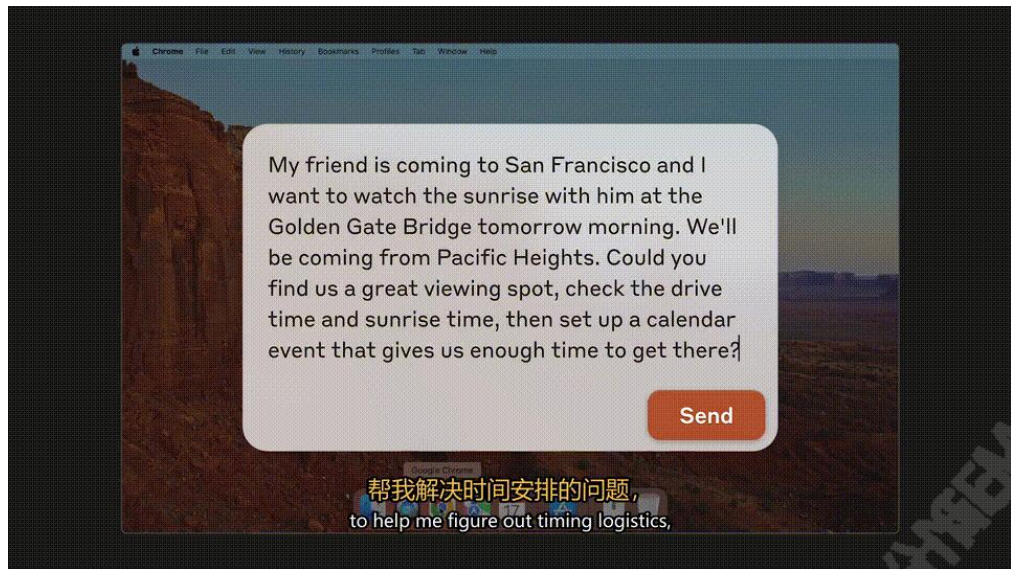
近期两类AI Agent应用

	GUI Agent	Code/CLI Agent
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
试用场景	[自动化] 操作主导型 简单、重复性任务 生活娱乐、自动化办公	[智能化]知识密集型 复杂创作任务、 办公生产力场景
示例	Operator、Apple Intelligence、Mobile-Agent、豆包手机	Code (Claude、OpenAI、Gemini、Qwen)
	GUI+CLI桌面端Agent (CodeX、Claude Code、Qwen Code)	

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。

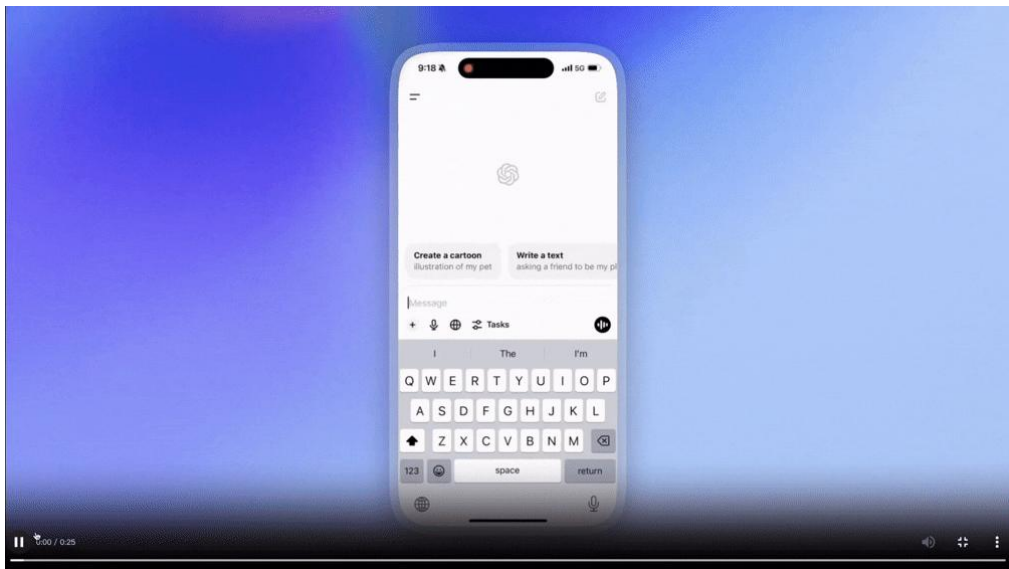
Claude (Computer)



Apple Intelligence



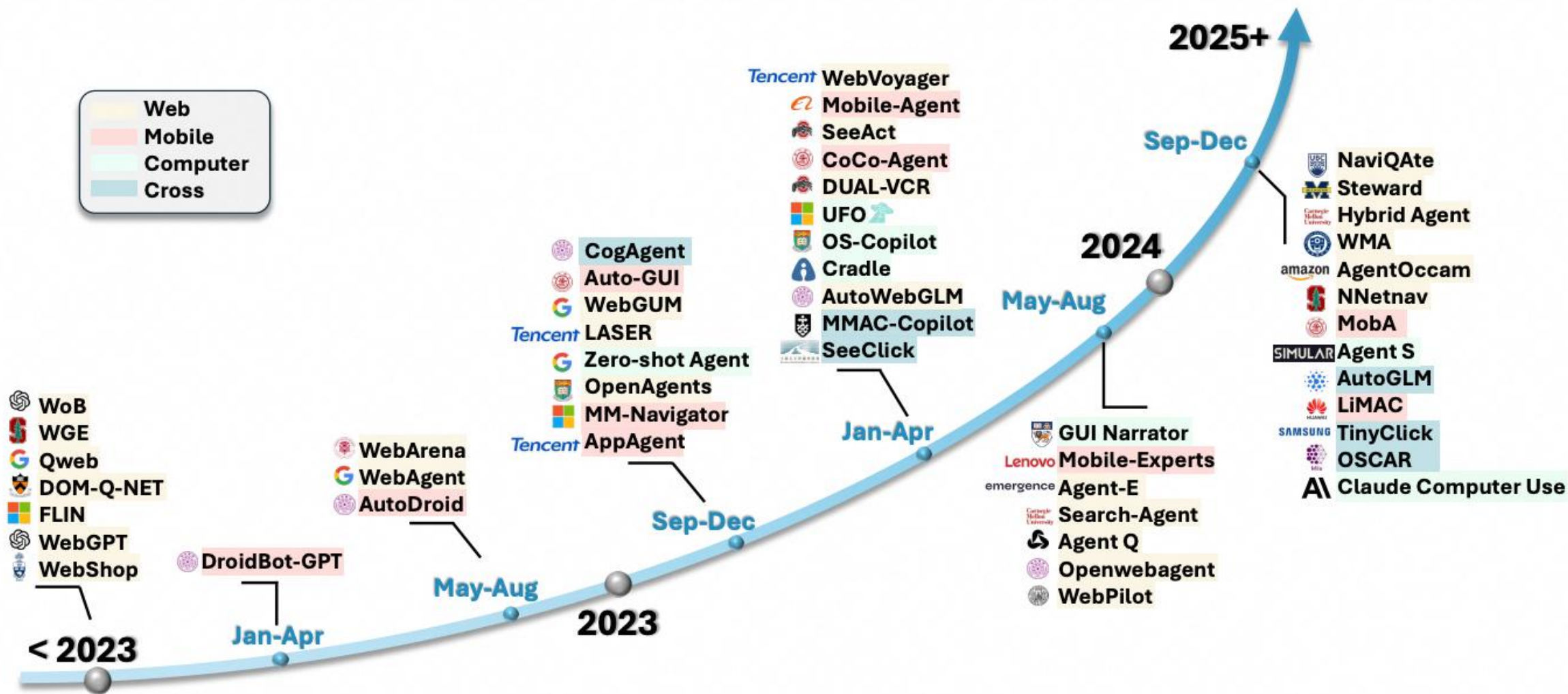
OpenAI Operator



豆包手机

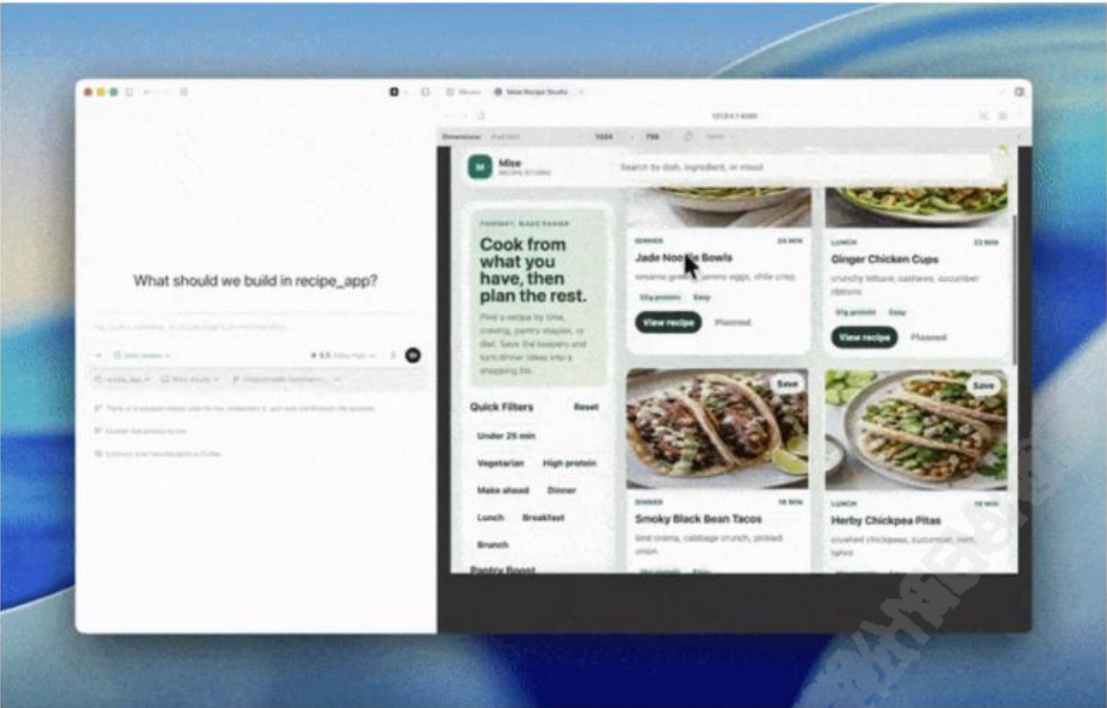


AutoGLM



GUI+CLI桌面端Agent

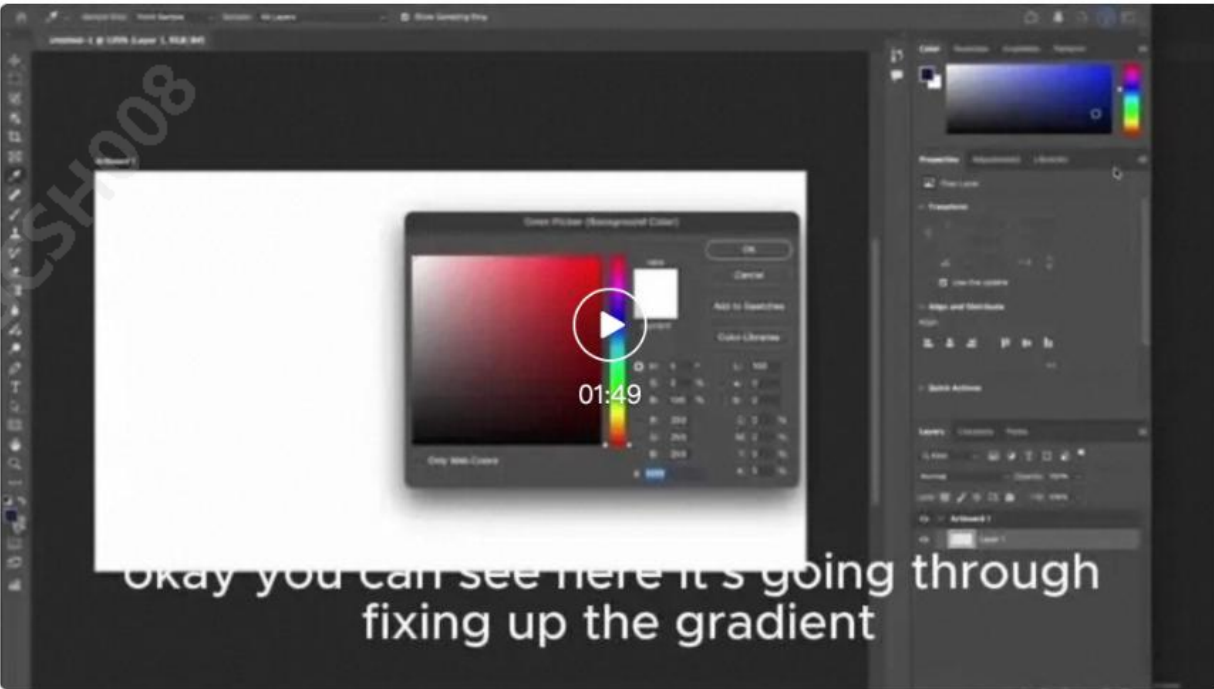
在Codex应用内浏览器中还添加了设备工具栏，让构建和测试响应式应用变得更加便捷——



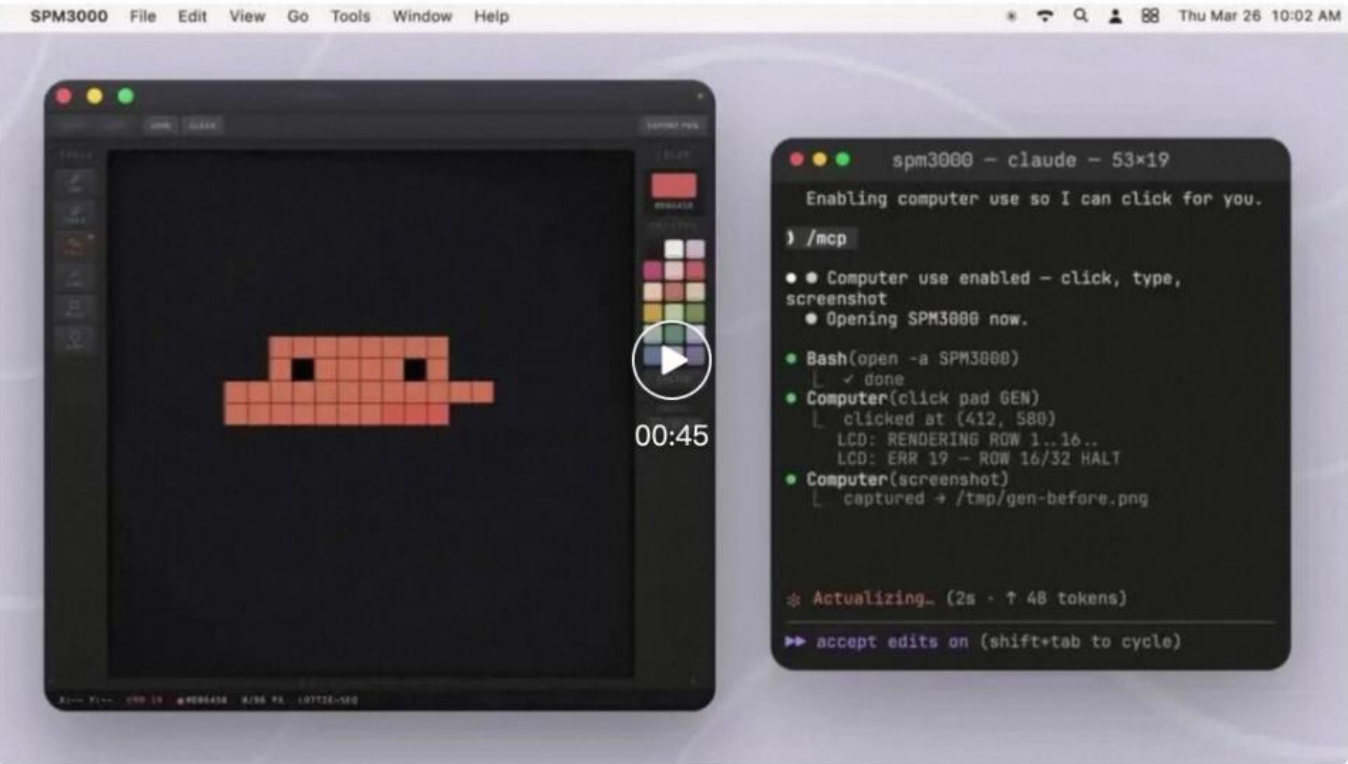
任务三：AI视频生成。Codex调用Adobe Firefly，根据文字描述生成视频素材片段，自动拼接、加转场。



任务二：播客封面设计。Codex打开Photoshop，根据播客主题自动选择配色方案、排版标题文字、调整图层混合模式，输出一张可以直接上传的封面图。

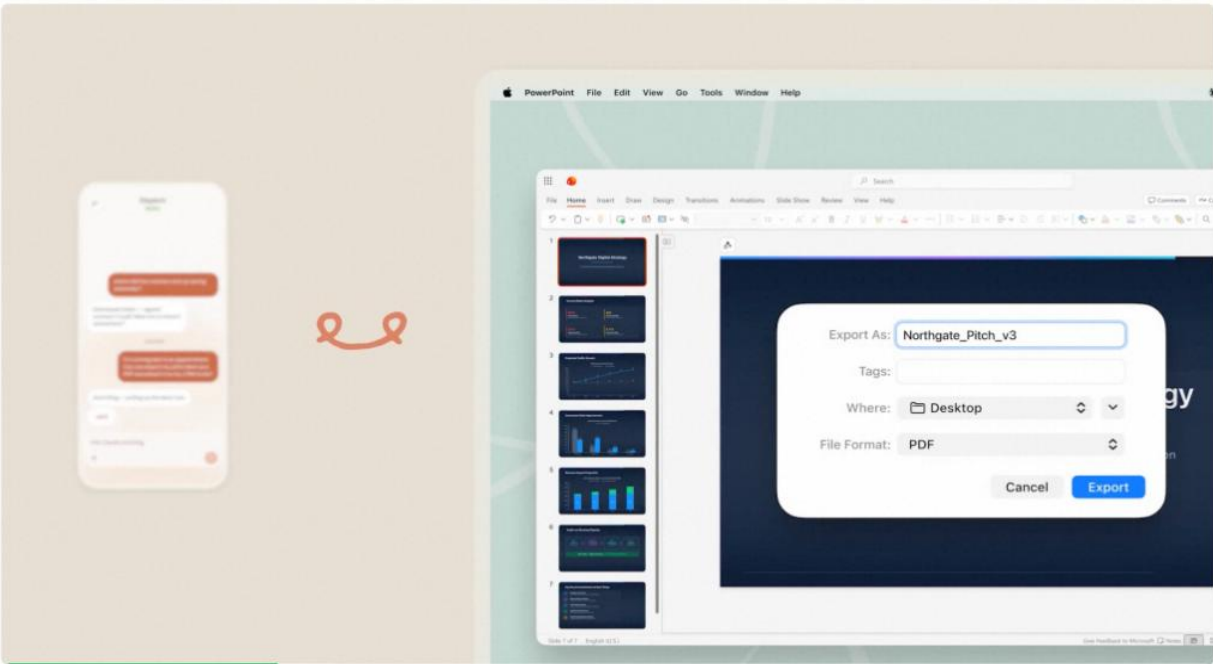


在官方演示中，只甩一个指令过去，AI就自己启动正在开发的应用，自己复现bug，自己修复，自己测试修复效果。



基于GUI的操作，像真人一样通过实时截图“看到”屏幕上的所有内容，然后模拟鼠标键盘操作。

和人类用电脑没有任何区别了。



基于 Qwen3.7-Plus 构建的 Hybrid-Agent 智能体系统，将大模型的代码生成能力与 GUI 自动化执行深度融合，实现了从需求分析到版本迭代的 APP 全链路开发。Agent持续稳定运行 11+ 小时，全程自动完成了一款英语单词学习 APP 的完整研发闭环。累计生成代码超过 10,000+ 行，触发 Agent 调用超过 1,000+ 次，覆盖软件开发全生命周期的核心环节：需求文档生成、代码自动编写、自动化安装部署、测试用例创建、GUI 自动化测试、多场景并行化测试、产品说明自动更新、自动版本迭代演进。

Hybrid-Agent

APP代码开发 & 图形化界面自动化测试
App Development & GUI Automation Testing

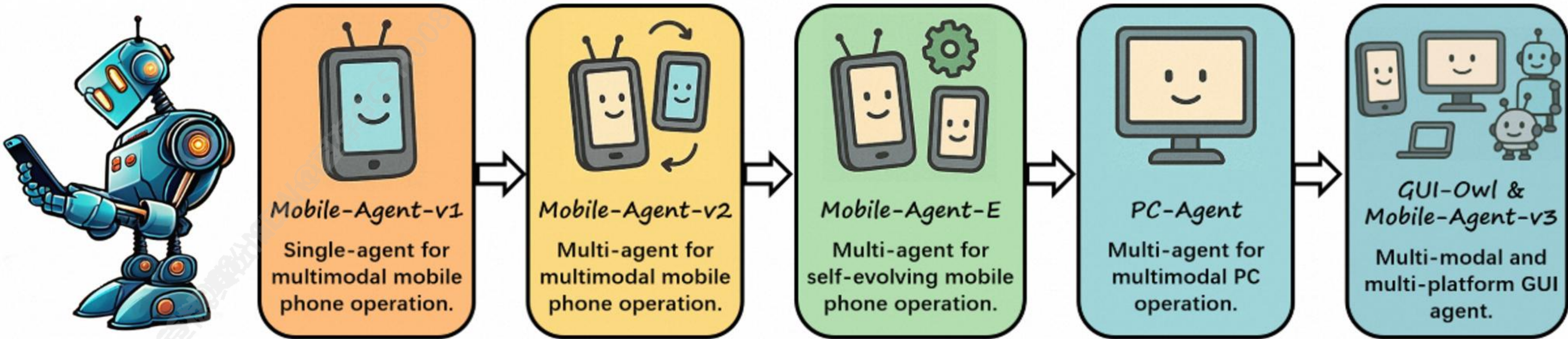
演进迭代
Continuous Iteration

02 多模态、多端智能体Mobile-Agent

多模态、多端Mobile-Agent

通义多模态、多端GUI智能体Mobile-Agent

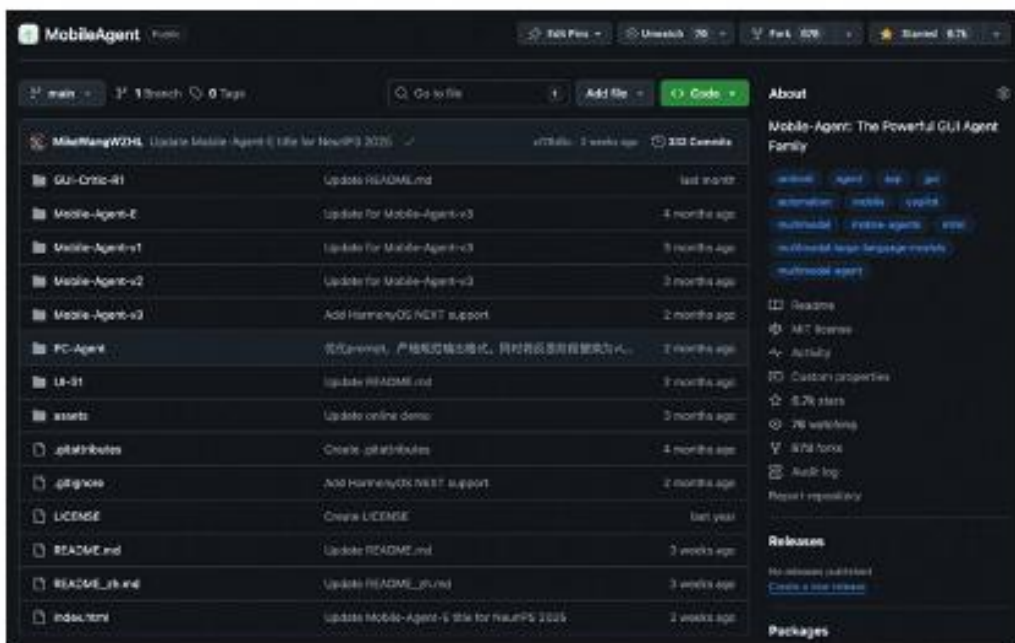
Mobile-Agent: 强大的GUI智能体家族



CCL2024, CCL 2025 Highlight System



Github 8.9k stars



多模态、多端Mobile-Agent



手机



平板



车机



PC & Web

核心挑战

UI界面感知

UI元素多样，不同系统
机型、APP，差异大；需要精
确定位精确操作

长程任务
规划反思

需要具有复杂任务规划、最优
路径选择、
多图理解推理能力

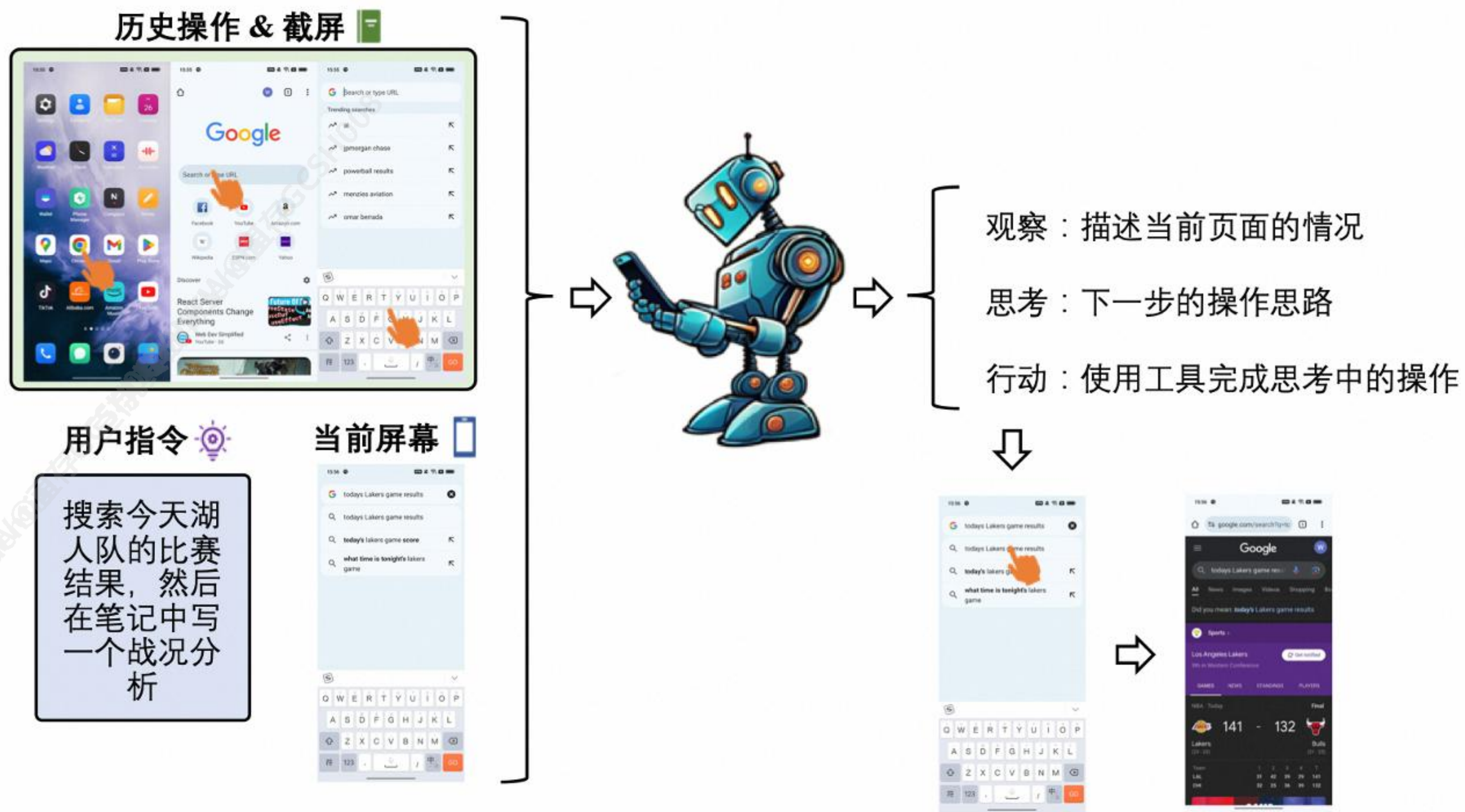
Foundation
Agent能力

工具/Memory/Agent框架
/Critic/CLI等能力

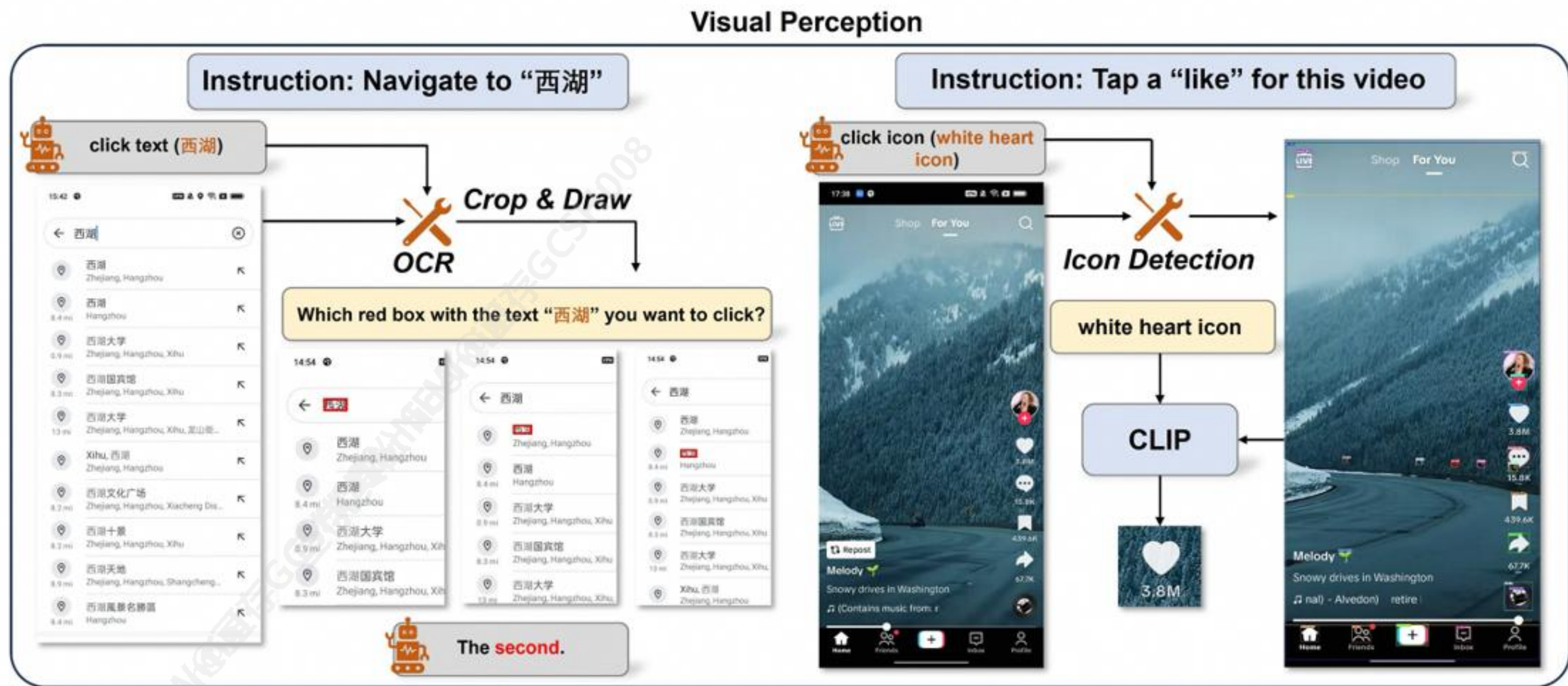
Long
Horizon路径
选择能力

Long Horizon最优路径选择路
由能力

多模态移动端智能体Mobile-Agent V1



多模态移动端智能体Mobile-Agent V1



行为空间

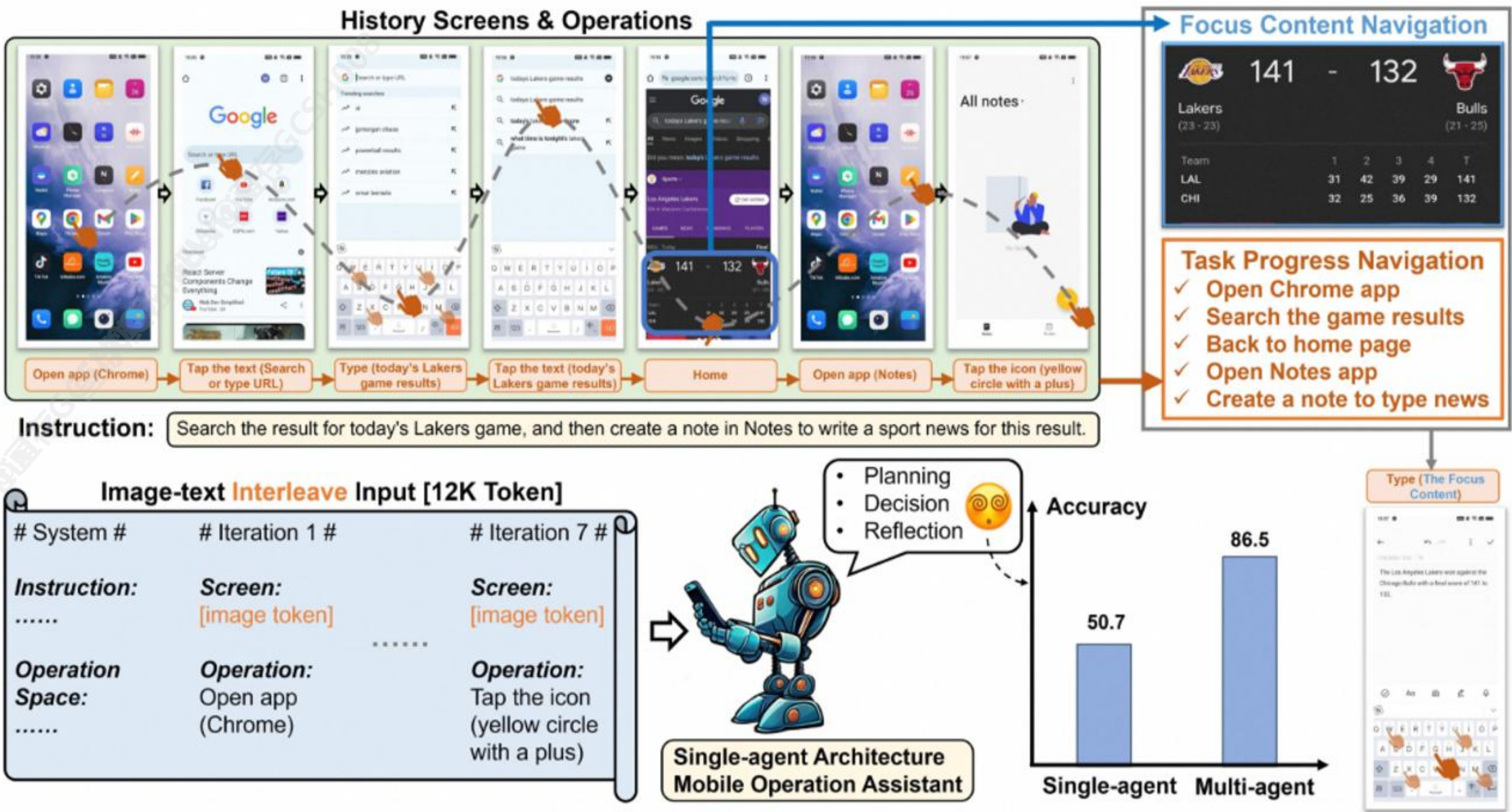
1. 点击文本
2. 点击图标
3. 打字
4. 上划 & 下划
5. 返回上一页面
6. 返回桌面
7. 结束

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

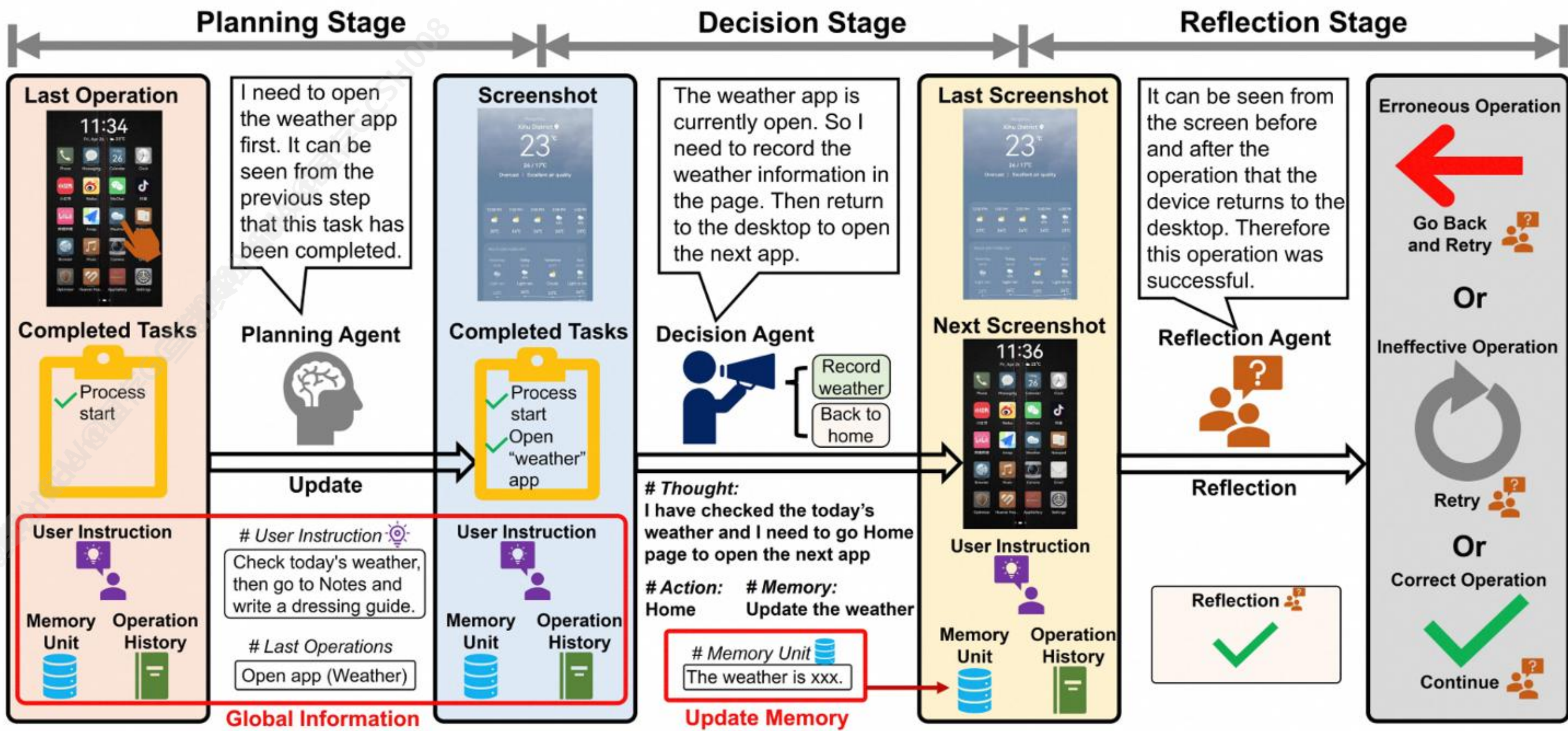
多模态多智能体Mobile-Agent V2

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

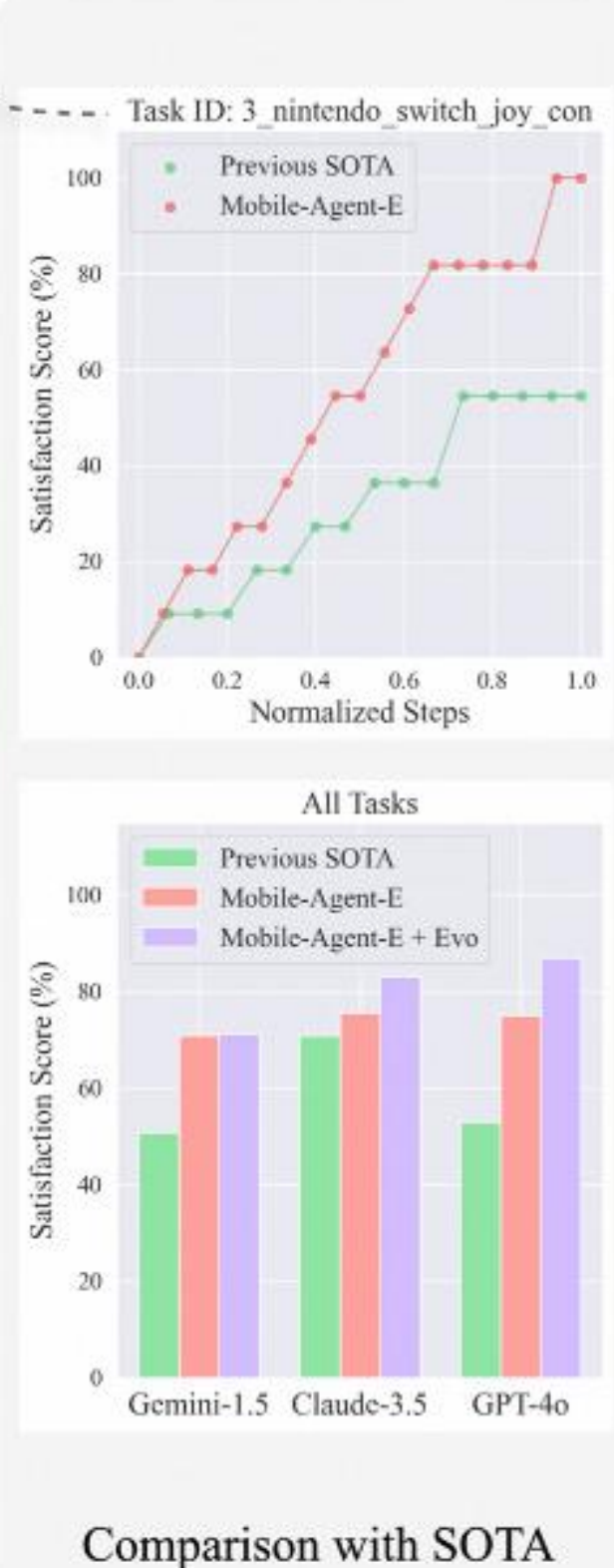
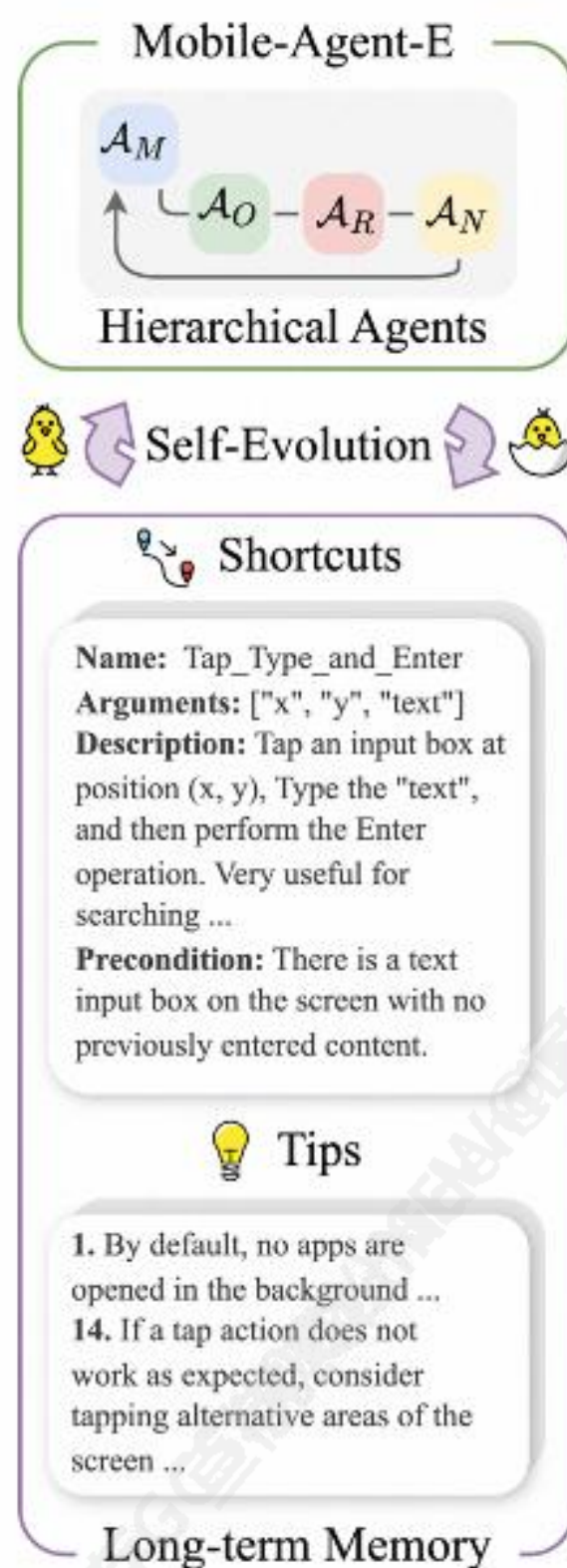


多模态多智能体Mobile-Agent V2

采用多智能体框架，包括Planning Agent、Decision Agent、Reflection Agent



Mobile-Agent-E: 解决复杂任务、自主进化

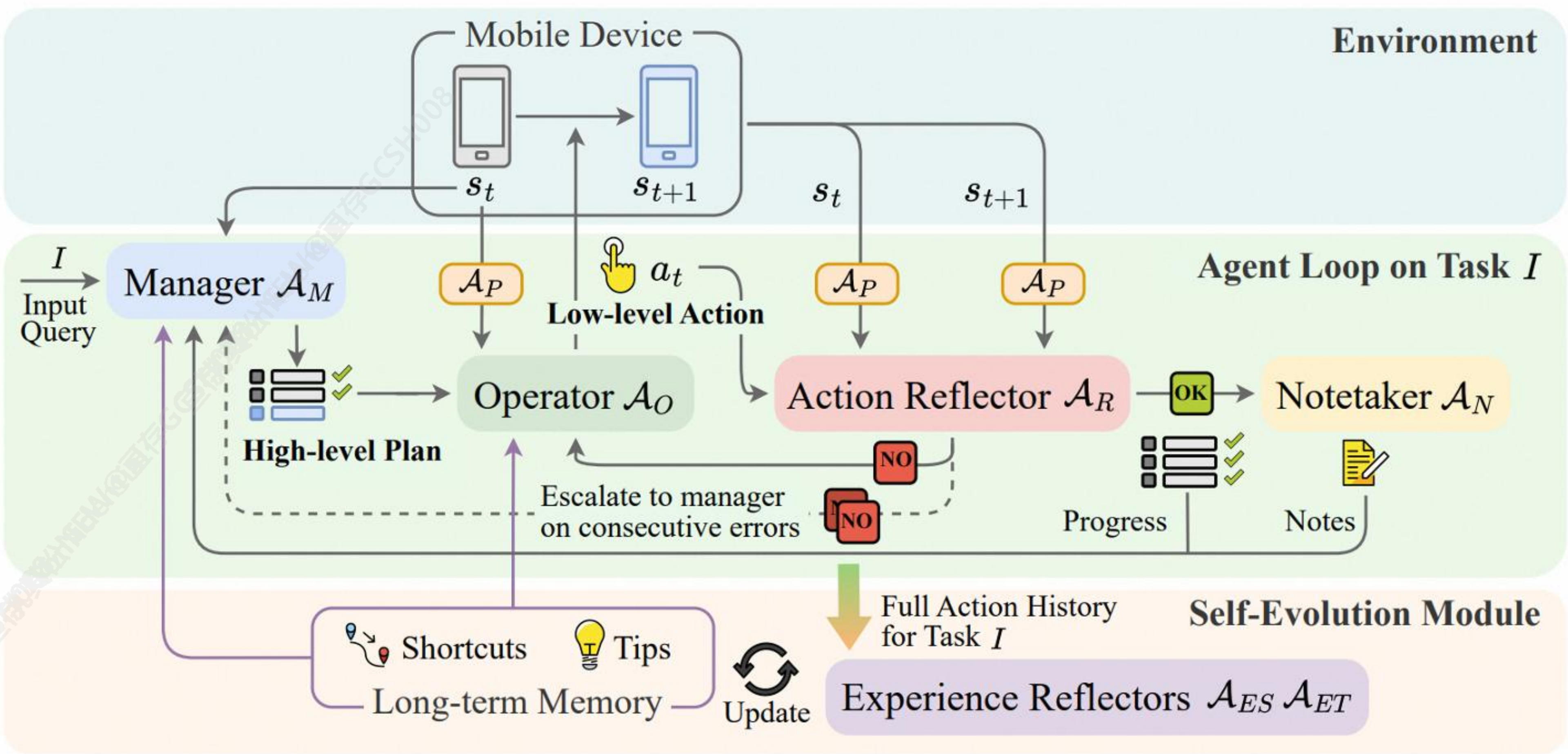


复杂指令:
执行复杂推理、多步规划以及跨App操作

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

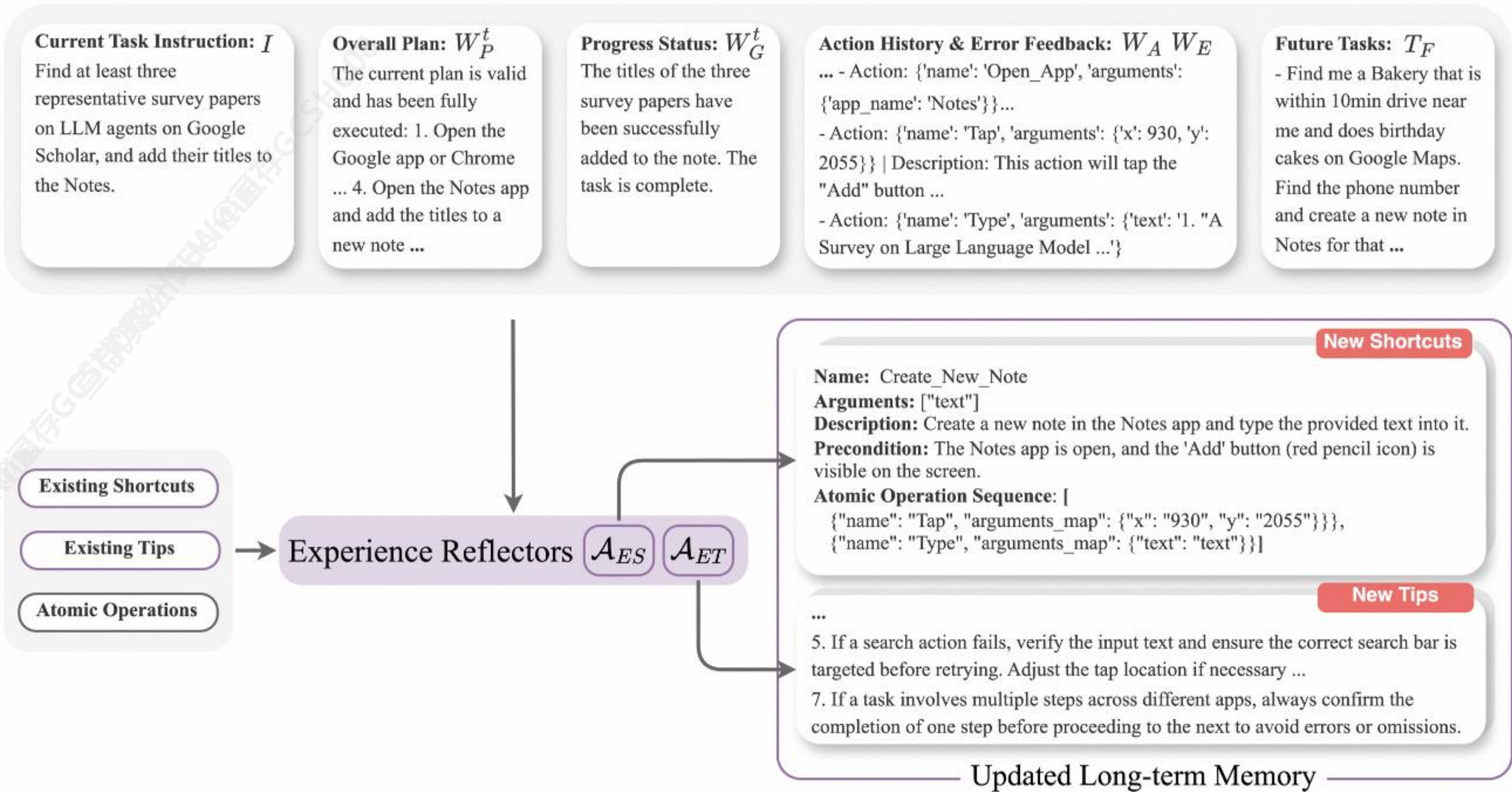
多平台比价任务示例

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

两个经验Experience Reflectors会根据当前任务的操作记录和错误日志等信息，对Tips和Shortcuts 进行可能的优化和更新

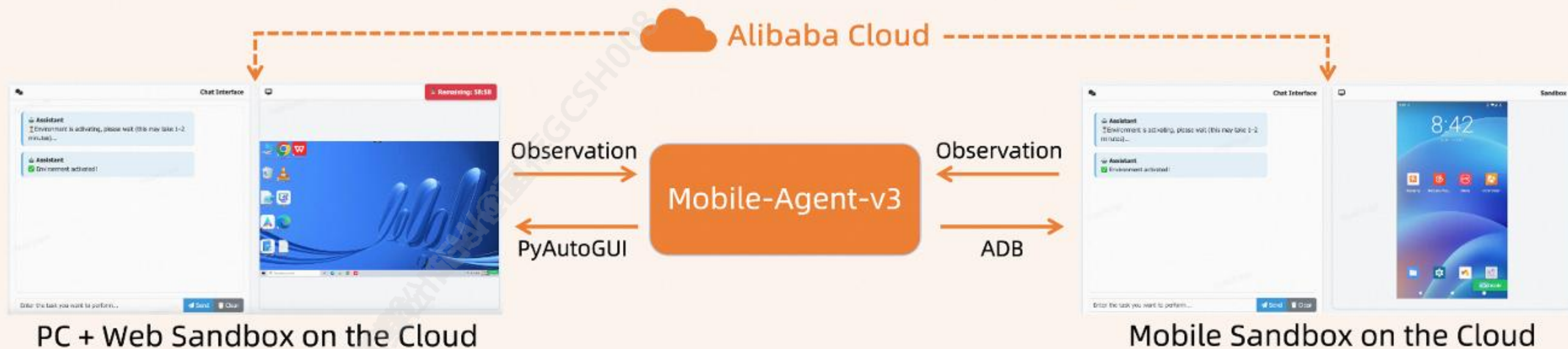


03 Foundation Agent for GUI

[illegible]

Mobile-Agent-v3 & GUI-Owl

Multi-Platform Environment Supporting



ECS集群

虚拟化环境

Android 模拟器

Windows/Linux 模拟器

Web 服务器

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

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 ↑

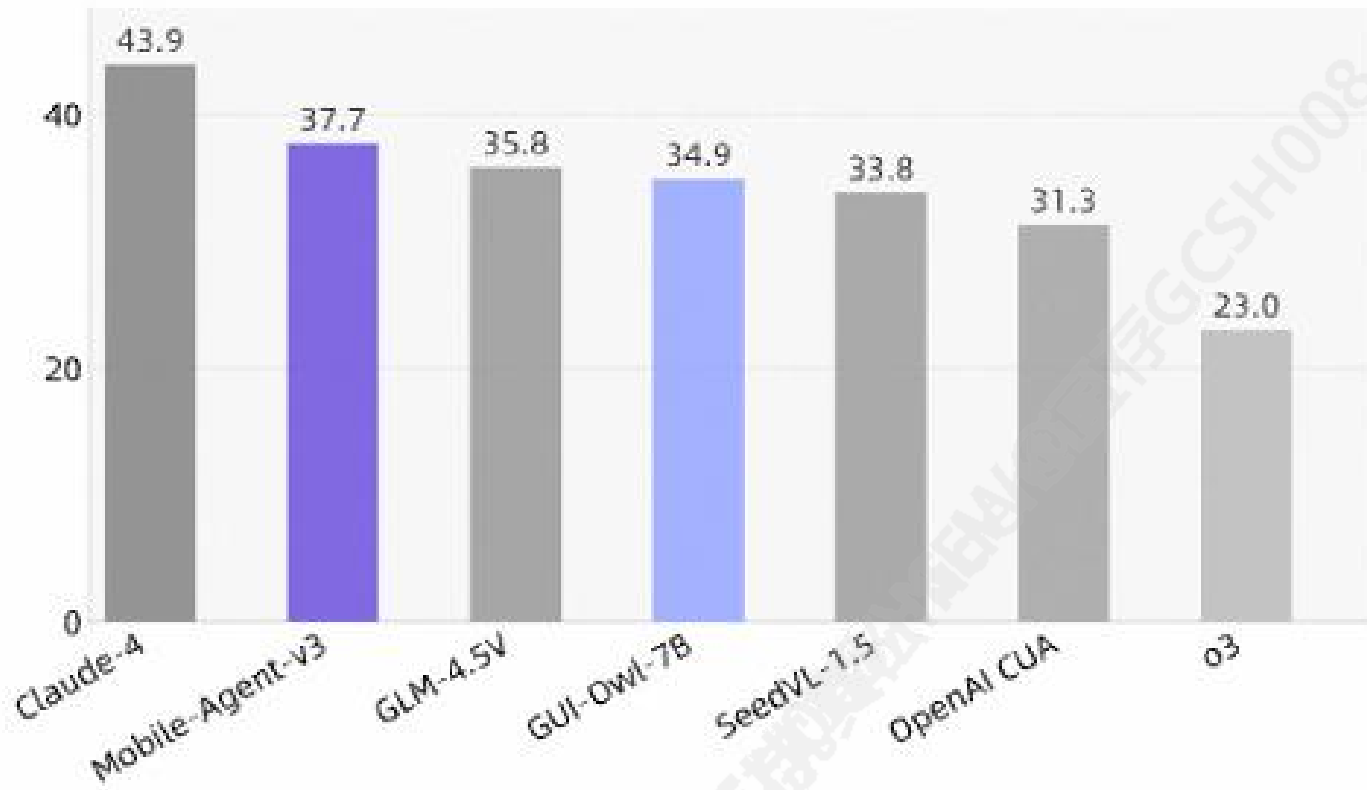
PAI GPU集群

SFT/RL 模型训练

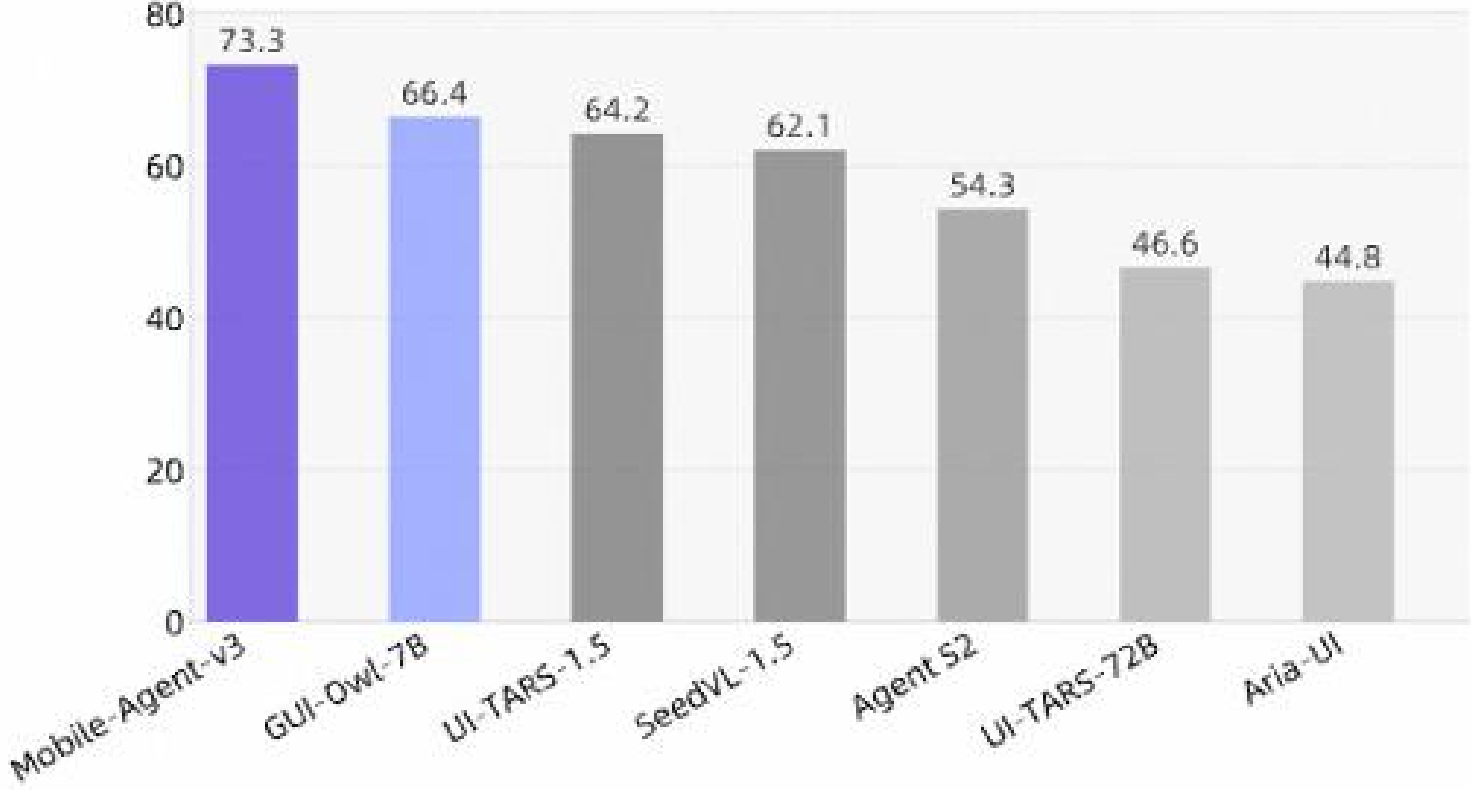
基建架构

GUI自动化开源新SOTA

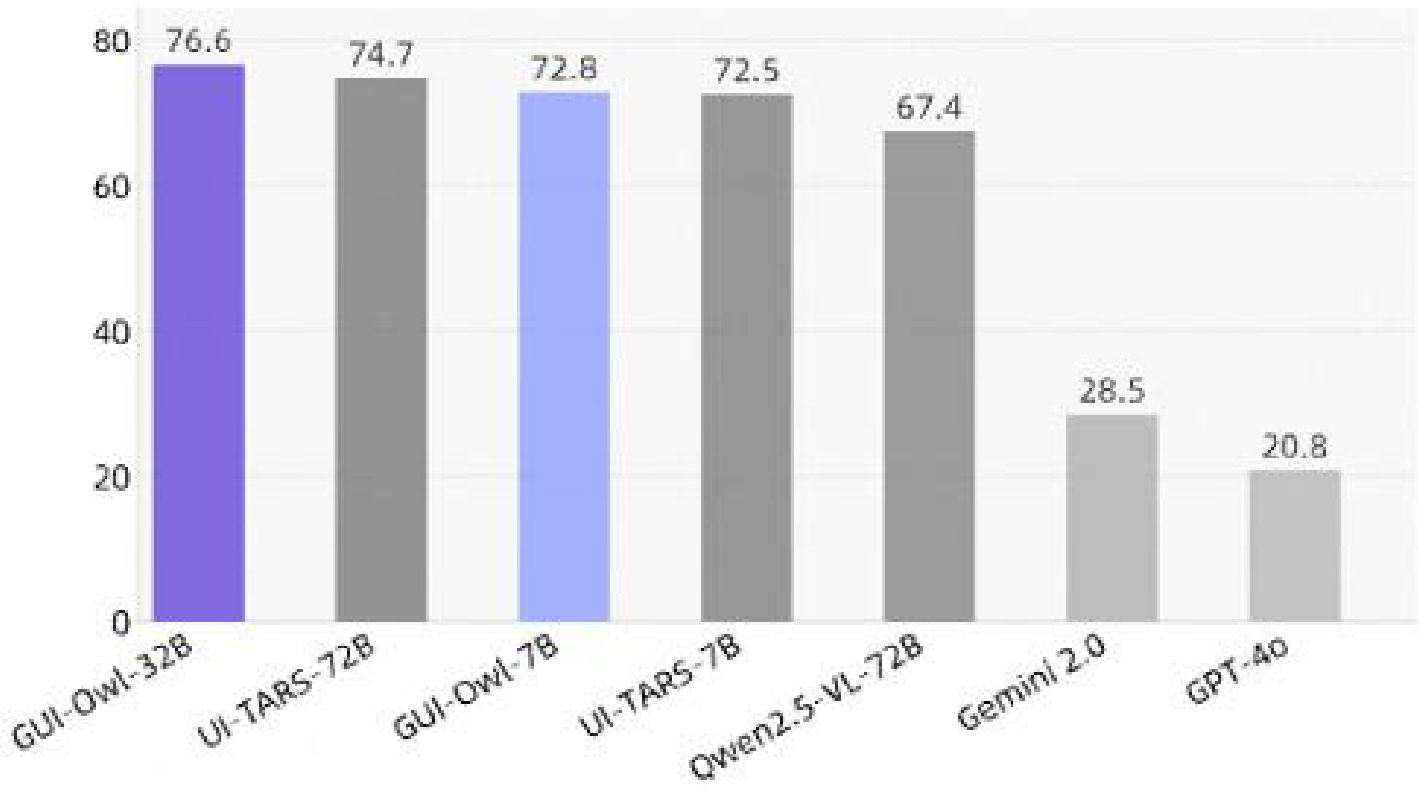
OSWorld



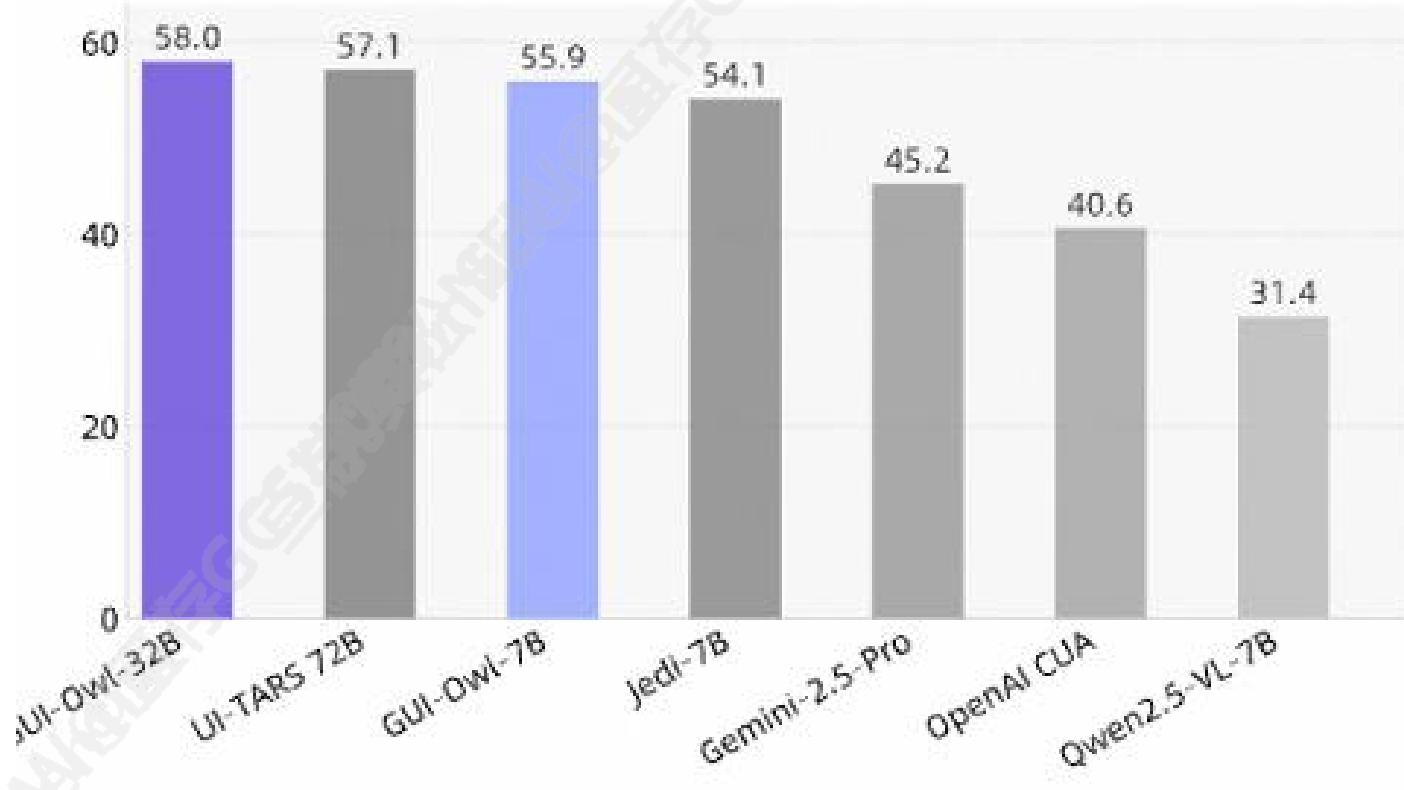
Android World



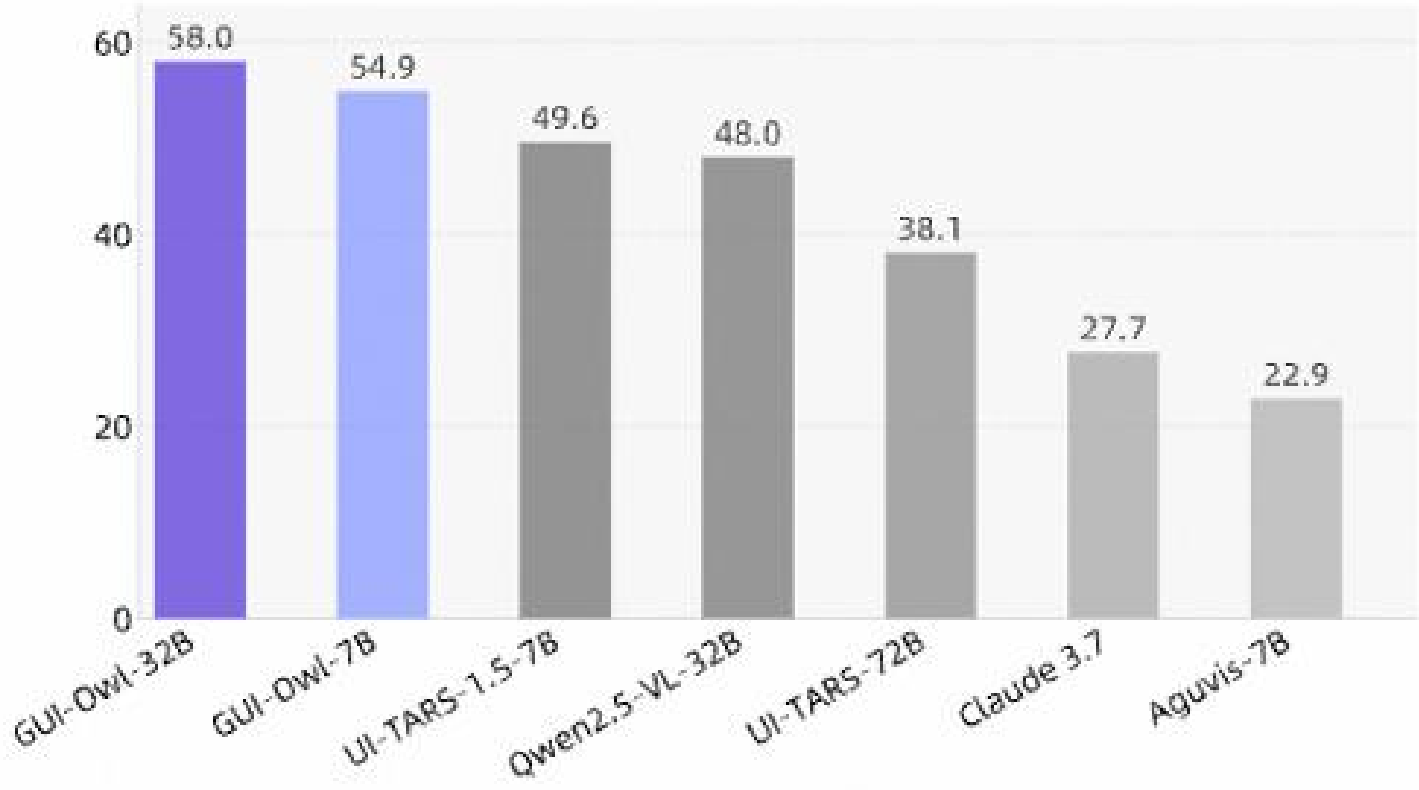
Android Control



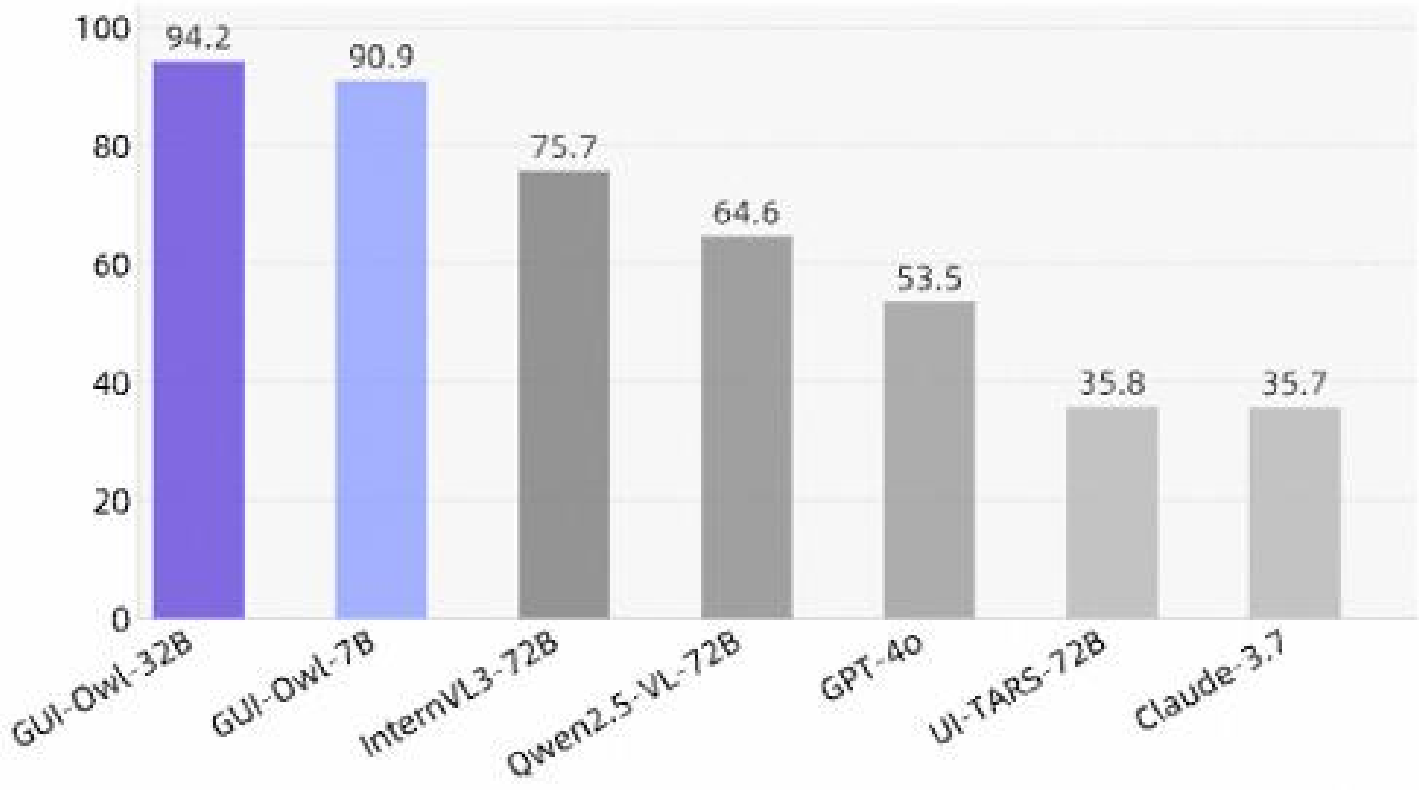
OSWorld-G



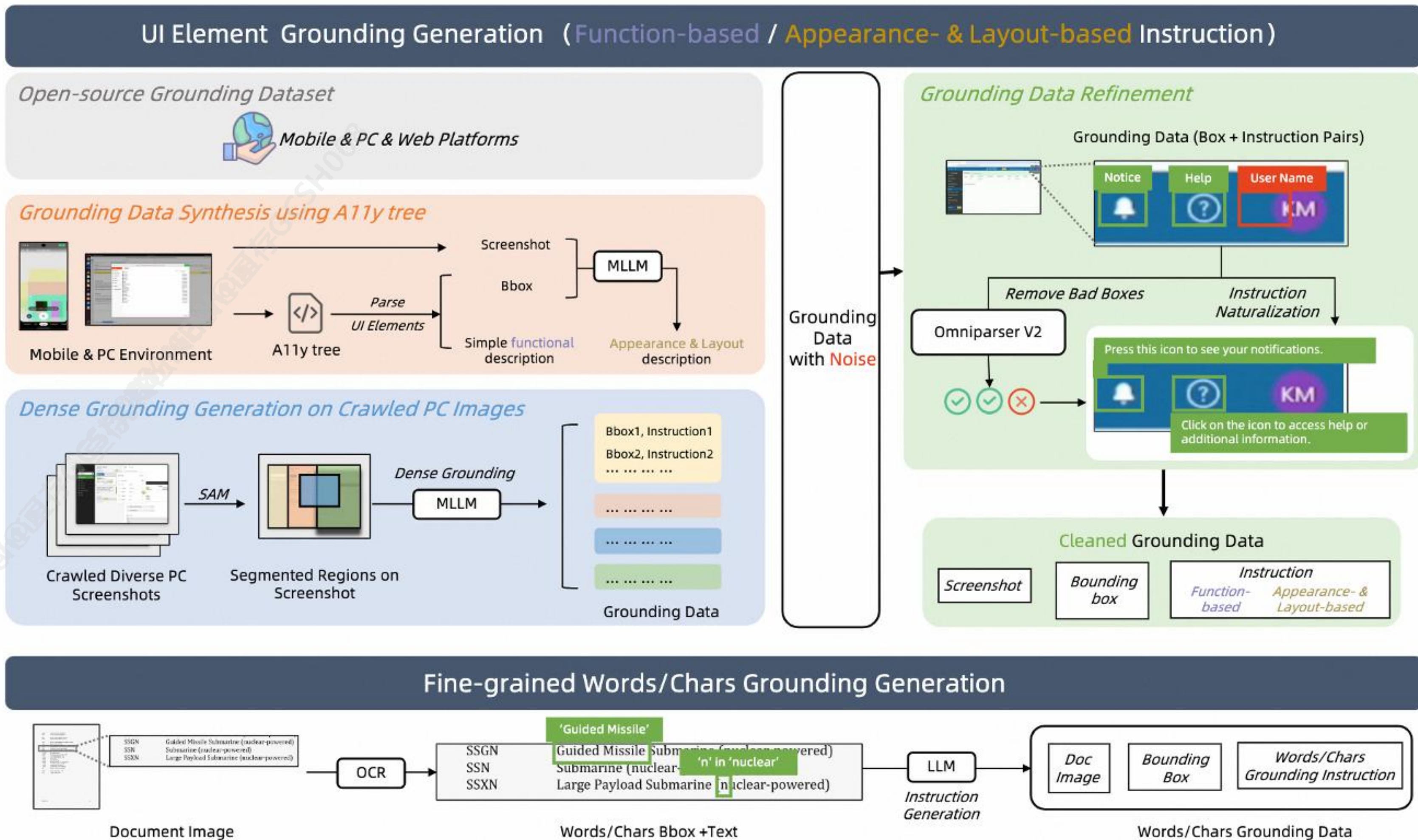
ScreenSpotPro



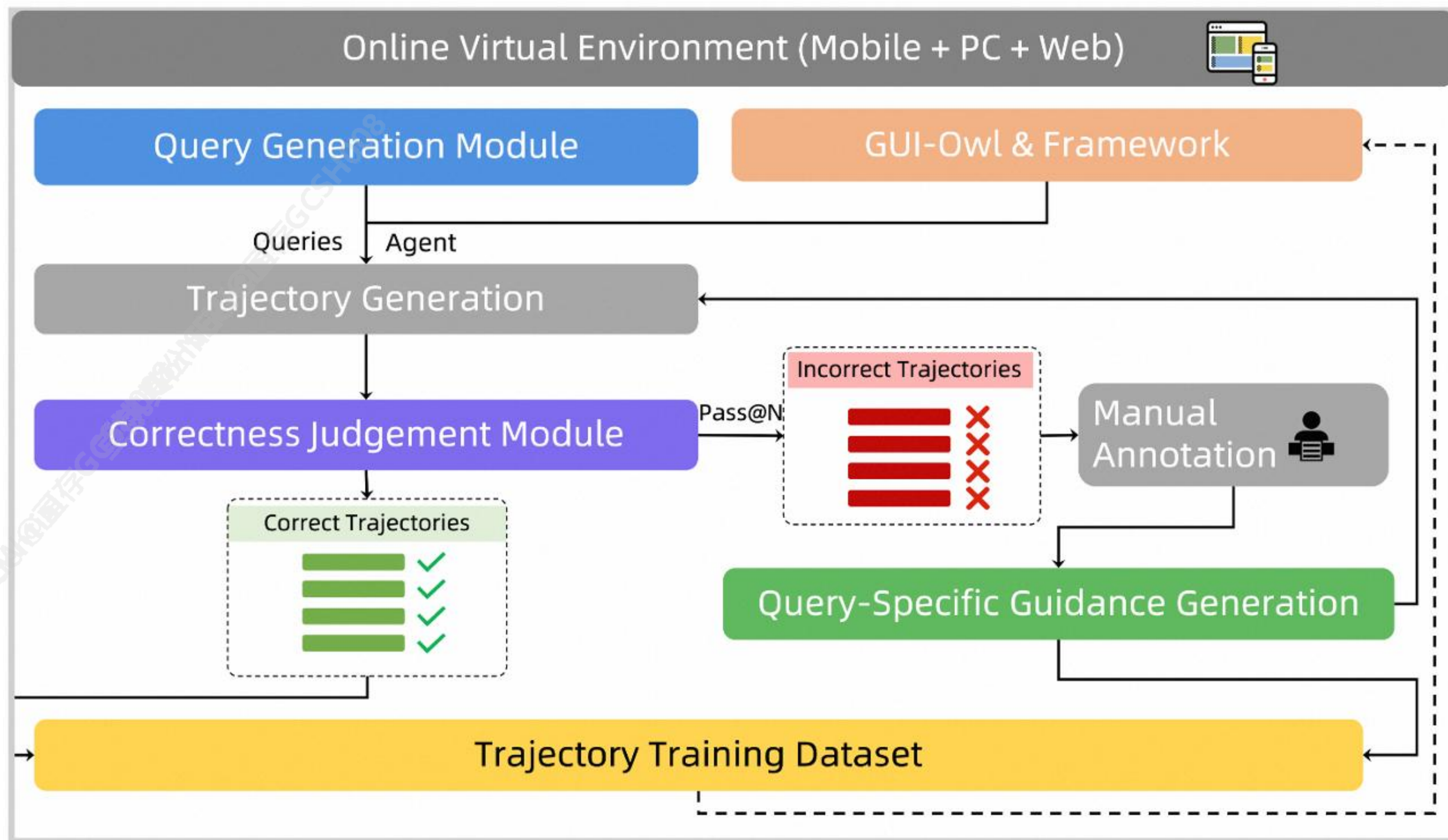
MMBench-GUI L1 (Hard)



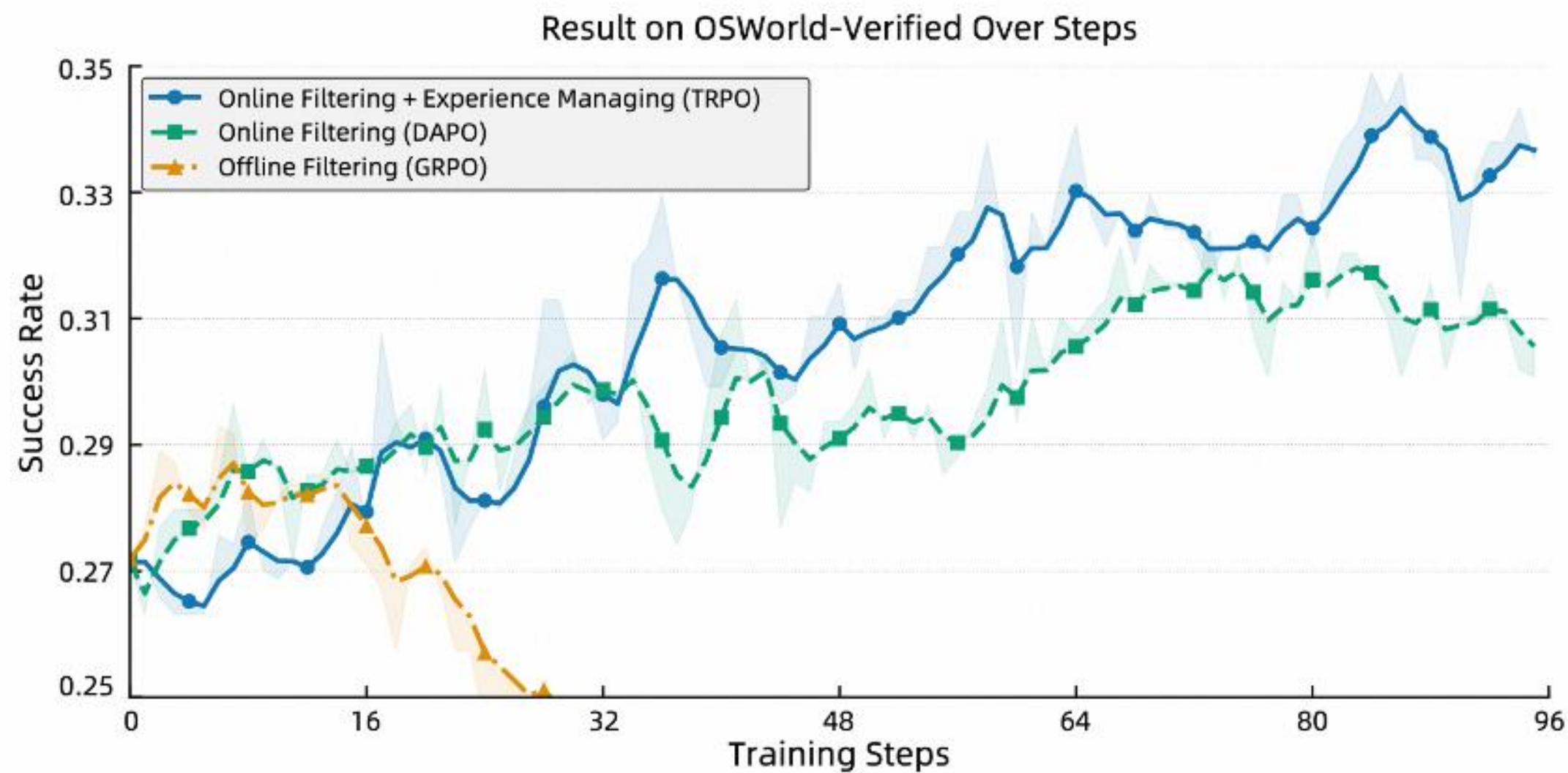
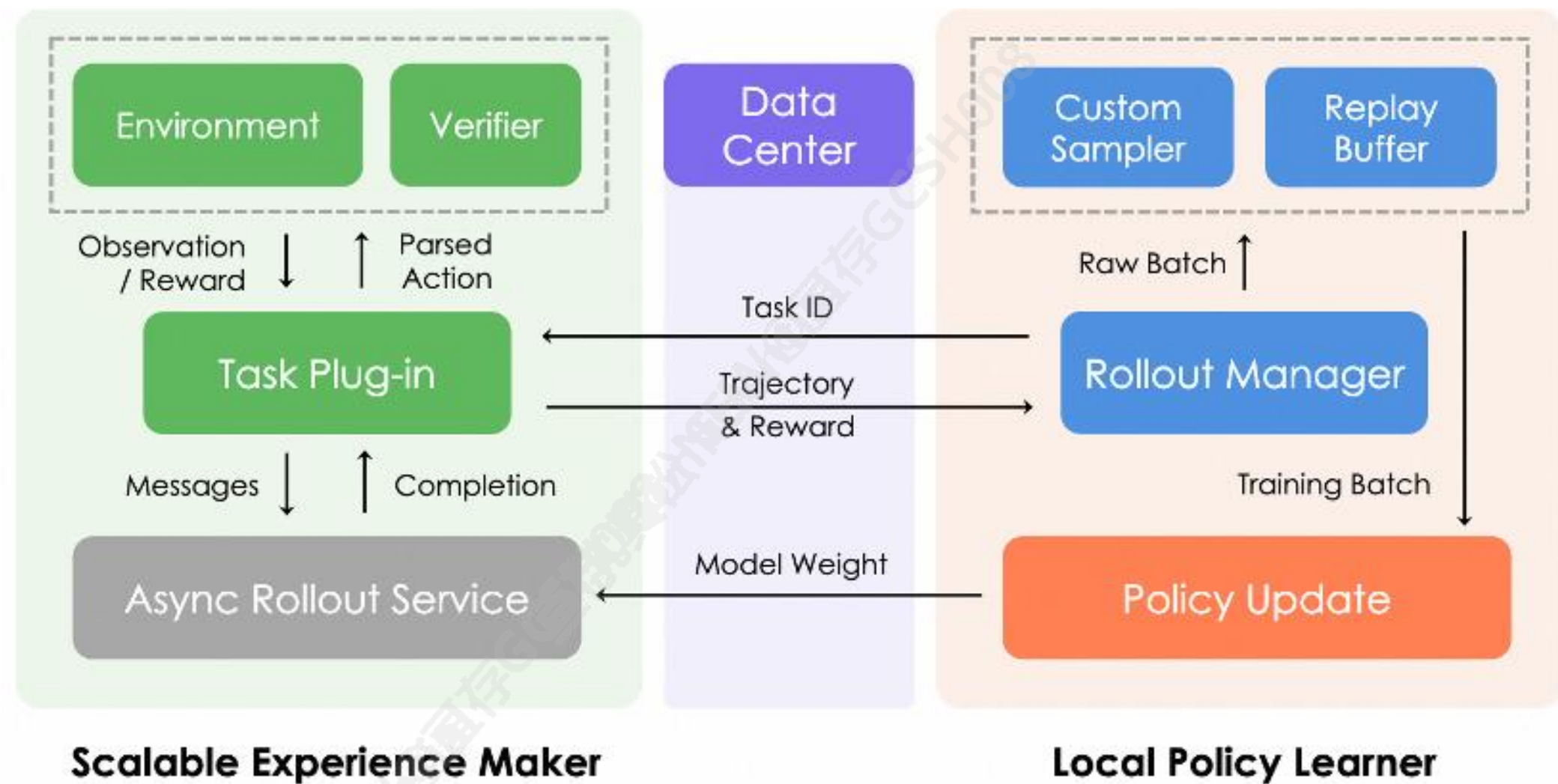
GUI-Owl数据合成链路-Grounding



GUI-Owl数据合成链路-自主进化轨迹合成

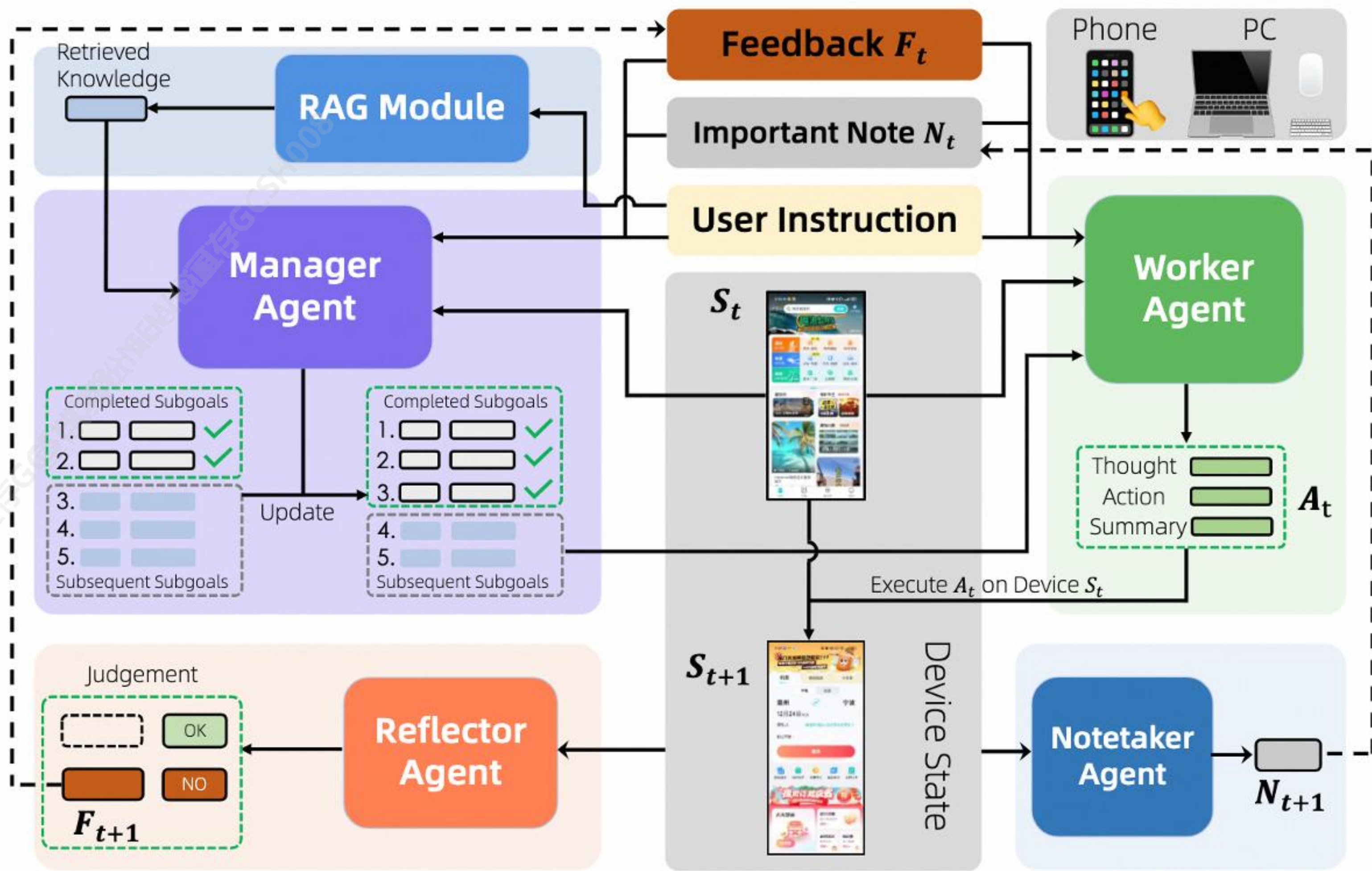


GUI-Owl Agentic RL能力提升



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

Mobile-Agent-V3 Agent框架



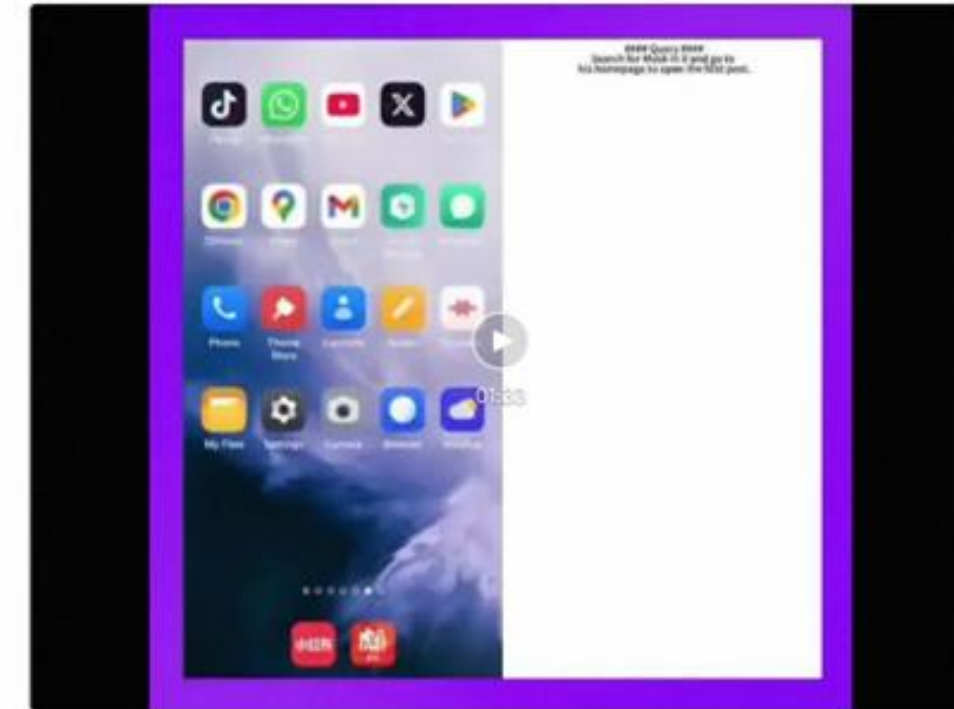
Qwen3-VL：明察、深思、广行

		Qwen3-VL 235B-A22B Instruct	Gemini2.5-Pro Thinking 128	GPT5 Mini Without Thinking	Claude-Opus-4.1 Without Thinking	Other Best Without Thinking
STEM&Puzzle	MMMU _{val}	78.7	80.9	74.4*	77.2	73.4* (Sawt. 1%)
	MMMU_Pro	68.1	71.2	62.7*	60.7	59.9* (Sawt. 1%)
	MathVista _{mini}	84.9	77.7	50.9	74.5	83.0* (Sawt. 1%)
	MathVista	66.5	66.0	45.8	57.7	65.5* (Sawt. 1%)
	MathVerse _{mini}	72.5	65.9	43.0	68.1	65.4* (Sawt. 1%)
General VQA	VisuLogic	29.9	26.9	27.2	27.2	33.0* (Sawt. 1%)
	MMBench_EN_V1.1 _{dev}	89.9	86.6	81.4	84.1	89.0* (Sawt. 1%)
	RealWorldQA	79.3	76.0	77.3	68.5	78.0* (Sawt. 1%)
	MMStar	78.4	78.5	65.2	71.0	76.2* (Sawt. 1%)
	SingleVQA	63.0	66.9	56.7	55.7	63.1* (Sawt. 1%)
Subjective Experience and Instruction Following	HallusionBench	63.2	60.9	53.7	55.1	60.0* (Sawt. 1%)
	MM_MT_Bench	8.5	7.6	7.5	7.9	7.8* (Sawt. 1%)
	MIABench	91.3	91.3	92.6	90.0	90.8 (Sawt. 1%)
	MMLangBench-Doc	57.0	51.2	42.4	48.1	50.1 (Sawt. 1%)
	DocVQA ₁₁₃₁	97.1	94.0	89.4	89.2	96.7* (Sawt. 1%)
Text Recognition and Chart/Document Understanding	InfoVQA ₁₁₃₁	89.2	82.9	69.9	60.9	87.3* (Sawt. 1%)
	A127 ₁₁₃₁	89.7	90.0	84.1	84.4	89.7* (Sawt. 1%)
	OCRBench	920.0	872.0	787.0	750.0	904* (Sawt. 1%)
	OCRBenchV2 _{gen/h}	67.1 / 61.8	55.2 / 53.1	48.2 / 37.7	47.2 / 38.0	61.5 / 63.7* (Sawt. 1%)
	CC_OCR	82.2	76.8	66.1	66.0	79.8* (Sawt. 1%)

		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* (Sawt. 1%)
	MMMU_Pro	69.3	68.8*	78.4*	64.8	70.1* (Sawt. 1%)
	MathVista _{mini}	85.8	82.7*	81.3	75.5	85.6* (Sawt. 1%)
	MathVista	74.6	73.3*	70.9	64.3	69.9* (Sawt. 1%)
	MathVerse _{mini}	85.0	82.9	84.1	70.6	72.1* (Sawt. 1%)
General VQA	ZEROBench	4.0	3.0	2.0	1.0	2.0* (Sawt. 1%)
	ZEROBench_Sub	27.7	33.8	24.6	26.3	30.8* (Sawt. 1%)
	VisuLogic	34.4	31.6	28.5	27.9	35.0* (Sawt. 1%)
	MMBench_EN_V1.1 _{dev}	90.6	90.1*	85.9	84.9	89.9* (Sawt. 1%)
	RealWorldQA	81.3	78.0*	82.8	69.9	79.1* (Sawt. 1%)
Subjective Experience and Instruction Following	MMStar	78.7	77.5*	76.4	72.1	77.9* (Sawt. 1%)
	SingleVQA	61.3	65.4*	61.8	56.7	63.4* (Sawt. 1%)
	HallusionBench	66.7	63.7*	65.7	60.4	65.4* (Sawt. 1%)
	MM_MT_Bench	8.5	8.4	7.6	7.8	7.7* (Sawt. 1%)
	MIABench	92.7	92.3	92.4	91.2	92.3 (Sawt. 1%)
Text Recognition and Chart/Document Understanding	MMLangBench-Doc	56.2	55.6	51.5	54.5	52.7 (Sawt. 1%)
	DocVQA ₁₁₃₁	96.5	92.6	91.5	92.5	96.9* (Sawt. 1%)
	InfoVQA ₁₁₃₁	89.5	84.2	79.0	69.4	88.4* (Sawt. 1%)
	A127 ₁₁₃₁	89.2	90.9	89.7	86.4	91.2* (Sawt. 1%)
	OCRBench	875.0	846.0	810.0	764.0	907.0* (Sawt. 1%)

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

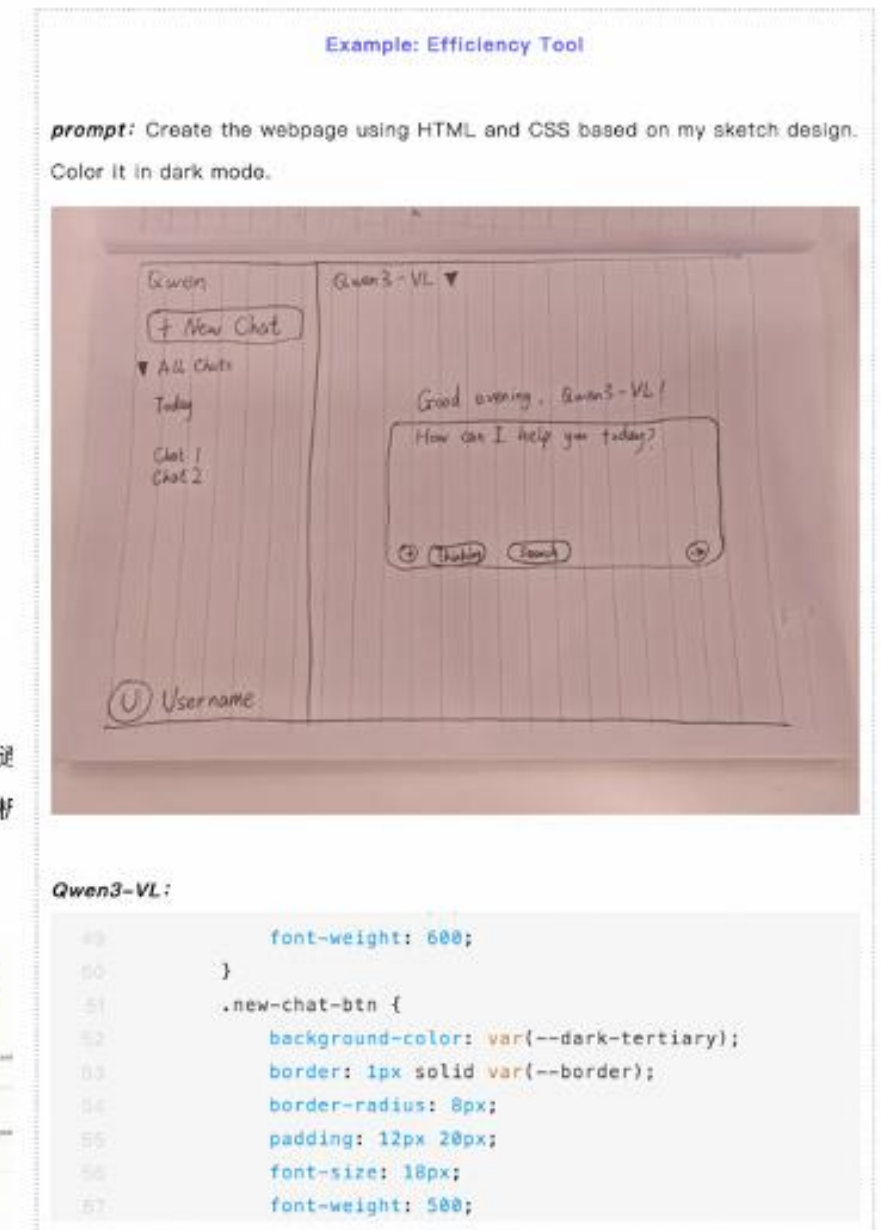
Qwen3-VL 可以像人类一样仔细观察图像的局部细节，并结合工具进行复杂推理。比如通过路边的路牌判断具体位置，或根据人物照片搜索相关信息，完成细粒度识别和逻辑分析任务。



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，能把手给草图转成网页代码，或帮助调试界面问题，提升开发效率。

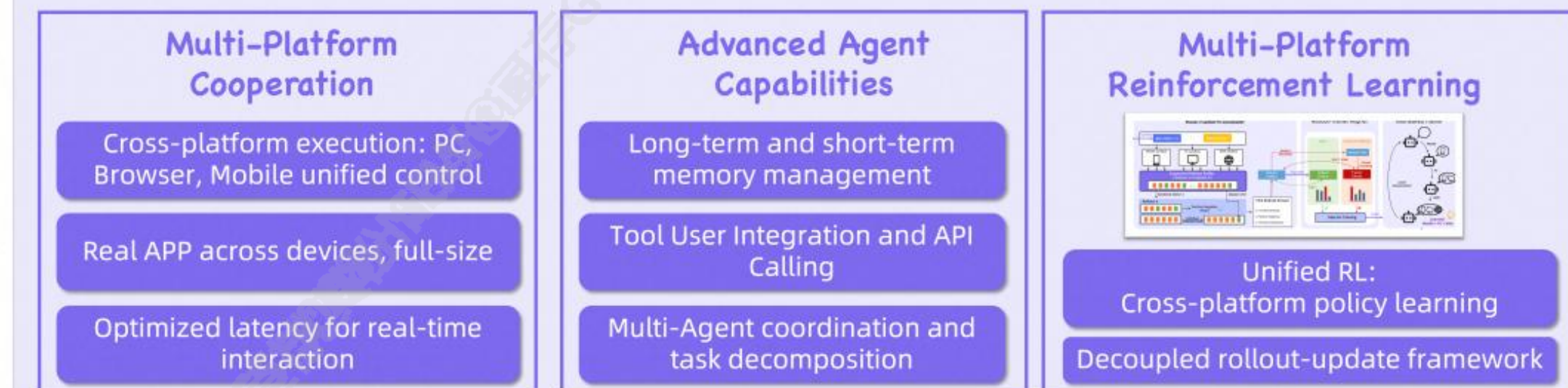


Mobile-Agent-v3.5 & GUI-Owl-1.5

Multi-Platform Environment Supporting

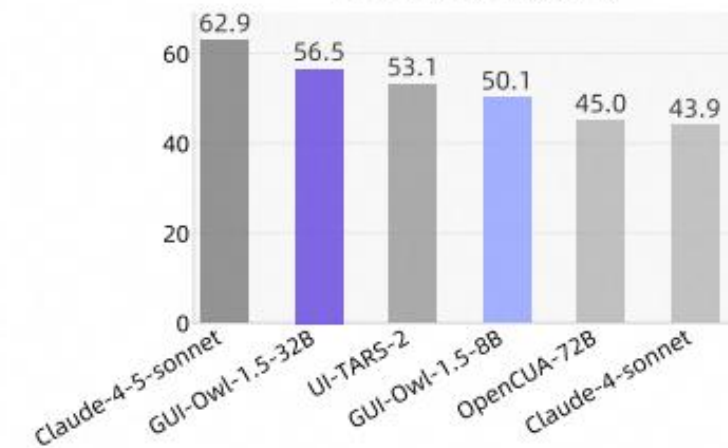


Mobile-Agent-v3.5 Highlight Capability

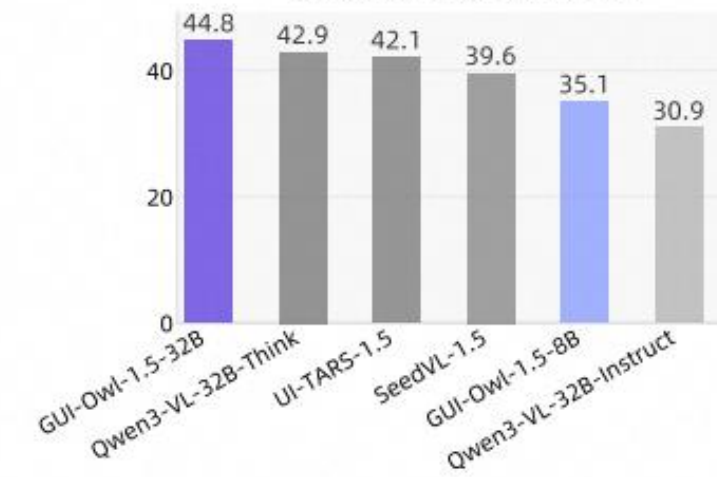


SOTA GUI-Agent Model across 20+ GUI benchmarks!

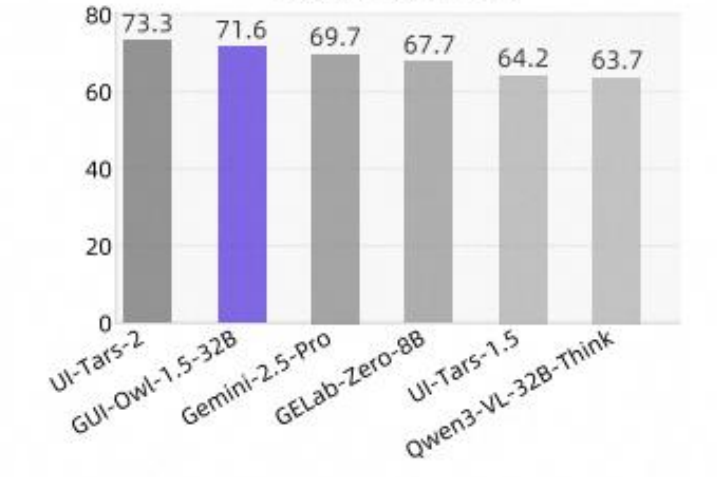
OSWorld-Verified



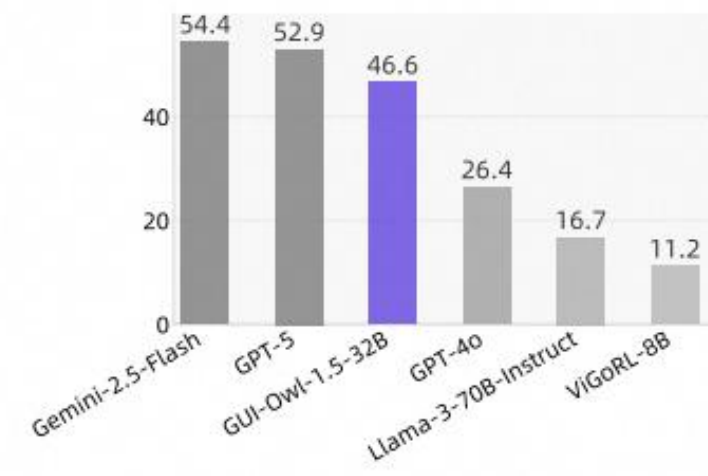
WindowsAgentArena



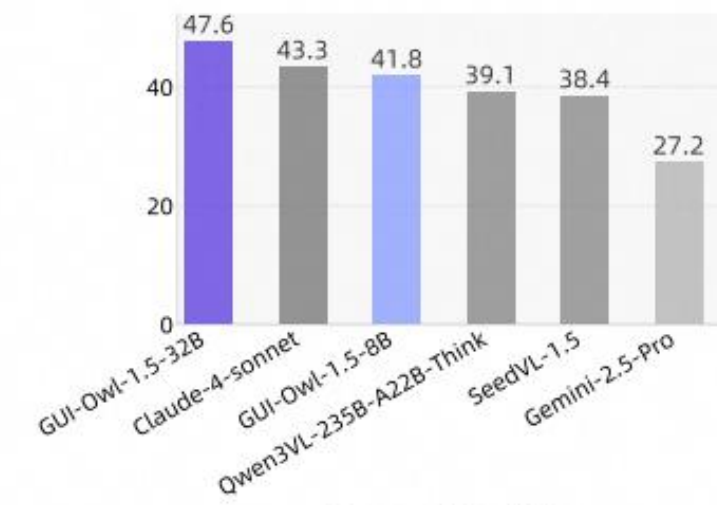
AndroidWorld



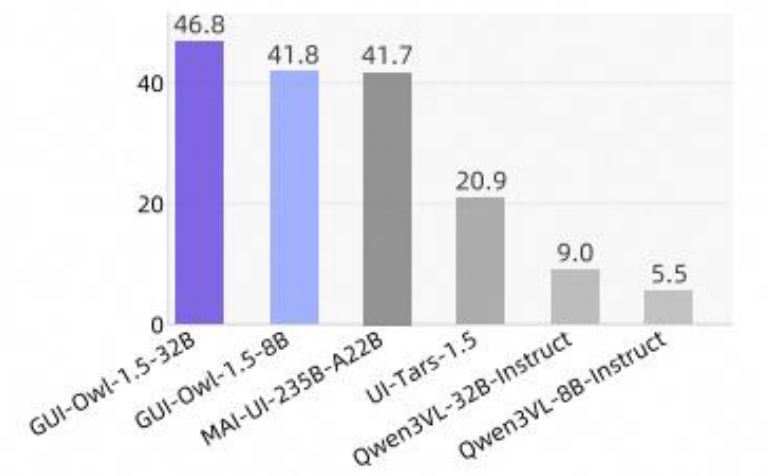
VisualWebArena



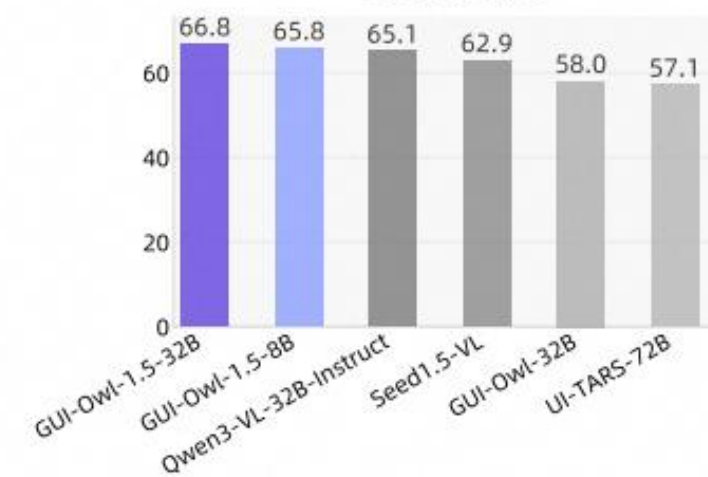
OSWorld-MCP



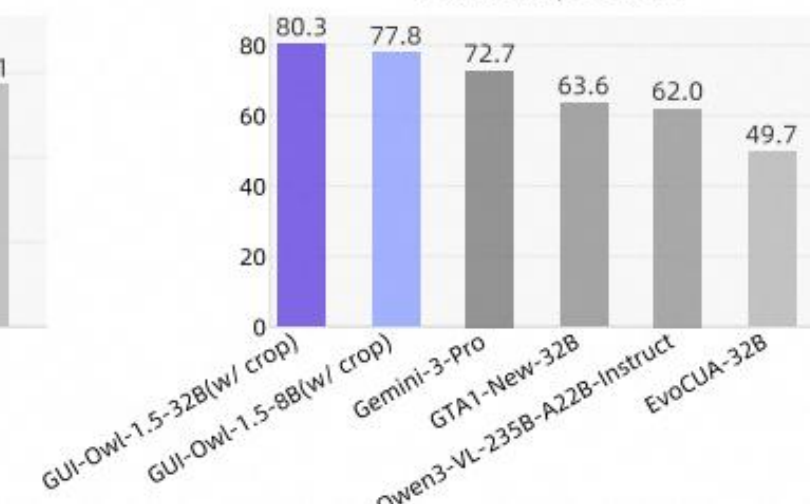
MobileWorld



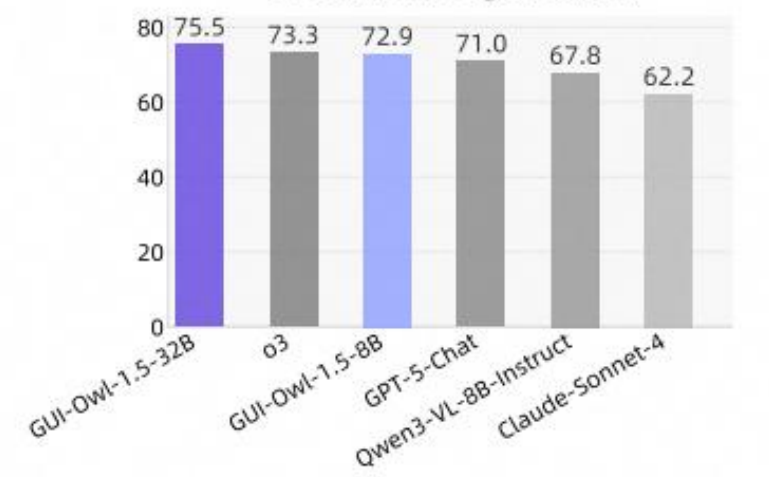
OSWorld-G



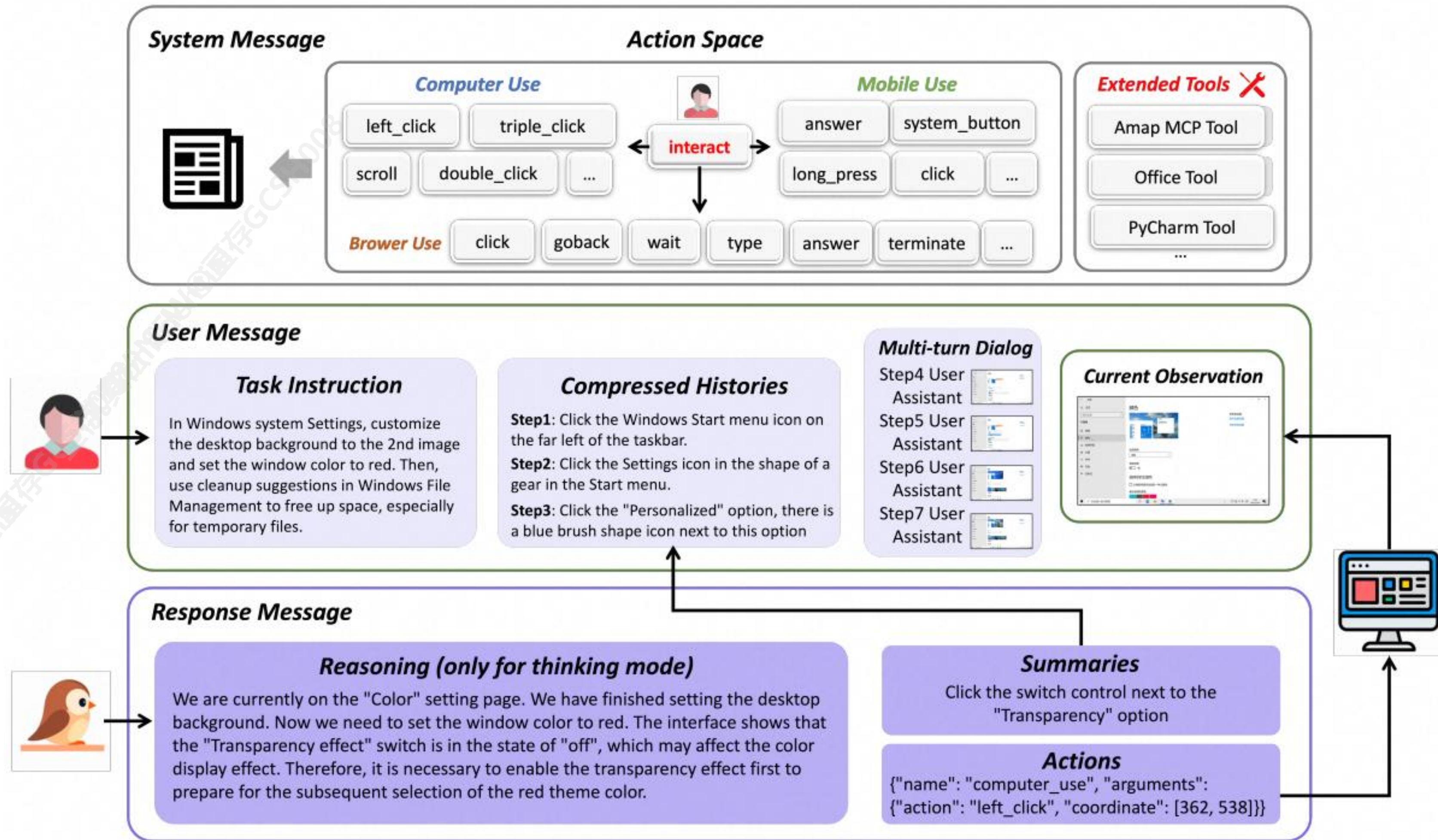
ScreenSpot-Pro



GUI Knowledge Bench



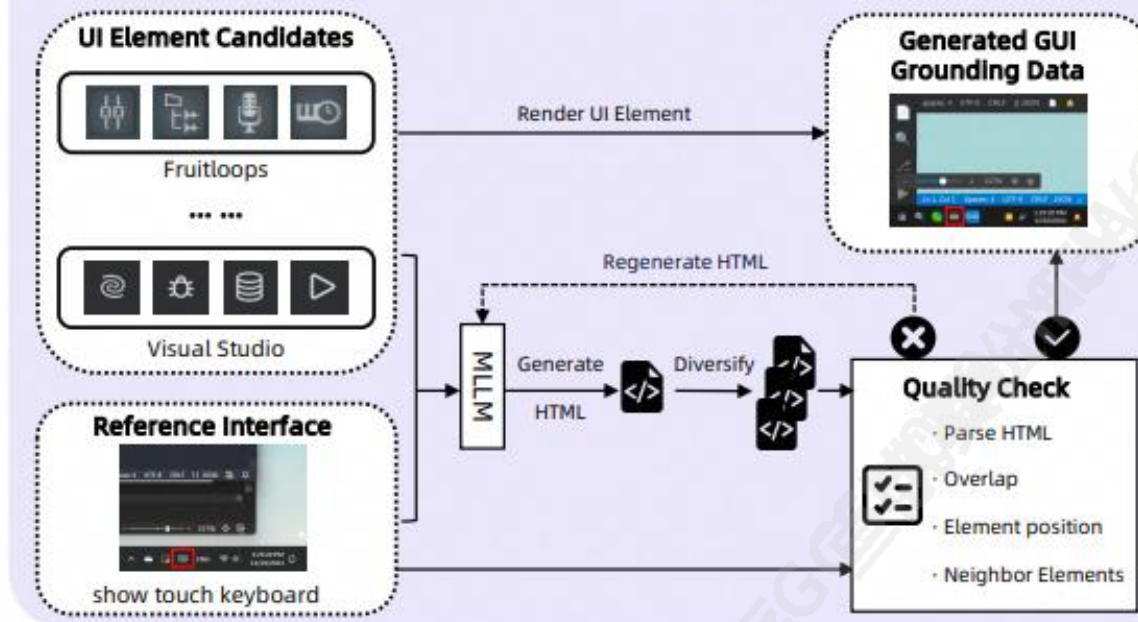
Mobile-Agent-v3.5 & GUI-Owl-1.5



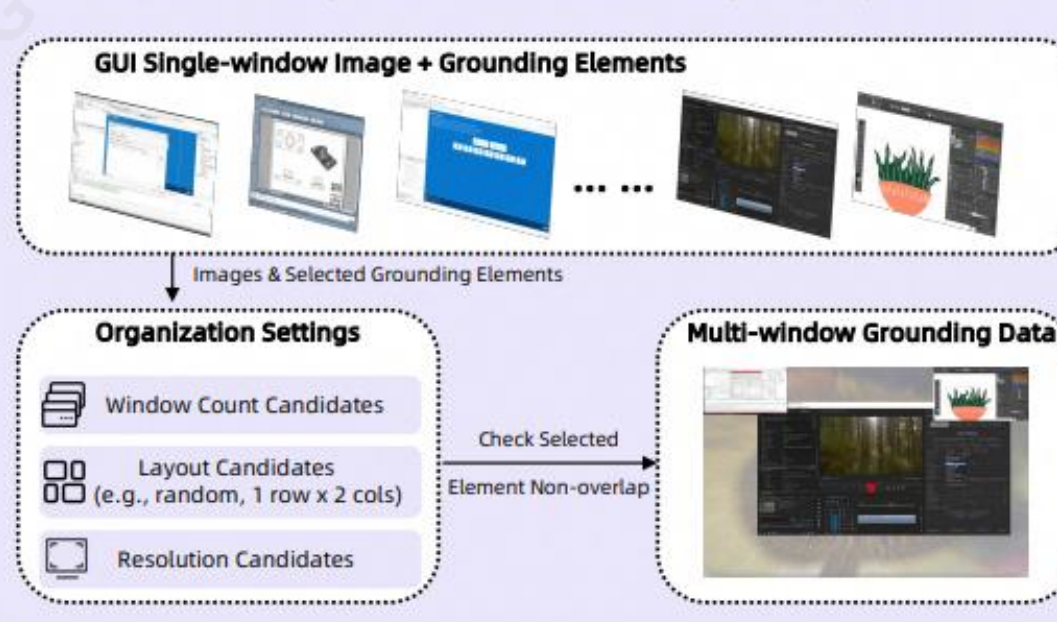
Mobile-Agent-v3.5 & GUI-Owl-1.5

Hard Grounding Data Synthesis

Challenging App GUI Grounding Data Synthesis



Multi-window High-resolution Grounding Data Synthesis

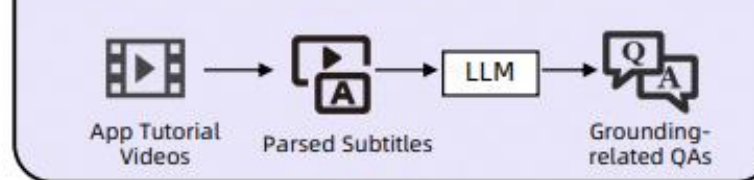


High-Quality Grounding Data Extension

Grounding Data Extraction from Trajectories



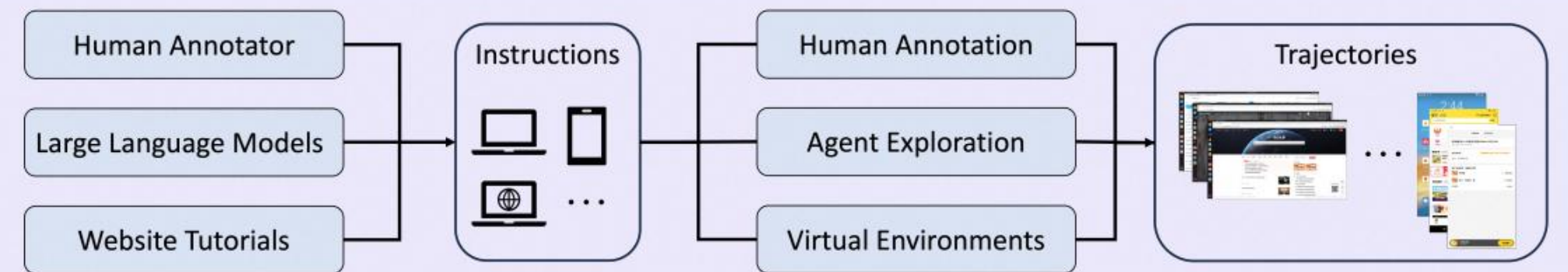
Grounding-related QAs Generation



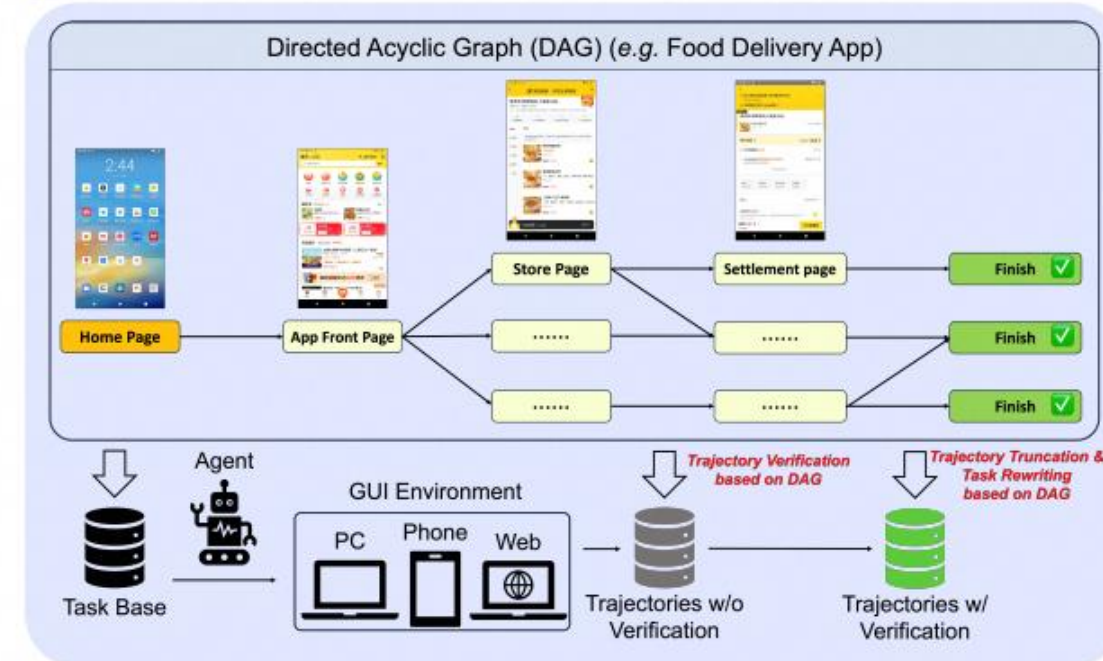
Infeasible Grounding Data Generation



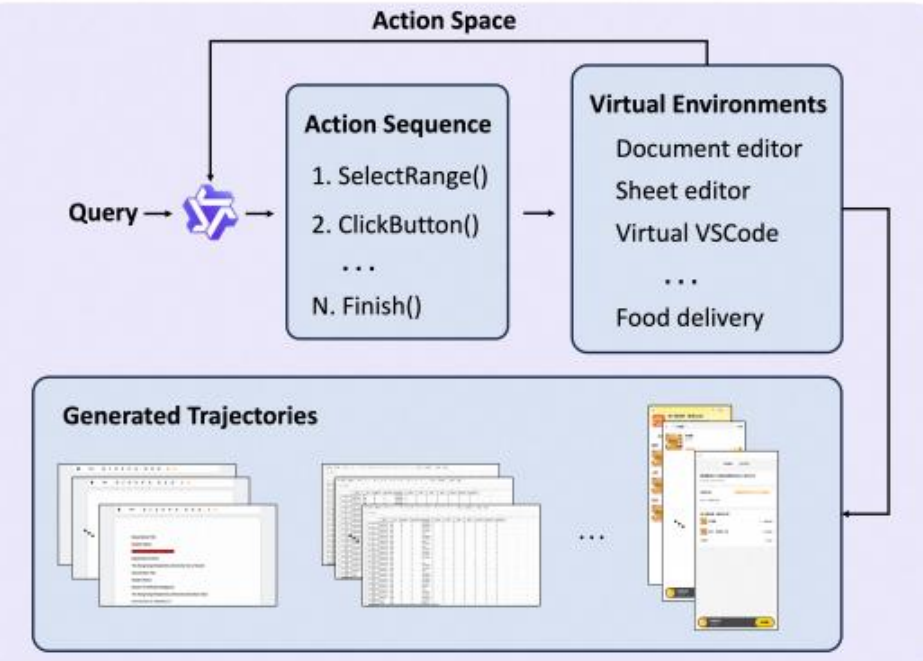
Trajectory Collection Pipeline



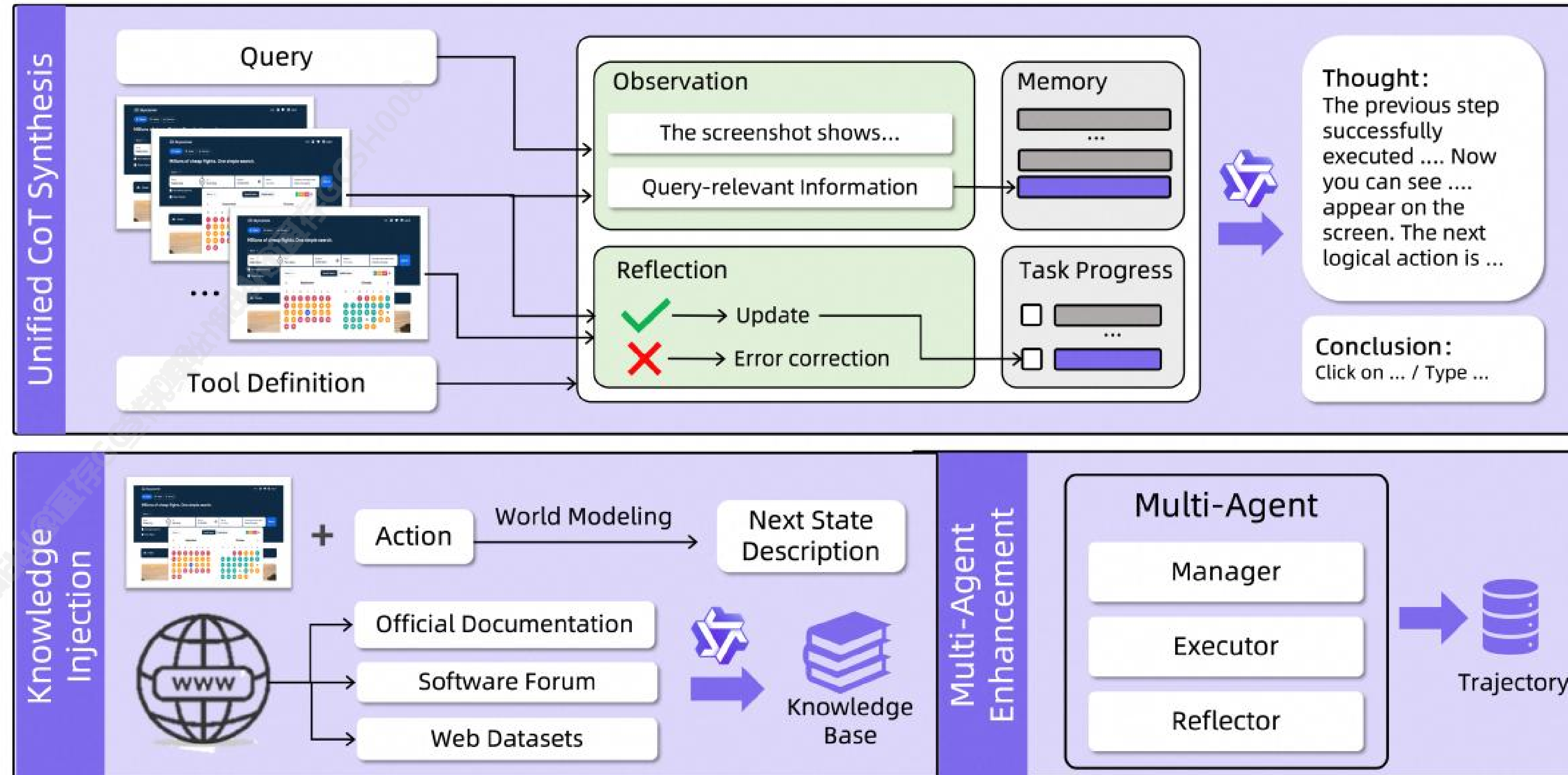
Query Generation & Trajectory Verification & Task Rewriting



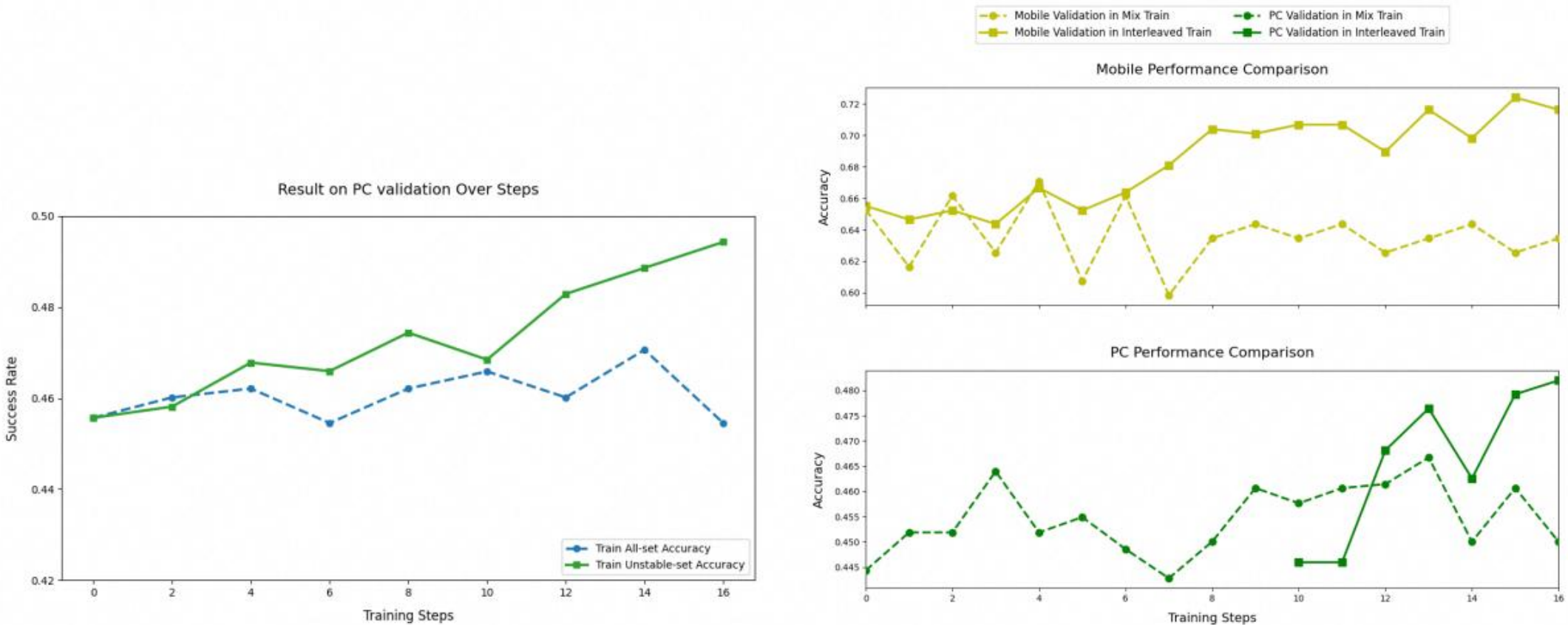
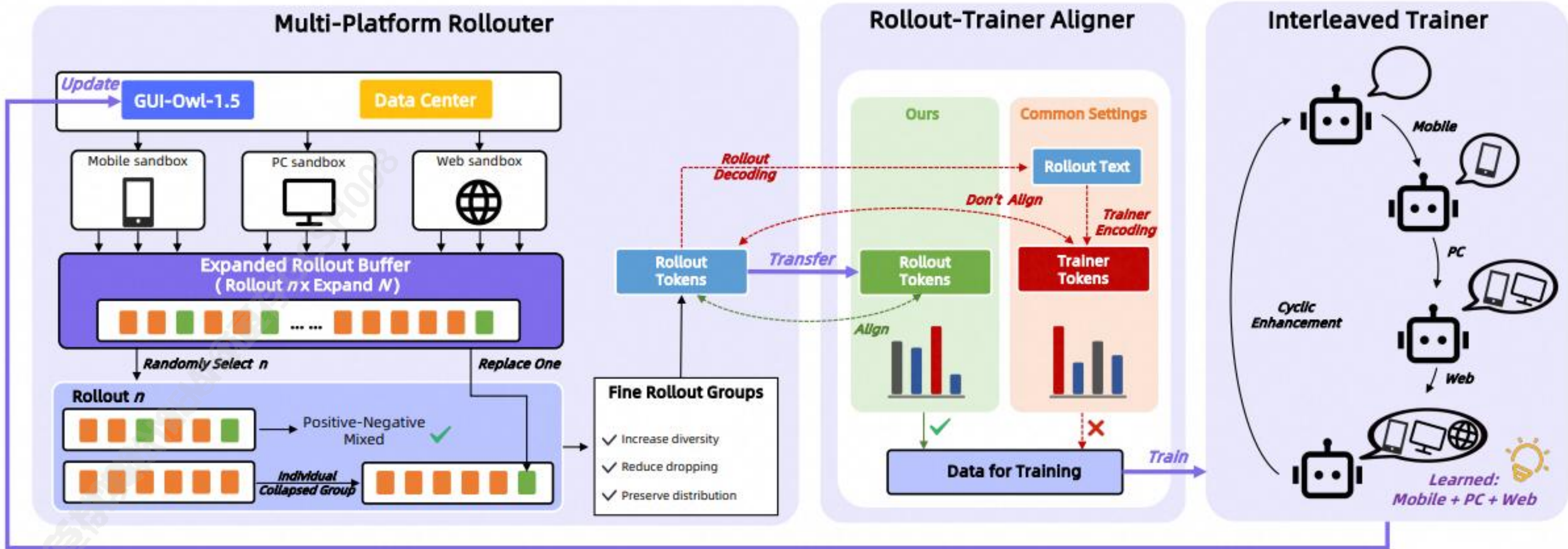
Virtual Environment



Mobile-Agent-v3.5 & GUI-Owl-1.5



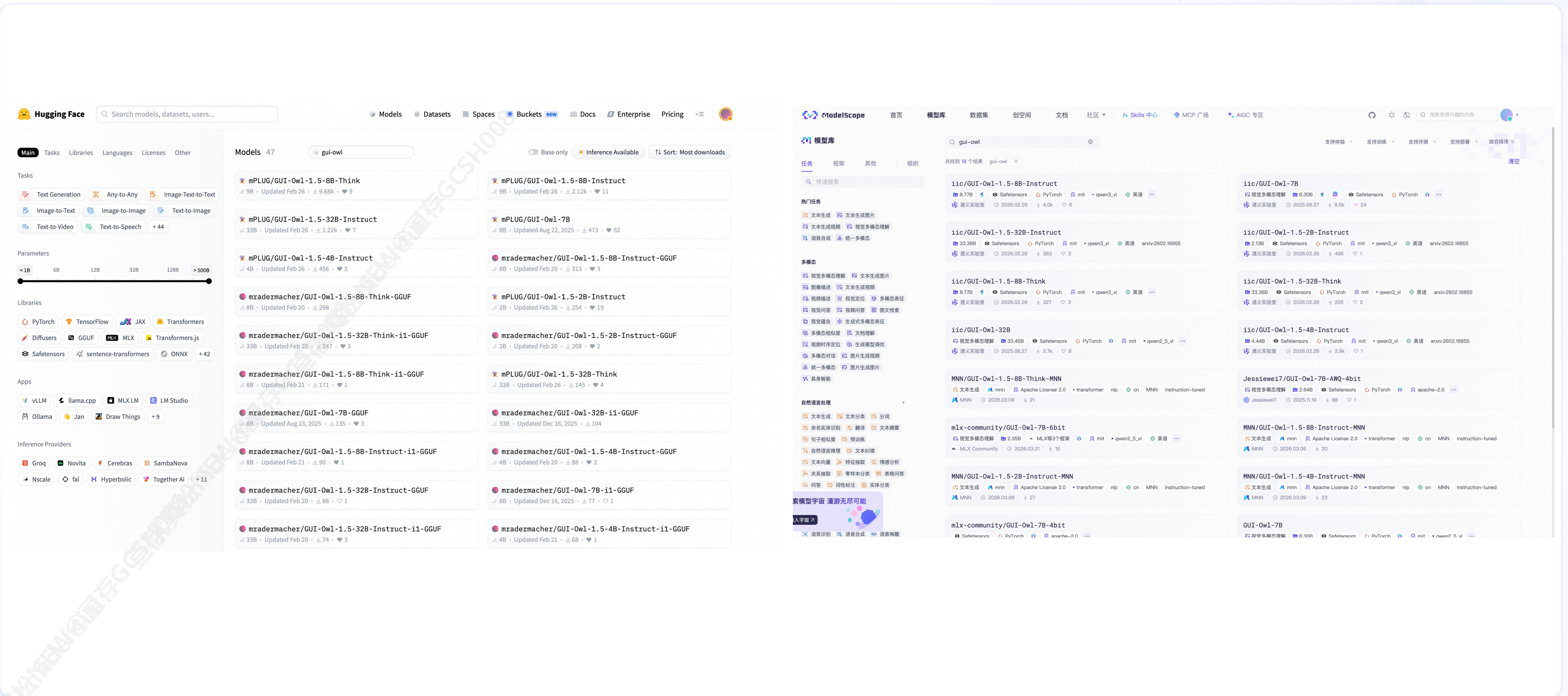
Mobile-Agent-v3.5 & GUI-Owl-1.5



(a) All-set Train VS Unstable-set Train

(b) Mix-platform Train VS Interleaved Train

Mobile-Agent-v3.5 & GUI-Owl-1.5开源应用



Mobile-Agent-v3.5 & GUI-Owl-1.5开源应用

MobileAgent

运行中

MobileAgentTest/computer_use

MobileAgent测试

36.3k次访问

2026-04-03 更新

关联内容 (1)

复制

16

+

合集

空间内容

交流反馈 1

空间管理

 通义 Mobile-Agent-V3.5

MobileAgent Git网址

百炼API在线调用

无影AI云手机

钉钉扫码加入交流群

微信扫码加入交流群

任务配置

设备选择

电脑

模型名称

gui-plus-2026-02-26

enable_thinking

False

激活环境

最大步数

10

补充信息

请输入PC_USE模型补充信息

聊天界面

欢迎使用Computer Use Agent!

请在下方输入您想要执行的任务。

示例任务: use the web browser to get the current weather in Hangzhou via Bing Search

无影

等待Sandbox启动

请在左侧输入任务以启动Computer Use Agent

Mobile-Agent-v3.5 & GUI-Owl-1.5开源应用

三

阿里云百炼

模型应用订阅Token Plan体验文档API 参考

模型广场

模型体验

文本模型

语音模型

视觉模型

全模态模型

向量模型

模型推理

TPM 预留

模型训练

数据管理

模型调优

我的模型

模型评测

模型压缩

模型广场 / GUI-Plus-2026-02-26

中国内地

填写问卷赢好礼

模型Code

gui-plus-2026-02-26

模型介绍

视觉理解

深度思考

GUI系列模型，相较于GUI-Plus的模型，支持思考模式与非思考模式，模型在处理跨平台、多app任务的理解和执行等效果得到了全方位的大幅度提升。支持工具调用。此版本为2026年2月26日快照。

模型能力

输入模态		输出模态	
模型体验	×	前缀续写 ①	×
function calling ①	✓	cache 缓存 ①	×
结构化输出 ①	×	批量推理	×
联网搜索 ①	×	模型调优	×

模型价格

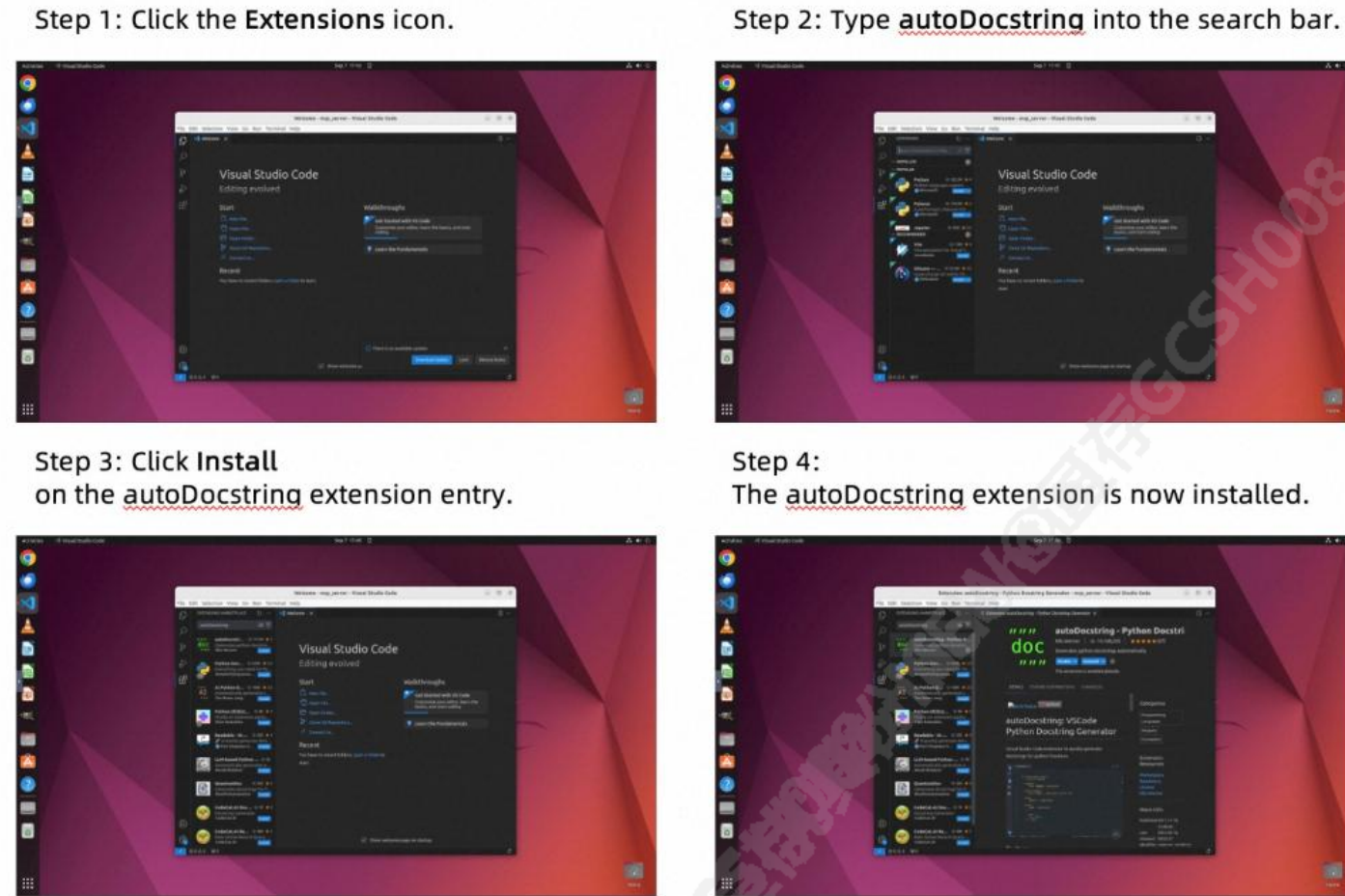
⇌ 切换千 tokens

输入	1.5	输出	4.5
	元/每百万 tokens		元/每百万 tokens

GUI工具调用&Benchmark

OSWorld-MCP (158高质量MCP工具)

Tool-CUA (GUI-Tool最优路径选择)



(a) a GUI-based approach

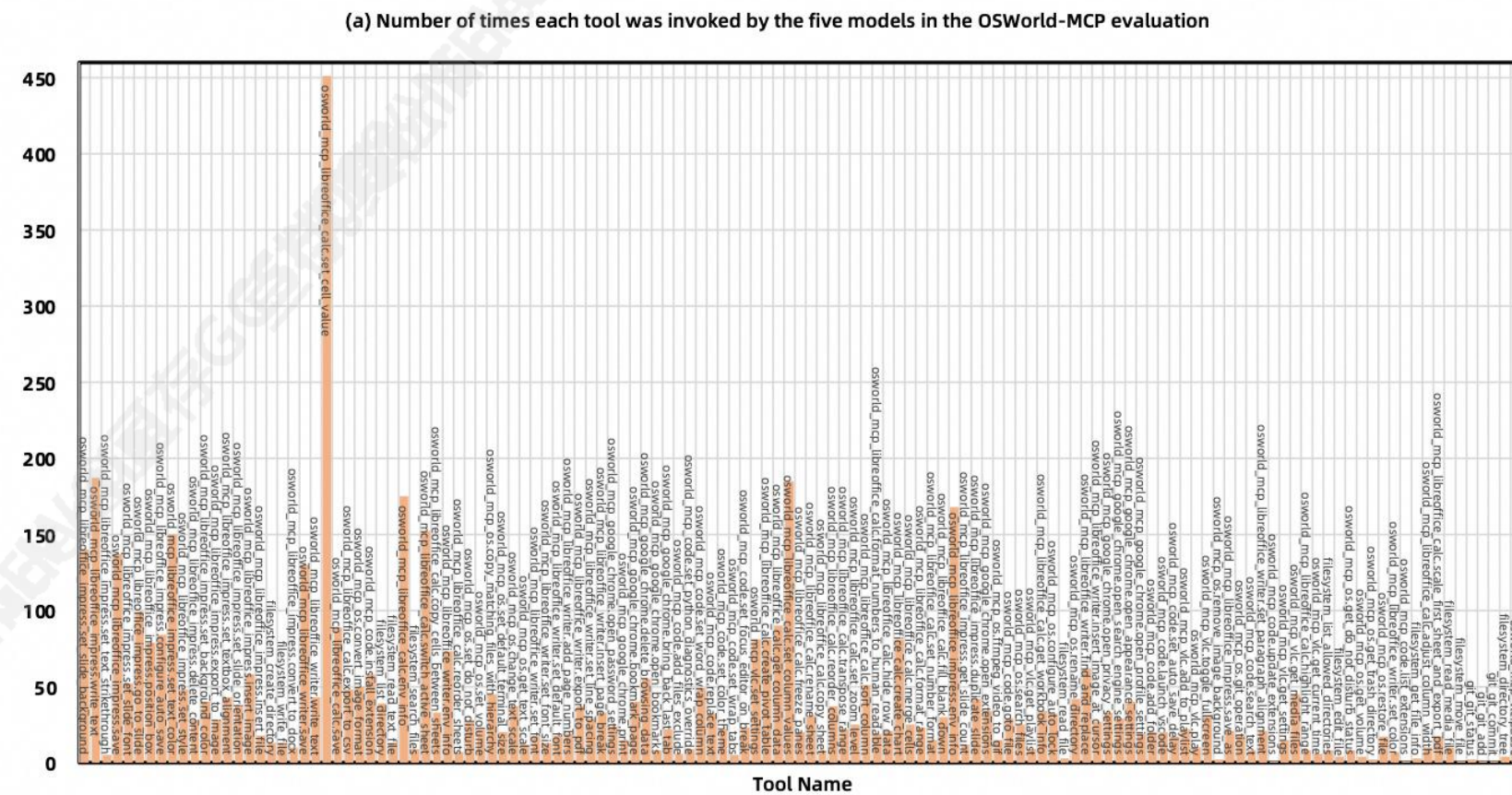
```

1 @mcp.tool
2 def install_extension(cls, extension_id,
3   pre_release=False):
4     """
5     Installs an extension or updates it in
6     VSCode.
7
8     Args:
9         extension_id (str)
10        pre_release (bool)
11    """
12    try:
13        command = ["code", "--install-
14        extension", extension_id]
15        if pre_release:
16            command.append("--pre-release")
17        subprocess.run(command, check=True)
18        ret = "Successfully installed_
19        extension"
20    except subprocess.CalledProcessError as
21    e:
22        ret = f"Error installing extension:_{
23        {e}"
24    except Exception as e:
25        ret = f"Unexpected_error:_{e}"
26    return ret

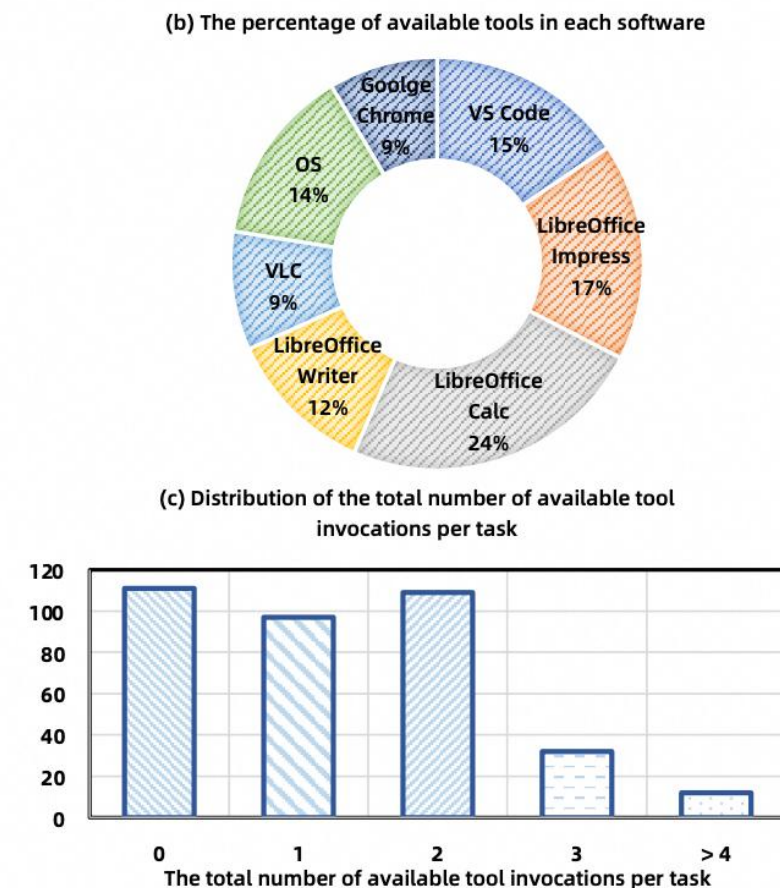
```

(b) MCP tool

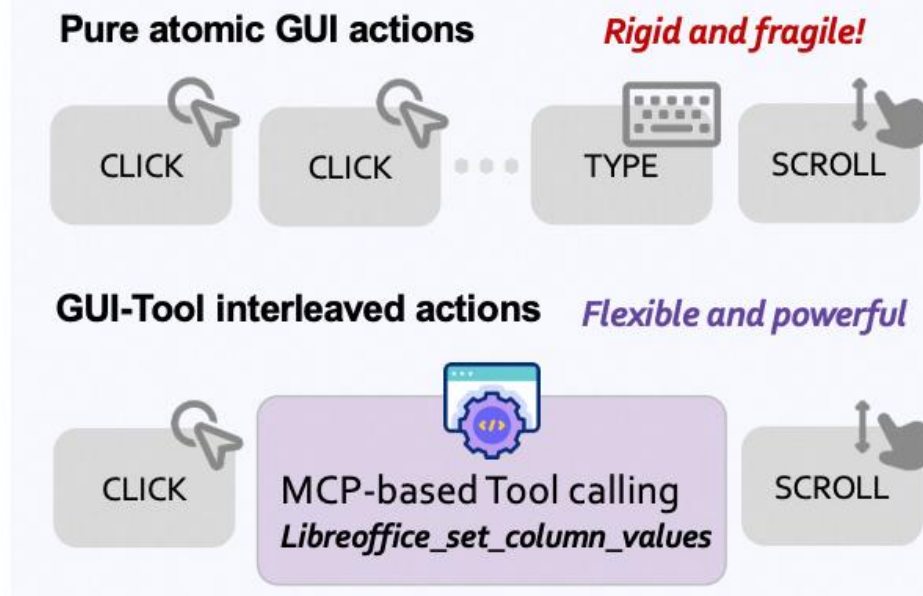
Figure 1: Task execution via GUI vs MCP Tool



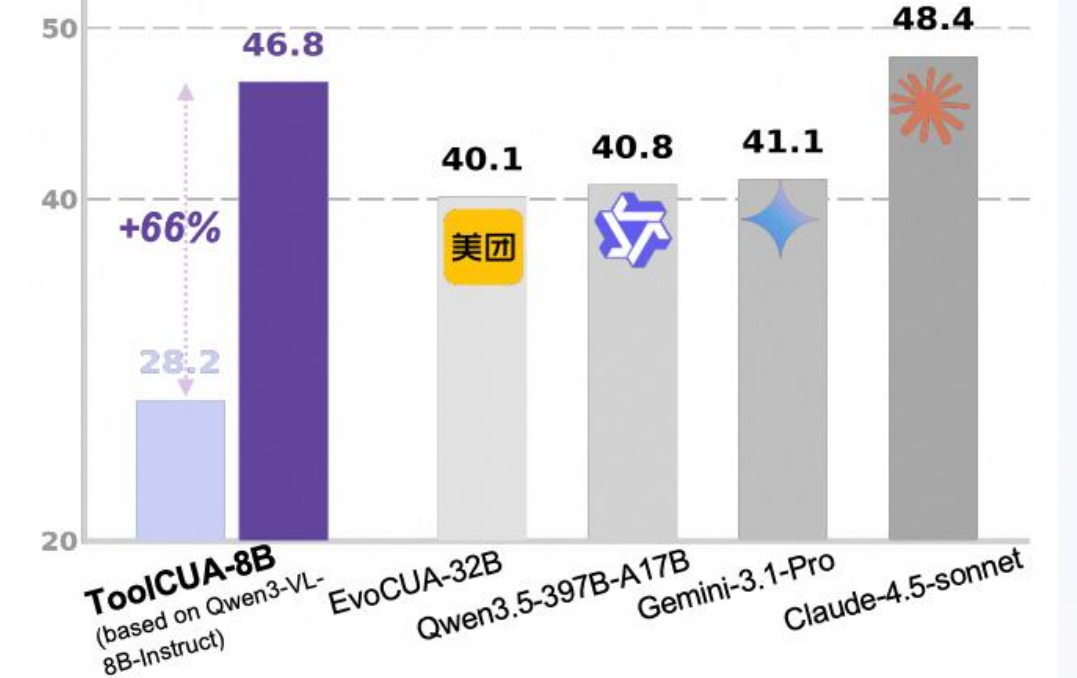
(a) Number of times each tool was invoked by the five models in the OSWorld-MCP evaluation



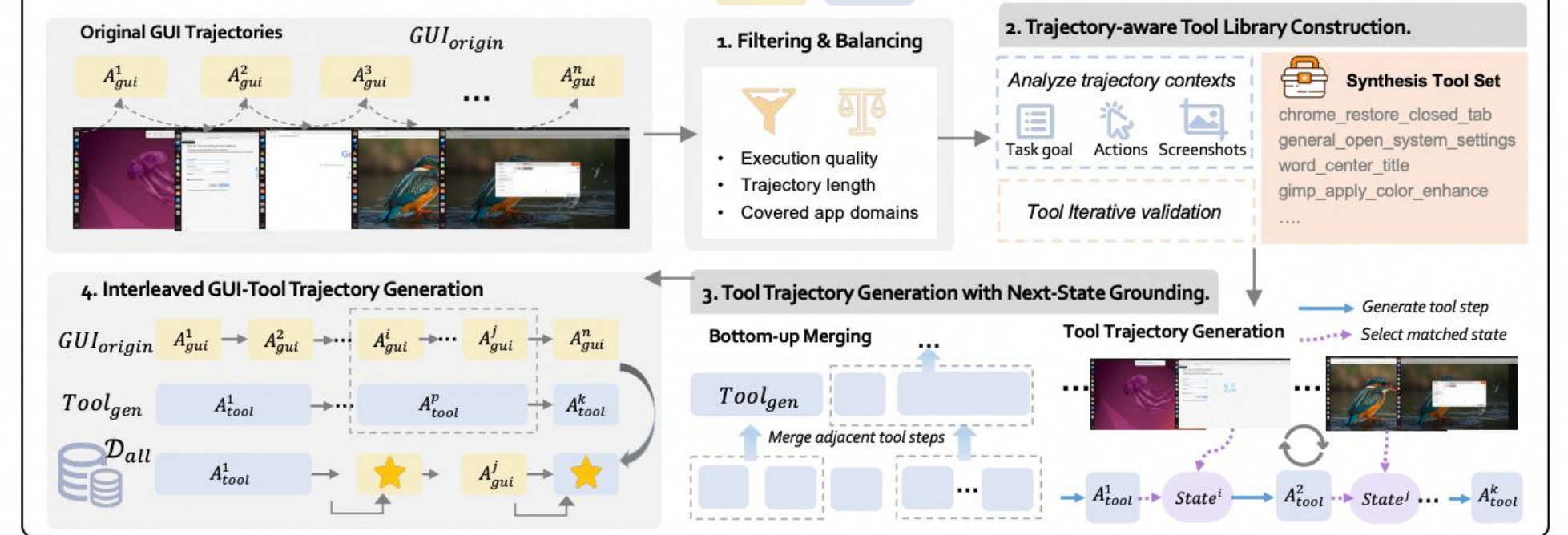
(a) Tool augmented GUI action for Execution Efficiency



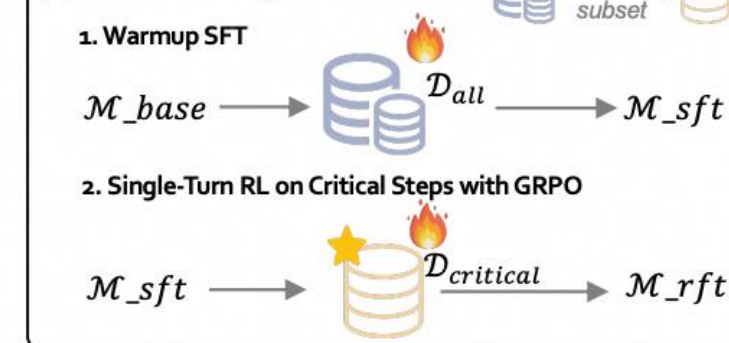
(b) ToolCUA Performance on OSWorld-MCP (feasible)



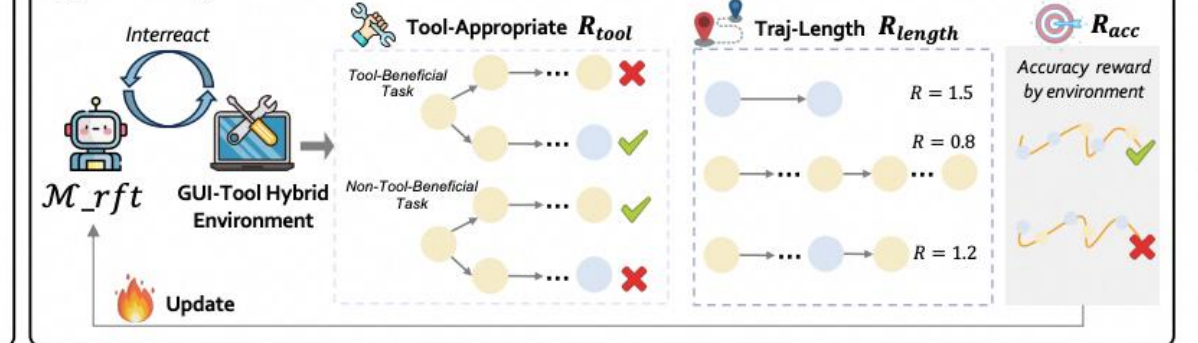
(a) Scaling Interleaved GUI-Tool Trajectories via Tool Synthesis



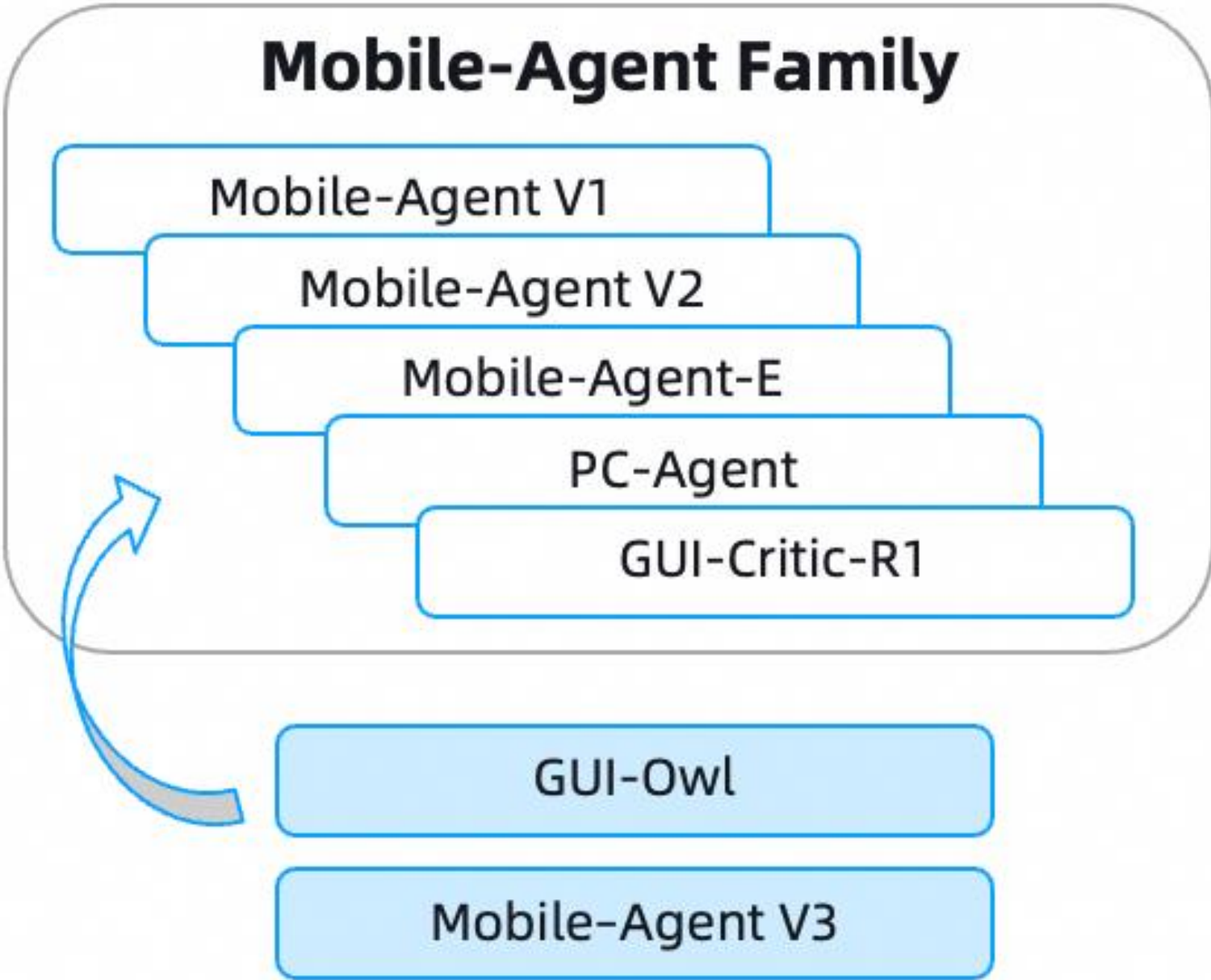
(b) Tool-Bootstrapped GUI RFT



(c) Online Agentic RL in GUI-MCP Environment



Mobile-Agent系列开源应用



体验Demo在魔搭、百炼上线并项目仓库开源...

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



Qwen3.5、3.6、3.7

GUI 智能体

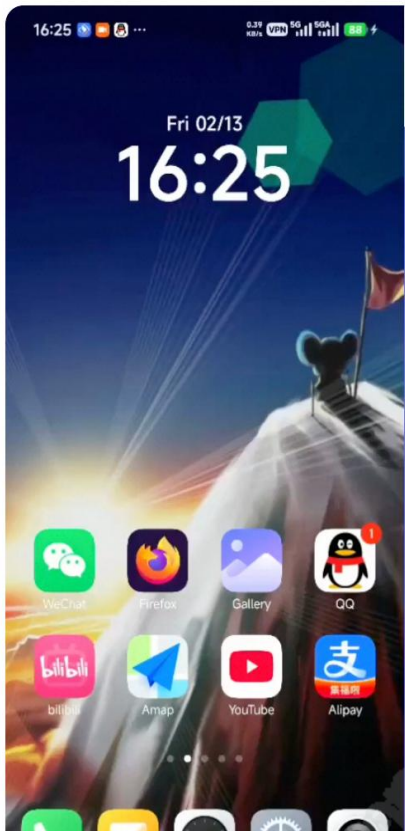
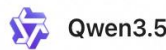
Qwen3.5 可作为视觉智能体，自主操作手机与电脑完成日常任务。在移动端，它已适配更多主流应用，支持自然语言指令驱动操作；在 PC 端，能处理跨应用的数据整理、多步骤流程自动化等复杂任务，有效减少重复性人工干预，提升工作效率。

展开所有演示

演示4 Searching, Liking, and Commenting

4 / 5

Search for qwen3vl on YouTube, then like the video, save it to watch later, and comment _it is helpful



Search for "qwen3vl" on YouTube, then like the video, save it to watch later, and comment "it is helpful".

等待执行指令...



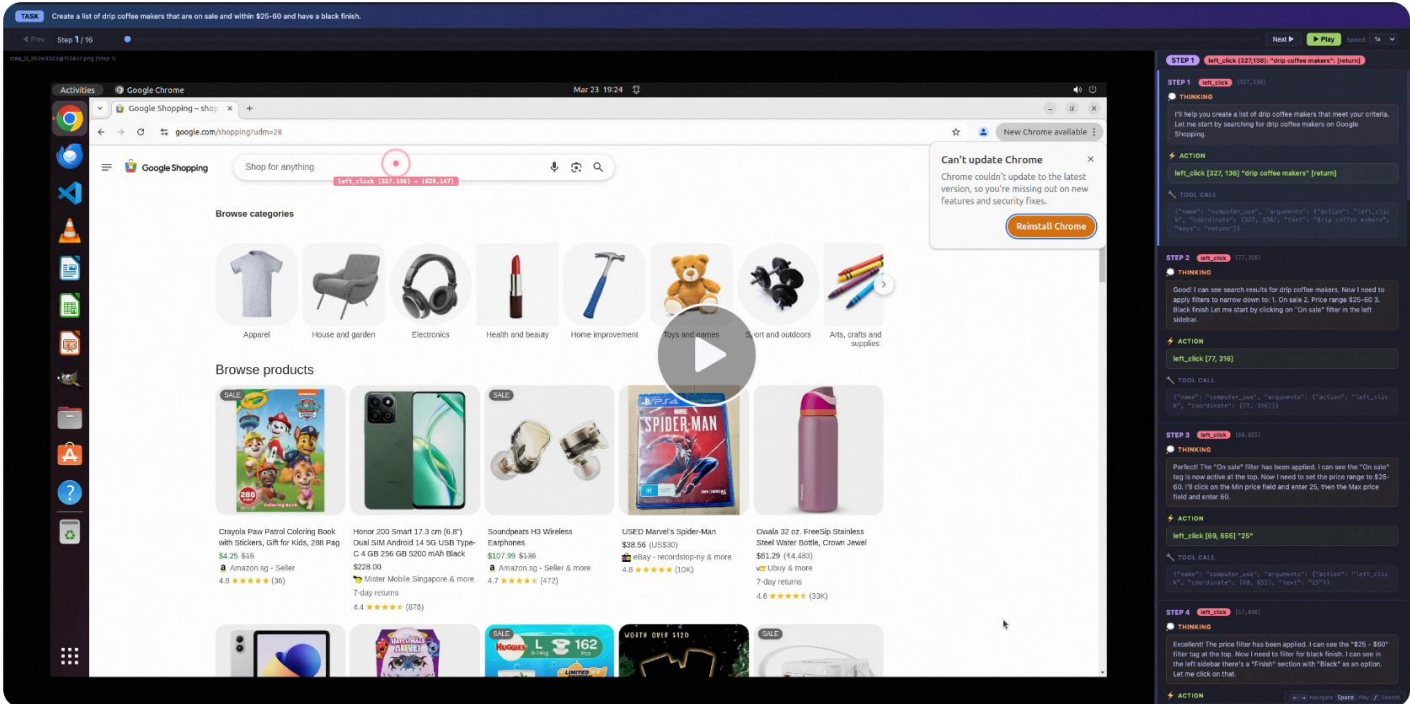
视觉Agent应用

我们关注的是模型如何在环境中持续感知、推理并采取行动。以 GUI Agent 为代表，模型可以基于屏幕内容理解当前界面状态，并结合规划能力执行下一步操作；而像 OpenClaw 这样的探索，则进一步展示了多模态模型在开放环境中完成复杂交互任务的潜力。结合 Claude Code 风格的工作流、多跳搜索、CI 与外部工具调用，模型能够逐步从单轮问答演进为面向真实任务的执行系统：先理解问题，再检索信息、生成方案、调用工具，并在反馈中持续迭代。

演示1 Computer-Using Agent

展开所有演示

Create a list of drip coffee makers that are on sale and within \$25-60 and have a black finish.



基于 Qwen3.7-Plus 构建的 Hybrid-Agent 智能体系统，将大模型的代码生成能力与 GUI 自动化执行深度融合，实现了从需求分析到版本迭代的 APP 全链路开发。Agent持续稳定运行 11+ 小时，全程自动完成了一款英语单词学习 APP 的完整研发闭环。累计生成代码超过 10,000+ 行，触发 Agent 调用超过 1,000+ 次，覆盖软件开发全生命周期的核心环节：需求文档生成、代码自动编写、自动化安装部署、测试用例创建、GUI 自动化测试、多场景并行化测试、产品说明自动更新、自动版本迭代演进。

Hybrid-Agent

APP代码开发 & 图形化界面自动化测试
App Development & GUI Automation Testing

演进迭代
Continuous Iteration

浏览器智能助手

我们基于 Qwen3.7-Plus 构建了浏览器智能助手，并通过 Qwen for Chrome 浏览器插件完成任务演示与录制。Qwen for Chrome 是一款嵌入 Chrome 浏览器的智能助手，用户可以在浏览器侧边栏中直接与 Qwen 对话，并在授权后切换至 Agent 模式。在该模式下，Qwen 能够感知当前网页内容、理解用户任务、规划操作步骤，并以 Browser Agent 的形式在真实浏览器环境中执行点击、输入、跳转、配置和验证等操作。

在此基础上，Qwen3.7 浏览器 Agent 将大模型的页面理解、任务规划与 GUI 自动化执行能力深度融合。面对非科班用户“采购一台最便宜 ECS 服务器”的需求，Agent 能够直接进入云控制台，完成实例规格比价、低成本选型、镜像与存储配置、安全组设置、订单确认等完整操作，并在价格变化、库存限制或购买受阻时主动反思和调整策略。随后，Agent 继续承接实例扩容与运维升级任务，自动完成停机、配置调整、磁盘扩容、服务恢复与结果验证，覆盖云服务器从采购到升级的真实使用链路。原本需要用户理解复杂控制台逻辑、反复切换页面并手动排查问题的流程，如今可以被 Agent 转化为连续、高效、可交付的浏览器自动化任务。

Computer-Use Agent

Qwen for Chrome

大模型通用智能体-未来展望

技术角度

- Agentic RL Scaling, 提升自主推理和知识进化;
- CLI+GUI相结合桌面端Agent, 自主解决问题、完成生产力任务;
- 个性化交互与记忆;

Agent!



Agentic RL
Scaling



个性化交互



CLI+GUI
桌面端Agent

谢谢