

QCon
全球软件开发大会

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

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

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

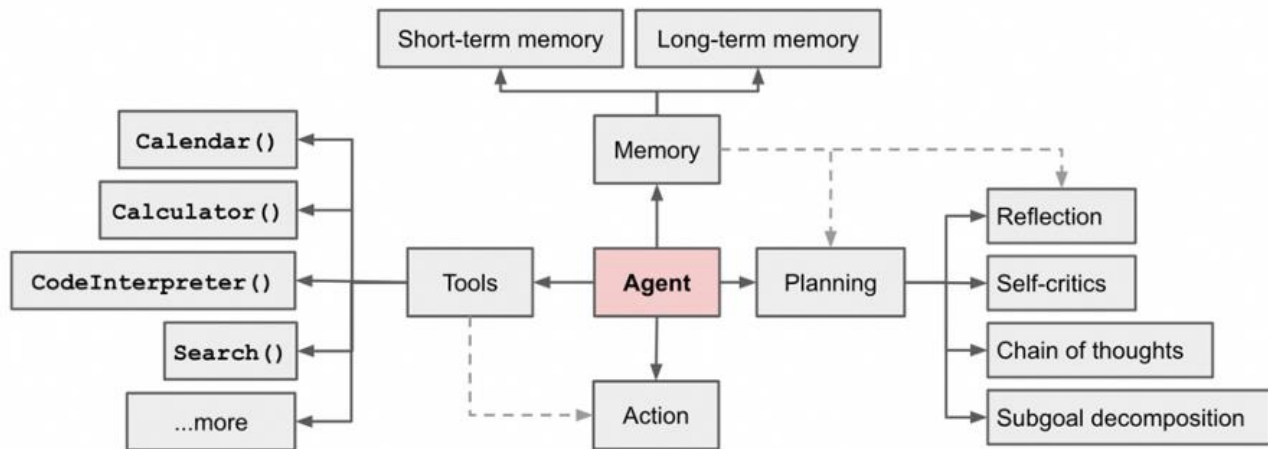
01

大模型智能体背景

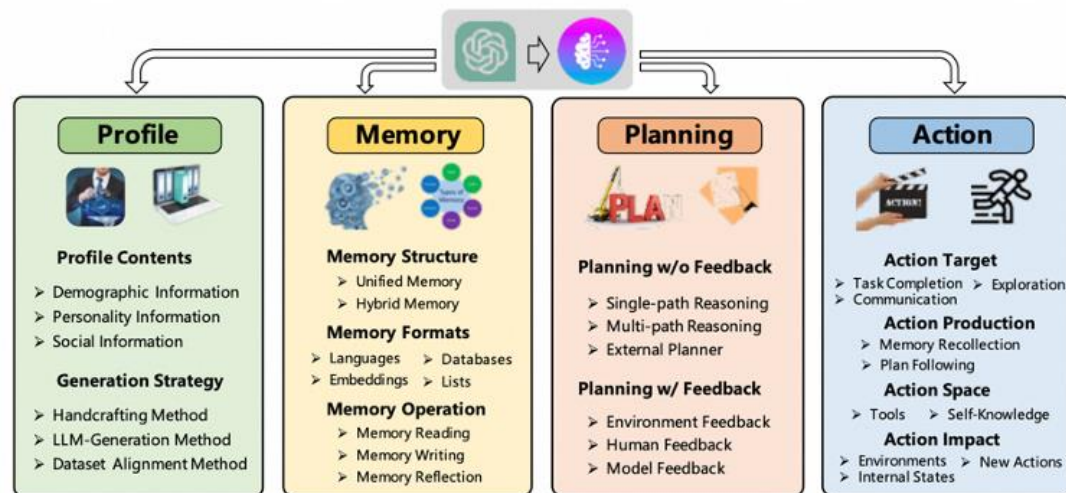
大模型智能体系统

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

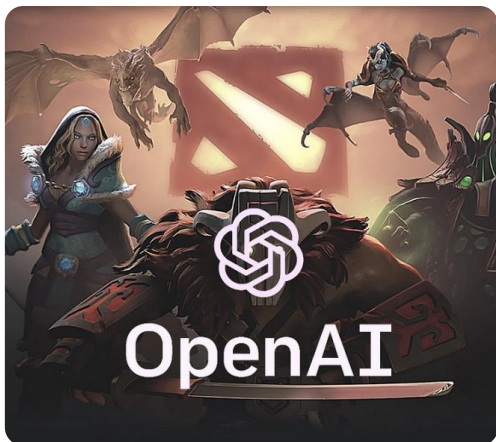
Agent System Overview from Lilian Weng's blog



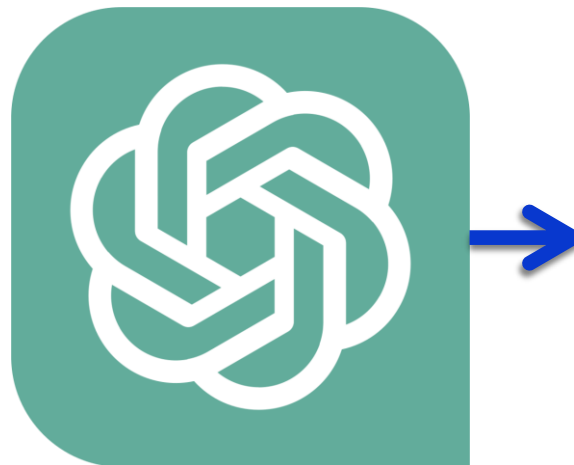
Wang et al. A Survey on Large Language Model based Autonomous Agents



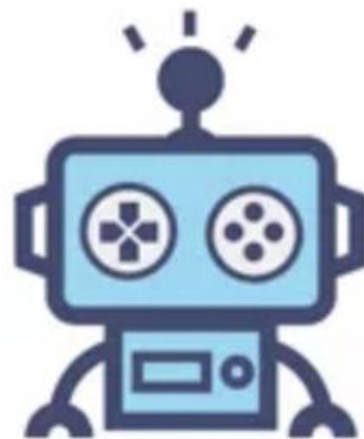
大模型智能体的优势



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

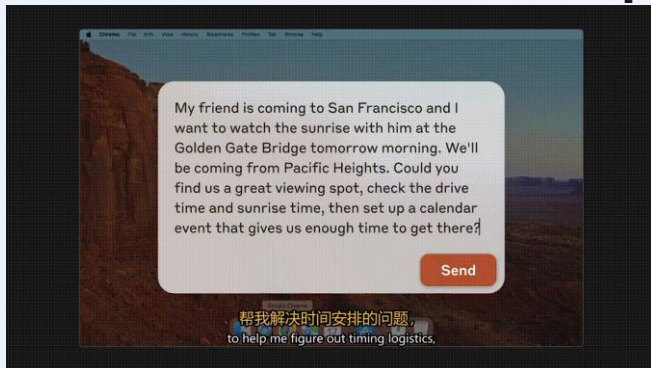
近期两类AI Agent应用

	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。

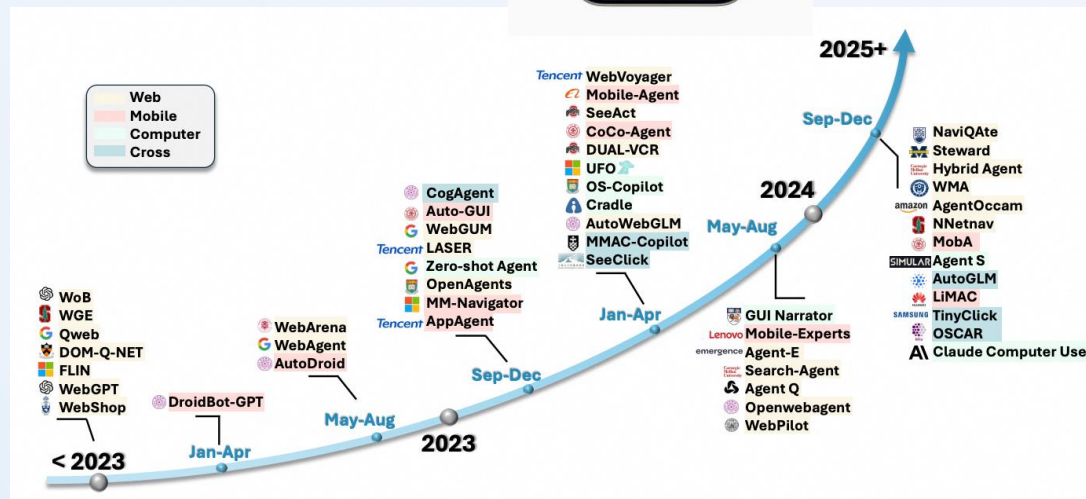
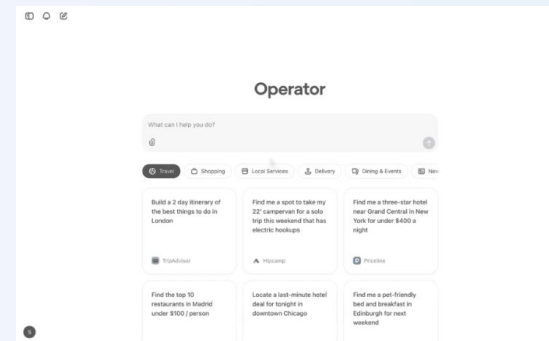
Claude 3.5 Sonnet (Computer)



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



Introduction

Recent advancements in large language models (LLMs) have significantly enhanced text generation capabilities, yet evaluating their performance in generative writing remains a challenge. Existing benchmarks primarily focus on generic text generation or are limited in writing tasks, failing to capture the diverse requirements of high-quality written content across various domains.

WritingBench addresses this gap by providing a comprehensive evaluation framework designed specifically for assessing writing capabilities across multiple dimensions, including style, format, and length requirements.

Key Contributions

- ① A comprehensive benchmark with 1,239 queries across 6 primary domains and 100 subdomains
- ② A query-dependent evaluation framework with instance-specific criteria generation
- ③ A fine-tuned critic model for criteria-aware scoring with 83% human alignment
- ④ Demonstration of framework effectiveness through data curation and model enhancement

Example Query from WritingBench

I need to write a research proposal focused on quantum computing applications in financial risk modeling. The proposal should follow the IEEE conference template and be approximately 2,500 words. Use a formal academic tone with precise technical language.

- Include the following sections:
- Abstract (500 words)
 - Introduction with problem statement
 - Literature Review
 - Methodology
 - Expected Results
 - Timeline
 - Budget Justification
 - References (using IEEE citation style)

Please reference these recent papers on quantum algorithms for financial modeling...

Content specificity | Formal requirement | Length constraint
Style requirement | Material reference

Each WritingBench query is designed with multiple requirements and evaluation dimensions, simulating real-world writing scenarios.

Benchmark Comparison

How WritingBench compares to existing writing evaluation benchmarks

BENCHMARK	QUERIES	PRIMARY DOMAINS	SECONDARY DOMAINS	STYLE REQ.	FORMAT REQ.	LENGTH REQ.	Avg. INPUT TOKENS
EQ-Bench	241	1	—	✗	✗	✗	130
LongBench-Write	120	7	—	✗	✗	✓	87
HelloBench	647	5	38	✗	✗	✓	1,210
WritingBench	1,239	6	100	✓	✓	✓	1,546

Comprehensive Coverage

WritingBench covers 6 primary domains and 100 secondary subdomains, ensuring a thorough evaluation of writing capabilities across diverse contexts.

Multi-Dimensional Requirements

Unique among benchmarks, WritingBench evaluates style, format, and length requirements, addressing the complexity of real-world writing tasks.

Extended Input Context

With an average input of 1,546 tokens and maximum of 19,361 tokens, WritingBench tests models' ability to handle complex, lengthy contexts.

Domain Coverage

WritingBench covers 6 primary domains and 100 secondary subdomains

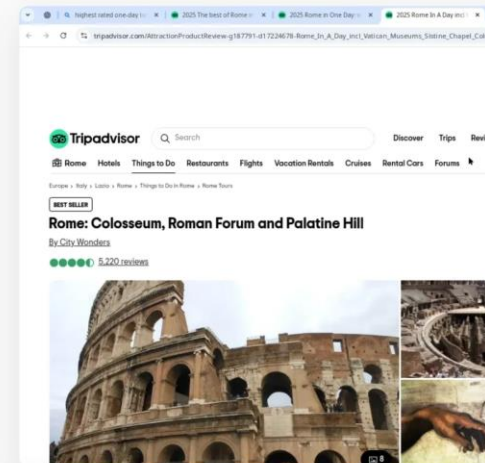


Find and book me the highest rated one-day tour of Rome on Tripadvisor.

I'll search for the highest-rated tour of historic Rome on TripAdvisor. Once I find a suitable option, I'll provide you with the details. Let's begin.

Worked for 2 minutes ^

Navigating to TripAdvisor website
Selecting "Things to Do" category
Searching for historic Rome tours
Closing pop-up, continuing tour search
Searching for Historic Rome tours
Exploring all historic Rome tour options
Closing Colosseum tab, resuming tour search
Closing tour pop-up, tab afterward
Exploring options for top-rated tours
Sorting results by tour ratings
Exploring filters for top-rated tours
Scrolling for sorting options, finding tours



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

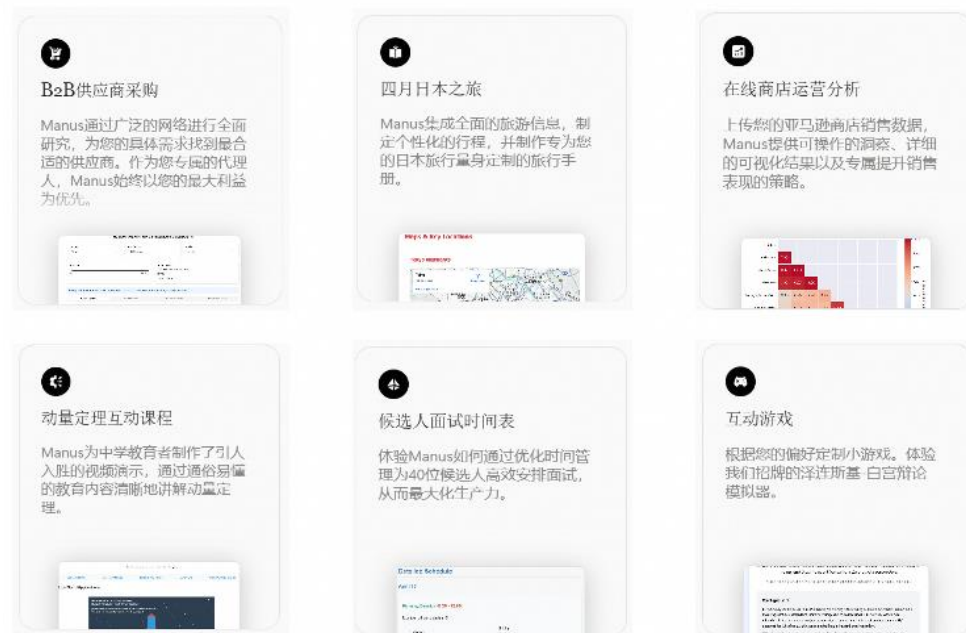
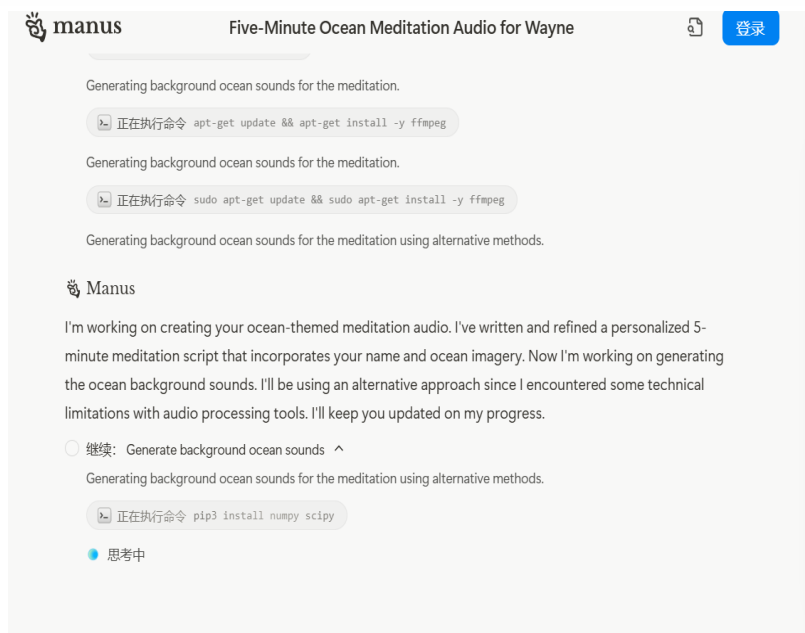
基于Computer-Using Agent 模型，结合GPT-4o的视觉理解能力和强化学习习得的推理能力，自动执行鼠标和键盘的组合操作，无需API，具备推理思维链和自动纠错能力



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

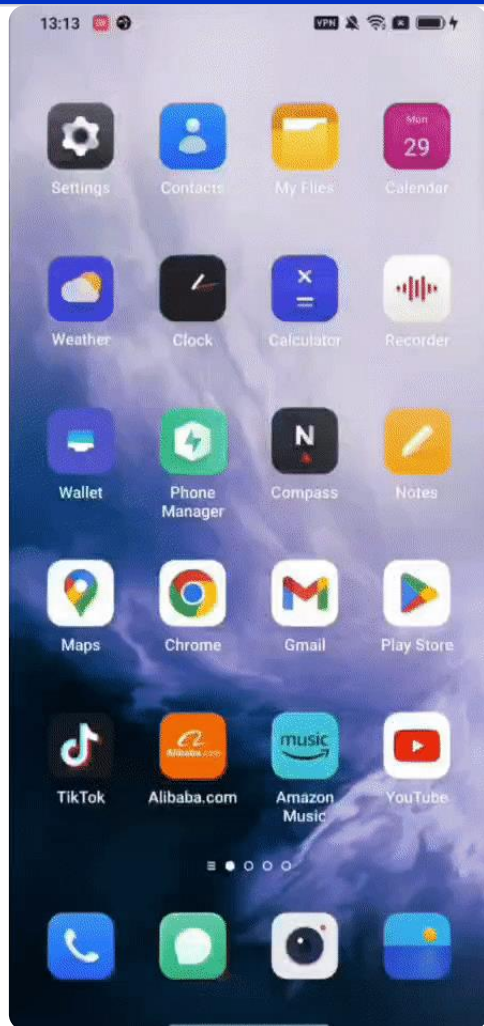
Manus强调“需求→规划→执行→交付”全流程自动化，无需用户持续指导便可能直接生成可交付成果，动态调整执行路径，在解决现实世界问题方面表现卓越

02

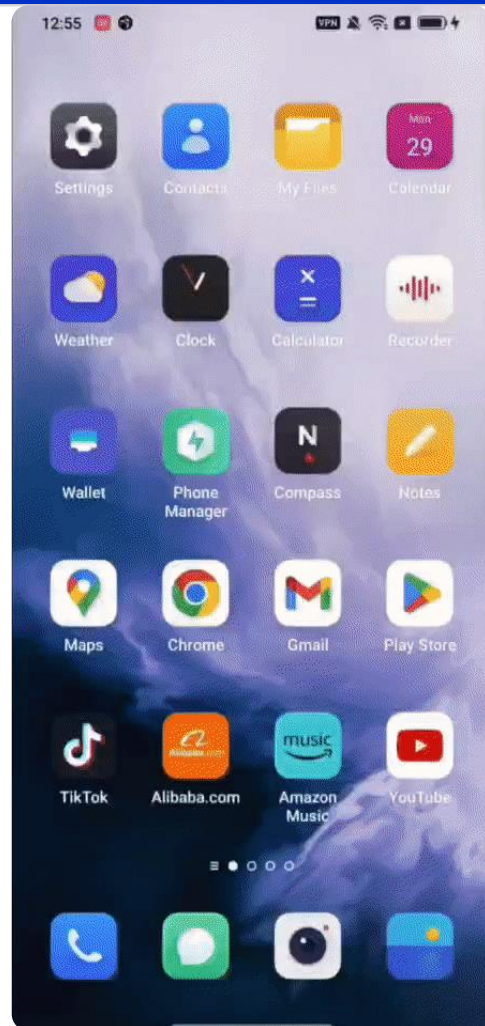
多模态多端智能体 Mobile-Agent

多模态多端智能体Mobile-Agent

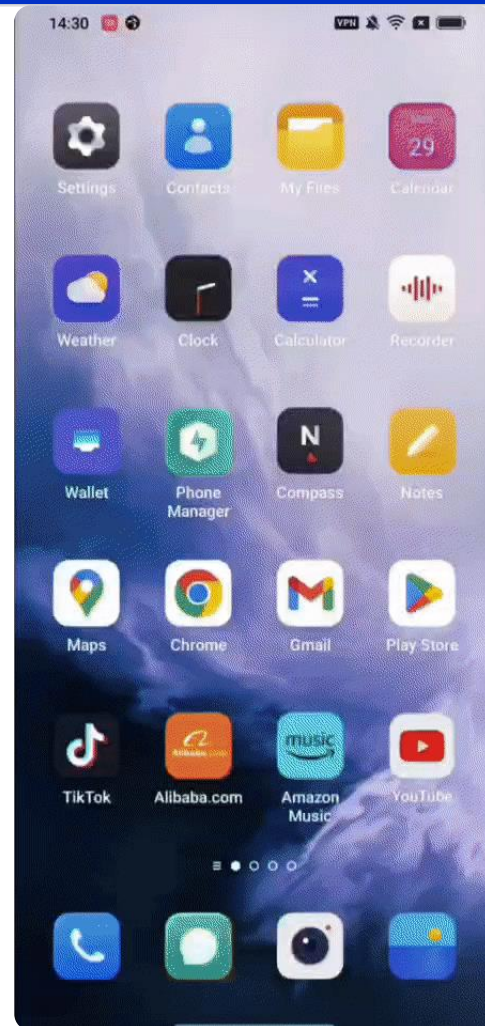
分析天气



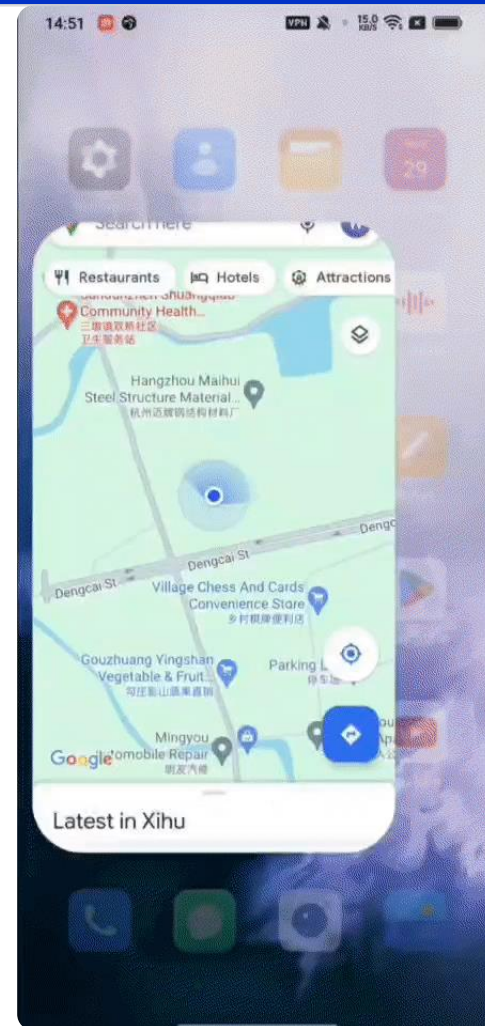
搜索视频并评论



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

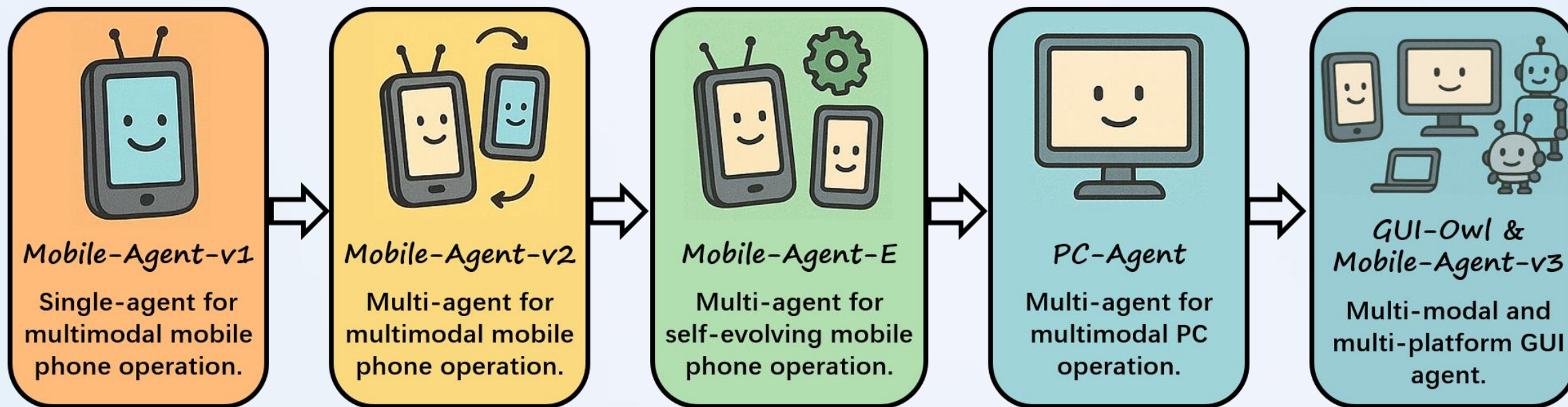
PC Agent

Institute for Intelligent Computing of Alibaba Group

<https://arxiv.org/abs/2502.14282>

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

UI界面理解

多步操作
规划反思



定位操作

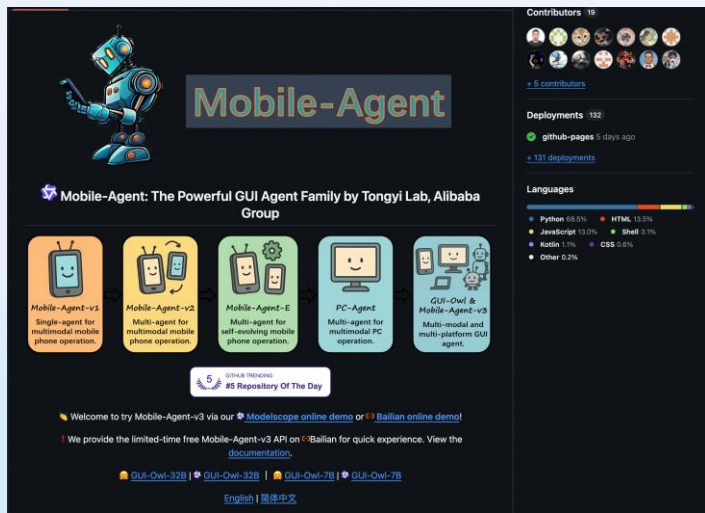
操作时延

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



多模态多端智能体Mobile-Agent-V1

行为空间

Visual Perception



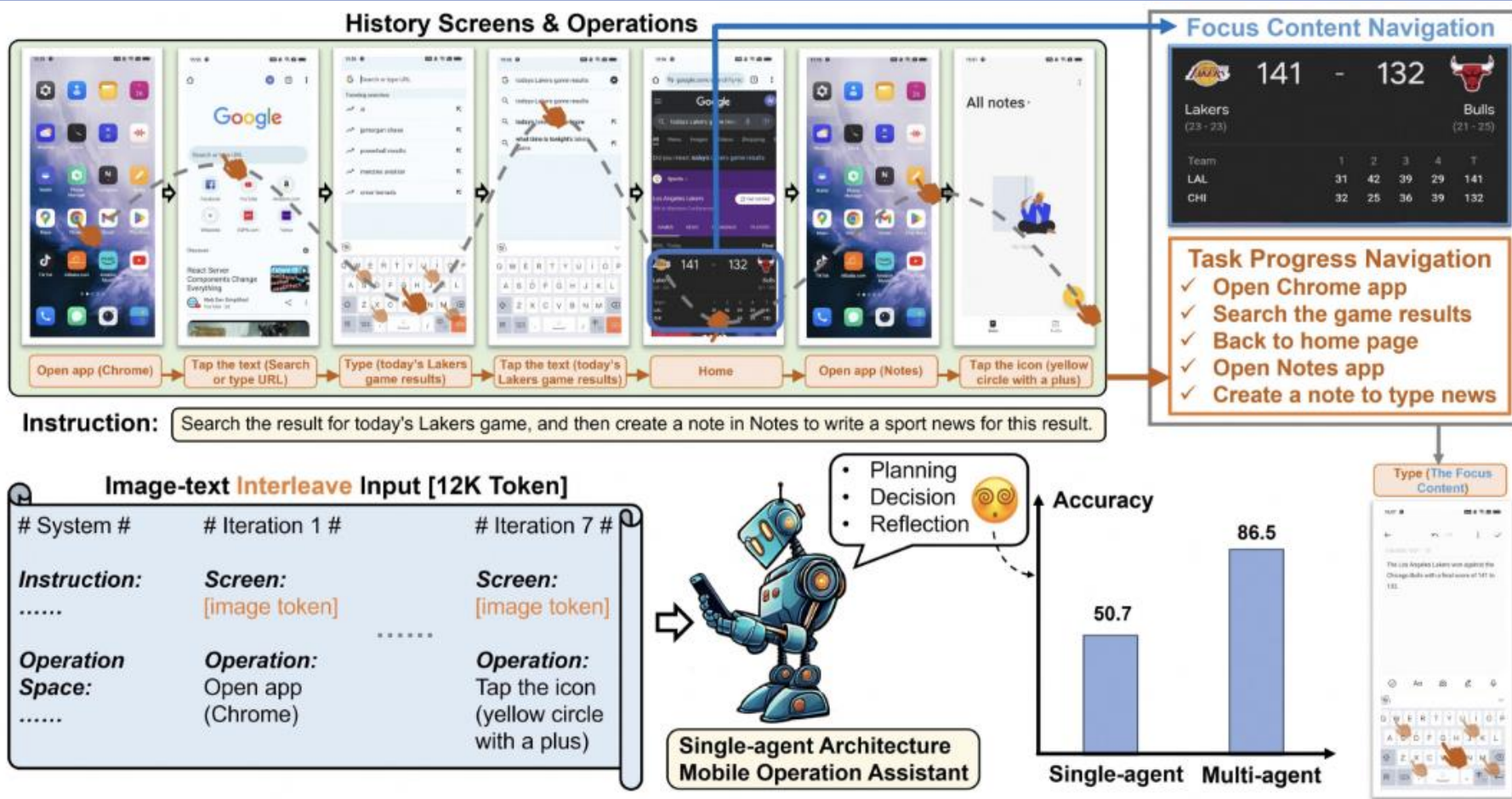
大模型缺乏输出精确坐标的grounding能力

- 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

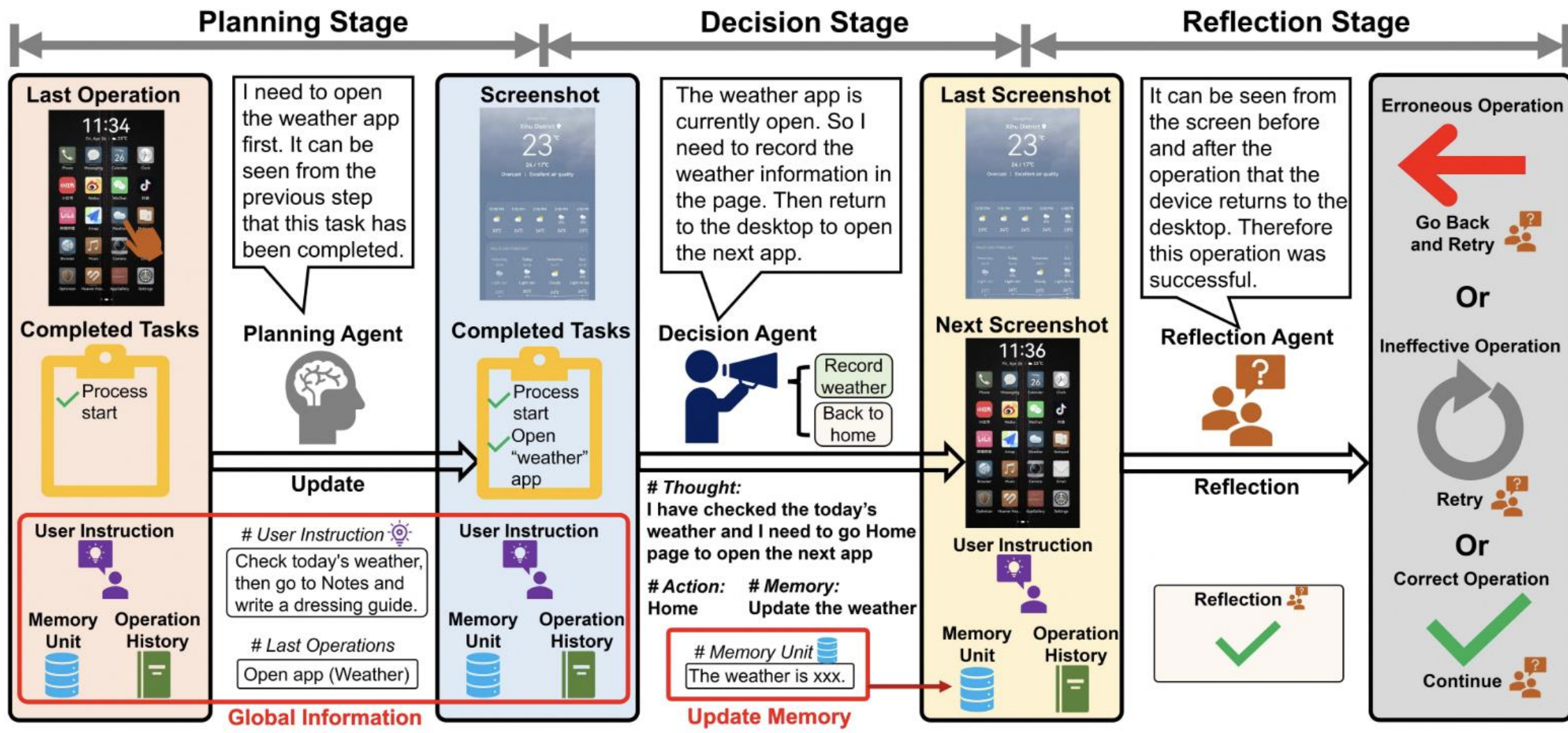
1. 点击文本
2. 点击图标
3. 打字
4. 上划 & 下划
5. 返回上一页面
6. 返回桌面
7. 结束

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
	<i>System app</i>							
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
	<i>External app</i>							
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
	<i>Multi-app</i>							
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.

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.

Method	Basic Instruction				Advanced Instruction			
	SR	CR	DA	RA	SR	CR	DA	RA
	<i>System app</i>							
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
	<i>External app</i>							
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
	<i>Multi-app</i>							
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.

Mobile-Agent-E: 解决复杂任务、自主进化



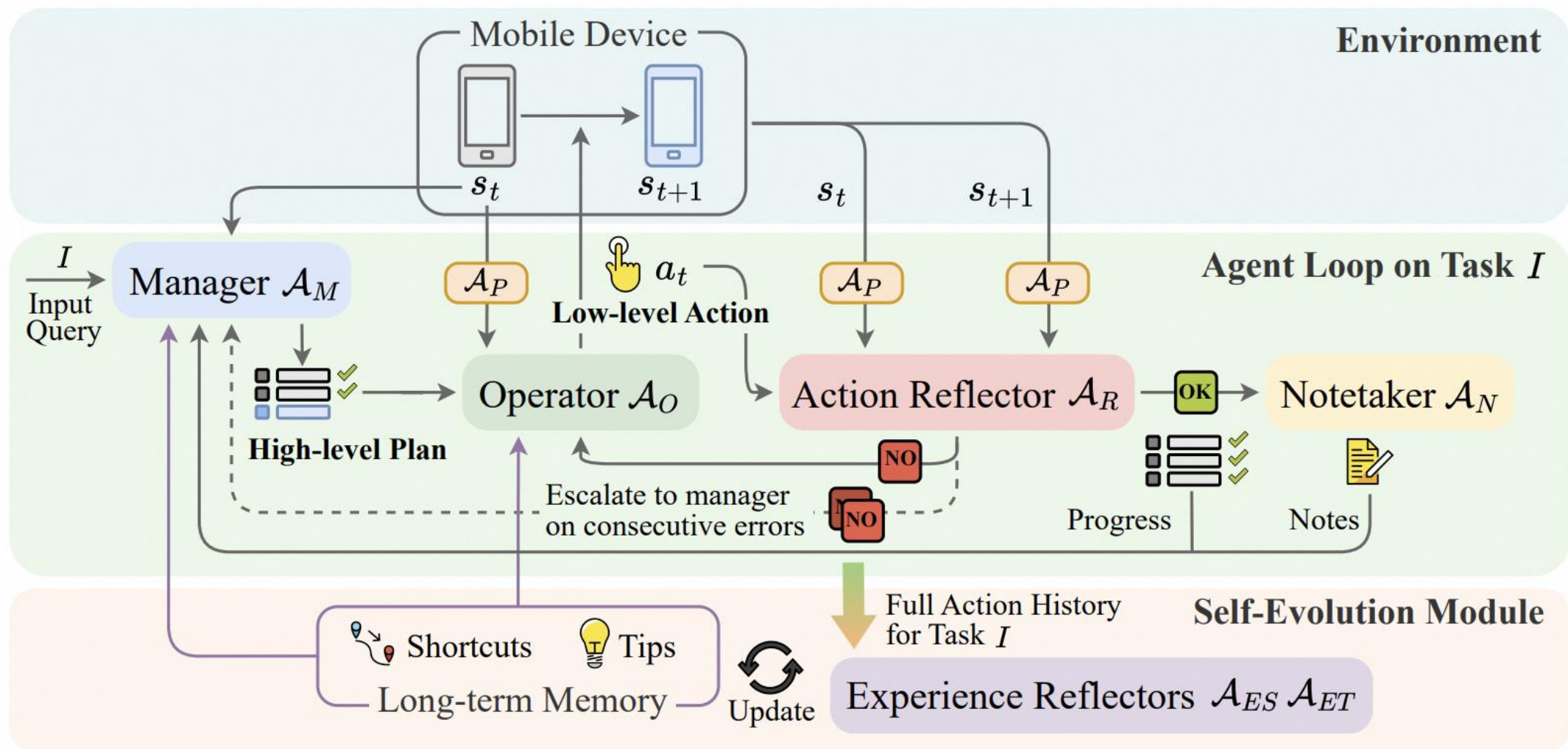
复杂指令:

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

03

Foundation Agent for GUI

User: What are these products? Please give their names in Chinese and English.

Swipe up

Click (Setting APP)

Click (connected devices)

Click (connection preference)

Click (Bluetooth)

Click (Button)

Finish

Output of the YOLOv2 model for the same image. The bounding boxes are labeled with the predicted class and confidence score. The labels are: "cyclist (not wearing helmet)" (blue), "motorcyclist (not)" (yellow), and "cyclist (not wearing helmet)" (red).

```

{"class": "cyclist (not wearing helmet)", "x1": 150, "y1": 250, "x2": 200, "y2": 350, "score": 0.95},
{"class": "motorcyclist (not)", "x1": 150, "y1": 450, "x2": 200, "y2": 550, "score": 0.95},
{"class": "cyclist (not wearing helmet)", "x1": 350, "y1": 450, "x2": 400, "y2": 550, "score": 0.95}

```

أبو منير
 لبيع وصيانة الروبوتات
 روبوتات ماء مكيف، دهانات

COOLING CAR SYSTEM

أبو منير
 052-204-5334
 050-831-0796
 محمد أبو منير
 016-851-8236

Qwen2.5-VL **COOLING** أبو منير لبيع وصيانة التبريد والتكييف - مكيف - نظافات **SMK**
 أبو منير 052-204-5334 محمد أبو سراج 059-831-0796
 056-811-8256


 arXiv

2020-06-16

Qwen2.5 Technical Report

Qwen Team

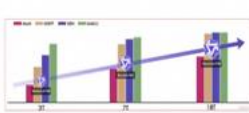


<https://github.com/QwenLM>
<https://arxiv.org/abs/2503.19152>
<https://qwenlm.github.io/>

Abstract

In this report, we introduce Qwen2.5, a comprehensive series of large language models (LLMs) designed to meet diverse needs for pre-training to post-training scenarios. Qwen 2.5 has been significantly improved during both the pre-training and post-training stages. In terms of pre-training, we have scaled the high-quality pre-training dataset from the previous 7 trillion tokens to 18 trillion tokens. This provides a better foundation for improving our expert knowledge, and reasoning capabilities. In terms of post-training, we employ iterative pre-training, adaptive distillation, and over 1 million samples, as well as multi-stage reinforcement learning, including off-policy training (OPX) and online training (OPXT) to further improve the model’s reasoning performance, and finally improve the generalization, structured data analysis, and instruction following.

Schedule density and cardinality are taken effectively by Qwen2.5 72B series in both pre-training and post-training. The open-weight offering in both 72B models and 72B series showed models of 0.8B, 1.8B, 7B, 72B, 72B, and 72B parameters. Quantized versions including 4-bit, 8-bit, 16-bit, and 32-bit are also available. Qwen2.5 72B series are available in both Hugging Face Hubs ModelScope and Kaggle. In addition, for better usability, we have also provided a locally accessible version of Qwen2.5 72B series. Qwen2.5 72B series are available from Alibaba Cloud Model Studio. Qwen2.5 is designed to help democratize top performance in a wide range of applications, from language understanding, reasoning, and code generation to text-to-image synthesis. In particular, the open-weight Qwen2.5 72B-series supports multi-modal input at open and proprietary models and demonstrates competitive performance in the state-of-the-art open-weight models. Qwen2.5 480B includes, which is around 1 times the number of Qwen2.5 72B series (Qwen2.5 72B series open-weight) with only training the capacity of open Qwen2.5 72B series. The state-of-the-art competitive results in the benchmarks, Qwen2.5 Instruct and Qwen2.5 Instruct-Chat, especially in the long-context, Qwen2.5 72B series, Math Qwen2.5, Qwen2.5 480B series (Qwen2.5 480B series, and multimodal models).



Model Size	Qwen 1.5 Pre-training (days)	Qwen 1.5 Post-training (days)	Qwen 2.5 Pre-training (days)	Qwen 2.5 Post-training (days)
0.5B	~10	~10	~10	~10
1.8B	~15	~15	~15	~15
7B	~25	~25	~25	~25
72B	~45	~45	~65	~65
480B	~55	~55	~85	~85

Figure 1: The training duration of the Qwen series, data scaling and pre-training cost. Qwen 2.5, 72B series, includes scaling for pre-training. Qwen2.5 480B series, includes scaling for pre-training series. Qwen2.5 series, especially in terms of data scaling, demonstrates the importance of scale in training large language models.

[illegible]



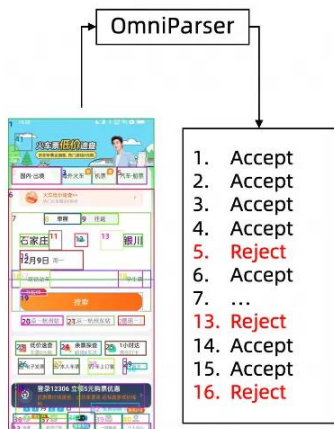
Qwen2.5-VL: Mobile-Agent能力提升



图片Hash去重



原始数据预处理



OmniParser - XML元素交叉清洗

Screenshot-Elements
自动标注集

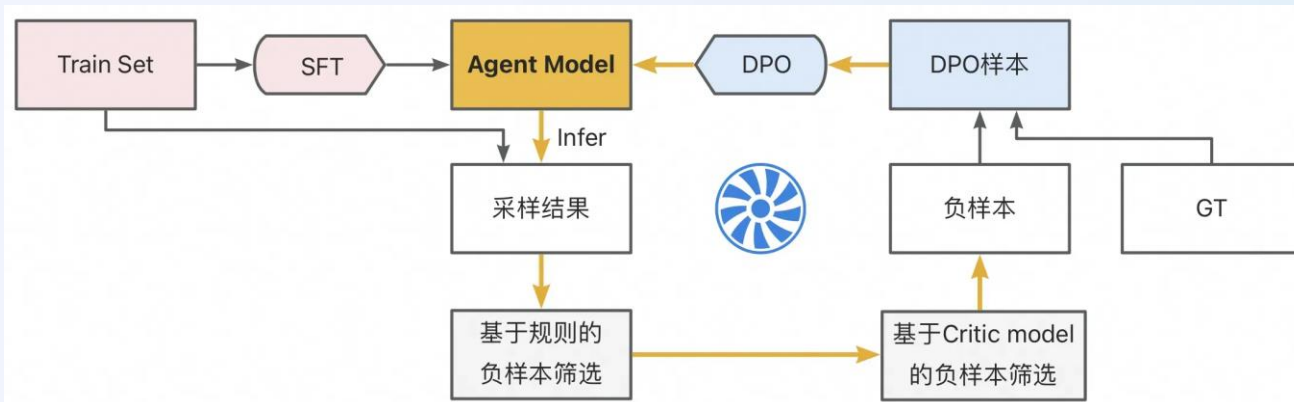
- GPT-4o/Claude 标注元素
功能性/外观性描述
- 人工二次校验

Grounding训练数据集



已产出146K图粒度标注数据

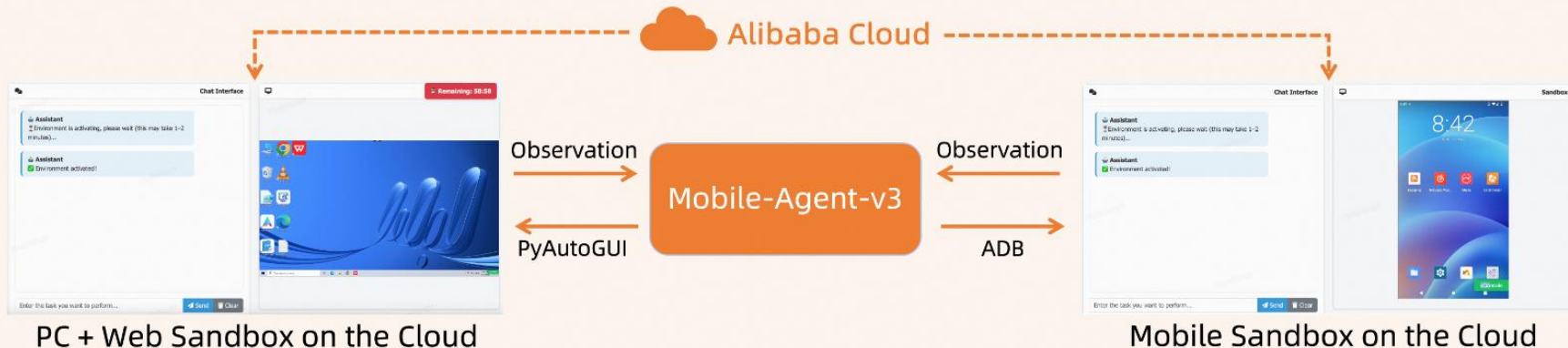
Grounding指令打标校验



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
AndroidWorld _{SR}	34.5% (SoM)	26% (SoM)	27.9%	26.1%	6% (SoM)	35%
MobileMiniWob++ _{SR}	61%	42% (SoM)	61% (SoM)	66%	50% (SoM)	68%
OSWorld	5.03	4.70	14.90	10.26	2.42	8.83

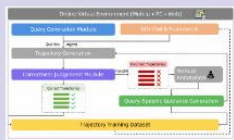
Mobile-Agent-V3 & GUI-Owl

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

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

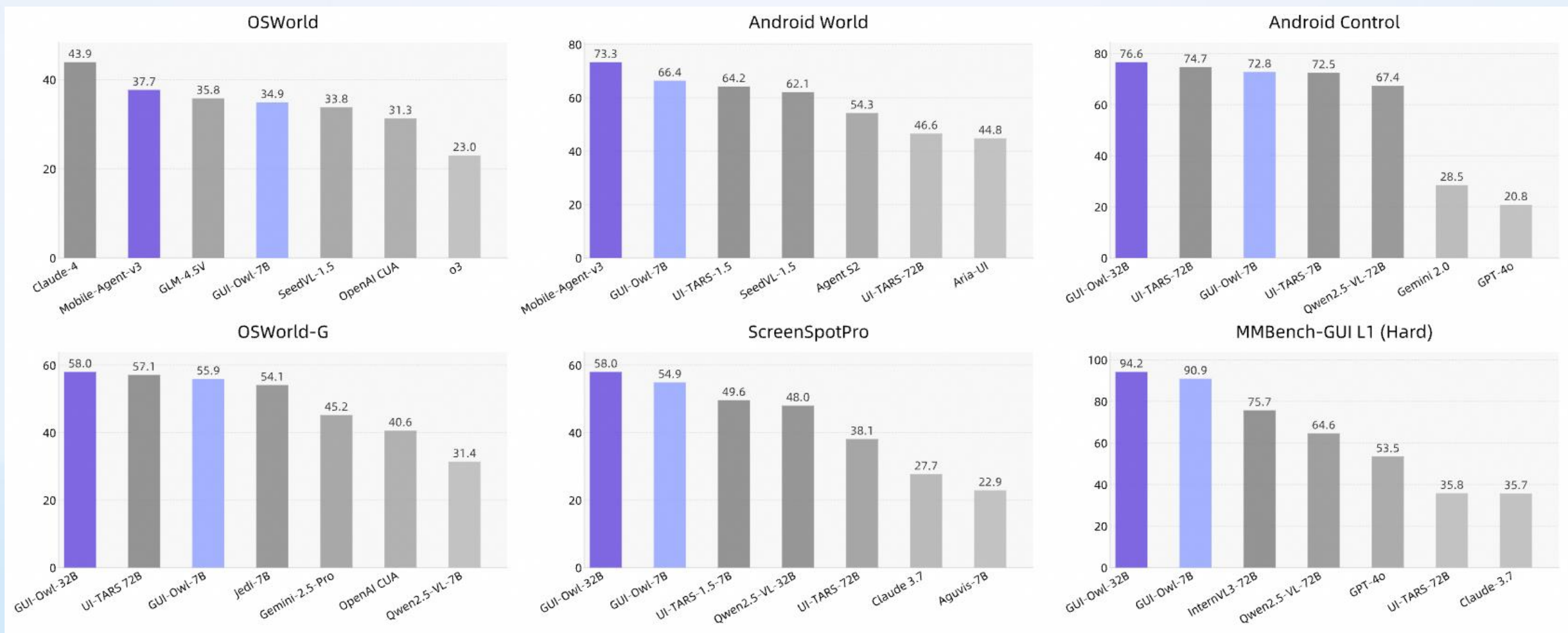
轨迹回流 ↓ 模型部署 ↑

PAI GPU集群

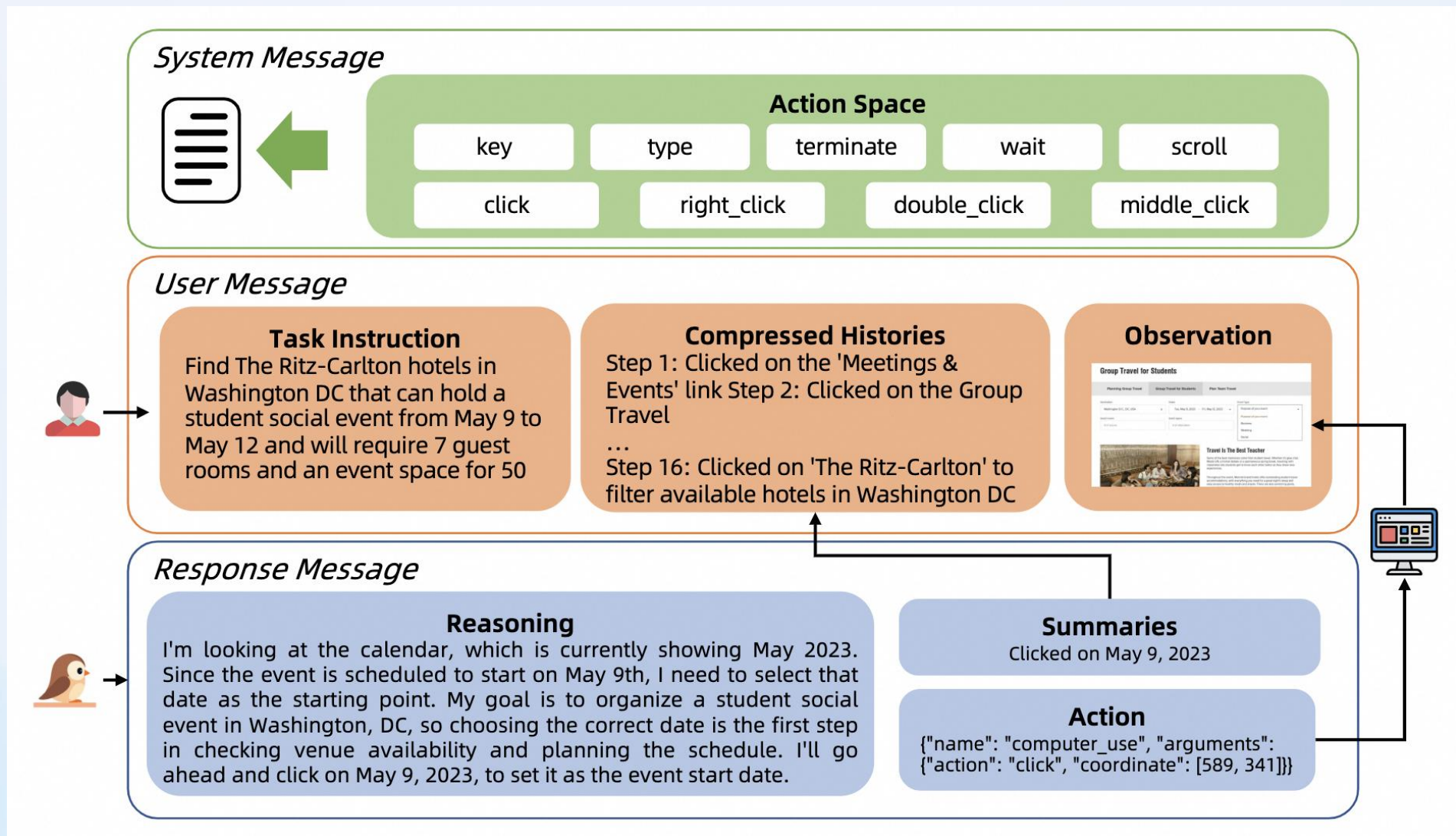
SFT/RL 模型训练

基建架构

Mobile-Agent-V3 & GUI-Owl



GUI-Owl整体交互flow





GUI-Owl数据合成链路-Grounding

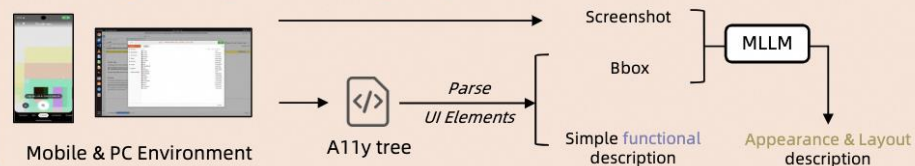
UI Element Grounding Generation (Function-based / Appearance- & Layout-based Instruction)

Open-source Grounding Dataset

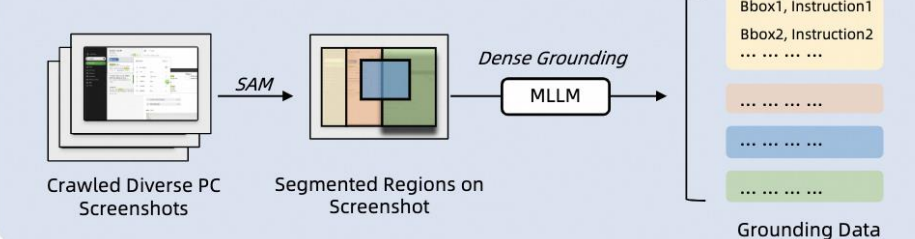


Mobile & PC & Web Platforms

Grounding Data Synthesis using A11y tree



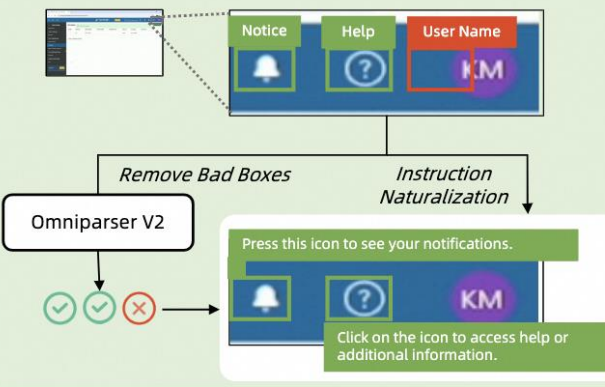
Dense Grounding Generation on Crawled PC Images



Grounding Data with Noise

Grounding Data Refinement

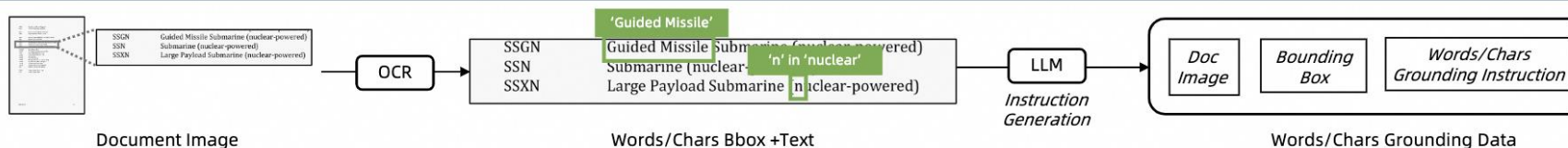
Grounding Data (Box + Instruction Pairs)



Cleaned Grounding Data

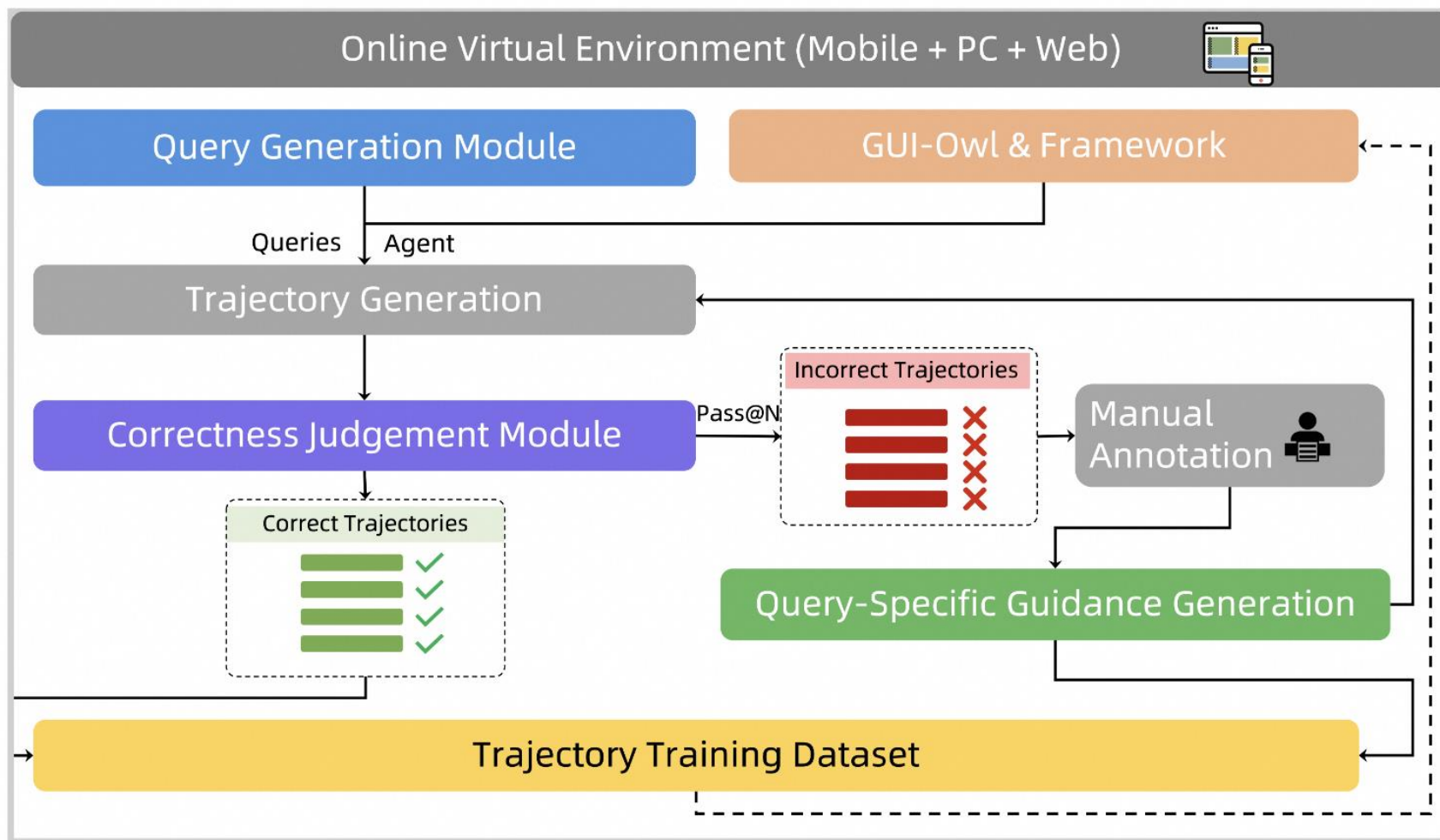


Fine-grained Words/Chars Grounding Generation

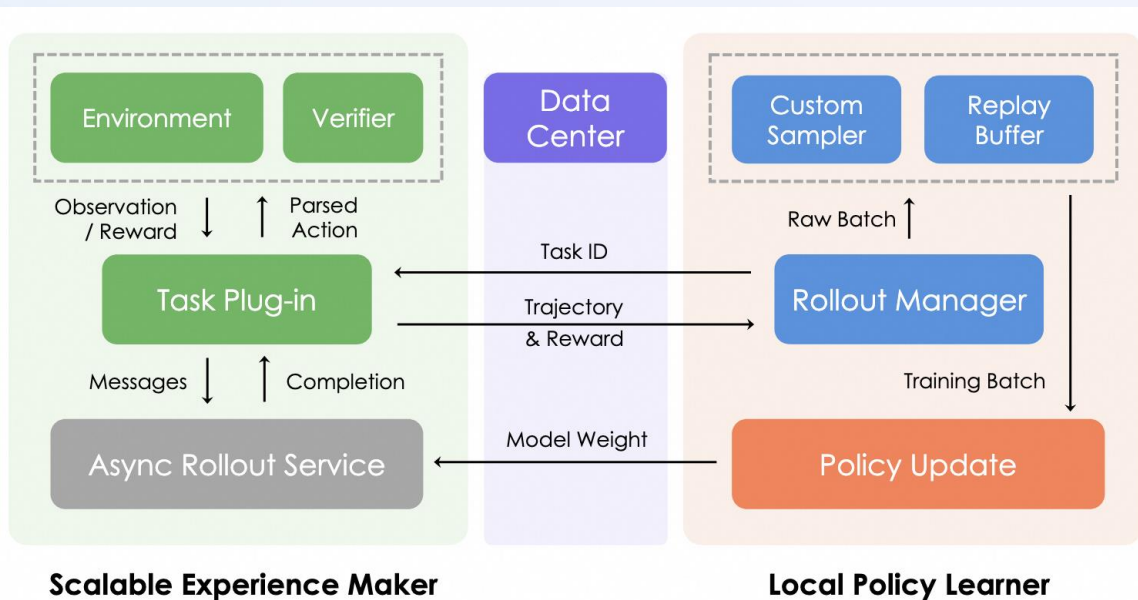




GUI-Owl数据合成链路-自进化轨迹合成

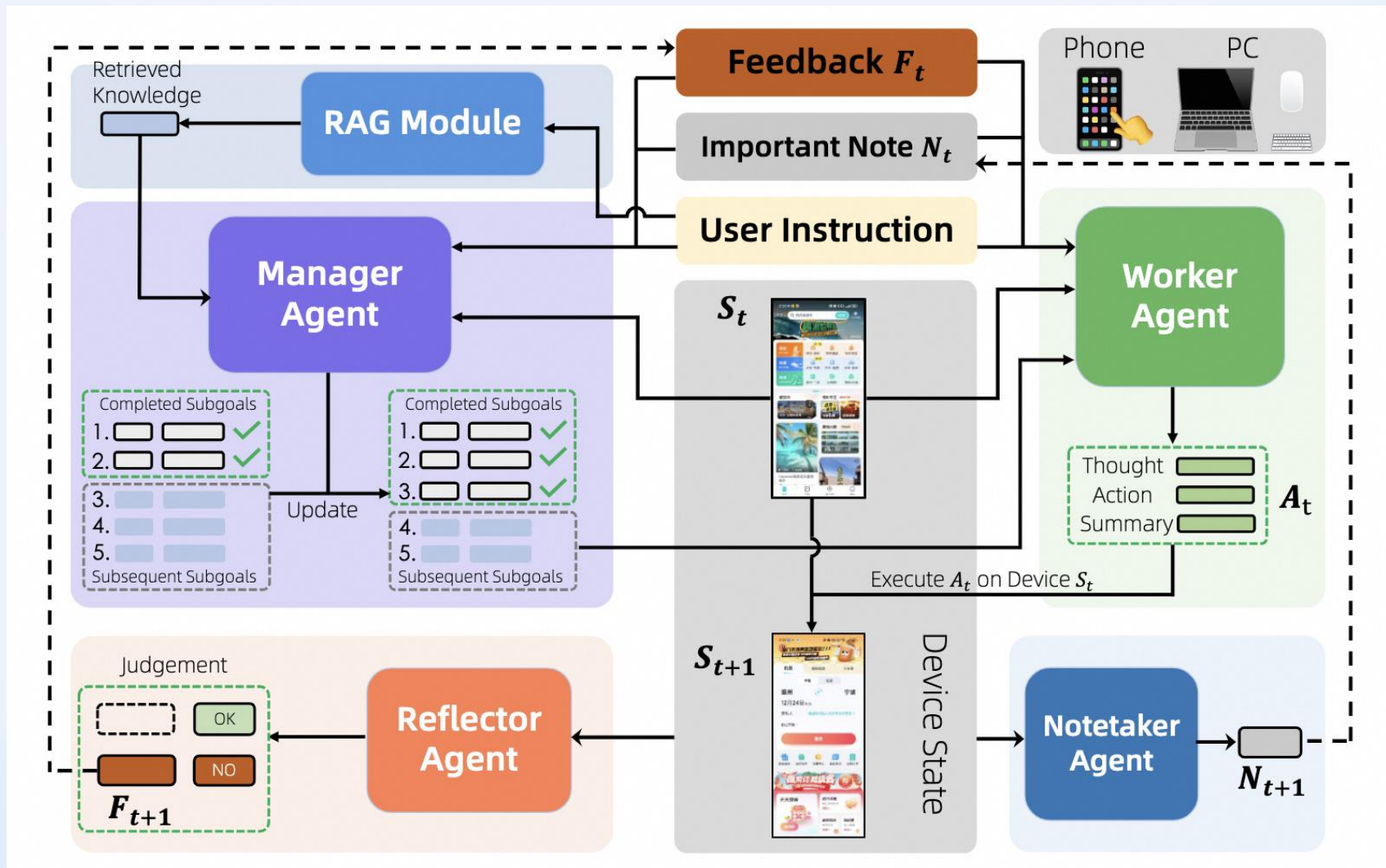


Mobile-Agent Agentic RL能力提升

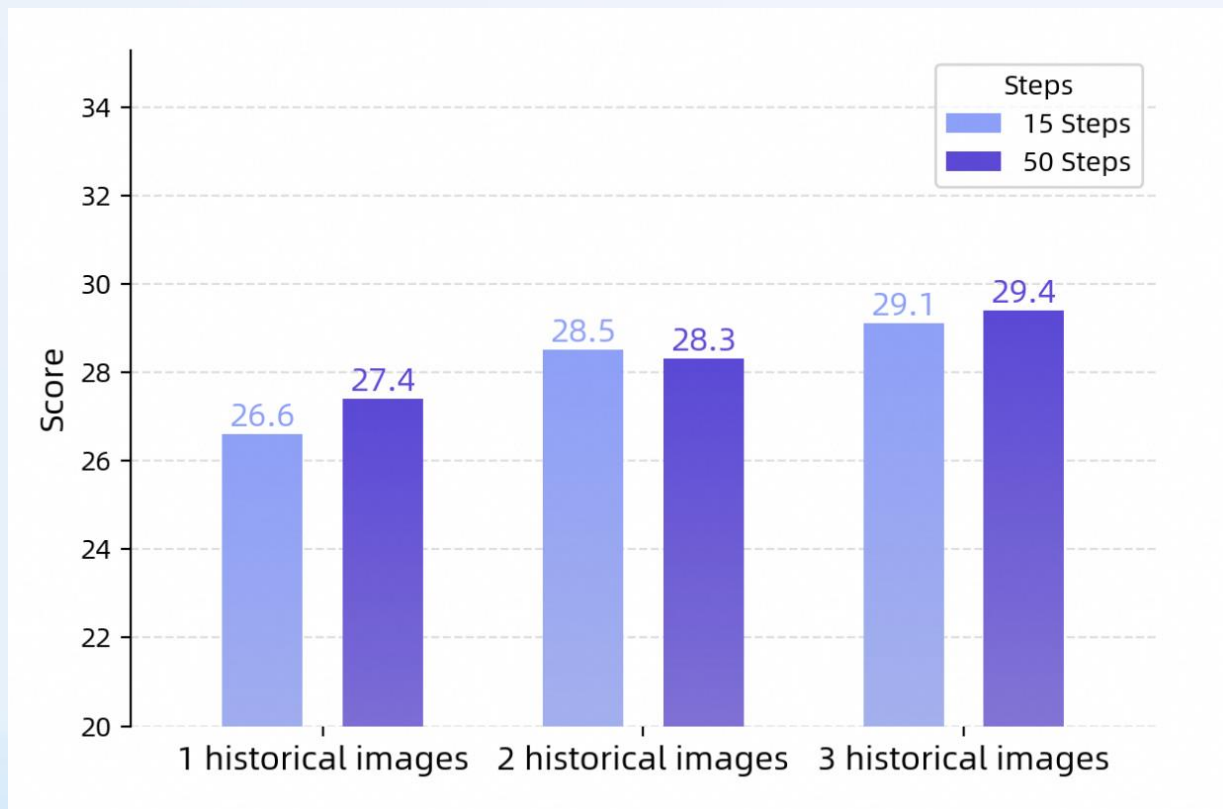


$$\mathcal{L}_{\text{TRPO}} = -\frac{1}{N} \sum_{i=1}^G \sum_{s=1}^{S_i} \sum_{t=1}^{|\mathbf{o}_{i,s}|} \left\{ \min \left[r_t(\theta) \hat{A}_{\tau_i}, \text{clip} \left(r_t(\theta), 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{\tau_i} \right] \right\}$$

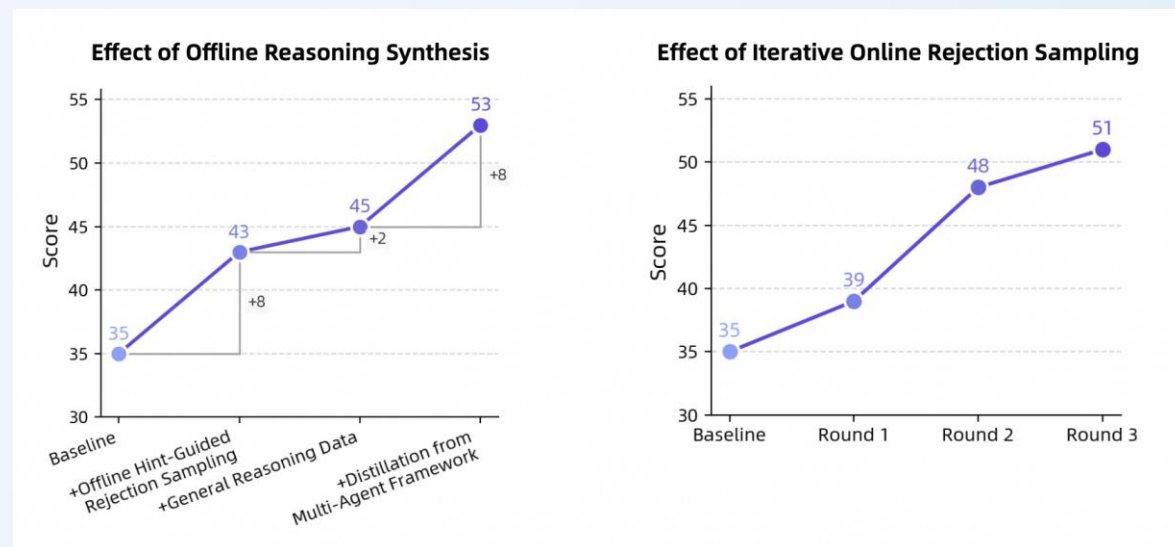
Mobile-Agent V3 Agent框架



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

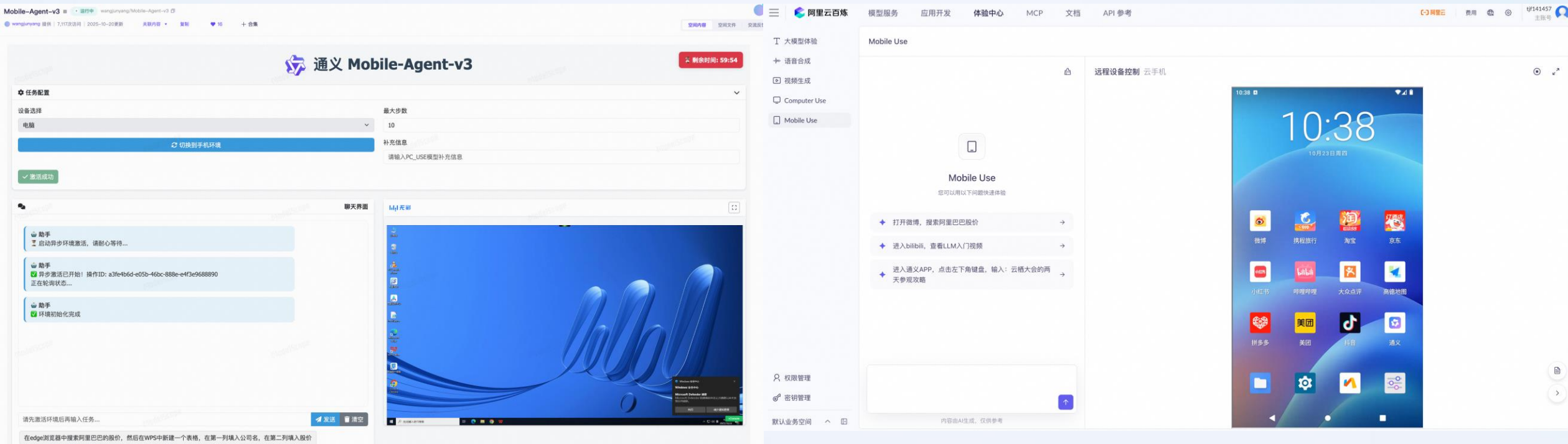
Model	Success Rate (%)	
	Mobile-Agent-E (Wang et al., 2025b) on AndroidWorld	Agent-S2 (Agashe et al., 2025) on a subset of OSWorld-Verified
<i>Baseline Models</i>		
UI-TARS-1.5 (Qin et al., 2025)	14.1	14.7
UI-TARS-72B (Qin et al., 2025)	14.8	19.0
Qwen2.5-VL-72B (Bai et al., 2025)	52.6	38.6
Seed-1.5-VL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

Performance comparison on agentic frameworks

Mobile-Agent V3云沙箱Demo



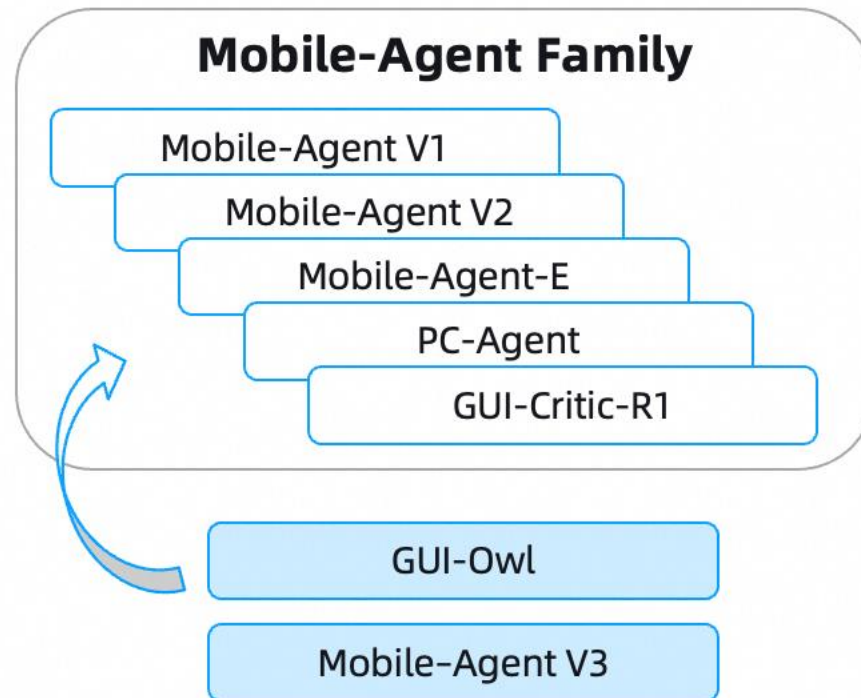
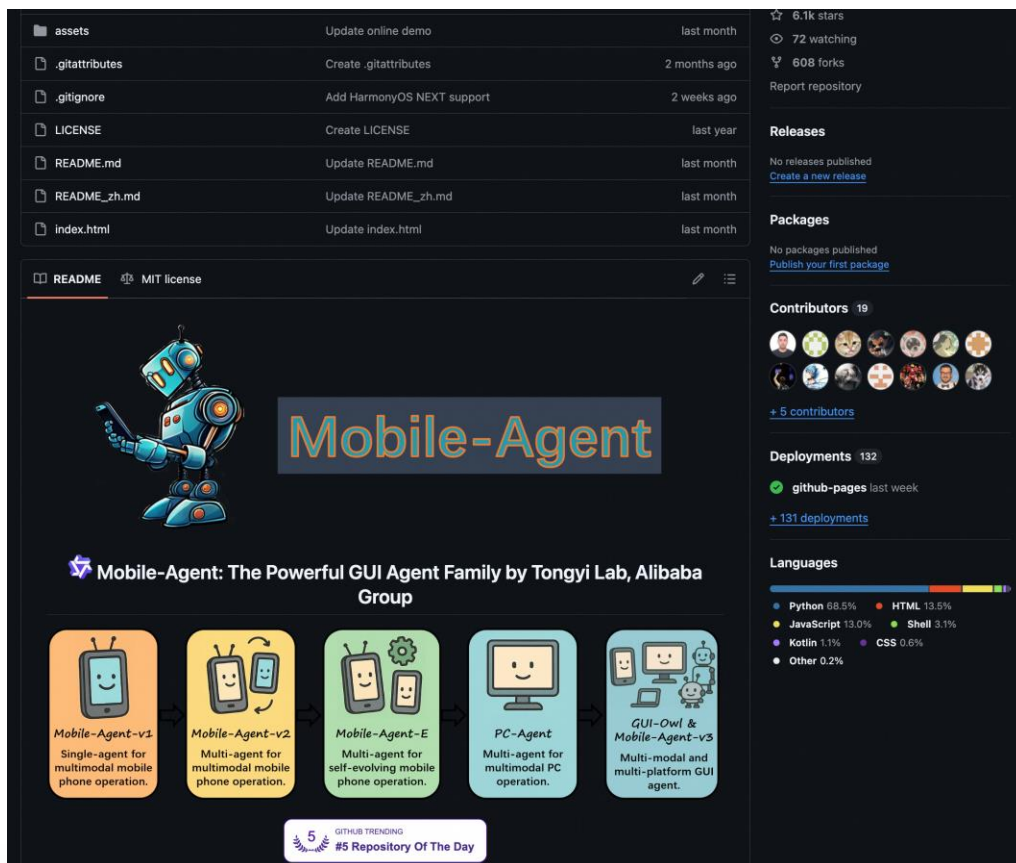
Mobile-Agent V3云沙箱Demo



<https://modelscope.cn/studios/wangjunyang/Mobile-Agent-v3>

<https://bailian.console.aliyun.com>

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>



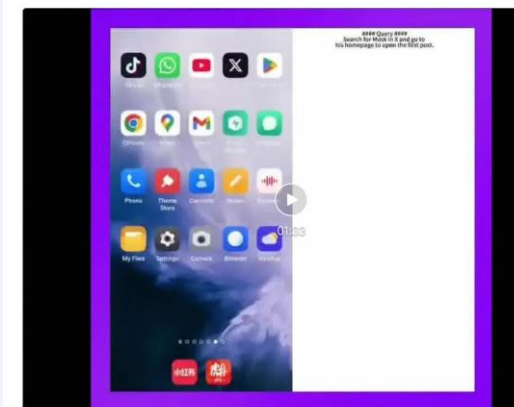
Qwen3-VL：明察、深思、广行

	Qwen3-VL 235B-A22B Instruct	Gemini2.5-Pro Thinking 1.5B	GPT5 Mini Without Ranking	Claude-Opus-4.1 Without Ranking	Other Best Without Ranking
STEM&Puzzle	MMMU_VAL	80.9	74.4*	77.2	73.6* (Seed1.5V)
	MMMU_Pro	68.1	71.2	62.7*	59.9* (Seed1.5V)
	MathVista _{mini}	84.9	77.7	50.9	83.0* (Seed1.5V)
	MathVision	66.5	66.0	45.8	57.7 (Seed1.5V)
	MathVerse _{mini}	72.5	65.9	43.0	68.1 (Seed1.5V)
General VQA	VideoLogic	29.9	26.9	27.2	27.2 (Seed1.5V)
	MMBench_EN_V1.1_dev	89.9	86.6	81.4	84.1 (Seed1.5V)
	RealWorldQA	79.3	76.0	77.3	68.5 (Seed1.5V)
	MMStar	78.4	78.5	65.2	71.0 (Seed1.5V)
Subjective Experience and Instruction Following	SimpleVQA	63.0	66.9	56.7	55.7 (Seed1.5V)
	HallusionBench	63.2	60.9	53.7	55.1 (Seed1.5V)
	MM_MT_Bench	8.5	7.6	7.5	7.9 (Seed1.5V)
	MIABench	91.3	91.3	92.6	90.0 (Seed1.5V)
Text Recognition and Chart/Document Understanding	MMLongBench-Doc	57.0	51.2	42.4	48.1 (Seed1.5V)
	DocVQA _{TEST}	97.1	94.0	89.6	89.2 (Seed1.5V)
	InfoVQA _{TEST}	89.2	82.9	69.9	60.9 (Seed1.5V)
	A1ZD _{TEST}	89.7	90.0	84.1	84.4 (Seed1.5V)
	OCRBench	920.0	872.0	787.0	750.0 (Seed1.5V)
2D/3D Grounding	OCRBenchV2 _(en/zh)	67.1 / 61.8	55.2 / 53.1	48.2 / 37.7	47.2 / 38.0 (Seed1.5V)
	CC_OCR	82.2	76.8	66.1	66.0 (Seed1.5V)
	CharXiv(RQ)	62.1	62.9	57.8*	60.2 (Seed1.5V)
	RefCOCO _{avg}	91.9	—	—	91.6* (Seed1.5V)
	CountBench	93.0	91.0	87.8	91.9 (Seed1.5V)
Multi-Image	ODinW13	48.6	34.5	—	40.6 (Seed1.5V)
	ARKiScenes	56.9	—	—	27.5 (Seed1.5V)
	Hypersim	13.0	—	—	9.6 (Seed1.5V)
	SUNRGBD	39.4	—	—	33.5* (Seed1.5V)
Embodied and Spatial Understanding	BLINK	70.7	70.0	62.8	62.9 (Seed1.5V)
	MUIRBENCH	72.8	74.0	66.5	— (Seed1.5V)
	ERQA	51.3	50.3	42.0*	26.0 (Seed1.5V)
	VSI-Bench	62.6	31.4	—	48.4* (Seed1.5V)
Video	EmbSpatialBench	83.1	73.3	75.1	66.0 (Seed1.5V)
	RefSpatialBench	65.5	35.4	23.1	— (Seed1.5V)
	RoboSpatialHome	69.5	49.2	43.6	— (Seed1.5V)
	VideoMME(w/o sub)	79.2	80.6	77.3	73.3 (Seed1.5V)
Agent	MLVU	84.3	81.2	78.3	71.2 (Seed1.5V)
	VLBench	67.7	69.0	—	— (Seed1.5V)
	CharadesSTA	64.8	—	—	— (Seed1.5V)
	VideoMMU	74.7	79.4	61.6*	70.1 (Seed1.5V)
Coding	ScreenSpot	95.4	—	—	— (Seed1.5V)
	ScreenSpot Pro	62.0	—	—	— (Seed1.5V)
	OSWorldG	66.7	—	—	— (Seed1.5V)
	AndroidWorld	63.7	—	—	— (Seed1.5V)
Other Best Without Ranking	Design2Code	92.0	90.3	88.9	85.3 (Seed1.5V)
	ChartMimic_v2_Direct	80.5	79.9*	—	— (Seed1.5V)
	UniSvg	69.3	67.9	—	— (Seed1.5V)

Note: The default evaluation is conducted through API calls, and * indicates scores from the report. Results on video understanding are measured using a 2560-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。比如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。



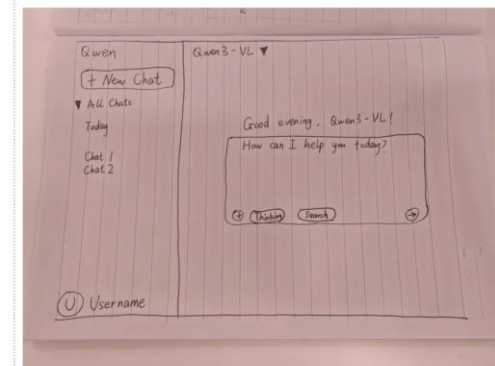
Example: Think with Image

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:

```
49 font-weight: 600;
50 }
51 .new-chat-btn {
52   background-color: var(--dark-tertiary);
53   border: 1px solid var(--border);
54   border-radius: 8px;
55   padding: 12px 20px;
56   font-size: 18px;
57   font-weight: 500;
```

● 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

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

THANKS

大模型正在重新定义软件

Large Language Model Is Redefining The Software

