多模态手机智能体Mobile-Agent技术解析及应用

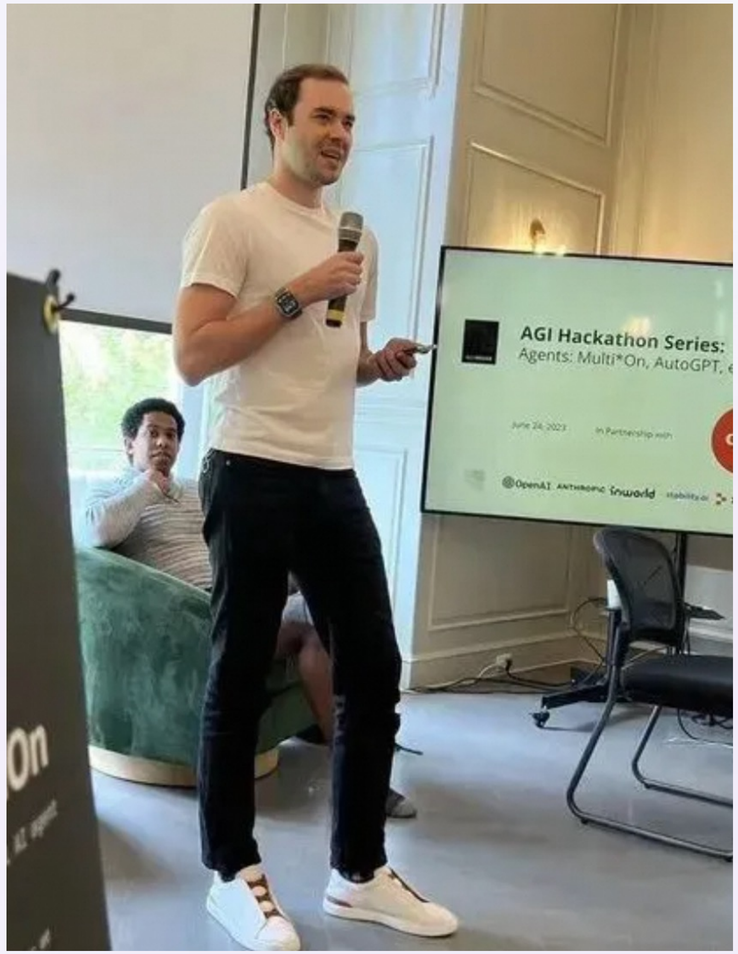
徐海洋

通义实验室自然语言智能团队

2024/06/19

ModelScope

大模型智能体是人工智能应用的未来



“如果一篇论文提出了某种不同的训练方法，我们内部的Slack上会嗤之以鼻，认为都是我们玩剩下的。但是当新的AI Agents论文出来的时候，我们会认真兴奋的讨论” — Andrej Karpathy

“AI Agent不仅会改变每个人与计算机交互方式。它还将颠覆软件行业，带来自我们从键入命令到点击图标以来最大的计算变革” — 比尔盖茨

大模型智能体的优势

OpenAI Five

DeepMind AlphaStar

LLM Agent with ChatGPT

传统基于RL的智能体的局限性

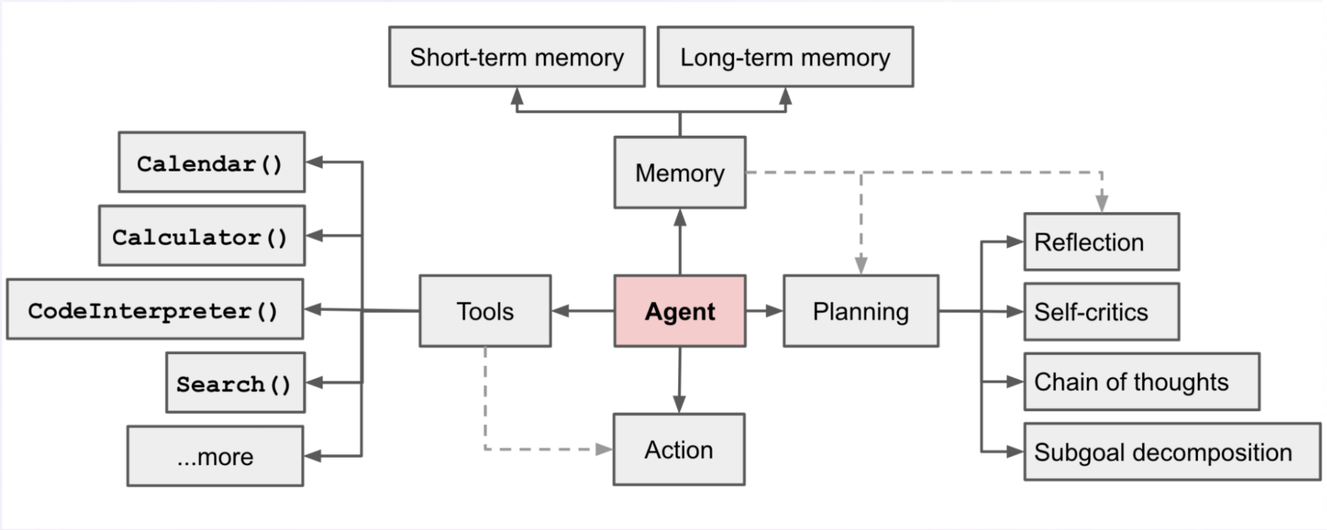
- 数据采样专有环境和低效
- 面向特定任务
- 稀疏奖励和长时段问题

大模型智能体的优势

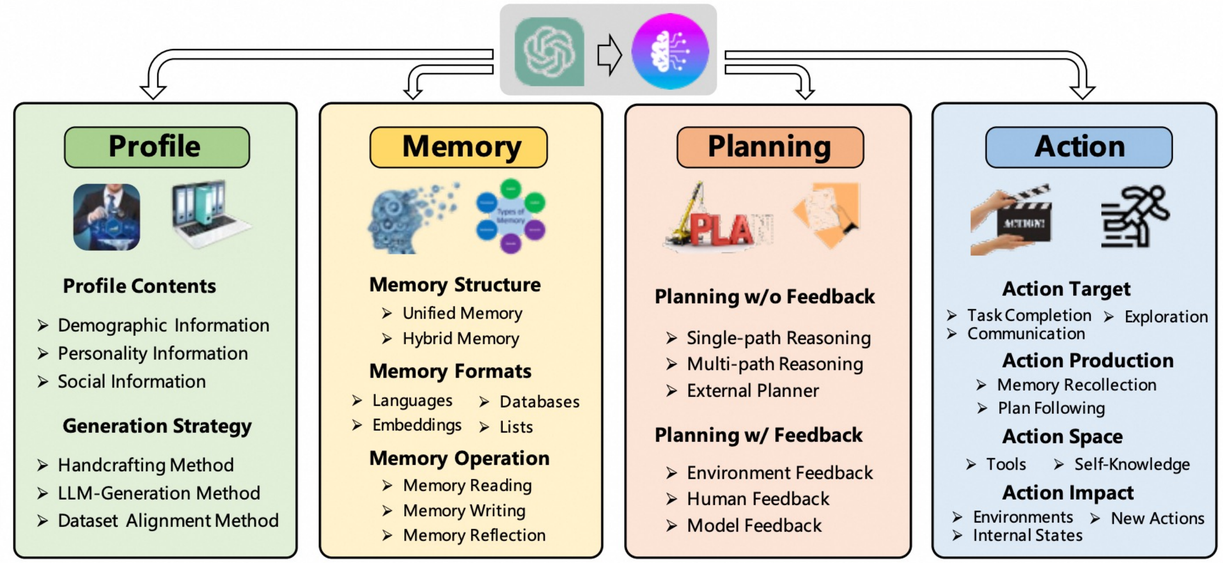
- 丰富的世界知识
- 推理/规划能力
- 工具使用（检索、code等）
- In-context Learning

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



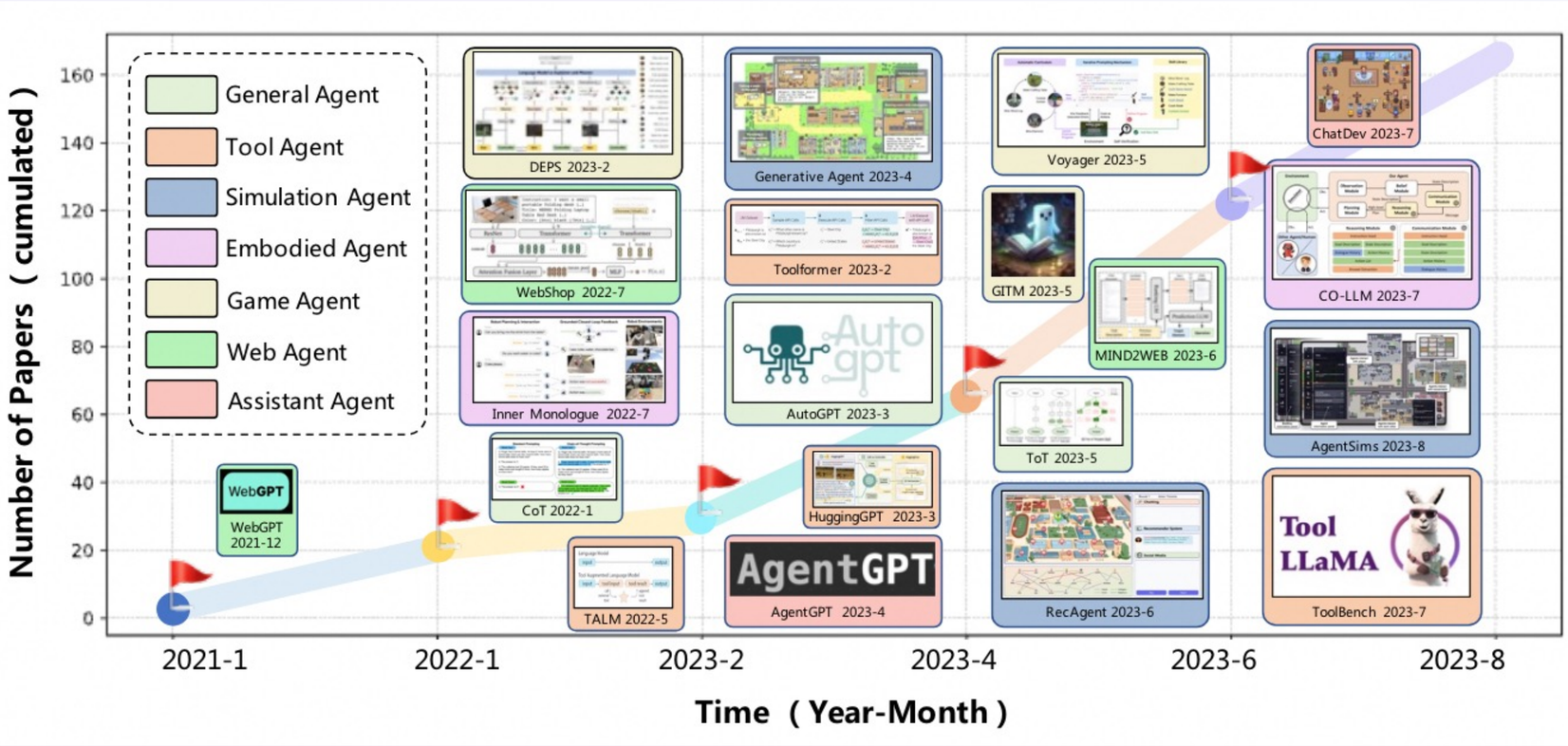
Agent System Overview from Lilian Weng's blog

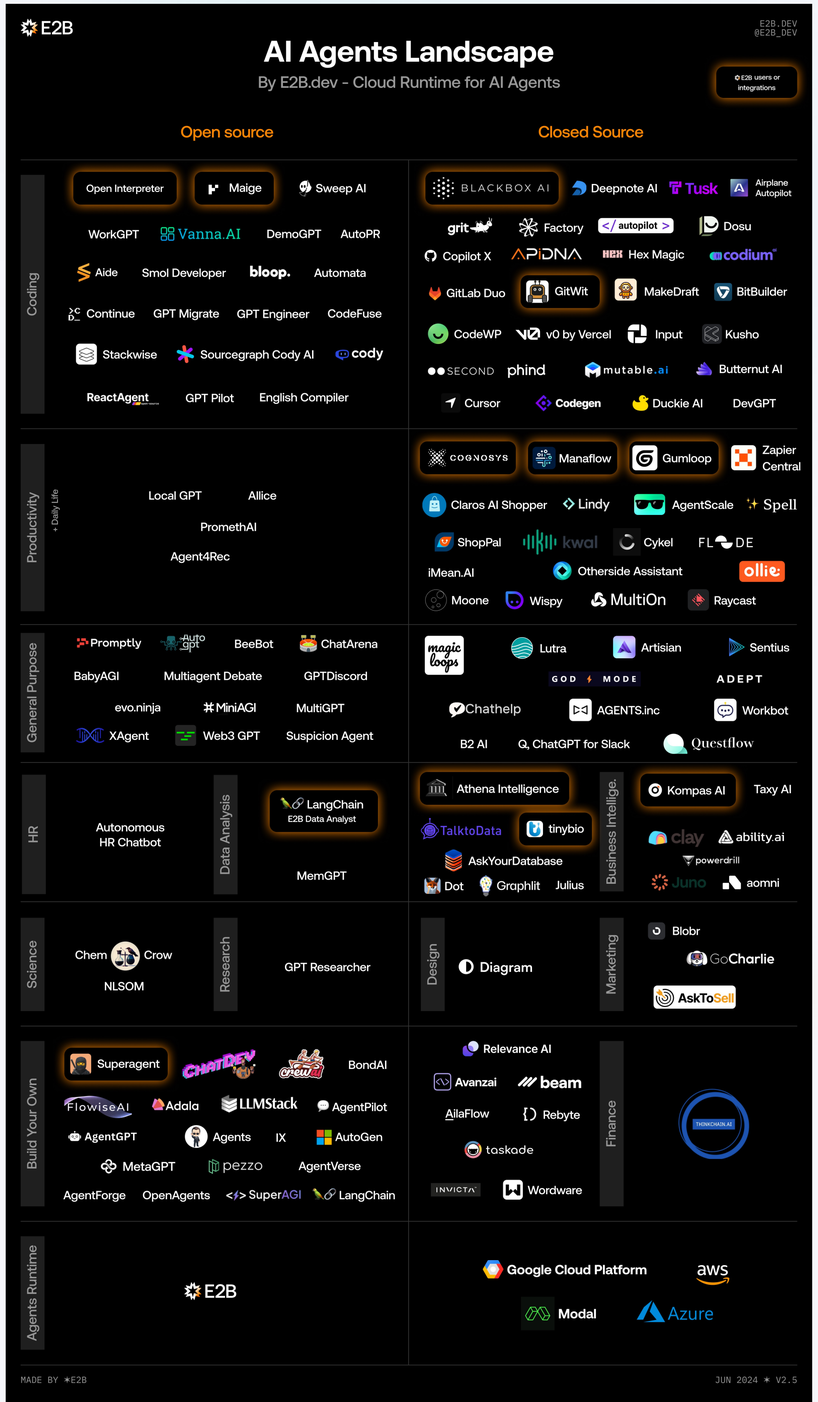


Wang et al. A Survey on Large Language Model bas Autonomous Agents

大模型智能体发展迅速

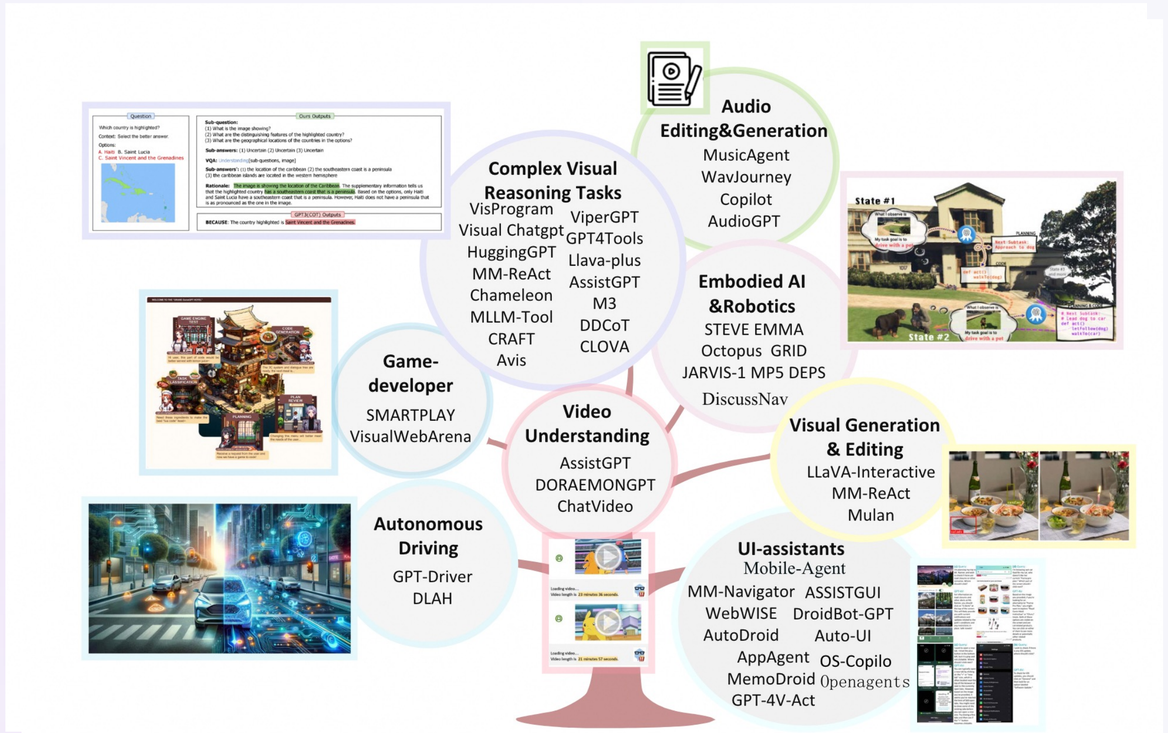
- 大模型广泛使用后，各类大模型智能体模型、框架、应用呈现井喷趋势

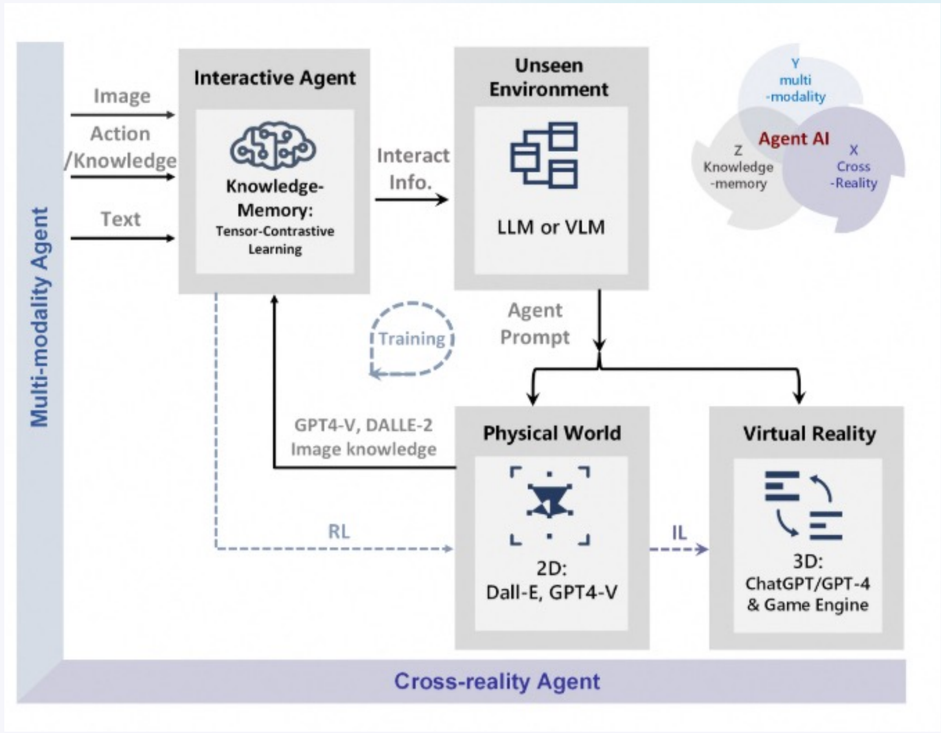


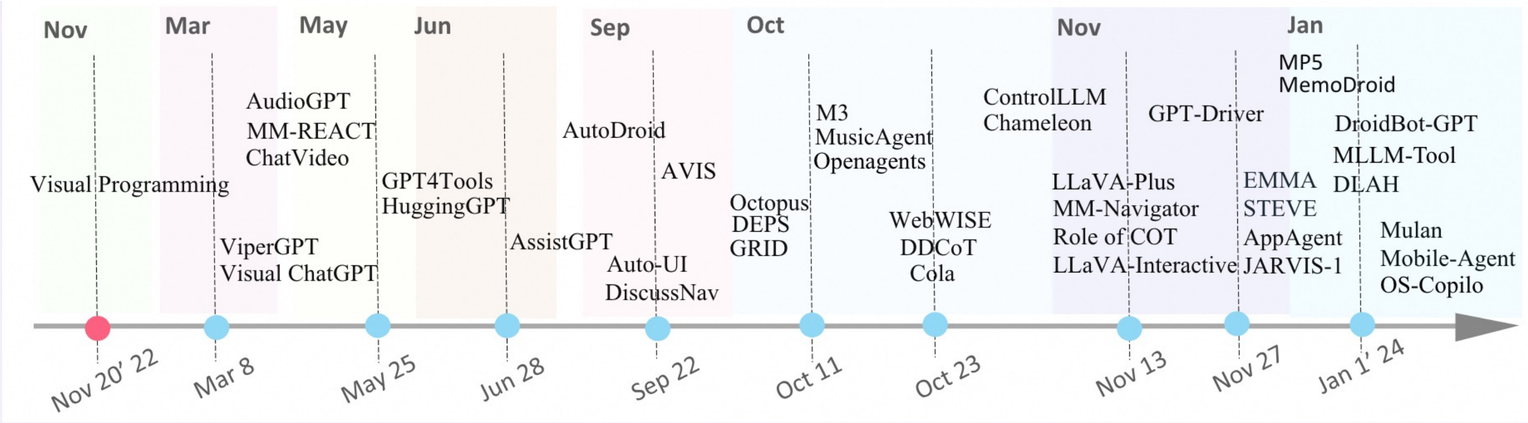


多模态大模型智能体

- 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





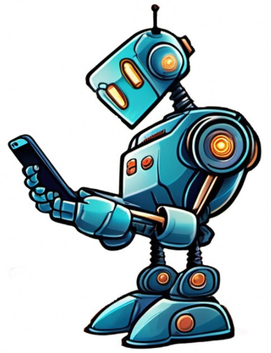


ModelScope

多模态手机智能体Mobile-Agent

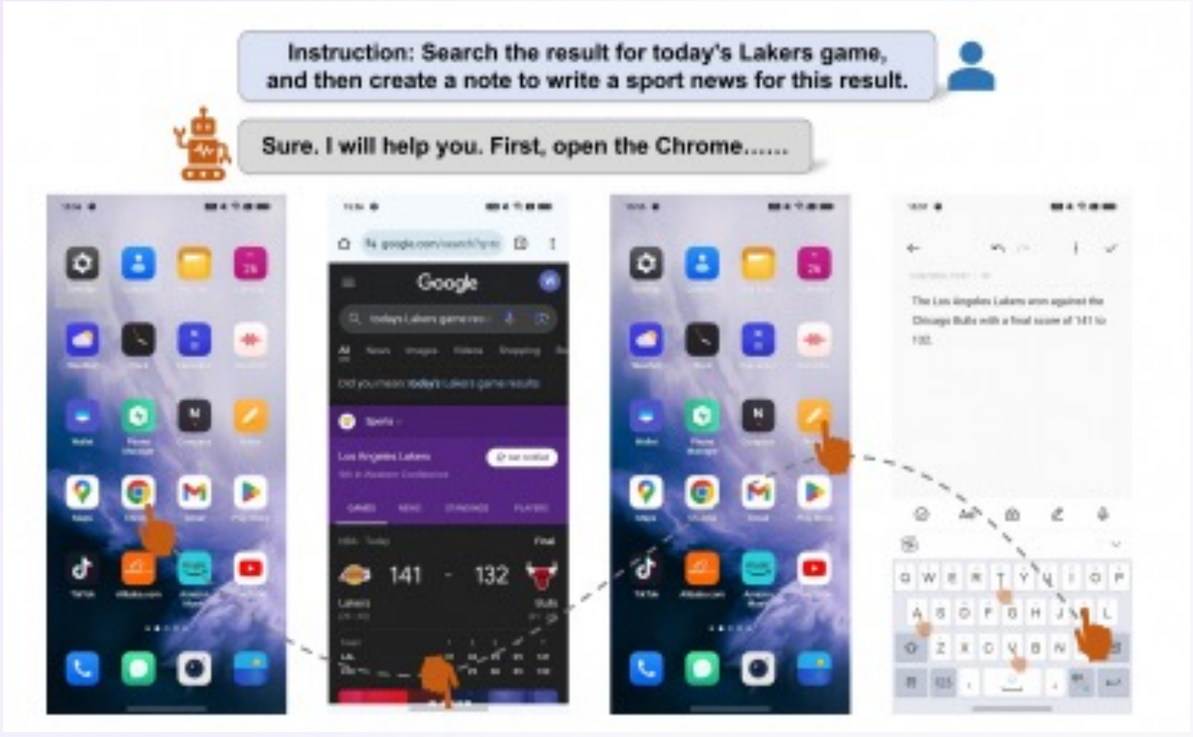
一句指令实现自动操作手机

1. 纯视觉方案，不依赖系统数据
2. 可以多个应用之间操作
3. 感知、规划、反思三者结合
4. 无需训练、即插即用



Mobile-Agent

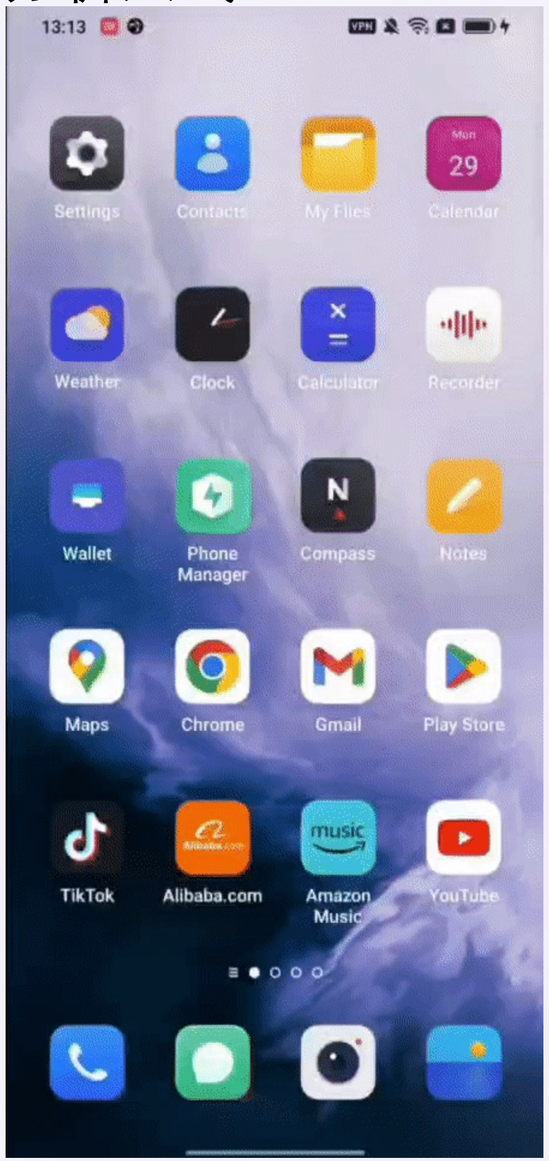
Mobile-Agent: The Powerful Mobile Device Operation Assistant Family



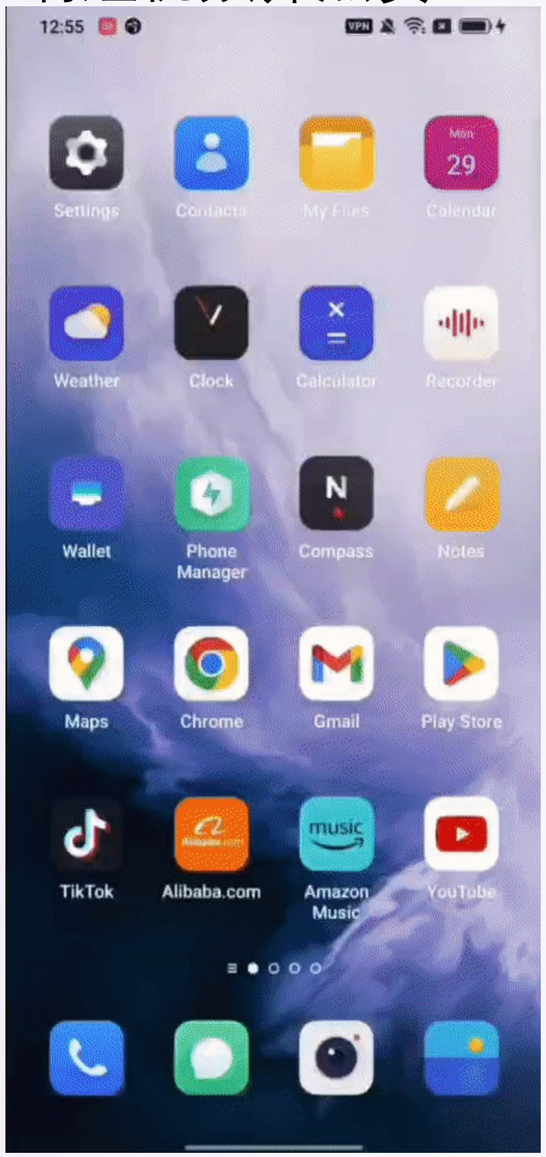
ModelScope

多模态手机智能体Mobile-Agent

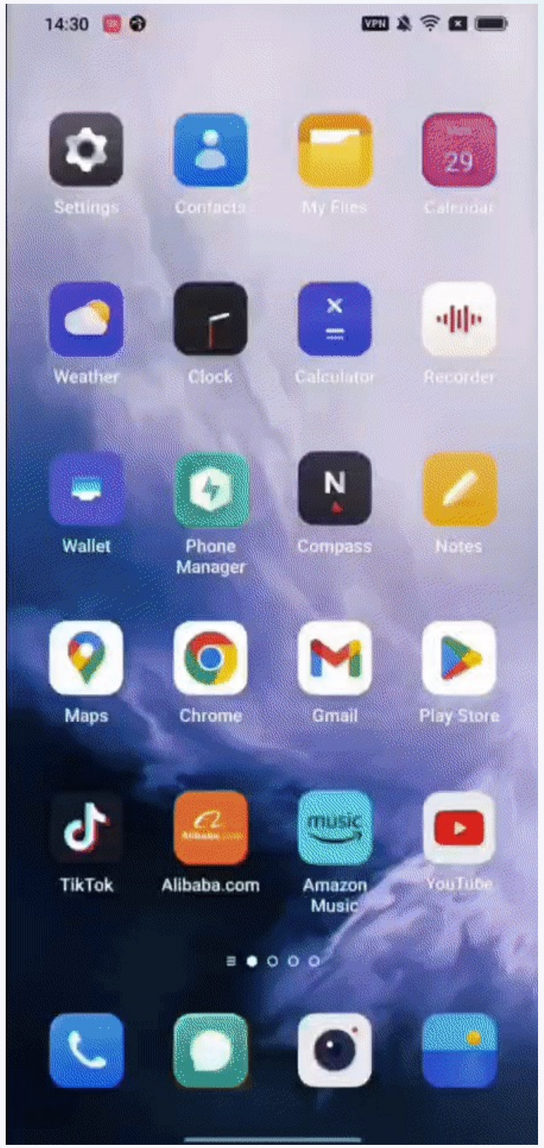
分析天气



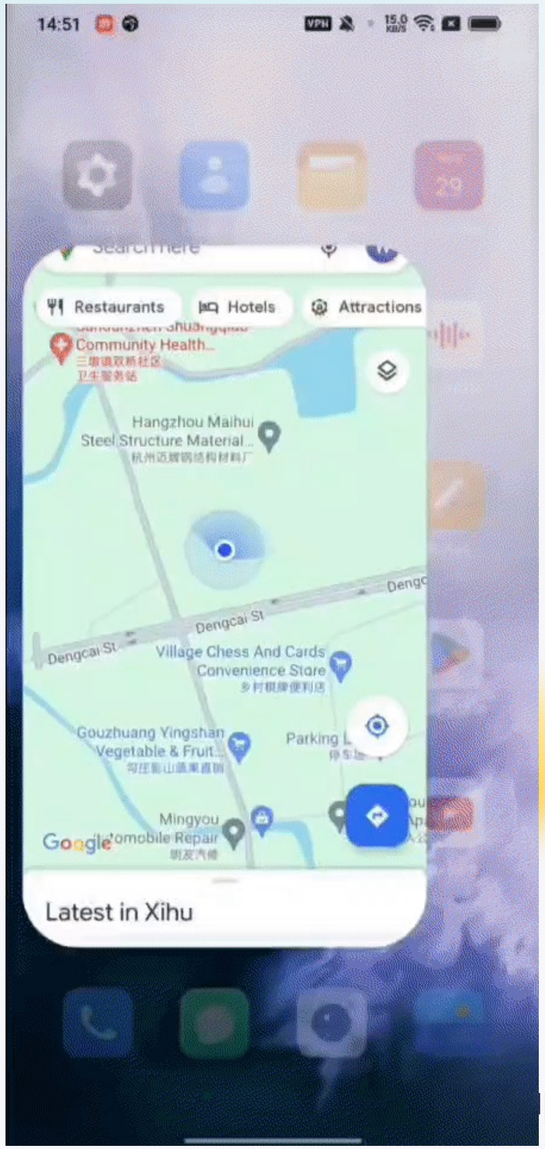
刷短视频并点赞



搜索视频并评论



导航



多模态手机智能体Mobile-Agent

Instruction: [Rules and operations of the game].
Help me play this game.

click text (5)

click text (出牌)

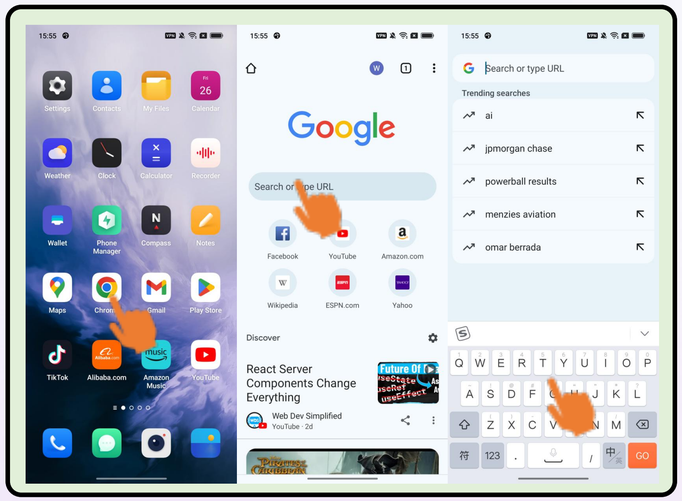
click text (出牌)

click text (7)

ModelScope

多模态手机智能体Mobile-Agent

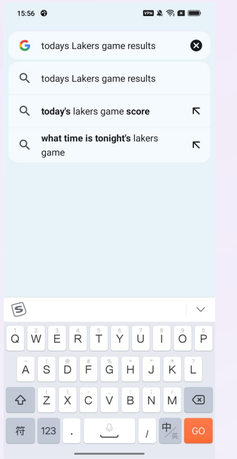
历史操作 & 截屏



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

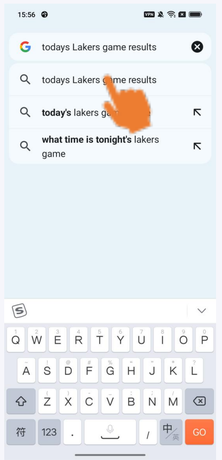
当前屏幕

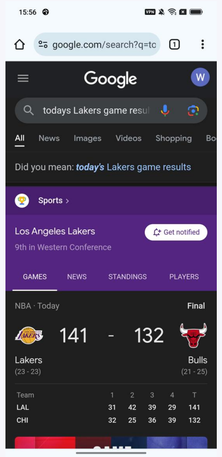


观察：描述当前页面的情况

思考：下一步的操作思路

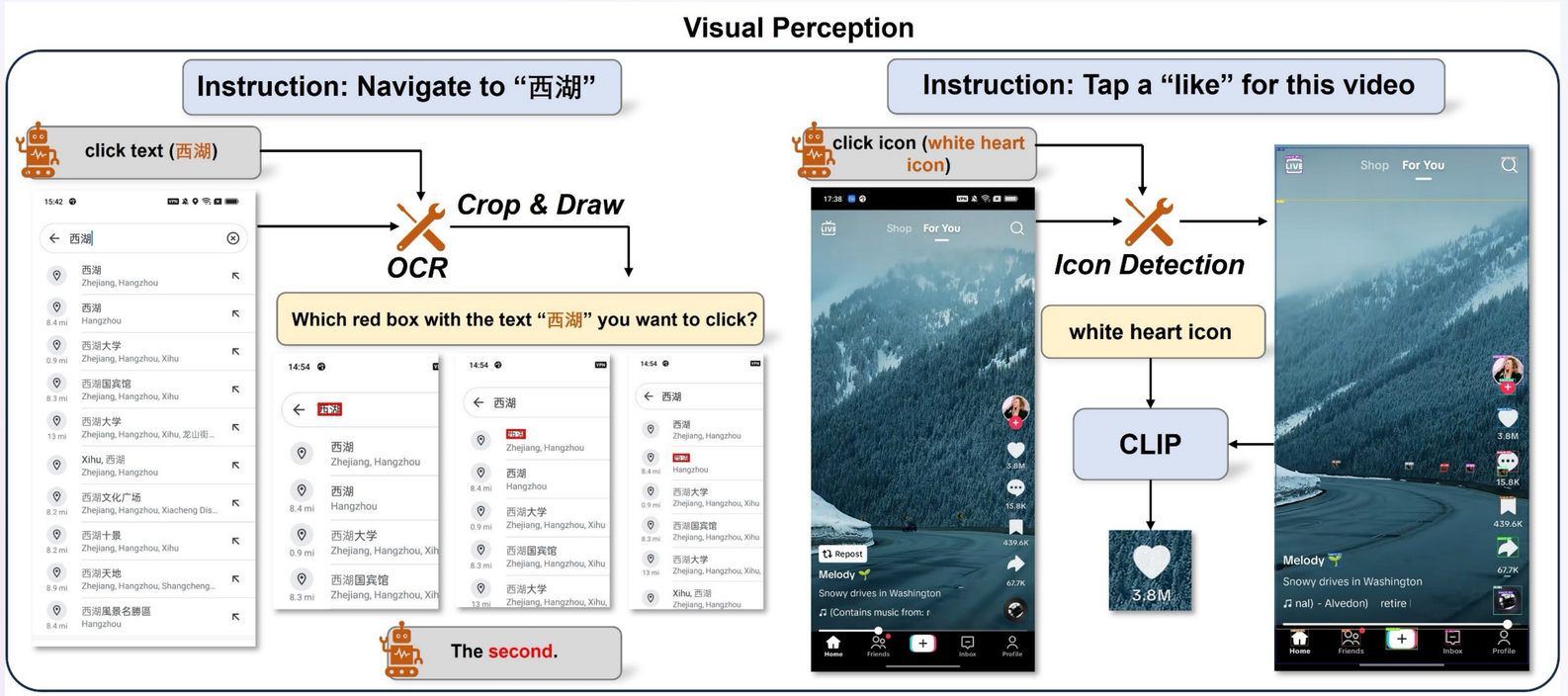
行动：使用工具完成思考中的操作





ModelScope

多模态手机智能体Mobile-Agent



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

ModelScope

多模态手机智能体Mobile-Agent

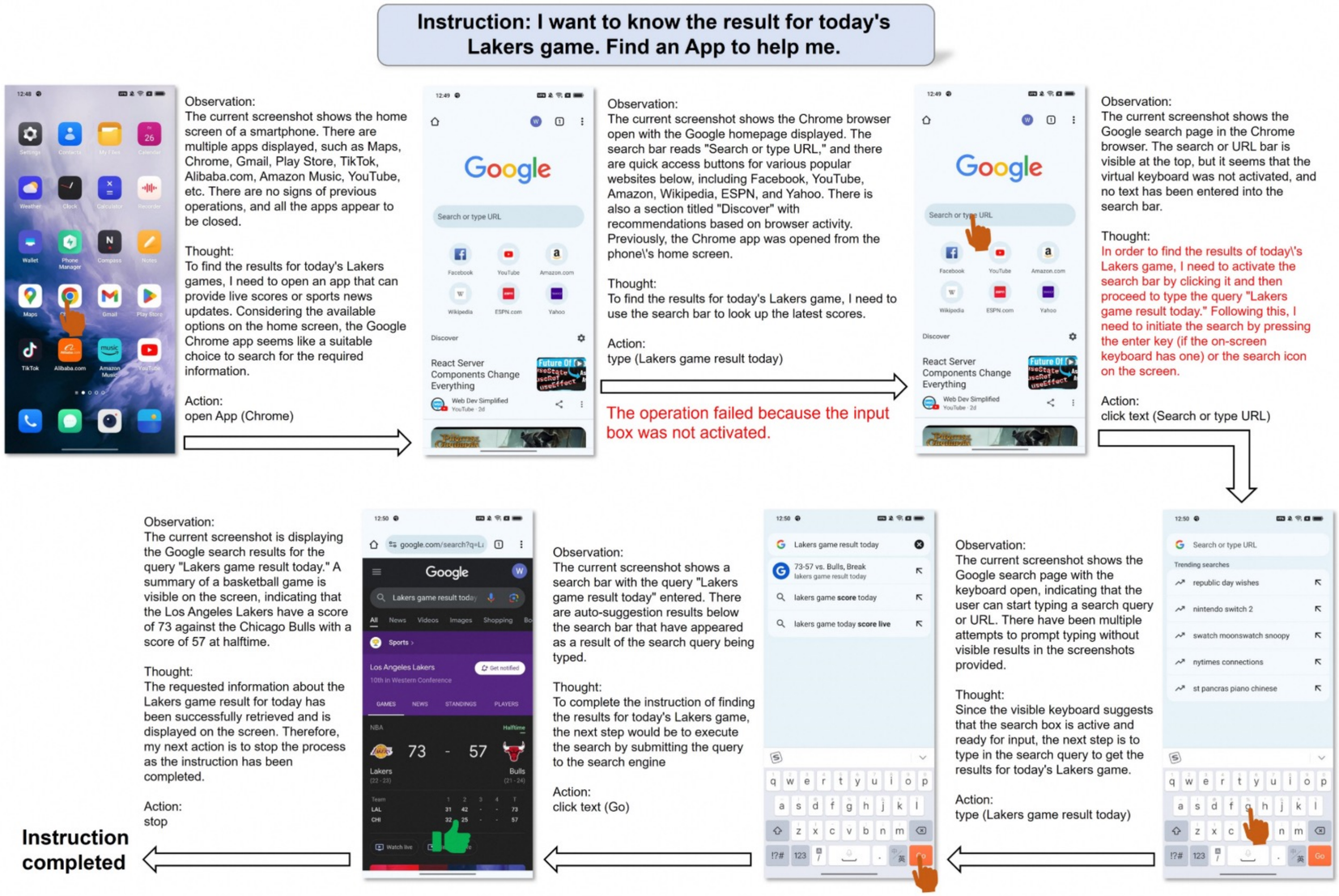


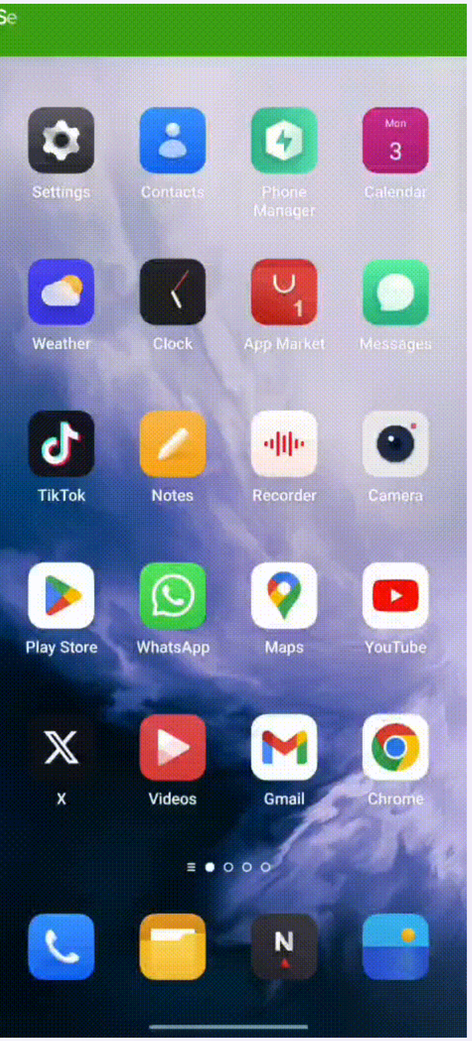
Figure 4: Case of self-reflection and error correction after using invalid operations.

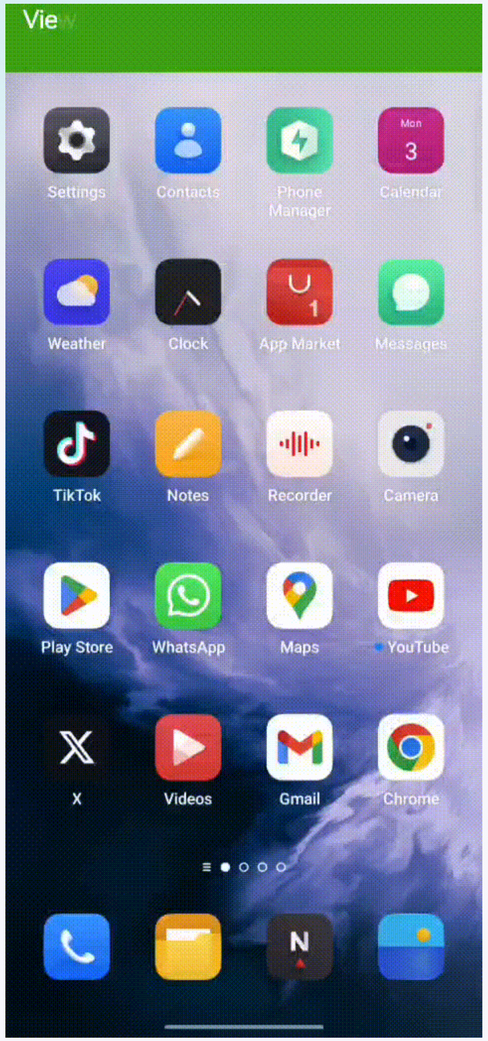
多模态手机智能体Mobile-Agent-V2

- 首次在手机操作任务上采用**多智能体架构**，并延续了一代的纯视觉方案
- 多智能体各司其职，实现了**更有效**的任务进度追踪、任务相关信息记忆和操作反思
- **更强大**的复杂指令拆解能力、跨应用操作能力和多语言场景操作能力

ModelScope

多模态手机智能体Mobile-Agent-V2

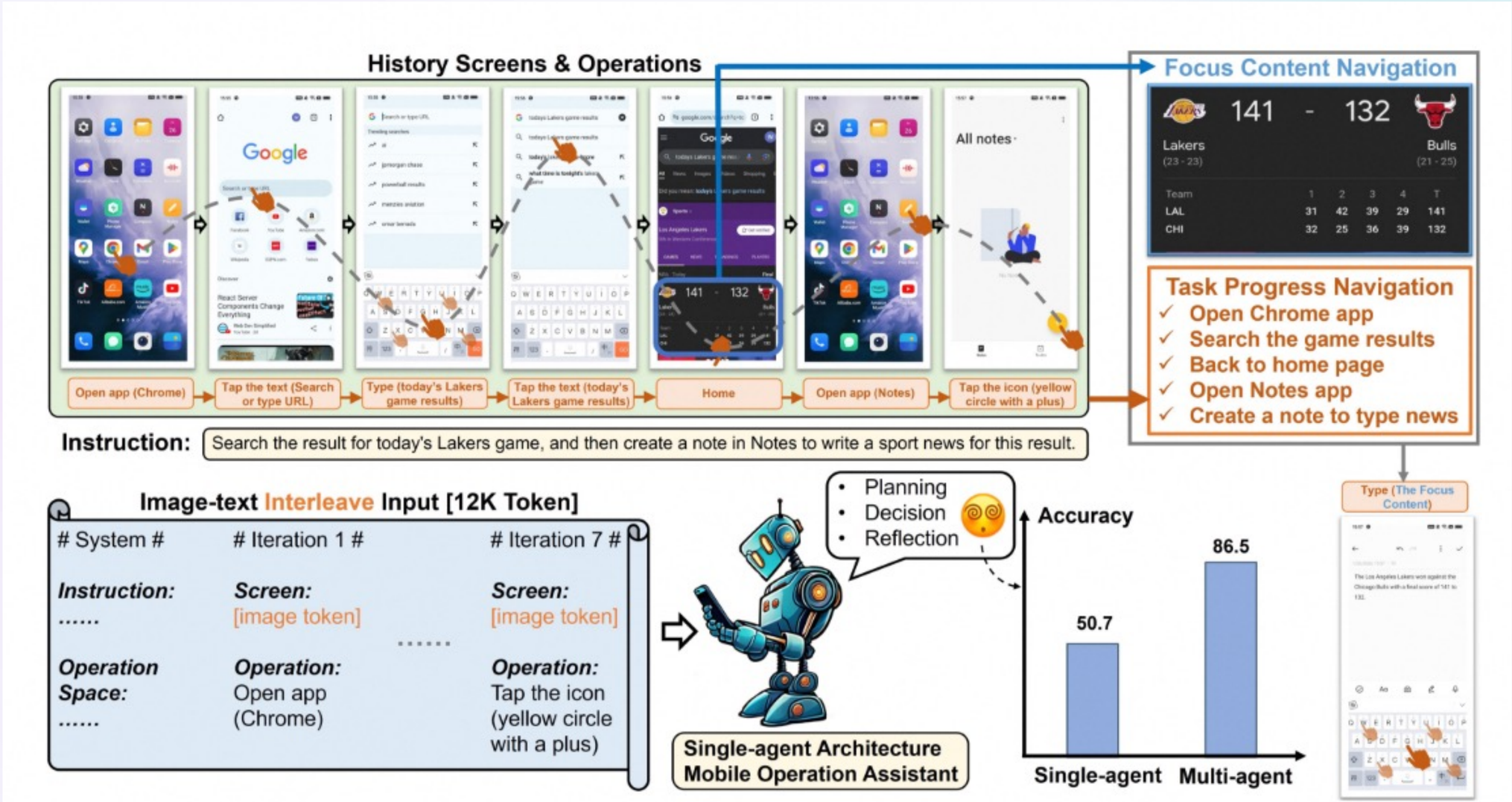




ModelScope

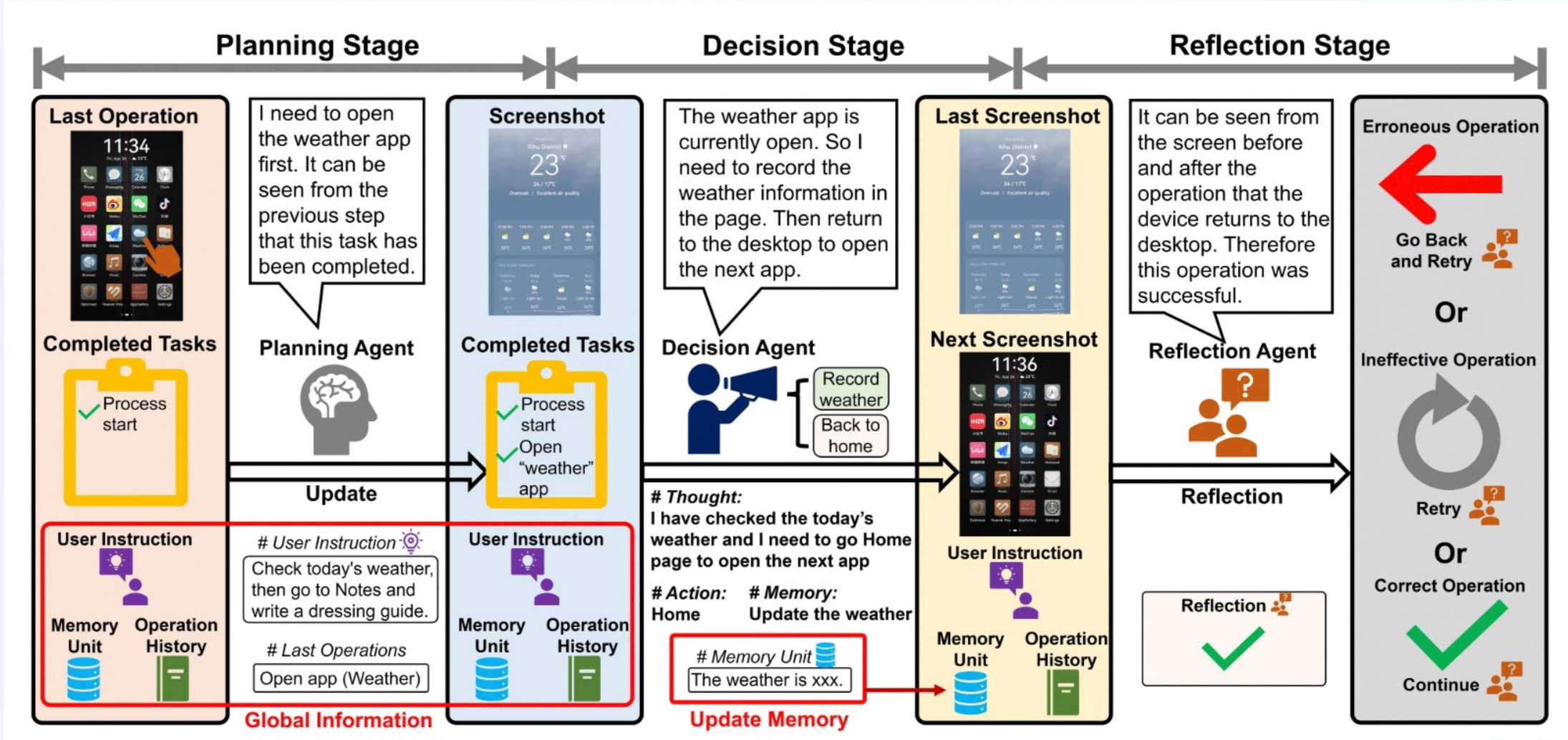
多模态手机智能体Mobile-Agent-V2

- 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



多模态手机智能体Mobile-Agent-V2

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



多模态手机智能体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.

多模态手机智能体Mobile-Agent-V2

Ablation Setting			Basic			Advanced		
Planning Agent	Reflection Agent	Memory Unit	SR	CR	DA	SR	CR	DA
	✓	✓	59.1	63.7	58.9	29.5	43.8	42.6
✓	✓		77.3	83.6	84.0	45.5	72.3	69.8
✓		✓	86.4	89.2	85.7	54.5	75.9	72.4
✓	✓	✓	**88.6**	**93.9**	**89.4**	**61.4**	**82.1**	**80.3**

Table 4: The results of the ablation study on planning agent, reflection agent, and memory unit.

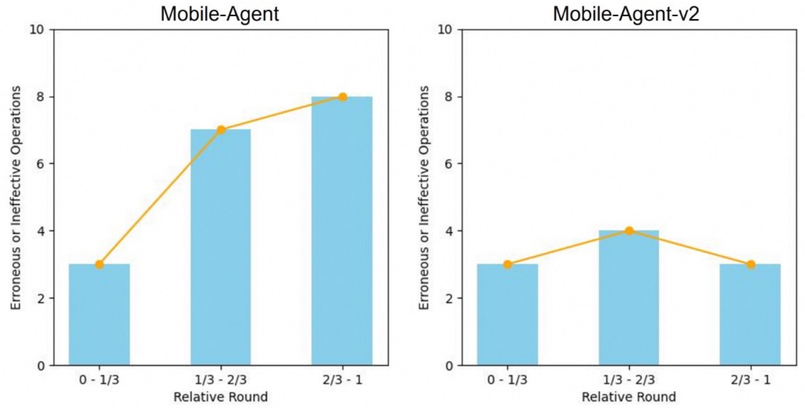
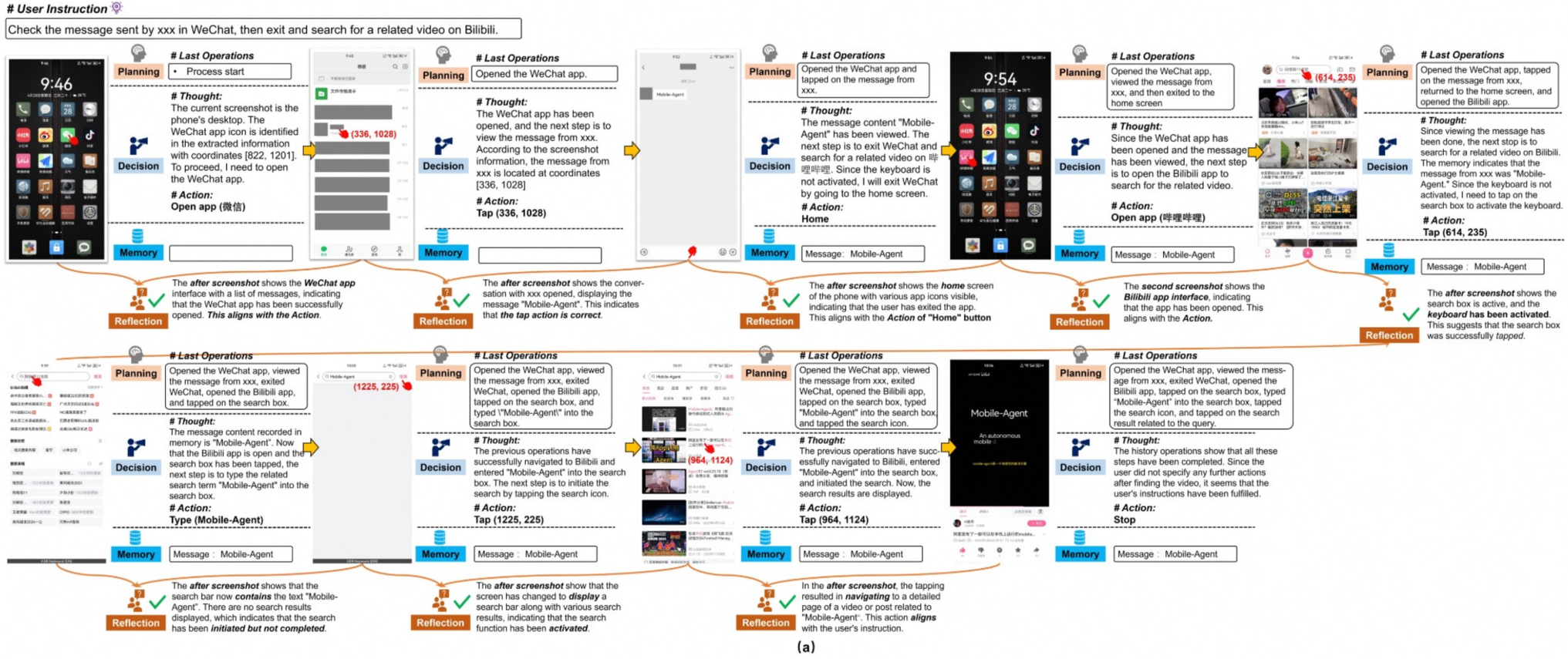


Figure 4: The relative positions of erroneous or ineffective operations in the operation sequence.

Model	Basic	Advanced
	SR&DA	SR&DA
GPT-4V w/o agent	2.7	0.9
Gemini-1.5-Pro	38.2	29.8
Qwen-VL-Max	42.1	33.6
GPT-4V	**92.7**	**83.5**

Table 3: Performance results of Mobile-Agent-v2 with different MLLMs. To better illustrate the differences, we converted all instructions to single-step forms and evaluated the success rate (which is the same as decision accuracy) of each single-step task.

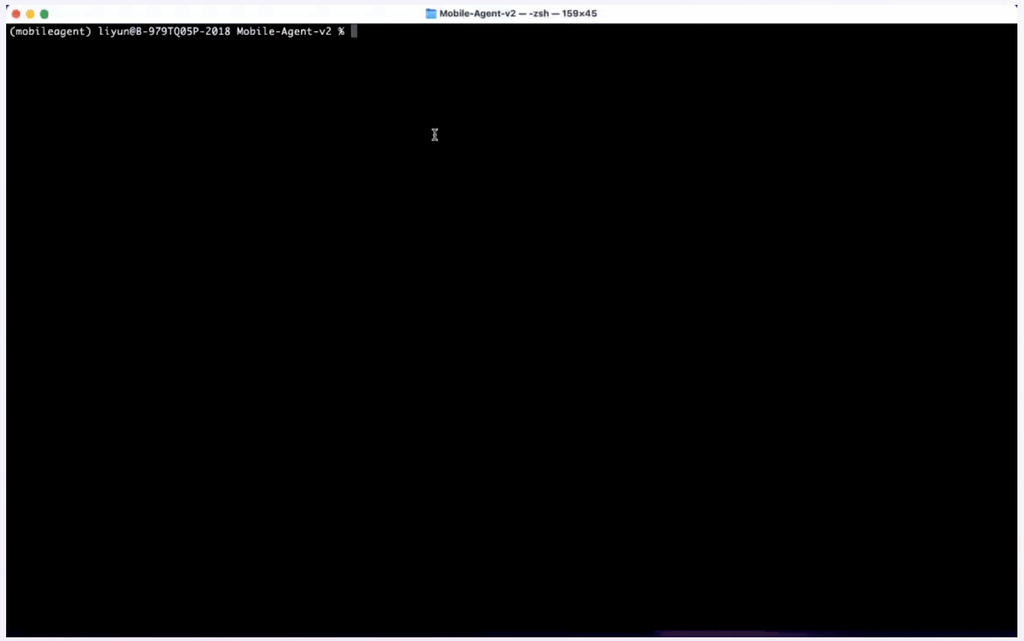
多模态手机智能体Mobile-Agent-V2



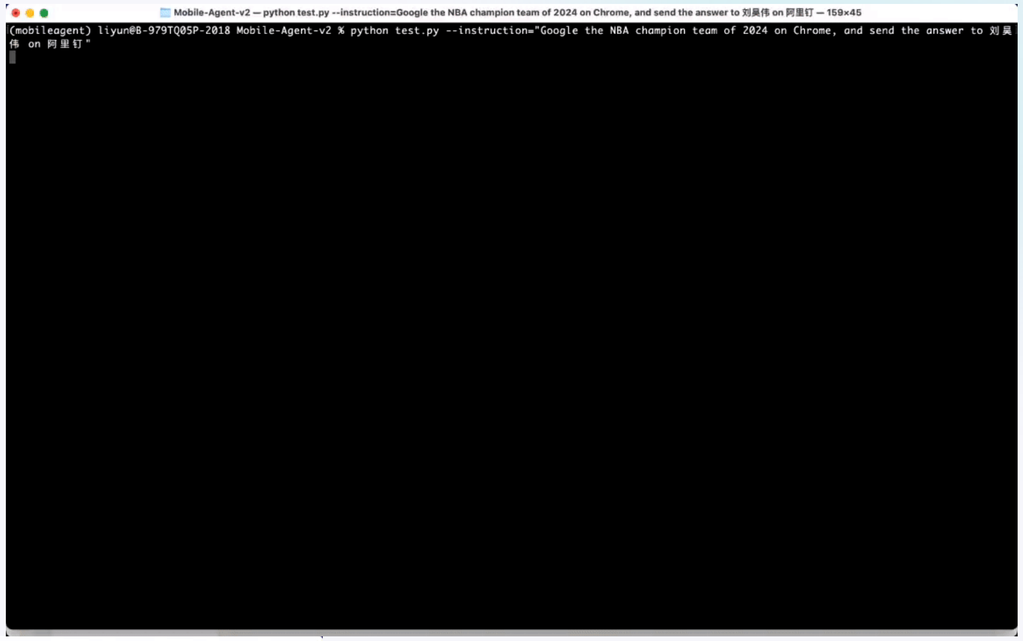
ModelScope

多模态电脑智能体PC-Agent

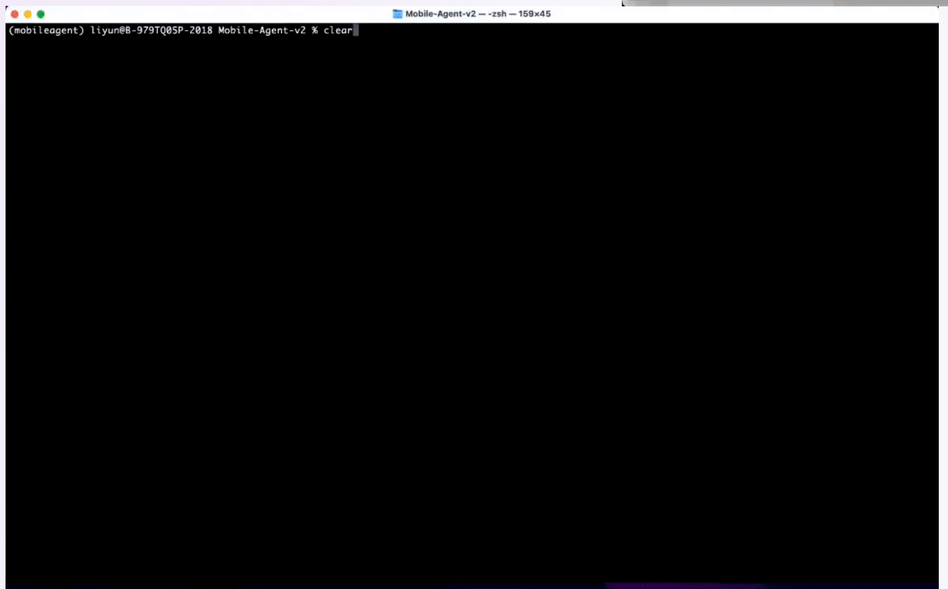
论文搜索下载



搜索新闻发送钉钉



Word编辑保存



ModelScope

多模态手机智能体Mobile-Agent开源

MobileAgent Public

Edit Pins | Unwatch 35 | Fork 179 | Starred 2.2k

main | 1 Branch | 0 Tags | Go to file t | Add file | Code

junyangwang0410 Update README.md 7e353ef · 6 hours ago 88 Commits

Mobile-Agent-v2	Update README_zh.md	6 hours ago
Mobile-Agent	Update README.md	6 hours ago
assets	Update for v2	2 weeks ago
LICENSE	Create LICENSE	5 months ago
README.md	Update README.md	2 weeks ago
README_zh.md	Update README_zh.md	2 weeks ago

README | MIT license

Mobile-Agent

Mobile-Agent: The Powerful Mobile Device Operation Assistant Family

About

Mobile-Agent: The Powerful Mobile Device Operation Assistant Family

arxiv.org/abs/2406.01014

android agent harmony ios app gui automation mobile copilot multimodal mobile-agents mllm multimodal-large-language-models gpt4v multimodal-agent

Readme
MIT license
Activity
Custom properties
2.2k stars
35 watching
179 forks
Report repository

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

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

modelscope-agent Public

Watch 29 | Fork 242 | Starred 2.1k

master | 41 Branches | 9 Tags | Go to file t | Add file | Code

mushenL and 杨堃 Add error throwing for the Ollama section regarding the ... c4428f3 · yesterday 360 Commits

.dev_scripts	Add tool manage service (#384)	last month
.github	allow_fork_pass_ci (#449)	last month
apps	feat/gradio_gc (#490)	4 days ago
config	update qwen-max -> qwen-max-1201 (#246)	5 months ago
docker	Feat: assistant api (#423)	last month
docs	remove complicated import which might lead import error...	5 days ago
examples	chore(doc): fix broken links and typos (#494)	yesterday
modelscope_agent	Add error throwing for the Ollama section regarding the d...	yesterday
modelscope_agent_servers	Doc/qwen2 (#473)	2 weeks ago
resources	Refactor with new framework (#267)	4 months ago
scripts	fix model_server conflict (#485)	5 days ago
tests	Feat/rag use llm (#483)	5 days ago
.gitignore	Add tool manage service (#384)	last month
.pre-commit-config.yaml	Feat/web browser (#169)	6 months ago
.pre-commit-config_local.yaml	modelscope agent	10 months ago
CODE_OF_CONDUCT.md	modelscope agent	10 months ago

About

ModelScope-Agent: An agent framework connecting models in ModelScope with the world

modelscope.cn/studios/agent

agent chatbot android-application multi-agents mobile-agents gpts llm multimodal-large-language-models qwen assistantapi chatglm-4 open-gpts mobile-agent

Readme
Apache-2.0 license
Code of conduct
Activity
Custom properties
2.1k stars
29 watching
242 forks
Report repository

Releases 8

v0.6.1 release: Rag Module U... Latest
last week

+ 7 releases

https://github.com/modelscope/modelscope-agent

ModelScope

多模态手机智能体Mobile-Agent开源

🔧开始

❗目前仅安卓和鸿蒙系统（版本号 <= 4）支持工具调试。其他系统如iOS暂时不支持使用Mobile-Agent。

安装依赖

```
pip install -r requirements.txt
```

准备通过ADB连接你的移动设备

1. 下载 Android Debug Bridge（ADB）。
2. 在你的移动设备上开启“USB调试”或“ADB调试”，它通常需要打开开发者选项并在其中开启。
3. 通过数据线连接移动设备和电脑，在手机的连接选项中选择“传输文件”。
4. 用下面的命令来测试你的连接是否成功: `/path/to/adb devices`。如果输出的结果显示你的设备列表不为空，则说明连接成功。
5. 如果你是用的是MacOS或者Linux，请先为 ADB 开启权限: `sudo chmod +x /path/to/adb`。
6. `/path/to/adb` 在Windows电脑上将是 `xx/xx/adb.exe` 的文件格式，而在MacOS或者Linux则是 `xx/xx/adb` 的文件格式。

在你的移动设备上安装 ADB 键盘

1. 下载 ADB 键盘的 apk 安装包。
2. 在设备上点击该 apk 来安装。
3. 在系统设置中将默认输入法切换为 “ADB Keyboard”。

选择适合的运行方式

1. 在 `run.py` 的22行起编辑你的设置， 并且输入你的 ADB 路径，指令，GPT-4 API URL 和 Token。

2.选择适合你的设备的图标描述模型的调用方法：

- 如果您的设备配备了高性能GPU，我们建议使用“local”方法。它是指在本地设备中部署图标描述模型。如果您的设备足够强大，则通常具有更好的效率。
- 如果您的设备不足以运行7B 大小的 LLM，请选择“ api”方法。我们使用并行调用来确保效率。

3.选择图标描述模型：

- 如果选择“local”方法，则需要在“qwen-vl-chat”和“qwen-vl-chat-int4”之间进行选择，其中“qwen-vl-chat”需要更多的GPU内存，但提供了更好的性能与“qwen-vl-chat-int4”相比。同时，“qwen_api”可以是空置的。
- 如果您选择“api”方法，则需要在“qwen-vl-plus”和“qwen-vl-max”之间进行选择，其中“qwen-vl-max”需要更多的费用，但与“qwen-vl-plus”相比提供了更好的性能。此外，您还需要申请Qwen-VL 的 API-KEY，并将其输入到“qwen_api”。

4.您可以在“add_info”中添加操作知识（例如，完成您需要的指令所需的特定步骤），以帮助更准确地运行移动设备。

5.如果您想进一步提高移动设备的效率，则可以将“ reflection_Switch”和“ memory_switch”设置为“ False”。

- “ reflection_switch”用于确定是否在此过程中添加“反思智能体”。这可能会导致操作陷入死周期。但是您可以将操作知识添加到“ add_info”中以避免它。
- “ memory_switch”用于决定是否将“内存单元”添加到该过程中。如果你的指令中不需要在后续操作中使用之前屏幕中的信息，则可以将其关闭。

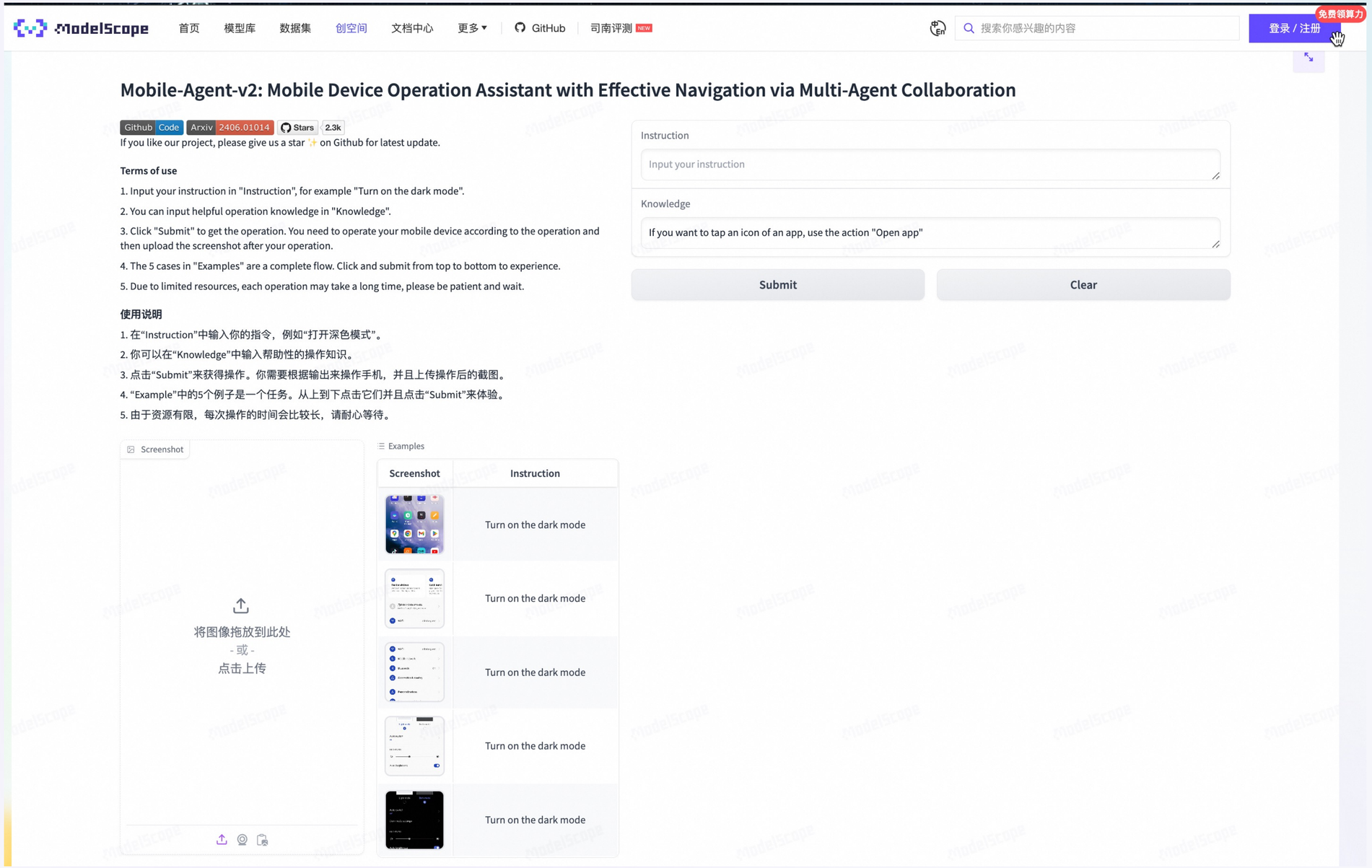
运行

```
python run.py
```

https://github.com/X-PLUG/MobileAgent/blob/main/Mobile-Agent-v2/README_zh.md

ModelScope

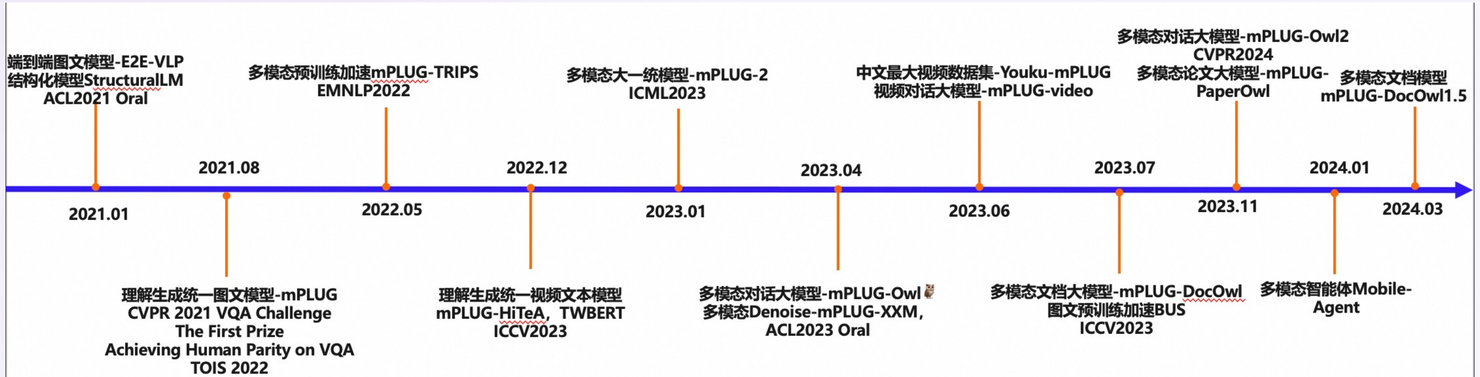
多模态手机智能体Mobile-Agent开源



https://modelscope.cn/studios/wangjunyang/Mobile-Agent/

ModelScope

通义mPLUG模块化多模态体系



ModelScope

通义mPLUG模块化多模态体系

MobileAgent Public

Mobile-Agent: The Powerful Mobile Device Operation Assistant Family

android · agent · harmony · ios · app · gui · automation · mobile · copilot · multimodal · + 5

Python · MIT License · 179 · 2.2k · 16 · 0 · Updated 15 minutes ago

mPLUG-DocOwl Public

mPLUG-DocOwl: Modularized Multimodal Large Language Model for Document Understanding

multimodal · table-understanding · document-understanding · mllm · multimodal-large-language-models · chart-understanding

Python · Apache License 2.0 · 64 · 1.1k · 27 · 1 · Updated 2 weeks ago

RoleInteract Public

RoleInteract: Evaluating the Social Interaction of Role-Playing Agents

Python · 4 · 34 · 1 · 0 · Updated 3 weeks ago

Multi-LLM-Agent Public

Python · 20 · 156 · 4 · 0 · Updated on Apr 23

mPLUG-Owl Public

mPLUG-Owl & mPLUG-Owl2: Modularized Multimodal Large Language Model

video · chatbot · dialogue · pytorch · transformer · llama · gpt · alpaca · visual-recognition · multimodal · + 10

Python · MIT License · 158 · 2k · 77 · 1 · Updated on Apr 6

SocialBench Public

SocialBench

Apache License 2.0 · 0 · 1 · 0 · 0 · Updated on Feb 15

mPLUG-HalOwl Public

mPLUG-HalOwl: Multimodal Hallucination Evaluation and Mitigating

benchmark · contrastive-learning · hallucinations · mllm · multimodal-large-language-models · multimodal-hallucination

Python · MIT License · 1 · 59 · 3 · 0 · Updated on Jan 29

https://github.com/orgs/X-PLUG/repositories

ngjunyang/Mobile-Agent/

ModelScope


下午5:21
XU
编辑个人资料
xuhaiyang-mPLUG
@xuhaiya2483846
Researcher @AlibabaGroup，Multimodal-mPLUG
github.com/X-PLUG
科技 出生于 1992年9月5日
2023年5月 加入
2,880 正在关注 506 关注者
帖子 回复 亮点 文章 媒体
你已转帖
xuhaiyang-mPLUG @xuhai... · 23年5月9日
🔥 [05.05] We released code and dataset for instruction tuning!
🔥 [05.05] Online demo on HuggingFace is available. Thank Huggingface for free computing resources!
🔥 [05.05] Online demo on HuggingFace now supports recieve video!
🔥 [05.05] We upload evaluation set OwlEval!
AK @_akhaliq · 23年5月4日
mPLUG-Owl🦉: Modularization Empowers Large Language Models with Multimodality
@Gradio demo is out on @huggingface
demo: huggingface.co/spaces/MAGA


下午5:22
回答
知乎
徐海洋-mPLUG
阿里通义实验室-多模态mPLUG
阿里巴巴 从业人员
45 创作
1161 赞同
335 感谢
分享二维码名片

ModelScope