

GUI-Agent: 研究、应用与展望

塔杰¹, 王亮², 刘进², 黄勃³, 王一鹤²

(1. 西藏自治区科技信息研究所, 西藏拉萨 850008; 2. 武汉大学计算机学院, 湖北武汉 430072;
3. 上海工程技术大学电子电气工程学院, 上海 201620)

摘要: GUI-Agent 作为人机交互的新范式, 在日常生活中扮演着越来越重要的角色。借助多模态大模型强大的模态交互能力, 其能够模拟人类用户与数字系统的交互方式, 执行点击、滑动、文本输入以及更复杂的任务。从 GUI-Agent 发展历程出发, 深入剖析其架构基础、优化策略、测试基准和评价指标, 搜集了常见应用场景中的实际应用案例, 旨在为相关研究人员提供一个更细粒度的综述, 帮助其快速、直观地了解 GUI-Agent 及其研究进展, 并总结了当前面临的问题与挑战, 讨论了未来的发展方向, 以期促进该领域的进一步发展。

关键词: 图形用户界面智能体; 大模型; 人机交互; 智能决策

DOI: 10.11907/rjdk.251140

中图分类号: TP181

文献标识码: A

开放科学(资源服务)标识码(OSID):

文章编号: 1672-7800(2025)008-0001-10



GUI-Agent: Research, Application and Prospect

TA Jie¹, WANG Liang², LIU Jin², HUANG Bo³, WANG Yihe²

(1. Institute of Science & Technology Information of Tibet Autonomous Region, Lasa 850008, China;

2. School of Computer Science, Wuhan University, Wuhan 430072, China;

3. School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: GUI-Agent, as a new paradigm of human-computer interaction, plays an increasingly important role in daily life. With the powerful modal interaction capabilities of the multimodal large model, it is able to simulate the way human users interact with digital systems, performing clicks, swipes, text input, and more complex tasks. This paper starts from the development history of GUI-Agent, deeply analyzes its architecture foundation, optimization strategy, test benchmark and evaluation index, and collects practical application cases in common application scenarios. This paper aims to provide a more granular review for relevant researchers, help them quickly and intuitively understand the current research progress of GUI-Agent, summarize the problems and challenges it faces, and discuss the future development direction to promote the further application and development of related fields.

Key Words: GUI-Agent; large model; human-computer interaction; intelligent decision

0 引言

图形用户界面智能体(Graphical User Interface Agent, GUI-Agent)作为人机交互领域的革命性技术,正通过自动化操作与智能决策重构数字生产力。从银行系统自动填单到医疗影像系统智能标注,从工业控制界面异常检测到跨平台办公流程自动化,GUI-Agent已在金融、医疗、制造等20余个行业实现规模化应用。

Agent可在生活的很多方面发挥其独有的作用。例

如,在工作中往往需要查阅大量资料,或者寻找多年前的工作报告。没有Agent时,用户通常需要花费大量时间阅读所有文档,或在互联网上搜索相关信息,从中筛选出有用的信息并汇总,该过程可能会重复多次。最后,再设计一份PPT用于工作汇报。而如今在Agent的帮助下,如图1所示,也许只需要输入一行指令:“阅读这些资料,汇总相关信息,总结为PPT。”

本文聚焦GUI-Agent的技术演进路径、核心支撑技术体系,以及大模型赋能的智能增强方向,探讨其未来技术突破与产业落地前景。具体来说,本文首先明确基于大模

收稿日期: 2024-03-01

扫描二维码阅读全文:

基金项目: 西藏自治区自然科学基金项目(XZ202401ZR0119)

作者简介: 塔杰(1991-),男,西藏自治区科技信息研究所助理研究员,研究方向为软件工程、人工智能等。



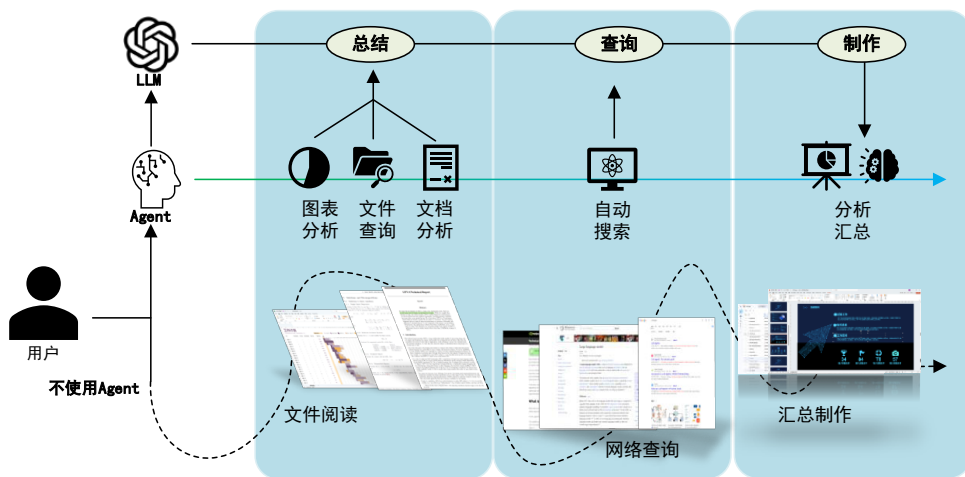


Fig. 1 Application of GUI-Agent in work

图1 GUI-Agent在工作中的应用

型的 GUI-Agent 的定义与结构,并介绍 GUI-Agent 工作的通常流程;然后总结不同的优化策略,给出当前常见的数据集与环境及其对应的评价指标;最后说明 GUI-Agent 的实际应用与存在的缺陷,并讨论 GUI-Agent 未来的发展方向与尚未解决的问题。

1 相关介绍

1.1 图形界面智能体原理

GUI-agent 是一种用于辅助与图形用户界面进行交互的人工智能代理,其能够模拟人类操作,通过自动化执行

任务实现与软件应用的互动,提升工作与学习效率^[1]。当前的 GUI-Agent 系统通常采用包含感知理解、规划决策和定位执行的 3 层架构体系^[2]。一般来说,GUI-agent 首先截取界面图像,通过图像识别、光学字符识别(Optical Character Recognition, OCR)等相关技术将其分解为可识别的元素。之后 GUI-agent 对识别到的元素进行分类,锁定目标,并根据预先设定的任务逻辑生成相应的操作指令序列。GUI-agent 基于相关指令执行具体操作,重复一系列用户操作,并监测执行结果以判断任务是否需要调整或重试,如图 2 所示。该交互模式使 GUI-agent 能够适应软件界面的变化,通过不断学习和适应实现更高的自动化水平。

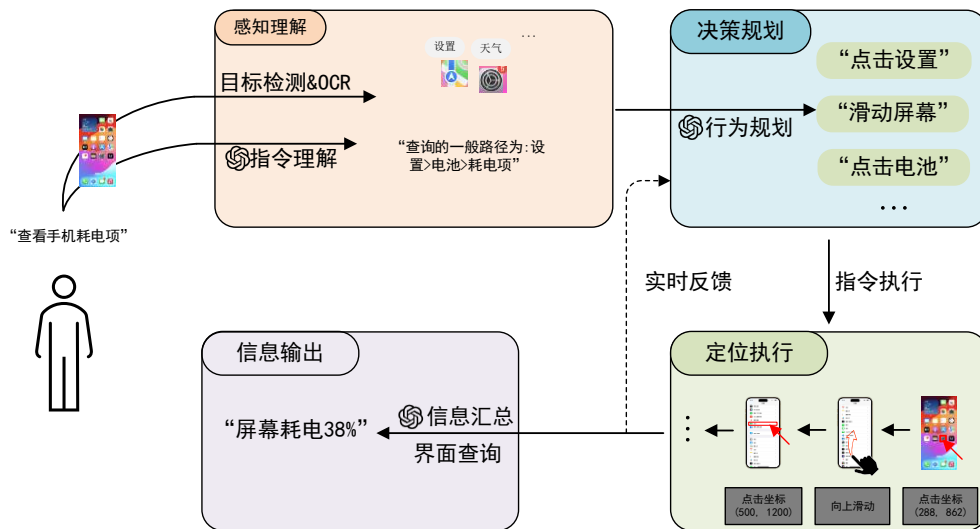


Fig. 2 GUI-Agent workflow

图2 GUI-Agent工作流程

1.2 图形界面智能体发展历程

图形用户界面(Graphical User Interface, GUI)长期以来一直是人机交互的核心。与传统命令行交互不同的是,其提供了一种以视觉驱动的直观方式帮助用户与数字系统进行交互,极大地降低了使用门槛^[3]。然而,随着技术的发展与互联网的普及,用户需要花费越来越多时间用于

掌握与使用工具界面,操作的繁琐性与复杂性逐渐显现。因此,GUI的交互自动化需求日益增长。

早期的 GUI 自动化研究主要聚焦于基于脚本或规则的方法,如 VBA (Visual Basic for Applications) 编辑器中的宏录制工具,能够实现对固定流程的模拟,简化在 Office 应用程序中的操作^[4-5]。该方法通过预定义的静态规则和逻辑

辑来执行具体任务,在很多简单、重复性强的任务上十分高效,且易于实现和理解,至今仍有一定的应用价值^[6]。但缺点在于其通用性较差,需要根据具体使用场景切换不同的脚本和API(Application Programming Interface)。而对于大量第三方开发的中小型应用和网页,逐个接入API并不现实。同时在任务难度上升时,需要进行大量的规则定义,提升了系统的复杂度,降低了可靠性。基于脚本语言(如Python、Java等)的Agent有着更高的灵活度,也能够处理更复杂的任务,但仍难以满足实际应用的需求^[7]。

21世纪以来,随着以图像识别和自然语言处理为代表的机器学习技术的兴起,Agent开始从基于规则的系统向着基于学习的系统转变^[8]。这种新方法能够处理更复杂的任务,并学习用户的交互逻辑,具备相当强大的任务泛化能力,在动态环境中也有着不错的表现。如Xu等^[9]提出的RUSS(Rapid Universal Support Service),将自然语言映射到特定的指令语言ThingTalk,检索Web页面上的实际元素ID。实验证明RUSS能够以76.7%的准确率执行从自然语言指令到Web操作的映射,表现出在开放环境中的应用潜力。

随着大语言模型的发展与广泛应用,GUI Agent具备了来自大模型的语义理解与强大的推理能力,作为节省时间与学习成本的工具登上时代舞台^[10-11]。因其具有直接与图形界面进行交互的特点,免去了对特殊API的需求,泛用性获得了极大的提升^[12]。但在GUI发展日益迅猛的当下,当前的识别方法对图形界面理解的准确度差强人意。因此,迭代开发新的GUI Agent势在必行。随着新的大型多模态模型的问世与发展,模型对于自然语言和UI元素的处理能力有了质的飞跃^[13-14]。这使得GUI Agent可以不使用或少使用后端API,而依靠对UI界面的识别完成更复杂的代理任务^[15]。

1.3 图形界面智能体结构

本文将当前主流的GUI-Agent系统定义为三级架构范式,其中的技术演进过程展示了从规则驱动到认知增强的范式转变。本节从感知理解、规划决策与定位执行3个维度展开论述。

1.3.1 感知理解模块

感知理解模块负责理解用户指令与解析界面,为后续规划决策提供基础。由于可将常见的语音/文本指令编码为结构化语义,大语言模型(如GPT系列^[11])已成为相当泛用的解决方案,而界面解析技术则经历了更复杂的发展过程^[16]。

图形界面理解相关技术经历了多次迭代。在语言大模型时期,Agent主要依靠纯文本形式的语言大模型,使用HTML或文档对象模型(Document Object Model, DOM)进行网页内容识别。对于其它操作环境,如PC中的其他应用和智能手机等,UI通过截图的方式传输到Agent中,并通过OCR等技术进行文本识别,再交给任务规划器^[17]。

OCR方法对文字的识别能力强,但丢失了布局和图标等诸多信息,精确度受限^[18]。随着相关技术的发展,研究人员开始利用新的大型多模态模型(LMMs)作为工具,不再依赖OCR方法,而是将截图直接输入到Agent中,通过对UI界面的整体识别直接得到文本信息和图标元素,并保留全部的布局信息^[19-21]。CoCo-Agent提出综合环境感知(Comprehensive Environment Perception, CEP)方法,结合历史截图和动作,以促进GUI感知多方面细粒度的信息^[22]。CogAgent提出视觉和语言双分支结构,将视觉元素映射到对应的文本标签上,增强了对用户意图的预测和执行能力,在降低计算成本的同时提升了对UI界面的理解能力^[23]。AutoWebGLM提出一种受人类浏览模式启发而开发的HTML简化算法,能够准确地保留重要的文本信息,通过删除冗余和无效信息提高智能体对网页的理解和交互能力,从而提升Agent执行任务的成功率^[24]。DUAL-VCR认为UI界面的设计总是存在内在的联系,并将每个Web元素与其在屏幕上的相邻元素进行关联,从而增强了Agent对UI界面的理解和交互能力^[25]。Ferret-UI增加了放大局部细节的Any-res技术,有效提升了对于细节处图标的识别理解能力^[26]。PTP(Patch-and-Text Prediction)通过大量消融实验探讨了掩码方法在界面感知中的应用潜力^[27]。上述方法对Agent运行流程中的不同步骤进行深度思考并优化,向着建立高实用性、强操作能力的Agent的目标推进,极大促进了Agent系统感知理解能力的提升。

感知理解领域相关工作通常采用模拟人类感知方式的注意力方法,将研究重点放在寻找与识别关键图标和区域上,从而改善大量信息造成的感知延迟与理解困难问题。但值得注意的是,非关键区域并非毫无意义。充分权衡重点区域与非重点区域的颗粒度,并选用更适配的基础模型,能够帮助GUI-Agent更好地感知与理解当前环境,进一步提升Agent系统对当前状态感知的准确度与定位能力。

1.3.2 规划决策模块

规划决策模块将用户需求与当前的环境和知识结合,生成具体操作流程,通过任务分解和资源调配优化操作步骤,提高执行效率。在充分理解用户意图与当前状态后,另一个LLM将综合当前状态与已有知识生成具体执行计划,并根据其他模块的反馈实时调整计划,保证指令能够准确地执行。规划决策过程至关重要,其决定了用户指令执行成功与否以及执行效率。

规划涉及将全局任务分解为多个子任务,这些子任务从初始状态开始逐步接近目标状态。同时,决策器提供的反馈也会影响任务规划器的推理过程。因此,需要引入记忆与检索模块,通过对人类神经系统记忆方式的模拟,能够显著增强Agent系统模拟人类进行人机交互的能力^[28-29]。RUSS设计了领域特定语言ThingTalk,将自然语言指令映射到ThingTalk语句,即能使Agent理解指令并执

行^[9]。SYNAPSE 利用行动记忆与动作相似性检测,存储和检索过去的行动序列以避免重复的规划计算,同时检测当前任务与历史任务间的相似性,从而快速找到最优的行动方案,有效降低了行动规划成本,显著提升了完成率^[30]。CoCo-Agent 提出条件动作预测(Conditional Action Prediction, CAP)方法,将动作预测分解为动作类型预测和以动作类型为条件的动作目标,实验结果证明了该方法的有效性^[22]。WebAgent 将复杂的任务分解为一系列对应单个指令的子任务,使 Agent 能够更好地理解与执行用户指令,并能够利用历史内存中的信息优化执行效率^[31]。Auto-GUI 开发了一种与思维链^[32]相似的行动链技术,该技术利用一系列行动历史记录和未来行动计划来辅助决策,模拟人类的思考方式,利用历史经验来指导当前的决策,能够充分改善信息丢失与错误问题,并增强了对抽象复杂指令的执行能力^[15]。AutoWebGLM 构建了强化学习框架,并提出基于拒绝采样的策略蒸馏算法(Rejection Sampling-based Policy Distillation, RS-PD)进行持续学习,在多项基准测试中证明了其对灾难性遗忘问题的改善效果,有效提升了规划决策能力^[24]。MM-Navigator 强化了反思模块,允许 Agent 在发现错误后回溯到前一步状态重新进行规划,从而提高了任务执行的准确性和成功率^[33-34]。WebVoyager 引入额外的外部批评者(GPT-4V)用于对规划进行自动化评估,首次实现了无需人工干预的交互式决策优化,证明了该策略的有效性。SEEACT 以文本形式生成计划,并将文本计划对应到 HTML 元素后执行操作,扩充了决策选项,可从更多维度进行判别^[35]。ScreenAgent 受体验式学习周期理论的启发,设计了优化的反思模块,使反思过程更接近人类的行为和思维过程^[36-37]。Cradle 学习用户的真实操作,并使用包含自我反思、任务推理、技能策划和动作规划等 4 个模块的决策规划器推断当前最合适的任务,部署在复杂的 3A 级游戏《荒野大镖客 II》中成功完成复杂的主线剧情,证明了该框架的通用性与潜力^[38]。AgentS 利用检索增强生成(Retrieval-augmented Generation, RAG)方法使代理能够获得更新、更可靠的外部知识辅助进行行为规划,规避了内部知识库与外部真实环境的不匹配^[39-40]。AWM (Agent Workflow Memory) 模拟人类对过去工作的总结思维,探索了工作流归纳方法,从过去执行任务的经验中提取可重用的工作流,并将其整合到 Agent 的内存中以指导未来的任务处理过程^[41]。

当前任务规划领域的相关研究主要利用持续学习或强化记忆的方法应对灾难性遗忘问题,对于不存在于训练数据中的领域外知识问题,通常使用检索增强生成方法对专业领域知识进行强化;对于决策反思问题,则通常选择模仿人类的行为逻辑对历史行为进行整合记忆,并在失败时进行多次尝试,积累经验以更好地完成后续工作。在 Agent 相关领域的技术研究中,规划决策模块并不直接参与端到端的交互,但负责整合端侧信息,因此有着至关重

要的意义与较高的研究价值。

1.3.3 定位执行模块

定位执行模块将接收到的操作指令实际部署到设备中,通过精确的元素定位和操作执行,确保指令的准确实施,完成用户任务。执行模块以细的粒度(通常精确到像素级坐标)表达决策模块通过的操作,同时需要处理更高级别的语义操作,例如键入文本、滚动或单击特定元素。

由于 UI 界面的高度复杂性,基于文本表示的执行策略易受到冗余和噪声影响^[42]。WebAgent 使用 DOM 或 HTML 表示来定位交互元素^[31]。SEEACT 提出基于元素、基于文本和基于图像的 3 类执行策略,深入讨论了不同执行策略的表现,分析了不同策略的性能与局限性。从实验结果来看,基于文本的方法在所有测试场景和指标上均表现最佳,接近监督微调,并显著优于仅使用文本的语言大模型。基于元素的方法由于依赖启发式的元素定位策略,在不包含文本的网页中表现较差。基于图像的方法在处理具有丰富语义和空间关系的网页截图时常出现严重的幻觉,影响定位的准确性。MobileAgent 将具有可访问性的 API 与 UI 截图相结合来识别移动平台上的 GUI 组件,展现了综合性方法的优势^[43]。MM-Navigator 利用多模态大模型生成下一步操作的文本输出,并将标签与界面中的元素相结合,辅助进行操作定位^[44]。

定位执行模块尽管在功能上相对简单,但考虑到用户体验,需要在准确度和速度之间进行抉择。同时,在实际使用过程中,还需要面对如机器人检测等阻碍。其作为落地实用的“最后一公里”问题,会直接影响着 Agent 用户的感受,因此在产品设计与相关研究中都至关重要。

1.4 图形界面智能体优化策略

大型基础模型在广泛的领域发挥了巨大作用,也有着优异表现,但在专业领域和特殊应用背景下的测试结果仍难以令人满意。本节对 Agent 在真实场景的应用以及更进一步的多智能体协同与领域模型进行了研究。

WebVoyager 与真实的专业网站建立联系,使用动态的真实案例对 Agent 进行训练,从而能够真实地与现实世界的使用场景进行交互^[45]。进一步的,OS-Copilot 构建了 FRIDAY,一个能够完成一般计算机任务的自优化 Agent,在进行行为规划时会首先针对任务进行内存检索,搜索合适的工具。若工具不存在或不合适则生成新工具,并将最终的选择存入内存,等待下一次的调用^[46]。Mobile-Agent-v2 引入多路径决策规划,对于多模态数据,使用 ConvNextViT-document 进行文本识别,Qwen-VL-Int4 进行图标描述,而纯文本信息则由 GPT-4 进行处理。

WebAgent 提出针对 HTML 文档的领域专家模型 HTML-T5 用于任务规划和条件 HTML 摘要生成,并结合 Flan-U-PaLM 进行多 Agent 合作^[31]。消融实验验证了该方法在处理真实网站任务时,相较于单一的 Agent,显著提升了任务的成功率。UFO (UI-Focused agent) 采用双代理框

架实现跨应用导航与操作,其中 HostAgent 负责选择应用程序来满足用户请求,AppAgent 负责在选定的应用程序上迭代执行操作,直到任务在特定应用程序中成功完成^[47]。Cradle 设计了多个模块,分别由一个 Agent 独立执行,有效增强了 Agent 的界面理解与任务规划能力,对于长期和复杂任务的执行效果更加显著。

Octopus 系列专注于解决专业化和本地轻量化问题,通过封装专业知识和特定功能作为外部知识库,利用多 Agent 协助工作,并探讨了云设备协作的落地问题^[48]。CoAT 引入深度的动作思维,考虑每个步骤产生的中间过程的影响与逻辑,需要执行哪些动作以及所选动作导致的结果,其中包含:屏幕上下文、动作思考、动作目标和动作结果,同时构建了第一个也是最大的细粒度数据集 AITZ^[49]。Agent-E 将计划形式化为有向无环图,其中每个节点代表一项任务,箭头代表任务之间的相互依赖关系,以确保任务按照正确的顺序进行,并利用蒸馏方法对 UI 界面进行简化,从而提高交互的效率和准确性^[50]。OS-Copilot 识别与分解复杂任务得到子任务间的关系,构建有向无环图,并行执行多个子任务,从而显著提高任务执行效率,并使用动态调度策略优化资源利用率^[46]。IOA (Internet of Agents) 通过引入智能体集成协议和即时通讯式的架构设计,尝试在多种不同的设备与环境进行泛用性的设计,实现异构智能体间的无缝协作^[51]。MobileFlow 引入 MoE 进行专业化的 Agent 设计^[52-53]。SheetCopilot 使用 Agent 进行 Excel 表格设计与制作^[54]。VideoGUI 认为应专注于解决用户执行时具有挑战的任务,如使用 GUI-Agent 进行更高级的规划任务,让 Agent 使用 Adobe Photoshop 对图片进行调整等^[55]。

2 测试基准与评价指标

2.1 图形界面智能体测试基准

训练数据是模型训练的基石,优质且针对任务的数据集能够显著提升 GUI-Agent 的性能表现。基于不同条件的测试基准能够帮助确定技术的最优化方向和潜在的改进空间。针对 GUI-Agent 繁杂的环境与复杂的流程,研究人员设计了大量基准测试。本文聚焦于 GUI-Agent 相关的数据集和测试方法。

随着 GUI 在人机交互中的广泛应用,自动化和智能化的 GUI 操作需求日益增长。为了有效评估和提升 GUI-Agent 的性能,高质量的数据集已成为研究的核心资源。一般来说,GUI-Agent 数据集包含用户指令、环境信息和操作序列,这些要素共同构成了训练模型所需的基础数据。同时,优质的数据集不仅能够提升模型的学习效果,而且能帮助研究人员识别技术的优化方向和潜在的改进空间。

MT-Mind2Web 是经典的网页端 Agent 测试基准,其中包含来自 137 个网站的 2 000 多个开放式任务,涵盖 31 个

不同领域,且每个任务都包含了用户与网站交互的具体行为序列^[56]。同时提出了对话式网页导航任务,该任务要求 Agent 在复杂的基于网络的环境中与用户进行多轮交互,并执行多步骤的指令,测试 Agent 处理和利用历史信息,以及在上下文限制条件下理解和执行用户连续指令的能力。WEBLINX 则针对执行多轮对话的真实环境 Web 导航任务,数据集中包含 2 337 个专家演示的对话式网页导航基准,覆盖了 155 个真实世界的网站和多种任务类别^[57]。同时提出了基于动作的评价指标,包括意图匹配 (Intent Match, IM)、元素相似度 (Element Similarity, ES) 和文本相似度 (Text Similarity, TS),并根据动作类型分为元素组和文本组进行分组评估,进一步细化了评估标准。Visual-WebBench 包含 7 类跨度很大的任务,涉及 1 500 个人工策划的实例,覆盖多个领域和子领域,并正式探讨了多模态大模型在相关领域的应用现状与前景,旨在全面评估多模态大模型的网站级、元素级和动作级理解能力,为后续研究打下基础^[58]。

在移动设备端,AndroidControl 是最著名的测试基准之一,包含 15 283 个在 Android 应用中完成日常任务的演示,涵盖了 833 个不同的 Android 应用,并且每个任务都提供了不同复杂度的人类指令,使研究人员能够探索模型在不同任务复杂度下的表现^[59]。标注者还在执行每个操作前提供低级别的自然语言描述,从而生成包含高级别和低级别指令的数据集。而 MobileAgentBench 定义了 100 个移动设备上的任务,覆盖 10 个开源应用程序,并按难度级别分类,支持在真实设备上进行测试,并实时判断任务状态,具有更强的实时性^[60]。GUI Odyssey 中包含 7 735 个实例,涵盖 6 种类型的跨应用程序任务,共计 201 个不同的应用程序,以及 1 399 个应用程序组合的任务,专门用于训练和评估移动设备跨应用程序的 Agent^[61]。

在桌面端方面,GUI-World 数据集中包含超 12 000 个屏幕录像和 YouTube 上的指导视频,专门用于评估和提高多模态大模型的 GUI 理解能力,涵盖一系列类别和场景,包括网站、桌面和移动应用程序、多窗口交互以及扩展现实环境,支持跨各种操作系统的任务设置、基于执行的评估和交互式学习^[62]。GUICourse 认为多模态大模型缺少对 GUI 元素功能和控制方法的全面理解,因此其提出的数据集分为 3 部分:GUIEnv、GUIAct 与 GUIChat^[63]。GUIEnv 数据集用于增强多模态大模型的 OCR 和定位能力,GUIAct 数据集用于丰富多模态大模型对 GUI 组件和交互的知识,GUIChat 数据集用于提升 GUI-Agent 的交互技能。通过 3 个数据集的交叉验证,能够更全面地对 Agent 进行评估。Spider2-V 针对解决数据科学和工程工作流程自动化问题设计,提出综合评估多模态智能体在编写代码和控制 GUI 方面的能力,包含了 494 个真实世界的任务,涵盖了 20 个企业级专业应用,是专业型测试基准的代表之一^[64]。OmniACT 中包含超过 9 800 对屏幕图像和自然语言任务描述

的数据集,覆盖了多种操作系统和网络应用,旨在通过结合视觉和语言理解来提高代理的多模态理解能力^[65]。Crab 提出支持跨环境任务的 Agent 基准测试框架,并结合了基于图的细粒度评估方法和图像评估器构建机制^[66]。通过将任务分解为多个子目标,并为每个子目标分配一个判断函数来检查其完整性,从而提供细粒度的评估。PTP 提出通过掩码并恢复截图中图像块与文本的方式来同时提升模型对视觉特征和语言理解的基准测试框架^[27]。LlamaTouch 认为现有评估方法依赖于人工验证或已建立的数据集将代理预测的动作与预定义的动作序列进行比较,不可扩展且不可靠^[67]。LlamaTouch 采用仅评估代理是否遍历了所有手动注释的基本应用程序/系统状态的方式进行评估。VAB(Visual Agent Bench)建立了一个包含 4 482 个真实轨迹的综合性测试平台,旨在更好地评估 Agent 在视觉层面更高层次的推理和规划能力,以及 Agent 在多任务并行环境下的表现^[68]。

此外,还有一些研究人员发现静态的数据集和测试基准难以满足飞速发展的 GUI-Agent 的需求,因此开发了能够与真实世界交互并更新,或更接近真实环境的数据集和环境。

WebCanvas 提出一个在线评估框架,用于动态、实时地评估 Agent 在每个关键动作节点是否完成任务了所需的关键中间动作或状态,并提出一个包含 542 个任务的实时在线基准测试数据集。该数据集中的每个任务都有人类专家对关键节点验证的注释,从而更好地验证 Agent 的表现^[69]。MMInA 通过在真实世界的网站(包括购物、旅行等领域)上操作,确保了结果的真实性和适用性^[70]。该研究提出一个包含 1 050 个任务的多跳多模态基准,涵盖多个领域,每个任务都需要 Agent 从网页中自主提取信息作为观察结果。此外,MMInA 还提出了记忆增强方法,通过重播过去的动作轨迹进行反思,从而优化后续动作。Visual-WebArena 中包含专门设计用于评估自主多模态 Agent 的多样化基准测试工具,包含分类广告、购物和 Reddit 网站相关任务,并对当前最先进的大型语言模型(LLM)和视觉—语言模型(Visual Language Model, VLM)进行了评估^[71]。

AndroidWorld 创建了一个动态并具有完整功能的 Android 环境,用于开发和评估 Agent。该环境提供了对 20 个 Android 应用的访问,并能够动态构建任务,从而允许在更广泛和现实的任务集合上进行测试^[72]。B-MoCA 在提供 60 个 Android 操作系统日常任务的同时引入了壁纸,通过对语言图标位置与大小的随机化调整,能够更好地评估 Agent 的泛化能力^[73]。AgentStudio 提供了一个轻量级的交互式环境,具有高度通用的观察和动作空间,可进行视频观察和 GUI/API 操作^[74]。其集成了用于创建在线基准任务、注释 GUI 元素和标记视频的操作工具。针对当前 Agent 相关研究缺乏系统化的基础设施来构建和测试代理,以及真实世界任务用于全面评估 Agent 基本能力的问

题,OS-World 构建了一个涵盖大部分桌面操作系统的评估平台,包含 369 个真实世界的常见任务,弥补了先前研究中对图像用户界面定位和操作方面的缺陷,帮助 Agent 能够更好地适应计算机系统的复杂性^[75]。

AER(Agent-Eval-Refine)提供了一个外部的自动评估器,无需依赖人类演示或专家即可对用户指令和代理行为轨迹进行评估并反馈,并探索了端到端评估与模块化评估的特点,通过推理时的策略优化和过滤行为克隆,能够显著提升代理的性能^[76]。CoQuest 对人机交互进行了理论建模,并通过实验证明了设计模式对用户体验的影响^[77]。在 Agent 落地工业界的前提条件下,与之相对应的特定环境下的 Agent 评估方法也应受到重视。

2.2 图形界面智能体评估指标

在 Agent 训练过程中,为了使决策器能够更好地进行判别,同时使研究人员更直观地了解实验结果以便后续进行调整,有效、直观的测试指标是必不可少的。近年来实验常用的测试指标见表 1。

Table 1 Graphical interface agent evaluation indicators

表 1 图形界面智能体评估指标

测试指标	内容
成功率	忽略中间过程,如果 Agent 完成指令,则视为成功,否则视为失败
步骤得分	Agent 成功到达关键节点并通过相关评估函数验证时,视为成功
步骤完成率	衡量成功完成任务需要的步骤数与任务总步骤数之比
决策准确度	在 Agent 执行过程中,所有决策中正确决策的比例
反思准确度	在 Agent 执行过程中,所有反思中正确反思的比例
相对效率	与其他解决方案(如人类操作)相比消耗的时间以及其他资源之比
完成率	Agent 能够完成的人类操作步骤数与人类采取的总步骤数之比
风险比率	Agent 在任务执行过程中的行为相关潜在风险大小
回合成功率	Agent 在交互过程中成功处理请求的轮次与所有轮次的比例

GUI-agent 的性能和有效性评估在确保其在复杂用户界面环境下的功能性与可靠性至关重要。同时,作为一种自动化工具,GUI-agent 的表现不仅由其任务执行的准确性与效率决定,而且取决于其能否灵活应对不同用户需求以及资源利用率等方面。因此,定义和选择合适的评估指标(Metrics)是衡量 GUI-agent 系统效率、稳定性及用户满意度的关键。

由于 GUI-Agent 可应用的范围广,环境十分多样,各研究针对的问题也有明显差异,学术界短期内并没有形成如计算机视觉领域的 ImageNet 一般具有较强代表性的评价标准。本文总结了部分具有共通性的工作,以 GPT-4 或 GPT-4V 作为 Agent 所使用的基础模型,尽可能客观且直观地对比展现这些工作的实验成果与研究重心,以期未来的研究人员提供思路,具体见表 2。

除这些固定的测试指标外,最近研究人员还创新性地引入了外部大型基础模型(如 GPT-4V)作为额外的评估器。利用这些先进模型强大的多模态理解能力,能够对

Table 2 Display of partial work experiment results
表 2 部分工作实验结果展示

benchmark	环境	评估标准	内容	平台	GPT-4 (GPT-4V) 成功率/%	特点
MT-Mind2Web	静态	元素准确率,操作 F1 分数,步骤成功率,任务成功率	137 个网站的 2 350 个任务	网页	40.20	使用真实的网站快照,提供未经过滤的动态数据
OmniACT	静态	序列得分,行动得分,成功率	200 多个桌面和网页上共 9 802 个数据点	桌面 网页	32.75	包含需要长期规划的复杂任务,人类用户的成功率为 82.23%
VisualWeb Arena	静态	成功率	314 个模板的 910 个任务	网页	16.37	通过视觉任务测试 Agent 的视觉理解和动作执行能力,人类用户的成功率为 88.70%
OSWorld	Ubuntu	成功率	Ubuntu 中的 369 个任务,其中 43 个可以在 Windows 上使用	桌面	12.24	提供了真实的交互环境,人类用户的成功率为 72.36%
MMInA	静态	成功率	跨网站的 1 050 个任务	网页	19.85	包含需要在多个网站跳转的复杂多步骤推理,人类用户的成功率为 96.25%
GUI Odyssey	静态	动作匹配得分,成功率	6 个移动设备的 7 735 个任务,以及 1 399 个应用程序组合	手机	14.04	针对需要跨应用的复杂任务
GUI-World	静态	外部模型(如 GPT-4)评估,成功率	6 种场景共 8 种类型任务,共 12 379 个视频	桌面 网页 手机	83.50	引入外部大模型作为评估标准
VisualWeb Bench	静态	成功率	139 个网站的 1 534 个任务	桌面 网页 手机	64.60	针对具备结构化布局和多样化交互元素的网页设计
VideoGUI	静态	成功率	86 个复杂任务和 463 个子任务	网页 桌面	33.90	其中的完整复杂任务,Agent 几乎不可能完成
MobileAgent Bench	安卓	成功率,步骤效率,假阴性,假阳性	10 个开源应用程序的 100 个任务	手机	40.00	简化了测试流程,能够在真实设备上实时得到测试结果
AndroidWorld	安卓	成功率	20 个应用程序的 116 个任务	手机	30.60	第一个真实环境的动态测试基准,人类用户的成功率为 80%
WebCanvas	静态	成功率,效率分数	Mind2Web 中挑选并更新的 542 个任务	手机 桌面	23.10	采用高效维护策略的在线评估框架
Crab	静态	成功率,执行效率,成本效率	29 个安卓任务、73 个 Ubuntu 任务和 18 个跨平台任务,共计 120 个任务	手机 桌面 网页	13.30	跨环境的多智能体评估
Spider2-V	静态	成功率	20 个专业应用的 494 个任务	手机	14.00	专注于专业数据科学和工程工作流程的基准测试
AITW	静态	成功率,任务完成率,正确步骤比例,步骤效率	159 个 Android 应用和 198 个网站的 715 142 个任务	手机 网页	42.50	移动环境最大且最常用的测试基准
sheetcopilot	静态	准确率,总步骤数	221 个表格任务,共 1 669 条指令	Excel 桌面	55.00	针对表格场景开发
ScreenSpot	静态	动作匹配得分,成功率	23 个应用程序的 1 172 种元素	网页 手机	16.20	针对小图标识识别困难的问题开发
WebArena	静态	成功率	241 个模板的 812 个任务	网页	14.41	关注于评估任务完成的功能性与正确性,人类用户的成功率为 78.24%

GUI-agent 的执行过程进行深度评估,并以自然语言的方式输出评价意见。该方法不仅允许对 GUI-agent 的操作和决策进行更智能化的分析,而且能捕捉到传统指标可能忽略的细微差异和复杂场景。随着人工智能技术的不断进步,如何更好地利用这些评估器将成为未来重要的研究方向,从而实现更准确、全面的评估机制,促进 GUI-agent 技术的发展和应用。

在这些测试基准之外,还有很多极具想象力和前瞻性的工作,如 Cradle 使用游戏完成度作为评估基准,VideoGUI 则开发了对于人类用户成功率极低的测试基准作为对未来 Agent 工作的期待。除数据化的测试基准外,文献

[78]还提出以用户为中心的能够从更高维度对 GUI-Agent 进行评估的设计思路,以进一步推动相关领域的发展。

3 GUI-Agent 实际应用

GUI-Agent 目前已成功应用于工业领域,一些互联网龙头企业在其产品中加入 Agent 以优化用户体验。虽然仍有很多问题亟待解决,但已经能够为用户提供很大的帮助。腾讯的 AppAgent 有较强的复杂任务执行能力,且能够从用户的自主操作中记录所使用的 UI 元素和操作,以作进一步的学习探索。微软提出的 MM-Navigator 同样具备

持续学习的能力,并且允许 Agent 进行试错。若操作错误,将返回上一步状态并尝试新方法^[71]。Google DeepMind 团队开发了 WebAgent,使用多智能体协作进行持续学习,并具有一定的多平台泛化能力^[31,79]。以下是近年来提出的具有一定代表性的智能体:

(1)SkyAgents(昆仑万维)。帮助用户通过自然语言和简单操作,无需编程即可快速构建和部署自己的 AI 智能体,实现感知环境、决策和执行动作等功能。

(2)超级小爱(小米)。超级小爱在原有小爱同学的基础上进行了全面升级,提供了更加智能、便捷的语音交互体验。其能够进行屏幕内容提取,帮助用户找到关键信息,并对屏幕上的内容进行总结和精炼。同时具有全局执行能力,可直接为用户进行跨应用和设备的操作。

(3)YOYO 智能体(荣耀)。全新的 YOYO 智能体能够结合用户习惯,理解复杂指令,自动控制 APP 和系统,轻松完成任务。如在美团上进行指定外卖的下单并进行备注,或启动支付宝等 APP 为用户进行查询、续费等操作。

(4)DeepResearch(OpenAI)。帮助用户执行耗时的在线研究任务,可研究领域从复杂的科学问题到汽车推荐,且可以开展跨学科研究。DeepResearch 可以在 5~30 min 内完成人类数小时才能完成的复杂任务,显著提升工作效率。但有研究指出,其仍存在胡编乱造、回复与提问不相干等幻觉现象,且可能难以区分权威信息和谣言。

(5)PPTAgent(中国科学院大学)。其能够完成 PPT 制作任务。具体来说,其首先结合大语言模型对输入的文档进行解析,生成完整的大纲与布局。然后确定每页的内容,执行文本和图片的定点插入操作,生成最终的 PPT。该方法可应用于多个领域,有着很高的应用价值与较大的开发潜力。

(6)Manus(Monica)。采用包含 GUI-Agent 在内的多智能体协同作业,能够进行深度规划并执行较复杂的任务。在单次执行中自主进行信息查询与筛选,构建工程文件,并交付完整的成果。Manus 引领了工业界在智能体设计方向的风潮,带来了新的方向与思路,并在一定程度上影响了相关领域研究人员的研究重心。

总体而言,这些研究主要针对 Web 场景、个人计算机场景与移动智能设备场景。早期研究主要集中在对各自的领域进行专门的设计,随着研究的进一步深入和大型基础模型的发展,一些研究开始向着通用性 Agent 方向发展,致力于开发出能够在任意 GUI 平台进行高精度导航操作的 Agent,以实现更高的工作效率。

4 结语

GUI-Agent 的研究目前虽然取得了显著进展,但仍面临一些关键问题和挑战。首先是多模态能力需要提升,尽管多模态模型在 GUI 理解方面取得了进展,但将视觉信息

与语言信息进行有效融合仍然是一个难题。其次,任务推理和规划能力的提高也是亟待解决的问题,现有的 LLM 在复杂任务推理和计划方面仍存在不足,尤其是面对动态、不可预见的场景时,可通过多智能体方案提升系统的鲁棒性来解决此类问题。同时,现有的 GUI-Agent 系统还存在响应延迟与逻辑断层问题,可能的解决方案是强化学习、记忆增强和轻量化实时推理。此外,实时性和用户体验也是一项重要的挑战,用户对响应速度有着较高要求,而当前的 Agent 在处理复杂任务时往往难以满足这一需求。

本文聚焦 GUI-Agent 的发展历程,从应用角度出发,分析其具体构成与优化策略。通过对当前研究的分析,未来的工作将集中于多模态体系构建,针对跨平台的动态适配、长周期任务持续性以及安全边界建模三大问题作进一步深入研究。随着技术的不断发展,未来人们与数字系统的交互方式将会迈向新阶段,生活和工作方式将与 Agent 系统深度结合。

参考文献:

- [1] JANSEN B J. The graphical user interface[J]. ACM SIGCHI Bulletin, 1998,30(2):22-26.
- [2] LIN K Q, LI L, GAO D, et al. ShowUI: one vision-language-action model for generalist gui agent[DB/OL]. <https://arxiv.org/abs/2411.17465>.
- [3] SAMPATH H, MERRICK A, MACVEAN A. Accessibility of command line interfaces[C]//Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021:1-10.
- [4] IVANČIĆ L, SUŠA V D, BOSILJ V V. Robotic process automation: systematic literature review[C]//Proceedings of BPM Blockchain and Central and Eastern Europe Forum, 2019:280-295.
- [5] HELLMANN T D, MAURER F. Rule-based exploratory testing of graphical user interfaces[C]//Proceedings of 2011 Agile Conference, 2011:107-116.
- [6] BANERJEE I, NGUYEN B, GAROUSHI V, et al. Graphical user interface (GUI) testing: systematic mapping and repository[J]. Information and Software Technology, 2013,55(10):1679-1694.
- [7] QURESHI I A, NADEEM A. GUI testing techniques: a survey[J]. International Journal of Future Computer and Communication, 2013, 2(2):142.
- [8] SAID K S, NIE L, AJIBODE A A, et al. GUI testing for mobile applications: objectives, approaches and challenges[C]//Proceedings of the 12th Asia-Pacific Symposium on Internetware, 2020:51-60.
- [9] XU N, DU M, CAMPAGNA G, et al. Grounding open-domain instructions to automate web support tasks[C]//Mexico:2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2021.
- [10] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33:1877-1901.
- [11] AI L B Y, AI C, AI R. GPT-4 technical report[DB/OL]. <https://arxiv.org/abs/2303.09774>.
- [12] HUANG X, LIU W, CHEN X, et al. Understanding the planning of LLM agents: a survey[DB/OL]. <https://arxiv.org/abs/2402.02716>.
- [13] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation[C]//Proceedings of International Conference on Machine Learn-

- ing, 2021: 8821–8831.
- [14] WANG P, BAI S, TAN S, et al. Qwen2-VL: enhancing vision-language model's perception of the world at any resolution[DB/OL]. <https://arxiv.org/abs/2409.12191>.
 - [15] ZHANG Z, ZHANG A. You only look at screens: multimodal chain-of-action agents[DB/OL]. <https://arxiv.org/abs/2309.11436>.
 - [16] ZHANG C, HE S, QIAN J, et al. Large language model-brained GUI agents: a survey[DB/OL]. <https://arxiv.org/abs/2411.18279>.
 - [17] HAMAD K, KAYA M. A detailed analysis of optical character recognition technology[J]. *International Journal of Applied Mathematics Electronics and Computers*, 2016, 4: 244–249.
 - [18] NGUYEN T T H, JATOWT A, COUSTATY M, et al. Survey of post-OCR processing approaches[J]. *ACM Computing Surveys*, 2021, 54(6): 1–37.
 - [19] NIU R, LI J, WANG S, et al. Screenagent: a vision language model-driven computer control agent[J]. *International Joint Conference on Artificial Intelligence*, 2024, 38(4): 3460–3467.
 - [20] CHENG K, SUN Q, CHU Y, et al. Seeclick: harnessing gui grounding for advanced visual GUI agents[C]//*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024: 9313–9332.
 - [21] HUANG Z, CHENG H Z, PAN J, et al. Spiritsight agent: advanced GUI agent with one look[DB/OL]. <https://arxiv.org/abs/2503.03196>.
 - [22] MA X, ZHANG Z, ZHAO H. Comprehensive cognitive LLM agent for smartphone GUI automation [DB/OL]. <https://arxiv.org/abs/2402.11941>.
 - [23] HONG W, WANG W, LV Q, et al. Cogagent: a visual language model for GUI agents[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024: 14281–14290.
 - [24] LAI H, LIU X, LONG I L, et al. AutoWebGLM: a large language model-based web navigating agent[C]//*Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024: 5295–5306.
 - [25] KIL J, SONG C H, ZHENG B, et al. Dual-view visual contextualization for web navigation[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024: 14445–14454.
 - [26] YOU K, ZHANG H, SCHOOP E, et al. Ferret-UI: grounded mobile UI understanding with multimodal LLMs[C]//*Proceedings of European Conference on Computer Vision*, 2025: 240–255.
 - [27] GAO T, WANG Z, BHASKAR A, et al. Improving language understanding from screenshots[DB/OL]. <https://arxiv.org/abs/2402.14073>.
 - [28] SHINN N, CASSANO F, GOPINATH A, et al. Reflexion: language agents with verbal reinforcement learning[J]. *Advances in Neural Information Processing Systems*, 2023, 36: 8634–8652.
 - [29] HOU Y, TAMOTO H, MIYASHITA H. "My agent understands me better": integrating dynamic human-like memory recall and consolidation in LLM-based agents[C]// *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024: 1–7.
 - [30] ZHENG L, WANG R, WANG X, et al. Synapse: trajectory-as-exemplar prompting with memory for computer control[DB/OL]. <https://arxiv.org/html/2306.07863>.
 - [31] GUR I, FURUTA H, HUANG A V, et al. A real-world webagent with planning, long context understanding, and program synthesis[DB/OL]. <https://arxiv.org/abs/2307.12856>.
 - [32] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 24824–24837.
 - [33] CHEN Z, LIU K, WANG Q, et al. Mindsearch: mimicking human minds elicits deep AI searcher [DB/OL]. <https://arxiv.org/abs/2407.20183>.
 - [34] DENG Y, ZHANG X, ZHANG W, et al. On the multi-turn instruction following for conversational web agents[DB/OL]. <https://arxiv.org/abs/2402.15057>.
 - [35] ZHENG B, GOU B, KIL J, et al. GPT-4V (ision) is a generalist web agent, if grounded[C]//*Proceedings of the 41st International Conference on Machine Learning*, 2024: 61349–61385.
 - [36] BAECHLER G, SUNKARA S, WANG M, et al. Screenai: a vision-language model for ui and infographics understanding[DB/OL]. <https://arxiv.org/abs/2402.04615>.
 - [37] KOLB D A. *Experiential learning: experience as the source of learning and development*[M]. New York: FT Press, 2014.
 - [38] TAN W, ZHANG W, XU X, et al. Cradle: empowering foundation agents towards general computer control[C]//*Vancouver: NeurIPS 2024 Workshop on Open-World Agents*, 2024.
 - [39] AGASHE S, HAN J, GAN S, et al. Agent S: an open agentic framework that uses computers like a human[C]//*Vancouver: NeurIPS 2024 Workshop on Open-World Agents*, 2024.
 - [40] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 9459–9474.
 - [41] WANG Z Z, MAO J, FRIED D, et al. Agent workflow memory [DB/OL]. <https://arxiv.org/abs/2409.07429>.
 - [42] GOU B, WANG R, ZHENG B, et al. Navigating the digital world as humans do: universal visual grounding for GUI agents[DB/OL]. <https://arxiv.org/abs/2410.05243>.
 - [43] WANG J, XU H, YE J, et al. Mobile-agent: autonomous multi-modal mobile device agent with visual perception[C]//*Vienna: ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.
 - [44] YAN A, YANG Z, ZHU W, et al. GPT-4V in wonderland: large multi-modal models for zero-shot smartphone GUI navigation[DB/OL]. <https://arxiv.org/abs/2311.07562>.
 - [45] HE H, YAO W, MA K, et al. WebVoyager: building an end-to-end web agent with Large Multimodal Models[C]//*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024: 6864–6890.
 - [46] WU Z, HAN C, DING Z, et al. OS-Copilot: towards generalist computer agents with self-improvement[C]//*Vienna: ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.
 - [47] ZHANG C, LI L, HE S, et al. UFO: a UI-focused agent for windows OS interaction[DB/OL]. <https://arxiv.org/abs/2402.07939>.
 - [48] YANG J, DONG Y, LIU S, et al. Octopus: embodied vision-language programmer from environmental feedback [DB/OL]. <https://arxiv.org/abs/2310.08588>.
 - [49] ZHANG J, WU J, TENG Y, et al. Android in the zoo: chain-of-action-thought for GUI agents[DB/OL]. <https://arxiv.org/abs/2403.02713>.
 - [50] ABUELSAAD T, AKKIL D, DEY P, et al. Agent-E: from autonomous web navigation to foundational design principles in agentic systems[C]// *Vancouver: NeurIPS 2024 Workshop on Open-World Agents*, 2024.
 - [51] CHEN W, YOU Z, LI R, et al. Internet of agents: weaving a web of heterogeneous agents for collaborative intelligence[C]//*Vienna: The Thirteenth International Conference on Learning Representations*, 2024.
 - [52] NONG S, ZHU J, WU R, et al. Mobileflow: a multimodal LLM for mo-

- mobile GUI agent [C]// Vancouver: NeurIPS 2024 Workshop on Open-World Agents, 2024.
- [53] MASOUDNIA S, EBRAHIMPOUR R. Mixture of experts: a literature survey[J]. *Artificial Intelligence Review*, 2014, 42: 275–293.
- [54] LI H, SU J, CHEN Y, et al. SheetCopilot: bringing software productivity to the next level through large language models[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023:4952–4984.
- [55] LIN K Q, LI L, GAO D, et al. VideoGUI: a benchmark for GUI automation from instructional videos[C]// Vancouver: Proceedings of the 38th International Conference on Neural Information Processing Systems, 2024.
- [56] DENG X, GU Y, ZHENG B, et al. MIND2WEB: towards a generalist agent for the web [C]// International Conference on Neural Information Processing Systems, 2023:28091–28114.
- [57] LÜ X H, KASNER Z, REDDY S. WebLINX: real-world website navigation with multi-turn dialogue [C]// Proceedings of the 37th International Conference on Neural Information Processing Systems, 2024: 33007–33056.
- [58] LIU J, SONG Y, LIN B Y, et al. VisualWebBench: how far have multimodal LLMs evolved in web page understanding and grounding [DB/OL]. <https://arxiv.org/abs/2404.05955>.
- [59] LI W, BISHOP W, LI A, et al. On the effects of data scale on computer control agents[DB/OL]. <https://arxiv.org/abs/2406.03679>.
- [60] WANG L, DENG Y, ZHA Y, et al. MobileAgentBench: an efficient and user-friendly benchmark for mobile LLM agents[DB/OL]. <https://arxiv.org/abs/2406.08184>.
- [61] LU Q, SHAO W, LIU Z, et al. GUI Odyssey: a comprehensive dataset for cross-app GUI navigation on mobile devices[DB/OL]. <https://arxiv.org/abs/2406.08451>.
- [62] CHEN D, HUANG Y, WU S, et al. GUI-WORLD: a dataset for GUI-oriented multimodal LLM-based agents[C]//Singapore: International Conference on Learning Representations, 2024.
- [63] CHEN W, CUI J, HU J, et al. GUICourse: from general vision language models to versatile GUI agents [DB/OL]. <https://arxiv.org/abs/2406.11317>.
- [64] CAO R. Spider2-V: how far are multimodal agents from automating data science and engineering workflows?[C]// Vancouver: International Conference on Neural Information Processing Systems, 2024.
- [65] KAPOOR R, BUTALA Y P, RUSSAK M, et al. Omniaact: a dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web[C]// European Conference on Computer Vision, 2025: 161–178.
- [66] XU T, CHEN L, WU D J, et al. CRAB: cross-environment agent benchmark for multimodal language model agents[DB/OL]. <https://arxiv.org/abs/2407.01511>.
- [67] ZHANG L, WANG S, JIA X, et al. LlamaTouch: a faithful and scalable testbed for mobile UI automation task evaluation [C]// Pittsburgh: The 37th Annual ACM Symposium on User Interface Software and Technology, 2024.
- [68] LIU X, ZHANG T, GU Y, et al. VisualAgentBench: towards large multimodal models as visual foundation agents [C]// Singapore: International Conference on Learning Representations, 2024.
- [69] PAN Y, KONG D, ZHOU S, et al. WebCanvas: benchmarking web agents in online environments [C]//Vienna: International Conference on Machine Learning, 2024.
- [70] ZHANG Z, TIAN S, CHEN L, et al. MMInA: benchmarking multihop multimodal internet agents[DB/OL]. <https://arxiv.org/abs/2404.09992>.
- [71] KOH J Y, LO R, JANG L, et al. VisualWebArena: evaluating multimodal agents on realistic visual web tasks [C]// Vienna: ICLR 2024 Workshop on Large Language Model (LLM) Agents, 2024.
- [72] RAWLES C, CLINCKEMAILLIE S, CHANG Y, et al. Androidworld: a dynamic benchmarking environment for autonomous agents [DB/OL]. <https://arxiv.org/abs/2405.14573>.
- [73] LEE J, MIN T, AN M, et al. Benchmarking mobile device control agents across diverse configurations[C]// Singapore: International Conference on Learning Representations, 2024.
- [74] ZHENG L, HUANG Z, XUE Z, et al. AgentStudio: a toolkit for building general virtual agents [C]//Singapore: International Conference on Learning Representations, 2024.
- [75] XIE T, ZHANG D, CHEN J, et al. OSWorld: benchmarking multimodal agents for open-ended tasks in real computer environments[J]. *Advances in Neural Information Processing Systems*, 2025, 37: 52040–52094.
- [76] PAN J, ZHANG Y, TOMLIN N, et al. Autonomous evaluation and refinement of digital agents [J]. Philadelphia: Conference on Language Modeling 2024, 2024.
- [77] LIU Y, CHEN S, CHENG H, et al. How AI processing delays foster creativity: exploring research question co-creation with an LLM-based agent [C]//Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024: 1–25.
- [78] RAEES M, MEIJERINK I, LYKOURENTZOU I, et al. From explainable to interactive AI: a literature review on current trends in human-AI interaction [J]. *International Journal of Human-Computer Studies*, 2024, 189: 103301.
- [79] ZOU Q J, JIANG Y J, GAO B, et al. An overview of cooperative multi-agent deep reinforcement learning [J]. *Aero Weaponry*, 2022, 29 (6) : 78–88.
- 邹启杰, 蒋亚军, 高兵, 等. 协作多智能体深度强化学习研究综述[J]. *航空兵器*, 2022, 29(6):78–88.

(责任编辑:黄健)