GUI Agents with Foundation Models: A Comprehensive Survey

基于基础模型的 GUI 代理: 全面综述

Shuai Wang[1], Weiwen Liu[1], Jingxuan Chen[1], Yuqi Zhou[2], Weinan Gan[1], Xingshan Zeng[1], Yuhan Che[1], Shuai Yu[1], Xinlong Hao[1], Kun Shao[1], Bin Wang[1], Chuhan Wu[1], Yasheng Wang[1], Ruiming Tang[1], Jianye Hao[1]

王帅[1]，刘伟文[1]，陈景轩[1]，周宇琦[2]，甘伟南[1]，曾兴山[1]，车宇涵[1]，余帅[1]，郝新龙[1]，邵坤[1]，王斌[1]，吴楚涵[1]，王亚胜[1]，唐瑞明[1]，郝建业[1]

[1] Huawei Noah’s Ark Lab [2] Renmin University of China

[1] 华为诺亚方舟实验室 [2] 中国人民大学

{wangshuai231, liuweiwen8}@huawei.com

{wangshuai231, liuweiwen8}@huawei.com

Abstract

摘要

Recent advances in foundation models, particularly Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs), have facilitated the development of intelligent agents capable of performing complex tasks. By leveraging the ability of (M)LLMs to process and interpret Graphical User Interfaces (GUIs), these agents can autonomously execute user instructions, simulating human-like interactions such as clicking and typing. This survey consolidates recent research on (M)LLM-based GUI agents, highlighting key innovations in data resources, frameworks, and applications. We begin by reviewing representative datasets and benchmarks, followed by an overview of a generalized, unified framework that encapsulates the essential components of prior studies, supported by a detailed taxonomy. Additionally, we explore relevant commercial applications. Drawing insights from existing work, we identify key challenges and propose future research directions. We hope this survey will inspire further advancements in the field of (M)LLM-based GUI agents.

近年来，基础模型，特别是大型语言模型 (LLMs) 和多模态大型语言模型 (MLLMs) 的进展，推动了能够执行复杂任务的智能代理的发展。通过利用 (M)LLMs 处理和理解图形用户界面 (GUI) 的能力，这些代理能够自主执行用户指令，模拟类似人类的交互行为，如点击和输入。本文综述了基于 (M)LLM 的 GUI 代理的最新研究，重点介绍了数据资源、框架和应用方面的关键创新。我们首先回顾了代表性的数据集和基准测试，随后概述了一个通用统一的框架，涵盖了以往研究的核心组成部分，并辅以详细的分类法。此外，我们还探讨了相关的商业应用。基于现有工作，我们识别了主要挑战并提出了未来的研究方向。希望本综述能激发基于 (M)LLM 的 GUI 代理领域的进一步发展。

1 Introduction

1 引言

Graphical User Interfaces (GUIs) are the primary medium through which humans interact with digital devices. From mobile phones to websites, people engage with GUIs daily, and well-designed GUI agents can significantly enhance the user experience. Thus, research on GUI agents has been extensive. However, traditional rule-based and reinforcement learning-based methods struggle with tasks requiring humanlike interactions [Liu et al., 2018], limiting their applicability.

图形用户界面 (GUI) 是人类与数字设备交互的主要媒介。从手机到网站，人们每天都在使用 GUI，而设计良好的 GUI 代理能够显著提升用户体验。因此，GUI 代理的研究一直十分广泛。然而，传统的基于规则和强化学习的方法在需要类人交互的任务中表现不佳 [Liu et al., 2018]，限制了其应用范围。

Recent advancements in Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) have significantly enhanced their capabilities in language understanding and cognitive processing [Achiam et al., 2024; Touvron et al., 2023; Yang et al., 2024a]. With improved natural language comprehension and enhanced reasoning abilities, (M)LLM-based agents can now effectively interpret and utilize human language, formulate detailed plans, and execute complex tasks. These breakthroughs provide new opportunities for researchers to address challenges previously considered highly difficult, such as automating tasks within GUIs.

大型语言模型 (LLMs) 和多模态大型语言模型 (MLLMs) 的最新进展显著提升了它们在语言理解和认知处理方面的能力 [Achiam et al., 2024; Touvron et al., 2023; Yang et al., 2024a]。凭借改进的自然语言理解和增强的推理能力，基于 (M)LLM 的代理现已能够有效解读和利用人类语言，制定详细计划并执行复杂任务。这些突破为研究者提供了新的机遇，以解决此前被认为极具挑战性的任务，如 GUI 内的自动化操作。

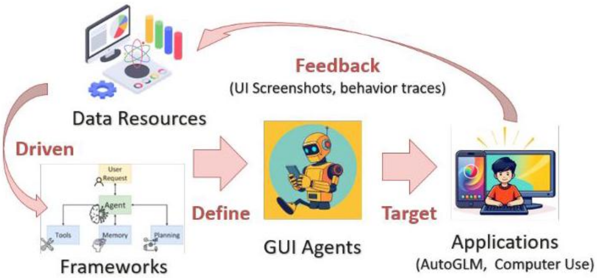

Figure 1: The foundational aspects and goals of GUI agents.

图 1:GUI 代理的基础方面与目标。

As shown in Figure 2, recent studies on GUI agents illustrate a shift from simple Transformer-based models to (M)LLM-based agentic frameworks. Their capabilities have expanded from single-modality interactions to multimodal processing, making them increasingly relevant to commercial applications. Given these advancements, we believe it is timely to systematically analyze the development trends of GUI agents, particularly from an application perspective.

如图 2 所示，近期关于 GUI 代理的研究显示出从简单的基于 Transformer 的模型向 (M)LLM 驱动的代理框架的转变。其能力已从单一模态交互扩展到多模态处理，使其在商业应用中日益重要。鉴于这些进展，我们认为现在正是系统分析 GUI 代理发展趋势，特别是从应用角度出发的合适时机。

This paper aims to provide a structured overview of the latest and influential work in the field of GUI agents. As depicted in Figure 1, we focus on the foundational aspects and goals of GUI agents. Data resources, such as user instructions, User Interface (UI) screenshots, and behavior traces, drive the design of GUI agents [Rawles et al., 2023; Lu et al., 2024a]. Frameworks define the underlying algorithms and models that enable intelligent decision-making [Li et al., 2024b; Wang et al., 2024a; Zhu et al., 2024]. Applications represent the optimized and practical goals [Lai et al., 2024; Liu et al., 2024]. The current state of these aspects reflects the maturity of the field and highlights future research priorities.

本文旨在提供 GUI 代理领域最新且有影响力工作的结构化综述。如图 1 所示，我们聚焦于 GUI 代理的基础方面和目标。数据资源，如用户指令、用户界面 (UI) 截图和行为轨迹，驱动了 GUI 代理的设计 [Rawles et al., 2023; Lu et al., 2024a]。框架定义了支持智能决策的底层算法和模型 [Li et al., 2024b; Wang et al., 2024a; Zhu et al., 2024]。应用则代表了优化后的实际目标 [Lai et al., 2024; Liu et al., 2024]。这些方面的现状反映了该领域的成熟度，并突显了未来的研究重点。

To this end, we organize this survey around three key areas: Data Resources, Frameworks, and Applications. The main contributions of this paper are: 1) a comprehensive summary of existing research and a detailed review of current data sources, providing a useful guide for newcomers to the field; 2) a unified and generalized GUI agent framework with clearly defined and categorized functional components to facilitate a structured review; 3) an analysis of trends in both research and commercial applications of GUI agents.

为此，我们围绕三个关键领域组织本综述: 数据资源、框架和应用。本文的主要贡献包括:1) 对现有研究的全面总结及当前数据源的详细回顾，为该领域新手提供有益指导；2) 提出一个统一且通用的 GUI 代理框架，明确划分和分类功能组件，便于结构化评述；3) 分析 GUI 代理在研究和商业应用中的发展趋势。

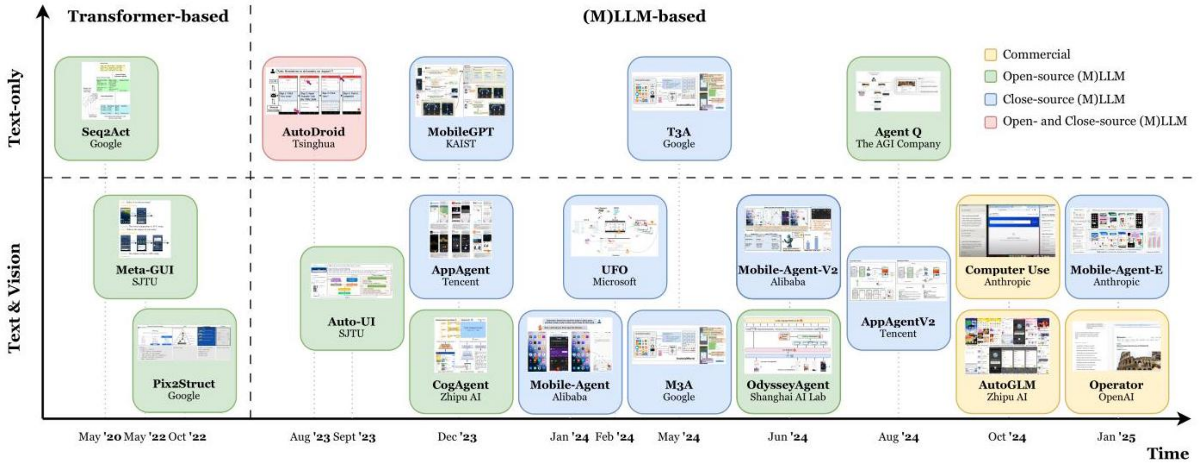

Figure 2: Illustration of the growth trend in the field of GUI agents with foundation models.

图 2: 基于基础模型的 GUI 代理领域增长趋势示意。

2 GUI Agent Data Resources

2 GUI 代理数据资源

Recent research has focused on developing datasets and benchmarks to train and evaluate the capabilities of (M)LLM-based GUI agents. A variety of datasets are available for training GUI agents. These agents employ different approaches to interact with environments. Additionally, multiple methods have been proposed for evaluation.

近期研究集中于开发数据集和基准测试，以训练和评估基于 (多模态) 大型语言模型 ((M)LLM)GUI 代理的能力。现有多种数据集可用于训练 GUI 代理，这些代理采用不同的方法与环境交互。此外，还提出了多种评估方法。

Dataset: Common datasets for training GUI agents typically contain natural language instructions that describe task goals, along with demonstration trajectories that include screenshots and action pairs. A pioneering work in this area is PIXELHELP [Li et al., 2020], which introduces a new class of problems focused on translating natural language instructions into actions on mobile user interfaces. In recent years, Android in the Wild [Rawles et al., 2023] has created a dataset featuring a variety of single-step and multi-step tasks. Aimed at advancing GUI navigation agent research, Android-In-The-Zoo [Zhang et al., 2024b] introduces a benchmark dataset with chained action reasoning annotations.

数据集: 用于训练 GUI 代理的常见数据集通常包含描述任务目标的自然语言指令，以及包括截图和动作对的演示轨迹。该领域的开创性工作是 PIXELHELP [Li et al., 2020]，其引入了一类新问题，聚焦于将自然语言指令转化为移动用户界面上的操作。近年来，Android in the Wild [Rawles et al., 2023] 创建了涵盖多种单步和多步任务的数据集。为推动 GUI 导航代理研究，Android-In-The-Zoo [Zhang et al., 2024b] 引入了带有链式动作推理注释的基准数据集。

Insight-UI [Shen et al., 2024] automatically constructs a GUI pre-training dataset that simulates multiple platforms across 312,000 domains. To assess model performance both within and beyond the scope of training data, AndroidCon-trol [Li et al., 2024a] includes demonstrations of daily tasks along with both high- and low-level human-generated instructions. The scope of mobile control datasets is further extended from single-application to cross-application scenarios by GUI-Odyssey [Lu et al., 2024a].

Insight-UI [Shen et al., 2024] 自动构建了一个模拟跨 312,000 个领域多平台的 GUI 预训练数据集。为了评估模型在训练数据内外的表现，AndroidControl [Li et al., 2024a] 包含了日常任务的演示以及高低层次的人类生成指令。GUI-Odyssey [Lu et al., 2024a] 进一步将移动控制数据集的范围从单应用扩展到跨应用场景。

Most of the aforementioned datasets are primarily limited to English and image-based tasks. However, UGIF Dataset [Venkatesh et al., 2024] covers eight languages, Mobile3M [Wu et al., 2024] focuses on Chinese, and GUI-WORLD [Chen et al., 2024a] includes video annotations, expanding the dataset landscape for broader multilingual and multimodal research.

上述大多数数据集主要限于英语和基于图像的任务。然而，UGIF Dataset [Venkatesh et al., 2024] 涵盖八种语言，Mobile3M [Wu et al., 2024] 专注于中文，GUI-WORLD [Chen et al., 2024a] 包含视频注释，拓展了多语言和多模态研究的数据集格局。

Environment: GUI agents require environments for task execution, which can be broadly categorized into three types. The first category is static environments, where the environment remains fixed as it was when developed. Agents in this category operate within predefined datasets without the ability to create new states.

环境:GUI 代理需要环境来执行任务，环境大致可分为三类。第一类是静态环境，环境保持开发时的固定状态。此类代理在预定义数据集中操作，无法创建新状态。

In contrast, the second and third categories involve dynamic environments, where new outcomes can emerge during agent execution. The key distinction between these categories lies in whether the dynamic environment is simulated or realistic. Simulations of real-world environments require additional implementation but are often cleaner and free of distractions, such as pop-ups and advertisements. We-bArena [Zhou et al., 2023] implements a versatile website covering e-commerce, social forums, collaborative software development, and content management. Similarly, GUI Testing Arena [Zhao et al., 2024] provides a standardized environment for testing GUI agents, including defect injection.

相比之下，第二类和第三类涉及动态环境，代理执行过程中可能产生新结果。这两类的关键区别在于动态环境是模拟的还是现实的。现实环境的模拟需要额外实现，但通常更简洁，无弹窗和广告等干扰。WebArena [Zhou et al., 2023] 实现了涵盖电商、社交论坛、协同软件开发和内容管理的多功能网站。类似地，GUI Testing Arena [Zhao et al., 2024] 提供了用于测试 GUI 代理的标准化环境，包括缺陷注入。

For realistic environments, agents interact directly with web or mobile platforms as human users do, better reflecting real-world conditions. SPA-Bench [Chen et al., 2024b] encompasses tasks that involve both system and third-party mobile applications, supporting single-app and cross-app scenarios in both English and Chinese.

对于现实环境，代理直接像人类用户一样与网页或移动平台交互，更真实地反映现实条件。SPA-Bench [Chen et al., 2024b] 涵盖系统及第三方移动应用任务，支持单应用和跨应用场景，包含英语和中文。

Evaluation: Another critical component of GUI agent datasets is the evaluation of agent performance. The most common and important metric is success rate, which measures how effectively an agent completes tasks. Additional metrics, such as efficiency, are sometimes considered as well.

评估:GUI 代理数据集的另一个关键组成部分是代理性能的评估。最常用且重要的指标是成功率，用以衡量代理完成任务的有效性。有时也会考虑效率等附加指标。

Evaluation methods are often closely tied to the environment type. In static environments, action matching is a widely used method that compares an agent's executed action sequence with a human demonstration (e.g., Rawles et al. [2023], Li et al. [2024a]). However, a major limitation of action matching is its inability to account for multiple successful execution paths, leading to false negatives when evaluating agent performance.

评估方法通常与环境类型密切相关。在静态环境中，动作匹配是一种广泛使用的方法，通过将代理执行的动作序列与人类演示进行比较 (如 Rawles et al. [2023]，Li et al. [2024a])。然而，动作匹配的主要局限在于无法考虑多条成功执行路径，导致评估代理性能时出现假阴性。

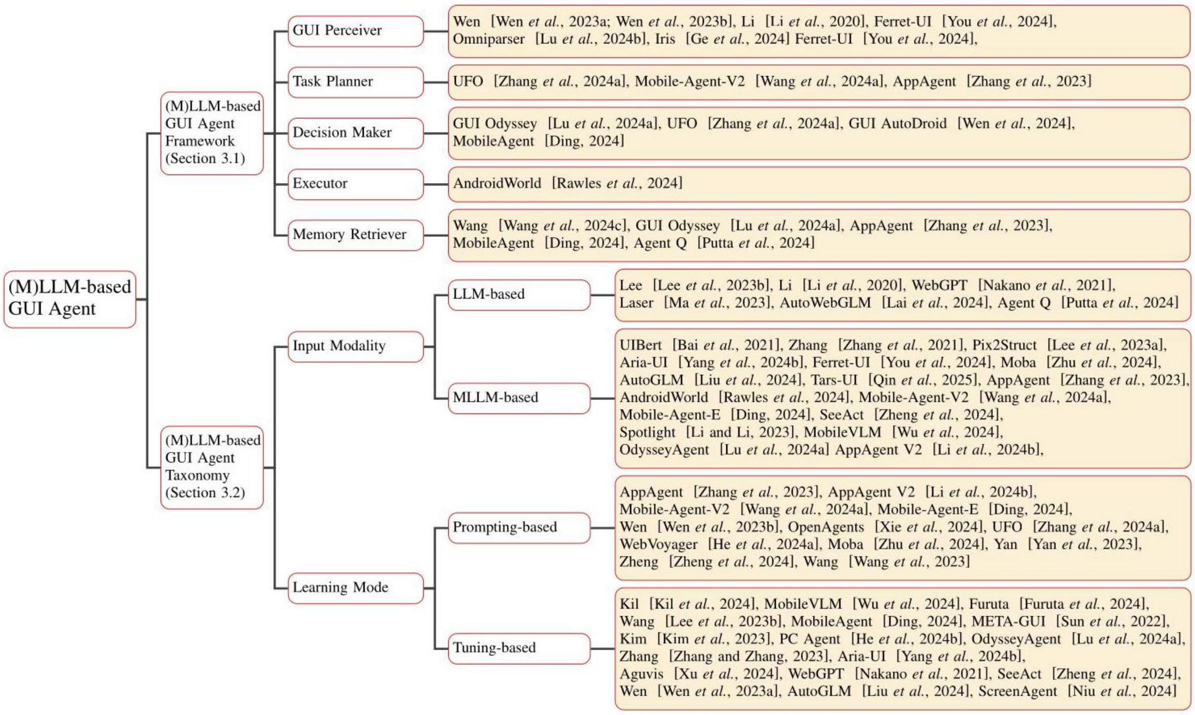
(M)LLM-based GUI Agent
(M)LLM-based GUI Agent Framework (Section 3.1)
GUI Perceiver
Wen [Wen *et al.*, 2023a; Wen *et al.*, 2023b], Li [Li *et al.*, 2020], Ferret-UI [You *et al.*, 2024], Omniparser [Lu *et al.*, 2024b], Iris [Ge *et al.*, 2024] Ferret-UI [You *et al.*, 2024],
Task Planner
UFO [Zhang *et al.*, 2024a], Mobile-Agent-V2 [Wang *et al.*, 2024a], AppAgent [Zhang *et al.*, 2023]
Decision Maker
GUI Odyssey [Lu *et al.*, 2024a], UFO [Zhang *et al.*, 2024a], GUI AutoDroid [Wen *et al.*, 2024], MobileAgent [Ding, 2024]
Executor
AndroidWorld [Rawles *et al.*, 2024]
Memory Retriever
Wang [Wang *et al.*, 2024c], GUI Odyssey [Lu *et al.*, 2024a], AppAgent [Zhang *et al.*, 2023], MobileAgent [Ding, 2024], Agent Q [Putta *et al.*, 2024]
(M)LLM-based GUI Agent Taxonomy (Section 3.2)
Input Modality
LLM-based
Lee [Lee *et al.*, 2023b], Li [Li *et al.*, 2020], WebGPT [Nakano *et al.*, 2021], Laser [Ma *et al.*, 2023], AutoWebGLM [Lai *et al.*, 2024], Agent Q [Putta *et al.*, 2024]
MLLM-based
UIBert [Bai *et al.*, 2021], Zhang [Zhang *et al.*, 2021], Pix2Struct [Lee *et al.*, 2023a], Aria-UI [Yang *et al.*, 2024b], Ferret-UI [You *et al.*, 2024], Moba [Zhu *et al.*, 2024], AutoGLM [Liu *et al.*, 2024], Tars-UI [Qin *et al.*, 2025], AppAgent [Zhang *et al.*, 2023], AndroidWorld [Rawles *et al.*, 2024], Mobile-Agent-V2 [Wang *et al.*, 2024a], Mobile-Agent-E [Ding, 2024], SeeAct [Zheng *et al.*, 2024], Spotlight [Li and Li, 2023], MobileVLM [Wu *et al.*, 2024], OdysseyAgent [Lu *et al.*, 2024a] AppAgent V2 [Li *et al.*, 2024b],
Learning Mode
Prompting-based
AppAgent [Zhang *et al.*, 2023], AppAgent V2 [Li *et al.*, 2024b], Mobile-Agent-V2 [Wang *et al.*, 2024a], Mobile-Agent-E [Ding, 2024], Wen [Wen *et al.*, 2023b], OpenAgents [Xie *et al.*, 2024], UFO [Zhang *et al.*, 2024a], WebVoyager [He *et al.*, 2024a], Moba [Zhu *et al.*, 2024], Yan [Yan *et al.*, 2023], Zheng [Zheng *et al.*, 2024], Wang [Wang *et al.*, 2023]
Tuning-based
Kil [Kil *et al.*, 2024], MobileVLM [Wu *et al.*, 2024], Furuta [Furuta *et al.*, 2024], Wang [Lee *et al.*, 2023b], MobileAgent [Ding, 2024], META-GUI [Sun *et al.*, 2022], Kim [Kim *et al.*, 2023], PC Agent [He *et al.*, 2024b], OdysseyAgent [Lu *et al.*, 2024a], Zhang [Zhang and Zhang, 2023], Aria-UI [Yang *et al.*, 2024b], Aguvis [Xu *et al.*, 2024], WebGPT [Nakano *et al.*, 2021], SeeAct [Zheng *et al.*, 2024], Wen [Wen *et al.*, 2023a], AutoGLM [Liu *et al.*, 2024], ScreenAgent [Niu *et al.*, 2024]

Figure 3: A comprehensive taxonomy of (M)LLM-based GUI Agents: frameworks, modality, and learning paradigms.

图 3:(多模态) 大型语言模型 ((M)LLM) 基础的 GUI 代理的全面分类: 框架、模态和学习范式。

Evaluating dynamic environments, whether simulated or realistic, presents additional challenges due to their uncertain conditions. Evaluation methods can range from fully human-dependent to semi-automated and fully automated approaches. Human evaluations require manual verification, making them non-reusable. In AppAgent [Li et al., 2024b] and MobileAgent [Ding, 2024], human evaluators assess whether each agent-executed task was successful. Semiautomated evaluations involve human-developed validation logic that can be reused for different execution trajectories of the same task. For example, WebArena [Zhou et al., 2023] and AndroidWorld [Rawles et al., 2024] incorporate handcrafted validation functions for task completion. Fully automated evaluations eliminate human involvement by relying on models for success detection. SPA-Bench [Chen et al., 2024b], for instance, employs MLLMs for evaluating task completion. Although reducing human labor is crucial for large-scale evaluation, balancing efficiency with accuracy remains a key research challenge.

评估动态环境 (无论是模拟还是现实) 因其不确定性条件而面临额外挑战。评估方法可从完全依赖人工到半自动和全自动不等。人工评估需手动验证，无法复用。在 AppAgent [Li et al., 2024b] 和 MobileAgent [Ding, 2024] 中，人工评估者判断每个代理执行的任务是否成功。半自动评估涉及人工开发的验证逻辑，可复用于同一任务的不同执行轨迹。例如，WebArena [Zhou et al., 2023] 和 AndroidWorld [Rawles et al., 2024] 包含手工编写的任务完成验证函数。全自动评估则通过模型检测成功，消除人工参与。SPA-Bench [Chen et al., 2024b] 采用多模态大型语言模型 (MLLM) 评估任务完成情况。尽管减少人工劳动对大规模评估至关重要，但如何在效率与准确性之间取得平衡仍是关键研究挑战。

3 (M)LLM-based GUI Agent

3 基于 (多模态) 大型语言模型 ((M)LLM) 的 GUI 代理

With the human-like capabilities of (M)LLMs, GUI agents aim to handle various tasks to meet users' needs. Organizing the frameworks of GUI agents and designing methods to optimize their performance is crucial to unlocking the full potential of (M)LLMs. As shown in Figure 3, we summarize a generalized Framework and discuss its components in relation to existing works in Section 3.1. Building on this foundation, we then review recent influential Methods for constructing and optimizing GUI agents, categorizing them with an exhaustive taxonomy in Section 3.2.

凭借类人能力的 (多模态) 大型语言模型 ((M)LLMs)，图形用户界面 (GUI) 代理旨在处理各种任务以满足用户需求。组织 GUI 代理的框架并设计优化其性能的方法，对于充分发挥 (M)LLMs 的潜力至关重要。如图 3 所示，我们总结了一个通用框架，并在第 3.1 节中结合现有工作讨论其组成部分。在此基础上，我们在第 3.2 节回顾了构建和优化 GUI 代理的最新重要方法，并通过详尽的分类法对其进行归类。

3.1 (M)LLM-based GUI Agent Framework

3.1 基于 (多模态) 大型语言模型的 GUI 代理框架

The goal of GUI agents is to automatically control a device to complete tasks defined by the user. Typically, GUI agents take a user's query and the device's UI status as inputs and generate a series of human-like actions to achieve the tasks.

GUI 代理的目标是自动控制设备以完成用户定义的任务。通常，GUI 代理以用户查询和设备的 UI 状态作为输入，生成一系列类人的操作以实现任务。

As shown in Figure 4, we present a generalized (M)LLM-based GUI agent framework, consisting of five components: GUI Perceiver, Task Planner, Decision Maker, Memory Retriever, and Executor. Many variations of this framework exist. For instance, Wang et al. [2024a] proposes a multi-agent GUI control framework comprising a planning agent, a decision agent, and a reflection agent to tackle navigation challenges in mobile device operations. This approach shares functional similarities with our proposed framework. A follow-up study by Wang et al.

[2025] further disentangles high-level planning from low-level action decisions by employing dedicated agents and introduces memory-based self-evolution to enhance performance.

如图 4 所示，我们提出了一个通用的基于 (多模态) 大型语言模型的 GUI 代理框架，包含五个组成部分:GUI 感知器、任务规划器、决策者、记忆检索器和执行器。该框架存在多种变体。例如，Wang 等人 [2024a] 提出了一个多代理 GUI 控制框架，包括规划代理、决策代理和反思代理，以解决移动设备操作中的导航难题。该方法在功能上与我们提出的框架相似。Wang 等人 [2025] 的后续研究通过专用代理进一步区分了高层规划与低层动作决策，并引入基于记忆的自我进化机制以提升性能。

GUI Perceiver: To effectively complete a device task, a GUI agent should accurately interpret user input and detect changes in the device's UI. Although language models excel in understanding user intent [Touvron et al., 2023; Achiam et al., 2024], navigating device UIs requires a reliable visual perception model to understand GUIs.

GUI 感知器: 为了有效完成设备任务，GUI 代理应准确解读用户输入并检测设备 UI 的变化。尽管语言模型在理解用户意图方面表现出色 [Touvron 等，2023；Achiam 等，2024]，但导航设备 UI 需要可靠的视觉感知模型来理解图形用户界面。

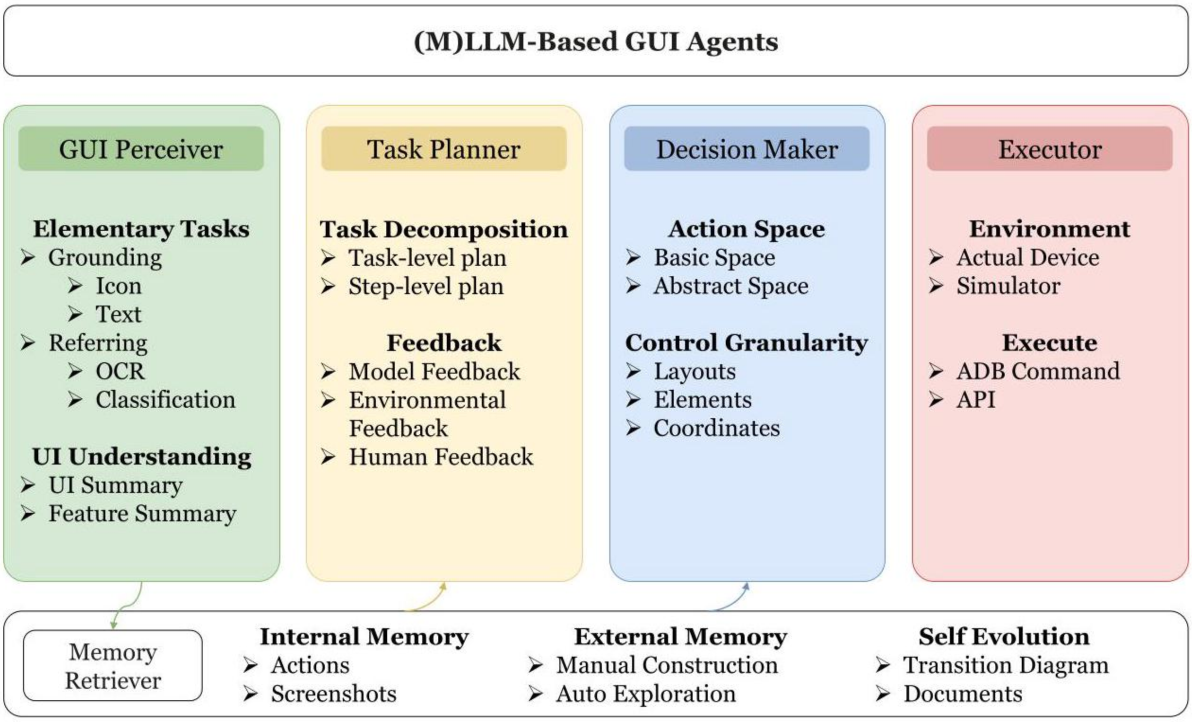

Figure 4: (M)LLM-based GUI agents: the generalized framework and key technologies.

图 4:(多模态) 大型语言模型驱动的 GUI 代理: 通用框架及关键技术。

A GUI Perceiver appears explicitly or implicitly in GUI agent frameworks. For agents based on single-modal LLMs [Wen et al., 2023a; Wen et al., 2023b; Li et al., 2020], a GUI Perceiver is usually an explicit module of the frameworks. However, for agents with multi-modal LLMs [Hong et al., 2024; Zhang et al., 2023; Wang et al., 2024b], UI perception is seen as a capability of the model itself.

GUI 感知器在 GUI 代理框架中显式或隐式存在。对于基于单模态大型语言模型的代理 [Wen 等，2023a；Wen 等，2023b；Li 等，2020]，GUI 感知器通常是框架中的显式模块。然而，对于具备多模态大型语言模型的代理 [Hong 等，2024；Zhang 等，2023；Wang 等，2024b]，UI 感知被视为模型自身的能力。

UI perception is also an important problem in GUI agent research, some work [You et al., 2024; Zhang et al., 2021; Lu et al., 2024b] focuses on understanding and processing UIs, rather than building the agent. For example, Pix2struct [Lee et al., 2023a] employs a ViT-based image-encoder-text-decoder architecture, which pre-trains on Screenshot-HTML data pairs and fine-tunes for specific tasks. This method has shown strong performance in web-based visual comprehension tasks. Screen2words [Wang et al., 2021] is a novel approach that encapsulates a UI screen into a coherent language representation, which is based on a transformer encoder-decoder architecture to process UIs and generate the representation. To address the defects of purely vision-based screen parsing methods, Ge et al. [2024] introduces Iris, a visual agent for GUI understanding, addressing challenges related to architectural limitations for heterogeneous GUI information and annotation bias in GUI training via two innovations: An information-sensitive architecture to prioritize high-density UI regions via edge detection, and a dual-learning strategy that refines visual/functional knowledge iteratively using unlabeled data, reducing annotation dependence.

UI 感知也是 GUI 代理研究中的重要问题，一些工作 [You 等，2024；Zhang 等，2021；Lu 等，2024b] 专注于理解和处理 UI，而非构建代理。例如，Pix2struct[Lee 等，2023a] 采用基于 ViT 的图像编码器-文本解码器架构，在截图-HTML 数据对上进行预训练，并针对特定任务进行微调。该方法在基于网页的视觉理解任务中表现优异。Screen2words[Wang 等，2021] 是一种新颖方法，将 UI 屏幕封装为连贯的语言表示，基于 Transformer 编码器-解码器架构处理 UI 并生成表示。为解决纯视觉屏幕解析方法的缺陷，Ge 等人 [2024] 提出了 Iris，一种用于 GUI 理解的视觉代理，通过两项创新应对异构 GUI 信息的架构限制和 GUI 训练中的标注偏差: 一种信息敏感架构通过边缘检测优先处理高密度 UI 区域；一种双重学习策略利用无标注数据迭代优化视觉与功能知识，减少对标注的依赖。

Task Planner: The GUI agent should effectively decompose complex tasks, often employing a Chain-of-Thought (CoT) approach. Due to the complexity of tasks, recent studies [Zhang et al., 2024a; Wang et al., 2024a] introduce an additional module to support more detailed planning.

任务规划器:GUI 代理应有效分解复杂任务，通常采用链式思维 (Chain-of-Thought, CoT) 方法。鉴于任务复杂性，近期研究 [Zhang 等，2024a；Wang 等，2024a] 引入额外模块以支持更细致的规划。

Throughout the GUI agent's process, plans may adapt dynamically based on decision feedback, typically achieved through a ReAct-style. For instance, Zhang et al. [2023] uses on-screen observations to enhance the CoT for improved decision-making, while Wang et al. [2024a] develops a reflection agent that provides feedback to refine plans.

在 GUI 代理的执行过程中，计划可能基于决策反馈动态调整，通常通过 ReAct 风格实现。例如，Zhang 等人 [2023] 利用屏幕观察增强 CoT 以改进决策，而 Wang 等人 [2024a] 开发了反思代理，提供反馈以优化计划。

Decision Maker: A Decision Maker provides the next operation(s) to control a device. Most studies [Lu et al., 2024a; Zhang et al., 2024a; Wen et al., 2024] define a set of UI-related actions—such as click, text, and scroll—as a basic action space. In a more complicated case, Ding [2024] encapsulates a sequence of actions to create

Standard Operating Procedures(SOPs) to guide further operations.

决策者: 决策者负责提供下一步操作以控制设备。大多数研究 [Lu 等，2024a；Zhang 等，2024a；Wen 等，2024] 定义了一组与 UI 相关的基本动作，如点击、输入文本和滚动，作为基本动作空间。在更复杂的情况下，Ding[2024] 将一系列动作封装为标准操作流程 (Standard Operating Procedures, SOPs)，以指导后续操作。

As the power of GUI agents improves, the granularity of operations becomes more refined. Recent work has progressed from element-level operations [Zhang et al., 2023; Wang et al., 2024b] to coordinate-level controls [Wang et al., 2024a; Hong et al., 2024].

随着 GUI 代理能力的提升，操作的粒度变得更加细化。近期的研究已从元素级操作 [Zhang et al., 2023; Wang et al., 2024b] 发展到坐标级控制 [Wang et al., 2024a; Hong et al., 2024]。

Executor: An Executor maps outputs to the relevant environments. While most studies use Android Debug Bridge (ADB) to control real devices [Li et al., 2024b; Wang et al., 2024a], Rawles et al. [2024] develops a simulator to access additional UI-related information.

执行器 (Executor): 执行器将输出映射到相关环境。尽管大多数研究使用 Android 调试桥 (Android Debug Bridge, ADB) 来控制真实设备 [Li et al., 2024b; Wang et al., 2024a]，Rawles et al. [2024] 开发了一个模拟器以访问更多与 UI 相关的信息。

Memory Retriever: A Memory Retriever is designed as an additional source of information to help agents perform tasks more effectively [Wang et al., 2024c].

记忆检索器 (Memory Retriever): 记忆检索器被设计为额外的信息来源，帮助代理更有效地执行任务 [Wang et al., 2024c]。

GUI agents' memory is typically divided into internal and external categories. Internal memory [Lu et al., 2024a] consists of prior actions, screenshots, and system states during execution, while external memory [Zhang et al., 2023; Ding, 2024] includes knowledge and rules related to the UI or task, providing additional inputs for the agent.

GUI 代理的记忆通常分为内部和外部两类。内部记忆 [Lu et al., 2024a] 包括先前的操作、截图和执行过程中的系统状态，而外部记忆 [Zhang et al., 2023; Ding, 2024] 则包含与 UI 或任务相关的知识和规则，为代理提供额外输入。

3.2 (M)LLM-based GUI Agent Taxonomy

3.2 基于 (多模态) 大语言模型 ((M)LLM) 的 GUI 代理分类

Consequently, this paper classifies existing work with the difference of input modality and learning mode in Figure 3.

因此，本文根据输入模态和学习模式的差异，在图 3 中对现有工作进行了分类。

GUI Agents with Different Input modality

具有不同输入模态的 GUI 代理

LLM-based GUI Agents: With the limited multimodal capability, earlier GUI agents [Lee et al., 2023b; Li et al., 2020; Ma et al., 2023; Lai et al., 2024; Putta et al., 2024; Nakano et al., 2021; Nakano et al., 2021] often require a GUI perceiver to convert GUI screens into text-based inputs.

基于大语言模型 (LLM) 的 GUI 代理: 由于多模态能力有限，早期的 GUI 代理 [Lee et al., 2023b; Li et al., 2020; Ma et al., 2023; Lai et al., 2024; Putta et al., 2024; Nakano et al., 2021] 通常需要 GUI 感知器将 GUI 界面转换为基于文本的输入。

So, parsing and grounding the GUI screens is the first step. For instance, Li et al. [2020] transforms the screen into a series of object descriptions and applies a transformer-based action mapping. The problem definitions and datasets have spurred further research. You et al. [2024] proposes a series of referring and grounding tasks, which provide valuable insights into the pre-training of GUIs. Lu et al. [2024b] proposes a screen parsing framework incorporating the local semantics of functionality with interactable region detection for better UI understanding and element grounding.

因此，解析和定位 GUI 界面是第一步。例如，Li et al. [2020] 将界面转换为一系列对象描述，并应用基于 Transformer 的动作映射。问题定义和数据集推动了进一步研究。You et al. [2024] 提出了一系列指代和定位任务，为 GUI 的预训练提供了宝贵见解。Lu et al. [2024b] 提出了一个屏幕解析框架，结合功能的局部语义和可交互区域检测，以实现更好的 UI 理解和元素定位。

Afterward, LLMs are used as the core of the agents. Wen et al. [2024] further converts GUI screenshots into a simplified HTML representation for compatibility with the LLMs. By combining GUI representation with app-specific knowledge, they build Auto-Droid, a GUI agent based on online GPT and on-device Vicuna. In the field of web automation, LASER [Ma et al., 2023] navigates web environments purely through text, treating web navigation as state-space exploration to enable flexible state transitions and error recovery. Similarly, AutoWebGLM [Lai et al., 2024] processes HTML text data without visual inputs, refining webpage structures to preserve key information for ChatGLM3-6B. Agent Q [Putta et al., 2024] further extends this paradigm by relying solely on HTML DOM text for reasoning and decision-making, emphasizing language models for planning and action execution. WebGPT [Nakano et al., 2021], a fine-tuned GPT-3 model, uses text-based web browsing (processing HTML content) to collect information via commands like searching and clicking. It generates answers supported by references and is optimized using human feedback and rejection sampling.

随后，LLM 被用作代理的核心。Wen et al. [2024] 进一步将 GUI 截图转换为简化的 HTML 表示，以兼容 LLM。通过结合 GUI 表示与应用特定知识，他们构建了基于在线 GPT 和设备端 Vicuna 的 GUI 代理 Auto-Droid。在网页自动化领域，LASER[Ma et al., 2023] 纯文本导航网页环境，将网页导航视为状态空间探索，实现灵活的状态转换和错误恢复。类似地，AutoWebGLM[Lai et al., 2024] 处理 HTML 文本数据，无需视觉输入，优化网页结构以保留关键信息供 ChatGLM3-6B 使用。Agent Q[Putta et al., 2024] 进一步扩展了这一范式，仅依赖 HTML DOM 文本进行推理和决策，强调语言模型在规划和执行动作中的作用。WebGPT[Nakano et al., 2021] 是一个微调的 GPT-3 模型，使用基于文本的网页浏览 (处理 HTML 内容) 通过搜索和点击等命令收集信息，生成带有参考支持的答案，并通过人类反馈和拒绝采样进行优化。

MLLM-based GUI Agents: Recent studies [Wang et al., 2024a; Bai et al., 2021; Zhang et al., 2023; Kim et al., 2023] utilize the multimodal capabilities of advanced (M)LLMs to improve GUI comprehension and task execution.

基于多模态大语言模型 (MLLM) 的 GUI 代理: 近期研究 [Wang et al., 2024a; Bai et al., 2021; Zhang et al., 2023; Kim et al., 2023] 利用先进 (M)LLM 的多模态能力提升 GUI 理解和任务执行效果。

Leveraging the visual understanding capabilities of MLLMs, recent studies [Wang et al., 2024a; Li and Li, 2023; Bai et al., 2021; Zhu et al., 2024; Qin et al., 2025] explore end-to-end frameworks for GUI device control. For example, Spotlight [Li and Li, 2023] proposes a Vision-Language model framework, pre-trained on Web/mobile data and fine-tuned for UI tasks. This model greatly improves the ability to understand UIs. By combining screenshots with a user focus as input, Spotlight outperforms previous methods on multiple UI understanding tasks, showing verified gains in downstream tasks. Likewise, VUT [Li et al., 2021] is proposed for GUI understanding and multi-modal UI input modeling, using two Transformers: one for encoding and fusing image, structural, and language inputs, and the other for linking three task heads to complete five distinct UI modeling tasks and learn downstream multiple tasks end-to-end. Experiments show that VUT's multi-task learning framework can achieve state-of-the-art (SOTA) performance on UI modeling tasks. UIbert [Bai et al., 2021] focuses on heterogeneous GUI features and considers that the multi-modal information in the GUI is self-aligned. UIbert is a transformer-based joint image-text model, which is pre-trained in large-scale unlabeled GUI data to learn the feature representation of UI elements. Zhu et al. [2024] presents a two-level agent structure for executing complex and dynamic GUI tasks. Moba's Global Agent handles high-level planning, while the Local Agent selects actions for sub-tasks, streamlining the decision-making process with improved efficiency. UI-TARS [Qin et al., 2025] navigates interfaces through screenshots, enabling human-like interactions via keyboard and mouse. Leveraging a large-scale GUI dataset, it achieves context-aware UI understanding and precise captioning.

利用多模态大语言模型 (MLLMs) 的视觉理解能力，近期研究 [Wang et al., 2024a; Li and Li, 2023; Bai et al., 2021; Zhu et al., 2024; Qin et al., 2025] 探索了用于图形用户界面 (GUI) 设备控制的端到端框架。例如，Spotlight [Li and Li, 2023] 提出了一个视觉-语言模型框架，在网络/移动数据上预训练并针对 UI 任务进行微调。该模型大幅提升了对 UI 的理解能力。通过结合带有用户关注点的截图作为输入，Spotlight 在多项 UI 理解任务中优于以往方法，并在下游任务中验证了性能提升。同样，VUT [Li et al., 2021] 被提出用于 GUI 理解和多模态 UI 输入建模，采用两个 Transformer: 一个用于编码和融合图像、结构和语言输入，另一个用于连接三个任务头以完成五个不同的 UI 建模任务，并端到端学习多个下游任务。实验表明，VUT 的多任务学习框架在 UI 建模任务上可实现最先进 (SOTA) 性能。UIbert [Bai et al., 2021] 聚焦于异构 GUI 特征，认为 GUI 中的多模态信息是自对齐的。UIbert 是基于 Transformer 的图文联合模型，在大规模无标签 GUI 数据上预训练，以学习 UI 元素的特征表示。Zhu et al. [2024] 提出了一个两级代理结构，用于执行复杂且动态的 GUI 任务。Moba 的全局代理负责高层规划，而本地代理选择子任务的动作，简化决策过程并提升效率。UI-TARS [Qin et al., 2025] 通过截图导航界面，实现类似人类的键盘和鼠标交互。利用大规模 GUI 数据集，它实现了上下文感知的 UI 理解和精准的描述生成。

To enhance performance, some studies [Zhang et al., 2023; Rawles et al., 2024] utilize additional invisible meta-data. For instance, AndroidWorld [Rawles et al., 2024] establishes a fully functional Android environment with real-world tasks, serving as a benchmark for evaluating GUI agents. They propose M3A, a zero-shot prompting agent that uses Set-of-Marks as input. Experiments with M3A variants assess how different input modalities —text, screenshots, and accessibility trees-affect GUI agent performance. Yang et al. [2024b] proposes a framework incorporating dynamic action history with both textual and interleaved text-image formats, which allows it to ground elements more effectively for dynamic, multi-step scenarios.

为提升性能，一些研究 [Zhang et al., 2023; Rawles et al., 2024] 利用了额外的隐形元数据。例如，AndroidWorld [Rawles et al., 2024] 构建了一个功能完备的 Android 环境，包含真实任务，作为评估 GUI 代理的基准。他们提出了 M3A，一种零样本提示代理，使用标记集 (Set-of-Marks) 作为输入。通过对 M3A 不同变体的实验，评估了文本、截图和辅助功能树等不同输入模态对 GUI 代理性能的影响。Yang et al. [2024b] 提出了一个框架，结合动态动作历史，采用文本和交错文本-图像格式，使其能更有效地定位元素，适用于动态多步骤场景。

GUI Agents with Different Learning Mode

不同学习模式的 GUI 代理

Prompting-based GUI Agents: Prompting is an effective approach to building agents with minimal extra computational overhead. Given the diversity of GUIs and tasks, numerous studies [Zhang et al., 2023; Li et al., 2024b; Wang et al., 2024a; Wen et al., 2023b; Xie et al., 2024; Zhang et al., 2024a; He et al., 2024a] use prompting to create GUI agents, adopting CoT or ReAct styles.

基于提示的 GUI 代理: 提示是一种构建代理的有效方法，几乎不增加额外计算开销。鉴于 GUI 和任务的多样性，众多研究 [Zhang et al., 2023; Li et al., 2024b; Wang et al., 2024a; Wen et al., 2023b; Xie et al., 2024; Zhang et al., 2024a; He et al., 2024a] 采用提示技术创建 GUI 代理，使用链式思维 (CoT) 或反应式 (ReAct) 风格。

Recent studies use prompting to build and simulate the functions of GUI agent components. For example, Yan et al. [2023] introduces MM-Navigator, which utilizes GPT-4V for zero-shot GUI understanding and navigation. For the first time, this work demonstrates the significant potential of LLMs, particularly GPT-4V, for zero-shot GUI tasks. Manual evaluations show that MM-Navigator achieves impressive performance in generating reasonable action descriptions and single-step instructions for iOS tasks. Additionally, Wen et al. [2023b] presents DroidBot-GPT, which summarizes the app's status, past actions, and tasks into a prompt, using Chat-GPT to choose the next action. Beyond mobile applications, prompting-based approaches have also been widely adopted in web-based GUI agents. Zheng et al. [2024] proposes See-Act, a GPT-4V-based generalist web agent. With screenshots as input, SeeAct generates action descriptions and converts them into executable actions with designed action grounding techniques. OpenAgents [Xie et al., 2024] leverages prompts to guide browser extensions in executing tasks such as web navigation and form filling, operating purely on the reasoning capabilities of LLMs without additional training. Similarly, WebVoyager [He et al., 2024a] integrates visual and textual information from screenshots and web pages, using prompts to interpret UI elements and execute interactions like clicking and typing. UFO [Zhang et al., 2024a] dynamically generates task plans and executes actions through prompting, allowing it to generalize across diverse web tasks without requiring task-specific adaptations.

近期研究利用提示构建和模拟 GUI 代理组件的功能。例如，Yan et al. [2023] 提出了 MM-Navigator，利用 GPT-4V 实现零样本 GUI 理解和导航。该工作首次展示了大型语言模型 (LLMs)，尤其是 GPT-4V，在零样本 GUI 任务中的巨大潜力。人工评估显示，MM-Navigator 在生成合理的动作描述和 iOS 任务的单步指令方面表现出色。此外，Wen et al. [2023b] 提出了 DroidBot-GPT，将应用状态、历史动作和任务总结为提示，利用 Chat-GPT 选择下一步动作。除了移动应用，基于提示的方法也被广泛应用于基于网页的 GUI 代理。Zheng et al. [2024] 提出了 See-Act，一种基于 GPT-4V 的通用网页代理。以截图为输入，See-Act 生成动作描述，并通过设计的动作定位技术将其转化为可执行动作。OpenAgents [Xie et al., 2024] 利用提示引导浏览器扩展执行网页导航和表单填写等任务，完全依赖 LLMs 的推理能力，无需额外训练。类似地，WebVoyager [He et al., 2024a] 整合截图和网页的视觉及文本信息，利用提示解释 UI 元素并执行点击、输入等交互。UFO [Zhang et al., 2024a] 通过提示动态生成任务计划并执行动作，使其能在多样化网页任务中泛化，无需针对特定任务调整。

Table 1: Overview of (M)LLM-Based GUI Agents.

表 1:(多模态) 大语言模型 (MLLM) 基础的 GUI 代理概览。

Model Name	Category	GUI Perceiver	Learning Method	Base Model	Scenarios
Prompting-based					
PaLM [Wang et al., 2023]	Single Step	HTML	Few-shot prompting	PaLM	Mobile
MM-Navigator [Yan et al., 2023]	Single Step	Screenshot	Zero-shot prompting	GPT-4V	Mobile
MemoDroid [Lee et al., 2023b]	End-to-End	HTML	Few-shot prompting	ChatGPT/GPT-4V	Mobile/Desktop
AutoTask [Pan et al., 2023]	End-to-End	Screenshot/API	Zero-shot prompting	GPT-4V	Mobile
AppAgent [Zhang et al., 2023]	End-to-End	Screenshot	Exploration-based/In-context learning	GPT4V	Mobile
DroidBot-GPT [Wen et al., 2023b]	End-to-End	Screenshot	Zero-shot prompting	ChatGPT	Mobile
Mobile-Agent-V2 [Wang et al., 2024a]	End-to-End	Screenshot	Zero-shot prompting	GPT4V	Mobile
SeeAct [Zheng et al., 2024]	End-to-End	Screenshot/HTML	Few-shot prompting	GPT-4V	Web
Mobile-Agent-E [Wang et al., 2025]	End-to-End	Screenshot	Zero-shot prompting	GPT-4o/Claude-3.5-Sonnet/Gemini-1.5-pro	Mobile
Learning-based					
Spotlight [Li and Li, 2023]	UI modeling	Screenshot	Pretrain/SFT	ViT	Mobile/Web
Pix2Struct [Lee et al., 2023a]	UI modeling	Screenshot	Pretrain/SFT	ViT	Web
VUT [Li et al., 2021]	UI modeling	Screenshot	SFT	Transformer	Mobile/Web
Screen Recognition [Zhang et al., 2021]	UI modeling	Screenshot	SFT	Faster R-CNN	Mobile
Screen2Words [Wang et al., 2021]	UI modeling	Screenshot	SFT	Transformer	Mobile
Aria-UI [Yang et al., 2024b]	UI modeling	Screenshot	Pretrain/SFT	Aria	Mobile/Web/Desktop
Ferret-UI [You et al., 2024]	UI modeling	Screenshot	Pretrain/SFT	Ferret	Mobile
AutoDroiá Twen et al., 2024]	End-to-End	HTML	Exploration-based/SFT	Vicuna-7B	Mobile
Seq2Act [Li et al., 2020]	End-to-End	Texts	Supervised learning	Transformer	Mobile
Meta-GUI [Sun et al., 2022]	End-to-End	Screenshot/XML	Supervised learning	Transformer	Mobile
Agent Q [Putta et al., 2024]	End-to-End	Screenshot/DOM	RL/BC Training	Transformer	Web
WebGUM [Furuta et al., 2024]	End-to-End	Screenshot/HTML	SFT	Flan-T5	Web
CogAgent [Hong et al., 2024]	End-to-End	Screenshot	SFT	CogVLM	Mobile/Desktop
MobileVLM [Wu et al., 2024]	End-to-End	XML/Screenshot	Pretrain/SFT	Owen-VL-Chat	Mobile
WebGPT [Nakano et al., 2021]	End-to-End	Texts	SFT	GPT-3	Web
AutoGLM [Liu et al., 2024]	End-to-End	Screenshot/HTML	Pretrain/SFT/RL	ChatGLM	Mobile/Web
OdysseyAgent [Lu et al., 2024a]	End-to-End	Screenshot	SFT	Qwen-VL	Mobile

模型名称	类别	GUI 感知器	学习方法	基础模型	应用场景
基于提示					
PaLM [Wang et al., 2023]	单步	HTML	少量示例提示	PaLM	移动端
MM-Navigator [Yan et al., 2023]	单步	截图	零样本提示	GPT-4V	移动端
MemoDroid [Lee et al., 2023b]	端到端	HTML	少量示例提示	ChatGPT/GPT-4V	移动端/桌面端
AutoTask [Pan et al., 2023]	端到端	截图/API	零样本提示	GPT-4V	移动端
AppAgent [Zhang et al., 2023]	端到端	截图	基于探索/上下文学习	GPT4V	移动端
DroidBot-GPT [Wen et al., 2023b]	端到端	截图	零样本提示	ChatGPT	移动端
Mobile-Agent-V2 [Wang et al., 2024a]	端到端	截图	零样本提示	GPT4V	移动端
SeeAct [Zheng et al., 2024]	端到端	截图/HTML	少量示例提示	GPT-4V	网页
Mobile-Agent-E [Wang et al., 2025]	端到端	截图	零样本提示	GPT-4o/Claude-3.5-Sonnet/Gemini-1.5-pro	移动端
基于学习					
Spotlight [Li and Li, 2023]	界面建模	截图	预训练/微调 (SFT)	ViT	移动端/网页
Pix2Struct [Lee et al., 2023a]	界面建模	截图	预训练/微调 (SFT)	ViT	网页
VUT [Li et al., 2021]	界面建模	截图	微调 (SFT)	Transformer	移动端/网页
屏幕识别 [Zhang et al., 2021]	界面建模	截图	微调 (SFT)	Faster R-CNN	移动端
Screen2Words [Wang et al., 2021]	界面建模	截图	微调 (SFT)	Transformer	移动端
Aria-UI [Yang et al., 2024b]	界面建模	截图	预训练/微调 (SFT)	Aria	移动端/网页/桌面端
Ferret-UI [You et al., 2024]	界面建模	截图	预训练/微调 (SFT)	Ferret	移动端
AutoDroiá Twen et al., 2024]	端到端	HTML	基于探索/微调 (SFT)	Vicuna-7B	移动端
Seq2Act [Li et al., 2020]	端到端	文本	监督学习	Transformer	移动端
Meta-GUI [Sun et al., 2022]	端到端	截图/XML	监督学习	Transformer	移动端
Agent Q [Putta et al., 2024]	端到端	截图/DOM	强化学习/行为克隆训练	Transformer	网页
WebGUM [Furuta et al., 2024]	端到端	截图/HTML	微调 (SFT)	Flan-T5	网页
CogAgent [Hong 等, 2024]	端到端	截图	微调 (SFT)	CogVLM	移动端/桌面端
MobileVLM [Wu 等, 2024]	端到端	XML/截图	预训练/微调 (SFT)	Owen-VL-Chat	移动端
WebGPT [Nakano 等, 2021]	端到端	文本	微调 (SFT)	GPT-3	网页
AutoGLM [Liu 等, 2024]	端到端	截图/HTML	预训练/微调/强化学习	ChatGLM	移动端/网页
OdysseyAgent [Lu 等, 2024a]	端到端	截图	微调 (SFT)	Qwen-VL	移动端

Some studies enable the GUI agent to fully leverage external knowledge through prompting to complete GUI tasks.

一些研究通过提示使 GUI 代理能够充分利用外部知识来完成 GUI 任务。

AppAgent [Zhang et al., 2023] proposes a multi-modal agent framework to simulate human-like mobile phone operations. The framework is divided into two phases: Exploration, where agents explore applications and document their operations, and Deployment, where these documents guide the agent in observing, thinking, acting, and summarizing tasks. This is the first work to claim human-like GUI automation capabilities. AppAgent V2 [Li et al., 2024b] further improves GUI parsing, document generation, and prompt integration by incorporating optical character recognition (OCR) and detection tools, moving beyond the limitations of off-the-shelf parsers for UI element identification. Wang et al. [2023] uses a pure in-context learning method to implement interaction between LLMs and mobile UIs. The method divides the conversations between agents and users into four categories from the originator and designs a series of structural CoT prompting to adapt an LLM to execute mobile UI tasks. MobileGPT [Lee et al., 2023b] emulates the cognitive processes of human use of applications to enhance the LLM-based agent with a human-like app memory. MobileGPT uses a random explorer to explore and generate screen-related subtasks on many apps and save them as app memory. During the execution, the related memory is recalled to complete tasks. SFT-based GUI Agents: Supervised fine-tuning (SFT) allows (M)LLMs to adapt to specific domains and perform customized tasks with high efficiency. Recent studies on GUI agents [Wen et al., 2023a; Furuta et al., 2024; Niu et al., 2024; He et al., 2024b; Kil et al., 2024] demonstrate the benefits of SFT for GUI agents to process new modal inputs, learn specific procedures, or execute specialized tasks.

AppAgent [Zhang et al., 2023]提出了一种多模态代理框架，以模拟类人手机操作。该框架分为两个阶段: 探索阶段，代理探索应用并记录其操作；部署阶段，利用这些文档指导代理进行观察、思考、行动和总结任务。这是首个声称具备类人GUI自动化能力的工作。AppAgent V2 [Li et al., 2024b]通过引入光学字符识别(OCR)和检测工具，进一步提升了GUI解析、文档生成和提示整合，突破了现成解析器在UI元素识别上的局限。Wang et al. [2023]采用纯上下文学习方法实现大型语言模型(LLMs)与移动UI的交互。该方法将代理与用户的对话从发起者角度划分为四类，并设计了一系列结构化链式推理(CoT)提示，使LLM能够执行移动UI任务。MobileGPT [Lee et al., 2023b]模拟人类使用应用的认知过程，增强基于LLM的代理，赋予其类人应用记忆。MobileGPT使用随机探索器在多个应用上探索并生成与屏幕相关的子任务，保存为应用记忆。执行时调用相关记忆以完成任务。基于监督微调(SFT)的GUI代理: 监督微调使(多模态)LLMs能够适应特定领域并高效执行定制任务。近期关于GUI代理的研究[Wen et al., 2023a; Furuta et al., 2024; Niu et al., 2024; He et al., 2024b; Kil et al., 2024]展示了SFT在处理新模态输入、学习特定流程或执行专业任务方面的优势。

For instance, Furuta et al. [2024] proposes WebGUM for web navigation. WebGUM is jointly fine-tuned with an instruction-optimized language model and a vision encoder, incorporating temporal and local perceptual capabilities. The evaluation results on MiniWoB show that WebGUM outperforms GPT-4-based agents. Zhang and Zhang [2023] introduces Auto-UI, a multimodal solution combining an image-language encoder-decoder architecture with a Chain of Actions policy, fine-tuned on the AitW dataset. This Chain of Actions captures intermediate previous action histories and future action plans. Yang et al. [2024b] proposes a data-centric pipeline to generate high-quality generalization data from publicly available data. This data is used to fine-tune the VLM for diverse instructions in various environments. Xu et al. [2024] introduces a two-stage training paradigm for AGU-VIS. In the first stage, the agent learns visual representations of GUI components through self-supervised learning. In the second stage, it fine-tunes interactive tasks using reinforcement learning, enabling efficient autonomous GUI interaction. On computer-based environments, ScreenAgent [Niu et al., 2024] fine-tunes the ScreenAgent dataset, mapping screenshots to action sequences. It operates via VNC, following a planning-acting-reflecting framework inspired by Kolb's experiential learning. PC-Agent [He et al., 2024b] employs a multi-agent architecture, fine-tuning a planning agent on cognitive trajectories collected via PC Tracker, enabling it to model human cognitive patterns. Additionally, Kil et al. [2024] fine-tunes DeBERTa for element ranking and Flan-T5 for action prediction, incorporating visual signals to enhance web navigation.

例如，Furuta et al. [2024] 提出用于网页导航的 WebGUM。WebGUM 与指令优化语言模型和视觉编码器联合微调，融合了时间和局部感知能力。在 MiniWoB 上的评测结果显示，WebGUM 优于基于 GPT-4 的代理。Zhang 和 Zhang [2023] 引入 Auto-UI，一种结合图像-语言编码解码架构与动作链策略的多模态解决方案，在 AitW 数据集上微调。该动作链捕捉了中间的历史动作和未来动作计划。Yang et al. [2024b] 提出数据中心化流程，从公开数据生成高质量泛化数据，用于微调视觉语言模型 (VLM)，以适应多环境下的多样化指令。Xu et al. [2024] 引入 AGU-VIS 的两阶段训练范式。第一阶段，代理通过自监督学习掌握 GUI 组件的视觉表示；第二阶段，利用强化学习微调交互任务，实现高效自主 GUI 交互。在基于计算机的环境中，ScreenAgent [Niu et al., 2024] 微调 ScreenAgent 数据集，将截图映射为动作序列。其通过 VNC 操作，遵循受 Kolb 体验学习启发的规划-执行-反思框架。PC-Agent [He et al., 2024b] 采用多代理架构，基于 PC Tracker 收集的认知轨迹微调规划代理，使其能够模拟人类认知模式。此外，Kil et al. [2024] 微调 DeBERTa 用于元素排序，微调 Flan-T5 用于动作预测，结合视觉信号提升网页导航性能。

In summary, we provide a systematic overview of recent influential research on (M)LLM-based GUI agents. We address their goal formulations, input perceptions, and learning paradigms, as shown in Table 1

总之，我们系统性地综述了近期基于 (多模态)LLM 的 GUI 代理的影响力研究，涵盖其目标设定、输入感知和学习范式，如表 1 所示。

4 Industrial Applications of (M)LLM-Based GUI Agents

4 基于 (多模态)LLM 的 GUI 代理的工业应用

GUI agents have been widely used in industrial settings, such as mobile assistants and search agents, demonstrating significant commercial value and potential.

GUI 代理已广泛应用于工业领域，如移动助手和搜索代理，展现出显著的商业价值和潜力。

Google Assistant for Android: By saying phrases like ”Hey Google, start a run on Example App,” users can use Google Assistant for Android to launch apps, perform tasks, and access content. App Actions, powered by built-in intents (BIIs), enhance app functionality by integrating with Google Assistant. This enables users to navigate apps and access features through voice queries, which the Assistant interprets to display the desired screen or widget.

Android 版 Google Assistant: 用户通过说“嘿 Google，在示例应用上开始跑步”等语句，可以使用 Android 版 Google Assistant 启动应用、执行任务和访问内容。由内置意图 (BIIs) 驱动的 App Actions 增强了应用功能，使其能与 Google Assistant 集成。用户通过语音查询导航应用并访问功能，助手解析后展示所需的界面或控件。

Apple Intelligence: Apple Intelligence is the suite of AI-powered features and services developed by Apple. This includes technologies such as machine learning, natural language processing, and computer vision that power features like Siri, facial recognition, and photo organization. Apple also integrates AI into its hardware and software ecosystem to improve device performance and user experience. Their focus on privacy means that much of this AI processing happens on-device, ensuring that user data remains secure.

Apple Intelligence:Apple Intelligence 是苹果开发的一套 AI 驱动功能和服务，包括机器学习、自然语言处理和计算机视觉等技术，支持 Siri、面部识别和照片整理等功能。苹果还将 AI 集成到其硬件和软件生态系统中，以提升设备性能和用户体验。其注重隐私，许多 AI 处理在设备端完成，确保用户数据安全。

New Bing: Microsoft's search engine is designed to offer users a more intuitive, efficient, and comprehensive search experience. Leveraging cutting-edge artificial intelligence and machine learning technologies, New Bing goes beyond traditional keyword searches to understand the context and intent behind user queries. With New Bing as an example, the LLM-based deep search engine is also an important form of GUI agents.

New Bing: 微软的搜索引擎旨在为用户提供更直观、高效且全面的搜索体验。利用尖端的人工智能和机器学习技术，New Bing 超越了传统的关键词搜索，能够理解用户查询背后的上下文和意图。以 New Bing 为例，基于大型语言模型 (LLM) 的深度搜索引擎也是图形用户界面 (GUI) 代理的重要形式。

Anthropic Computer Use: Anthropic's ”Computer Use” feature enables Claude to interact with tools and manipulate a desktop environment. By understanding and executing commands, Computer-Using Agent(CUA) can perform the necessary actions to complete tasks, much like a human.

Anthropic 计算机使用:Anthropic 的“计算机使用”功能使 Claude 能够与工具交互并操作桌面环境。通过理解和执行命令，计算机使用代理 (CUA) 能够像人类一样执行完成任务所需的操作。

OpenAI Operator: OpenAI recently introduced Operator, an AI agent capable of autonomously performing tasks using its own browser. This agent leverages the CUA model, which combines GPT-4o's vision capabilities with advanced reasoning through reinforcement learning. Operator can interpret screenshots and interact with GUIs —such as buttons, menus, and text fields—just as humans do. This development marks a significant advancement in AI capabilities, enabling more efficient and autonomous interactions with digital interfaces.

OpenAI 操作员:OpenAI 最近推出了 Operator，一种能够自主使用自身浏览器执行任务的 AI 代理。该代理采用了 CUA 模型，结合了 GPT-4o 的视觉能力和通过强化学习实现的高级推理。Operator 能够解读截图并与图形用户界面 (如按钮、菜单和文本框) 交互，方式与人类相同。这一发展标志着 AI 能力的重大进步，使得与数字界面的交互更加高效和自主。

Microsoft Copilot: An AI tool in Microsoft 365 apps for productivity with GPT-based suggestions, task automation, and content generation. Enhances workflows, creativity, and decision-making with real-time insights.

微软 Copilot: 微软 365 应用中的一款 AI 工具，基于 GPT 提供建议、任务自动化和内容生成，提升工作流程、创造力和决策能力，提供实时洞察。

AutoGLM: AutoGLM [Liu et al., 2024] is designed for autonomous mission completion via GUIs on platforms like phones and the web. Its Android capability allows it to understand user instructions autonomously without manual input, enabling it to handle complex tasks such as ordering takeout, editing comments, shopping, and summarizing articles.

AutoGLM:AutoGLM [Liu et al., 2024] 旨在通过手机和网页等平台上的图形用户界面自主完成任务。其安卓功能使其能够自主理解用户指令，无需人工输入，从而处理复杂任务，如订外卖、编辑评论、购物和文章摘要。

MagicOS 9.0 YOYO: An advanced assistant with four main features: natural language and vision processing, user behavior learning, intent recognition and decision-making, and seamless app integration. It understands user habits to autonomously fulfill requests, such as ordering coffee through voice commands, by navigating apps and services.

MagicOS 9.0 YOYO: 一款先进助手，具备四大核心功能: 自然语言与视觉处理、用户行为学习、意图识别与决策，以及无缝应用集成。它理解用户习惯，能够自主完成请求，例如通过语音命令点咖啡，自动导航应用和服务。

5 Challenges

5 个挑战

Due to the rapid development of this field, we summarize several key research questions that require urgent attention:

鉴于该领域的快速发展，我们总结了几个亟需关注的关键研究问题:

Personalized GUI Agents: Due to the personal nature of user devices, GUI agents inherently interact with personalized information. As an example, users may commute from home to work during weekdays, while walking to their favorite restaurants and cafes on weekends. The integration of personalized information would clearly enhance the user experience with GUI agents. As the capabilities of (M)LLMs continue to improve, personalized GUI agents have become a priority. Effectively collecting and utilizing personal information to deliver a more intelligent experience for users is an essential topic for future research and applications.

个性化 GUI 代理: 由于用户设备的个人属性，GUI 代理天然涉及个性化信息。例如，用户在工作日可能从家到公司通勤, 而周末则步行前往喜欢的餐厅和咖啡馆。整合个性化信息显然能提升 GUI 代理的用户体验。随着 (M)LLM 能力的持续提升，个性化 GUI 代理已成为优先发展方向。有效收集和利用个人信息，为用户提供更智能的体验，是未来研究和应用的重要课题。

Security of GUI Agents: GUI devices play a crucial role in modern life, making the idea of allowing GUI agents to take control a significant concern for users. For instance, improper operations in financial apps could lead to substantial financial losses, while inappropriate comments on social media apps could damage one's reputation and privacy. Ensuring that GUI agents are not only highly efficient and capable of generalizing but also uphold user-specific security and provide transparency about their actions is an urgent research challenge. This is a critical issue, as it directly impacts the viability of applying GUI agents in real-world scenarios.

GUI 代理的安全性:GUI 设备在现代生活中扮演关键角色，允许 GUI 代理控制设备的想法令用户高度关注。例如，金融应用中的不当操作可能导致重大经济损失，社交媒体应用中的不当评论可能损害声誉和隐私。确保 GUI 代理不仅高效且具备泛化能力，同时维护用户特定的安全性并对其行为保持透明，是一项紧迫的研究挑战。这一问题至关重要，直接影响 GUI 代理在现实场景中的应用可行性。

Inference Efficiency: Humans are highly sensitive to GUI response time, which significantly impacts the user experience. Current (M)LLM-based GUI agents still face notable drawbacks with inference latency. Additionally, communication delay is also an important consideration in real-world applications. As a result, efficient device-cloud collaboration strategies and effective device-side (M)LLM research will become critical areas of focus in the future.

推理效率: 人类对 GUI 响应时间极为敏感，这对用户体验有显著影响。目前基于 (M)LLM 的 GUI 代理在推理延迟方面仍存在明显不足。此外，通信延迟也是现实应用中的重要考量。因此，高效的设备-云协作策略和有效的设备端 (M)LLM 研究将成为未来的关键关注点。

6 Conclusion

6 结论

In this paper, we provide a comprehensive review of the rapidly evolving field of (M)LLM-based GUI Agents. The review is organized into three main perspectives: Data Resources, Frameworks, and Applications. Additionally, we present a detailed taxonomy that connects existing research and highlights key techniques. We also discuss several challenges and propose potential future directions for GUI Agents that leverage foundation models.

本文全面回顾了快速发展的基于 (M)LLM 的 GUI 代理领域。综述从数据资源、框架和应用三个主要视角展开，此外还提出了一个详细的分类体系，连接现有研究并突出关键技术。我们还讨论了若干挑战，并提出了利用基础模型的 GUI 代理未来可能的发展方向。

References

参考文献

[Achiam et al., 2024] Josh Achiam, Steven Adler, et al. Gpt- 4 technical report, 2024.

Josh Achiam, Steven Adler 等。Gpt-4 技术报告，2024 年。

[Bai et al., 2021] Chen Bai, Xiaoyu Zang, Yan Xu, Srinivas Sunkara, Abhinav Rastogi, and Jieshan Chen. Uibert: Learning generic multimodal representations for ui understanding, 2021.

陈白、臧晓宇、徐岩、Srinivas Sunkara、Abhinav Rastogi 和陈杰山。Uibert: 用于 UI 理解的通用多模态表示学习，2021 年。

[Chen et al., 2024a] Dongping Chen, Yue Huang, Siyuan Wu, et al. Gui-world: A dataset for gui-oriented multimodal llm-based agents. arXiv preprint arXiv:2406.10819, 2024.

陈东平, 黄越, 吴思远等. Gui-world: 面向图形用户界面 (GUI) 的多模态大语言模型 (LLM) 代理数据集. arXiv 预印本 arXiv:2406.10819, 2024.

[Chen et al., 2024b] Jingxuan Chen, Derek Yuen, Bin Xie, et al. Spa-bench: A comprehensive benchmark for smart-phone agent evaluation. In NeurIPS 2024 Workshop on Open-World Agents, 2024.

陈景轩, Derek Yuen, 谢斌等. Spa-bench: 智能手机代理评估的综合基准. 载于 NeurIPS 2024 开放世界代理研讨会, 2024.

[Ding, 2024] Tinghe Ding. Mobileagent: enhancing mobile control via human-machine interaction and sop integration. arXiv preprint arXiv:2401.04124, 2024.

丁廷和. Mobileagent: 通过人机交互和标准操作程序 (SOP) 集成增强移动控制. arXiv 预印本 arXiv:2401.04124, 2024.

[Furuta et al., 2024] Hiroki Furuta, Kuang-Huei Lee, Ofir Nachum, et al. Multimodal web navigation with instruction-finetuned foundation models. In ICLR, 2024.

古田浩树, 李光辉, Ofir Nachum 等. 基于指令微调基础模型的多模态网页导航. 载于 ICLR, 2024.

[Ge et al., 2024] Zhiqi Ge, Juncheng Li, Xinglei Pang, et al. Iris: Breaking gui complexity with adaptive focus and self-refining. arXiv preprint arXiv:2412.10342, 2024.

葛志奇, 李俊成, 庞兴磊等. Iris: 通过自适应聚焦与自我优化破解图形用户界面复杂性. arXiv 预印本 arXiv:2412.10342, 2024.

[He et al., 2024a] Hongliang He, Wenlin Yao, Kaixin Ma, et al. Webvoyager: Building an end-to-end web agent with large multimodal models. arXiv preprint arXiv:2401.13919, 2024.

何洪亮, 姚文林, 马凯鑫等. Webvoyager: 利用大型多模态模型构建端到端网页代理. arXiv 预印本 arXiv:2401.13919, 2024.

[He et al., 2024b] Yanheng He, Jiahe Jin, Shijie Xia, et al. Pc agent: While you sleep, ai works-a cognitive journey into digital world. arXiv preprint arXiv:2412.17589, 2024.

何彦恒, 金佳禾, 夏世杰等. PC 代理: 当你沉睡时，人工智能在数字世界中认知探索. arXiv 预印本 arXiv:2412.17589, 2024.

[Hong et al., 2024] Wenyi Hong, Weihan Wang, Qingsong Lv, et al. Cogagent: A visual language model for gui agents. In CVPR, pages 14281-14290, 2024.

洪文怡, 王伟涵, 吕庆松等. Cogagent: 面向图形用户界面代理的视觉语言模型. 载于 CVPR, 页码 14281-14290, 2024.

[Kil et al., 2024] Jihyung Kil, Chan Hee Song, Boyuan Zheng, Xiang Deng, Yu Su, and Wei-Lun Chao. Dual-view visual contextualization for web navigation. In CVPR, pages 14445-14454, 2024.

吉炯, 宋灿熙, 郑博远, 邓翔, 苏宇, 赵伟伦. 双视角视觉上下文用于网页导航. 载于 CVPR, 页码 14445-14454, 2024.

[Kim et al., 2023] Geunwoo Kim, Pierre Baldi, and Stephen McAleer. Language models can solve computer tasks. In NIPS, pages 39648-39677, 2023.

金根宇, Pierre Baldi, Stephen McAleer. 语言模型能够解决计算机任务. 载于 NIPS, 页码 39648-39677, 2023.

[Lai et al., 2024] Hanyu Lai, Xiao Liu, Iat Long Iong, et al. Autowebglm: A large language model-based web navigating agent. In SIGKDD, pages 5295-5306, 2024.

赖涵宇, 刘晓, Iat Long Iong 等. Autowebglm: 基于大型语言模型的网页导航代理. 载于 SIGKDD, 页码 5295-5306, 2024.

[Lee et al., 2023a] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, et al. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In ICML, pages 18893- 18912, 2023.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc 等. Pix2struct: 将截图解析作为视觉语言理解的预训练. 载于 ICML, 页码 18893-18912, 2023.

[Lee et al., 2023b] Sunjae Lee, Junyoung Choi, Jungjae Lee, et al. Explore, select, derive, and recall: Augmenting llm with human-like memory for mobile task automation. arXiv preprint arXiv:2312.03003, 2023.

Sunjae Lee, Junyoung Choi, Jungjae Lee 等. 探索、选择、推导与回忆: 以类人记忆增强大语言模型实现移动任务自动化. arXiv 预印本 arXiv:2312.03003, 2023.

[Li and Li, 2023] Gang Li and Yang Li. Spotlight: Mobile ui understanding using vision-language models with a focus. In ICLR, 2023.

李刚, 李洋. Spotlight: 利用视觉语言模型聚焦移动用户界面理解. 载于 ICLR, 2023.

[Li et al., 2020] Yang Li, Jiacong He, Xin Zhou, Yuan Zhang, and Jason Baldridge. Mapping natural language instructions to mobile ui action sequences. In ACL , pages 8198-8210, 2020.

李洋, 何家聪, 周鑫, 张远, Jason Baldridge. 将自然语言指令映射为移动用户界面操作序列. 载于 ACL , 页码 8198-8210, 2020.

[Li et al., 2021] Yang Li, Gang Li, Xin Zhou, Mostafa De-hghani, and Alexey Gritsenko. Vut: Versatile ui transformer for multi-modal multi-task user interface modeling. arXiv preprint arXiv:2112.05692, 2021.

李洋, 李刚, 周鑫, Mostafa Dehghani, Alexey Gritsenko. VUT: 多模态多任务用户界面建模的多功能用户界面变换器. arXiv 预印本 arXiv:2112.05692, 2021.

[Li et al., 2024a] Wei Li, William Bishop, Alice Li, et al. On the effects of data scale on computer control agents. arXiv preprint arXiv:2406.03679, 2024.

李伟，威廉·毕晓普，爱丽丝·李等。数据规模对计算机控制代理影响的研究。arXiv 预印本 arXiv:2406.03679，2024 年。

[Li et al., 2024b] Yanda Li, Chi Zhang, Wanqi Yang, et al. Appagent v2: Advanced agent for flexible mobile interactions. arXiv preprint arXiv:2408.11824, 2024.

李艳达, 张驰, 杨万琦等。Appagent v2: 用于灵活移动交互的高级代理。arXiv 预印本 arXiv:2408.11824, 2024 年。

[Liu et al., 2018] Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, and Percy Liang. Reinforcement learning on web interfaces using workflow-guided exploration. In ICLR, 2018.

刘泽然，Kelvin Guu，潘普蓬·帕苏帕特，石天林，梁珀西。基于工作流引导探索的网页界面强化学习。发表于 ICLR，2018 年。

[Liu et al., 2024] Xiao Liu, Bo Qin, Dongzhu Liang, et al. Autoglm: Autonomous foundation agents for guis. arXiv preprint arXiv:2411.00820, 2024.

刘晓，秦博，梁东竹等。AutoGLM: 面向图形用户界面 (GUI) 的自主基础代理。arXiv 预印本 arXiv:2411.00820，2024 年。

[Lu et al., 2024a] Quanfeng Lu, Wenqi Shao, Zitao Liu, et al. Gui odyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. arXiv preprint arXiv:2406.08451, 2024.

陆全峰，邵文琦，刘子涛等。GUI Odyssey: 面向移动设备跨应用 GUI 导航的综合数据集。arXiv 预印本 arXiv:2406.08451，2024 年。

[Lu et al., 2024b] Yadong Lu, Jianwei Yang, Yelong Shen, and Ahmed Awadallah. Omniparser for pure vision based gui agent. arXiv preprint arXiv:2408.00203, 2024.

陆亚东，杨建伟，沈业龙，Ahmed Awadallah。Omniparser: 基于纯视觉的 GUI 代理。arXiv 预印本 arXiv:2408.00203，2024 年。

[Ma et al., 2023] Kaixin Ma, Hongming Zhang, Hongwei Wang, Xiaoman Pan, and Dong Yu. Laser: Llm agent with state-space exploration for web navigation. In NeurIPS Workshop, 2023.

马凯新，张宏明，王宏伟，潘晓曼，余东。LASER: 结合状态空间探索的语言模型 (LLM) 代理用于网页导航。发表于 NeurIPS 研讨会，2023 年。

[Nakano et al., 2021] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, et al. Webgpt: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332, 2021.

中野礼一郎, 雅各布·希尔顿, Suchir Balaji 等。WebGPT: 基于浏览器辅助的人类反馈问答系统。arXiv 预印本 arXiv:2112.09332，2021 年。

[Niu et al., 2024] Runliang Niu, Jindong Li, Shiqi Wang, et al. Screenagent: A vision language model-driven computer control agent. arXiv preprint arXiv:2402.07945, 2024.

牛润良，李金东，王世奇等。ScreenAgent: 基于视觉语言模型的计算机控制代理。arXiv 预印本 arXiv:2402.07945，2024 年。

[Pan et al., 2023] Lihang Pan, Bowen Wang, Chun Yu, Yux-uan Chen, Xiangyu Zhang, and Yuanchun Shi. Autotask: Executing arbitrary voice commands by exploring and learning from mobile gui. arXiv preprint arXiv:2312.16062, 2023.

潘立航，王博文，余春，陈宇轩，张翔宇，史元春。AutoTask: 通过探索和学习移动 GUI 执行任意语音命令。arXiv 预印本 arXiv:2312.16062，2023 年。

[Putta et al., 2024] Pranav Putta, Edmund Mills, Naman Garg, et al. Agent q: Advanced reasoning and learning for autonomous ai agents. arXiv preprint arXiv:2408.07199, 2024.

普拉纳夫·普塔, 埃德蒙·米尔斯, 纳曼·加格等。Agent Q: 面向自主 AI 代理的高级推理与学习。arXiv 预印本 arXiv:2408.07199，2024 年。

[Qin et al., 2025] Yujia Qin, Yining Ye, Junjie Fang, et al. Ui-tars: Pioneering automated gui interaction with native agents. arXiv preprint arXiv:2501.12326, 2025.

秦宇佳，叶一宁，方俊杰等。UI-TARS: 开创性的原生代理自动化 GUI 交互。arXiv 预印本 arXiv:2501.12326，2025 年。

[Rawles et al., 2023] Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy P Lillicrap. Androidinthewild: A large-scale dataset for android device control. In NIPS Datasets and Benchmarks Track, 2023.

克里斯托弗·罗尔斯，爱丽丝·李，丹尼尔·罗德里格斯，奥里亚娜·里瓦，蒂莫西·P·利利克拉普。AndroidInTheWild: 面向安卓设备控制的大规模数据集。发表于 NIPS 数据集与基准赛道，2023 年。

[Rawles et al., 2024] Christopher Rawles, Sarah Clincke-maillie, Yifan Chang, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. arXiv preprint arXiv:2405.14573, 2024.

克里斯托弗·罗尔斯，Sarah Clincke-Maillie，张一凡等。AndroidWorld: 面向自主代理的动态基准测试环境。arXiv 预印本 arXiv:2405.14573，2024 年。

[Shen et al., 2024] Huawen Shen, Chang Liu, Gengluo Li, et al. Falcon-ui: Understanding gui before following user instructions. arXiv preprint arXiv:2412.09362, 2024.

沈华文，刘畅，李耿洛等。Falcon-UI: 在执行用户指令前理解 GUI。arXiv 预印本 arXiv:2412.09362，2024 年。

[Sun et al., 2022] Liangtai Sun, Xingyu Chen, Lu Chen, Tianle Dai, Zichen Zhu, and Kai Yu. Meta-gui: Towards multi-modal conversational agents on mobile gui. In EMNLP, pages 6699-6712, 2022.

孙良泰，陈星宇，陈璐，戴天乐，朱子辰，余凯。Meta-GUI: 迈向移动 GUI 上的多模态对话代理。发表于 EMNLP，页码 6699-6712，2022 年。

[Touvron et al., 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard 等. Llama: 开放且高效的基础语言模型. arXiv 预印本 arXiv:2302.13971, 2023.

[Venkatesh et al., 2024] Sagar Gubbi Venkatesh, Partha Talukdar, and Srini Narayanan. Ugif-dataset: A new dataset for cross-lingual, cross-modal sequential actions on the ui. In Findings of NAACL, pages 1390-1399, 2024.

Sagar Gubbi Venkatesh, Partha Talukdar, Srini Narayanan. Ugif-dataset: 用于跨语言、跨模态用户界面顺序动作的新数据集. 发表在 NAACL 研究成果, 页码 1390-1399, 2024.

[Wang et al., 2021] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. Screen2words: Automatic mobile ui summarization with multimodal learning. In UIST, pages 498-510, 2021.

Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, Yang Li. Screen2words: 基于多模态学习的自动移动界面摘要生成. 发表在 UIST, 页码 498-510, 2021.

[Wang et al., 2023] Bryan Wang, Gang Li, and Yang Li. Enabling conversational interaction with mobile ui using large language models. In CHI, pages 1-17, 2023.

Bryan Wang, Gang Li, Yang Li. 利用大型语言模型实现移动界面的对话式交互. 发表在 CHI, 页码 1-17, 2023.

[Wang et al., 2024a] Junyang Wang, Haiyang Xu, Haitao Jia, et al. Mobile-agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration. In NIPS, 2024.

Junyang Wang, Haiyang Xu, Haitao Jia 等. Mobile-agent-v2: 通过多智能体协作实现高效导航的移动设备操作助手. 发表在 NIPS, 2024.

[Wang et al., 2024b] Junyang Wang, Haiyang Xu, Jiabo Ye, et al. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. arXiv preprint arXiv:2401.16158, 2024.

Junyang Wang, Haiyang Xu, Jiabo Ye 等. Mobile-agent: 具备视觉感知的自主多模态移动设备代理. arXiv 预印本 arXiv:2401.16158, 2024.

[Wang et al., 2024c] Lei Wang, Chen Ma, Xueyang Feng, et al. A survey on large language model based autonomous agents. FCS, 18(6):186345, 2024.

Lei Wang, Chen Ma, Xueyang Feng 等. 基于大型语言模型的自主智能体综述. FCS, 18(6):186345, 2024.

[Wang et al., 2025] Zhenhailong Wang, Haiyang Xu, Jun-yang Wang, et al. Mobile-agent-e: Self-evolving mobile assistant for complex tasks. arXiv preprint arXiv:2501.11733, 2025.

Zhenhailong Wang, Haiyang Xu, Jun-yang Wang 等. Mobile-agent-e: 面向复杂任务的自我进化移动助手. arXiv 预印本 arXiv:2501.11733, 2025.

[Wen et al., 2023a] Hao Wen, Yuanchun Li, Guohong Liu, et al. Empowering LLM to use Smartphone for Intelligent Task Automation. arXiv preprint arXiv:2308.15272, 2023.

Hao Wen, Yuanchun Li, Guohong Liu 等. 赋能大型语言模型使用智能手机实现智能任务自动化. arXiv 预印本 arXiv:2308.15272, 2023.

[Wen et al., 2023b] Hao Wen, Hongming Wang, Jiaxuan Liu, and Yuanchun Li. Droidbot-gpt: Gpt-powered ui automation for android. arXiv preprint arXiv:2304.07061, 2023.

Hao Wen, Hongming Wang, Jiaxuan Liu, Yuanchun Li. Droidbot-gpt: 基于 GPT 的安卓界面自动化工具. arXiv 预印本 arXiv:2304.07061, 2023.

[Wen et al., 2024] Hao Wen, Yuanchun Li, Guohong Liu, et al. Autodroid: Llm-powered task automation in android. In MobiCom, pages 543-557, 2024.

Hao Wen, Yuanchun Li, Guohong Liu 等. Autodroid: 基于大型语言模型的安卓任务自动化. 发表在 MobiCom, 页码 543-557, 2024.

[Wu et al., 2024] Qinzhuo Wu, Weikai Xu, Wei Liu, et al. Mobilevlm: A vision-language model for better intra-and inter-ui understanding. In EMNLP, pages 10231-10251, 2024.

Qinzhuo Wu, Weikai Xu, Wei Liu 等. Mobilevlm: 提升界面内外理解的视觉语言模型. 发表在 EMNLP, 页码 10231-10251, 2024.

[Xie et al., 2024] Tianbao Xie, Fan Zhou, et al. Openagents: An open platform for language agents in the wild. In ICLR 2024 Workshop on Large Language Model (LLM) Agents, 2024.

Tianbao Xie, Fan Zhou 等. Openagents: 面向实际应用的开放语言智能体平台. 发表在 ICLR 2024 大型语言模型 (LLM) 智能体研讨会, 2024.

[Xu et al., 2024] Yiheng Xu, Zekun Wang, Junli Wang, et al. Aguvis: Unified pure vision agents for autonomous gui interaction. arXiv preprint arXiv:2412.04454, 2024.

Yiheng Xu, Zekun Wang, Junli Wang 等. Aguvis: 统一的纯视觉智能体实现自主图形界面交互. arXiv 预印本 arXiv:2412.04454, 2024.

[Yan et al., 2023] An Yan, Zhengyuan Yang, Wanrong Zhu, et al. Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation. arXiv preprint arXiv:2311.07562, 2023.

An Yan, Zhengyuan Yang, Wanrong Zhu 等. GPT-4V 奇境: 用于零样本智能手机图形界面导航的大型多模态模型. arXiv 预印本 arXiv:2311.07562, 2023.

[Yang et al., 2024a] An Yang, Baosong Yang, et al. Qwen2 technical report, 2024.

An Yang, Baosong Yang 等. Qwen2 技术报告, 2024.

[Yang et al., 2024b] Yuhao Yang, Yue Wang, Dongxu Li, et al. Aria-ui: Visual grounding for gui instructions. arXiv preprint arXiv:2412.16256, 2024.

Yuhao Yang, Yue Wang, Dongxu Li 等. Aria-ui: 用于图形用户界面指令的视觉定位。arXiv 预印本 arXiv:2412.16256, 2024.

[You et al., 2024] Keen You, Haotian Zhang, Eldon Schoop, et al. Ferret-ui: Grounded mobile ui understanding with multimodal llms. In ECCV, pages 240-255, 2024.

Keen You, Haotian Zhang, Eldon Schoop 等. Ferret-ui: 基于多模态大语言模型的移动界面理解。发表于 ECCV, 页码 240-255, 2024.

[Zhang and Zhang, 2023] Zhuosheng Zhang and Aston Zhang. You only look at screens: Multimodal chain-of-action agents. arXiv preprint arXiv:2309.11436, 2023.

Zhuosheng Zhang 和 Aston Zhang. 你只需看屏幕: 多模态动作链代理。arXiv 预印本 arXiv:2309.11436, 2023.

[Zhang et al., 2021] Xiaoyi Zhang, Lilian De Greef, Amanda Swearngin, et al. Screen recognition: Creating accessibility metadata for mobile applications from pixels. In CHI, pages 1-15, 2021.

Xiaoyi Zhang, Lilian De Greef, Amanda Swearngin 等. 屏幕识别: 从像素创建移动应用的无障碍元数据。发表于 CHI, 页码 1-15, 2021.

[Zhang et al., 2023] Chi Zhang, Zhao Yang, Jiaxuan Liu, et al. AppAgent: Multimodal Agents as Smartphone Users. arXiv preprint arXiv:2312.13771, 2023.

Chi Zhang, Zhao Yang, Jiaxuan Liu 等. AppAgent: 作为智能手机用户的多模态代理。arXiv 预印本 arXiv:2312.13771, 2023.

[Zhang et al., 2024a] Chaoyun Zhang, Liqun Li, Shilin He, et al. Ufo: A ui-focused agent for windows os interaction. arXiv preprint arXiv:2402.07939, 2024.

Chaoyun Zhang, Liqun Li, Shilin He 等. UFO: 面向 Windows 操作系统交互的界面代理。arXiv 预印本 arXiv:2402.07939, 2024.

[Zhang et al., 2024b] Jiwen Zhang, Jihao Wu, Yihua Teng, et al. Android in the zoo: Chain-of-action-thought for gui agents. arXiv preprint arXiv:2403.02713, 2024.

Jiwen Zhang, Jihao Wu, Yihua Teng 等. Android 动物园: 用于图形用户界面代理的动作-思考链。arXiv 预印本 arXiv:2403.02713, 2024.

[Zhao et al., 2024] Kangjia Zhao, Jiahui Song, Leigang Sha, et al. Gui testing arena: A unified benchmark for advancing autonomous gui testing agent. arXiv preprint arXiv:2412.18426, 2024.

Kangjia Zhao, Jiahui Song, Leigang Sha 等. GUI 测试竞技场: 推进自主 GUI 测试代理的统一基准。arXiv 预印本 arXiv:2412.18426, 2024.

[Zheng et al., 2024] Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v (ision) is a generalist web agent, if grounded. arXiv preprint arXiv:2401.01614, 2024.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun 和 Yu Su. GPT-4V(视觉) 是一个通用的网络代理，前提是有定位支持。arXiv 预印本 arXiv:2401.01614, 2024.

[Zhou et al., 2023] Shuyan Zhou, Frank F Xu, Hao Zhu, et al. Webarena: A realistic web environment for building autonomous agents. In ICLR, 2023.

Shuyan Zhou, Frank F Xu, Hao Zhu 等. Webarena: 构建自主代理的真实网络环境。发表于 ICLR, 2023.

[Zhu et al., 2024] Zichen Zhu, Hao Tang, Yansi Li, et al. Moba: A two-level agent system for efficient mobile task automation. arXiv preprint arXiv:2410.13757, 2024.

Zichen Zhu, Hao Tang, Yansi Li 等. MOBA: 高效移动任务自动化的两级代理系统。arXiv 预印本 arXiv:2410.13757, 2024.