CS598 JY2 Final Survey Report - Multimodal Web Agents

Vincent Han *vh12@illinois.edu*
vh12

Rui Pan *ruip4@illinois.edu*
ruip4

Rui Yang *ry21@illinois.edu*
ry21

Abstract

In recent years, the rapid advancement of multimodal large language models (MLLMs) has driven significant progress in developing autonomous agents capable of interacting with dynamic web environments. Multimodal web agents represent a promising research direction, leveraging MLLMs to process and integrate diverse inputs—such as text and images—to perform complex web-based tasks, including navigation, information retrieval, and plan execution. This paper provides a comprehensive survey of related works, encompassing MLLMs, web agent benchmarks, and multimodal web agent algorithms. Specifically, we begin by outlining the foundational background of these agents, with a focus on MLLMs and web agent. We then introduce benchmark datasets and evaluation metrics designed to assess performance in real-world web interactions. Furthermore, we systematically review and summarize current multimodal web agent algorithms, categorizing them into prompting-based and learning-based approaches. We also discuss key techniques utilized in multimodal web agent development, such as grounding, trajectory data curation, multi-stage fine-tuning, and reinforcement learning. This review aims to provide a comprehensive foundation for researchers and practitioners working at the intersection of MLLMs, human-computer interaction, and web agents.

1 Introduction

As the concept of Artificial General Intelligence (AGI) gains increasingly widespread attention, numerous attempts have been made towards achieving this ultimate goal. Based on OpenAI's planned five steps to AGI and the latest status of state-of-the-art models (Guo et al., 2025; OpenAI, 2025a; Chervonyi et al., 2025), the research field is currently in the transition phase from strong reasoning models (Level 2) to powerful autonomous agents (Level 3), where the model is capable of taking actions to complete certain tasks. Among them, GUI-based agents (Qin et al., 2025; OpenAI, 2025b) have shown great promise, given the advantages of unrestricted interactable environments offered by the Internet and desktop applications. On a personal note, these agents greatly boost the productivity of computer-based tasks, allowing a much more efficient way for information retrieval, presentation preparation, ticket booking, online social interaction, and numerous other commercial applications. On a professional level, almost all the existing GUI-based agents have a strong dependence on the multimodal large language models and web-based agents (Xie et al., 2024), calling for further investigation and improvements of the underlying techniques.

2 Background

2.1 Multimodal LLMs

Multimodal large language models take input besides text, allowing the processing and understanding of more diverse data sources, such as images, audio, videos, and tabular data (Liang et al., 2024) (Zhang

et al., 2024). The realization of multimodal LLM's represents a pivotal advancement towards the attainment of AGI, since for an agent to achieve full autonomy, it is necessary for it to understand and process a full variety of input sources that are not confined to mere text. One of the most important early attempts in this direction can be traced back to CLIP (Radford et al., 2021), which learns the similarity between arbitrary image and paired text caption via contrastive learning, leading to shared embedding representations for text and image. Flamingo (Alayrac et al., 2022) further extends CLIP by replacing its text-encoder with a language model, empowering the model to generate text outputs based on given images. However, Flamingo's full-parameter training paradigm turns out to be less efficient, where most modern approaches (Li et al., 2023; Zhu et al., 2023; Liu et al., 2024b) adopt a more parameter-efficient tuning fashion with a tunable MLP projector, leaving the pretrained image-encoder and text-decoder frozen.

CLIP marks the beginning of a sequence of advances in multimodal LLMs. In addition to being trained on a vast corpus of image-text pairs, multimodal LLMs have extended themselves to incorporate video and audio processing, with the emergence of key architecture such as Minigpt-4 (Zhu et al., 2023), LLaVA (Liu et al., 2024b), and Video-LLaVA (Sally Paganin, 2023). This paradigm has been generalized to handle more modality (Wu et al., 2023), enabling modern LLMs to accept various types of inputs and outputs. These models further prove to be a powerful advancement in the concept of multimodality, as they likewise incorporate novel training techniques, such as the use of better pretraining datasets, and few-shot and in-context learning capabilities.

2.2 Agents

Traditional LLMs have made huge progresses towards AGI, and excel in conversation-styled interactions, such as instruction-following (Zhou et al., 2023a) and question-answering tasks (Hendrycks et al., 2020; Zhong et al., 2023). However, this kind of ability fall short in solving domain-specific problems that require complex arithmetics, recursive factual look-up, or other API-dependent functionalities. To address this shortcoming, tool-based approaches have been proposed (Qin et al., 2023; Schick et al., 2023; Patil et al., 2023), endowing LLMs the ability to call external functions, such as Python executor for math and coding problems (Jimenez et al., 2023), or web API for factual look-up (Yao et al., 2022). In particular, web-based approaches (Zhou et al., 2023c; Yang et al., 2023) allow agents to interact with the Internet, enabling powerful agentic applications with practical benefits. Nevertheless, most web agents in this direction are still text-based, leaving room for the development of multimodal web agents.

A key distinction that is necessary for the advancement of AGI is the adjective "agenic", which distinguishes LLMs from LLM agents by introducing an additional metric as to how "agent-like" a model is. It is not sufficient to define an "agent" to simply be a model that is able to passively respond to the environment (with the thermometer being the obvious counterexample): instead, there are numerous metrics by which we can understand how "agenic" a model is. Generally, AI agents that are "agenic" tend to be situated in more complex environments and pursue sophisticated goals, are able to act in a more autonomous manner instead of being fully dictated by natural language instructions, and exploit use of planning and system design (Kapoor et al., 2024). This distinguishes an agent from a reactive chatbot. For instance, AutoGPT and ReAct (Yao et al., 2022) showcase agents that generate subgoals and iterate through reasoning-action loops. Similarly, browser-based agents demonstrate how language models can be embedded in event-driven frameworks, receiving visual or DOM-based inputs and generating action sequences.

However, many current web agents remain text-only, processing HTML or DOM representations without accessing visual or multimodal context. This bottleneck limits their robustness in real-world web navigation, where interactive components (e.g., buttons, banners, dropdowns) are often visual or dynamically rendered (Xue et al., 2025). In this context, multimodal agents represent the next leap in agent development. These agents combine the reasoning power of LLMs with the perceptual capacity of vision-language models, enabling more generalizable and human-like interaction with digital environments (Zheng et al., 2024). In short, as LLMs evolve into true agents, research focus is shifting from purely linguistic proficiency to behavioral competence, situational awareness, and decision-making—hallmarks of artificial general intelligence.

2.3 Text-Only Web Agents

Text-only web agents are agents that, using an LLM as an intermediary layer of reasoning, browse the internet and perform text-only tasks such as crawling, parsing, and extracting (Zhou et al., 2023b). Text-only web agents, though constrained by their singularity in modality, nevertheless do represent a major advancement towards the realization of AGI, as it could access, navigate, and interact with websites on WWW using natural language instructions (Deng et al., 2023a; Xue et al., 2025). They exploit the strength of LLMs in parsing content on websites, the vast majority of which is in the modality of natural language text, which indeed turns out to be the strength of LLMs, as LLMs have already demonstrated great prowess in tasks such as question-answering, instruction following, summarization, and conversational interactions (Zhou et al., 2023a). Thus, text-only web agents already could handle complex tasks such as finding the cheapest flight, booking a hotel, or extracting useful information from encyclopedic content online (Tur et al., 2025). Indeed, web agents themselves already mark an important step towards AGI, as gaining access to the world wide web, the largest information system in the world, enables a plethora of functionality and actions once considered beyond the reach of language models. But, as the name "text-only" suggests, the prowess of this type of model is necessarily constrained only to the modality of text and, thus, cannot handle images, videos, or other forms of content, which likewise are featured abundantly online (Xue et al., 2025; Zhu et al., 2025).

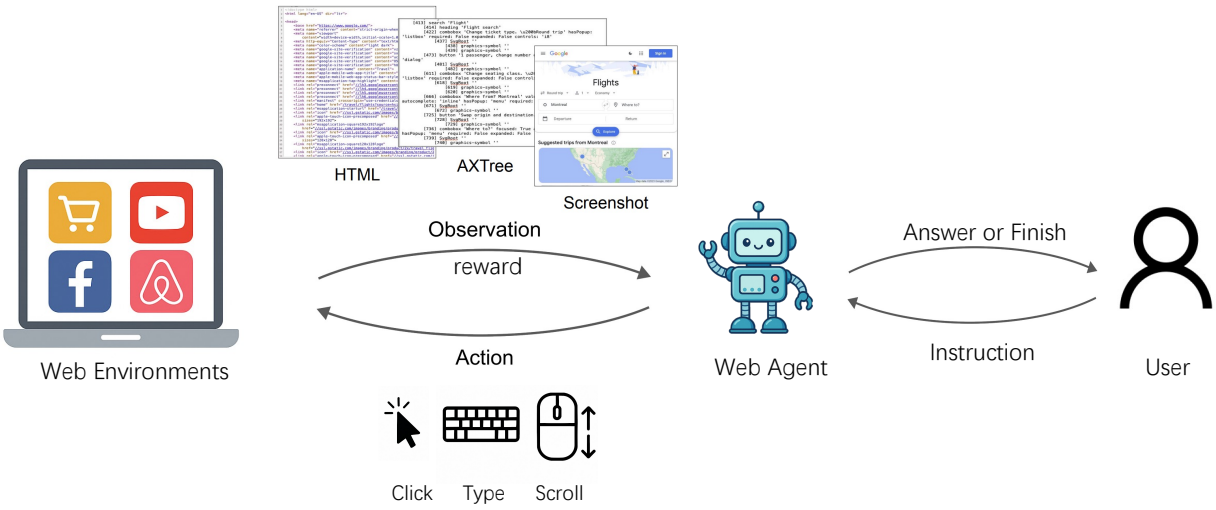

Figure 1: Illustration of the web agent interaction framework.

3 Datasets and Benchmarks for Multimodal Web Agents

To facilitate the training and evaluation of web agents, this section reviews recent datasets and benchmarks designed for multimodal web agents. For clarity, we first provide an overview of the agent interaction pipeline and the common observation types, as illustrated in Figure 1 and Figure 2.

3.1 Datasets

Recent advances in datasets for multimodal web agents focus on realistic environments with increasingly large-scale data that enable visual grounding, planning, and reasoning. Table 1 provides a summary and comparison of key datasets.

Multimodal Mind2Web (Zheng et al., 2024) is a foundational large-scale static dataset for training generalist agents. It contains 2,350 tasks (1009 for train and 1341 for test) across 137 websites in 31 domains, each paired with crowdsourced action sequences and natural language instructions. Compared with the previous Mind2Web dataset (Deng et al., 2023b), this multimodal variant dataset adds webpage screenshots, enabling both DOM-based and vision-based decision-making. Actions include high-level browser

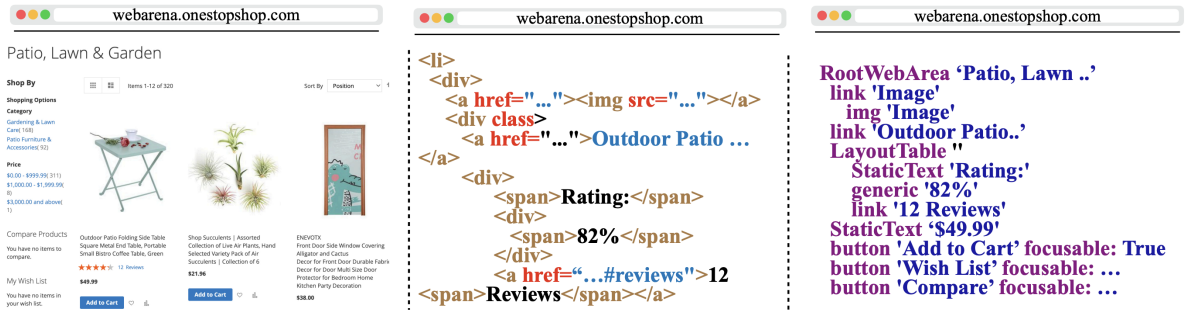

Figure 2: Illustration of the three observation types in WebArena (Zhou et al., 2023c): screenshot (left), HTML DOM tree (center), and accessibility tree (right).

Table 1: Comparison of Datasets for Multimodal Web Agents

Dataset	Task/Trajectory Number	Observation	Avg Steps	Release Time (Venue)	Description
Multimodal Mind2Web (Zheng et al., 2024)	1,009 train; 1,341 test	HTML, Screenshot	7.7	2024.01 (ICML 2024)	Large-scale static dataset of human demonstration
WebLinx (Lù et al., 2024)	1404 train; 140 valid; 1838 test	Screenshot, DOM tree	43	2024.02 (ICML 2024)	Conversational web navigation
GUIAct (Chen et al., 2024)	67K 1-step, 15K multi-step	Screenshot	1 (1-step); 7.5 (multi-step)	2024.06 (preprint)	Massive visual grounding and instruction following
UGround (Gou et al., 2024)	1.3M	Screenshot	1	2024.10 (ICLR 2025)	Synthetic data for GUI visaul grounding
AGUVIS (Xu et al., 2024b)	1M (stage 1), 35K (stage 2)	Screenshot	1 (stage 1), 8.4 (stage 2)	2024.12 (ICLR 2025)	Stage 1: grounding, Stage 2: planning
AgentTrek (Xu et al., 2024a)	10398	Screenshot, HTML, accessibility tree	12.1	2024.12 (ICLR 2025)	Synthetic large-scale data generated from web tutorials

operations such as `CLICK`, `TYPE`, and `SCROLL`, and task performance is evaluated via element success rate and trajectory success rate. With defined train/test splits, this dataset has supported a range of training and evaluation pipelines and serves as a key resource in general web agent research.

Beyond simple browser interactions, **WebLinx** (Lù et al., 2024) expands the scope to conversational web navigation, featuring significantly longer interactions—averaging 43 dialogue turns per task. The dataset includes 1,404 demonstrations for training, 140 for validation, and 1,838 for testing, collected from 155 real-world websites. Each example includes browser screenshots, DOM trees, and natural language dialogues between a user (instructor) and an assistant (navigator). This rich, multimodal setup enables the study of contextual reasoning in dynamic, dialogue-driven environments.

Focusing on large-scale grounding, **UGround** (Gou et al., 2024) contains 1.3 million GUI elements across screenshots, with each sample annotated for visual grounding. With one-step interactions, it provides synthetic supervision for pretraining models that map language instruction and screenshots to functional GUI actions, forming the foundation for versatile GUI agents.

GUIAct (Chen et al., 2024) further emphasizes vision-based multimodal learning, offering 82,000 tasks covering both desktop and mobile GUIs. These include 67K single-step and 15K multi-step instruction-following tasks grounded in screenshots with optional OCR. Agents execute structured action scripts (e.g.,

bounding box clicks or text input), and success requires precise execution of each step. The dataset supports structured generalization and long-horizon reasoning in visual UI settings.

To address GUI control across platforms, **AGUVIS** (Xu et al., 2024b) provides tens of thousands of vision-only demonstrations spanning web, desktop, and mobile interfaces. It structures data into two stages: GUI grounding (stage 1) and planning (stage 2), supporting modular training and evaluation. The dataset includes both synthetic data and human demonstrations, relying on screenshots and language instruction as input and unified GUI actions (e.g., click, type) as output. AGUVIS can thus handle purely vision-based GUI control across diverse interfaces.

Finally, **AgentTrek** (Xu et al., 2024a) introduces a large-scale synthetic dataset built from real-world web tutorials collected from the Internet. It contains 10398 trajectories covering tasks such as online shopping and form submission. Each trajectory follows a step-by-step instructional format and is represented through either HTML DOM or screenshot-based sequences. AgentTrek has proven highly effective for pretraining, showing strong downstream improvements on benchmarks like WebArena. By bridging the gap between data scale and task diversity, it offers a scalable pipeline for multimodal web agent data generation and serves as a strong foundation for training instruction-tuned web agents.

Summary. Although web navigation datasets are still limited compared to those used to pretraining foundation models, recent datasets for multimodal web agents have significantly advanced in scale, realism, and task diversity, enabling robust training across vision-language-action grounding, planning, and conversational interaction. Foundational resources like Multimodal Mind2Web provide large-scale static dataset, while UGround and GUIAct focus on precise visual grounding for GUI actions. Cross-platform datasets such as AGUVIS unify control across web, desktop, and mobile interfaces, and synthetic collections like AgentTrek demonstrate how scalable data generation can enhance pretraining and generalization. Together, these datasets address critical gaps—from low-level action execution to long-horizon reasoning. These resources move the frontier of web agent training and instrumental in developing generalist agents capable of navigating the dynamic, visually rich real-world web and executing GUI tasks.

3.2 Benchmarks

As the demand for more robust evaluation of multimodal web agents grows, a diverse set of benchmarks has emerged. These benchmarks differ in complexity, modality, and task structure, offering a rich landscape for assessing both generalist and specialized web agents. A comparison of different benchmarks is shown in Table 2.

WebArena (Zhou et al., 2023c) introduces 812 long-horizon tasks generated from 241 templates across four domains: E-commerce, Forums, Collaborative Coding, and Content Management. Agents interact with offline-hosted websites using high-level browser actions and receive DOM-based textual observations, which can optionally include the accessibility tree and screenshots. Tasks are evaluated based on functional task success, with a limit of 30 steps per task.

Visual WebArena (Koh et al., 2024a) builds on WebArena by introducing visual grounding. It includes 910 tasks across three visually rich simulated sites, including Classifieds, Shopping, and Reddit, requiring agents to interpret text and images in both user instructions and webpages. Observation, action space, and evaluation are similar to WebArena.

Multimodal Mind2Web (Zheng et al., 2024) has already been introduced above as a foundational large-scale static dataset for training. Moreover, it can also be used for evaluating GUI models with those 1341 test samples, each paired with crowdsourced action sequences and natural language instructions. The performance is evaluated via element-level and trajectory-level success rates.

Moving beyond simulated environments, **WebVoyager** (He et al., 2024a) evaluates agents on 643 tasks across 15 real-world websites, including Amazon and Wikipedia. Tasks range from simple information lookups to complex multi-step procedures, requiring agents to interact with dynamic content using native mouse and keyboard actions. Success is measured using a GPT-4V-based evaluator, which is proven to align closely with human judgment.

Table 2: Comparison of Representative Benchmarks for Multimodal Web Agents

Benchmark	Online/Offline Hosted	Task Size	Evaluation Metrics	Observation	Release Time (Venue)	Description
WebArena (Zhou et al., 2023c)	Offline	812	Functional task success	Screenshot, HTML, accessibility tree	2023.07 (ICLR 2024)	Realistic simulated websites across four domains
Visual WebArena (Koh et al., 2024a)	Offline	910	Functional task success	Screenshot, HTML, accessibility tree	2024.01 (ACL 2024)	Extension of WebArena with visually grounded tasks
WebVoyager (He et al., 2024a)	Online	643	GPT-4V-based success judgment	Screenshot	2024.01 (ACL 2024)	Open-web online benchmark across 15 websites
Multimodal Mind2Web (Zheng et al., 2024)	Static	1341	Action matching score	Screenshot, HTML	2024.01 (ICML 2024)	Non-interactable benchmark comparing with human demonstrations
WorkArena (Drouin et al., 2024)	Online	33	Functional task success	Screenshot,HTML, accessibility tree	2024.03 (ICML 2024)	Enterprise web tasks simulating realistic knowledge-work scenarios
Mind2Web-Live (Pan et al., 2024b)	Online	542	Functional step and task success	Screenshot, HTML	2024.06 (ICLR 2025)	Online benchmark with intermediate and final success measure
AssistantBench (Yoran et al., 2024)	Online	214	Answer matching	Screenshot, HTML	2024.07 (EMNLP 2024)	Complex and realistic web tasks that can be automatically verified
WorkArena++ (Boisvert et al., 2024)	Online	682	Functional task success	Screenshot, HTML, accessibility tree	2024.07 (NeurIPS 2024)	Large-scale benchmark for reasoning-intensive, multi-step enterprise workflows

WebCanvas (Pan et al., 2024b) introduces **Mind2Web-Live**, an online benchmark adapted from the original Mind2Web tasks. It contains 542 real-world web tasks designed for continual evaluation in live environments. Unlike static benchmarks, agents interact with dynamic, evolving contents. A novel evaluation metric rewards agents for matching key intermediate states, not just final outcomes, providing a more granular evaluation. A potential limitation, however, is the variability of live websites over time, which may affect the consistency and accuracy of evaluations.

Following previous benchmarks, a series of recent works focus on more challenging tasks in realistic settings or enterprise environments. **AssistantBench** (Yoran et al., 2024) addresses open-ended, time-consuming information tasks sourced from real user needs. It includes 214 tasks requiring agents to browse more than 258 websites. Agents must navigate multiple web pages and reason to complete complex objectives—such as monitoring real-estate markets. Performance is evaluated using metrics like Accuracy, Precision, and Exact Match. **WorkArena** (Drouin et al., 2024) focuses on enterprise workflows within the ServiceNow platform, featuring 29 tasks such as timesheet updates and report generation. These are carried out via a real remote interface through the BrowserGym API, with observations including structured DOM data and screenshots; task success is measured by full task completion. **WorkArena++** builds on WorkArena by introducing 682 procedurally generated tasks that demand more advanced capabilities such as planning, logical and arithmetic reasoning, retrieval, and contextual understanding. It maintains the same environment and action space as WorkArena while significantly increasing task complexity.

Summary. These benchmarks for multimodal web agents progressively advance in realism, scale, and difficulty. Some (e.g., WebArena, Visual WebArena, Multimodal Mind2Web) use simulated environments for controlled testing, while others (WebVoyager, AssistantBench, Mind2Web-Live) leverage real or live websites to assess generalization. Overall, they prioritize realistic task execution, multimodal reasoning, and robust evaluation. As models improve, these benchmarks will be crucial for tracking progress toward general-purpose, reliable multimodal agents.

4 Multimodal Web Agents

Unlike conventional text-based web agents, multimodal web agents are capable of accepting visual information from web pages and possess the potential of extending to more general scenarios. Early stages of multimodal web agents involve proper application of existing general agent frameworks like ReAct (Yao et al., 2022) to web planning tasks, together with closed-source multimodal LLMs backbones such as GPT-4V (Achiam et al., 2023), Claude (Anthropic., 2024a), and Gemini (Team et al., 2023). Later, the tendency shifted towards learning-based multimodal web agents, where the trajectory dataset quality, training algorithms, and planning strategies started to play a more important role.

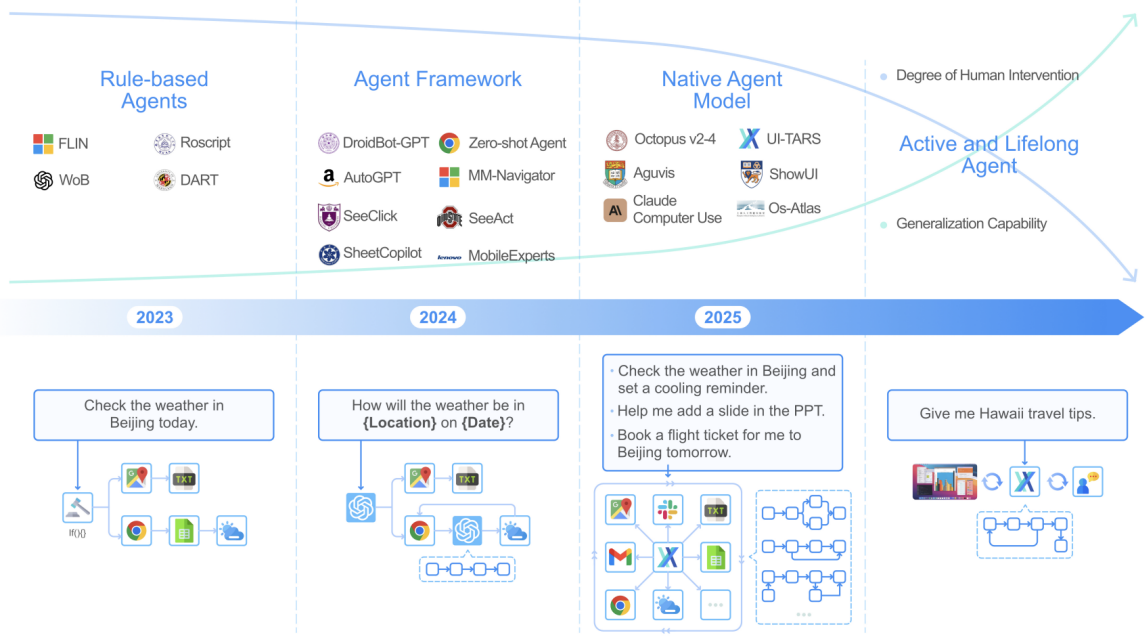

Figure 3: Evolution of multimodal web agents (Qin et al., 2025).

4.1 Prompt-based Methods

Prompt-based multimodal web agents majorly elicit the planning ability from closed-source large language models (LLMs), guiding them by carefully crafted prompts that describe the web environment and task at hand. A typical pipeline provides the LLM with context such as webpage text or DOM structure, screenshots, and interaction history, along with a user instruction. The LLM then outputs the next action (e.g., clicking a button or typing text), which the agent executes in the browser. This cycle repeats iteratively until the task is complete or a step limit is reached. A series of prompt-based multimodal web agents have been developed on top of this paradigm.

WebVoyager (He et al., 2024a) is one of the earliest works that explores this direction, which extends the capabilities of autonomous web navigation by harnessing the power of multimodal inputs, enabling seamless task completion across real-world web environments. By employing GPT-4V (Achiam et al., 2023) to interpret both visual inputs, such as screenshots, and textual data like HTML elements, WebVoyager can effectively interact with dynamic and highly interactive web pages. This includes managing content that is rendered in real time and navigating complex user interfaces. Its multimodal framework allows it to outperform traditional text-only systems, setting a new benchmark in the field of web-based task automation.

SeeAct (Zheng et al., 2024) points out that *grounding* strategies (Chandu et al., 2021; Gu et al., 2022) play a crucial role in enhancing web agents' performance by explicitly specifying interactable web elements with bounding boxes and index labels. This allows a more compact representation of oracle action trajectories

from human annotators, which is much easier for the backbone GPT-4V model to understand and generalize. The importance of the grounding techniques for closed-sourced LLMs has also been observed in Cheng et al. (2024), where a corresponding benchmark and datasets are curated for further improving the open-source models' ability.

Search-Agent (Koh et al., 2024b) brings a fresh perspective to multi-step reasoning in interactive web settings by employing a best-first search algorithm. This method allows the agent to evaluate various action trajectories, leading to more informed decisions and significantly improved performance—boosting success rates by as much as 39% on benchmarks such as WebArena (Zhou et al., 2023c). Its seamless integration with current multimodal LLMs highlights the power of search-driven strategies in tackling intricate web-based tasks.

Hybrid Agent (Song et al., 2024) is exemplified by its additional access to API calls, which intelligently alternates between direct API interactions and traditional web browsing, depending on the nature of the task at hand. Leveraging APIs for structured data retrieval allows the agent to bypass cumbersome interface elements, resulting in faster and more accurate task execution. This fusion of programmatic precision with flexible, browser-like behavior illustrates how combining different modalities can significantly enhance the effectiveness of AI-driven web agents.

WMA (Chae et al., 2024) advances predictive modeling by introducing a learned world model that simulates user interface interactions and anticipates their outcomes. Centered around transition-based observations, this approach enables agents to forecast the impact of their actions before execution, enhancing overall task efficiency. Its strength lies in handling complex, long-horizon tasks where precision is critical. By evaluating possible outcomes beforehand, WMA reduces unnecessary steps and demonstrates strong results on benchmarks such as WebArena (Zhou et al., 2023c) and Mind2Web (Zheng et al., 2024).

Several other works in this domain, including **NaviQAte** (Shahbandeh et al., 2024), **AdaptAgent** (Verma et al., 2024), **ECLAIR** (Wornow et al., 2024), and **PAE** (Zhou et al., 2024), employ comparable techniques while introducing variations in task representations, or demonstration examples and strategies. Despite the absence of standardized benchmarking across these studies—which makes direct, one-to-one comparisons challenging—most leverage the Mind2Web benchmark (Deng et al., 2023b) or its multimodal extension (Zheng et al., 2024). This allows a high-level understanding of progress and emerging trends within the field, as shown in Table 3.

Summary. Prompt-based web agents treat task execution as a policy encoded within a language model, activated through the right prompts. They rely on key factors of planning strategies, high-quality few-shot examples, and compact representation of the state/action space, rather than model re-training. This enables fast development iteration and wide task coverage, though often at the expense of performance in domain-specific tasks.

4.2 Learning-based Methods

Similar to prompt-based methods, state/action representations and planning strategies also play a crucial role in learning-based approaches. Moreover, learning-based methods further improve the performance of multimodal web agents by introducing enhanced datasets and training algorithms.

Better State Representation Like SeeAct, **SeeClick** (Cheng et al., 2024) also identifies the importance of *grounding*, where a pre-training dataset is constructed to enpower Qwen-VL (Bai et al., 2023) with the ability of pinpointing interactable web elements. Similarly, **AutoWebGLM** (Lai et al., 2024) addresses complex web navigation challenges by streamlining HTML content to highlight only essential webpage elements, leading to more accurate task execution. Through the use of reinforcement learning and rejection sampling for fine-tuning, it demonstrates strong performance across intricate tasks on websites in both English and Chinese. With a bilingual dataset and modular design that separates perception and action, AutoWebGLM proves highly effective for cross-domain web environments, showcasing the value of efficiency and adaptability in handling diverse online tasks.

Table 3: Comparison of Multimodal Web Agents, where * means only a subset of tasks are evaluated.

Method	Mind2Web (Cross-Task) Step Success Rate (%)	WebVoyager Task Success Rate (%)	MiniWoB++ Accuracy (%)	Date
LLaVA-1.5 (Liu et al., 2024a)	8.1	-	-	23.10
Qwen-VL (Bai et al., 2023)	13.3	-	-	23.08
GPT-3.5	16.8	-	39.6	22.03
OpenWebVoyager (8B) (He et al., 2024b)	20.5*	27.4	-	24.10
SeeClick (Cheng et al., 2024)	25.5	-	-	24.01
GPT-4 (all tool mode) (Achiam et al., 2023)	32.3	30.8	53.0	23.03
SeeAct (Zheng et al., 2024) + GPT-4V	40.2	-	-	24.01
Magma-8B (Yang et al., 2025)	43.4	-	-	25.02
Explorer-7B (Pahuja et al., 2025)	53.2	-	53.3	25.02
Claude* (Anthropic., 2024b)	53.5	-	-	24.10
AGUVIS-7B (Xu et al., 2024b)	60.4	-	-	24.12
AdaptAgent (Verma et al., 2024) + SeeAct (Zheng et al., 2024) + GPT-4o	61.9	-	-	24.11
AGUVIS-72B (Xu et al., 2024b)	64.0	-	-	24.12
WMA (Chae et al., 2024) + AWM (Wang et al., 2024) + GPT-3.5-Turbo	67.0	-	-	24.10
UI-TARS-7B (Qin et al., 2025)	67.1	-	-	25.01
UI-TARS-72B (Qin et al., 2025)	68.6	-	-	25.01

A more sophisticated state representation involves directly learning a perception model specifically designed for web pages, which is the approach adopted by Magma. **Magma** (Yang et al., 2025) serves as a unified foundation model designed for multimodal AI agents, bridging language, vision, and action comprehension to support both user interface navigation and robotic manipulation. Departing from earlier models that focus exclusively on either web automation or robotics, Magma is trained on a diverse dataset of approximately 39 million samples, including UI screenshots, web navigation sequences, robot motion data, and instructional videos. It utilizes techniques like SoM and Trace-of-Mark to strengthen its ability to identify and predict actionable elements, thereby improving the precision of task execution across varied domains.

Better Trajectory Dataset **Explorer** (Pahuja et al., 2025) introduces a scalable framework for synthesizing the largest and most diverse trajectory-level dataset for training multimodal web agents. The dataset contains over 94K successful web trajectories spanning 49K unique URLs, 720K screenshots, and 33 million web elements, enabling broad coverage of real-world tasks. In particular, a scalable exploration-driven method is proposed to generate trajectories at 28 cents per successful instance, making large-scale data collection affordable.

Better Training Strategies **WebDreamer** (Gu et al., 2024) presents a novel application of large language models (LLMs) for planning-driven web navigation. Rather than acting reactively, it leverages LLMs to simulate and assess the outcomes of multiple action sequences in advance—much like a "dreamer" envisioning possible futures. This forward-looking strategy allows WebDreamer to identify and pursue the course of action with the highest anticipated reward. By forecasting the value of different action paths before execution, the system effectively navigates the complexities of dynamic web environments while addressing key challenges such as safety and decision robustness. Its planning-based method has outperformed reactive agents in evaluations on benchmarks like VisualWebArena (Koh et al., 2024a) and Mind2Web-live (Pan et al., 2024a).

AGUVIS (Xu et al., 2024b) adopts a vision-only strategy to automate interactions with graphical user interfaces, bypassing the constraints associated with text-dependent methods such as HTML parsing or accessibility tree analysis. Its platform-independent architecture enables seamless operation across web, desktop, and mobile environments while also optimizing inference efficiency. The system undergoes a two-phase training regimen—initially focusing on grounding visual elements in GUIs, followed by the integration of reasoning and planning into a cohesive model. This comprehensive approach results in cutting-edge performance in both real-time and offline tasks, enhancing the speed and accuracy of decision-making processes.

Agent Q (Putta et al., 2024) enhances performance in complex web-based tasks like product searches and booking systems by integrating Monte Carlo Tree Search (MCTS) with self-evaluation techniques and Direct Preference Optimization (DPO) (Rafailov et al., 2023). It builds on the capabilities of the LLaMA-3 70B model (Grattafiori et al., 2024), refining it to interpret HTML DOM structures and produce coherent action plans, contextual reasoning steps, and tailored commands for various environments. This approach highlights the effectiveness of combining strategic search, reflective feedback, and iterative model tuning to advance the development of intelligent, autonomous agents.

OpenWebVoyager (He et al., 2024b) introduces a framework designed for continual learning and broad generalization across diverse web environments. It integrates imitation learning with a cyclical process of exploration, feedback, and optimization to refine performance over time. Utilizing existing multimodal models like Idefics2-8B (Laurençon et al., 2024), the agent autonomously navigates real-world websites. Rather than depending on synthetic setups, it is trained on rich, varied datasets and fine-tuned using trajectory feedback validated by GPT-4. This methodology marks a significant step forward for GUI-based agents, showcasing robust adaptability across a wide range of web tasks and domains.

UI-TARS (Qin et al., 2025) represents the current state of the art in open-source multimodal web agents, which even outperforms OpenAI's **Operator** (OpenAI, 2025b) in several aspects beyond web applications, such as computer-use tasks (Xie et al., 2024). It is a 72B-parameter vision-language model trained from scratch—without relying on external LLMs at inference—to control native GUIs using only pixels. UI-TARS outperforms models like GPT-4 and Claude on tough benchmarks. These gains majorly come from several innovations:

- A massive dataset of GUI interactions teaching precise perception and grounding
- A unified action space enabling consistent control across applications
- Embedded reasoning strategies like task decomposition and reflection
- Iterative training using the model's own failures to improve robustness

UI-TARS shows that with sufficient data and training, a learned agent can match or exceed prompt-based models on complex tasks. The trade-off is the resource-intensive training pipeline, which only a few organizations can support.

Summary. Learning-based approaches invest heavily in task-specific optimization to build agents with better datasets and training strategies. While early agents were limited to toy environments, recent models trained on rich datasets have closed the performance gap, rendering the obtained agents more controllable and customizable. Several key trends can be observed in this field, including: 1) a shift from prompt-based to learning-based approaches, 2) the integration of more advanced training algorithms, such as self-reflection in reinforcement learning, to enhance planning capabilities, and 3) scaling through larger, high-quality datasets and more powerful backbone models.

5 Conclusion

In conclusion, the growing capabilities of multimodal large language models have opened new avenues for building autonomous agents that can effectively navigate and interact with complex web environments. By synthesizing insights from a wide range of studies, this survey highlights the current state of multimodal

web agents, from foundational MLLMs to advanced algorithmic strategies and evaluation methodologies. Through the categorization of prompting- and learning-based approaches, as well as an exploration of key techniques such as grounding and reinforcement learning, we offer a structured understanding of the field. This work serves as a valuable resource for guiding future research and development in multimodal web agents, laying the groundwork for more adaptive, intelligent, and generalizable human-computer interaction systems.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

AI Anthropic. Introducing the next generation of claude., 2024a.

AI Anthropic. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku., 2024b.

J Bai, S Bai, S Yang, S Wang, S Tan, P Wang, J Lin, C Zhou, and J Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. arxiv 2023. *arXiv preprint arXiv:2308.12966*, 2023.

Léo Boisvert, Megh Thakkar, Maxime Gasse, Massimo Caccia, Thibault de Chezelles, Quentin Cappart, Nicolas Chapados, Alexandre Lacoste, and Alexandre Drouin. Workarena++: Towards compositional planning and reasoning-based common knowledge work tasks. *Advances in Neural Information Processing Systems*, 37:5996–6051, 2024.

Hyungjoo Chae, Namyoung Kim, Kai Tzu-iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. Web agents with world models: Learning and leveraging environment dynamics in web navigation. *arXiv preprint arXiv:2410.13232*, 2024.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. Grounding'grounding'in nlp. *arXiv preprint arXiv:2106.02192*, 2021.

Wentong Chen, Junbo Cui, Jinyi Hu, Yujia Qin, Junjie Fang, Yue Zhao, Chongyi Wang, Jun Liu, Guirong Chen, Yupeng Huo, et al. Guicourse: From general vision language models to versatile gui agents. *arXiv preprint arXiv:2406.11317*, 2024.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*, 2024.

Yuri Chervonyi, Trieu H Trinh, Miroslav Olšák, Xiaomeng Yang, Hoang Nguyen, Marcelo Menegali, Junehyuk Jung, Vikas Verma, Quoc V Le, and Thang Luong. Gold-medalist performance in solving olympiad geometry with alphageometry2. *arXiv preprint arXiv:2502.03544*, 2025.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023a.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023b.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, et al. Workarena: How capable are web agents at solving common knowledge work tasks? *arXiv preprint arXiv:2403.07718*, 2024.

Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. Navigating the digital world as humans do: Universal visual grounding for gui agents. *arXiv preprint arXiv:2410.05243*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yu Gu, Xiang Deng, and Yu Su. Don't generate, discriminate: A proposal for grounding language models to real-world environments. *arXiv preprint arXiv:2212.09736*, 2022.

Yu Gu, Boyuan Zheng, Boyu Gou, Kai Zhang, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, Huan Sun, and Yu Su. Is your llm secretly a world model of the internet? model-based planning for web agents. *arXiv preprint arXiv:2411.06559*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919*, 2024a.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Hongming Zhang, Tianqing Fang, Zhenzhong Lan, and Dong Yu. Openwebvoyager: Building multimodal web agents via iterative real-world exploration, feedback and optimization. *arXiv preprint arXiv:2410.19609*, 2024b.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. Ai agents that matter. *arXiv preprint arXiv:2407.01502*, 2024.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*, 2024a.

Jing Yu Koh, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. Tree search for language model agents. *arXiv preprint arXiv:2407.01476*, 2024b.

Hanyu Lai, Xiao Liu, Iat Long Iong, Shuntian Yao, Yuxuan Chen, Pengbo Shen, Hao Yu, Hanchen Zhang, Xiaohan Zhang, Yuxiao Dong, et al. Autowebglm: A large language model-based web navigating agent. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5295–5306, 2024.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pp. 405–409, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Xing Han Lù, Zdeněk Kasner, and Siva Reddy. Weblinx: Real-world website navigation with multi-turn dialogue. *arXiv preprint arXiv:2402.05930*, 2024.

OpenAI. Openai o3-mini, 2025a. URL `https://openai.com/index/openai-o3-mini/`.

OpenAI. Introducing operator, 2025b. URL `https://openai.com/index/introducing-operator/`.

Vardaan Pahuja, Yadong Lu, Corby Rosset, Boyu Gou, Arindam Mitra, Spencer Whitehead, Yu Su, and Ahmed Awadallah. Explorer: Scaling exploration-driven web trajectory synthesis for multimodal web agents. *arXiv preprint arXiv:2502.11357*, 2025.

Yichen Pan, Dehan Kong, Sida Zhou, Cheng Cui, Yifei Leng, Bing Jiang, Hangyu Liu, Yanyi Shang, Shuyan Zhou, Tongshuang Wu, et al. Webcanvas: Benchmarking web agents in online environments. *arXiv preprint arXiv:2406.12373*, 2024a.

Yichen Pan, Dehan Kong, Sida Zhou, Cheng Cui, Yifei Leng, Bing Jiang, Hangyu Liu, Yanyi Shang, Shuyan Zhou, Tongshuang Wu, et al. Webcanvas: Benchmarking web agents in online environments. *arXiv preprint arXiv:2406.12373*, 2024b.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023.

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent q: Advanced reasoning and learning for autonomous ai agents. *arXiv preprint arXiv:2408.07199*, 2024.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Fernando Andrés Quintana Sally Paganin, Garritt L. Page. Informed random partition models with temporal dependence. *arXiv preprint arXiv:2311.14502*, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.

Mobina Shahbandeh, Parsa Alian, Noor Nashid, and Ali Mesbah. Naviqate: Functionality-guided web application navigation. *arXiv preprint arXiv:2409.10741*, 2024.

Yueqi Song, Frank Xu, Shuyan Zhou, and Graham Neubig. Beyond browsing: Api-based web agents. *arXiv preprint arXiv:2410.16464*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Ada Defne Tur, Nicholas Meade, Xing Han Lù, Alejandra Zambrano, Arkil Patel, Esin Durmus, Spandana Gella, Karolina Stańczak, and Siva Reddy. Safearena: Evaluating the safety of autonomous web agents. *arXiv preprint arXiv:2503.04957*, 2025.

Gaurav Verma, Rachneet Kaur, Nishan Srishankar, Zhen Zeng, Tucker Balch, and Manuela Veloso. Adaptagent: Adapting multimodal web agents with few-shot learning from human demonstrations. *arXiv preprint arXiv:2411.13451*, 2024.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024.

Michael Wornow, Avanika Narayan, Krista Opsahl-Ong, Quinn McIntyre, Nigam H Shah, and Christopher Re. Automating the enterprise with foundation models. *arXiv preprint arXiv:2405.03710*, 2024.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *arXiv preprint arXiv:2404.07972*, 2024.

Yiheng Xu, Dunjie Lu, Zhennan Shen, Junli Wang, Zekun Wang, Yuchen Mao, Caiming Xiong, and Tao Yu. Agenttrek: Agent trajectory synthesis via guiding replay with web tutorials. *arXiv preprint arXiv:2412.09605*, 2024a.

Yiheng Xu, Zekun Wang, Junli Wang, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu, and Caiming Xiong. Aguvis: Unified pure vision agents for autonomous gui interaction. *arXiv preprint arXiv:2412.04454*, 2024b.

Tianci Xue, Weijian Qi, Tianneng Shi, Chan Hee Song, Boyu Gou, Dawn Song, Huan Sun, and Yu Su. An illusion of progress? assessing the current state of web agents. *arXiv preprint arXiv:2504.01382*, 2025.

Hui Yang, Sifu Yue, and Yunzhong He. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*, 2023.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. *arXiv preprint arXiv:2502.13130*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. Assistantbench: Can web agents solve realistic and time-consuming tasks? *arXiv preprint arXiv:2407.15711*, 2024.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 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.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023a.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023b.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023c.

Yifei Zhou, Qianlan Yang, Kaixiang Lin, Min Bai, Xiong Zhou, Yu-Xiong Wang, Sergey Levine, and Erran Li. Proposer-agent-evaluator (pae): Autonomous skill discovery for foundation model internet agents. *arXiv preprint arXiv:2412.13194*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Yuxuan Zhu, Antony Kellermann, Dylan Bowman, Philip Li, Akul Gupta, Adarsh Danda, Richard Fang, Conner Jensen, Eric Ihli, Jason Benn, et al. Cve-bench: A benchmark for ai agents' ability to exploit real-world web application vulnerabilities. *arXiv preprint arXiv:2503.17332*, 2025.