MMInA: Benchmarking Multihop Multimodal Internet Agents

Ziniu Zhang[*] **Shulin Tian**[*] **Liangyu Chen**[*,†] **Ziwei Liu** ✉

S-Lab, Nanyang Technological University

{michaelzhangziniu}@gmail.com

{stian006, lchen025, ziwei.liu}@ntu.edu.sg

Abstract

Autonomous embodied agents live on an Internet of multimedia websites. Can they hop around multimodal websites to complete complex user tasks? Existing benchmarks fail to assess them in a realistic, evolving environment for their embodiment across websites. To answer this question, we present **MMInA**, a multihop and multimodal benchmark to evaluate the embodied agents for compositional Internet tasks, with several appealing properties: **1)** *Evolving real-world multimodal websites.* Our benchmark uniquely operates on evolving real-world websites, ensuring a high degree of realism and applicability to natural user tasks. Our data includes 1,050 human-written tasks covering various domains such as shopping and travel, with each task requiring the agent to autonomously extract multimodal information from web pages as observations; **2)** *Multihop web browsing.* Our dataset features naturally compositional tasks that require information from or actions on multiple websites to solve, to assess long-range reasoning capabilities on web tasks; **3)** *Holistic evaluation.* We propose a novel protocol for evaluating an agent's progress in completing multihop tasks. We experiment with both standalone (multimodal) language models and heuristic-based web agents. Extensive experiments demonstrate that while long-chain multihop web tasks are easy for humans, they remain challenging for state-of-the-art web agents. We identify that agents are more likely to fail on the early hops when solving tasks of more hops, which results in lower task success rates. To address this issue, we propose a simple memory augmentation approach replaying past action trajectories to reflect. Our method significantly improved both the single-hop and multihop web browsing abilities of agents. See our code and data at https://mmina.cliangyu.com.

1 Introduction

Building embodied agents capable of autonomous behaviors navigating in various environments has been a longstanding and intricate challenge in the realm of artificial intelligence research [27, 54, 9, 41]. One common scenario that necessitates automation involves the interaction with digital interfaces [33, 43], with a particular emphasis on the automation of actions performed on rich Internet websites [40, 50, 14]. Real-world web tasks are usually compositional, involving multihop actions across multiple websites. Achieving this goal requires endowing agents with both *long-range planning* and *multimodal reasoning* capabilities. This includes understanding high-level instructions derived from user inputs, planning multihop actions within the web browser environment, encompassing HTML content and *visual cues* from web resources, and making practical predictions based on the observations. While web agents show promise in addressing single-hop tasks that require only one website according to existing benchmarks [21, 24, 25, 53, 17], we observe that they encounter

---

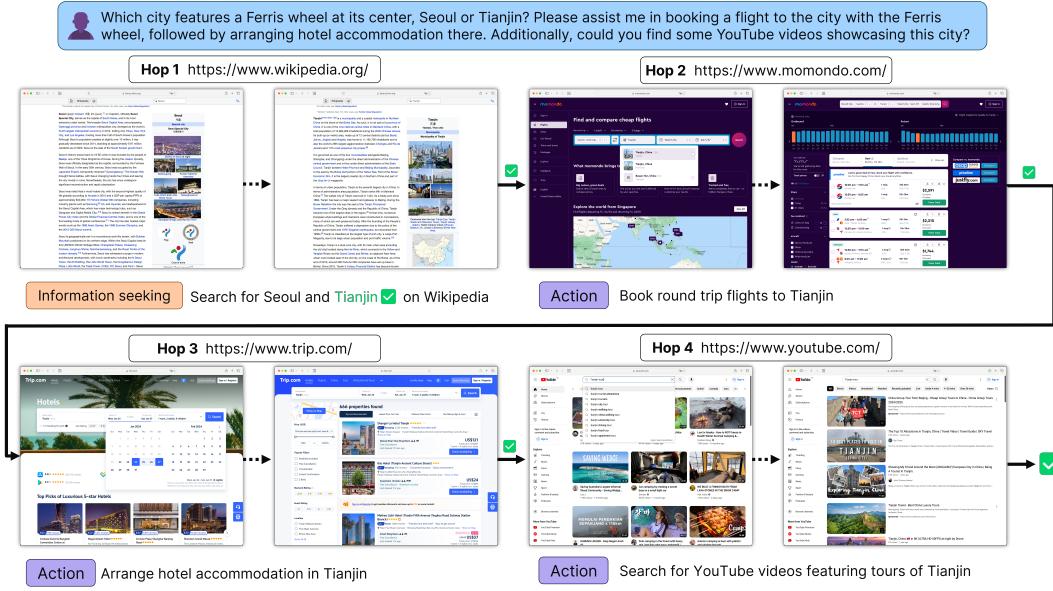[*]Equal Contribution.   [†]Project Lead.   ✉Corresponding Author.

Figure 1: **An example task from MMInA**. To evaluate an Internet agent's ability to carry out complex tasks, we make it navigate through a variety of websites to gather information and execute actions. In our proposed holistic evaluation protocol, each phase of the compositional task (defined as a *hop*) and the overall task are assessed for performance. Our benchmark includes 1,050 varied human-written multimodal tasks that require an average of 2.85 hops between websites and 12.9 actions to complete. The longest compositional task takes 10 hops.

challenge to solve multihop web tasks, which are prevalent in most realistic scenarios that a user requires information from or acts on multiple websites for a high-level task. (Tab. 3). This gap motivates us to establish a multihop web browsing benchmark to assess the usefulness of Internet agents in natural multihop tasks.

Another gap in web agent research is multimodality. Existing benchmarks pose autonomous agent tasks that rely solely on textual inputs and textual information [53, 8, 50]. However, in real-world scenarios, visual information often plays an indispensable role and cannot be disregarded. For instance, consider the task of "Help me purchase a blue cotton shirt", where the color attribute, derived from visual information, becomes crucial in fulfilling the user's request. What's more, there is a notable lack of current web-based agent benchmarks for emphasis on assessing the capabilities of comprehending and interacting with both textual and visual inputs, where a significant number of them primarily concentrate on tasks that involve text-based interactions.

To address the two issues above, we present MMInA, a novel benchmark designed to advance the field of multihop and multimodal Internet task solving. Our benchmark uniquely operates on evolving real-world websites (Fig. 1), ensuring a high degree of realism and applicability to natural scenarios, (Tab. 2). Central to MMInA is its focus on realistic tasks that users commonly engage in, such as navigating e-commerce platforms, extracting and synthesizing information from content-rich sites like Wikipedia, and performing comparative analysis across multiple web sources. Our carefully designed human-written 1,050 tasks not only challenge the agents to comprehend multimodal inputs at multiple website hops but also to execute sophisticated, multi-step reasoning – a significant leap from the simpler tasks commonly seen in existing benchmarks.

Our extensive experiments with state-of-the-art agents reveal that while there is significant progress in handling simple textual tasks, the integrated and sequential nature of tasks in MMInA poses a substantial challenge. For instance, the best-performing standalone model, GPT-4V [29], achieves an overall success rate of 21.8% across tasks. This is a notable improvement over textual agent baselines but still lags behind the human performance (96.3%). We identify that agents are more likely to fail on the early hops when solving tasks of more hops (Tab. 4), which results in lower task success rates. These results underscore the complexities of real-world web navigation and decision-making, emphasizing the need for further advancements in multimodal and multihop reasoning. To bridge

Benchmark	Multi-modal	Max / Avg. Hops	Website Type	Dynamic Interaction	# Websites
MiniWoB++ [22]	✓	1 / 1.00	Static simplified websites	✓ (Open-ended)	100
WebShop [49]	✓	1 / 1.00	Static simplified websites	✓ (Open-ended)	1
Mind2Web [8]	✗	1 / 1.00	Static real-world websites	✗ (MC)	131
RUSS [47]	✗	2 / 1.10	Static real-world websites	✗ (MC)	22
WebArena [53]	✗	2 / 1.06	Static real-world websites	✓ (Open-ended)	6
VWA [17]	✓	2 / 1.05	Static real-world websites	✓ (Open-ended)	3
MMInA	✓	10 / 2.85	Evolving real-world websites	✓ (Open-ended)	14

Table 1: **Comparison between the MMInA benchmark and related benchmarks.** MMInA employs a flexible environment that supports agents to generate open-ended actions. We selected 14 evolving real-world websites to benchmark multihop multimodal Internet agents, which can be easily expanded for future deployments.

this gap, we propose a simple memory augmentation approach replaying past action trajectories to reflect. Our method significantly improved both the single-hop and multihop web browsing abilities of agents. Memory augmentation is a flexible model-agnostic technique, which can be lent to more large multimodal models as agents in the future.

In summary, our contributions are as follows:

- We introduce MMInA, an Internet agent benchmark encompassing 1,050 multihop multi-modal tasks spanning 14 diverse websites with evolving and realistic features. The inclusion of multihop tasks enables the assessment of more intricate and human-like actions, closely mirroring real-world problem-solving trajectories. We benchmark the current representative large language models (LLMs) and large multimodal models (LMMs) as agents on the benchmark.

- We propose a holistic evaluation method for multihop tasks. Based on experimental observations, the task success rates of completing multihop tasks for existing agents are exceedingly low. Our new protocol evaluates the completion results by hops, providing a fine-grained method to assess an agent's ability to complete multihop tasks.

- We propose a memory-augmented method that can effectively improve agent performance by reflecting on the action histories of past tasks. The method is simple and flexible, yet effective; significantly improving LMMs as Internet agents.

2   Related Works

Agent Benchmarks and Environments   Most existing works evaluate autonomous agents on curated textual I/O interfaces, leaving a gap in assessing their performances on real-world automation tasks. APIBench [21], introduced by Gorilla[32], is a tool-augmented LLM benchmark to assess the tool utilization abilities of agents for code generation tasks. AgentBench [24] stepped forward to provide a more general toolkit with lots of closed-box environments to assess agents' performances in answering user queries. BOLAA [25] is another benchmark that coordinates multiple autonomous agents to make collective decisions. OpenAGI [12] and GAIA [28] are multimodal benchmarks crafted for generalist agents that define multi-step tasks across multiple modalities. However, none of the above benchmarks explored the usage of LLMs or LMMs in web browsing environments or posed an effective evaluation metric specifically tailored for web agent tasks.

Web Agents   WoB [40] and its tool extension MiniWoB++ [22] established a platform of website widgets where agents can complete online tasks through basic keyboard and mouse operations. Webshop [49] was a simulated e-commerce environment featuring 1.18 million real-world products, complemented by 12,087 crowdsourced textual instructions. It converted the action prediction task into a choice-based imitation learning process, which facilitated the accuracy and ability for task execution. However, this approach failed to evaluate open-ended agent actions in the real world. It was also limited by its monotonous design of a one-website environment, which resulted in only a single category of web browsing tasks. Mind2Web [8] tried to construct a generalist web agent, which created a dataset for crafting and benchmarking the web agents by the ability of instruction following. It proposed a two-stage training to convert the action prediction problem into MCQs. SeeAct [52] was a following work that enabled multimodal information for visually understanding rendered webpages

a **Source websites of the 2,991 hops.**

b **Intent types of hops.**
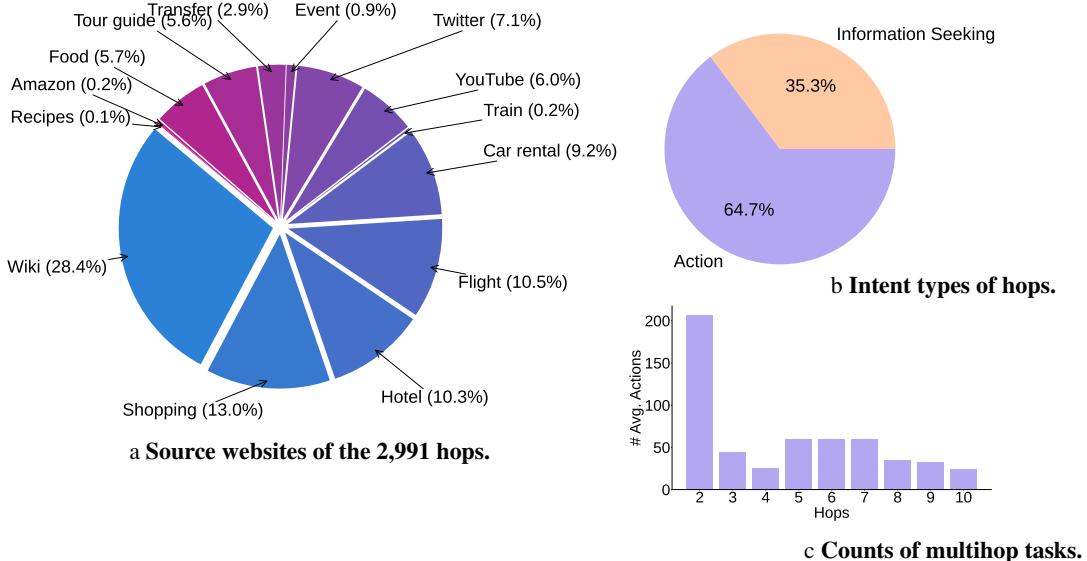

c **Counts of multihop tasks.**

Figure 2: **Statistics of the MMInA benchmark.** A web browsing task is composed of one or multiple hops between websites. MMInA covers diverse intents and domains of hops to resemble naturally compositional user tasks.

and generating more accurate action plans. WebVoyager [13] was capable of capturing screenshots of web pages and then using JavaScript tools to automatically identify interactive elements based on the types of webpage elements. WebArena [53] deployed a standalone set of multicategory websites in an interactive environment. VisualWebArena [17] was a subsequent project that built upon the foundation of WebArena, introducing the reliance on visual cues into the benchmark's design. The tasks of existing benchmarks are oversimplified whose completions require a single website, which is highly diverged from the natural web browsing tasks and should originally be designed to multihop over a long-horizon setting.

3 MMInA Benchmark

3.1 Environment

Following [53], we formulate web browsing as a partially observable Markov decision process $\langle S, A, P, R \rangle$. The state space S is the whole Internet content, the status of the browser environment and agent, exceeding representable expressions in practice. Therefore, we pass the partial observation space Ω of S to the agent. At each time step t, an agent arrives at a certain state of a particular web page. The accessibility tree of the screenshot with linked images, with the action/state histories, forms a partial observation $o_t \in \Omega$ for the agent. Then the web agent takes an action a_t sampled from the action space A, being either an executable operation on the web page or a textual output as the answer to a question (Sec. 3.4). The state transition probability matrix $P : S \times A \to S'$ is implicitly encoded as the world knowledge of the Internet environment, which can be inferred or learned by a web agent. Our reward function R is expressed with language output, PASS or FAIL, as a result of each hop, not step. Naturally, we define hop as a subtask that is completed on a specific website. For example, in the task of Fig. 1, the agent receives a PASS if it finds the correct destination in the first hop, another PASS for arriving at the desired flight search page at the second hop, and so on. The task is successful only if the agent receives PASS at each hop.

Action Space To achieve a more faithful replication of human actions on the web, we follow [17] to condense the potential agent-executed actions into a set of 12 summarized actions. Leveraging the playwright library, we simulate web pages on an X graphics server, employing a diverse array of actions to interact with the pages. These actions span a broad spectrum of behaviors mirroring human interactions with web pages, encompassing actions such as clicking on links, scrolling up and down using the scroll wheel, typing with the keyboard, and more. From the hop counts shown in Fig. 3, we observe that as the number of hops increases, the number of actions required by the
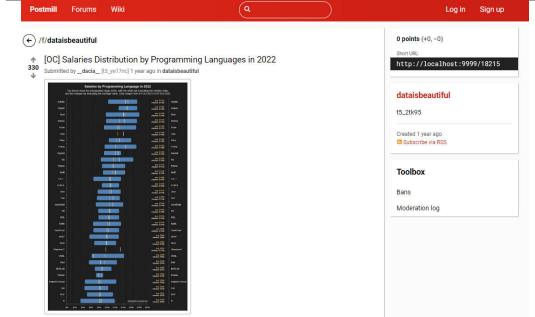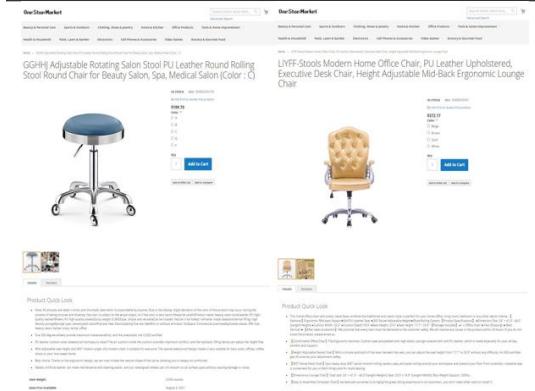
4

Websites	Task
VWA	
	When did programming language that has the largest variance in salary first appear? Answer using the information from the Wikipedia site in the second tab.
MMInA	
	Do both LIYFF-Stools Modern Home Office Chair and GGHHJ Adjustable Rotating Salon Stool PU Leather Round Rolling Stool Round Chair have armrest?

Table 2: **Comparison between VWA [17] and MMInA.** MMInA requires multimodal inputs at multiple steps to accomplish the task, which makes a more challenging multimodal benchmark.

agent also increases. However, it's worth noting that the average number of actions required for 5-hop data is lower than that for 4-hop data. In our dataset, 4-hop content involves comparative operations, such as "Which one has a Ferris wheel in the center of the city, Tianjin or Chengdu?" Since our definition of multi-hop tasks involves navigation across different web pages, we do not categorize these comparative questions as 5-hop tasks. On average, an MMInA task takes 12.9 actions to complete.

Observation Space   The observation space Ω usually embeds partial observations of the Internet to simulate real web browsing experiences. Observations include the task descriptions and the website contents. To represent web content in a reasonable length, we primarily use accessibility trees that provide a structured and unified layout of the web page content. Each node of the accessibility tree has an element ID, the type of the element, and the textual content of the element. If the element is an image, the environment downloads the image and paints the element ID on the image as a reference for the multimodal agent.

3.2   Multimodal Web Content

Our work at MMInA focuses on multimodality-reliant tasks, which require both images and textual data to complete. For example, the task "Which one is more furry, Hi&Yeah Comfy Faux Fur Cute Desk Chair or Armen Living Diamond Office Chair?" requires the agent to locate and compare specified items on referenced web pages, analyzing the images and textual web page content to provide an answer.

MMInA's approach contrasts with VWA, as all tasks in our framework necessitate the processing of both visual and textual information in multiple turns (Tab. 2). With richer multimodal content, MMInA becomes a more challenging yet realistic benchmark.

To tackle these multimodal tasks effectively, MMInA integrates an automated process for extracting the accessibility tree from web pages, encompassing both images and text. This enables the agent to download referenced images and utilize them alongside the accessibility tree as inputs for a multimodal agent. Such an approach underscores the intricate interplay between visual and textual data in solving real-world tasks within a multimodal framework.

3.3 Multihop Cross-website Browsing

MMInA dataset features multihop tasks across 14 distinct websites, encompassing a diverse range of domains such as shopping, ticket booking, travel guide exploration, and local food discovery (refer to Fig. 2). We define a "multihop task" as a task that necessitates actions across multiple websites. It requires the agent to automatically go to the next website after finishing a hop. This setting emphasizes the complexity and real-world relevance of multihop web browsing, as they more accurately simulate the sequence of actions a human user typically performs when facing a compositional high-level task. In practice, we provide the links to the available websites (see details in Appendix A) in the task descriptions.

Figure 3: **Average actions in multihop tasks.**

3.4 Evaluation

Single-hop Evaluation   Following [53], we have utilized two distinct evaluation methodologies for single-hop tasks, based on the semantics and effectiveness of the predicted actions. Each method is tailored to suit different task characteristics within the MMInA dataset, providing either strict or loose bounds for the evaluation process.

The first method, `"must_include"`, is grounded in a keyword-based approach. Here, we establish a set of compositional keywords for each task. The agent's response is deemed successful (`PASS`) only if it incorporates all these specified keywords. Conversely, the omission of even a single keyword results in the task being classified as a failure (`FAIL`). This method ensures a stringent, keyword-focused evaluation of the agent's performance.

The second method, designated as `"fuzzy_match"`, leverages the advanced capabilities of large language models, exemplified by GPT-3.5-Turbo. This method involves inputting both the reference answer and the agent's response into the language model. Subsequently, the model is prompted with a query framed as follows: "Given the statement {`pred`}, would it be correct to infer {`reference`}? Yes or No". Here {`pred`} represents for agent's response and {`reference`} means the reference answer. The evaluation outcome is determined based on the model's response: a task is adjudged as (`PASS`) if the model confirms the inference (Yes), and as (`FAIL`) if it does not (No). This approach offers a more nuanced and flexible evaluation framework, effectively accommodating the subtleties and complexities inherent in language interpretation. To illustrate, in a task where color identification is involved, if the reference answer is "gold" and the agent's response is "yellow", the `"fuzzy_match"` method will try to grasp and assess the semantic meaning of the context and provide the conclusive matching result.

Multihop Evaluation   In our experiments with multihop tasks(Tab. 3), we often observe a remarkably low completion rate for the entire task, if not zero. To provide a more holistic evaluation, we propose an evaluation method tailored for N-hop problems:

The evaluation involves maintaining a queue containing the conditions of each hop's completion. In particular, the last element of a queue is always an "END" marker that signifies the whole multihop task is completed, making the queue's length $N + 1$. An agent succeeds at a hop once it finds out the required information (*e.g.*, an answer string) or reaches the desired state (*e.g.*, a specific URL). For
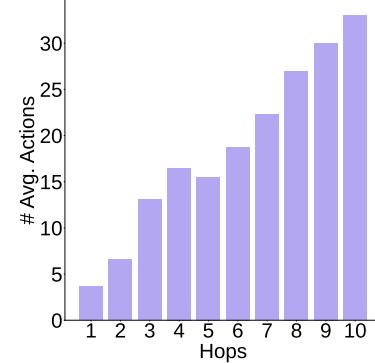
6

Model	Agent	Inputs	Hop Success Rate (↑)				Task Success Rate (↑)			
			1 hop	2-4 hops	5+ hops	overall	1 hop	2-4 hops	5+ hops	overall
Text-only	Fuyu-8B	Acc. Tree	0	0	0	0	0	0	0	0
	CodeLLaMA-7B		1.18	0	0	0.29	1.18	0	0	0.58
	WebShop		20.67	0	0	4.17	20.67	0	0	10.12
	Gemini-Pro		19.09	**34.12**	2.13	11.85	19.09	0.76	0	9.54
	GPT-4		14.37	30.56	5.23	12.26	14.37	9.09	0	9.34
Caption	CodeLLaMA-7B w/ caps	Acc. Tree + Caps	5.71	0	0	1.61	5.71	0	0	2.79
	WebShop w/ caps		29.72	0.00	0.00	5.61	29.72	0	0	14.55
	Gemini-Pro w/ caps		30.12	11.09	0.05	12.38	30.12	1.52	0.38	15.22
	GPT-4 w/ caps		38.58	20.70	3.43	13.50	38.58	3.79	0	19.85
Multimodal	Fuyu-8B	Images + Acc. Tree	27.36	0	0	5.52	27.36	0	0	13.39
	Gemini-Pro-Vision		28.94	16.38	4.03	10.66	28.94	1.51	**1.13**	18.40
	GPT-4V		**42.91**	21.23	3.99	13.89	**42.91**	3.03	0	**21.77**
Multimodal (Memory)	Gemini-Pro-Vision	Images + Acc. Tree	39.17	**23.93**	**4.78**	**14.27**	39.17	**10.61**	1.13	20.13
Human	-	Webpage	99.02	97.91	93.77	98.43	99.02	95.34	88.12	96.25

Table 3: **MMInA Benchmark Results.** We evaluated 4 types of agents on the proposed MMInA benchmark: 1) LLM Agents; 2) LMM Agents; 3) Heuristic-Based Web Agents; 4) Human Baselines. The hop success rate is defined by the percentage (%) of successful visits of the targeted websites; while the task success rate is calculated by the overall percentage (%) of successful tasks from the whole task set.

simplicity, our benchmark enforces that the agent completes tasks in sequence, *i.e.*, the agent is only allowed to proceed to the next hop if the current hop is correctly completed. A task is completed only if all the hops are correctly completed in sequence.

Our single-hop and multihop evaluation methods aim to provide a systematic and insightful approach to assess the performance of agents in tackling multihop tasks, addressing the challenges posed by such tasks' complexity.

4 Experiments

4.1 Baselines

A variety of state-of-the-art LLMs and LMMs, along with the adapted forms of web-oriented models, were employed to evaluate their performance in the MMInA benchmark. The default parameters were used from either open-sourced pretrained models or API-based models for evaluation. For the web-trained agent WebShop, which was built on the specific, static environment, we utilized the GPT-3.5-turbo model to generate formatted queries to replace the phrases that should have originally been queried from the built-in environment. In text-only baselines, only textual inputs are given the LMM Fuyu-8B. We follow the display setting from [53] in a viewport 1280×2048, and the webpage accessibility tree as the text input for the models. For text-based models, we performed the tasks into 2 settings: 1) *text-only*: only the available textual information was given while neglecting the image information; 2) *caption-augmented*: using BLIP-2 [20] model to translate the information of images into text format to combine with the original text information as the input to models. For multimodal models, both the images and text information from the website were provided. We categorized the models as follows:

LLMs as Agents   Several works have shown that LLMs can act as powerful agents that have the ability to predict feasible actions and interact with the environment by prompting [25, 53, 28]. As the textual input is the accessibility tree representation of the webpage, we categorized the text-based agents into 4 groups: 1) pretrained open-source LLMs, like CodeLLaMA[37]; 2) text decoders from pretrained open-source LMM, like Fuyu-8B [4]; 3) API-based LLMs, like GPT-4 [29] and Gemini-Pro [42]; 4) pretrained web-based agents with only the language module enabled, like WebShop [49].

LMMs as Agents   With the evolution of LMMs, particularly in their remarkable capabilities for multimodal comprehension and reasoning, we embarked on a series of experiments to integrate multimodal information into task evaluation. Our experiments involved prominent LMMs such as Fuyu-8B [4], Gemini-Pro-Vision [1], and GPT-4V [2].

SR(%) / H.C.	1st	2nd	3rd	4th	5th	6th
2	56.50	11.00	-	-	-	-
3	22.73	4.55	0.00	-	-	-
4	12.50	0.00	0.00	0.00	-	-
5	12.28	1.75	0.00	0.00	0.00	-
6	16.67	0.00	0.00	0.00	0.00	0.00

a GPT-4V

SR(%) / H.C.	1st	2nd	3rd	4th	5th	6th
2	69.28	8.43	-	-	-	-
3	32.56	0.00	0.00	-	-	-
4	40.00	0.00	0.00	0.00	-	-
5	41.67	5.00	0.00	0.00	0.00	-
6	31.03	1.72	0.00	0.00	0.00	0.00

b Gemini-Pro-Vision

Table 4: **Evaluation by Hops.** Tables (a) and (b) display the hop success rates (SR), distinguished by hop counts (H.C.) of the tasks ranging from 2 to 6, for the models GPT-4V and Gemini-Pro-Vision respectively. Higher success rates are marked with darker colors. Both agents fail on the early hops when solving tasks of more hops.

Heuristic-Based Web Agents   Previously, several heuristic-based web agents were specifically crafted with the intention of navigation and completion of web-based tasks [49, 8, 52]. To test the generalization ability of this category of agents, we adapted WebShop in both LLM and LMM settings with the help of GPT-3.5-turbo to format the input into the required format.

Human Baselines   We conducted a comparison of hop and task performances within the same settings as an average of three human test takers. The test takers come from various socioeconomic backgrounds, without information on the tasks before the evaluation. Human baselines consistently outperform all existing web agents with significant margins.

4.2   Main Results

The results for the different models are shown in Tab. 3, where the hop performance and task performance are evaluated respectively. The hop success rate was calculated by the percentage of successful visits for the targeted websites, and the task success rate was evaluated based on the final task completion results, *i.e.*, the percentage of successful tasks completed by agents with the total number of tasks.

From the experimental results, we found current state-of-the-art models showed significantly degraded performance regarding the multihop tasks, which reflected their inability to effectively recognize and comprehend structured, long-context information from the web that encapsulates the fundamental details of web content. Moreover, current agents exhibited huge performance drops with the increased number of hops, revealing their limitations in long-chain reasoning capabilities.

The hop success rate, which counts every successful task completion at a website, serves as an auxiliary metric to more accurately represent the procedural performance of each agent. On single-hop tasks, GPT-4V outperformed all other models, showcasing its superior capabilities in image and context comprehension, as well as planning. However, for tasks with a hop count ranging from 2 to 4, we observed unexpected performances. Specifically, Gemini-Pro without captions and GPT-4 without captions exhibited higher performances compared to their counterparts that are augmented with captions. Further analysis of the agents' trajectories revealed that when agents are given relatively simple tasks while being under-informed or lacking sufficient information, they tend to "wander" through the given hops, which often results in a non-ending loop and ultimately leads to failure. This phenomenon explains why some agents exhibit a higher hop success rate while simultaneously having a low task success rate. The insight also justifies the need for a holistic evaluation protocol with MMInA.

The experimental results showed that: 1) *Multimodality-reliance*: Multimodal models exhibit overall higher performance in both hop and task performance that make more accurate predictions on the proposed benchmark; 2) *Context window length*: Language models specifically trained to comprehend structured and long contexts, like CodeLLaMA and GPT series, which are tailored for interpreting a highly structured text format, are more suitable to execute web-based tasks that utilize structured representations of webpages; 3) *Web-based models*: the models that trained on web-based content (*e.g.*, Fuyu-8B, WebShop) still exhibit the versatility and adaptability in unfamiliar environments.
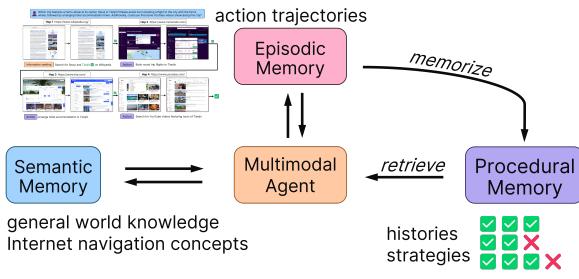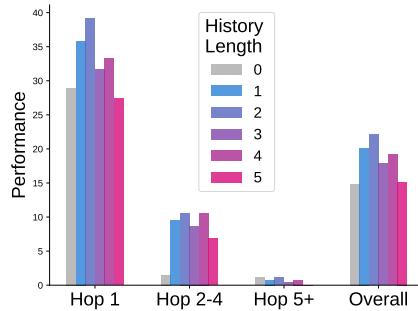
Figure 4: **Memory-augmented agents.** Our method complements LMMs by enhancing procedural memory with action trajectories on similar tasks.

Figure 5: **Success rates of memory-augmented agents, by history lengths.**

4.3 Analysis: Why are Multihop Web Tasks Challenging?

Search Space Agents often exhibit poor performance on multihop tasks, struggling to complete even the initial hops. However, when each hop is isolated into a separate single-hop task, agents generally perform well on them all.

Upon analyzing the result, we discovered that for single-hop tasks, where only one reference URL is provided, agents typically explore solely within the specific website, eventually completing the task through trial and error. In contrast, for multi-hop tasks, we provide a prompt containing all the websites necessary to complete the task. When agents fail to complete a task within the expected website, they may navigate to other websites in an attempt to proceed. This excessive amount of observation can confuse the agent's focus, leading to a catastrophic decrease in task completion rates.

For instance, in a task like "Help me book a flight to Tokyo, find a tour guide, search for YouTube videos, rent a car, and book a hotel," the agent may struggle with the initial step of entering the destination and clicking "search" on a flight booking website. It might then attempt to navigate to a "trip guide" webpage, and realize it cannot complete the task. Despite returning to the flight booking website, the agent's limited memory prevents it from recalling previous actions in multi-step tasks. As a result, the agent may repeat actions in a cycle within the task, illustrating the challenges of multi-step tasks.

Additionally, we provide the agent with the termination condition for each hop as the prompt. However, the agent occasionally struggles to extract this crucial information, leading to a lack of awareness of the correct endpoint. As a result, it may become stuck in the current hop even after completing it, failing to proceed to the next hop to accomplish the task.

Agent Input Length The total hop count of a multihop task defines the length of the task. It is assumed that the success rate of each hop is contingent upon the successful completion of the preceding hop, and the overall number of hops should not influence the success rate of any individual hop, provided that the hop is defined within a specific domain. However, after aligning the first hop semantically across all tasks, our empirical findings in Tab. 4 indicated an unexpected pattern that was contrary to the assumption above. We observed that agents performed better on tasks with fewer total hops, achieving higher success rates in completing the first hop. Conversely, as the total hop count increased, there was a noticeable decline in the success rate within the first hop. We attribute this phenomenon to the enlarged search space and the agent's weak zero-shot long-context reasoning capabilities, which we resolve in Sec. 4.4.

This observation underscores the complexity inherent in multihop tasks, which extends beyond merely accumulating the performance of single-hop tasks. It poses a unique challenge that requires the coordination, handling, and control of the entire task flow. This complexity highlights the need for web agents to develop enhanced planning and reasoning capabilities to adeptly manage and execute these intricate tasks.

4.4 Memory-augmented Agents

Agents operating within dynamic environments must execute actions that are informed by not only real-time environment observations and user queries but also their historical action patterns. Based on the previous experimental results, the complexity of predicting actions lies in the requirement for diverse types of memory at different phases of this process, emphasizing the importance of retaining information through various tasks, actions, and web interactions. We propose memory-augmented web agents, characterized by three specific memory systems: semantic memory, episodic memory, and procedural memory. Semantic memory comprises the agent's general world knowledge, which can be continually sourced and updated from the Internet or comprehensive knowledge bases. In practice, semantic memory is encoded in LLM weights. Episodic memory acts as a repository to temporarily store the captured step-wise action trajectories of individual tasks, allowing agents to methodically record and subsequently recall each predictive step when predicting the next possible action for the specific task. In practice, it is usually in the form of the previous context for an autoregressive model or in-context examples. Procedural memory, activated upon the completion of a task's action sequence, encodes the full trajectory and outcomes (success or failure) to distill historical strategies. In this work, we highlight the importance of procedural memory that significantly improves a web agent's performance by replaying the agent's action trajectories on past tasks similar to the current one (Tab. 3). With the three memory systems integrated, multimodal agents are equipped to retrieve a wealth of relevant information, using it as a reference point to navigate and adapt to forthcoming actions, thereby ensuring a sophisticated, context-aware approach to complex task resolution.

Implementation   We implement memory-augmented agents on top of LMMs by concatenating the action trajectories (including the task descriptions and the web contents as observations) of the last K tasks to the prompt of LMMs in a sequence of tasks. The replayed experience potentially grounds the agent's reasoning, *i.e.*, prunes the search spaces of the agent. However, it extends the input length by K times, which poses a challenge to LLMs that are trained on short corpora. To balance both sides, we explore the optimal K value to construct the procedural memory. As shown in Fig. 5, the optimal performance is usually obtained at $K = 2$. Our experiments revealed that agents benefiting from procedural memory references displayed superior performance in both action prediction and execution. Nevertheless, our findings also highlighted a non-linear relationship between performance and the number of historical references. In simpler tasks, such as those within the shopping or Wikipedia domain, superior performance was observed with a small number of history references, particularly when the history number was set at 1 and 2 respectively, where the performance with a higher history number showed limited improvements from the base conditions. With the higher history number settings, the unprioritized history will introduce biases and turbulence inside the model decision process as well, causing a lower increase in the model's final performance. Notably, we performed experiments with Gemini-Pro-Vision, but the proposed method is model-agnostic to be flexibly deployed with any LMM or LLM.

5 Conclusion, Challenges, and Outlook

We present MMInA, a benchmark featuring three properties: **1)** we benchmark agents on real-world websites and propose 1,050 multimodal multihop tasks among 14 various websites. We experiment with current SOTA LLMs and LMMs and also provide human baselines; **2)** we propose a novel holistic evaluation method for multihop tasks that assess both task and hop success rates; **3)** we propose a flexible memory-augmented method to improve the performance by enhancing procedural memory of agents.

Limitations   Due to the protection mechanisms employed by web pages, it's exceptionally challenging to find a website that allows us to directly fetch images from HTML files. So one of the websites we utilized is an offline standalone website and the other is an open-source website.

Future Work   With the increasing prevalence of mobile devices, we aim to expand the task domain to include mobile platforms. Additionally, to enhance the capabilities of the agent, we will introduce a mechanism for long-term memory, allowing the agent to selectively remember useful actions taken during a task. For multihop tasks, our current evaluation methods primarily rely on keyword queries within website URLs or indirect approaches for assessment. Moving forward, we will consider employing an evaluation method focused on actions, which will directly guide the agent's operations.

Potential Broader Impact Our benchmark provides a testing bed for subsequent agent research and identifies certain issues that can guide future related studies. We do not identify any potential negative impact of this work.

References

[1] Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. 2023. 7

[2] Gpt-4v(ision) system card. 2023. 7

[3] 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. 19

[4] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. 7

[5] 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. 20

[6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 19

[7] Biplab Deka, Zifeng Huang, and Ranjitha Kumar. Erica: Interaction mining mobile apps. In *Proceedings of the 29th annual symposium on user interface software and technology*, pages 767–776, 2016. 20

[8] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *arXiv preprint arXiv:2306.06070*, 2023. 2, 3, 8

[9] Razvan V Florian. Autonomous artificial intelligent agents. *Center for Cognitive and Neural Studies (Coneural), Cluj-Napoca, Romania*, 2003. 1

[10] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 19

[11] Tianyu Gao, Zirui Wang, Adithya Bhaskar, and Danqi Chen. Improving language understanding from screenshots, 2024. 20

[12] Yingqiang Ge, Wenyue Hua, Jianchao Ji, Juntao Tan, Shuyuan Xu, and Yongfeng Zhang. Openagi: When llm meets domain experts. *arXiv preprint arXiv:2304.04370*, 2023. 3

[13] 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*, 2024. 4

[14] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2023. 1

[15] Raghav Kapoor, Yash Parag Butala, Melisa Russak, Jing Yu Koh, Kiran Kamble, Waseem Alshikh, and Ruslan Salakhutdinov. Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. *arXiv preprint arXiv:2402.17553*, 2024. 20

[16] Jihyung Kil, Chan Hee Song, Boyuan Zheng, Xiang Deng, Yu Su, and Wei-Lun Chao. Dual-view visual contextualization for web navigation. *arXiv preprint arXiv:2402.04476*, 2024. 20

[17] 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*, 2024. 1, 3, 4, 5

[18] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023. 20

[19] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 19

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 7

[21] Minghao Li, Feifan Song, Bowen Yu, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. Api-bank: A benchmark for tool-augmented llms. *arXiv preprint arXiv:2304.08244*, 2023. 1, 3

[22] Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, and Percy Liang. Reinforcement learning on web interfaces using workflow-guided exploration. *arXiv preprint arXiv:1802.08802*, 2018. 3

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 19

[24] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023. 1, 3

[25] Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, et al. Bolaa: Benchmarking and orchestrating llm-augmented autonomous agents. *arXiv preprint arXiv:2308.05960*, 2023. 1, 3, 7

[26] Xing Han Lù, Zdeněk Kasner, and Siva Reddy. Weblinx: Real-world website navigation with multi-turn dialogue. *arXiv preprint arXiv:2402.05930*, 2024. 20

[27] Pattie Maes. Modeling adaptive autonomous agents. *Artificial life*, 1(1_2):135–162, 1993. 1

[28] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023. 3, 7

[29] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan,

Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. 2, 7

[30] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 19

[31] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv. org/abs/2203.02155*, 13, 2022. 19

[32] 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. 3

[33] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018. 1

[34] 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. *arXiv preprint arXiv:2103.00020*, 2021. 19

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 19

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 19

[37] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023. 7

[38] Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. Language modelling with pixels, 2023. 20

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 19

[40] Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. World of bits: An open-domain platform for web-based agents. In *International Conference on Machine Learning*, pages 3135–3144. PMLR, 2017. 1, 3

[41] Luc Steels and Rodney Brooks. *The artificial life route to artificial intelligence: Building embodied, situated agents*. Routledge, 2018. 1

[42] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan

Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova,

15

Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2023. 7

[43] Daniel Toyama, Philippe Hamel, Anita Gergely, Gheorghe Comanici, Amelia Glaese, Zafarali Ahmed, Tyler Jackson, Shibl Mourad, and Doina Precup. Androidenv: A reinforcement learning platform for android. *arXiv preprint arXiv:2105.13231*, 2021. 1

[44] Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *arXiv preprint arXiv:2401.16158*, 2024. 20

[45] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 19

[46] Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. Travelplanner: A benchmark for real-world planning with language agents. *arXiv preprint arXiv:2402.01622*, 2024. 20

[47] Nancy Xu, Sam Masling, Michael Du, Giovanni Campagna, Larry Heck, James Landay, and Monica S Lam. Grounding open-domain instructions to automate web support tasks. *arXiv preprint arXiv:2103.16057*, 2021. 3

[48] Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*, 2023. 20

[49] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022. 3, 7, 8

[50] 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. 1, 2

[51] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*, 2023. 19

[52] 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. 3, 8

[53] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023. 1, 2, 3, 4, 6, 7

[54] Tom Ziemke. Adaptive behavior in autonomous agents. *Presence*, 7(6):564–587, 1998. 1

A MMInA Benchmark Details

A.1 Website Links

Table A1: **Links to MMInA websites.** All of them are evolving real-world websites with content updated over time except Shopping. Our flexible evaluation protocol facilitates supporting more websites in the future.

Description	URL
Wikipedia[1]	https://library.kiwix.org/viewer#wikipedia_en_all_maxi_2024-01/A/User%3AThe_other_Kiwix_guy/Landing
Car renting	https://www.trip.com/carhire/
Flight booking	https://www.momondo.com/
Hotel booking	https://www.trip.com/hotels/
Event searching	https://www.eventbrite.com/
Twitter	https://twitter.com/home
Amazon	https://www.amazon.com/
YouTube	https://www.youtube.com/
Find food	https://www.timeout.com/
Exchange dollars	https://www.xe.com/
Travel guide	https://www.nomadicmatt.com
Recipes	https://www.allrecipes.com/
Train booking	https://www.trip.com/trains/
Shopping	OneStopMarket (an offline standalone website)

[1] Since libraries in Kiwix may update, resulting in URLs with advanced dates, it's advisable to verify the Wikipedia library on the official Kiwix page. However, this doesn't affect our experiments.

B Experiment Details

B.1 Supplementary Results

Hop Analysis We follow the previous settings of hop analysis in the main paper, illustrating the agent performance of a GPT-4V agent in Table A2. We observed again that agents performed better on tasks with fewer total hops, achieving higher success rates in completing the first hop. Conversely, as the total hop count increased, there was a noticeable decline in the success rate within the first hop. Because there are fewer long-range (>7 hops) tasks, the success rates fluctuate due to randomness.

GPT4V	count	sr1	sr2	sr3	sr4	sr5	sr6	sr7	sr8	sr9	sr10
2-h	200	56.50	11.00								
3-h	44	22.73	4.55	0.00							
4-h	16	12.50	0.00	0.00	0.00						
5-h	57	12.28	1.75	0.00	0.00	0.00					
6-h	60	16.67	0.00	0.00	0.00	0.00	0.00				
7-h	59	25.42	0.00	0.00	0.00	0.00	0.00	0.00			
8-h	35	40.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		
9-h	30	56.67	20.00	3.33	0.00	0.00	0.00	0.00	0.00	0.00	
10-h	19	52.63	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table A2: **Performance of GPT-4V on multihop tasks based on the hop number ranging from 2 to 10.** The success rate (sr) is calculated based on single-hop evaluation results over the whole completed hop numbers.

B.2 More Related Works

The literature review will be divided into several key sections, each focusing on a critical aspect of research related to the development of multimodal autonomous web agents. This comprehensive

overview will delve into multimodal datasets, large language/multimodal models as backbones, and various types of autonomous agents, including embodied agents and web agents. The aim is to provide a thorough understanding of the current state of research in these areas, as well as to identify gaps and provide insights for future work.

Multimodal Datasets Recent progress in multimodal learning, showcased by models like CLIP[34], DALL-E[35], Stable Diffusion[36], Flamingo[3], and GPT series, has led to significant improvements in areas such as zero-shot classification, image creation, and in-context learning. Although these models employ various algorithmic approaches, including contrastive learning, diffusion techniques, and auto-regressive models, they share a fundamental reliance on large datasets comprising image-text pairs. This commonality underscores the importance of such datasets in driving advancements in multimodal AI capabilities.

Webdataset[1] is a commonly used dataset as it contains thousands of image-text pairs data scraped from websites; LAINON-5b[39] is a dataset that contains 5.85 billion CLIP-filtered image-text pairs, where 2.32B contain English language; MIMIC-IT[19] is a dataset consists of 2.8 million multimodal instruction-response pairs equipped with rich in-context information, with 2.2 million unique instructions derived from images and videos; DataComp[10] is a newly brought dataset which consists of four stages: A) Deciding on a scale that fits within resource limitations. B) Creating a dataset, opting for either the filtering approach or the Bring Your Own Data (BYOD) track. C) Using a set architecture and specific hyperparameters to train a CLIP model on the created dataset. D) Assessing the performance of the trained model across a variety of downstream tasks.

Large Language/Multimodal Models as Backbones Instruction tuning is a common method used in LLMs training, which involves refining pre-trained LLMs using datasets formatted as instructions. This approach enhances the model's ability to perform new, unseen tasks by simply following directions, thereby improving its zero-shot capabilities. Some notable models like ChatGPT[30], InstructGPT[31], FLAN[45, 6] are built on top of instruction tuning methods.

Inherited the success from LLMs, LMM training is also extended to the instruction-tuning methods by utilizing the multimodal instruction data, which contains: a textual `<instruction>` to describe the task; `<image>`, `<text>` pair as input to enable the multimodalities; the model output with a token `<output><EOS>` to identify the end of the output. A multimodal instruction sample can be denoted in a triplet form, i.e., I, M, R, where I, M, R represent the instruction, the multimodal input, and the ground truth response, respectively. The LMM predicts an answer given the instruction and the multimodal input:

$$A = f(I, M; \theta)$$

, and the optimizing objective can be formulated as:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log p(R_i | I, R_{<i}; \theta) \tag{1}$$

1) Tranformation The efficacy of instruction tuning in the training of LMMs is significantly constrained by the limitations in length and type of data available in current Visual Question Answering (VQA) datasets. To address this, some researchers have opted to adapt the provided instructions, transforming the succinct answer data into extended sentences enriched with semantic details [51]. Other studies, such as in, reconstructed the answer by prompting ChatGPT to emulate the capabilities of advanced language models.

2) Self-Instruct LLaVA [23] extends the multimodal approach by converting images into descriptive texts and outlines of bounding boxes, then uses GPT-4 to create additional data within the context provided by initial examples.

Autonomous Agents in Virtual World Agents designed for Graphical User Interfaces (GUIs) are crafted to streamline complex activities on digital devices like smartphones and desktops. These GUI agents may employ HTML as inputs, or alternatively, use screenshots to facilitate task execution in a broader context. Traditionally, research has revolved around training these agents in restrictive, static environments, a practice that deviates from human learning and hinders the agents' ability to

---

[1] https://github.com/webdataset/webdataset

make decisions akin to humans. However, the emergence of large language models (LLMs) and large multimodal models (LMMs) equipped with vast web knowledge marks a pivotal shift towards achieving a more human-like intellect in agents, sparking a surge in research on LLM/LMM-enhanced autonomous agents. This section aims to explore the latest state-of-the-art (SOTA) developments in autonomous agents, examining both web GUI agents and mobile GUI agents.

1) GUI Agents - Web Agents Despite the current progress of web agents discussed in the main paper, several works also explored the development of web agents. TravelPlanner [46] proposed a benchmark that provides a sandbox environment with tools for accessing nearly four million data records. It includes 1,225 planning intents and reference plans to evaluate the planning strategies of language agents by using tools; OmniACT [15] presented a dataset and benchmark for assessing an agent's capability to generate executable programs for computer tasks. It uses the PyAutoGUI Python library to automate mouse and keyboard operations across different operating systems and web domains. It addresses the limitations of HTML-based agents by providing a multimodal challenge where visual cues are crucial, thus enabling a more robust understanding of UI elements, but it still showed the inability to handle native desktop applications or multi-application tasks; WEBLINX [26] also proposed a benchmark for conversational web navigation, addressing the problem of enabling a digital agent to control a web browser and follow user instructions in a multi-turn dialogue fashion. The method involves a retrieval-inspired model that prunes HTML pages by ranking relevant elements, addressing the issue of LLMs not being able to process entire web pages in real-time. The technology used includes a dense markup ranker for element selection and multimodal models that combine screenshots, action history, and textual website representation. The performance is evaluated on tasks like creating a task on Google Calendar, with the model's ability to replicate human behavior when navigating the web; DUAL-VCR [16], leverages the "dual view" of HTML elements in webpage screenshots, contextualizing each element with its visual neighbors. This approach uses both textual and visual features to create more informative representations for decision-making.

2) GUI Agents - Mobile Agents Besides web agents, mobile GUI agents are gaining more and more popularity, which are developed to handle intricate tasks automatically on digital devices like smartphones. ERICA [7] defines a system for interaction mining in Android applications, it employs a human-computer interaction approach to capture the data, making it scalable and capable of capturing a wide range of interactions. PIXEL [38] and Pix2Struct [18] showed promising capability in multilingual transfer and UI navigation respectively, but they struggle with language understanding tasks compared to text-only LMs like BERT, limiting their utility. Patch-and-Text Prediction (PTP) proposed in [11] leads to better language understanding capabilities by masking and recovering both image patches and text within screenshots; AppAgent [48] presented a multimodal agent that operates smartphone apps through low-level actions like tapping and swiping, mimicking human interactions. The agent learns app functionalities through exploration, either autonomously or by observing human demonstrations, and then applies this knowledge to execute tasks; Mobile-Agent [44], which uses visual perception tools to locate operations on a mobile device using screenshots. It involves OCR models for identifying visual and textual elements, while realizing self-planning and self-reflection to autonomously navigate mobile apps, with a benchmark called Mobile-Eval introduced for performance evaluation; SeeClick [5] is a visual GUI agent that operates solely on screenshots, bypassing the need for structured text. It employs Large Vision-Language Models (LVLMs) enhanced with GUI grounding pre-training to accurately locate screen elements based on instructions. The method involves automating the curation of GUI grounding data and creating a GUI grounding benchmark named ScreenSpot. It adapts universally to various GUI platforms and its reliance on screenshots. simplifying action space to clicking and typing.