https://doi.org/10.1007/s44267-025-00093-y

Visual Intelligence

RESEARCH **Open Access**

Large multimodal agents: a survey

Xie Junlin[1], Zhihong Chen[1], Ruifei Zhang[1] and Guanbin Li[2*]

Abstract

Large language models (LLMs) have achieved superior performance in powering text-based AI agents, endowing them with decision-making and reasoning abilities that are analogous to those exhibited by humans. Concurrently, an emerging research trend is focused on extending these LLM-powered AI agents into the *multimodal* domain. This extension facilitates the interpretation and response of AI agents to diverse multimodal user queries, thereby handling more intricate and nuanced tasks. In this paper, we conduct a systematic review of LLM-driven multimodal agents, which we refer to as *large multimodal agents* (`LMAs` for short). First, we introduce the essential components involved in developing `LMAs` and categorize the current body of research into four distinct types. Subsequently, we review the collaborative frameworks that integrate multiple `LMAs`, with the aim of enhancing collective efficacy. One of the critical challenges in this field is the diverse evaluation methods used across existing studies, which impedes effective comparison among different `LMAs`. Therefore, we compile these evaluation methodologies and establish a comprehensive framework to bridge the gaps. This framework aims to standardize evaluations, facilitating more meaningful comparisons. Concluding our review, we highlight the extensive applications of `LMAs` and propose potential future research directions. Our discussion aims to provide valuable insights and guidelines for future research in this rapidly evolving field.

Keywords: Large multimodal agents, Comprehensive framework, AI agents

1 Introduction

An *agent* is a system capable of perceiving its environment and making decisions based on these perceptions to achieve specific goals [1]. While proficient in narrow domains, early agents [2, 3] often lack adaptability and generalization, highlighting a significant disparity with human intelligence. Recent advancements in large language models (LLMs) have begun to bridge this gap, where LLMs enhance their capabilities in command interpretation, knowledge assimilation [4, 5], and mimicry of human reasoning and learning [6, 7]. These agents primarily utilize LLMs as their decision-making tool and are further enhanced with critical human-like attributes, such as memory. This enhancement allows them to handle a variety of natural language processing tasks and interact with the environment using language [8, 9]. However, real-world scenarios often involve information that spans beyond text, encompassing multiple modalities, with a significant emphasis on the visual aspect. Consequently, the next evolutionary step for LLM-powered intelligent agents is to acquire the capability to process and generate *multimodal* information, particularly visual data. This ability is essential for these agents to evolve into more sophisticated AI entities that exhibit human-level intelligence. Agents equipped with this capability are referred to as *large multimodal agents* (`LMAs`) in our paper.[1] Typically, they face more sophisticated challenges than language-only agents. Take web searching for example, an `LMA` first requires the input of a user requirements to look up relevant information through a search bar. Subsequently, it navigates to web pages through mouse clicks and scrolls to browse real-time content on those pages. Finally, the `LMA` needs to process

*Correspondence: liguanbin@mail.sysu.edu.cn
[2]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, 510275, China
Full list of author information is available at the end of the article

[1] The name derives from "large multimodal models" (LMMs).

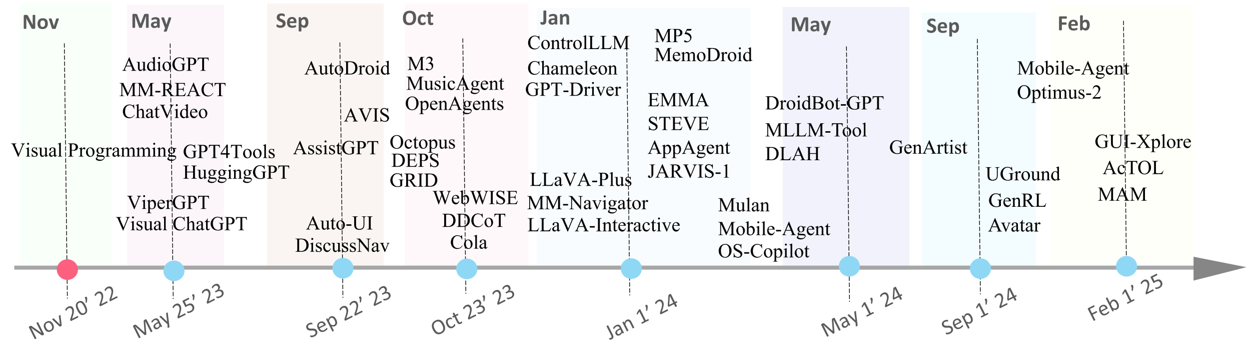


Figure 1 Representative research papers from top AI conferences on LLM-powered multimodal agents. These papers, published between November 2022 and February 2025, are categorized by their publication date. These papers include Visual Programming [13], AudioGPT [14], MM-REACT [15], ChatVideo [16], ViperGPT [17], Visual ChatGPT [18], GPT4Tools [19], HuggingGPT [20], AutoDroid [21], AssistGPT [22], AVIS [23], Auto-UI [24], DiscussNav [25], M3 [26], MusicAgent [27], OpenAgents [28], Octopus [29], DEPS [30], GRID [31], WebWISE [32], DDCoT [33], Cola [11], ControlLLM [34], Chameleon [35], GPT-Driver [36], LLaVA-Plus [37], MM-Navigator [38], LLaVA-Interactive [39], MP5 [40], MemoDroid [41], EMMA [42], STEVE [43], AppAgent [44], JARVIS-1 [45], DroidBot-GPT [46], MLLM-Tool [47], DLAH [48], GenArtist [49], UGround [50], GenRL [51], Avatar [52], Mobile-Agent [53], Optimus-2 [54], GUI-Xplore [55], AcTOL [56], and MAM [57]

multimodal data (e.g., text, videos, and images) and perform multi-step reasoning, including extracting key information from web articles, video reports, and social media updates, and integrating this information to respond to the user's query. Figure 1 shows the representative LLM-powered multimodal agents. We note that existing studies in LMAs were conducted in isolation, and therefore it is necessary to further advance the field by summarizing and comparing existing frameworks. There exist several surveys related to LLM-powered agents [10–12] while few of them focused on the multimodal aspects.

In this paper, we aim to fill the gap by summarizing the main developments of LMAs (in Fig. 1). First, we give an introduction about the core components (Sect. 2) and propose a new taxonomy for existing studies (Sect. 3) with further discussion on existing collaborative frameworks (Sect. 4). Regarding the evaluation, we outline the existing methodologies for assessing the performance of LMAs, followed by a comprehensive summary (Sect. 5). Then, the application section provides an exhaustive overview of the broad real-world applications of multimodal agents and their related tasks (Sect. 6). We conclude this work by discussing and suggesting possible future directions for LMAs to provide useful research guidance.

2 The core components of LMAs

In this section, we detail four core elements of LMAs including perception, planning, action, and memory.

Perception Perception is a complex cognitive process that enables humans to collect and interpret environmental information. In LMAs, the perception component primarily focuses on processing multimodal information from diverse environments. As illustrated in Table 1, LMAs in different tasks involve various modalities. They require extracting key information from these different modalities that is most beneficial for task completion, thereby facilitating more effective planning and execution of the tasks.

Early research [15, 17, 18, 22] on processing multimodal information often relies on simple correlation models or tools that convert images or audio into text descriptions. However, this conversion approach tends to generate a large amount of irrelevant and redundant information, particularly for complex modalities (e.g., video). Along with the input length constraint, LLMs frequently face challenges in effectively extracting pertinent information for planning. To address this issue, recent studies [16, 65] have introduced the concept of sub-task tools, which are designed to handle sophisticated data types. In an environment resembling the real world (i.e., open-world games), Wang et al. [45] proposed a novel method for processing non-textual modal information. This approach begins by extracting key visual vocabulary from the environment and then employs the GPT model to further refine this vocabulary into a series of descriptive sentences. As LLMs perceive visual modalities within the environment, they use visual modalities to retrieve the most relevant descriptive sentences, which effectively enhances their understanding of the surroundings.

Planning Planners play a central role in LMAs, akin to the function of the human brain. They are responsible for thinking deeply about the current task and formulating corresponding plans. Compared to language-only agents, LMAs operate in a more complicated environment, making it more challenging to devise reasonable plans. We detail planners from four perspectives (models, format, inspection & reflection, and planning methods).

Table 1 Core components of large multimodal agents (LMAs). This presentation delineates the component details of all `LMAs`, encompassing their task-specific modalities, the models outlined by planners, the methodologies and formats employed in planning, the variety of actions involved, the extent of multi-agent collaboration, and the incorporation of long-term memory. "V"represents the virtual action, "T" indicates the use of a tool, and "E" embodies the physical action

Type	Model	Task focus				Planner				Action		Multi-agent	Long memory
		Text	Image	Video	Audio	Model	Format	Inspect	Planning method	Action type	Action learning		
Type I	VisProg [13]	✓	✓	✗	✗	GPT3.5	Program	✗	Fix	T (VFMs & Python)	Prompt	✗	✗
	ControlLLM [34]	✓	✓	✗	✗	GPT4	Language	✓	Fix	T (VFMs & Python)	Prompt	✗	✗
	Visual ChatGPT [18]	✗	✓	✗	✗	GPT3.5	Language	✗	Fix	T (VFMs)	Prompt	✗	✗
	ViperGPT [17]	✓	✓	✓	✗	GPT3.5	Program	✗	Fix	T (VFMs & API & Python)	Prompt	✗	✗
	MM-REACT [15]	✗	✓	✓	✗	ChatGPT & GPT3.5	Language	✓	Fix	T (Web & API)	Prompt	✗	✗
	Chameleon [35]	✓	✓	✗	✗	GPT3.5	Language	✗	Fix	T (VFMs & API & Python & Web)	Prompt	✗	✗
	HuggingGPT [20]	✗	✓	✓	✗	GPT3.5	Language	✗	Fix	T (VFMs)	Prompt	✗	✗
	CLOVA [58]	✓	✓	✗	✗	GPT4	Language	✓	Dynamic	T (VFMs & API)	Prompt	✗	✗
	CRAFT [59]	✓	✓	✗	✗	GPT4	Program	✗	Fix	T (Custom tools)	Prompt	✗	✗
	Cola [60]	✓	✓	✓	✗	ChatGPT	Language	✗	Fix	T (VFMs)	Prompt	✗	✗
	M3 [26]	✗	✓	✓	✗	GPT3.5	Language	✓	Dynamic	T (VFMs & API)	Prompt	✗	✗
	DEPS [61]	✓	✓	✗	✗	GPT-4	Language	✓	Dynamic	E	Prompt	✗	✗
	GRID [31]	✓	✓	✗	✗	GPT-4	Language	✓	Dynamic	T (VFMs & API)	Prompt	✗	✗
	DroidBot-GPT [46]	✓	✓	✗	✗	ChatGPT	Language	✗	Dynamic	V	Prompt	✗	✗
	ASSISTGUI [62]	✓	✓	✗	✗	GPT-4	Language	✓	Dynamic	V(GUI parser)	Prompt	✗	✗
	GPT-Driver [36]	✓	✓	✗	✗	GPT-3.5	Language	✗	Dynamic	E	Prompt	✗	✗
	LLaVA-Interactive [39]	✓	✓	✗	✗	GPT-4	Language	✗	Dynamic	T (VFMs)	Prompt	✗	✗
	MusicAgent [27]	✓	✗	✗	✓	ChatGPT	Language	✗	Dynamic	T (Music-Models)	Prompt	✗	✗
	AudioGPT [14]	✓	✗	✗	✓	GPT-4(V)	Language	✗	Fix	T (API)	Prompt	✗	✗
	AssistGPT [22]	✗	✓	✓	✗	GPT3.5	Lang. & Prog.	✓	Dynamic	T (VFMs & API)	Prompt	✗	✗
	MuLan [63]	✓	✓	✗	✗	GPT3.5	Language	✓	Dynamic	T (VFMs & Python)	Prompt	✗	✗
	Mobile-Agent [53]	✓	✓	✗	✗	GPT4	Language	✓	Dynamic	V & T(VFMs)	Prompt	✗	✗

Table 1 (*Continued*)

Type	Model	Task focus				Planner				Action		Multi-agent	Long memory
		Text	Image	Video	Audio	Model	Format	Inspect	Planning method	Action type	Action learning		
Type II	GPT-Driver [36]	✗	✓	✓	✗	GPT4	Language	✗	Fix	E	Learning	✗	✗
	LLAVA-PLUS [37]	✗	✓	✓	✗	Llava	Language	✓	Dynamic	T (VFMs)	Learning	✗	✗
	GPT4Tools [19]	✗	✓	✓	✗	Llama	Language	✓	Dynamic	T (VFMs)	Learning	✗	✗
	Tool-LMM [47]	✓	✓	✓	✓	Vicuna	Language	✗	Dynamic	T (VFMs & API)	Learning	✗	✗
	STEVE [43]	✓	✓	✗	✗	STEVE-13B	Program	✗	Fix	E	Learning	✗	✗
	EMMA [42]	✓	✓	✗	✗	LMDecoder	Language	✗	Fix	E	Learning	✗	✗
	Auto-UI [64]	✓	✓	✗	✗	LMDecoder	Language	✗	Dynamic	E	Learning	✗	✗
	WebWISE [32]	✓	✓	✗	✗	LMDecoder	Language	✓	Dynamic	E	Learning	✗	✗
	AcTOL [56]	✓	✓	✗	✗	LMDecoder	Language	✓	Dynamic	E	Learning	✗	✗
Type III	DoraemonGPT [65]	✓	✗	✓	✗	GPT4	Language	✗	Dynamic	T (VFMs)	Prompt	✗	✓
	ChatVideo [16]	✓	✓	✓	✗	ChatGPT	Language	✗	Fix	T (VFMs)	Prompt	✗	✓
	OS-Copilot [66]	✓	✓	✗	✗	GPT4	Language	✓	Dynamic	V	Prompt	✗	✓
Type IV	OpenAgents [28]	✓	✓	✓	✗	GPT3.5 & GPT4	Language	✓	Dynamic	V & T	Prompt	✗	✓
	MEIA [67]	✓	✓	✗	✗	GPT3.5 & GPT4	Language	✗	Fix	E	Prompt	✗	✓
	JARVIS-1 [45]	✓	✓	✗	✗	GPT-4	Language	✓	Dynamic	E	Prompt	✗	✓
	AppAgent [44]	✓	✓	✗	✗	GPT-4(V)	Language	✓	Dynamic	E	Prompt	✗	✓
	MM-Navigator [38]	✓	✓	✗	✗	GPT-4(V)	Language	✗	Dynamic	T (API)	Prompt	✗	✓
	DLAH [48]	✓	✓	✗	✗	GPT-3.5	Language	✗	Dynamic	V(Simulator-interfaces)	Prompt	✗	✓
	Copilot [68]	✓	✗	✗	✓	GPT-3.5	Language	✓	Dynamic	T (Music-Models)	Prompt	✗	✓
	Wavjourney [69]	✓	✗	✗	✓	GPT-4	Program	✗	Dynamic	T (Music-Models)	Prompt	✗	✓
	Optimus-2 [54]	✓	✓	✓	✗	GPT-4V	Text	✗	Dynamic	E	Prompt	✗	✓
	Wavjourney [69]	✓	✗	✗	✓	GPT-4	Program	✗	Dynamic	T (Music-Models)	Prompt	✗	✓
	Optimus-2 [54]	✓	✓	✓	✗	GPT-4V	Text	✗	Dynamic	E	Dynamic	✗	✓
	GUI-Xplore [55]	✓	✓	✗	✗	GPT-4V	Text	✓	Dynamic	V(Simulator-interfaces)	Dynamic	✗	✓
Multi-agent	AVIS [23]	✗	✓	✓	✗	GPT4	Language	✓	Dynamic	T (VFMs)	Prompt	✓	✗
	MP5 [40]	✓	✓	✗	✗	GPT-4	Language	✓	Dynamic	E	Prompt	✓	✓
	MemoDroid [41]	✓	✓	✗	✗	GPT-4	Language	✓	Dynamic	V(VFMs & API)	Prompt	✓	✓
	DiscussNav [25]	✓	✓	✗	✗	GPT-4 & ChatGPT	Language	✗	Dynamic	E	Prompt	✓	✗

Table 2 A summary of different tools. Their corresponding modalities, skills, and available sources are presented

Modality	Skill	Tools	Source
Image	VQA	BLIP2 [70]	Github, Hugging Face
	Grounding/Detection	G-DINO [71]	Github, Hugging Face
	Image caption	BLIP [72],BLIP2 [70], InstructBLIP [73]	Github, Hugging Face, API
	OCR	EasyOCR, Umi-OCR	Github, API
	Image editing	Instruct P2P [74]	Github, Hugging Face, API
	Image generation	Stable Diffusion [75], DALLE-3 [76]	Github, Hugging Face, API
	Image segmentation	SAM [77], PaddleSeg [78]	Github, Hugging Face, API
Text	Knowledge retrieval	Bing Search	Website, API
	Programming related skill	PyLint, PyChecker	Python, API
Video	Video editing	Editly	Github, API
	Object tracking	OSTrack [79]	Github, Hugging Face, API
Audio	Speech to text	Whisper [80]	Github, Hugging Face, API
	Text to speech	StyleTTS 2 [6]	Github, API

1) **Models**: As shown in Table 1, existing studies employ different models as planners. Among them, the most popular ones are GPT-3.5 or GPT-4 [17, 18, 20, 22, 35, 45]. Yet, these models are not publicly available and therefore some studies have begun shifting towards using open-source models, such as LLaMA [19] and LLaVA [37], where the latter can directly process information of multiple modalities, enhancing their ability to make more optimal plans.
2) **Format**: It represents how to formulate the plans made by planners. As shown in Table 1, there are two formatting ways. The first one is natural language. For example, in Ref. [20], the planning content obtained is "*The first thing I did was use OpenCV's openpose control model to analyze the pose of the boy in the image....*", where the plan made is to use "*OpenCV's openpose control model*". The second one is in the form of programs, like "*image_patch = ImagePatch(image)*" as described in Ref. [17], which invokes the *ImagePatch* function to execute the planning. There are also hybrid forms, such as Ref. [22].
3) **Inspection & Reflection**: It is challenging for an LMA to consistently make meaningful and task-completing plans in a complex multimodal environment. This component aims at enhancing robustness and adaptability. Some research methods [45, 61] store successful experiences in long-term memory, including the multimodal states, to guide planning. During the planning process, they first retrieve relevant experiences, aiding planners in thoughtful deliberation to reduce uncertainty. Additionally, Ref. [23] utilizes plans made by humans in different states while performing the same tasks. When encountering similar states, planners can refer to these "standard answers" for contemplation, leading to more rational plans. Moreover, Ref. [65] employs more complex planning methods, like Monte Carlo, to expand the scope of planning search to find the optimal planning strategy.
4) **Planning Methods**: Existing planning strategies can be categorized into two types: static and dynamic planning as shown in Table 1. The former [15, 17, 18, 20, 35] refers to decomposing the goal into a series of sub-plans based on the initial input, similar to chain of thought (CoT) [33], where plans are not reformulated even if errors occur during the process. The latter [22, 26, 45, 65] implies that each plan is formulated based on the current environmental information or feedback. If errors are detected in the plan, it will revert to the original state for re-planning [23].

Action The action component in multimodal agent systems is responsible for executing the plans and decisions formulated by the planner. It translates these plans into specific actions, such as the use of tools, physical movements, or interactions with interfaces, thereby ensuring that the agent can achieve its goals and interact with the environment accurately and efficiently. Our discussion focuses on two aspects: types and approaches.

Actions in Table 1 are classified into three categories: tool use (T), embodied actions (E), and virtual actions (V), where tools includes visual foundation models (VFMs), APIs and Python (as listed in Table 2). Embodied actions are performed by physical entities like robots [36, 48] or virtual characters [31, 42, 45, 61]. Virtual actions [32, 46, 62, 81] include web tasks (e.g., clicking links, scrolling, and keyboard use). In terms of approaches, as shown in Table 1, there are primarily two types. The first type involves using prompts to provide agents with information about executable actions, such as the tools available at the moment and their functions. The second type involves collecting data on actions and leveraging this information to self-instruct the fine-tuning process of open-source large models, such as LLaVA [37]. These data are typically generated by advanced models, such as GPT-4. Compared to language-only agents, the complexity of information and data related to actions requires more sophisticated methods to optimize the learning strategy.

Memory Early studies show that memory mechanisms play a vital role in the operation of general-purpose agents. Similar to humans, memory in agents can be categorized into long and short memory. In a simple environment, short memory is sufficient for an agent to handle tasks at hand. However, in more complex and realistic settings, long memory becomes essential. In Table 1, we can see that only a minority of LMAs incorporate long memory. Unlike language-only agents, these multimodal agents require long memory capable of storing information across various modalities. In some studies [16, 44, 48, 65], all modalities are converted into textual formats for storage. However, in Ref. [45], a multimodal long memory system is proposed, designed specifically to archive previous successful experiences. Specifically, these memories are stored as key-value pairs, where the key is the multimodal state and the value is the successful plan. Upon encountering a new multimodal state x, the most analogous examples t are retrieved based on their encoded similarity:

$$p(t|x) \propto \mathrm{CLIP}_{\mathrm{v}}(k_t)^{\top}\mathrm{CLIP}_{\mathrm{v}}(k_x) \tag{1}$$

where k_t represents the key's visual information encoded via the CLIP model, compared for similarity with the current visual state k_x, also encoded by CLIP.

3 The taxonomy of LMAs

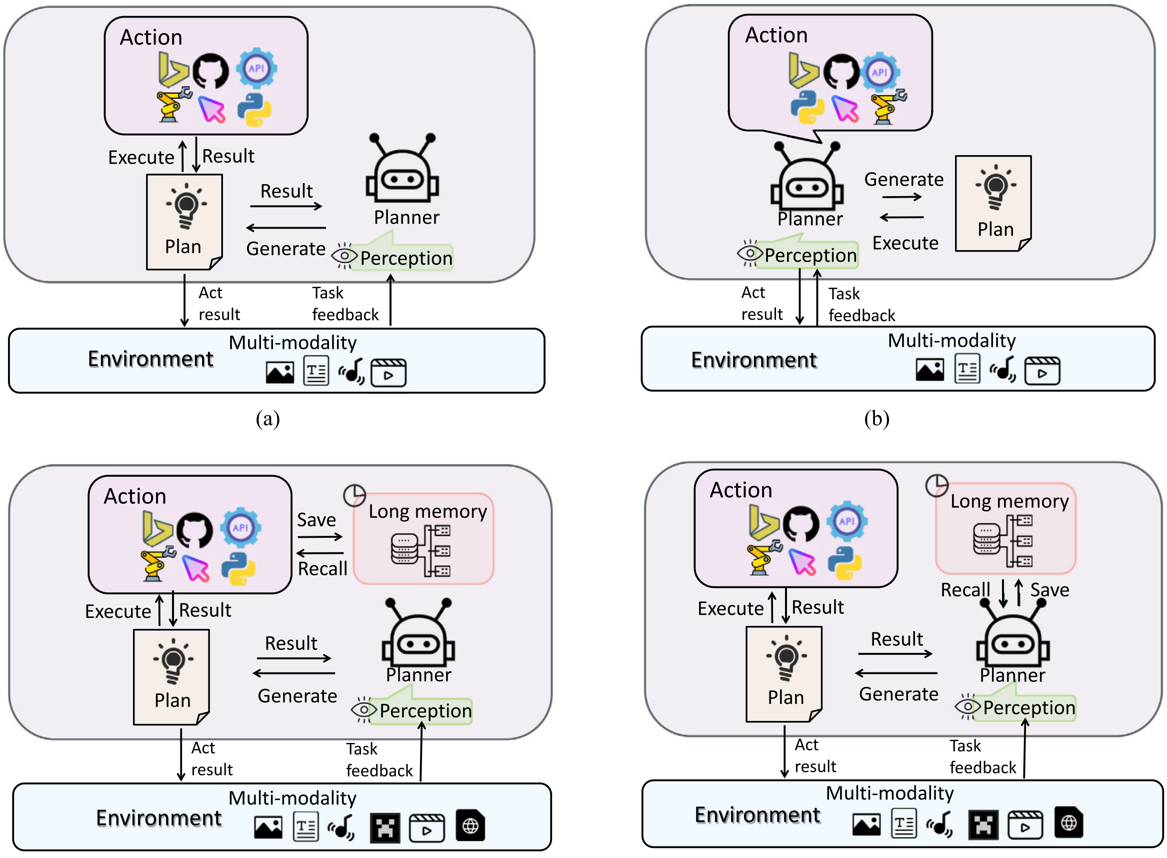
In this section, we present a taxonomy of existing studies by classifying them into four types.

Type I: closed-source LLMs as planners w/o long-term memory Early studies [13, 17, 18, 20, 22, 26] employed prompts to utilize closed-source large language models (e.g., GPT-3.5) as the planner for inference and planning as illustrated in Fig. 2(a). Depending on the specific environment or task requirements, the execution of these plans may be carried out by downstream toolkits or through direct interaction with the environment using physical devices like mice or robotic arms. LMAs of this type typically operate in simpler settings, undertaking conventional tasks such as image editing, visual grounding, and visual question answering (VQA).

Type II: fine-tuned LLMs as planners w/o long-term memory LMAs of this type involve collecting multimodal instruction-following data or employing self-instruction to fine-tune open-source large language models (such as LLaMA) [19] or multimodal models (like LLaVA) [37, 47], as illustrated in Fig. 2(b). This enhancement not only allows the models to serve as the central "brain" for reasoning and planning but also to execute these plans. The environments and tasks faced by Type II LMAs are similar to those in Type I, typically involving traditional visual or multimodal tasks. Compared to canonical scenarios characterized by relatively simple dynamics, closed environments, and basic tasks, LMAs in open-world games like Minecraft are required to execute precise planning in dynamic contexts, handle tasks of high complexity, and engage in lifelong learning to adapt to new challenges. Therefore, building upon the foundation of Type I and Type II, Type III and Type IV LMAs integrate a memory component, showing great promise in developing towards a generalist agent in the field of artificial intelligence.

Type III: planners with indirect long-term memory For Type III LMAs [16, 65], as illustrated in Fig. 2(c), LLMs function as the central planner and are equipped with long memory. These planners access and retrieve long memories by invoking relevant tools, leveraging these memories for enhanced reasoning and planning. For example, the multimodal agent framework developed in Ref. [65] is tailored for dynamic tasks such as video processing. This framework consists of a planner, a toolkit, and a task-relevant memory bank that catalogues spatial and temporal attributes. The planner employs specialized sub-task tools to query the memory bank for spatiotemporal attributes related to the video content, enabling inference on task-relevant temporal and spatial data. Stored within the toolkit, each tool is designed for specific types of spatiotemporal reasoning and acts as an executor within the framework.

Type IV: planners with native long-term memory Different from Type III, Type IV LMAs [40, 45, 48, 81] feature LLMs directly interacting with long memory, bypassing the need for tools to access long memories, as illustrated in Fig. 2(d). For example, the multimodal agent proposed in Ref. [45] demonstrates proficiency in completing over 200 distinct tasks within the open-world context of Minecraft. In their multimodal agent design, the interactive planner, merging a multimodal foundation model with an LLM, first translates environmental multimodal inputs into text. The planner further employs a self-check mechanism to anticipate and assess each step in execution, proactively spotting potential flaws and, combined with environmental feedback and self-explanation, swiftly corrects and refines plans without extra information. Moreover, this multimodal agent framework includes a novel multimodal memory. Successful task plans and their initial multimodal states are stored, and the planner retrieves



(a) (b) (c) (d)

Figure 2 Illustrations on four types of LMAs. (a) Type I: Closed-source LLMs as planners w/o long-term memory. They mainly use prompt techniques to guide closed-source LLMs in decision-making and planning to complete tasks without long memory. (b) Type II: Fine-tuned LLMs as planners w/o long-term memory. They use action-related data to fine-tune existing open-source large models, enabling them to achieve decision-making, planning, and tool invocation capabilities comparable to closed-source LLMs. Unlike (a) and (b), (c) and (d) introduce long-term memory functions, further enhancing their generalization and adaptation abilities in environments closer to the real world. However, because their planners use different methods to retrieve memories, they can be further divided into: (c) Type III: Planners with indirect long-term memory; (d) Type IV: Planners with native long-term memory

similar states from this database for new tasks, using accumulated experiences for faster, more efficient task completion.

4 Multi-agent collaboration

We further introduce the collaborative framework for LMAs beyond the discussion within isolated agents in this section.

As shown in Fig. 3(a) and Fig. 3(b), these frameworks employ multiple LMAs working collaboratively. The key distinction between the two frameworks lies in the presence or absence of a memory component, but their underlying principle is consistent: multiple LMAs have different roles and responsibilities, enabling them to coordinate actions to collectively achieve a common goal. This structure alleviates the burden on a single agent, thereby enhancing task performance [23, 25, 40, 41].

For example, in Table 1, in the multimodal agent framework proposed by Qin et al. [40], a perceiver agent is introduced to sense the multimodal environment, comprised of large multimodal models. An agent, designated as Patroller, is responsible for engaging in multiple interactions with the perceiver agent, conducting real-time checks and feedback on the perceived environmental data to ensure the accuracy of current plans and actions. When execution failures are detected or reevaluation is necessitated, Patroller provides pertinent information to the planner, prompting a reorganization or update of the action sequences under the sub-goals. The MemoDroid framework [41] comprises several key agents that collaboratively work to automate mobile tasks. The Exploration Agent is responsible for offline analysis of the target application interface, generating a list of potential sub-tasks based on UI elements, which are then stored in the ap-

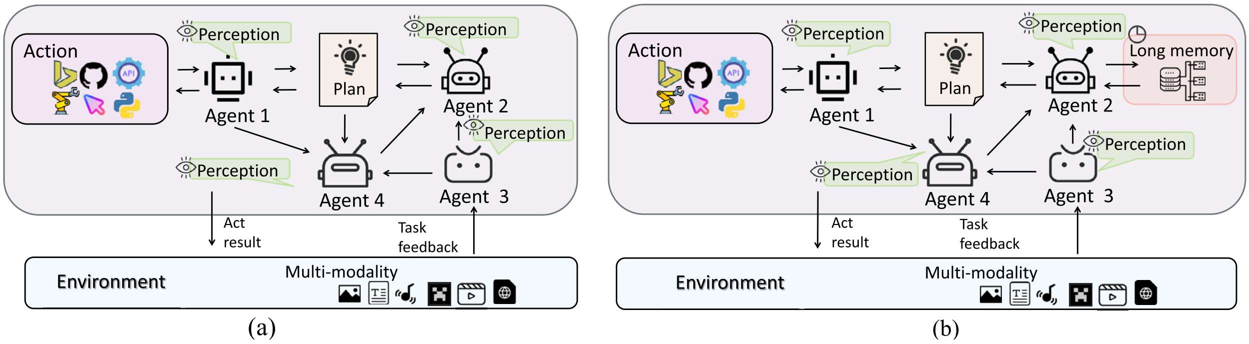


Figure 3 Illustrations on two types of multi-agent frameworks. In these two frameworks, completing tasks or instructions from the environment relies on the cooperation of multiple agents. Each agent is responsible for a specific duty, which may involve processing environmental information or handling decision-making and planning, thus distributing the pressure that would otherwise be borne by a single-agent to complete the task. The unique aspect of framework (b) is its long-term memory capability

plication memory. During the online execution phase, the Selection Agent determines specific sub-tasks to execute from the explored set, based on user commands and the current screen state. The Deduction Agent further identifies and completes the underlying action sequences required for the selected sub-tasks by prompting an LLM. Concurrently, the Recall Agent, upon encountering tasks similar to those previously learned, can directly invoke and execute the corresponding sub-tasks and action sequences from memory. The MAM [57] proposes a multi-agent framework, which decomposes medical diagnosis into specialized roles (e.g., general practitioners, radiologists) through modular design and collaborative mechanisms, overcoming the limitations of traditional unified models. Its advantages include: 1) Its modular architecture enables localized knowledge updates, thus avoiding the high cost of global retraining; 2) Flexible integration of existing specialized models simulates real-world medical team collaboration; 3) Experimental validation demonstrates superior performance over unimodal models on multimodal medical data. MetaDesigner [82] proposes a multi-agent framework for artistic font generation, where three specialized agents—Pipeline, Glyph, and Texture—collaborate to achieve end-to-end design from semantic enhancement to texture detailing. Its key advantages include: 1) Multi-agent coordination separately handles design flow, glyph structure, and texture style, improving generation precision; 2) A dynamic feedback mechanism integrates user preferences and multimodal evaluation for automatic parameter optimization; 3) Experimental validation demonstrates the system's superior performance in visual quality and semantic consistency.

5 Evaluation

The predominant focus of research is on enhancing the capabilities of current LMAs. However, limited efforts are devoted to developing methodologies for the assessment and evaluation of these agents. The majority of research continues to depend on conventional metrics for evaluating performance, clearly illustrating the challenges inherent in assessing LMAs. This also underscores the necessity of developing pragmatic assessment criteria and establishing benchmark datasets in this domain. This section summarizes existing evaluations of LMAs and offers perspectives on future developments.

5.1 Subjective evaluation

Subjective evaluation mainly refers to using humans to assess the capabilities of these LMAs. Our ultimate goal is to create a LMA that can comprehend the world like humans and autonomously execute a variety of tasks. Therefore, it is crucial to adopt subjective evaluations of human users on the capabilities of LMAs. The main evaluation metrics include versatility, user-friendliness, scalability, and value and safety.

Versatility Versatility denotes the capacity of an LMA to adeptly utilize diverse tools, execute both physical and virtual actions, and manage assorted tasks. Ref. [35] proposed comparing the scale and types of tools utilized in existing LMAs, as well as assessing the diversity of their capabilities.

User-friendliness User-friendliness involves user satisfaction with the outcomes of tasks completed by LMAs, including efficiency, accuracy, and the richness of the results. This type of assessment is relatively subjective. In Ref. [38], human evaluation of the LMA is essential to precisely assess its effectiveness in interpreting and executing user instructions.

Scalability Scalability fundamentally evaluates the capability of LMAs to assimilate new competencies and address emerging challenges. Given the dynamic nature of human requirements, it is imperative to rigorously assess the adaptability and lifelong learning potential of LMAs. For example, the evaluation in Ref. [37] focuses on the proficiency of agents in using previously unseen tools to complete tasks.

Value and safety The "value and safety" metric plays a critical role in determining the practical significance and safety of agents for human users. While many current evaluations overlook this metric, it is essential to consider the "value and safety" of LMAs. Compared to language agents, LMAs can handle a wider range of task categories, making it even more important for them to follow ethical and moral principles consistent with human societal values.

However, there are three critical limitations to current subjective evaluation methods for LMAs. First, conventional user-friendliness assessments focus solely on final task completion while neglecting the evaluation of action sequence rationality (e.g., logical errors in tool invocation order). Second, tool diversity metrics fail to effectively assess cross-modal coordination capabilities (e.g., performance when handling conflicts between visual and voice commands). Finally, safety evaluations predominantly employ predefined static scenarios, making them inadequate for identifying emergent risks from dynamic tool combinations (e.g., the potential for misinformation to propagate through the combined use of image generation tools and social media APIs).

5.2 Objective evaluation

Objective evaluation, distinct from subjective assessment, relies on quantitative metrics to comprehensively, systematically, and standardly assess the capabilities of LMAs. It is currently the most widely adopted evaluation method in multimodal agent research.

Metrics Metrics play a crucial role in objective assessment. In current multimodal agent research [15, 17, 18, 22, 23, 35, 65], specific task-related metrics are employed, such as the accuracy of answers generated by the agent in tasks like visual question answering (VQA) [17, 58]. However, the traditional task metrics established prior to the emergence of LLMs are not sufficiently effective in evaluating LLM-powered LMAs. As a result, an increasing number of research efforts are directed towards identifying more appropriate metrics for assessment. For instance, in VisualWebArena [83], a specialized assessment metric is designed to evaluate the performance of LMAs in handling visually guided tasks. This includes measuring the accuracy of the agent's visual understanding of webpage content, such as the ability to recognize and utilize interactable elements marked by Set-of-Marks for operations and achieving state transitions based on task objectives, as defined by a manually designed reward function. Besides, it encompasses the accuracy of responses to specific visual scene questions and the alignment of actions executed based on visual information.

Benchmarks Benchmark [84] represents a testing environment that encompasses a suite of evaluation standards, datasets, and tasks. It is utilized to assess and compare the performance of different algorithms or systems. Compared to benchmarks for conventional tasks [18, 23, 35, 37], SmartPlay [85] utilizes a carefully designed set of games to comprehensively measure various abilities of LMAs, establishing detailed evaluation metrics and challenge levels for each capability. Contrasting with the approach of using games to evaluate, GAIA [86] has developed a test set comprising 466 questions and their answers. These questions require AI systems to possess a range of fundamental abilities, such as reasoning, processing multimodal information, web navigation, and proficient tool use. Diverging from the current trend of creating increasingly difficult tasks for humans, it focuses on conceptually simple yet challenging questions for existing advanced AI systems. These questions involve real-world scenarios that necessitate the precise execution of complex operational sequences, with outputs that are easy to verify. Similarly, VisualWebArena [83] is a benchmark test suite designed to assess and advance the capabilities of LMAs in processing visual and textual understanding tasks on real web pages. There are also other benchmarks [87, 88] that have effectively tested the capabilities of agents.

In recent years, a growing number of benchmarks have shifted toward dynamic and realistic evaluation environments that closely mimic real-world conditions, as shown in Table 3. Moving beyond early approaches that focused solely on final task success rates, current research emphasizes multi-dimensional and fine-grained capability assessments. For instance, by simulating complex operational sequences (e.g., web navigation, multimodal information processing) or designing tasks requiring long-term planning and tool usage, these benchmarks provide a more comprehensive measure of an agent's adaptability in real-world scenarios. Furthermore, evaluation metrics have expanded from mere success rates to encompass dimensions such as task completion efficiency, robustness, and interpretability, thereby offering a more precise reflection of an agent's overall performance.

6 Application

LMAs, proficient in processing diverse data modalities, surpass language-only agents in decision-making and response generation across varied scenarios. Their adaptability makes them exceptionally useful in real-world, multisensory environments, as illustrated in Fig. 4.

Table 3 Comparison of multimodal agent benchmarks

Test task types	Benchmark	Evaluation dimensions	Task complexity
GUI tasks	Mind2web [89]	Final success rate	Closed-rule
	WebArena [90]	Final success rate and basic evaluation	Closed-rule
	VisualWebArena [83]	Final success rate and basic evaluation	Closed-rule
	Spa-Bench [91]	Final success rate and complex evaluation	Dynamic complex reasoning
	DSGBench [30]	Final success rate and complex evaluation	Closed-rule
	Balrog [92]	Final success rate and complex evaluation	Dynamic complex reasoning
Game tasks	SmartPlay [85]	Final success rate	Closed-rule and dynamic complex reasoning
	DSGBench [30]	Final success rate and complex evaluation	Closed-rule
General real-world tasks	TravelPlanner [88]	Final success rate	Closed-rule
	RiOSWorld [93]	Final success rate and basic reasoning evaluation	Dynamic complex reasoning
	GAIA [86]	Final success rate and basic reasoning evaluation	Closed-rule and dynamic complex reasoning

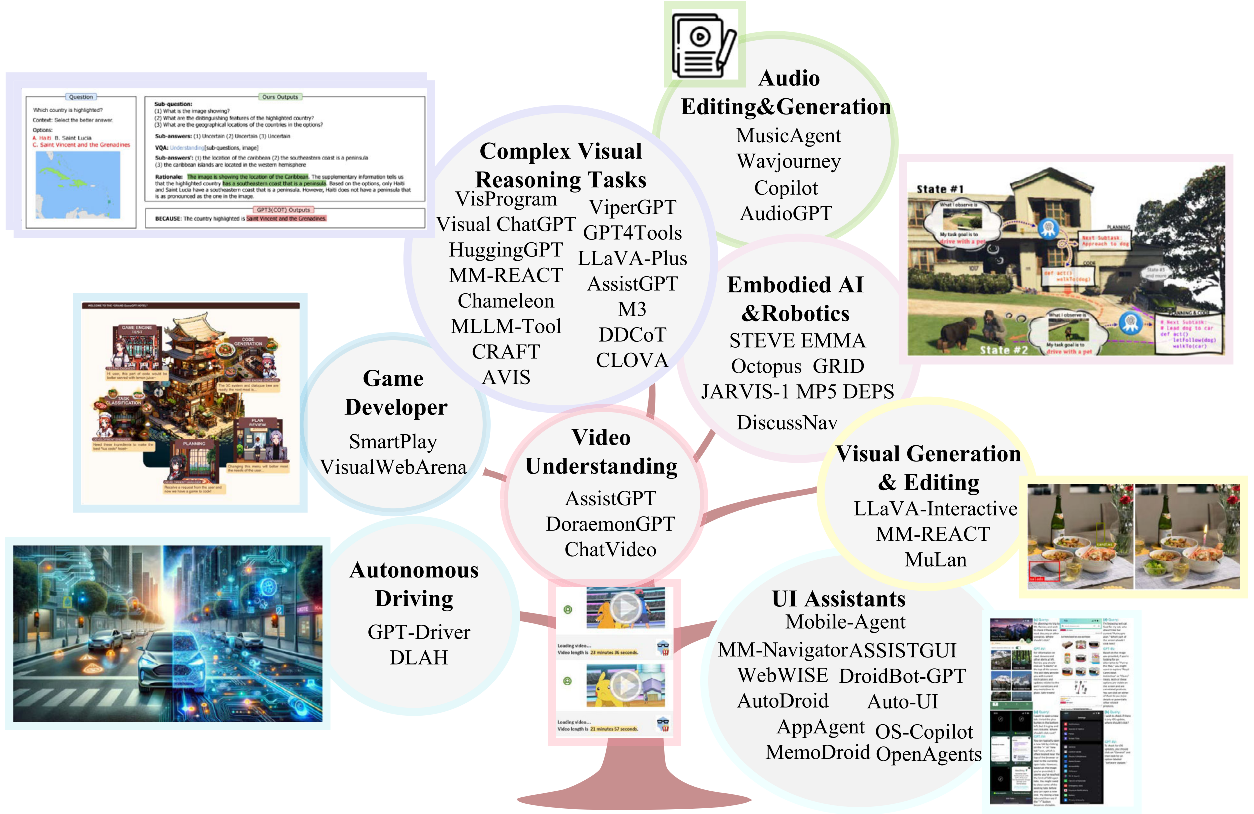


Figure 4 A variety of applications of LMAs. These applications include MusicAgent [27], Wavjourney [69], Copilot [66], AudioGPT [14], ViperGPT [17], Visual ChatGPT [18], GPT4Tools [19], HuggingGPT [20], LLaVA-Plus [37], MM-REACT [15], AssistGPT [22], Chameleon [35], MLLM-Tool [47], M3 [26], DDCoT [33], CRAFT [59], CLOVA [58], AVIS [23], STEVE [43], EMMA [42], Octopus [29], GRID [31], JARVIS-1 [45], MP5 [40], DEPS [30], DiscussNav [25], SmartPlay [85], VisualWebArena [83], AssistGPT [22], DoraemonGPT [65], LLaVA-Interactive [39], MuLan [63], GPT-Driver [36], DLAH [48], Mobile-Agent [53], MM-Navigator [38], ASSISTGUI [62], WebWISE [32], DroidBot-GPT [46], AutoDroid [21], Auto-UI [24], AppAgent [44], OS-Copilot [66], MemoDroid [41], and OpenAgents [28]

GUI automation In this application, the objective of LMAs is to understand and simulate human actions within user interfaces, enabling the execution of repetitive tasks, navigation across multiple applications, and the simplification of complex workflows. This automation holds the potential to save users' time and energy, allowing them to focus on the more critical and creative aspects of their work [32, 38, 41, 44, 46, 55, 62, 64, 84, 91, 94–97]. For example, GPT-4V-Act, is an advanced AI that combines GPT-4V's capabilities with web browsing to improve human-computer interactions. Its main goal is to make user interfaces more accessible, simplify workflow automation, and enhance automated UI testing. This AI is especially beneficial for people with disabilities or limited tech skills, helping them navigate complex interfaces more easily.

Robotics and embodied AI This application [29, 31, 40, 42, 43, 45, 51, 54, 56, 61, 71, 98–100] focuses on integrating the perceptual, reasoning, and action capabilities of robots with physical interactions in their environments. Employing a multimodal agent, robots are enabled to utilize diverse sensory channels, such as vision, hearing, and touch, to acquire comprehensive environmental data. For example, the MP5 system [40] is a cutting-edge, multimodal entity system used in Minecraft that utilizes active perception to smartly break down and carry out extensive, indefinite tasks with large language models.

Game developement Game AI [83, 85] endeavors to design and implement these agents to exhibit intelligence and realism, thereby providing engaging and challenging player experiences. The successful integration of agent technology in games has led to the creation of more sophisticated and interactive virtual environments.

Autonomous driving Traditional approaches to autonomous vehicles [21] face obstacles in effectively perceiving and interpreting complex scenarios. Recent progress in multimodal agent-based technologies, notably driven by LLMs, marks a substantial advancement in overcoming these challenges and bridging the perception gap [36, 48, 101, 102]. Ref. [36] presents GPT-Driver, a pioneering approach that employs the OpenAI GPT-3.5 model as a reliable motion planner for autonomous vehicles, with a specific focus on generating safe and comfortable driving trajectories. Harnessing the inherent reasoning capabilities of LLMs, their method provides a promising solution to the issue of limited generalization in novel driving scenarios.

Video understanding The video understanding agents [22, 65, 103–105] are artificial intelligence systems specifically designed for analyzing and comprehending video content. They utilize deep learning techniques to extract essential information from videos, identifying objects, actions, and scenes to enhance understanding of the video content.

Visual generation & editing Some applications[15, 16, 39] are designed for the creation and manipulation of visual content. Using advanced technologies, these tools effortlessly create and modify images, offering users a flexible option for creative projects. For instance, LLaVA-Interactive [39] is an open-source multimodal interactive system that amalgamates the capabilities of pre-trained AI models to facilitate multi-turn dialogues with visual cues and generate edited images, thereby realizing a cost-effective, flexible, and intuitive AI-assisted visual content creation experience. MovieAgent [106] introduces an automated movie generation framework using multi-agent chain of thought (CoT) planning. It autonomously structures scenes, cinematography, and character interactions, ensuring script fidelity, character consistency, and narrative coherence. By simulating roles like director and screenwriter, it streamlines production, achieving state-of-the-art results in automated long-form video generation. MM-StoryAgent [107] proposes an open-source multi-agent framework leveraging LLMs and generative tools to create immersive narrated video storybooks, enhancing story attractiveness and expressiveness through refined plots, role-consistent visuals, and multi-channel audio, validated by objective and subjective evaluations. MotionAgent [108] introduces a motion field agent to enable fine-grained motion control in text-guided image-to-video generation by converting text-based motion descriptions into explicit motion fields, achieving precise control over object and camera motion.

Complex visual reasoning tasks This area is a key focus in multimodal agent research, mainly emphasizing the analysis of multimodal content. This prevalence is attributed to the superior cognitive capabilities of LLMs in comprehending and reasoning through knowledge-based queries, surpassing the capabilities of previous models [26, 33, 109–112]. Within these applications, the primary focus is on QA tasks [15, 18, 20, 35]. This entails leveraging visual modalities (images or videos) and textual modalities (questions or questions with accompanying documents) to generate reasoned responses.

Audio editing & generation The LMAs in this application integrate foundational expert models in the audio domain, making music editing and creation efficient [14, 27, 68, 69].

7 Limitation

Computational costs The computational overhead of LMAs predominantly stems from three critical operational dimensions. In the realm of cross-modal data processing, the visual information pipeline proves particularly resource-intensive, as exemplified by conventional ViT encoders requiring the generation of over 100,000 visual tokens when processing 4K-resolution imagery, resulting in a substantial 18 GB memory footprint, while empirical studies demonstrate that merely 12% of these tokens typically contain semantically meaningful information [113]. The compounded computational burden of tool orchestration manifests most acutely in complex workflows, with a standard multimodal report generation task – involving sequential invocations of image synthesis, textual refinement, and audio proofing modules – routinely consuming in excess of 100,000 tokens per operation [114]. Perhaps most critically, the memory management challenges inherent to dynamic operational environments present extraordinary scaling difficulties, as evidenced by operating system agents sustaining 48 GB memory utilization across merely 100 operational steps due to persistent caching requirements for both screen captures and action histories [24]. These computational bottlenecks yield tangible performance constraints in applied settings, with even state-of-the-art implementations exceeding the 15 W power envelope by 37% in clinical robotic applications [57].

Ethical concerns Contemporary research reveals three principal ethical challenges in LMAs systems. The primary concern involves cross-modal alignment and value consistency, where vision-language architectures exhibit 68% semantic disjunction risk during multimodal processing [93]. Particularly concerning are healthcare applications where LMAs generate clinically coherent yet ethically non-compliant outputs, such as end-of-life recommendations violating patient autonomy principles [115]. The second challenge manifests in systematic bias amplification, with multimodal systems demonstrating 37% reduced STEM field recommendations for women when processing gender-stereotyped inputs [116]. Diagnostic accuracy disparities further exacerbate, showing 23% higher error rates for minority populations compared to unimodal systems. Most critically, the simulation-to-reality transition presents compound challenges: 67% of deployed LMAs exhibit behavioral deviation from training objectives [117], while continual learning architectures render 98% of ethical violations untraceable [118].

Scalability issues The scalability of LMAs faces several fundamental limitations. First, the combinatorial explosion in task planning complexity leads to deteriorating planning efficiency as task scenarios expand [119]. Second, the inherent difficulty in aligning multimodal perception (e.g., vision, language) with physical actions poses significant challenges for scalable learning [42]. Finally, substantial system integration overhead arises from the need for extensive engineering efforts to coordinate perception, planning, and execution modules [120].

8 Conclusion and future research

In this survey, we provide a thorough overview of the latest research on multimodal agents driven by LLMs (LMAs). We start by introducing the core components of LMAs (i.e., perception, planning, action, and memory) and classify existing studies into four categories. Subsequently, we compile existing methodologies for evaluating LMAs and devise a comprehensive evaluation framework. Finally, we spotlight a range of current and significant application scenarios within the realm of LMAs. Despite the notable progress, this field still faces many unresolved challenges, and there is considerable room for improvement. We finally highlight several promising directions based on the reviewed progress:

1) On frameworks: The future frameworks of LMAs may evolve from two distinct perspectives. From the perspective of an individual agent, progress should be made towards creating a more *unified* system. This entails planners directly interacting with multimodal environments [65], utilizing a comprehensive set of tools [35], and manipulating memory directly [45]. From the perspective of multiple agents, advancing the effective coordination among multiple multimodal agents to execute collective tasks emerges as a critical research trajectory. This encompasses essential aspects such as collaborative mechanisms, communication protocols, and strategic task distribution.
2) On evaluation: Systematic and standard evaluation frameworks are highly desired for this field. An ideal evaluation framework should encompass a spectrum of assessment tasks [83, 85], varying from straightforward to intricate, each bearing significant relevance and utility for humans. It ought to incorporate lucid and judicious evaluation metrics, meticulously designed to evaluate the diverse capabilities of an LMA in a comprehensive, yet non-repetitive manner. Moreover, the dataset used for evaluation should be meticulously curated to reflect a closer resemblance to real-world scenarios.
3) On application: The potential applications of LMAs in the real world are substantial, offering solutions to problems that were previously challenging for conventional models, such as web browsing. Furthermore, the intersection of LMAs with the field of human-computer interaction [32, 46] represents one of the significant directions for future applications. Their ability to process and understand information from various modalities enables them to perform more complex and nuanced tasks, thereby enhancing their utility in real-world scenarios and improving the interaction between humans and machines.

Abbreviations

CoT, chain of thought; LMAs, large multimodal agents; VFMs, visual foundation models.

Author contributions

The conception of this review was contributed by JX, ZC and RZ. The literature search and initial manuscript drafting were performed by JX. All authors contributed to manuscript revision and approved the submitted version.

Funding information

This work is supported by the National Key R&D Program of China (Nos.2024YFB3908503 and 2024YFB3908500), and the National Natural Science Foundation of China (No. 62322608). This work is also sponsored by CIE-Tencent Robotics X Rhino-Bird Focused Research Program.

Data availability

An up-to-date resource list is available at https://github.com/jun0wanan/awesome-large-multimodal-agents.

Declarations

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

Author details

[1]Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Guangdong, 518172, China. [2]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, 510275, China.

Received: 27 April 2025 Revised: 12 October 2025
Accepted: 16 October 2025 Published online: 26 November 2025

References

1. Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: theory and practice. *Knowledge Engineering Review*, *10*(2), 115–152.
2. Osoba, O.A., Vardavas, R., Grana, J., Zutshi, R., & Jaycocks, A. (2020). Policy-focused agent-based modeling using RL behavioral models. arXiv preprint. arXiv:2006.05048.
3. Wang, X., & Su, H. (2020). Completely model-free RL-based consensus of continuous-time multi-agent systems. *Applied Mathematics and Computation*, *382*, 125312.
4. Pan, J. Z., Razniewski, S., Kalo, J.-C., Singhania, S., Chen, J., Dietze, S., Jabeen, H., Omeliyanenko, J., Zhang, W., Lissandrini, M., et al. (2023). Large language models and knowledge graphs: opportunities and challenges. *Transactions on Graph Data and Knowledge*, *1*(1), 1–38.
5. Zhang, Z., Fang, M., Chen, L., Namazi-Rad, M.-R., & Wang, J. (2023). How do large language models capture the ever-changing world knowledge? A review of recent advances. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 8289–8311). Stroudsburg: ACL.
6. Li, Y. A., Han, C., Raghavan, V., Mischler, G., & Mesgarani, N. (2023). Styletts 2: towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 37th international conference on neural information processing systems* (pp. 19594–19621). Red Hook: Curran Associates.
7. Yang, L., Zhang, S., Yu, Z., Bao, G., Wang, Y., Wang, J., Xu, R., Ye, W., Xie, X., Chen, W., et al. (2024). Supervised knowledge makes large language models better in-context learners. In *Proceedings of the 12th international conference on learning representations* (pp. 1–17). Retrieved August 7, 2025, from https://openreview.net/forum?id=bAMPOUF227.
8. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: language models can teach themselves to use tools. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 37th international conference on neural information processing systems* (pp. 68539–68551). Red Hook: Curran Associates.
9. Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., et al. (2024). Toolllm: facilitating large language models to master 16000+ real-world APIs. In *Proceedings of the 12th international conference on learning representations* (pp. 1–23). Retrieved August 7, 2025, from https://openreview.net/pdf?id=dHng2O0Jjr.
10. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al. (2025). The rise and potential of large language model based agents: a survey. *Science China. Information Sciences*, *68*(2), 121101.
11. Sumers, T., Yao, S., Narasimhan, K., & Griffiths, T. (2024). Cognitive architectures for language agents. *Transactions on Machine Learning Research*, *2*, 1–32.
12. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z.-Y., Tang, J., Chen, X., Lin, Y., et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*, 186345.
13. Gupta, T., & Kembhavi, A. (2023). Visual programming: compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14953–14962). Piscataway: IEEE.
14. Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J., Liu, J., et al. (2024). AudioGPT: understanding and generating speech, music, sound, and talking head. In *Proceedings of the 38th AAAI conference on artificial intelligence and 36th conference on innovative applications of artificial intelligence and 14th symposium on educational advances in artificial intelligence* (pp. 23802–23804). Palo Alto: AAAI Press.
15. Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., & Wang, L. (2023). MM-REACT: prompting ChatGPT for multimodal reasoning and action. arXiv preprint. arXiv:2303.11381.
16. Wang, J., Chen, D., Luo, C., Dai, X., Yuan, L., Wu, Z., & Jiang, Y.-G. (2023). ChatVideo: a tracklet-centric multimodal and versatile video understanding system. arXiv preprint. arXiv:2304.14407.
17. Surís, D., Menon, S., & Vondrick, C. (2023). ViperGPT: visual inference via Python execution for reasoning. In *Proceedings of the 2023 IEEE/CVF international conference on computer vision* (pp. 11854–11864). Piscataway: IEEE.
18. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., & Duan, N. (2023). Visual ChatGPT: talking, drawing and editing with visual foundation models. arXiv preprint. arXiv:2303.04671.
19. Yang, R., Song, L., Li, Y., Zhao, S., Ge, Y., Li, X., & Shan, Y. (2023). GPT4Tools: teaching large language model to use tools via self-instruction. In *Proceedings of the 37th international conference on neural information processing systems* (pp. 71995–72007). Red Hook: Curran Associates.
20. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., & Zhuang, Y. (2023). HuggingGPT: solving AI tasks with ChatGPT and its friends in huggingface. arXiv preprint. arXiv:2303.17580.
21. Maurer, M., Gerdes, J.C., Lenz, B., & Winner, H. (2016). *Autonomous driving: technical, legal and social aspects*. Berlin: Springer.
22. Gao, D., Ji, L., Zhou, L., Lin, K.Q., Chen, J., Fan, Z., & Shou, M.Z. (2023). AssistGPT: a general multi-modal assistant that can plan, execute, inspect, and learn. arXiv preprint. arXiv:2306.08640.
23. Hu, Z., Iscen, A., Sun, C., Chang, K.-W., Sun, Y., Ross, D., Schmid, C., & Fathi, A. (2023). AVIS: autonomous visual information seeking with large language model agent. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 37th international conference on neural information processing systems* (pp. 867–878). Red Hook: Curran Associates.
24. Cai, D., Wang, S., Peng, C., Zhang, Z., Lu, Z., Qi, T., Lane, N. D., & Xu, M. (2025). Ubiquitous memory augmentation via mobile multimodal embedding system. *Nature Communications*, *16*(1), 1–12.
25. Long, Y., Li, X., Cai, W., & Dong, H. (2024). Discuss before moving: visual language navigation via multi-expert discussions. In *Proceedings of the IEEE international conference on robotics and automation* (pp. 17380–17387). Piscataway: IEEE.
26. Liu, X., Li, R., Ji, W., & Lin, T. (2024). Towards robust multi-modal reasoning via model selection. In *Proceedings of the 12th international conference on learning representations* (pp. 1–22). Retrieved August 7, 2025, from https://openreview.net/forum?id=KTf4DGAzus.
27. Yu, D., Song, K., Lu, P., He, T., Tan, X., Ye, W., Zhang, S., & Bian, J. (2023). MusicAgent: an AI agent for music understanding and generation with large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations* (pp. 246–255). Stroudsburg: ACL.
28. Xie, T., Zhou, F., Cheng, Z., Shi, P., Weng, L., Liu, Y., Hua, T.J., Zhao, J., Liu, Q., Liu, C., et al. (2023). OpenAgents: an open platform for language agents in the wild. arXiv preprint. arXiv:2310.10634.
29. Yang, J., Dong, Y., Liu, S., Li, B., Wang, Z., Tan, H., Jiang, C., Kang, J., Zhang, Y., Zhou, K., et al. (2024). Octopus: embodied vision-language programmer from environmental feedback. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Proceedings of the 18th European conference on computer vision* (pp. 20–38). Cham: Springer.

30. Tang, W., Zhou, Y., Xu, E., Cheng, K., Li, M., & Xiao, L. (2025). DSGBench: a diverse strategic game benchmark for evaluating LLM-based agents in complex decision-making environments. arXiv preprint. arXiv:2503.06047.
31. Vemprala, S., Chen, S., Shukla, A., Narayanan, D., & Kapoor, A. (2023). GRID: a platform for general robot intelligence development. arXiv preprint. arXiv:2310.00887.
32. Tao, H., Sethuraman, T. V., Shlapentokh-Rothman, M., Hoiem, D., & Ji, H. (2023). WebWISE: web interface control and sequential exploration with large language models. arXiv preprint. arXiv:2310.16042.
33. Zheng, G., Yang, B., Tang, J., Zhou, H.-Y., & Yang, S. (2023). DDCoT: duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 37th international conference on neural information processing systems* (pp. 5168–5191). Red Hook: Curran Associates.
34. Liu, Z., Lai, Z., Gao, Z., Cui, E., Li, Z., Zhu, X., Lu, L., Chen, Q., Qiao, Y., Dai, J., et al. (2024). ControlLLM: augment language models with tools by searching on graphs. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Proceedings of the 18th European conference on computer vision* (pp. 89–105). Cham: Springer.
35. Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.-W., Wu, Y.N., Zhu, S.-C., & Gao, J. (2023). Chameleon: plug-and-play compositional reasoning with large language models. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 37th international conference on neural information processing systems* (pp. 43447–43478). Red Hook: Curran Associates.
36. Mao, J., Qian, Y., Zhao, H., & Wang, Y. (2023). GPT-Driver: learning to drive with GPT. arXiv preprint. arXiv:2310.01415.
37. Liu, S., Cheng, H., Liu, H., Zhang, H., Li, F., Ren, T., Zou, X., Yang, J., Su, H., Zhu, J., et al. (2024). LLaVA-Plus: learning to use tools for creating multimodal agents. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Proceedings of the 18th European conference on computer vision* (pp. 126–142). Cham: Springer.
38. Yan, A., Yang, Z., Zhu, W., Lin, K., Li, L., Wang, J., Yang, J., Zhong, Y., McAuley, J., Gao, J., et al. (2023). GPT-4v in wonderland: large multimodal models for zero-shot smartphone GUI navigation. arXiv preprint. arXiv:2311.07562.
39. Chen, W.-G., Spiridonova, I., Yang, J., Gao, J., & Li, C. (2023). LLaVA-Interactive: an all-in-one demo for image chat, segmentation, generation and editing. arXiv preprint. arXiv:2311.00571.
40. Qin, Y., Zhou, E., Liu, Q., Yin, Z., Sheng, L., Zhang, R., Qiao, Y., & Shao, J. (2024). MP5: a multi-modal open-ended embodied system in minecraft via active perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16307–16316). Piscataway: IEEE.
41. Lee, S., Choi, J., Lee, J., Choi, H., Ko, S. Y., Oh, S., & Shin, I. (2023). Explore, select, derive, and recall: augmenting LLM with human-like memory for mobile task automation. arXiv preprint. arXiv:2312.03003.
42. Yang, Y., Zhou, T., Li, K., Tao, D., Li, L., Shen, L., He, X., Jiang, J., & Shi, Y. (2024). Embodied multi-modal agent trained by an LLM from a parallel textworld. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 26275–26285). Piscataway: IEEE.
43. Zhao, Z., Chai, W., Wang, X., Li, B., Hao, S., Cao, S., Ye, T., & Wang, G. (2024). See and think: embodied agent in virtual environment. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Proceedings of the 18th European conference on computer vision* (pp. 187–204). Cham: Springer.
44. Zhang, C., Yang, Z., Liu, J., Li, Y., Han, Y., Chen, X., Huang, Z., Fu, B., & Yu, G. (2025). AppAgent: multimodal agents as smartphone users. In *Proceedings of the 2025 CHI conference on human factors in computing systems* (pp. 1–20). New York: ACM.
45. Wang, Z., Cai, S., Liu, A., Jin, Y., Hou, J., Zhang, B., Lin, H., He, Z., Zheng, Z., Yang, Y., et al. (2025). JARVIS-1: open-world multi-task agents with memory-augmented multimodal language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(3), 1894–1907.
46. Wen, H., Wang, H., Liu, J., & Li, Y. (2023). DroidBot-GPT: GPT-powered UI automation for Android. arXiv preprint. arXiv:2304.07061.
47. Wang, C., Luo, W., Dong, S., Xuan, X., Li, Z., Ma, L., & Gao, S. (2025). MLLM-Tool: a multimodal large language model for tool agent learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 6678–6687). Piscataway: IEEE.
48. Fu, D., Li, X., Wen, L., Dou, M., Cai, P., Shi, B., & Qiao, Y. (2024). Drive like a human: rethinking autonomous driving with large language models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops* (pp. 910–919). Piscataway: IEEE.
49. Wang, Z., Li, A., Li, Z., & Liu, X. (2024). GenArtist: multimodal LLM as an agent for unified image generation and editing. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, & C. Zhang (Eds.), *Proceedings of the 38th international conference on neural information processing systems* (pp. 128374–128395). Red Hook: Curran Associates.
50. Gou, B., Wang, R., Zheng, B., Xie, Y., Chang, C., Shu, Y., Sun, H., & Su, Y. (2024). Navigating the digital world as humans do: universal visual grounding for GUI agents. In *Proceedings of the 13th international conference on learning representations* (pp. 1–33). Retrieved September 5, 2025, from https://openreview.net/forum?id=kxnoqaisCT.
51. Mazzaglia, P., Verbelen, T., Dhoedt, B., Courville, A. C., & Mudumba, S. R. (2024). GenRL: multimodal-foundation world models for generalization in embodied agents. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, & C. Zhang (Eds.), *Proceedings of the 38th international conference on neural information processing systems* (pp. 27529–27555). Red Hook: Curran Associates.
52. Wu, S., Zhao, S., Huang, Q., Huang, K., Yasunaga, M., Cao, K., Ioannidis, V., Subbian, K., Leskovec, J., & Zou, J. Y. (2024). Avatar: optimizing LLM agents for tool usage via contrastive reasoning. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, & C. Zhang (Eds.), *Proceedings of the 38th international conference on neural information processing systems* (pp. 25981–26010). Red Hook: Curran Associates.
53. Wang, J., Xu, H., Ye, J., Yan, M., Shen, W., Zhang, J., Huang, F., & Sang, J. (2024). Mobile-Agent: autonomous multi-modal mobile device agent with visual perception. arXiv preprint. arXiv:2401.16158.
54. Li, Z., Xie, Y., Shao, R., Chen, G., Jiang, D., & Nie, L. (2025). Optimus-2: multimodal minecraft agent with goal-observation-action conditioned policy. In *Proceedings of the computer vision and pattern recognition conference* (pp. 9039–9049). Piscataway: IEEE.
55. Sun, Y., Zhao, S., Yu, T., Wen, H., Va, S., Xu, M., Li, Y., & Zhang, C. (2025). GUI-Xplore: empowering generalizable GUI agents with one exploration. In *Proceedings of the computer vision and pattern recognition conference* (pp. 19477–19486). Piscataway: IEEE.
56. Zhang, Z., Zhu, L., Fang, Z., Huang, Z., & Luo, Y. (2025). Provable ordering and continuity in vision-language pretraining for generalizable embodied agents. arXiv preprint. arXiv:2502.01218.
57. Zhou, Y., Song, L., & Mam, J. S. (2025). Modular multi-agent framework for multi-modal medical diagnosis via role-specialized collaboration. arXiv preprint. arXiv:2506.19835.
58. Gao, Z., Du, Y., Zhang, X., Ma, X., Han, W., Zhu, S.-C., & Li, Q. (2024). CLOVA: a closed-loop visual assistant with tool usage and update. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13258–13268). Piscataway: IEEE.
59. Yuan, L., Chen, Y., Wang, X., Fung, Y. R., Peng, H., & Ji, H. (2024). CRAFT: customizing LLMs by creating and retrieving from specialized toolsets. In *Proceedings of the 12th international conference on learning representations* (pp. 1–29). Retrieved August 7, 2025, from https://openreview.net/forum?id=G0vdDSt9XM.
60. Chen, L., Li, B., Shen, S., Yang, J., Li, C., Keutzer, K., Darrell, T., & Liu, Z. (2023). Large language models are visual reasoning coordinators. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 37th international conference on neural information processing systems* (pp. 70115–70140). Red Hook: Curran Associates.
61. Wang, Z., Cai, S., Liu, A., Ma, X., & Liang, Y. (2023). Describe, explain, plan and select: interactive planning with large language models enables open-world multi-task agents. arXiv preprint. arXiv:2302.01560.
62. Gao, D., Ji, L., Bai, Z., Ouyang, M., Li, P., Mao, D., Wu, Q., Zhang, W., Wang, P., Guo, X., et al. (2023). ASSISTGUI: task-oriented desktop graphical user interface automation. arXiv preprint. arXiv:2312.13108.
63. Li, S., Wang, R., Hsieh, C.-J., Cheng, M., & Zhou, T. (2024). MuLan: multimodal-LLM agent for progressive multi-object diffusion. arXiv preprint. arXiv:2402.12741.
64. Zhang, Z., & Zhang, A. (2024). You only look at screens: multimodal chain-of-action agents. In *Findings of the Association for Computational Linguistics* (pp. 3132–3149). Stroudsburg: ACL.
65. Yang, Z., Chen, G., Li, X., Wang, W., & Yang, Y. (2024). DoraemonGPT: toward understanding dynamic scenes with large language models (exemplified as a video agent). In *Proceedings of the international*

conference on machine learning (pp. 55976–55997). Retrieved August 8, 2025, from https://openreview.net/forum?id=QMy2RLnxGN.

66. Wu, Z., Han, C., Ding, Z., Weng, Z., Liu, Z., Yao, S., Yu, T., & Kong, L. (2024). OS-Copilot: towards generalist computer agents with self-improvement. arXiv preprint. arXiv:2402.07456.
67. Liu, Y., Song, X., Jiang, K., Chen, W., Luo, J., Li, G., & Lin, L. (2024). Multimodal embodied interactive agent for cafe scene. arXiv preprint. arXiv:2402.00290.
68. Zhang, Y., Maezawa, A., Xia, G., Yamamoto, K., & Dixon, S. (2023). Loop copilot: conducting AI ensembles for music generation and iterative editing. arXiv preprint. arXiv:2310.12404.
69. Liu, X., Zhu, Z., Liu, H., Yuan, Y., Cui, M., Huang, Q., Liang, J., Cao, Y., Kong, Q., Plumbley, M. D., et al. (2023). Wavjourney: compositional audio creation with large language models. arXiv preprint. arXiv:2307.14335.
70. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the international conference on machine learning* (pp. 19730–19742). Retrieved August 7, 2025, from https://proceedings.mlr.press/v202/li23q.html.
71. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al. (2024). Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Proceedings of the 18th European conference on computer vision* (pp. 38–55). Cham: Springer.
72. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the international conference on machine learning* (pp. 12888–12900). Retrieved August 7, 2025, from https://proceedings.mlr.press/v162/li22n.html.
73. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., & Hoi, S. C. H. (2023). InstructBLIP: towards general-purpose vision-language models with instruction tuning. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 37th international conference on neural information processing systems* (pp. 1–18). Red Hook: Curran Associates.
74. Xu, J., Wang, X., Cao, Y.-P., Cheng, W., Shan, Y., & Gao, S. (2023). InstructP2P: learning to edit 3D point clouds with text instructions. arXiv preprint. arXiv:2306.07154.
75. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684–10695). Piscataway: IEEE.
76. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. Retrieved August 7, 2025, from https://cdn.openai.com/papers/dall-e-3.pdf.
77. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015–4026). Piscataway: IEEE.
78. Liu, Y., Chu, L., Chen, G., Wu, Z., Chen, Z., Lai, B., & Hao, Y. (2021). PaddleSeg: a high-efficient development toolkit for image segmentation. arXiv preprint. arXiv:2101.06175.
79. Ye, B., Chang, H., Ma, B., Shan, S., & Chen, X. (2022). Joint feature learning and relation modeling for tracking: a one-stream framework. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Proceedings of the 17th European conference on computer vision* (pp. 341–357). Cham: Springer.
80. Cao, N., Lin, Y.-R., Sun, X., Lazer, D., Liu, S., & Qu, H. (2012). Whisper: tracing the spatiotemporal process of information diffusion in real time. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2649–2658.
81. Zhang, J., Zhang, J., Pertsch, K., Liu, Z., Ren, X., Chang, M., Sun, S.H., & Lim, J. J. (2023). Bootstrap your own skills: learning to solve new tasks with large language model guidance. In *Proceedings of the conference on robot learning* (pp. 302–325). Retrieved August 7, 2025, from https://proceedings.mlr.press/v229/zhang23a.html.
82. He, J.-Y., Cheng, Z.-Q., Li, C., Sun, J., He, Q., Xiang, W., Chen, H., Lan, J.-P., Lin, X., Zhu, K., et al. (2025). MetaDesigner: advancing artistic typography through AI-driven, user-centric, and multilingual wordart synthesis. In *Proceedings of the 13th international conference on learning representations* (pp. 1–24). Retrieved August 7, 2025, from https://openreview.net/forum?id=Mv3GAYJGcW.
83. Koh, J. Y., Lo, R., Jang, L., Duvvur, V., Lim, M., Huang, P.-Y., Neubig, G., Zhou, S., Salakhutdinov, R., & Fried, D. (2024). VisualWebArena: evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics* (pp. 881–905). Stroudsburg: ACL.
84. Zheng, L., Huang, Z., Xue, Z., Wang, X., An, B., & Yan, S. (2024). Agentstudio: a toolkit for building general virtual agents. In *Proceedings of the 13th international conference on learning representations* (pp. 1–42). Retrieved August 7, 2025, from https://openreview.net/forum?id=axUf8BOjnH.
85. Wu, Y., Tang, X., Mitchell, T. M., & Li, Y. (2024). SmartPlay: a benchmark for LLMs as intelligent agents. In *Proceedings of the 12th international conference on learning representations* (pp. 1–19). Retrieved August 7, 2025, from https://openreview.net/forum?id=S2oTVrlcp3.
86. Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., & Scialom, T. (2024). GAIA: a benchmark for general AI assistants. In *Proceedings of the 12th international conference on learning representations* (pp. 1–25). Retrieved August 7, 2025, from https://openreview.net/forum?id=fibxvahvs3.
87. Lù, X.H., Kasner, Z., & Reddy, S. (2024). Weblinx: real-world website navigation with multi-turn dialogue. arXiv preprint. arXiv:2402.05930.
88. Xie, J., Zhang, K., Chen, J., Zhu, T., Lou, R., Tian, Y., Xiao, Y., & Su, Y. (2024). Travelplanner: a benchmark for real-world planning with language agents. In *Proceedings of the 41st international conference on machine learning* (pp. 1–24). Retrieved August 7, 2025, from https://openreview.net/forum?id=l5XQzNkAOe.
89. Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., & Su, Y. (2023). Mind2web: towards a generalist agent for the web. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 37th international conference on neural information processing systems* (pp. 28091–28114). Red Hook: Curran Associates.
90. Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., et al. (2024). Webarena: a realistic web environment for building autonomous agents. In *Proceedings of the 12th international conference on learning representations*. Retrieved August 7, 2025, from https://openreview.net/forum?id=oKn9c6ytLx.
91. Chen, J., Yuen, D., Xie, B., Yang, Y., Chen, G., Wu, Z., Yixing, L., Zhou, X., Liu, W., Wang, S., et al. (2024). Spa-Bench: a comprehensive benchmark for smartphone agent evaluation. In *Proceedings of the 13th international conference on learning representations* (pp. 1–38). Retrieved August 7, 2025, from https://openreview.net/forum?id=OZbFRNhpwr.
92. Paglieri, D., Cupiał, B., Coward, S., Piterbarg, U., Wolczyk, M., Khan, A., Pignatelli, E., Kuciński, Ł., Pinto, L., Fergus, R., et al. (2024). BALROG: benchmarking agentic LLM and VLM reasoning on games. arXiv preprint. arXiv:2411.13543.
93. Yang, J., Shao, S., Liu, D., & Shao, J. (2025). RiOSWorld: benchmarking the risk of multimodal compter-use agents. arXiv preprint. arXiv:2506.00618.
94. Wen, H., Li, Y., Liu, G., Zhao, S., Yu, T., Li, T. J.-J., Jiang, S., Liu, Y., Zhang, Y., & Liu, Y. (2023). Empowering LLM to use smartphone for intelligent task automation. arXiv preprint. arXiv:2308.15272.
95. Huq, F., Wang, Z. Z., Xu, F. F., Ou, T., Zhou, S., Bigham, J. P., & Neubig, G. (2025). Cowpilot: a framework for autonomous and human-agent collaborative web navigation. arXiv preprint. arXiv:2501.16609.
96. Wang, J., Xu, H., Zhang, X., Yan, M., Zhang, J., Huang, F., & Sang, J. (2025). Mobile-Agent-V: learning mobile device operation through video-guided multi-agent collaboration. arXiv preprint. arXiv:2502.17110.
97. Huang, Z., Cheng, Z., Pan, J., Hou, Z., & Zhan, M. (2025). Spiritsight agent: advanced GUI agent with one look. In *Proceedings of the computer vision and pattern recognition conference* (pp. 29490–29500). Piscataway: IEEE.
98. Kim, J., Kim, M.-S., Chung, J., Cho, J., Kim, J., Kim, S., Sim, G., & Yu, Y. (2025). Egospeak: learning when to speak for egocentric conversational agents in the wild. In *Findings of the Association for Computational Linguistics* (pp. 2990–3005). Stroudsburg: ACL.
99. Wang, H., Liu, P., Cai, W., Wu, M., Qian, Z., & Dong, H. (2024). MO-DDN: a coarse-to-fine attribute-based exploration agent for multi-object demand-driven navigation. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, & C. Zhang (Eds.), *Proceedings of the 38th international conference on neural information processing systems* (pp. 1–39). Red Hook: Curran Associates.
100. Matthews, M. T., Beukman, M., Lu, C., & Foerster, J. N. (2024). Kinetix: investigating the training of general agents through open-ended physics-based control tasks. In *Proceedings of the 13th international conference on learning representations* (pp. 1–50). Retrieved August 7, 2025, from https://openreview.net/forum?id=zCxGCdzreM.2025.

101. Zhou, X., Liu, M., Zagar, B.L., Yurtsever, E., & Knoll, A. C. (2023). Vision language models in autonomous driving and intelligent transportation systems. arXiv preprint. arXiv:2310.14414.
102. Wen, L., Yang, X., Fu, D., Wang, X., Cai, P., Li, X., Ma, T., Li, Y., Xu, L., Shang, D., et al. (2023). On the road with GPT-4v (ision): early explorations of visual-language model on autonomous driving. arXiv preprint. arXiv:2311.05332.
103. Wei, K., Zhou, Z., Wang, B., Araki, J., Lange, L., Huang, r., & Feng, z. (2025). Premind: multi-agent video understanding for advanced indexing of presentation-style videos. arXiv preprint. arXiv:2503.00162.
104. Sun, G., Jin, M., Wang, Z., Wang, C.-L., Ma, S., Wang, Q., Geng, T., Wu, Y.N., Zhang, Y., & Liu, D. (2025). Visual agents as fast and slow thinkers. In *Proceedings of the 13th international conference on learning representations* (pp. 1–26). Retrieved August 7, 2025, from https://openreview.net/forum?id=ncCuiD3KJQ.
105. Fan, S., Guo, M.-H., & Yang, S. (2025). Agentic keyframe search for video question answering. arXiv preprint. arXiv:2503.16032.
106. Wu, W., Zhu, Z., & Mike, S.Z. (2025). Automated movie generation via multi-agent cot planning. arXiv preprint. arXiv:2503.07314.
107. Xu, X., Mei, J., Li, C., Wu, Y., Yan, M., Lai, S., Zhang, J., & Wu, M. (2025). MM-StoryAgent: immersive narrated storybook video generation with a multi-agent paradigm across text, image and audio. arXiv preprint. arXiv:2503.05242.
108. Liao, X., Zeng, X., Wang, L., Yu, G., Lin, G., & Zhang, C. (2025). Motionagent: fine-grained controllable video generation via motion field agent. arXiv preprint. arXiv:2502.03207.
109. Karthik, S., Roth, K., Mancini, M., & Akata, Z. (2024). Vision-by-language for training-free compositional image retrieval. In *Proceedings of the 12th international conference on learning representations* (pp. 1–16). Retrieved August 7, 2025, from https://openreview.net/forum?id=EDPxCjXzSb.
110. Guo, Y., Zhuang, S., Li, K., Qiao, Y., & Wang, Y. (2024). Transagent: transfer vision-language foundation models with heterogeneous agent collaboration. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, & C. Zhang (Eds.), *Proceedings of the 38th international conference on neural information processing systems* (pp. 1–26). Red Hook: Curran Associates.
111. Gao, Z., Zhang, B., Li, P., Ma, X., Yuan, T., Fan, Y., Wu, Y., Jia, Y., Zhu, S.-C., & Li, Q. (2024). Multi-modal agent tuning: building a VLM-driven agent for efficient tool usage. arXiv preprint. arXiv:2412.15606.
112. Yan, Y., Wang, S., Huo, J., Yu, P. S., Hu, X., & Wen, Q. (2025). Mathagent: leveraging a mixture-of-math-agent framework for real-world multimodal mathematical error detection. arXiv preprint. arXiv:2503.18132.
113. Zhang, Z., Pham, P., Zhao, W., Wan, K., Li, Y.-J., Zhou, J., Miranda, D., Kale, A., & Xu, C. (2024). Treat visual tokens as text? But your MLLM only needs fewer efforts to see. arXiv preprint. arXiv:2410.06169.
114. Zhang, H., Guo, H., Guo, S., Cao, M., Huang, W., Liu, J., & Zhang, G. (2024). ING-VP: MLLMs cannot play easy vision-based games yet. arXiv preprint. arXiv:2410.06555.
115. Vandemeulebroucke, T. (2025). The ethics of artificial intelligence systems in healthcare and medicine: from a local to a global perspective, and back. *Pflügers Archiv-European Journal of Physiology*, *477*(4), 591–601.
116. Luo, Y., Shi, M., Khan, M.O., Afzal, M.M., Huang, H., Yuan, S., Tian, Y., Song, L., Kouhana, A., Elze, T., et al. (2024). Fairclip: harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12289–12301). Piscataway: IEEE.
117. Yang, J., Tan, R., Wu, Q., Zheng, R., Peng, B., Liang, Y., Gu, Y., Cai, M., Ye, S., Jang, J., et al. (2025). Magma: a foundation model for multimodal AI agents. In *Proceedings of the computer vision and pattern recognition conference* (pp. 14203–14214). Piscataway: IEEE.
118. Leturc, C., & Bonnet, G. (2024). Using n-ary multi-modal logics in argumentation frameworks to reason about ethics. *AI Communications*, *37*(3), 323–355.
119. Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., et al. (2025). Why do multi-agent LLM systems fail? arXiv preprint. arXiv:2503.13657.
120. Durante, Z., Huang, Q., Wake, N., Gong, R., Park, J. S., Sarkar, B., Taori, R., Noda, Y., Terzopoulos, D., Choi, Y., et al. (2024). Agent AI: surveying the horizons of multimodal interaction. arXiv preprint. arXiv:2401.03568.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.