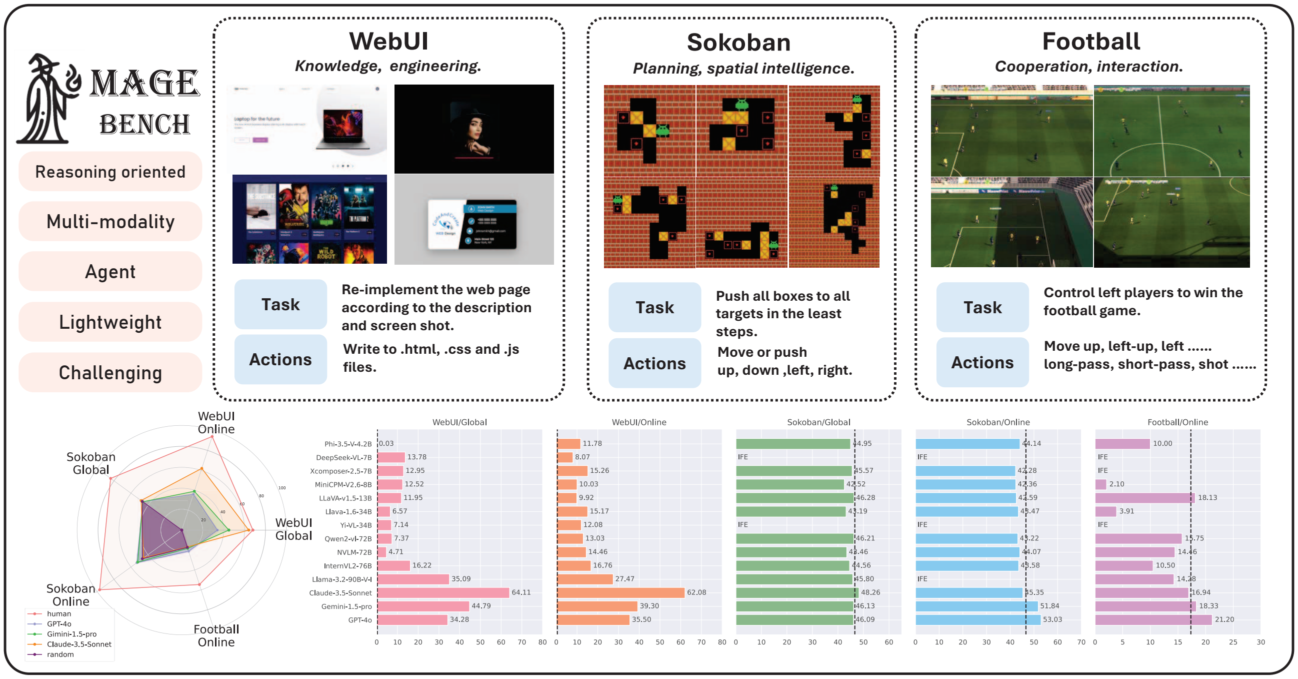
MageBench: Bridging Large Multimodal Models to Agents

Miaosen Zhang[1†] Qi Dai[2‡] Yifan Yang[2] Jianmin Bao[2] Dongdong Chen[2]
Kai Qiu[2] Chong Luo[2] Xin Geng[1‡] Baining Guo[1,2‡]

[1]Southeast University [2]Microsoft



Model	WebUI/Global	WebUI/Online	Sokoban/Global	Sokoban/Online	Football/Online
Phi-3.5-V-4.2B	0.03	11.78	44.95	44.14	10.00
DeepSeek-VL-7B	13.78	8.07	IFE	IFE	IFE
Xcomposer-2.5-7B	12.95	15.26	45.57	42.28	IFE
MiniCPM-V2.6-8B	12.52	10.03	42.52	42.36	2.10
LLaVA-v1.5-13B	11.95	9.92	46.28	42.59	18.13
Llava-1.6-34B	6.57	15.17	43.19	43.47	3.91
Yi-VL-34B	7.14	12.08	IFE	IFE	IFE
Qwen2-vl-72B	7.37	13.03	46.21	43.22	15.75
NVLM-72B	4.71	14.46	43.46	44.07	14.46
InternVL2-76B	16.22	16.76	44.56	43.58	10.50
Llama-3.2-90B-V-I	35.09	27.47	45.80	IFE	14.28
Claude-3.5-Sonnet	64.11	62.08	48.26	45.35	16.94
Gemini-1.5-pro	44.79	39.30	46.13	51.84	18.33
GPT-4o	34.28	35.50	46.09	53.03	21.20

Figure 1. Overview of the MageBench. MageBench is a multi-modality reasoning benchmark built upon lightweight agent environments. It currently contains three environments: WebUI, Sokoban, and Football. The results indicate that the existing models are still far from reaching human-level performance on agentic reasoning tasks. Only a few models outperform the results of random actions, represented by the black dashed line in the bar chart.

Abstract

*Recent models like OpenAI's O1 and DeepSeek's R1, which utilize test-time scaling techniques, have demonstrated remarkable improvements in reasoning capabilities. We anticipate that in the near future, multimodal models will also experience significant breakthroughs in multimodal reasoning. This will require some highly challenging and specialized evaluations. As one of the most crucial real-world applications of multimodal models, visual agents require complex and comprehensive capabilities such as spatial planning and vision-in-the-chain type reasoning. These capabilities are currently lacking in existing multimodal benchmarks. In this paper, we introduce **MageBench**, a **M**ultimodal reasoning benchmark built upon light-weight **AGE**nt environments that pose significant reasoning challenges and hold substantial practical value. The results show that only a few product-level models are better than random acting, and all of them are far inferior to human level. We analyze and summarize their errors and capability gaps in visual planning. Furthermore, we found that rule-based RL can significantly boost visual reasoning capabilities. This highlights that our benchmark could serve as a valuable testing ground for the emerging field of agentic RL research.*

[†]The work is completed during internship at Microsoft Research Asia.
[‡]Corresponding authors.

1. Introduction

The advent of Large Language Models (LLMs) [11, 17, 22, 84, 85, 118], and Large Multimodal Models (LMMs) [6, 8, 52, 83] has revolutionized the fields of natural language processing and computer vision. These models have demonstrated remarkable capabilities across a variety of classical tasks, including translation [64, 90, 102, 114], summarization [7, 68, 119], VQA [13, 27, 67, 75], captioning [86, 103], etc. The more recent OpenAI o1 [4] and DeepSeek's R1 [29] stand out due to its exceptional reasoning abilities, particularly on math and coding. The leap in reasoning capability of LLMs has paved the way for the development of LLM-based agents, which harness the power of these models to autonomously perform a range of sophisticated tasks.

Compared to the reasoning in LLMs, the reasoning and test time scaling in LMMs are more complex. This increased complexity arises because, in many real-world LMM applications (such as virtual agents and robotics), perception tasks and reasoning tasks are intertwined and influence each other. Therefore, we need a more frequent and complex interaction-based evaluation method for visual reasoning. Unfortunately, rare efforts have been made – existing benchmarks for LMMs mainly focus on the simple VQA problems [27, 38, 43, 55, 107, 109]; their reasoning assessment generally relies on the language part, which does not require interleaved involvement of visual signals [25, 58, 63, 115, 121, 123] .

In this work, we attempt to introduce more complex reasoning paradigms into the evaluation of visual reasoning. When defining 'complex reasoning paradigms', we expect not only the reasoning of initial visual input, but also the continuous understanding of visual feedback throughout the entire process. These tasks require models to dynamically interact with visual information, continually updating their understanding and decisions based on new visual cues, much like a human would. We refer to this reasoning paradigm concept, which integrates other modality (vision) into the reasoning chain, as **Vision-in-the-Chain (ViC)**, as illustrated in the last block of Figure 2. Note that, ViC is not a novel implementation or method as all LMM empowered agent solutions are leveraging LMM's ViC abilities. We restrict this definition as a novel concept specifically in the field of LMM reasoning.

Technically, the ViC paradigm is fundamentally different from previous reasoning paradigms, *e.g.* text chain-of-thought (CoT) [40, 94, 95, 111, 120] and visual CoT [25, 63, 115, 121, 123]. The latter two paradigms only perform text-based reasoning with multiple intermediate steps, without incorporating the visual signals at each step, as shown in Figure 2. The continuous integration of visual feedback in ViC ensures that models can handle intricate tasks, *e.g.* navigation and driving, which is more aligned with the needs of agents [30, 92, 100] and robotics [87, 113].

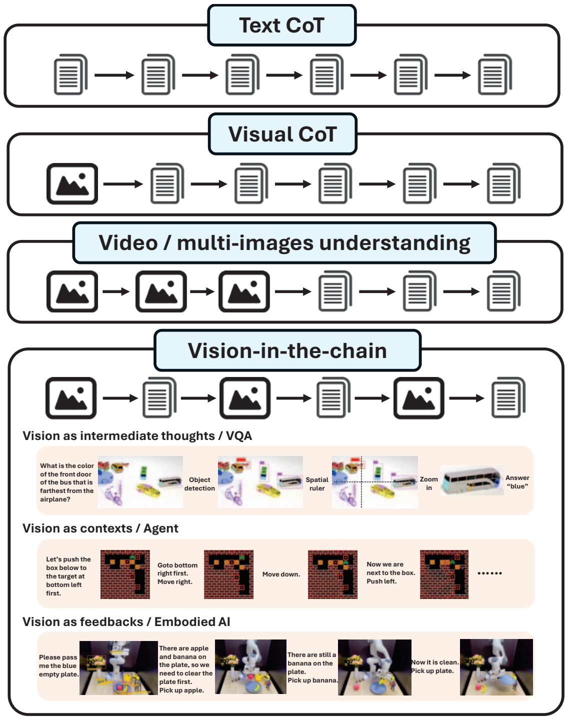

Text CoT
Visual CoT
Video / multi-images understanding
Vision-in-the-chain
Vision as intermediate thoughts / VQA
What is the color of the front door of the bus that is farthest from the airplane?
Object detection
Spatial ruler
Zoom in
Answer "blue"
Vision as contexts / Agent
Let's push the box below to the target at bottom left first.
Goto bottom right first. Move right.
Move down.
Now we are next to the box. Push left.
......
Vision as feedbacks / Embodied AI
Please pass me the blue empty plate.
There are apple and banana on the plate, so we need to clear the plate first. Pick up apple.
There are still a banana on the plate. Pick up banana.
Now it is clean. Pick up plate.

Figure 2. The difference between vision-in-the-chain reasoning and existing reasoning paradigm. Example images are adapted from [36, 71, 125].

Evidently, Agent environments are inherently the optimal scenarios for ViC-type reasoning. However, existing Agent environments [15, 88, 126] are not suitable for testing the intrinsic ViC capabilities of models. This is because the research and outcomes associated with complex agent scenarios are strongly coupled with the design of the agent systems themselves, such as prompts and pipelines, obscuring the model's inherent capabilities. With the above background, we present **MageBench**, a **M**ultimodal reasoning benchmark built upon light-weight **AGE**nt environments that using fixed minimal design of agent system in order to access the reasoning ability and potential for LMMs to be general agents. MageBench poses significant reasoning challenges and holds substantial practical value.

During environment selection, we prioritize the visual abilities required by the tasks rather than the relationships inherent to the environments themselves. Additionally, to accommodate the requirements of RL and scaling, we ensure that the environments are as lightweight as possible. Ultimately, we established WebUI to reflect cross-modal knowledge and engineering capabilities; Sokoban to represent the spatial intelligence and planning abilities required in the robotics domain; and Football to demonstrate social and interactive capabilities in multi-agent scenarios, as social and interactive capabilities are fundamental characteristics of intelligent agents [26, 97, 100].

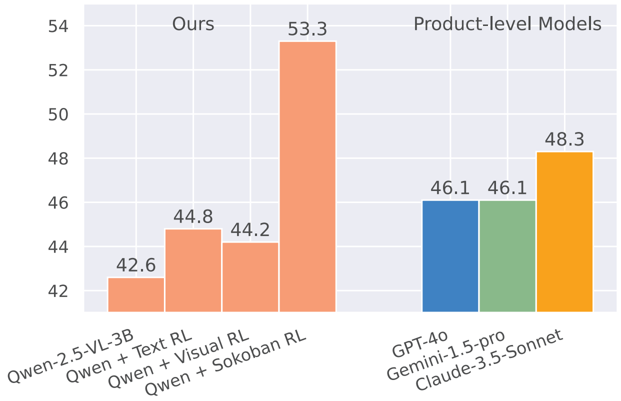


Figure 3. The Impact of Rule-based Reinforcement Learning on Sokoban-Global mini Results. This shows that our benchmark is well suited as a test scenario for LMM agentic RL studies.

We propose two baseline agent setting: Global (the model only observes the initial state and gives all actions) and Online (the model interacts with the environment to continuously obtain image observations and output actions), which correspond to Visual CoT and ViC types of reasoning, respectively. We tested 14 strongest open-source and close-source LMMs selected from each model family, and the level of human performance in Tab. 2, and more models in Supp. C.1. We found that in the Online setting, only GPT-4o and Gemini-1.5-pro outperformed the random level, and all of them are far inferior to human level. This shows that they severely lack ViC-type reasoning capabilities, making existing LMMs far from ideal for agent and robotics applications. This may inspire us to recognize that existing LMMs lack visual progressive training, such as that involving video. In addition, the existing models do a good job in the Global setting of WebUI. Claude can even approach human level. However, they failed to boost the result with browser's rendering feedback, but human can continually adapt their codes to nearly perfection. This is possibly caused by the shortcomings of interleaved image-text long context handing.

To mutually verify the scalability of our benchmark, we adopted a strategy similar to DeepSeek R1 Zero [29]. We employed rule-based RL and various datasets to train Qwen-2.5-VL-3B [9]. The results indicated that, regardless of whether it was trained on pure text, multimodal data, or in-domain environments, the model demonstrated enhanced reasoning capabilities (see Figure 3). Notably, after training with Sokoban, it exhibited performance in the Sokoban-Global setting that surpassed the results of product-level large models. This paper does not focus on innovative RL algorithms; however, the results suggest that our benchmark can effectively serve as a testing ground for the burgeoning research in agentic RL.

2. Related Work

Large Multimodal Models. The advent of large language models (LLMs) [6, 11, 17, 22] has demonstrated remarkable reasoning capabilities [94, 95] and the potential for general intelligence [23]. By employing a single model with different prompts, a multitude of tasks can be accomplished [21, 74]. A natural extension of this concept is to apply similar methods to other modalities to achieve general multimodal intelligence. Flamingo [8] was among the first to explore multimodal in-context learning [98, 103, 127], followed by the emergence of numerous large multimodal models [1–3, 6, 52, 83]. These models employ various technical approaches [42, 49–52, 82]. As technology has progressed, product-level multimodal large models such as GPT-4V [6], GPT-4O [2], Gemini [83], Claude [1], and Grok-2 [3] have showcased state-of-the-art performance.

Visual Reasoning. Chain-of-thought prompting [40, 94, 95], flow engineering [73], self-reflection [79, 105], and their various variants [124] have demonstrated significant improvements. In visual tasks, the primary evaluation datasets for visual reasoning are those based on VQA tasks, such as ScienceQA [58] and MathVista [59]. Due to the limitations of these evaluation datasets, many existing studies [25, 63, 115, 121, 123] on visual reasoning using “CoT" as a keyword mainly focus on extracting information from multimodal problems, and then utilize text-based intermediate processes such as captioning [121], rationales [58], relational graphs [63], and question tables [123].

Some recent works have attempted to incorporate procedural information from other modalities, leveraging ViC type reasoning to fulfill certain tasks, such as Image-of-thought [125] and DetToolChain [99] However, they are constrained to classical vision tasks and hence are not suitable for benchmarking.

Rule based RL and Test Time Scaling. In the context of reasoning scaling for LLMs[10, 72], the community has explored various approaches such as process reward models [47, 122] and MCTS-based decoding [28, 70]. With the successes of R1[29], rule-based RL on verifiable math[18, 33] and coding[14, 32, 45] datasets, has emerged as a standout technique. Through extensive community replication and research [35, 101], it has been observed that different optimization algorithms (e.g., PPO [77], GRPO [78], Reinforce++ [34]) are not the primary determinants of test-time scaling success. Instead, the key factor lies in the verifiable datasets and rule-based rewards, which offer more accurate, stable, and unhackable rewards compared to any previous reward models.

Recently, the visual research community has demonstrated a growing interest in visual test-time scaling. Beyond visual math problems [57–59], our research tasks serve as an excellent testing ground with verifiable, strong reasoning, and easy-to-scale features.

Table 1. Comparing existing LMM benchmark types and ours.

Bench. Type	Works	CoT type	Target
Perception	[24, 39, 43, 62, 96, 129]	None	Accessing LMMs' capability to seek information from images.
Knowledge and Reasoning	[31, 58, 59, 110, 116]	text	Accessing LMMs' general ability including perception and reasoning. Usually first percept and then reason by text.
Agent	[53, 54]	ViC/text	Accessing agent system design and LMMs' perception and decision making. Result and planning rely heavily on system and pipeline design.
Robotic	[12, 61, 65, 81]	ViC	Same with agent but require more real world commen sence.
Ours (Knowledge and Reasoning)	-	ViC	Accessing LMM it-selves planning and perception with fixed and simplest system design via light-weight agent environment.

Benchmarking LMMs There are numerous evaluation datasets for LMMs that comprehensively assess various capabilities, We summarize them into three categories in Table 1. For perception-oriented benchmarks [24, 39, 62, 96, 129], CoT generally cannot boost the result, as they do not require reasoning. Existing vision reasoning [27, 31, 38, 116] and knowledge-oriented benchmarks [110] benefit from text-based CoT. Most of these evaluations are presented in the form of multiple-choice questions, which simplify and abstract real-world problems [44]. Another approach to evaluating models is to deploy them on agents for task-level end-to-end assessments [53, 54]. However, not all meaningful environments can prohibit meaningful and representative reasoning skills of LMMs, as the planning progress can be largely utilized by the system design (like work flow, specific system prompt, etc.). We will elaborate more on this in Sec.3 and Supp. A.4.

As shown in Table 1, our work is built upon three lightweight agent environments with minimal design of work flow and system prompts, but it is not an agent benchmark. The difference is that we ensure the system prompts do not leak any planning clues and do not allow system design. This will amplify the model's intrinsic planning capabilities. In addition, our environments selection is reasoning ability-oriented, instead of environment-oriented as agent benchmarks do. Hence, it is not and not necessarily highly related between environments, but they are complementary in terms of reasoning abilities.

3. Introducing MageBench

3.1. Environment selection

MageBench aims to select the most simple, representative environments from the perspective of reasoning and with generalization ability. We investigate dozens of environments and select those meet the criteria below:

- **Representativeness on Reasoning**: Considering the reasoning abilities required for LMMs to become general agents, we believe they need at least real-world engineering knowledge to assist humen (WebUI), spatial understanding and planning (Sokoban) and skill to cooperate for future advanced multi-agent system (Football). We investigate and exclude many robotics simulation environments(*e.g.*, [37, 69, 112] and OmniGibson in [54]), Virtual reality game (*e.g.*, [15, 48, 128]) and app manipulation (*e.g.*, [20, 46, 66, 88, 91]), although they are more practical with complex actions, their high level planning are actually simple and direct. For example, to buy an commodity in a webpage, there is a clear high-level "search-browse-click-..." pipeline can be systematically designed and the model only needs to fulfill low-level tasks like percept and act, although they are still not easy.
- **Visual Feedback**: We excluded environments where images are not the sole form of feedback [104, 126].
- **Simplicity**: The environments are highly streamlined, with minimal database, library, memory and hardware requirements. This is crucial for serving as a testing ground for the future rule-based on-policy RL trainings.
- **Discrete Action Space**: Since expecting a LMM to output continuous values such as angles is clearly impractical and can lead to significant errors, we require that the environment's inputs be limited to finite and discrete actions, such as the up, down, left, and right movements in Sokoban. This requirement excludes many Embodied AI environments like ALFRED [80], Habitat [76], VirtualHome [69] and etc. Although we could use predefined actions, doing so might introduce additional constraints.

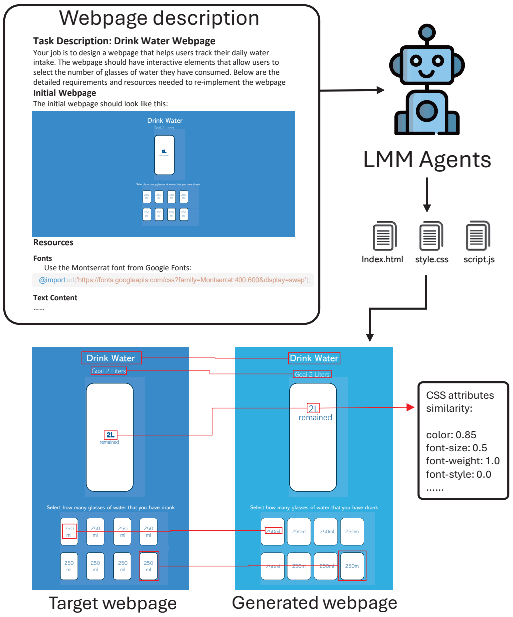
Based on the aforementioned criteria, we selected representative environments for each capability, ultimately choosing Sokoban, WebUI, and Football to form our benchmark. It is worth noting that for each capability, there might be alternative environments available; however, we opted for the relatively simpler ones. Our goal is for our benchmark to cover all capabilities, without requiring each environment to be irreplaceable or to encompass all similar environments.

3.2. Agent settings

During testing, we treat LMM as agent and utilize two fixed standard settings:

- **Global Planner Agent.** The model only observes the minimal designed system prompt(p_{sys}) and the initial environment(o_0) once and continuously makes all subsequent decisions($\mathbf{a_i}$), formulated as

$$\pi_\theta(p_{sys}, o_0) \to \mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_T}.$$



Task Description: Drink Water Webpage

Your job is to design a webpage that helps users track their daily water intake. The webpage should have interactive elements that allow users to select the number of glasses of water they have consumed. Below are the detailed requirements and resources needed to re-implement the webpage

Initial Webpage

The initial webpage should look like this:

Resources

Fonts

Use the Montserrat font from Google Fonts:

```
@import url('https://fonts.googleapis.com/css?family=Montserrat:400,600&display=swap');
```

Text Content

……

CSS attributes similarity:

color: 0.85
font-size: 0.5
font-weight: 1.0
font-style: 0.0
……

Figure 4. An overview of WebUI and its evaluation. LMM Agents are required to re-generate the webpage according to the description. We match the generated elements with the atomic elements in the ground truth. Then we compare the CSS attributes to obtain a similarity score. A specific example of task description can be found in Supp. G.1.3. Technique details of evaluation can be found in Supp. A.1.3.

- **Online Planner Agent.** The model observes and analyzes on each step and takes actions online, which forms the ViC-type reasoning, for $t = 1 \rightarrow T$:

$$\pi_\theta(p_{sys}, ..., \mathbf{a_{t-1}}, \mathbf{o_{t-1}}, \mathbf{a_t}, \mathbf{o_t}) \rightarrow \mathbf{a_{t+1}}.$$

There will be more details about both settings (e.g., agent memory), we left them to Supp. B. We will introduce each of our environments in detail below.

3.3. WebUI

We collected minimal web projects from GitHub, strictly adhering to the corresponding licenses, consisting of only a few HTML, JavaScript, and CSS files. For each webpage, we will create a Markdown-formatted webpage description. The webpage description is a image-text interleaved document that provides sufficient information to fully reconstruct the website. It includes detailed webpage descriptions, external resources, and screenshots before and after various interactions and etc. The task of WebUI is to reconstruct the website based on the description and a Google Chrome web driver.

The evaluation of WebUI is based on comparing the CSS properties of atomic elements. We first define **atomic elements** as follows: Suppose webpage B is a reproduction of webpage A. An HTML tag in webpage A is considered atomic if any attribute it contains is guaranteed to appear in a corresponding tag in webpage B. For example, all headings, texts, and images in webpage A may exist in webpage B as different types of tags (*e.g.*, both h1 and span tags can display text). During evaluation, we first match the atomic elements in the target webpage with the elements in the webpage generated by the agent using a carefully designed matching algorithm. Next, we compare the CSS property similarity of the successfully matched elements, typically using metrics such as relative error or checking if the values are equal. Finally, we provide a weighted similarity score as the evaluation result. We refer to this metric as **Atomic Element Similarity (AES)**. The technical details involved are extensive and will be presented in the Supp. A.1.3.

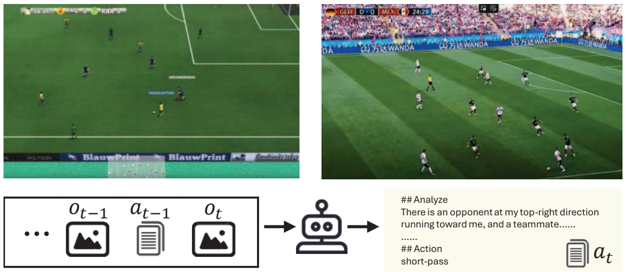

Analyze
There is an opponent at my top-right direction running toward me, and a teammate……
……
Action
short-pass

Figure 5. The segment from the Germany vs. Mexico match in the 2018 FIFA World Cup (right), and the initial game scene inspired by it (left). Model analyzes and generates one of the actions (bottom), similar process for the Sokoban-Online.

3.4. Sokoban

Sokoban is a well-known logic video game where the task is to maneuver a character to push all boxes onto designated target areas. The game is highly challenging due to the presence of numerous losing states and traps, necessitating strong planning abilities [71]. The planning and reasoning capabilities of LMMs may effectively mitigate this issue, making this environment ideal for testing an agent's path finding, planning, error correction, and foresight abilities. We utilize the rendering environment provided by [71]. We generated and stored 182 levels of varying difficulty. Further details can be found in Supp. A.2.

We also adapt the reward value defined in [71] to evaluate the LMMs. However, unlike their approach, we use the historically optimal reward throughout the trajectory rather than the final reward. This is because the reward includes a penalty for the number of steps taken. Based on extensive testing, we found that given the current capabilities of LMMs, using the final reward tends to be dominated by factors such as the model's output length, the number of steps we set (for the online setting), the length penalties and etc. This is an outcome we aim to avoid.

3.5. Football

Football, as one of the most competitive and cooperative sports, can fully demonstrate an agent's spatial intelligence and collective intelligence, potentially providing a research foundation for future LMM multi-agent systems. We chose the rendering platform provided by Google Football Research [41] for our study. We generated 108 scenarios as initial states, with each initial scenario serving as a level (analogous to a level in Sokoban). These levels cover different areas of the football field and are categorized into personal (scenarios where good passing routes are unavailable, requiring players to showcase individual skills), teamwork (scenarios suitable for demonstrating team collaboration and passing), and real-world (scenarios from actual World Cup matches). The agent will use text outputs to perform 18 actions (including moving, long passing, and shot), starting from each scenario and simulating up to 400 frames until a goal is scored or the ball is intercepted. More details can be found in Supp. A.3. We also designed an automatic rendering algorithm (see Supp. A.3.4.) that reduces the average number of API calls needed per scenario from 80 to less than 20, without affecting the results.

During the simulation process, the LMM agent will control only one player (always the player in possession of the ball), while the other players are controlled by built-in AI bots. The presence of numerous agents results in a highly stochastic environment simulation. Using metrics such as win-rate leads to high variance and requires extensive repetitions. To address this issue, we carefully designed a more dense reward system to comprehensively evaluate the model's capabilities. The design of this reward system is as follows:

$$
\begin{aligned}
R^{(t)} =& \lambda_1 S_{move}^{(t)} + \lambda_2 S_{oppo}^{(t)} + \lambda_3 \delta_{scored}^{(t)} \\
&+ \lambda_4 \delta_{stole}^{(t)} \frac{t}{T} + \lambda_5 \delta_{pass}^{(t)} S_{pass}^{(t)} \\
&+ \lambda_6 \delta_{shot}^{(t)} S_{shot}^{(t)},
\end{aligned} \tag{1}
$$

where $\delta_{event}^{(t)}$ is a indicator function that event happened at time step t. $S_{move}^{(t)}$ and $S_{oppo}^{(t)}$ represent the reward values obtained after processing the distance the ball has been moved forward and the number of opponents surpassed, respectively. $S_{pass}^{(t)}$ and $S_{shot}^{(t)}$ are metrics that quantify the quality of passing and shot. For additional details and specific parameters, please refer to Supp. A.3.3.

4. Rule-based RL for LMMs

Since **RL is not the primary focus of this study** and is only used to complement and validate our benchmark and the synergy of multimodal test time scaling, we did not invest substantial effort in algorithmic design. We employed the PPO algorithm [77] and utilized a rule-based reward system similar to MageBench-Sokoban. We conducted experiments with KL divergence and implemented a repeat penalty on the reward, as illustrated in Figure 6. The RL experiments were designed to demonstrate that our benchmark is forward-looking and can effectively validate the emerging research in Agentic RL. Details of the RL implementation are provided in the Suppl. E.

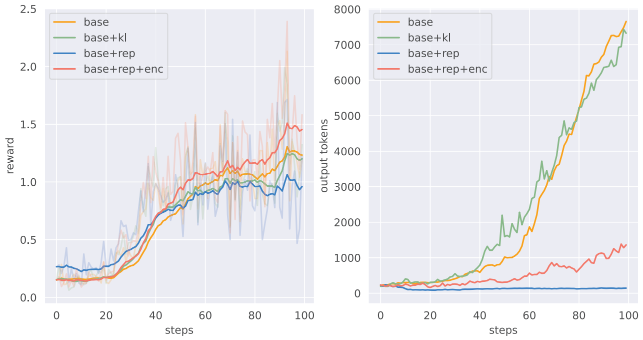


Figure 6. Task rewards (left) and output tokens (right) curves when changing reward strategies. "+kl" means we will add kl regularization with coefficient 0.001. "+rep" stands for using a repeat penalty that will minus 0.2 from reward if repetition detected. "+enc" means encourage the model with $len(tokens) \times 0.001$ only when output length is less than 200 tokens.

5. Results and Analysis

5.1. Standard setting

We use the MageBench and the minimal and unified agent setting as the carrier to study what kind of LMMs have the potential to become a genral agent. It is worth stating that these models may perform better in MageBench with specially designed agents, prompts, and settings, but for the sake of a fair comparison, we will use the same standard settings below.

- Online planner will be able to see 5 history actions and 1 history image, so that we can fairly evaluate models without multi-images capability. According to our tests, these two types of memory do not have a significant impact on the performance for current models, See Supp. C.2.
- Agent-based evaluation always has a large stochastic variance. Hence, we require all experiments to be repeated and averaged. For MageBench mini set, Sokoban and WebUI should repeat 3 times and Football repeat 10 times. For MageBench complete set (in Supp. C.1), all experiments reported are averaged over 3 repetition.

We select LMMs that are trained for general usages and support flexible interleaved image-text inputs, and have at least 4096 context length to evaluate. Table 2 presents the test results under the standard settings on the MageBench mini subset. We evaluated the strongest models from each

Table 2. Evaluation on MageBench test-mini subset with unified prompt. IFE stands for "Instruction Following Error". It is defined as follows: if more than 90% of the outputs are not parsed into valid actions, or if 90% of the actions are the same (indicating that the model is repeating a certain action), it is considered an IFE. δ represents the significance difference derived from repeated experiments.

Model	WebUI AES (%)		Sokoban Reward		Football Reward
	Global ($\delta = \pm 2.0$)	Online ($\delta = \pm 2.6$)	Global ($\delta = \pm 1.9$)	Online ($\delta = \pm 1.8$)	Online ($\delta = \pm 2.2$)
Phi-3.5-V-4.2B [5]	0.03	11.78	44.95	44.14	10.00
DeepSeek-VL-7B [56]	13.78	8.07	IFE	IFE	IFE
Xcomposer-2.5-7B [117]	12.95	15.26	45.57	42.28	IFE
MiniCPM-V2.6-8B [106]	12.52	10.03	42.52	42.36	2.10
LLaVA-v1.5-13B [50]	11.95	9.92	46.28	42.59	18.13
Llava-1.6-34B [51]	6.57	15.17	43.19	43.47	3.91
Yi-VL-34B [108]	7.14	12.08	IFE	IFE	IFE
Qwen2-vl-72B [93]	7.37	13.03	46.21	43.22	15.75
NVLM-72B [19]	4.71	14.46	43.46	44.07	14.46
InternVL2-76B-LLaMA3 [16]	16.22	16.76	44.56	43.58	10.50
Qwen2.5-vl-72B [9]	24.57	25.13	45.13	50.88	19.03
Llama-3.2-90B-Vision [22]	35.09	27.47	45.80	IFE	14.28
Claude-3.5-Sonnet [1]	**64.11**	**62.08**	**48.26**	45.35	16.94
Gemini-1.5-pro [83]	44.79	39.30	46.13	51.84	18.33
GPT-4o [2]	34.28	35.50	46.09	**53.03**	**21.20**
Idle Baseline	0.00	0.00	41.18	41.18	2.53
Random Baseline	0.00	0.00	46.61	46.61	17.33
Human	68.71	94.32	83.63	96.85	54.68

open-source LMM series (first block) and the results of three closed-source product-level models (second block). We also provided idle and random baselines, as well as human-level results in the third block. For Sokoban and Football in the idle baseline, no actions were taken (an idle action is available in the football environment). The random baseline refers to randomly selecting a possible action. During human-level testing, annotators were selected from several PhD candidates with strong reasoning abilities. The testing conditions for the human annotators were completely fair when compared to the models. For example, in the results for Sokoban-Global, humans could not control the player and could only observe the initial screen and record all actions using their imagination. In WebUI-Global, humans were not allowed to view the browser's rendered output while writing code, whereas in the WebUI-Online setting, humans were permitted to observe the rendered screen.

Overall, we found that although open-source models have achieved performance levels comparable to closed-source models on many VQA tasks, they still fall significantly short of the requirements for AI agents. In the Sokoban and Football, only GPT-4o and Gemini performed better than the random baseline under the online setting. This may be attributed to the optimization of product-level models for multi-turn dialogue and multi-image scenarios. Claude's performance under the Global setting was very impressive, being the only model that could work in the Global setting for Sokoban, but it still lagged far behind human-level performance. This demonstrates that humans possess strong imaginative and think-ahead abilities, which are substantially lacking in current LMMs.

We were pleasantly surprised to observe that Claude demonstrated performance close to that of computer science PhD candidates in the WebUI-Global results. However, while humans can modify webpage code based on rendered screen to make it almost identical to the target webpage, current models fail to achieve this. We tried several prompt and self-reflection types for WebUI-Online, and left the details and results in Supp. C.4.

The prompts and details of all environment and agent settings can be found in Supp. A and G.

5.2. Best-of-N Result

We use best-of-N scaling curves to investigate the potential of the models in relevant tasks in Fig. 7. Firstly, we observe that in the Football environment, many models can surpass human performance by computing the best-of-N, with a steep upward slope. This indicates that there is a significant opportunity to achieve substantial improvements in this task using RL algorithms. However, in the Sokoban task, the best-of-N curve exhibits slow growth, suggesting that it will be difficulty to generate valuable trajectories through trial and error and experience accumulation. During our RL training, by leveraging repeat penalty and length encourage in reward, we speed up this process. The best-of-N curve on WebUI is very similar to the pass@N metric in code generation. We can observe that both model enhancement and increasing N can lead to substantial gains.

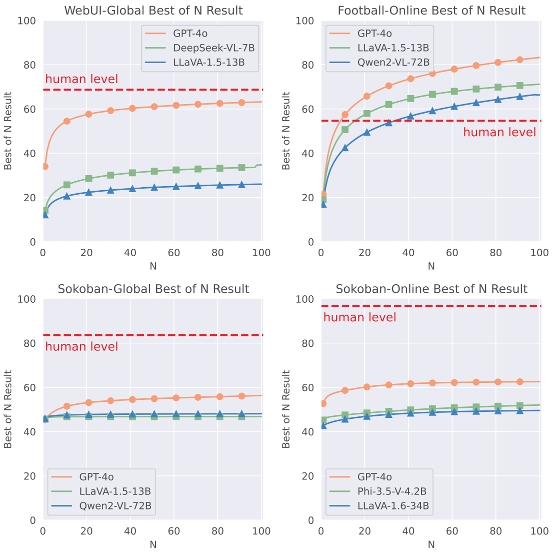


Figure 7. Best-of-N results of selected models.

5.3. Error Statistics and Analysis

We categorize the types of errors for different models on Sokoban and Football, finding that models suffer from repeating actions and instruction following errors. This indicates that existing models have deficiencies in training with the Vision-in-the-Chain type of data studied in this paper. We left the detailed statistics in Supp. C.3 .

Figure 8 illustrates the composition of lost scores for different models on the WebUI. The AES section represents the scores obtained by the models. If a parsing error (unable to recognize code) or a render error (unable to render the initial web page due to some compilation issues) occurs, the model loses all the scores. If interaction error happened, the model loses the scores of the sub-webpage. The attribute similarity is only calculated when a model does not encounter "Par.", "Ren.", or "Act." errors and successfully matches the corresponding HTML tags.

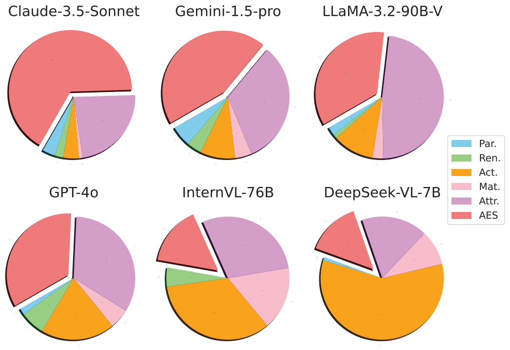


Figure 8. WebUI error construction. Each part of the pie graph is the corresponding score that model lost (or earned for AES part). "Par."=Parsing Error (or invalid actions); "Ren." = Rendering Error; "Act." = Webpage Interaction Error; "Mat." = HTML Tag Matching Error; "Attr." = Attribute Similarity Lost.

From Figure 8, it can be observed that stronger models tend to have fewer "Par.," "Ren.," and "Act." type errors (functional errors) caused by syntax and formatting issues, with a higher proportion of attribute setting errors instead. These models require stronger visual grounding capabilities to achieve further improvements. Conversely, weaker models have a higher proportion of functional errors, indicating their insufficient knowledge regarding web pages.

5.4. RL training results

Table 3 presents the results of training Qwen-2.5-instruct-vl. The results for "Qwen + Text" RL are obtained by using rule-based RL on the purely text-based reasoning dataset DeepScaler[60]. In contrast, the results for "Qwen + Visual RL" are based on training with the visual reasoning dataset MathV[89]. We left more details in Supp. E.

Table 3. MageBench results for RL trainings on different data.

Model	Sokoban-G	WebUI-G	Football-O
Claude-3.5-Sonnet	48.26	64.11	16.94
Gemini-1.5-pro	46.13	44.79	18.33
GPT-4o	46.09	34.28	21.20
Qwen-VL-2.5-3B-Instruct	42.35	11.20	15.36
Qwen + Text RL	44.81	11.80	18.46
Qwen + Visual RL	44.20	12.77	17.39
Qwen + Sokoban RL	**53.30**	10.52	19.19

The results first demonstrate that in an in-domain scenario, conducting agent-level RL can significantly enhance the performance of LMMs in the corresponding environment. Even a 3B model, after training, can achieve results that surpass those of product-level large models, which also attests to the scalability of our benchmark. Secondly, we were surprised to find that out-of-domain RL exhibited a certain degree of generalization capability. Specifically, RL training on text-based reasoning, visual reasoning data, and Sokoban all led to improvements in both Sokoban and Football. This suggests that there are certain connections between different agents, hinting at the potential emergence of future general intelligent agents. However, there was no improvement observed for WebUI, likely because WebUI primarily assesses the model's knowledge application, and reinforcement learning does not directly endow the model with new knowledge. This indicates that our benchmark environment selection is fairly comprehensive.

6. Summarization and Limitations

In this paper, we introduce a new benchmark called MageBench. We conducted tests on a wide range of both open-source and close-source LMMs. The results indicate that current models lack ViC type reasoning abilities. Our current environment is relatively limited and simple. We plan to incorporate more comprehensive content in the future. We hope to offer LMM developers valuable insights and optimization directions.

References

[1] Build with Claude, 2024. 3, 7

[2] Hello GPT-4o, 2024. 3, 7

[3] Grok-2 beta release, 2024. 3

[4] Introducing OpenAI o1-preview, 2024. 2

[5] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 7

[6] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3

[7] Toufique Ahmed and Premkumar Devanbu. Few-shot training llms for project-specific code-summarization. In *ASE*, 2022. 2

[8] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 2, 3

[9] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 7

[10] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, and Kai Dong. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024. 3

[11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, et al. Language models are few-shot learners. In *NeurIPS*, 2023. 2, 3

[12] Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, et al. Partnr: A benchmark for planning and reasoning in embodied multi-agent tasks. *ICLR*, 2024. 4

[13] Chaofeng Chen, Sensen Yang, Haoning Wu, Liang Liao, Zicheng Zhang, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-Ground: Image quality grounding with large multi-modality models. In *ACM MM*, 2024. 2

[14] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, and Yuris Burda. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. 3

[15] Peng Chen, Pi Bu, Jun Song, Yuan Gao, and Bo Zheng. Can vlms play action role-playing games? take black myth wukong as a study case. *arXiv preprint arXiv:2409.12889*, 2024. 2, 4

[16] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 7

[17] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, et al. Palm: Scaling language modeling with pathways. *JMLR*, 2023. 2, 3

[18] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, and Jerry Tworek. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 3

[19] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024. 7

[20] Shihan Deng, Weikai Xu, Hongda Sun, Wei Liu, Tao Tan, Jianfeng Liu, Ang Li, Jian Luan, et al. Mobile-bench: An evaluation benchmark for llm-based mobile agents. *arXiv preprint arXiv:2407.00993*, 2024. 4

[21] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, et al. A survey on in-context learning. In *EMNLP*, 2024. 3

[22] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 3, 7

[23] Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. How far are we from agi. *arXiv preprint arXiv:2405.10313*, 2024. 3

[24] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 4

[25] Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, et al. Cantor: Inspiring multimodal chain-of-thought of mllm. In *ACM MM*, 2024. 2, 3

[26] Richard Goodwin. Formalizing properties of agents. *JLC*, 1995. 2

[27] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 2, 4

[28] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rStar-Math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*, 2025. 3

[29] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, and Shirong Ma. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 3

[30] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024. 2

[31] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024. 4

[32] Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, and Samir Puranik. Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938*, 2021. 3

[33] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021. 3

[34] Jian Hu. REINFORCE++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025. 3

[35] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. Open-Reasoner-Zero: An open source approach to scaling reinforcement learning on the base model. https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero, 2025. 3

[36] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*, 2023. 2

[37] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022. 4

[38] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 2, 4

[39] Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. TableVQA-Bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*, 2024. 4

[40] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022. 2, 3

[41] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, et al. Google research football: A novel reinforcement learning environment. In *AAAI*, 2020. 6

[42] Fei Lei, Zhongqi Cao, Yuning Yang, Yibo Ding, and Cong Zhang. Learning the user's deeper preferences for multimodal recommendation systems. *TOMM*, 2023. 3

[43] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *CVPR*, 2024. 2, 4

[44] Jinming Li, Yichen Zhu, Zhiyuan Xu, Jindong Gu, Minjie Zhu, Xin Liu, Ning Liu, Yaxin Peng, Feifei Feng, and Jian Tang. Mmro: Are multimodal llms eligible as the brain for in-home robotics? *arXiv preprint arXiv:2406.19693*, 2024. 4

[45] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, and James Keeling. Competition-level code generation with alphacode. *Science*, 2022. 3

[46] Yanda Li, Chi Zhang, Wanqi Yang, Bin Fu, Pei Cheng, Xin Chen, Ling Chen, and Yunchao Wei. Appagent v2: Advanced agent for flexible mobile interactions. *arXiv preprint arXiv:2408.11824*, 2024. 4

[47] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. 2023. 3

[48] Haowei Lin, Zihao Wang, Jianzhu Ma, and Yitao Liang. Mcu: A task-centric framework for open-ended agent evaluation in minecraft. *arXiv preprint arXiv:2310.08367*, 2023. 4

[49] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *ACL*, 2023. 3

[50] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 7

[51] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, ocr, and world knowledge, 2024. 7

[52] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 2, 3

[53] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023. 4

[54] Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yifan Xu, Xixuan Song, Shudan Zhang, Hanyu Lai, et al. Visualagentbench: Towards large multimodal models as visual foundation agents. *arXiv preprint arXiv:2408.06327*, 2024. 4

[55] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2025. 2

[56] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 7

[57] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. 2021. 3

[58] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022. 2, 3, 4

[59] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 3, 4

[60] Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. DeepScaleR: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-

Preview-with-a-1-5B-Model-by-Scaling-RL - 19681902c1468005bed8ca303013a4e2, 2025. 8

[61] Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models. In *ICRA*, 2024. 4

[62] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, 2022. 4

[63] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *CVPR*, 2024. 2, 3

[64] Yasmin Moslem, Rejwanul Haque, John D Kelleher, and Andy Way. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*, 2023. 2

[65] Sid Nayak, Adelmo Morrison Orozco, Marina Have, Jackson Zhang, Vittal Thirumalai, Darren Chen, Aditya Kapoor, Eric Robinson, Karthik Gopalakrishnan, James Harrison, et al. Long-horizon planning for multi-agent robots in partially observable environments. *NeurIPS*, 2024. 4

[66] Songqin Nong, Jiali Zhu, Rui Wu, Jiongchao Jin, Shuo Shan, Xiutian Huang, and Wenhao Xu. Mobileflow: A multimodal llm for mobile gui agent. *arXiv preprint arXiv:2407.04346*, 2024. 4

[67] Yingzhe Peng, Chenduo Hao, Xu Yang, Jiawei Peng, Xinting Hu, and Xin Geng. Learnable in-context vector for visual question answering. *arXiv preprint arXiv:2406.13185*, 2024. 2

[68] Xiao Pu, Mingqi Gao, and Xiaojun Wan. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*, 2023. 2

[69] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *CVPR*, 2018. 4

[70] Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. Mutual reasoning makes smaller llms stronger problem-solvers. *arXiv preprint arXiv:2408.06195*, 2024. 3

[71] Sébastien Racanière, Théophane Weber, David Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adrià Puigdomènech Badia, Oriol Vinyals, et al. Imagination-augmented agents for deep reinforcement learning. *NeurIPS*, 2017. 2, 5

[72] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, and Sarah Henderson. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 3

[73] Tal Ridnik, Dedy Kredo, and Itamar Friedman. Code generation with alphacodium: From prompt engineering to flow engineering. *arXiv preprint arXiv:2401.08500*, 2024. 3

[74] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024. 3

[75] Ander Salaberria, Gorka Azkune, Oier Lopez de Lacalle, Aitor Soroa, and Eneko Agirre. Image captioning for effective use of language models in knowledge-based visual question answering. *ESA*, 2023. 2

[76] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *CVPR*, 2019. 4

[77] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3, 6

[78] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, and Mingchuan Zhang. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 3

[79] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *NeurIPS*, 2024. 3

[80] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *CVPR*, 2020. 4

[81] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *ICCV*, 2023. 4

[82] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 3

[83] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 3, 7

[84] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[85] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[86] Jiale Wang, Gee Wah Ng, Lee Onn Mak, Randall Cher, Ng Ding Hei Ryan, and Davis Wang. QCaption: Video captioning and q&a through fusion of large multimodal models. In *FUSION*, 2024. 2

[87] Jiaqi Wang, Zihao Wu, Yiwei Li, Hanqi Jiang, Peng Shu, Enze Shi, Huawen Hu, Chong Ma, et al. Large language models for robotics: Opportunities, challenges, and perspectives. *arXiv preprint arXiv:2401.04334*, 2024. 2

[88] 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. 2, 4

[89] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *NeurIPS*, 2024. 8

[90] Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*, 2023. 2

[91] Luyuan Wang, Yongyu Deng, Yiwei Zha, Guodong Mao, Qinmin Wang, Tianchen Min, Wei Chen, and Shoufa Chen. MobileAgentBench: An efficient and user-friendly benchmark for mobile llm agents. *arXiv preprint arXiv:2406.08184*, 2024. 4

[92] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, et al. A survey on large language model based autonomous agents. *FCS*, 2024. 2

[93] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 7

[94] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 2, 3

[95] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022. 2, 3

[96] Jingxuan Wei, Nan Xu, Guiyong Chang, Yin Luo, Bi-Hui Yu, and Ruifeng Guo. mChartQA: A universal benchmark for multimodal chart question answer based on vision-language alignment and reasoning. *arXiv preprint arXiv:2404.01548*, 2024. 4

[97] Michael Wooldridge and Nicholas R Jennings. Intelligent agents: Theory and practice. *KER*, 1995. 2

[98] Yongliang Wu and Xu Yang. A glance at in-context learning. *FCS*, 2024. 3

[99] Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. *arXiv preprint arXiv:2403.12488*, 2024. 3

[100] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023. 2

[101] Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-RL: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*, 2025. 3

[102] Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*, 2023. 2

[103] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. 2024. 2, 3

[104] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *NeurIPS*, 2022. 4

[105] 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. 3

[106] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 7

[107] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 2

[108] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 7

[109] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2

[110] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, , et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 4

[111] Eric Zelikman, YH Wu, Jesse Mu, and Noah D Goodman. STaR: Self-taught reasoner bootstrapping reasoning with reasoning. In *NeurIPS*, 2024. 2

[112] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, et al. Transporter networks: Rearranging the visual world for robotic manipulation. In *CoRL*, 2021. 4

[113] Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*, 2023. 2

[114] Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study. 2023. 2

[115] Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *arXiv preprint arXiv:2401.02582*, 2024. 2, 3

[116] Jiaxin Zhang, Zhongzhi Li, Mingliang Zhang, Fei Yin, Chenglin Liu, and Yashar Moshfeghi. GeoEval: benchmark for evaluating llms and multi-modal models on geometry problem-solving. *arXiv preprint arXiv:2402.10104*, 2024. 4

[117] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 7

[118] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2

[119] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. Benchmarking large language models for news summarization. *ACL*, 2024. 2

[120] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022. 2

[121] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 2, 3

[122] Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*, 2025. 3

[123] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. 2023. 2, 3

[124] Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models. *arXiv preprint arXiv:2310.04406*, 2023. 3

[125] Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*, 2024. 2, 3

[126] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023. 2, 4

[127] Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. *arXiv preprint arXiv:2402.11574*, 2024. 3

[128] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, et al. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023. 4

[129] Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. MultiChartQA: Benchmarking vision-language models on multi-chart problems. *arXiv preprint arXiv:2410.14179*, 2024. 4