OWLViz: An Open-World Benchmark for Visual Question Answering

Thuy Nguyen
Reasoning Foundation
Washington, D.C., USA

Dang Nguyen
University of Maryland
College Park, USA

Hoang Nguyen
Posts and Telecommunications
Institute of Technology
Ha Noi, Viet Nam

Thuan Luong
Posts and Telecommunications
Institute of Technology
Ha Noi, Viet Nam

Franck Dernoncourt
Adobe Research
Seattle, USA

Long Dang Hoang
Posts and Telecommunications
Institute of Technology
Ha Noi, Viet Nam
longdh@ptit.edu.vn

Viet Dac Lai
Adobe Research
San Jose, USA

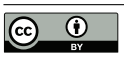
ABSTRACT

We present OWLViz, a challenging benchmark for **O**pen **W**or**L**d **VIS**ual question answering that evaluates multimodal AI systems on realistic, practical tasks. OWLViz features 248 carefully curated questions requiring the integration of multiple capabilities: common-sense knowledge, visual understanding, web exploration, and specialized tool usage. The benchmark specifically challenges models with visually degraded inputs, complex multi-step reasoning involving counting and measurement operations, and knowledge-intensive queries requiring external information retrieval from minimal visual cues. While humans achieve 69.2% accuracy on these intuitive tasks in under one minute, even state-of-the-art VLMs struggle dramatically, with the best model, Gemini 2.5 Pro, achieving only 27.09% accuracy. Current tool-calling agents and GUI agents, which rely on vision and vision-language models as tools, perform even worse, often failing to engage with available tools effectively. This substantial performance gap reveals critical limitations in multimodal systems' ability to select appropriate tools, coordinate heterogeneous resources, and execute complex reasoning sequences. OWLViz establishes new directions for advancing practical, open-world AI research and agent development.

KEYWORDS

Multimodal Agents, Visual Question Answering Benchmark, Open-World Reasoning

ACM Reference Format:
Thuy Nguyen, Dang Nguyen, Hoang Nguyen, Thuan Luong, Franck Dernoncourt, Long Dang Hoang, and Viet Dac Lai. 2026. OWLViz: An Open-World Benchmark for Visual Question Answering. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/EPVO9609



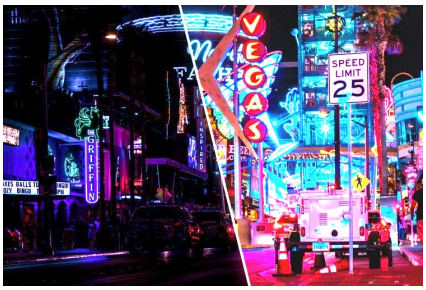
Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/EPVO9609

1 INTRODUCTION

Large Vision Language Models (VLMs) have recently demonstrated impressive visual understanding and reasoning capabilities across numerous tasks [22]. Equipped with these advanced capabilities, VLMs are rapidly surpassing existing AI benchmarks [4]. As the AI community pursues increasingly challenging problems, benchmark tasks are shifting from conventional fundamentals toward more human-centric challenges [3, 12, 32]. These human-centric tasks typically demand broader world knowledge, situation awareness, and more complex reasoning.

While there are many benchmarks for VQA that have been created, existing efforts face several notable problems. Many early benchmarks only focus on visual understanding tasks such as entity detection and entity attribution [9, 25, 31]. As a result, these benchmarks can only be used to evaluate the VLM's grounding capability, while many questions require additional capabilities such as reasoning. Recent benchmarks have shifted the focus towards complex questions where compositional reasoning and spatial reasoning are evaluated [2, 7, 13]. Even though tool calling capability was investigated in a close-environment [19, 24, 29], these datasets only addressed a limited set of tools. As a result, these benchmarks fail to fully capture the complexity of real human queries that often require interpreting complex visual contexts, combining heterogeneous sources of knowledge, and reasoning over long, multi-step chains of actions [27]. As such, current evaluations do not sufficiently evaluate a model's ability to act as an agent that can meaningfully invoke and coordinate tools to solve practical, open-ended questions grounded in the visual world.

To address this gap, we introduce OWLViz, a novel benchmark dataset specifically designed to evaluate vision-language models' ability to utilize tools in complex, multimodal reasoning tasks. OWLViz necessitates three distinct skill sets that challenge both VLMs and agentic systems. **First**, the dataset incorporates visually degraded inputs featuring low brightness, poor contrast, or



(a) Question: *"How many people are visible on the left side of the white line that cuts across the photo? Provide a numeric answer."*
Answer: 2
Skills: using external API, human recognition.
Difficulty level: 1.

(b) Question: *"How many umbrellas have 3 or more colors? Provide a numeric answer."*
Answer: 2.
Skills: object recognition, attribute identification, counting, object detection.
Difficulty level: 2.

(c) Question: *"This is in Fairfax, Virginia. What is the name of the road shown in the photo?"*
Answer: Shadowridge Dr; Shadowridge drive; Shadowridge.
Skills: OCR, knowledge search, knowledge retrieval, GUI, comparison, spatial relationships.
Difficulty level: 3.

Figure 1: Examples of the three core challenges in our OWLViz dataset. (a) Challenging visual conditions require image enhancement or specialized recognition tools to count people on a white line in a low-contrast night scene. (b) Complex reasoning tasks demanding object detection, attribute identification, and precise counting of multi-colored umbrellas in a dynamic street scene. (c) Knowledge-intensive queries require internet exploration and external data retrieval to identify specific locations based on minimal visual cues.

blur, scenarios where visual enhancement tools may be required to improve image quality for accurate processing. **Second**, tasks demand sophisticated reasoning capabilities to solve complex problems involving counting, projection, and measurement operations. **Third**, certain challenges require models to explore the internet and retrieve external data to answer questions based on minimal visual cues. Figure 1 illustrates representative examples of these challenging scenarios.

OWLViz introduces a benchmark for Agentic AI Assistant featuring 248 carefully annotated questions and answers where big proprietary VLM models failed to answer. OWLViz dataset is easy to understand, challenging for both human and AI and featuring extensive tool-use skills. Yet, evaluation on this dataset is simple and can be done automatically.

Despite their success in many visual grounding tasks, even the best-performing VLMs demonstrated surprisingly poor performance on our benchmark. Most models achieved less than 20% accuracy in exact-match evaluation and below 30% in LLM-match evaluation, indicating struggles with both precise answer generation and semantic understanding. In stark contrast, college students easily completed these tasks with 69.2% accuracy in under one minute, demonstrating that the questions assess intuitive reasoning abilities that humans possess but current AI systems lack. This significant performance gap reveals critical weaknesses in tool selection, multi-step planning, and information integration. OWLViz presents novel challenges and opportunities for advancing research in open-world visual understanding. We keep our dataset private to minimize data contamination. Additional information on how to access the data is available upon request.

2 RELATED WORK

VQA Dataset. Early visual question answering (VQA) datasets primarily evaluated fundamental visual understanding abilities such as object recognition and attribute identification [25, 31]. Subsequent benchmarks introduced more challenging reasoning tasks, including relational reasoning between objects [1] and compositional reasoning over spatial relationships [13, 17]. These datasets shifted the focus from simple recognition toward structured reasoning that requires integrating multiple visual cues and logical steps across a scene.

More recent benchmarks extend this setting by incorporating external knowledge and multimodal reasoning. For example, some datasets require commonsense knowledge beyond the visual content [26], while others combine textual and visual evidence to support reasoning [2, 7]. Despite these advances, existing VQA datasets still cover a relatively narrow range of real-world queries. In practical scenarios, users often ask questions that require multi-step reasoning, interaction with external tools, or retrieval of additional information beyond the image itself. Moreover, real-world visual inputs may be incomplete or degraded, which introduces challenges that remain largely underexplored in current benchmarks.

VLM Benchmarks. Recent advances in vision-language models (VLMs) have substantially improved multimodal understanding. Modern VLMs demonstrate strong performance across a range of visual perception tasks, from fine-grained object recognition to holistic scene understanding [11, 18, 38, 41]. Improvements in instruction understanding [16] and reasoning capabilities [5, 37] further enable models to perform multi-step inference over visual inputs. In addition, emerging tool-use capabilities [23, 28] allow VLMs to interact with external systems, gradually transforming

them from passive perception models into interactive problem-solving agents.

These advances have led to strong performance on several established multimodal benchmarks [10, 26, 27]. As a result, recent research increasingly explores evaluation settings that better reflect real-world applications and human-centered considerations such as reliability and alignment [8, 12, 15]. However, many existing benchmarks still assume controlled environments with predefined tools and clean visual inputs. Such assumptions limit their ability to evaluate autonomous systems that must coordinate heterogeneous tools, interpret ambiguous instructions, and operate under imperfect visual conditions.

Agentic Datasets. Recent benchmarks for tool-using agents commonly evaluate systems in curated environments with restricted tool sets [19, 24, 29]. These settings enable systematic benchmarking and reproducible evaluation, but they restrict the diversity of tools and interaction patterns available to agents. In practice, real-world systems must operate in environments where the space of available tools is large, heterogeneous, and constantly evolving, making tool selection and planning significantly more challenging.

More recent efforts move toward open-domain evaluation with broader tool access, such as GAIA [27]. While these benchmarks better approximate real-world assistant scenarios, many tasks still provide structured hints or intermediate guidance, which primarily evaluates tool integration rather than autonomous planning or tool discovery [28]. Consequently, they offer limited insight into how agents explore unknown tool spaces or adapt their strategies in open-ended environments.

Furthermore, most existing agentic datasets focus on text-based reasoning and rarely evaluate multimodal settings where visual understanding influences tool usage decisions. They also seldom assess robustness to degraded visual inputs or context-dependent tool selection. In addition, current benchmarks typically employ simplified success metrics based mainly on task completion accuracy, which may overlook practical trade-offs between accuracy, efficiency, and robustness that arise in real-world deployments. These limitations highlight the need for benchmarks that better evaluate multimodal agents operating in realistic and less constrained environments.

3 DATASET CREATION

OWLViz is designed to evaluate models' capabilities in comprehending image content and leveraging external tools to answer image-related questions. The dataset comprises 248 human-designed and annotated questions, each associated with an image and an unambiguous ground truth answer, allowing exact-match evaluation.

This section presents our methodology for constructing a dataset that more effectively evaluates Visual Question Answering (VQA) systems' performance on open-world visual queries. The dataset development process encompasses five principal phases: (1) systematic image acquisition, (2) open-world question design, (3) establishment of reasoning paths to ensure reliable answer derivation, (4) comprehensive data annotation, and (5) implementation of standardized answer formatting protocols.

3.1 Image Acquisition

Images were collected from a range of publicly accessible sources. For questions requiring only visual interpretation, we intentionally selected dense and detailed images to ensure sufficient visual complexity. For questions designed to prompt external searches, images were chosen to provide minimal but sufficient cues such as brand names, logos, or partial text while avoiding overexposure of information that could reduce the challenge of the task.

The five most frequent image sources are *redfin.com*, *istockphoto.com*, *zillowstatic.com*, *rdcpix.com*, *shutterstock.com*, and *pexels.com*. Together with 17 screenshots, these sources account for approximately half of the dataset. The 17 author-generated images were created in cases where direct image URLs were unavailable, such as screenshots taken from PDF reports, social media posts, or frames from YouTube videos.

3.2 Questions Design

The dataset shares several design principles with GAIA, including (1) targeting questions that are conceptually simple yet practically useful and challenging for contemporary AI systems, (2) ensuring interpretability, (3) maintaining robustness against memorization, and (4) facilitating ease of evaluation. However, OWLViz distinguishes itself in several key aspects: (5) exclusive focus on images and (6) concise and practical.

First, every question in OWLViz is directly linked to a photograph. Questions can be answered using (i) image content alone, (ii) a combination of image content and metadata embedded within the image, or (iii) an integration of image content, metadata, and external knowledge inferred from visual cues. Unlike other datasets where questions can be excessively lengthy and unlikely to reflect the way people naturally inquire about information, our dataset prioritizes practicality.

Second, OWLViz is designed to capture questions that closely mirror the types of inquiries people naturally make when interpreting images in everyday contexts. These include tasks such as counting objects, identifying locations or addresses, checking property costs, extracting key details, or interpreting relationships between elements in a scene. This emphasis on practical, context-driven questions ensures the dataset's relevance to real-world applications and everyday problem-solving scenarios.

We develop an initial set of skills that was incorporated into the annotation web interface; the skill list includes three main categories: visual skills, reasoning skills, and tool skills (See Table 1). During the annotation, we allow annotators to add new skills if needed, which results in some less frequent skills as shown in Figure 4.

3.3 Data Annotation

All questions in OWLViz were designed and annotated by the authors. Each question was carefully crafted and reviewed to ensure clarity, relevance, and alignment with a corresponding image. An exact-match answer was provided for each question, establishing a clear ground truth.

To ensure the quality, solvability, and objectivity of the dataset, the annotation process was divided into three distinct phases. In the first phase, one author was responsible for collecting the images and

Table 1: Taxonomy of skills required for our dataset, categorized into visual, reasoning, and tool-based capabilities. Each skill is accompanied by a functional description indicating the specific capability it represents in the context of visual question answering tasks.

	Skill	Description
Visual	human recognition	Identify humans and their locations in the image
	object recognition	Identity non-human objects and their location in the image
	identify 2 endpoints	Identify two specific endpoints of an object in an image
	object detection	Identifies what objects are present and where they are located by using a bounding box
	object segmentation	Provides pixel-level understanding of what and where all objects are
	spatial relationships	Understands 3D spatial relations between objects in the image.
	attribute identification	Specifies attributes of objects (e.g., color, shape, size)
Reasoning	object measuring	Measures object dimensions or distances in the image
	arithmetic calculation	Comprises operations such as Addition, Subtraction, Multiplication and Division
	counting	Handles counting queries such as 'how many...'
	comparison	Handles comparison queries
	logical operations	Handles logical operations such as AND (intersection) and OR (union) on sets
Tools	QR code scanning	Scans and decodes QR codes from the image
	OCR	Optical Character Recognition — extracts text and their locations in the image
	GUI	Handles graphical user interface-based tasks
	knowledge search	Searches for external knowledge beyond the image itself
	using external API	Uses an external API such as image enhancement and reading metadata.
	knowledge retrieval	Retrieve the necessary information in the metadata of images

constructing the initial set of questions. In the second phase, five other authors independently reviewed the questions and provided feedback to refine and clarify them where necessary. In the third phase, to promote annotation consistency and minimize bias, an internal tool was developed to randomly assign each question to at least two reviewers. Reviewers answered the questions without any additional context beyond the image and the question text.

Any question that could not be reliably answered by the reviewers without additional input from the original author was removed from the dataset. This three-phase process ensured that the final dataset includes only questions that are independently answerable and clearly grounded in the accompanying visual content.

3.4 Answer Format Standardization

We explicitly specify the format of the expected answer for each question. This includes defining whether the answer should be a *yes/no response*, a *multiple-choice selection*, or a *short exact-match answer*. For some questions, multiple answers are allowed. Acceptable answers are separated by semicolons. The exact output format instruction is provided in Appendix 3.2. By standardizing the answer format, we facilitate straightforward evaluation, enabling seamless testing of both exact-match answers and multiple-choice responses. This design ensures that the dataset is both practical and robust, catering to diverse testing and benchmarking needs while maintaining ease of implementation.

It is important to note that, in order to conform to an exact-match evaluation format, questions were transformed into constrained response types such as multiple choice, yes/no, single numerical values, or short text answers limited to a few words. While this approach facilitates consistent evaluation and benchmarking across models, it may also increase the likelihood of correct responses, as it narrows the range of possible outputs.

Preliminary experiments using free-form, open-ended answers reveal significantly higher failure rates, suggesting that exact-match formats may overestimate model performance by simplifying the response space. This is a trade-off between evaluation consistency and the complexity of real-world language understanding.

3.5 Difficulty Level

Following [27], we categorize questions into three levels of increasing difficulty based on the size of unique skills needed to answer the question. Figure 2 shows the a number of unique skills used per question.

We broadly define difficulty levels as follows:

- Level 1: Typically involves no more than 2 unique skills and at most 1 external tool.
- Level 2: Involves a greater number of skills - generally between 3 and 5 skills, typically includes a combination of two tools.
- Level 3: Designed for an ideal general-purpose assistant, these questions may require arbitrarily long sequences of actions, unrestricted use of tools, and general access to the whole Internet.

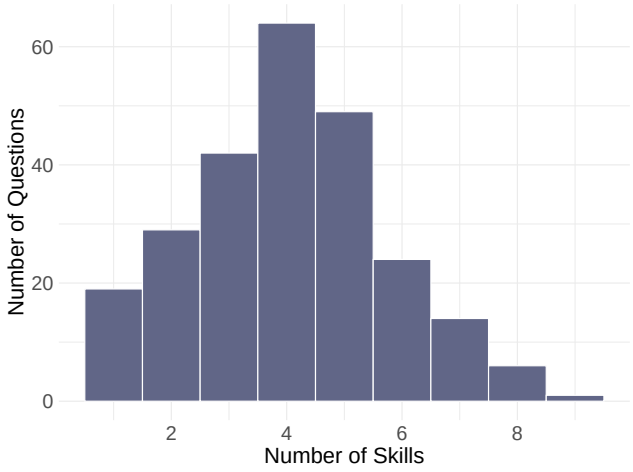


Figure 2: Number of unique skills used per question.

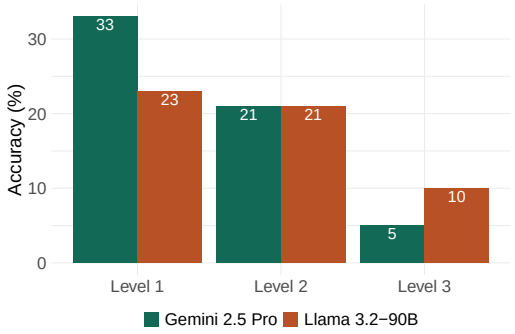


Figure 3: Model performance degradation across difficulty levels. Results show both Gemini 2.5 Pro and Llama 3.2-70B exhibit declining accuracy as task difficulty increases, illustrating current limitations in complex visual reasoning.

Figure 3 shows the performance of Gemini Pro and Llama 3.2 90B grouped by difficulty level. Overall, these VLMs' performance deteriorates, highlighting the persistent gap between current AI capabilities and human-like reasoning.

We also displays the distribution of skills needed that were generated by annotators in Figure 4 . Among visual skills, object detection, OCR, and spatial reasoning are most common. In reasoning skills, counting and knowledge retrieval dominate. For tool-use skills, knowledge search and GUI interaction are most frequent.

4 EXPERIMENTS

In this session, we assess the capabilities of three powerful methodological approaches on OWLViz: Vanilla VLMs, Tool-Calling Agents and GUI Agents. This comparative analysis allows us to identify and examine the key challenges presented by our dataset.

4.1 Evaluation Metrics

We employ two metrics to evaluate model performance: *Exact Match* (**EM**) and *LLM-based Match* (**LM**) [40]. The EM metric requires the

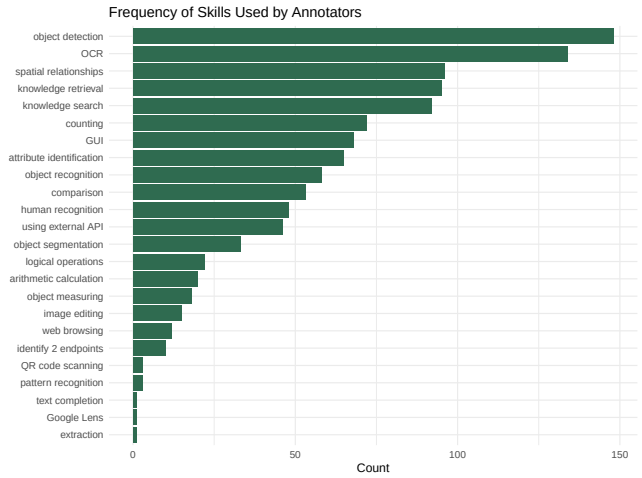


Figure 4: Distribution of skills annotated by human. Noted that annotators are allowed to add additional skills or tools that they used (e.g. Google Lens).

model's output to be identical to the ground truth answer, allowing for minor variations in capitalization and whitespace. In practice, some models still fail to follow output format instructions, leading to false negative errors. To allow for more flexible evaluation, we use LM, where GPT-4o acts as a judge to determine semantic equivalence between the predicted and ground truth answers (See Appendix 3.3).

4.2 Vanilla VLMs

Models. We evaluate a comprehensive set of vision-language models across three categories. For small open-source models (2.8B-17B parameters), we include DeepSeek-VL [34], a mixture-of-experts architecture designed for advanced multimodal understanding; Qwen-VL [35], which enhances vision-language perception at any resolution; InternVL [39], known for its strong performance on multimodal benchmarks; LLaVA [21], one of the pioneering open-source vision instruction tuning models; and Molmo [6], which provides open weights and training data for reproducible research. For large open-source models (32B-90B parameters), we evaluate scaled versions of Qwen-VL, InternVL, and Llama-3.2-Vision-Instruct, which generally offer improved capabilities through increased model capacity. Finally, we assess proprietary models including Anthropic's Claude-3.5-Sonnet, OpenAI's GPT-4V and GPT-4o representing different generations of multimodal capabilities, and Google's Gemini family (1.5-Pro, 2.0-Flash, 2.5-Pro) which demonstrates state-of-the-art performance on many vision-language tasks. This diverse selection allows us to evaluate how model scale, architecture choices, and training methodologies affect performance on our challenging benchmark.

Results. Table 2 presents our evaluation of VLMs across three categories: small open-source, large open-source, and proprietary models. The results demonstrate that OWLViz poses significant challenges even for state-of-the-art VLMs. The best-performing models, Gemini-2.5-pro-preview and Gemini-2.0-flash, achieve only

Table 2: Model performance breakdown into 3 groups ordered by EM performance. The best and second-best of each group are bolded and underlined, respectively.

	Model	EM	LM
	Human	**69.21**	
Small Open Source	DeepSeek-VL2-small (2.8B)	11.16	12.75
	DeepSeek-VL2 (4.5B active)	11.16	14.34
	Qwen2-VL-7B-Instruct	12.75	17.93
	Qwen2.5-VL-7B-Instruct	13.94	19.52
	InternVL3-8B	14.34	21.12
	LLaVa-v1.6-mistral-7B	14.74	15.54
	Llama-3.2-11B-Vision-Instruct	14.74	**25.10**
	InternVL2.5-8B	14.74	18.73
	LLaVa-v1.5-13B	16.33	16.33
	Molmo-7B-D-0924	17.13	20.32
	LLaVa-v1.5-7B	**18.33**	19.92
Large Open Source	Qwen2.5-VL-32B-Instruct	2.79	25.90
	InternVL2.5-38B	13.94	19.52
	InternVL3-78B	15.54	20.72
	Molmo-72B-0924	15.94	22.71
	InternVL2.5-78B	15.94	21.91
	InternVL3-38B	16.73	23.11
	Qwen2-VL-72B-Instruct	19.92	25.90
	Qwen2.5-VL-72B-Instruct	20.32	**26.29**
	Llama-3.2-90B-Vision-Instruct	**20.72**	24.70
Proprietary	Claude-3-5-sonnet-20241022	11.55	19.92
	GPT-4V	14.34	20.00
	Gemini-2.5-Flash	15.54	25.50
	GPT-4o (`2024-11-20`)	16.33	19.52
	Gemini-1.5-Pro	19.52	21.91
	Gemini-2.0-Flash	**21.51**	24.30
	Gemini-2.5-Pro	**21.51**	**27.09**

21.51% EM accuracy, indicating substantial room for improvement on this benchmark. This relatively low performance across all models suggests that our dataset effectively probes the limitations of current VLMs in handling complex visual reasoning tasks when not equipped with tools.

Within the open-source category, larger models generally perform better than their smaller counterparts, but the improvement is modest. For instance, Llama-3.2-11B-Vision-Instruct tops the small model leaderboard at 25.10% LM, its larger variant Llama-3.2-90B-Vision-Instruct yields a lower LM score (24.70%). This suggests that simply scaling up model size may not be sufficient to address the challenging nature of our benchmark. Among proprietary models, while Gemini-2.0-Flash sets the current state-of-the-art, its performance (21.51% EM, 27.09% LM) still falls significantly short of human-level understanding, highlighting the substantial gap between current AI capabilities and human visual reasoning abilities. Notably, the consistent gap between EM and LM scores across all models indicates that even when models grasp the correct concept, they often struggle to express it in the exact required format.

4.3 Tool-Calling Agents

Models. Tool-Calling Agents extend traditional VLMs by integrating external tools and action capabilities to solve complex visual reasoning tasks. Our evaluation examines six representative systems: LLaVa-Plus [23], which enhances LLaVA [21] with pretrained vision tools for improved reasoning; ViperGPT [33], which uses GPT-4o to generate Python code orchestrating multiple vision models; GPT4Tools [36], built on Vicuna with instruction tuning for visual tool control; HYDRA [14], employing deep reinforcement learning to fine-tune LLMs for dynamic visual reasoning; HF Agent includes predefined tools for visual tasks and web browsing; and DynaSaur [28], supporting real-time generation and composition of actions with capabilities for continual learning through action storage and reuse.

Table 3: Performance of Agentic models with tool-uses. The best and second-best of each group are bolded and underlined, respectively.

Model	MLLM	EM	LM
LLaVa-Plus	gpt-4o-2024-11-20	0.00	2.50
ViperGPT	gpt-4o-2024-11-20	7.56	12.35
GPT4Tools	vicuna-7b-v1.5	11.15	14.34
HYDRA	gpt-4o-2024-11-20	10.75	12.35
HF Agent	gpt-4o-2024-11-20	**18.32**	24.08
DynaSaur	gpt-4o-2024-11-20	16.23	**26.67**

Results. Table 3 presents the performance comparison of various large language models (LLM) agents equipped with tool-use capability. Among the models, HF Agent achieves the highest EM score (18.32%), indicating superior accuracy in exact task execution, followed closely by DynaSaur (16.23%). DynaSaur leads in LM score (26.67%), suggesting strong overall language understanding and generation capabilities. In contrast, LLava-Plus performs poorly across both metrics, particularly with an EM score of 0.00, highlighting limitations in task precision.

Table 4: Performance of LLM with tool-use

Model	EM
DynaSaur (w/o incentive)	9.0
DynaSaur (w/ incentive)	11.0

Tool Exploration Incentives. Our analysis reveals that the original tool-calling agent baselines demonstrate insufficient motivation to utilize external models or tools for visual question-answering tasks, instead defaulting to generating code comments for reasoning before directly producing answers. To mitigate this limitation, we implemented an explicit instruction requiring agents to employ at least one external tool for answer verification prior to submission. We subsequently conducted a comprehensive analysis of all libraries and machine learning models generated by DynaSaur across 100 sample instances. Figure 5 illustrates the comparative frequency of tool utilization between models with and without these incentives. This strategic modification yielded a 2% improvement in Exact Match (EM) score. For detailed results, readers may refer to Table 4 in the appendix.

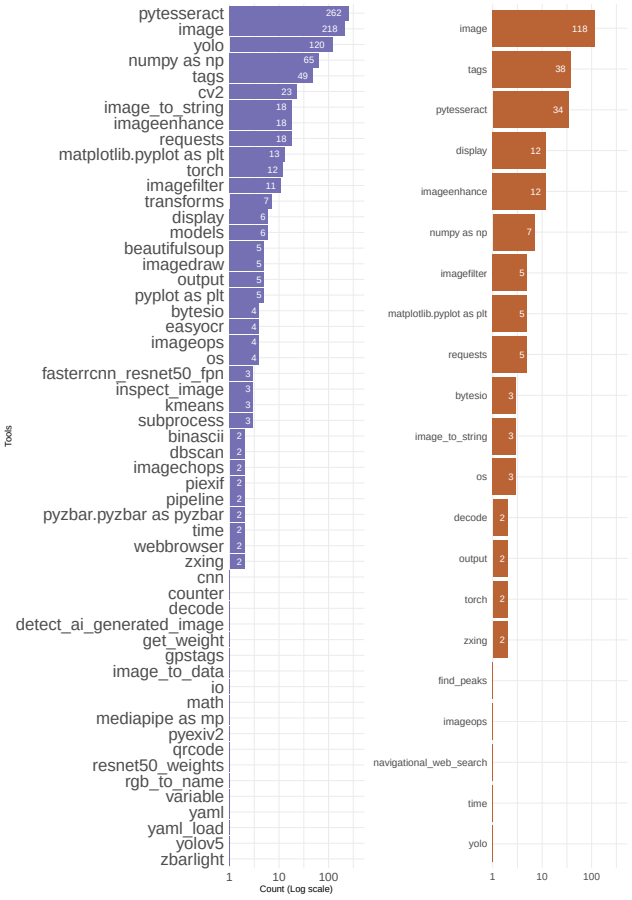


Figure 5: Comparison of libraries and ML models used by DynaSaur with (left) and without incentive (right).

In the no-incentive scenario (Figure 5), tool usage is more concentrated and limited in diversity, with image (118), tags (38), and pytesseract (34) being the most frequently used tools. Most other tools appear only a handful of times, reflecting a poor coverage of the tools being used. In contrast, models with incentives show both a significant increase in tool usage and a broader variety of tools. pytesseract leads with 262 mentions, followed by image (218), and YOLO (120). Additionally, a wide array of tools such as cv2, torch, transforms, and various deep learning components (e.g., fasterrcnn_resnet50_fpn, cnn) appear in the incentivized setting. This indicates that the presence of incentives not only increases effort but also encourages a broader adoption of more sophisticated, task-specific tools.

4.4 GUI Agents

Models. Our evaluation includes two GUI interaction baselines combining GPT-4o (2024-11-20) with specialized action determination components. UItars [30] pairs GPT-4o's reasoning capabilities with UI-TARS-2B-SFT for action determination in graphical interface tasks, enabling structured human-computer interactions. Similarly, ShowUI [20] employs the same GPT-4o model for reasoning but integrates ShowUI-2B for GUI action determination, facilitating efficient multimodal instruction-following in computer interfaces.

Results. Table 5 presents the performance of two GUI agents—UI-TARS and ShowUI—evaluated in task accuracy and the average number of mouse actions (Click, Hover, Scroll). Both agents fail to achieve any correct task as indicated by their EM scores of 0.00, highlighting a complete lack of task-following capabilities in output format. While LM scores are modest, with ShowUI slightly better than UI-TARS (12.80% vs. 12.31%).

Notably, the total number of actions performed by these agents is extremely low, reflecting minimal engagement with the interface. UI-TARS averaged fewer than 1 click (0.91), hover (0.51), and scroll (0.68) actions per task, while ShowUI executed slightly more clicks (0.97) but significantly fewer hovers (0.19) and scrolls (0.10). These numbers indicate that these GUI agents barely interact with the user interface. This somehow explains the low performance of the GUI agents, even when the model is allowed to explore freely.

Table 5: The performance in EM and LM of GUI Agents on OWLViz with the average number of mouse actions.

Model	EM	LM	Click	Hover	Scroll
UI-TARS	0.00	12.31	0.91	0.51	0.68
ShowUI	0.00	12.80	0.97	0.19	0.10

4.5 Comparison between orchestrations

In our evaluation of model performance across vision-language models (VLMs), agent-based systems, and GUI agents, we observe distinct capability profiles across these paradigms. Vision-language models, particularly proprietary systems such as Gemini-2.5-Pro, demonstrate the strongest overall performance, achieving the highest exact match (EM) and LLM match (LM) scores. Agent-based systems that incorporate tool use, such as HF Agent and DynaSaur powered by GPT-4o, show a marked improvement in EM over their base models (e.g., HF Agent achieves 18.32 EM vs. 16.33 for standalone GPT-4o), suggesting that tool augmentation enables more effective task execution and grounded reasoning. However, GUI agents such as UI-TARS and ShowUI exhibit significant limitations,

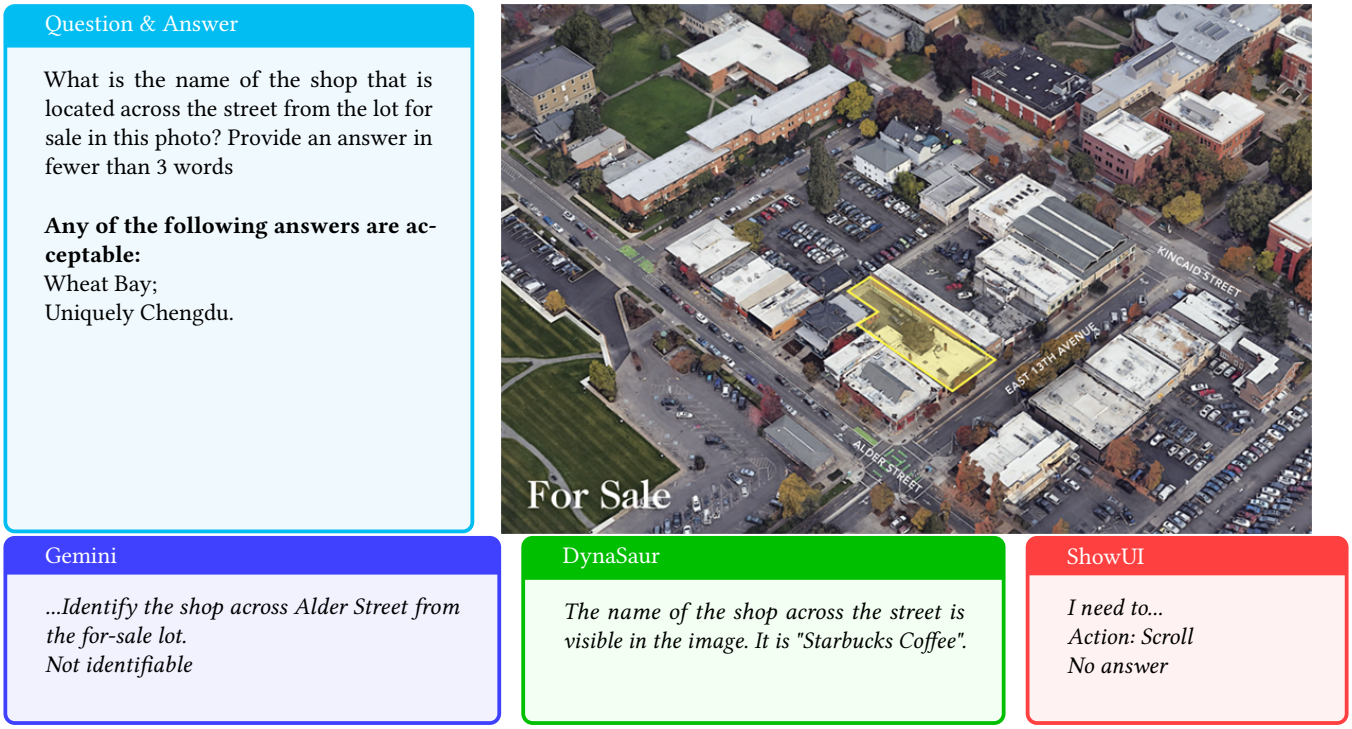


Figure 6: Qualitative results comparing different model capabilities on OWLViz. Results demonstrate varying capabilities across model types: Gemini (vanilla VLM) fails to identify the target, DynaSaur (tool-calling agent) produces an incorrect answer despite external search capabilities, and ShowUI (GUI agent) provides no answer.

with EM scores remaining at 0.00 and LM scores peaking at only 12.80%. Their low interaction metrics across click, hover, and scroll actions further highlight their current inability to perform even basic interface-driven tasks reliably.

4.6 Qualitative Example

We present a qualitative example that demonstrates the varying capabilities of different model types (see Figure 6). Please refer to Section 4 in the supplementary materials for detailed output.

- **Gemini**, representing vanilla VLMs, attempts to locate the "For Sale" lot but fails to identify the street name and the requested business, resulting in a conclusion "Not identifiable".
- **DynaSaur**, the tool-calling agent, leverages OCR and Search to identify the correct lot address ("821-825 E 13th Ave, Eugene, OR") and employs external search tools to find businesses in the vicinity. Through progressive refinement of search queries and cross-referencing, DynaSaur identifies "Starbucks Coffee" as the shop across from the lot. However, this is an incorrect answer, highlighting challenges in accurate spatial reasoning and external knowledge integration (e.g., maps with street view).
- **ShowUI**, representing GUI agents, demonstrates the limitations of restricted action capabilities. With only two available actions, the model attempts to scroll but ultimately fails to produce an answer. This illustrates how constraints on interaction modalities can significantly impact performance on complex visual reasoning tasks.

5 CONCLUSION

We introduced OWLViz, a challenging benchmark for evaluating AI models' visual understanding, reasoning, and tool use in practical, open-world scenarios. By integrating real-world tasks requiring image comprehension, metadata extraction, web exploration, and external tool use, OWLViz highlights the significant limitations of current multimodal AI systems across multiple dimensions of capability. Our comprehensive evaluation across vanilla VLMs, tool-use agents, and GUI agents demonstrates that these systems fall substantially short of human performance on intuitive tasks that humans complete easily and quickly.

Our results reveal that AI struggles with multi-step reasoning sequences, appropriate tool selection and coordination, and practical integration of heterogeneous capabilities required for real-world deployment. Even state-of-the-art models achieve less than 30% accuracy while humans reach nearly 70%, exposing fundamental gaps in current approaches to multimodal reasoning. These findings underscore the critical need for further advancements in developing agents that can autonomously navigate complex visual reasoning tasks, effectively leverage external tools when appropriate, and handle the degraded or ambiguous visual conditions common in real-world applications.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE Int'l Conf. on Computer Vision (ICCV)*.

[2] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595* (2023).

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818* (2023).

[4] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271* (2024).

[5] Long Hoang Dang, Thao Minh Le, Vuong Le, Tu Minh Phuong, and Truyen Tran. 2024. SADL: An Effective In-Context Learning Method for Compositional Visual QA. *CoRR* (2024).

[6] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146* (2024).

[7] Difei Gao, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2023. CRIC: A VQA Dataset for Compositional Reasoning on Vision and Commonsense. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5 (2023), 5561–5578. https://doi.org/10.1109/TPAMI.2022.3210780

[8] Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. 2024. FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI. *CoRR* (2024). https://doi.org/10.48550/ARXIV.2411.04872 arXiv:2411.04872

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conf. on Computer Vision and Pattern Recognition (CVPR)*.

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

[11] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2024. Visual Program Distillation: Distilling Tools and Programmatic Reasoning into Vision-Language Models. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 9590–9601. https://doi.org/10.1109/CVPR52733.2024.00916

[12] Zhe Hu, Yixiao Ren, Jing Li, and Yu Yin. 2024. VIVA: A Benchmark for Vision-Grounded Decision-Making with Human Values. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 2294–2311. https://doi.org/10.18653/v1/2024.emnlp-main.137

[13] Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[14] Brian Jalaian, Nathaniel D Bastian, et al. 2025. Hydra: An Agentic Reasoning Approach for Enhancing Adversarial Robustness and Mitigating Hallucinations in Vision-Language Models. *arXiv preprint arXiv:2504.14395* (2025).

[15] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. 2024. SWE-bench: Can Language Models Resolve Real-world Github Issues?. In *The Twelfth Int'l Conf. on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=VTF8yNQM66

[16] Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. FaithScore: Fine-grained Evaluations of Hallucinations in Large Vision-Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 5042–5063. https://doi.org/10.18653/v1/2024.findings-emnlp.290

[17] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conf. on computer vision and pattern recognition*. 2901–2910.

[18] Quang-Hung Le, Long Hoang Dang, Ngan Le, Truyen Tran, and Thao Minh Le. 2024. Progressive Multi-granular Alignments for Grounded Reasoning in Large Vision-Language Models. *arXiv preprint arXiv:2412.08125* (2024).

[19] Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023. API-Bank: A Comprehensive Benchmark for Tool-Augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3102–3116. https://doi.org/10.18653/v1/2023.emnlp-main.187

[20] Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Shiwei Wu, Zechen Bai, Weixian Lei, Lijuan Wang, and Mike Zheng Shou. 2024. ShowUI: One Vision-Language-Action Model for GUI Visual Agent. arXiv:2411.17465 [cs.CV] https://arxiv.org/abs/2411.17465

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning.

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

[23] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. 2025. Llava-plus: Learning to use tools for creating multimodal agents. In *European Conf. on Computer Vision*. Springer, 126–142.

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

[25] Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems* (2014).

[26] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf Conf. on computer vision and pattern recognition*. 3195–3204.

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

[28] Dang Nguyen, Viet Dac Lai, Seunghyun Yoon, Ryan A Rossi, Handong Zhao, Ruiyi Zhang, Puneet Mathur, Nedim Lipka, Yu Wang, Trung Bui, et al. 2024. DynaSaur: Large Language Agents Beyond Predefined Actions. *arXiv preprint arXiv:2411.01747* (2024). https://arxiv.org/pdf/2411.01747

[29] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2024. Gorilla: Large language model connected with massive apis. *Advances in Neural Information Processing Systems* (2024), 126544–126565.

[30] Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. 2025. UI-TARS: Pioneering Automated GUI Interaction with Native Agents. *arXiv preprint arXiv:2501.12326* (2025).

[31] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. *Advances in neural information processing systems* (2015).

[32] Alexis Roger, Esma Aïmeur, and Irina Rish. 2023. Towards ethical multimodal systems. *arXiv preprint arXiv:2304.13765* (2023).

[33] Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF Int'l Conf. on Computer Vision*. 11888–11898.

[34] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302* (2024).

[35] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115* (2024).

[36] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems* (2023), 71995–72007.

[37] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 9556–9567.

[38] Huaxiang Zhang, Yaojia Mu, Guo-Niu Zhu, and Zhongxue Gan. 2024. InsightSee: Advancing Multi-agent Vision-Language Models for Enhanced Visual Understanding. *CoRR* (2024). https://doi.org/10.48550/ARXIV.2405.20795 arXiv:2405.20795

[39] Linhai Zhang and Deyu Zhou. 2022. Temporal Knowledge Graph Completion with Approximated Gaussian Process Embedding. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo

Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 4697–4706. https://aclanthology.org/2022.coling-1.416/

[40] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* (2023), 46595–46623.

[41] Kaiwen Zhou, Kwonjoon Lee, Teruhisa Misu, and Xin Wang. 2024. ViCor: Bridging Visual Understanding and Commonsense Reasoning with Large Language Models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 10783–10795. https://doi.org/10.18653/V1/2024.FINDINGS-ACL.640