$\mathcal{R}$Bench-V: A Primary Assessment for Visual Reasoning Models with Multi-modal Outputs

Meng-Hao Guo[1], **Xuanyu Chu**[1], **Qianrui Yang**[1], **Zhe-Han Mo**[1], **Yiqing Shen**[1]

Pei-Lin Li[1], **Xinjie Lin**[1], **Jinnian Zhang**[2], **Xin-Sheng Chen**[1], **Yi Zhang**[1]

Kiyohiro Nakayama[3], **Zhengyang Geng**[4], **Houwen Peng**[2], **Han Hu**[2], **Shi-Min Hu**[1] *

[1] Tsinghua University, [2] Tencent Hunyuan Team, [3] Stanford University, [4] Carnegie Mellon University

Abstract

The rapid advancement of native multi-modal models and omni-models, exemplified by GPT-4o, Gemini and o3 with their capability to **process and generate** content across modalities such as text and images, marks a significant milestone in the evolution of intelligence. Systematic evaluation of their multi-modal output capabilities in visual thinking process (*a.k.a.,* multi-modal chain of thought, M-CoT) becomes critically important. However, existing benchmarks for evaluating multi-modal models primarily focus on assessing multi-modal inputs and text-only reasoning process while neglecting the importance of reasoning through multi-modal outputs. In this paper, we present a benchmark, dubbed as $\mathcal{R}$Bench-V, designed to assess models' vision-indispensable reasoning. To conduct $\mathcal{R}$Bench-V, we carefully hand-pick 803 questions covering math, physics, counting and games. Unlike problems in previous benchmarks, which typically specify certain input modalities, $\mathcal{R}$Bench-V presents problems centered on multi-modal outputs, which require image manipulation, such as generating novel images and constructing auxiliary lines to support reasoning process. We evaluate numerous open- and closed-source models on $\mathcal{R}$Bench-V, including o3, Gemini 2.5 pro, Qwen2.5-VL, etc. Even the best-performing model, o3, achieves only 25.8% accuracy on $\mathcal{R}$Bench-V, far below the human score of 82.3%, which shows current models struggle to leverage multi-modal reasoning. Data and code are available at `https://evalmodels.github.io/rbenchv`.

1 Introduction

"What I cannot create, I do not understand." — Richard Feynman.

Whether adults or children, when faced with complex problems, they sometimes turn to drawing or diagramming to organize their thoughts, support reasoning, and seek solutions. As highlighted by the quote 1 and findings in neuroscience Goldschmidt [1991], Pylyshyn [2001], Edwards [2012], Fan et al. [2023], the capability to draw during problem-solving is not only a hallmark of cognitive development but also an expression of human intelligence. But what about intelligent models? Can they also learn to draw in order to reason and solve problems?

Recently, researchers have made great progress toward equipping foundation models with above capability, and the landscape of foundation models has undergone a profound transformation, driven

*Shi-Min Hu is the corresponding author. E-mail: shimin@tsinghua.edu.cn.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

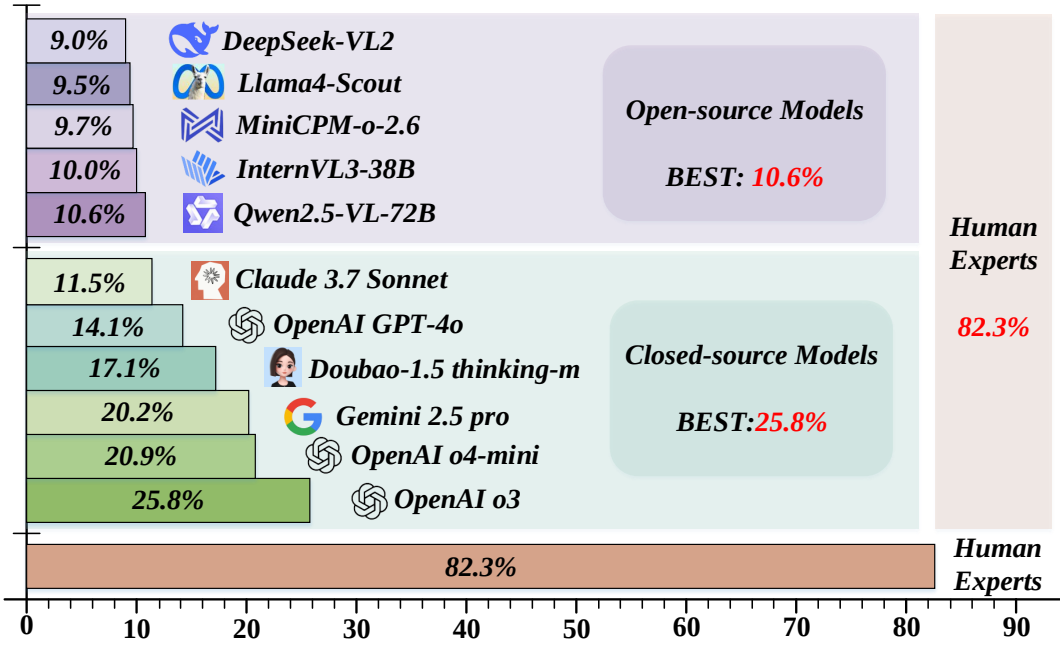

Figure 1: The comparison between open-source models, closed-source models and human experts on $\mathcal{R}$Bench-V. It reveals there remains a significant gap between models and human experts in visual reasoning with multi-modal outputs.

by two major converging trends. First, *modal convergence* reflects the evolution from single-modality language models, such as ChatGPT OpenAI [2022], to omni-modal systems capable of both multi-modal input and output, exemplified by GPT-4o OpenAI [2024a]. Second, *cognitive convergence* captures the transition from chat-oriented models to reasoning-driven models, as evidenced by the progression from ChatGPT OpenAI [2022] to more advanced systems such as OpenAI o1/o3 OpenAI [2024b] and DeepSeek R1 DeepSeek [2025].

As model input and output modalities converge, the evaluation frameworks for these leading foundation models must also evolve accordingly. Existing benchmarks such as MMMU Yue et al. [2024a] and MMLU Hendrycks et al. [2021], have played a key role in advancing the field by providing frameworks to evaluate model capabilities. However, these benchmarks are predominantly input-oriented, focusing on the model's ability to interpret, understand, and reason over multi-modal inputs, while overlooking an equally critical aspect: the modality of outputs. This refers to the model's ability to generate contextually appropriate multi-modal responses during the problem-solving process, whether through language, visual content, or other formats.

In this paper, we present $\mathcal{R}$Bench-V, an early exploration for multi-modal output-oriented reasoning benchmark. To build $\mathcal{R}$Bench-V, we carefully and strictly hand-pick and design 803 question-answer pairs, covering math, physics, counting, and games. In Fig 3, we clearly illustrate the differences between $\mathcal{R}$Bench-V and other classic language model benchmarks, MMLU Hendrycks et al. [2021], and the multi-modal input-oriented benchmark, MMMU Yue et al. [2024a]. It can be seen that the main difference from previous benchmarks is that in $\mathcal{R}$Bench-V, each question requires the model to produce multi-modal outputs during the reasoning process, particularly modifications on the images, such as drawing images, adding auxiliary lines, and so on.

We evaluate a wide range of open- and closed-source multi-modal large language models (MLLMs) and omni models on the $\mathcal{R}$Bench-V, including GPT-4o OpenAI [2024a], Gemini Google et al. [2023], Qwen2.5VL Bai et al. [2025], Claude3.5 Anthropic [2024], DeepSeek-VL2 Wu et al. [2024], *etc*. Besides, we also organize human to conduct tests on $\mathcal{R}$Bench-V. Our main observations and findings from the experiments are highlighted as follows. For more details, please refer to the experiment section.

- If models, such as the InternVL or Qwen-VL series, lack multi-modal CoT, merely increasing their model sizes will not effectively resolve the challenges of vision-indisperential reasoning.

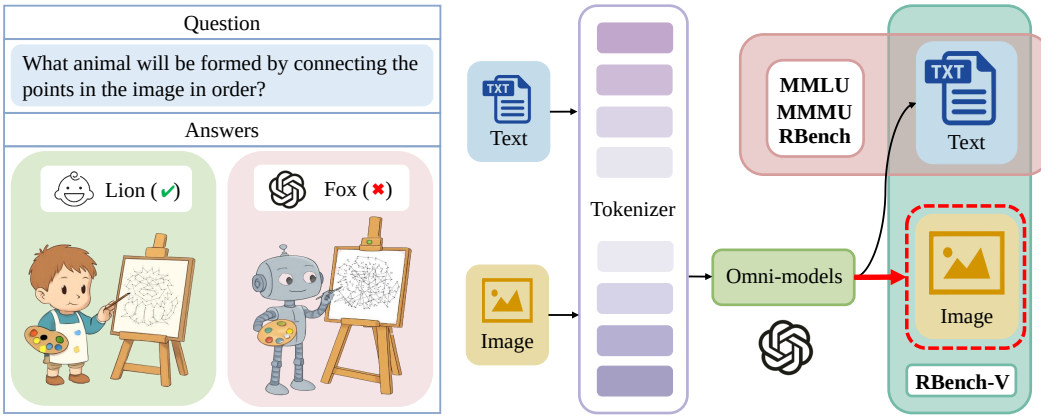

Figure 2: The motivation of $\mathcal{R}$Bench-V. Left: An illustration showing both humans and the GPT-4o model being asked a game-related question from $\mathcal{R}$Bench-V. Right: This part shows common benchmarks such as MMLU, MMMU, and Rench focus on multi-modal inputs and textual outputs, whereas $\mathcal{R}$Bench-V emphasizes not only multi-modal inputs but also multi-modal outputs.

It might be necessary to explore new paradigms, potentially incorporating M-CoT or agent-based reasoning frameworks, to solve visual-necessary complex problems.

- Despite the diverse capabilities of these models, even the highest-performing one, *i.e.*, o3 OpenAI [2025c], achieves only 25.8% accuracy on $\mathcal{R}$Bench-V, which is significantly lower than the human score of 82.3%. This stark contrast highlights that, while impressive in many areas, the models still struggle to generate and integrate appropriate multi-modal responses in the visual thinking process. Besides, the o3 model has achieved significant progress in visual reasoning, outperforming previous state-of-the-art models by a substantial margin (+4.9%). $\mathcal{R}$Bench-V effectively captures this advancement and offers an automated framework for evaluating multi-modal output capabilities in visual reasoning.
- The reasoning thought of models differ from that of human experts in math. We find that while models perform well on math questions, this does not necessarily indicate that they have learned to multi-modal reasoning. Instead, models often convert certain geometry problems into algebraic ones and solving them through textual reasoning. In contrast, human experts tend to prefer geometric solutions. This highlights a fundamental difference between the intelligence exhibited by current models and that of humans.

2 Related Work

2.1 Foundation Models

Foundation models have evolved rapidly along two axes: the expansion from language understanding to omni-models, and the progression from chat models to advanced reasoning models.

From language models to omni-models: It represents progress along the modality axis. Early foundation models—such as ChatGPT OpenAI [2022], LLaMA Touvron et al. [2023], Qwen Bai et al. [2023], Mistral Jiang et al. [2023], GLM Zeng et al. [2022], *etc*, are limited to text-based dialogue. Researchers then begin exploring models with multi-modal inputs, including GPT-4V OpenAI [2023], LLaVA Liu et al. [2024], miniGPT-4 Zhu et al. [2023], Claude3.5 Anthropic [2024], *etc*. Recently, attention has shifted toward omni-models, which not only receive multi-modal inputs but also generate flexible multi-modal outputs (e.g., GPT-4o OpenAI [2024a], Gemini 2.5 Pro Google [2025], Qwen2.5-Omni Xu et al. [2025] and Emu3 Wang et al. [2024b]).

From chat models to reasoning models: Another critical axis of advancement lies in reasoning capabilities. Early chat-based foundation models OpenAI [2022], Touvron et al. [2023], Abdin et al. [2024] primarily focus on fluent and context-aware dialogue generation. Recently, researchers have begun to push the boundaries of model reasoning OpenAI [2024b], DeepSeek [2025], Google [2025],

aiming to equip models with the ability to synthesize existing knowledge and solve complex, novel problems. For more, readers are referred to this survey Wang et al. [2025].

2.2 Benchmarking Foundation Models

Benchmarks serve as an essential tool for evaluating foundation models and providing a guiding light for their further development. Existing benchmarks primarily focus on multi-modal inputs such as text Hendrycks et al. [2021], Guo et al. [2025], Wang et al. [2024c], Chen et al. [2021], Contributors [2023] and multimodality Yue et al. [2024a,b], Lu et al. [2023], Wang et al. [2024a], Lu et al. [2023], Gao et al. [2024], but their outputs are all textual, overlooking the evaluation of models' multi-modal output capabilities. Besides, some studies Heusel et al. [2017], Kastryulin et al. [2022], You et al. [2024] have also focused on image generation quality, but they primarily emphasize aesthetic metrics.

As mentioned in Sec. 2.1, we believe that the next generation of powerful models are omni-models with strong textual and visual reasoning capabilities. Thus, in this paper, we present $\mathcal{R}$Bench-V, a benchmark for omni reasoning models. To the best of our knowledge, this is the first attempt to design benchmark to assess models' multi-modal generation capability in the visual thinking process.

3 $\mathcal{R}$Bench-V

3.1 Data Collection of $\mathcal{R}$Bench-V

The central challenge in developing $\mathcal{R}$Bench-V lies in designing and curating questions that assess models' ability to generate multi-modal outputs during visual reasoning. Clear examples are shown in Fig. 3, solving problems in $\mathcal{R}$Bench-V requires producing outputs beyond text, such as drawing geometric figures (top-right example) or tracing path through a maze (bottom-right example).

To build $\mathcal{R}$Bench-V, our principle in designing or collecting questions is that their solution should involve creating new visual content, such as creating images or modification of existing images, during the problem-solving process. We can imagine that numerous real-world scenarios, such as GUI operation and drawing, rely on multi-modal outputs for visual reasoning in daily life. In this work, $\mathcal{R}$Bench-V primarily focuses on math, physics, counting, and games. To curate high-quality questions in math and physics, we collaborate with domain experts. For counting and games, we conduct a rigorous rule to create, collect, and filter questions. The collection criteria for different domains are detailed as follows.

• **Math:** For math, we primarily focus on geometric and graph theory problems, including transformation geometry, planar geometry, solid geometry, *etc*. **Transformation Geometry:** The problems in $\mathcal{R}$Bench-V mainly involve translations, reflections and rotations, requiring the model to draw the resulting figures after applying these transformations. in order to arrive at the correct answer. **Planar Geometry:** These problems evaluate whether the model can construct appropriate auxiliary lines to aid in reasoning. **Solid Geometry:** The solid geometry tasks in $\mathcal{R}$Bench-V assess models to assemble 3D shapes from 2D components according to specific rules and to draw the resulting solid before answering the question. **Graph Theory:** $\mathcal{R}$Bench-V requires models to first complete the graph based on given constraints, mark the leaf nodes, and then reason to arrive at the correct answer.

• **Physics:** We primarily focus on optics, mechanics, electromagnetism, and thermodynamics. Not all above problems meet our criteria and we specifically select those that require visual reasoning. **Optics:** Tasks emphasize geometric optics, requiring models to trace light trajectories involving reflection, refraction, and diffraction. Precise visualization of light paths is essential for deriving optical principles. **Mechanics:** This category includes statics, kinematics, and dynamics, involving complex physical constraints. Models must interpret and construct geometric relationships using free-body diagrams and motion trajectories to analyze force interactions, motion paths, and equilibrium conditions. **Electromagnetism:** This area comprises two subcategories. Circuit analysis tasks require models to identify current paths and simplify circuit diagrams in complex scenarios. Dynamic problems combine electromagnetic fields with mechanics, necessitating the visualization of electric and magnetic field lines to analyze particle motion. **Thermodynamics:** Tasks primarily involve fluid force analysis, where models must visually represent dynamic changes in liquid surfaces and force distributions to solve problems related to surface tension, hydrostatic pressure, and buoyancy.

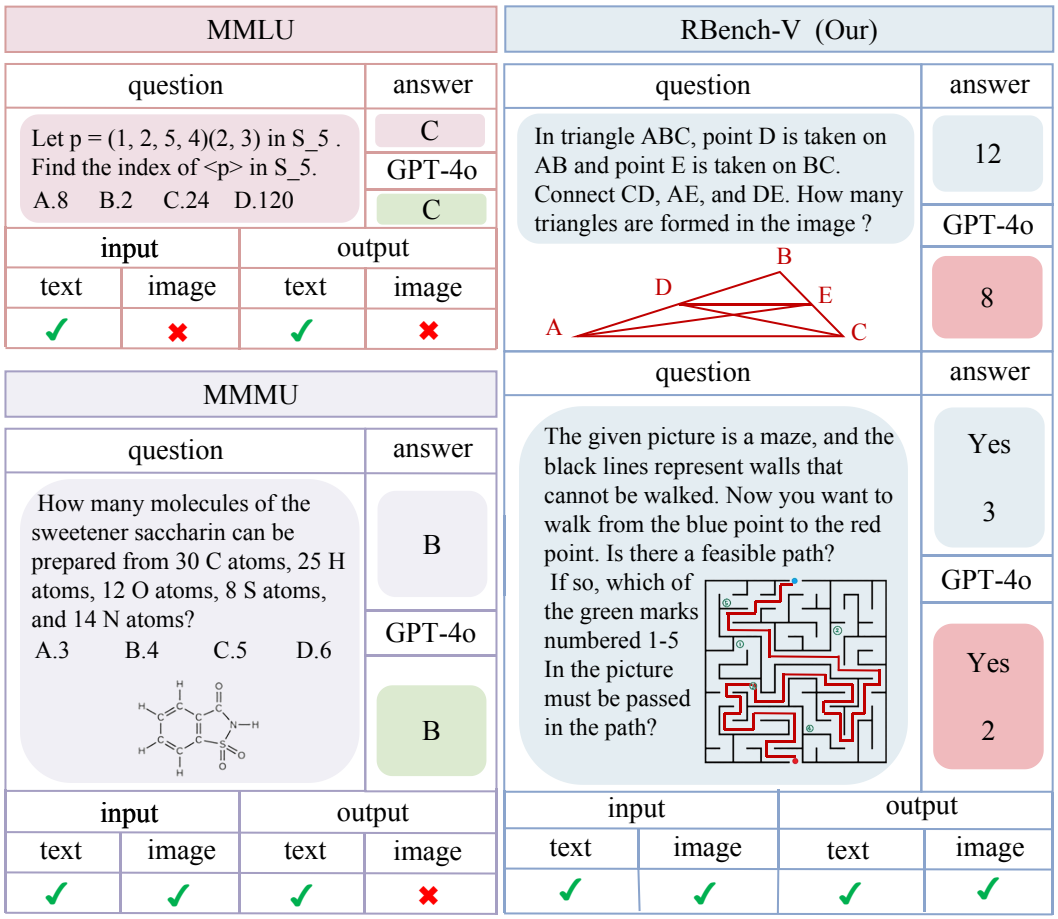

Figure 3: A visual comparison with MMLU, MMMU and $\mathcal{R}$Bench-V. It shows that solving problems in MMLU and MMMU mainly requires understanding multi-modal inputs and generating textual outputs, whereas solving problems in $\mathcal{R}$Bench-V demands not only understanding multi-modal inputs but also generating multi-modal outputs. The **red** lines shown in the figure are **not** part of the original questions and represent the multi-modal reasoning process when solving problems in $\mathcal{R}$Bench-V, such as drawing geometric shapes or tracing paths through a maze.

• **Counting:** Counting problems typically require no image modification. For instance, identifying two people in a clearly depicted photo needs no interaction. The counting problems in $\mathcal{R}$Bench-V differ from the simple example above and can be broadly categorized into the following three types: **Firstly,** problems require drawing geometric shapes based on descriptions or connecting lines within the image before answering questions such as how many triangles are present (an example is shown in the top-right corner of Fig. 3). **Secondly,** questions involve images with lots of targets and chaotic backgrounds. Solving such problems requires models to carefully check the images, mark the targets it has counted, and then reason to answer the total. (examples are provided in the supplementary materials.) **Thirdly,** problems demand an understanding of spatial relationships and imagination. Models need to mentally manipulate or move 3D objects and visualize the resulting state after movement, in order to complete the counting task.

• **Games:** We primarily focus on several types of games that require multi-modal outputs in the visual reasoning process: **Connect-the-dots**: Models need to connect a sequence of dots to reveal an image and identify what object in the image. **Mazes:** Models should trace the correct path through the maze and answer questions based on the trajectory. **Dart-and-balloon, and Gold Miner:** These require models to accurately draw the trajectory of darts and hooks, and determine their intersection with the target objects. **Puzzles:** The task involves moving different pieces to complete the full puzzle. **Ball-and-brick:** It requires drawing the trajectory of the ball, which may collide and bounce against the wall multiple times.

Table 2: Models and experts scores of multi-modal output requirements across different benchmarks.

(a) $\mathcal{R}$Bench-V vs. MMLU on text-only questions.

	$\mathcal{R}$Bench-V win	MMLU win	Tie
Model	92.4	3.8	3.8
Human	86.4	4.5	9.1

(b) $\mathcal{R}$Bench-V vs. MMMU on multi-modal questions.

	$\mathcal{R}$Bench-V win	MMMU win	Tie
Model	94.1	5.9	0.0
Human	83.4	6.6	10.0

3.2 Statistics of $\mathcal{R}$Bench-V

We conduct statistical analysis on $\mathcal{R}$Bench-V with the results presented in Tab. 1. It presents that $\mathcal{R}$Bench-V includes 803 questions across four areas, with 176 math questions, 157 physics questions, 195 counting problems, and 275 game-related questions, comprising 356 multiple-choice questions (We categorize questions with clearly limited answer choices as multiple-choice questions, such as the maze problem in the bottom right of Fig. 3.) and 447 open-ended questions. It is worth noting that since we primarily focus on multi-modal outputs rather than inputs, so $\mathcal{R}$Bench-V includes both text-only and multi-modal input questions. categorized as 40 and 763, respectively. As an early exploration into visual reasoning and multi-modal outputs, this paper focuses on text and image modalities, aiming to offer insights for foundation models. As for more modalities such as video and audio outputs, we expect more work to be done in the future.

Table 1: Statistics on $\mathcal{R}$Bench-V. MC denotes multiple-choice.

Statistics	**Number**
Total Questions	803
Math Questions	176
Physics Questions	157
Counting Questions	195
Games Questions	275
MC Questions	356
Open Questions	447
Text-only Questions	40
Multi-modal Questions	763

4 Experiments

After developing $\mathcal{R}$Bench-V, we comprehensively evaluate many open- and closed-source MLLMs (*e.g.,* Qwen2.5VL Bai et al. [2025], Claude3.7 Anthropic [2024], LLaVA-OneVision Li et al. [2024], *etc*) , omni-models (*e.g.,* GPT-4o OpenAI [2024a], Gemini2.5pro Google et al. [2024], Qwen2.5-Omni Xu et al. [2025], *etc*) , thinking models (*e.g.,* OpenAI o3 OpenAI [2025c] and Doubao-1.5-thinking-pro-m Seed [2025b]) and humam experts on $\mathcal{R}$Bench-V. We list all the tested models in Tab. 3.

By default, all evaluations are conducted in a zero-shot setting. Besides, since $\mathcal{R}$Bench-V includes both multiple-choice and open-ended questions, we adopt a unified LLM-as-a-Judge framework, with the judge model being GPT-4o. We report Top-1 accuracy (%) as our default evaluation metric.

4.1 Comparison with other benchmarks on multi-modal output capability requirements

Here, we compare $\mathcal{R}$Bench-V with other benchmarks (MMLU, MMMU) in terms of multi-modal output evaluation. As we know, it is challenging to find a quantitative method to assess the property for multi-modal outputs in existing benchmarks. Therefore, we construct expert scores and model scores to measure their multi-modal reasoning property. Specifically, we randomly sample 30 examples from each of $\mathcal{R}$Bench-V, MMLU, and MMMU, and instruct either human experts or the models (o3 and Doubao-1.5-thinking-m) to compare whether requires drawing during the thinking process.

We summarize the win rates in Table 2. Results from both human experts and models consistently show that $\mathcal{R}$Bench-V imposes a significantly higher requirement for multi-modal outputs during the reasoning process compared to MMLU and MMMU. This highlights that $\mathcal{R}$Bench-V is specifically designed to assess multi-modal output capabilities and visual reasoning skills.

4.2 Evaluating visual reasoning and multi-modal outputs of different models

We assess various open- and closed-source models, along with human experts, on $\mathcal{R}$Bench-V. The specific models are listed in Tab. 3. For open-source models, we use vLLM Kwon et al. [2023] and VLMEvalKit Duan et al. [2024] for deployment, setting the temperature as 0 while leaving all other

Table 3: Performance (%) of different models and human experts on $\mathcal{R}$Bench-V. † means long chain thinking model. * represents the omni-model. The highest scores are highlighted in **red**, and the second-highest scores are highlighted in **blue**.

Name	Overall	w/o Math	Math	Physics	Counting	Games
Open-source models						
Qwen2.5-Omni-7B* Xu et al. [2025]	7.7	4.5	11.4	1.9	2.1	7.7
InternVL-3-14B Zhu et al. [2025]	8.0	7.0	11.4	1.3	5.1	11.6
InternVL-3-8B Zhu et al. [2025]	8.2	6.0	15.9	1.9	5.6	8.7
Qwen2.5VL-7B Bai et al. [2025]	8.3	7.0	13.1	2.5	3.6	12.0
LLaVA-OneVision-7B Li et al. [2024]	8.5	6.8	14.2	2.5	4.6	10.9
DeepSeek-VL2 Wu et al. [2024]	9.0	7.0	15.9	0.6	5.6	11.6
LLaVA-OneVision-72B Li et al. [2024]	9.0	8.9	9.1	4.5	4.6	14.5
MiniCPM-2.6-V Yao et al. [2024]	9.1	7.2	15.9	1.3	6.2	11.3
Llama4-Scout (109B MoE) Meta [2025]	9.5	6.9	18.8	3.2	4.1	10.9
MiniCPM-2.6-o* Yao et al. [2024]	9.7	7.5	17.6	1.3	3.6	13.8
Qwen2.5VL-32B Bai et al. [2025]	10.0	6.4	22.7	2.5	4.1	10.2
InternVL-3-38B Zhu et al. [2025]	10.0	7.2	20.5	0.6	5.1	12.4
Qwen2.5VL-72B Bai et al. [2025]	10.6	9.2	15.3	3.8	6.2	14.5
Closed-source models						
QVQ-Max Qwen [2025]	11.0	8.1	21.0	5.7	6.2	10.9
Claude-3.7-sonnet Anthropic [2025]	11.5	9.1	19.9	3.8	8.7	12.4
OpenAI GPT-4.5 OpenAI [2025b]	12.6	11.0	18.2	2.5	11.8	15.3
Step-R1-V-Mini† StepFun [2025]	13.2	8.8	29.0	6.4	10.3	9.1
OpenAI GPT-4.1 OpenAI [2025a]	13.6	11.7	20.5	5.7	11.3	15.3
OpenAI GPT-4o-20250327* OpenAI [2024a]	14.1	11.2	24.4	3.2	13.3	14.2
Doubao-1.5-vision-pro Seed [2025a]	15.6	11.5	30.1	8.9	12.8	12.0
OpenAI o1† OpenAI [2024b]	16.2	11.0	34.7	5.7	12.3	13.1
Doubao-1.5-thinking-pro-m† Seed [2025b]	17.1	11.0	38.6	13.4	9.7	10.5
Gemini 2.5 pro-preview-0506 * Google [2025]	20.2	13.9	42.6	9.6	19.0	12.7
OpenAI o4-mini† OpenAI [2025c]	20.9	14.6	43.2	12.7	17.4	13.8
OpenAI o3† OpenAI [2025c]	**25.8**	**19.5**	**48.3**	**20.4**	**22.1**	**17.1**
Human Experts						
Human Experts Score	**82.3**	**81.7**	**84.7**	**69.4**	**81.0**	**89.1**

parameters at their default values. The above experiments are conducted on 8 x NVIDIA H20 GPUs. For closed-source models, we follow the official API usage guidelines provided for each model. If the official API allows setting the temperature parameter, we set it to 0; all other parameters are kept as recommended by the official documentation. For the human expert score, we invite senior undergraduate students to serve as human experts. For physics and math, we recruit some senior undergraduates for each major, assigning them different sets of questions. For games and counting tasks, we similarly invite some senior undergraduates, without restricting their academic backgrounds. The experimental results are summarized in Tab. 3, and a comprehensive analysis is presented in Sec. 4.4.

4.3 Visualization

Here, we visualize a correct example and an incorrect example: the correct case is from planar geometry in math, while the incorrect case is from games. The answer comes from the representative model, o3, the results are shown in Fig. 4.

From the results, we observe that although o3 arrives at the correct answer for the planar geometry question, its solution is based on an algebraic approach by establishing a coordinate system, rather than using a typical geometric method commonly adopted by humans. This suggests that the model tends to favor algebraic solutions with text-only reasoning thought over multi-modal geometric reasoning path when both approaches are available. It indicates that improvements in mathematical

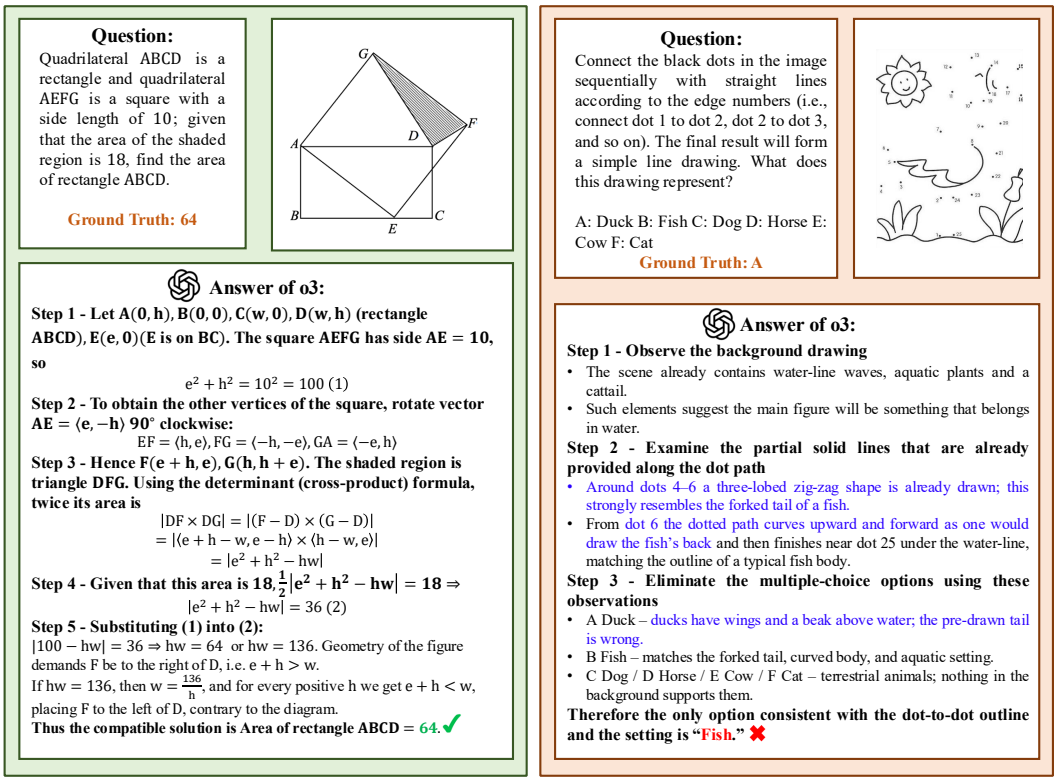

Figure 4: Examples of o3's responses to math and game questions in $\mathcal{R}$Bench-V. Left: o3 correctly answers a math question in $\mathcal{R}$Bench-V by transforming the geometry problem into an algebraic one using a coordinate system, whereas humans typically solve it using geometric methods. Right: o3 fails to answer a game question correctly. The blue highlights indicate the cause of the error and the key issue is that the model fails to follow the instructions to draw the required connections.

performance do not necessarily reflect genuine advancements in multimodal reasoning ability, but rather suggest that models have learned certain "multi-modal reasoning shortcuts". Experts in mathematics have validated this hypothesis, pointing out that most geometry problems can be solved using algebraic methods. In contrast, counting, and games do not exhibit such "multimodal reasoning shortcuts". Therefore, we also report the performance under the "w/o math" setting in Table 3, which may serve as a better indicator of a model's true multimodal reasoning capability.

In the second question, derived from a connect-the-dots task in the games category, o3 fails to generate a correct answer. Analysis reveals that the errors here mainly stem from o3 merely attempting to describe the points in the diagram, rather than actually connecting them as required by the question. Due to space limitations, we are unable to present more examples, but our analysis shows that the majority of model failures are caused by this limitation.

4.4 Observation and findings

If models, such as the InternVL or Qwen-VL series, lack multi-modal CoT, merely increasing their model sizes will not effectively resolve the challenge of visual reasoning. As shown in Tab. 3, increasing the parameter size of the Qwen2.5VL model from 7B to 72B does not result in a clear performance improvement on $\mathcal{R}$Bench-V. A similar phenomenon is also observed in the InternVL and LLaVA-OneVision series. It suggests that the scaling law may be insufficient to address the challenges of multi-modal output in visual reasoning. Furthermore, we question whether foundation models trained primarily via next-token prediction are inherently limited in their ability to handle such tasks. While this training paradigm is well-suited for text generation, it may be fundamentally inadequate for detailed and precise multi-modal generation and understanding such as precisely tracing curve trajectory in mazes.

Omni-models and long text-only CoT approaches also do not show significant improvements on this task. As shown in Table 3, a comparison between Qwen2.5VL-7B and Qwen2.5VL-Omni-7B, as well as between MiniCPM-V-2.6 and MiniCPM-o-2.6, indicates that simply incorporating images during output decoding, as done in omni-models, fails to effectively address the multimodal reasoning challenge. In addition, typical long text-only thinking models also show only marginal gains on $\mathcal{R}$Bench-V, as evidenced by the comparison between Double1.5-vision-pro and Doubao1.5-thinking-pro-m. Combining our analysis on scaling laws, omni-models, and long text-only CoT approaches, indicating that for current foundation models, novel techniques such as M-CoT and agents can be required to effectively solve visual reasoning tasks involving precise multimodal outputs.

Foundation models still fall well short of human expert performance in generating multi-modal outputs during visual reasoning. As shown in Table 3, even the best-performing model to date, o3, achieves only an overall accuracy of 25.8%, which remains significantly behind the human expert score of 82.3%. This substantial performance gap underscores the limitations of current foundation models in handling tasks that demand precise multi-modal outputs in the visual reasoning process. This phenomenon is clearly illustrated by the bar chart in Fig. 1. The results emphasize that, despite recent progress, there is still considerable room for advancement in multimodal reasoning.

The methods used by human experts and models to solve problems are not consistent. As shown in the Tab. 3, various models tend to perform best on the mathematics subject. We analyze and present representative examples from mathematics in Fig. 4, revealing that models often convert geometric problems into algebraic ones by constructing coordinate systems. This approach differs significantly from that of human experts, who typically solve such problems using geometric reasoning. It suggests that the intelligence of current models differs from that of humans. Therefore, to avoid such "multi-modal reasoning shortcuts," we also report the accuracy after removing math-related questions in Tab 3. The results show that excluding math further amplifies the performance gap between models and human experts.

OpenAI o3 has made substantial progress in visual reasoning with multi-modal output. The release of o3 attracted widespread attention, largely due to its impressive ability to handle complex visual reasoning tasks—a capability that has been challenging for previous models. We observed the same phenomenon in our proposed $\mathcal{R}$Bench-V, where o3 significantly outperformed all other models in tasks that require accurate and coherent multimodal outputs. This performance lead suggests that o3 has undergone deliberate and effective enhancements specifically aimed at improving its visual reasoning and output alignment capabilities. Notably, the results also validate the design of $\mathcal{R}$Bench-V itself. It demonstrates $\mathcal{R}$Bench-V can serve as a reliable benchmark for evaluating progress and tracking how models are evolving toward human-level multimodal reasoning.

Open-source models still lag far behind closed-source models. Although open-source models such as Qwen2.5VL Bai et al. [2025] and LLaMA 4 Meta [2025] are making continuous progress, there remains a noticeable gap (10.6% vs. 25.8%) between open-source and closed-source models in visual reasoning tasks that require multi-modal outputs. In addition, we find that the performance of current open-source models is quite similar, with low accuracy rates mostly ranging between 8% and 10%. It suggests that current open-source models exhibit only minimal capability in multi-modal reasoning. We hope the community will develop new techniques based on open-source models to enhance their multi-modal output capabilities and ultimately close the gap with closed-source models on $\mathcal{R}$Bench-V.

Text-based shortcuts remain a pervasive confound in assessing visual reasoning. Despite high aggregate scores, our analyses reveal that models can often bypass genuine visual reasoning by leveraging prior knowledge and algebraic reformulations, thereby inflating performance on ostensibly "visual" tasks. This also reflects a fundamental challenge in the field today: the persistent conflation of knowledge and reasoning. Many models solve problems through memorization rather than genuine reasoning. This issue is particularly evident in $\mathcal{R}$Bench-V, where most models lack the ability to draw or visualize solutions yet can still solve certain problem, strong evidence of this shortcut phenomenon. Meanwhile, this may also reflect the misalignment between the thinking patterns of foundation models and humans. foundation models tend to prefer solving problems through textual reasoning rather than visual representations, which highlights a key challenge currently faced by MLLMs.

5 Conclusion

In this work, we carefully hand-pick 803 questions across 4 topics and propose $\mathcal{R}$Bench-V, a benchmark specifically designed to evaluate models' multimodal output capabilities in the visual reasoning process. It systematically assesses the current performance of models, highlights the progress made by the o3 model in this domain, and reveals the significant gap between current intelligent models and human experts. Besides, according to our observation, the current technologies such as scaling law, long text-only CoT and joint text-visual decoding, fail to effectively address the challenges posed by $\mathcal{R}$Bench-V.

Looking ahead, we plan to further advance foundation models toward omni reasoning models, enabling more comprehensive and robust reasoning capabilities across modalities and tasks, and achieving stronger performance on $\mathcal{R}$Bench-V. In parallel, we will explore M-CoT (Multi-modal Chain-of-Thought) and agent-based strategies to enhance the reasoning depth and adaptability of these models. Furthermore, $\mathcal{R}$Bench-V is limited to text and image modalities. Future work will extend it to support richer output types, including audio, video, and other modalities, enabling more realistic and challenging multi-modal reasoning scenarios.

6 Acknowledgements

This work was supported by Beijing National Research Center for Information Science and Technology (BNRist). the National Natural Science Foundation of China (project No. 62495060, 623B2057), the Research Grant of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

References

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

Anthropic. Introducing claude 3.5 sonnet. *https://www.anthropic.com/news/claude-3-5-sonnet*, 2024.

Anthropic. Claude 3.7 sonnet. *https://www.anthropic.com/claude/sonnet*, 2025.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

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.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`, 2023.

DeepSeek. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024.

Betty Edwards. *Drawing on the right side of the brain: The definitive*. Penguin, 2012.

Judith E Fan, Wilma A Bainbridge, Rebecca Chamberlain, and Jeffrey D Wammes. Drawing as a versatile cognitive tool. *Nature Reviews Psychology*, 2(9):556–568, 2023.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*, 2024.

Gabriela Goldschmidt. The dialectics of sketching. *Creativity research journal*, 4(2):123–143, 1991.

Google. Gemini 2.5: Our most intelligent ai model. *https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking*, 2025.

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

Google, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Meng-Hao Guo, Jiajun Xu, Yi Zhang, Jiaxi Song, Haoyang Peng, Yi-Xuan Deng, Xinzhi Dong, Kiyohiro Nakayama, Zhengyang Geng, Chen Wang, Bolin Ni, Guo-Wei Yang, Yongming Rao, Houwen Peng, Han Hu, Gordon Wetzstein, and Shi min Hu. R-bench: Graduate-level multi-disciplinary benchmarks for llm & mllm complex reasoning evaluation, 2025. URL `https://arxiv.org/abs/2505.02018`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=d7KBjmI3GmQ`.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Sergey Kastryulin, Jamil Zakirov, Denis Prokopenko, and Dmitry V Dylov. Pytorch image quality: Metrics for image quality assessment. *arXiv preprint arXiv:2208.14818*, 2022.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

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

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.

Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *https://ai.meta.com/blog/llama-4-multimodal-intelligence/*, 2025.

OpenAI. Introducing chatgpt. *https://openai.com/index/chatgpt/*, 2022.

OpenAI. Gpt-4v(ision) system card. *https://openai.com/index/gpt-4v-system-card/*, 2023.

OpenAI. Gpt-4o system card. *https://openai.com/index/gpt-4o-system-card/*, 2024a.

OpenAI. Learning to reason with llms. *https://openai.com/index/learning-to-reason-with-llms/*, 2024b.

OpenAI. Introducing gpt-4.1 in the api. *https://openai.com/index/gpt-4-1/*, 2025a.

OpenAI. Introducing gpt 4.5. *https://openai.com/index/introducing-gpt-4-5/*, 2025b.

OpenAI. Openai o3 and o4-mini system card. *https://openai.com/index/o3-o4-mini-system-card/*, 2025c.

Zenon W Pylyshyn. Visual indexes, preconceptual objects, and situated vision. *Cognition*, 80(1-2): 127–158, 2001.

Qwen. Qvq-max: Think with evidence. *https://qwenlm.github.io/blog/qvq-max-preview/*, 2025.

Seed. Doubao-1.5-pro. *https://seed.bytedance.com/en/special/doubao_1_5_pro/*, 2025a.

Seed. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning, 2025b. URL https://arxiv.org/abs/2504.13914.

StepFun. Step-r1-v-mini. *https://www.stepfun.com/chats/new*, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024c.

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

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

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

Zhiyuan You, Jinjin Gu, Zheyuan Li, Xin Cai, Kaiwen Zhu, Chao Dong, and Tianfan Xue. Descriptive image quality assessment in the wild. *arXiv preprint arXiv:2405.18842*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024a.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024b.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's contributions—proposing RBench-V to evaluate visual reasoning with multi-modal outputs, and conducting systematic experiments. These claims are consistent with the rest of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: In the conclusion section, we discussed the limitation of our work and presented a challenge to the community, but did not yet propose a solution to this challenge.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This is a benchmark paper and does not include theoretical results or formal proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide code and data in abstract to reproduce our results.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and data can be access at `https://evalmodels.github.io/rbenchv`

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experiments section outlines key settings, including zero-shot evaluation, Top-1 accuracy, model deployment via vLLM and VLMEvalKit, and temperature settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Although Top-1 accuracy is reported, the paper does not include error bars or confidence intervals.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As mentioned in the experiment section, the experiments were conducted on 8 NVIDIA H20 GPUs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research does not involve sensitive populations or misuse and complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This study is a benchmark study and does not include sensitive data such as human faces.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The answer NA means that the paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper clearly attributes all third-party models and datasets in reference sections.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: RBench-V is a newly introduced dataset, and documentation and download links are provided via the project website.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Human experts participated in evaluations.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: Although the paper involves the participation of experts, it does not involve ethical issues.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: Our paper does not involve the use of large language models (LLMs) in any important, original, or non-standard component of the core methods.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.