BLINK-Twice: You see, but do you observe? A Reasoning Benchmark on Visual Perception

Junyan Ye[1,2], Dongzhi Jiang[3], Jun He[1], Baichuan Zhou[2], Zilong Huang[1], Zhiyuan Yan[4], Hongsheng Li[3], Conghui He[2], Weijia Li[1,†]
[1]Sun Yat-sen University, [2]Shanghai Artificial Intelligence Laboratory,
[3]CUHK MMLab, [4]Peking University
[†] Corresponding author: liweij29@mail.sysu.edu.cn

Abstract

Recently, Multimodal Large Language Models (MLLMs) have made rapid progress, particularly in enhancing their reasoning capabilities. However, existing reasoning benchmarks still primarily assess language-based reasoning, often treating visual input as replaceable context. To address this gap, we introduce BLINK-Twice, a vision-centric reasoning benchmark grounded in challenging perceptual tasks. Instead of relying on external knowledge, our tasks require models to reason from visual content alone, shifting the focus from language-based to image-grounded reasoning. Compared to prior perception benchmarks, it moves beyond shallow perception ("see") and requires fine-grained observation and analytical reasoning ("observe"). BLINK-Twice integrates three core components: seven types of visual challenges for testing visual reasoning, natural adversarial image pairs that enforce reliance on visual content, and annotated reasoning chains for fine-grained evaluation of the reasoning process rather than final answers alone. We evaluate 20 leading MLLMs, including 12 foundation models and 8 reasoning-enhanced models. BLINK-Twice poses a significant challenge to current models. While existing reasoning strategies in the language space—such as chain-of-thought or self-criticism can improve performance, they often result in unstable and redundant reasoning. We observe that repeated image observation improves performance across models, and active visual interaction, as demonstrated by models like o3, highlights the need for a new paradigm for vision reasoning. The dataset is publicly available at https://github.com/PicoTrex/BLINK-Twice.

1 Introduction

"You see, but you do not observe." — *Sherlock Holmes*

In recent years, Multimodal Large Language Models (MLLMs) built upon Large Language Models (LLMs) and advanced vision encoders have rapidly progressed. Both closed models like GPT-4o and Gemini, and open-source systems such as LLaVA, InternVL, and QwenVL, have demonstrated impressive visual perception, even surpassing human performance on certain tasks [36, 35, 50, 28, 63, 27, 52]. With the emergence of reasoning-augmented models such as OpenAI's o1 [37] and Deepseek-R1 [15], which leverage chain-of-thought (CoT) [47] and reinforcement learning techniques, reasoning has become a growing focus. Notably, this shift extends beyond language reasoning into the multimodal domain, as evidenced by the strong reasoning capabilities of models like Visual-RFT[30] and the recently released o3 [38].

To quantify the reasoning capabilities of MLLMs, the research community has proposed various multimodal reasoning benchmarks [59, 39, 58, 5, 17]. For instance, MMMU [57] evaluates models

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

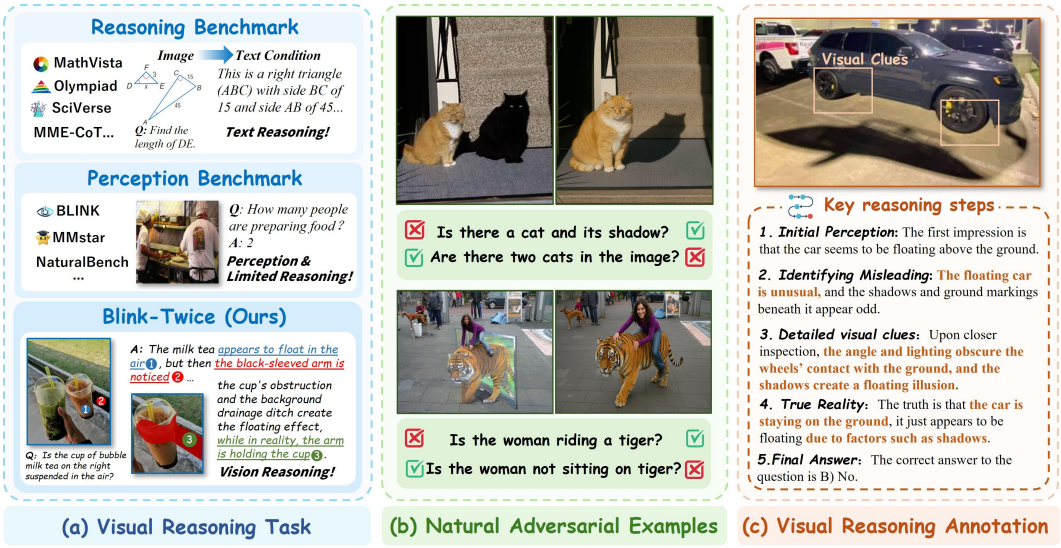

Figure 1: BLINK-Twice Task Overview: (a) Visual Reasoning task requiring detailed observation and careful reasoning; (b) Natural adversarial samples with similar appearance but opposite semantics, forcing models to rely on visual input; (c) Reasoning step annotation including detailed visual clues and true reality to evaluate thought chain output.

on college-level questions to assess knowledge-based reasoning; MathVerse [58] and OlympiadBench [18] focus on mathematical and physical reasoning challenges; MME-CoT [22] specifically targets chain-of-thought reasoning capabilities. However, most existing benchmarks remain centered on textual knowledge and logical reasoning, with visual input serving merely as auxiliary context—sometimes even replaceable by simple textual cues. These benchmarks rely more on the language model's knowledge and logical reasoning, while overlooking in-depth understanding and reasoning based on the visual content itself. Hence, there is an urgent need for a multimodal benchmark centered on vision-driven reasoning.

To this end, we propose the BLINK-Twice: A Reasoning Benchmark on Visual Perception. It starts from fundamental visual perception tasks, ensuring that answers depend on image content rather than prior knowledge or mathematical reasoning. By introducing more challenging visual perception and reasoning tasks, BLINK-Twice emphasizes the need to not only "see" but to truly "observe"—simple perception is no longer sufficient, and models must consciously attend to visual details and reason to fully comprehend the image. This aligns with the principle: "You see, but you do not observe." This turns visual perception tasks into a test of reasoning ability. This differs from previous perception benchmak such as BLINK [11] and NaturalBench [25], which primarily focus on direct perception tasks with limited reasoning requirements.

As illustrated in Figure 1, BLINK-Twice incorporates three key aspects: **i.** Our visual reasoning tasks span seven carefully curated **visual challenging** collection, such as visual dislocation, forced perspective, and motion illusion, enabling comprehensive evaluation of models' perception and reasoning capabilities; **ii.** Leveraging GPT-4o's powerful image editing capabilities, we construct **natural adversarial image pairs**—visually similar yet semantically distinct—forcing models to rely on detailed visual perception; **iii.** We provide **annotated reasoning chains and key detail scoring points**, enabling fine-grained analysis of reasoning quality and efficiency. Together, these designs establish BLINK-Twice as a strong framework for advancing the evaluation and development of multimodal reasoning systems beyond solely relying on final answer accuracy.

We systematically evaluate 20 leading MLLMs, including 12 foundation MLLMs and 8 reasoning-enhanced models incorporating chain-of-thought mechanisms. Our key findings are as follows:

- ***Current MLLMs exhibits potential flaw in visual reasoning, often "see" but fail to "observe".*** Even GPT-4o and Gemini-2.5 Pro perform suboptimally in both final answers (I-Acc, G-Acc) and reasoning processes (CoT-Score).

- ***Step-by-step reasoning or self-criticism helps reasoning models tackle complex visual reasoning challenges.*** QVQ, Claude-3.7-Thinking, and Gemini-2.0-Flash-Thinking outperform their base models by a notable margin.
- ***Reasoning models often overextend their reasoning chains, leading to redundanct reasoning.*** Efficiency analysis reveals that models like QVQ often over-criticize, producing redundant steps even after finding the correct answer.
- ***MLLM reasoning require new paradigms of visual reasoning, rather than relying solely on inference in the text space.*** Compared to single-pass perception followed by language reasoning, multi-turn dialogues with repeated image observation significantly enhance performance. The latest o3 model further demonstrates a novel paradigm through active visual reasoning via dynamic image cropping and transformation.

2 Related Work

Multimodal Large Language Models and Reasoning Model. In recent years, multimodal large language models (MLLMs) have achieved remarkable performance across both general tasks [28, 54, 42, 64] and specialized domains [48, 24, 55, 49, 16, 32]. Notable open-source models include LLaVA [28], and Qwen-VL [42], while closed-source counterparts such as GPT-4o [36] and Gemini [12] excel in advanced visual understanding and reasoning. With the release of o1 [37], the development of large models has increasingly shifted towards enhancing reasoning capabilities. DeepSeek R1 [15] and QwQ [41] enhance reasoning performance by generating intermediate reasoning steps, or Chains of Thought (CoT)[47], before final answers. Similarly, in the domain of MLLMs, early studies [60] explored adapting the CoT reasoning paradigm to vision-language tasks such as VQA and chart interpretation. Recent methods such as MM-EUREKA [34], Visual-RFT [30], and Skywork-R1V [40] employ large-scale rule-based RL to enhance multimodal reasoning [45, 21]. Meanwhile, closed-source models like Gemini [12, 13] and Claude [1, 2] have also introduced experimental versions designed specifically to enhance reasoning. OpenAI's latest o3 [38] model further advances image-level reasoning, enabling more meticulous observation and inference through operations like cropping and rotation. As multimodal reasoning models advance, effective benchmarks are needed to better evaluate their visual reasoning capabilities.

Multimodal & Reasoning Benchmarks. Meanwhile, numerous benchmarks have been proposed to evaluate the capabilities of MLLMs [29, 20, 56, 10, 9, 23, 33]. For instance, benchmarks such as BLINK [11], MM-Star [4], and NaturalBench [25] focus on assessing perception and understanding abilities through tasks like image captioning, counting, and basic spatial reasoning, effectively measuring fundamental visual recognition skills. In contrast, many benchmarks focus on evaluating the reasoning capabilities of MLLMs in domains such as expert knowledge, mathematics, and science [59, 39, 58, 5, 17, 19, 6]. For example, MMMU [57] uses college-level questions to assess models' mastery of expert knowledge and complex reasoning. MathVista [31] and OlympiadBench [18] further provide challenging problems in mathematics and physics to evaluate logical reasoning capabilities.

Compared to prior benchmarks focused on visual perception [11, 4, 25, 61, 10], BLINK-Twice moves beyond "See and Get" tasks by emphasizing reasoning and interpretation of image semantics. Unlike MathVista [31] and OlympiadBench [18], which focus on mathematical or scientific reasoning, our task design centers on visual reasoning. Most existing reasoning benchmarks treat visual input as auxiliary context for assisting reasoning occurring in the language domain. In some cases, visual input can even be replaced by textual cues, suggesting that it is not essential for reasoning. In contrast, BLINK-Twice highlights multimodal reasoning driven by image content. It also incorporates natural adversarial samples that compel the model to perform in-depth image analysis, along with detailed reasoning chain annotations to evaluate the quality, stability, and efficiency of reasoning—beyond mere answer accuracy. Additional dataset comparisons are in the supplementary materials.

3 Dataset

We introduce BLINK-Twice, a benchmark designed to evaluate models' visual reasoning capabilities. It contains 345 challenging base images across 7 types of visual challenges. These images were initially collected from over 650 samples across multiple internet platforms. Due to the benchmark's

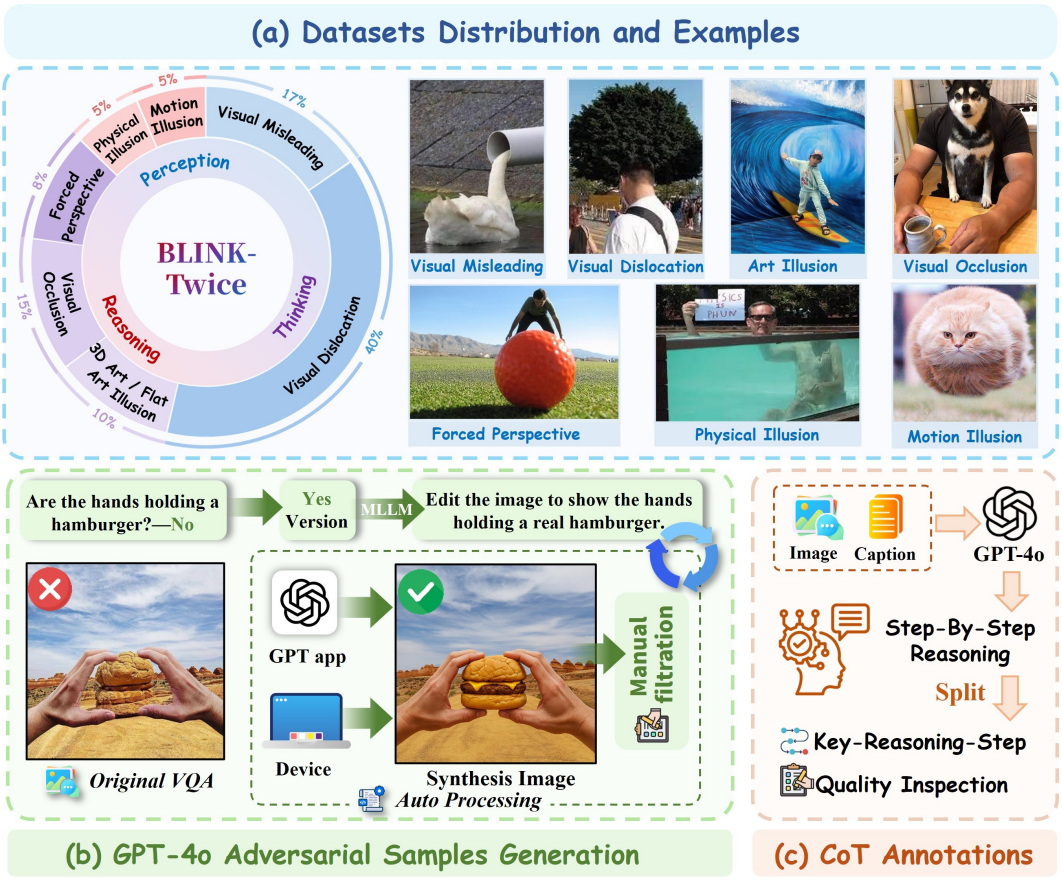

Figure 2: Overview of BLINK-Twice Dataset. (a) Distribution and examples of different visual challenges; (b) Pipeline for automatic adversarial sample generation and (c) reasoning chain annotation.

high requirements on visual ambiguity, scene diversity, and reasoning complexity, the data collection and filtering process was particularly demanding. Ultimately, only images that truly serve reasoning evaluation purposes were retained. Collection sources are detailed in the supplementary materials.

Additionally, leveraging the powerful image-editing capability of GPT-4o [36], we produce 103 natural adversarial samples. These samples are manually curated to ensure they are visually similar yet fundamentally different in factual content. To assess MLLMs' performance, we have curated 896 manually crafted VQA questions. Furthermore, the dataset includes 1,725 annotated reasoning step, generated through GPT-4o [36] and human-constructed prompts, highlighting two critical scoring aspects: detailed visual cues and true reality.

3.1 Visual Challenge Classification

BLINK-Twice comprises real-world, visually challenging images collected from the internet, categorized into seven fine-grained types, each representing distinct origins and mechanisms of visual misperception. Figure 2(a) shows the dataset distribution, with a brief description of each category provided below:

Visual Misleading: Errors in object recognition caused by coincidental alignments of color, shape, or composition — e.g., a swan's head in a pipe mistaken for a water splash due to color similarity.

Visual Dislocation: Spatial coincidence between foreground and background elements creates positional ambiguity — e.g., a man standing in front of a tree appears to have an exaggerated "afro" due to the alignment of the foliage with his head.

Art Illusion: Flat paintings or landscape art simulate three-dimensional effects — e.g., a boy standing on a painted surface appears to be surfing on ocean waves.

Visual Occlusion: Partial occlusion leads to identity or structural misjudgment — e.g., a dog blocking a man's face gives the illusion that the man is wearing a "dog head mask."

Forced Perspective: Manipulated camera angles and depth cues create size distortions — e.g., a golf ball close to the lens looks enormous, while a distant man appears to be holding it.

Physical Illusion: Natural physical phenomena such as reflection, refraction, or lighting distort visual interpretation — e.g., water refraction makes a man's head appear detached from his body.

Motion Illusion: Static images capture high-speed movement, creating a false sense of motion — e.g., a mid-air kitten appears to be floating due to its dynamic pose frozen at the peak of a leap.

3.2 Adversarial Samples Generation

To encourage image-grounded reasoning, we constructed natural adversarial samples, where each question pairs visually similar but semantically contrasting images with opposite answers, reducing reliance on commonsense and enhancing visual understanding. This setup discourages shortcut learning based on commonsense priors and compels the model to engage in genuine visual understanding. Unlike prior methods that rely on semi-automated CLIP-based filtering [25], we leverage the state-of-the-art image editing capabilities of GPT-4o to generate semantically altered yet locally consistent adversarial examples. This method not only simplifies data collection but also enables fine-grained and controllable semantic modifications tailored to specific reasoning challenges.

As illustrated in Fig. 2(b), we start with an original VQA sample (e.g., "Are the hands holding a hamburger? — No") and aim to generate a factual image yielding the opposite answer "Yes". We then utilize MLLM to interpret the image and question, producing structured editing instructions (e.g., region, object type, reference). For instance, "Edit the image to show the hands holding a real hamburger." Finally, GPT-4o is employed to perform image editing and synthesize the natural adversarial sample. Notably, since OpenAI's official image editing API was not available at the time, we used an automated scripting approach [51] to batch process and retrieve the edited images. Due to the strict image generation policies of OpenAI, we obtained only 243 initial samples, which were further manually filtered to retain only those meeting our quality and specification requirements.

3.3 Reasoning-step Annotation & Review

As shown in Fig. 2(c), we illustrate the reasoning annotation process. We start with human-labeled image facts (e.g., "The image looks like [A], but is actually [B]"), where [A] denotes the misleading appearance and [B] the ground truth, and then generate corresponding questions. GPT-4o is then guided to produce a step-by-step reasoning chain across five stages: Initial Perception, Identifying Misleading, Detailed Visual Clues, True Reality, and Final Answer. The "Detailed Visual Clues" and "True Reality" steps are defined as key reasoning steps and used to evaluate CoT scores. Finally, the annotated reasoning chains are manually reviewed for quality.

To mitigate potential hallucinations in GPT-4o's reasoning path annotations, we performed human validation for all samples in this task. Each response involving GPT was reviewed in at least two independent rounds. A total of five annotators participated in the verification process, with an additional 60 hours spent on this phase. In addition, all generated natural adversarial samples and their corresponding VQA questions are manually verified, taking approximately 50 hours in total.

4 Experiments

In this section, we evaluate a range of MLLMs, including both open and closed-source models, on our proposed benchmark. All evaluations are conducted under a zero-shot setting. We begin by introducing the evaluated models and our evaluation protocol. We then analyze the performance of existing MLLMs on challenging visual reasoning tasks, with a particular focus on their reasoning capabilities under our task configuration. Finally, we discuss potential directions toward advancing multimodal reasoning in future work.

Table 1: Evaluation results of open-and closed-source MLLM across multiple metrics. ☆ indicates CoT-enhanced variants; **bold** indicates the best result, and underlined denotes the second best.

Model	No-Acc	Yes-Acc	Q-Acc	I-Acc	G-Acc	CoT Score
Open-source MLLMs						
InternVL2-8B [8]	0.367	0.596	0.478	0.194	0.083	0.194
InternVL2-26B [8]	0.529	0.325	0.429	0.188	0.120	0.288
InternVL2-40B [8]	0.514	0.466	0.491	0.276	0.140	0.301
InternVL2.5-8B [7]	0.350	0.582	0.463	0.199	0.099	0.287
MM-Eureka-8B [34] ☆	0.319	0.610	0.461	0.176	0.078	0.285
Qwen2.5-VL-7B [53]	0.410	0.543	0.475	0.262	0.078	0.340
MM-Eureka-Qwen-7B [34] ☆	0.452	0.507	0.479	0.265	0.109	0.339
Qwen2-VL-72B [52]	0.372	0.614	0.491	0.233	0.061	0.341
QVQ-72B [45] ☆	0.517	**0.637**	0.575	0.336	0.067	**0.438**
Qwen-2.5-VL-32B [53] ☆	0.631	0.523	**0.578**	**0.353**	**0.158**	0.328
Qwen-2.5-VL-72B [53]	**0.653**	0.380	0.520	0.261	0.152	0.360
Closed-source MLLMs						
Claude-3.5-sonnet [1]	0.693	0.282	0.496	0.190	0.076	0.539
Claude-3.7-sonnet [2]	0.680	0.134	0.414	0.085	0.035	0.526
Claude-3.7-sonnet-thinking [2] ☆	0.717	0.274	0.502	0.189	0.101	0.536
Gemini-1.5-flash [44]	0.410	0.591	0.499	0.250	0.130	0.365
Gemini-2.0-flash [13]	0.360	0.694	0.525	0.242	0.071	0.469
Gemini-2.0-flash-thinking [13] ☆	0.503	0.583	0.542	0.353	0.156	0.470
GPT-4o [36]	0.616	0.523	0.571	0.351	0.198	**0.601**
o1 [37] ☆	0.710	0.503	0.608	0.392	0.186	–
Gemini-2.5-pro [14] ☆	**0.729**	**0.600**	**0.667**	**0.470**	**0.269**	0.584

4.1 Experimental Setup

Evaluated Models: For open-source models, we evaluate representative and strong MLLMs such as InternVL [8, 7, 3] and Qwen series [52, 53, 45], as well as the reasoning-oriented MM-EUREKA [34] model built upon them. For closed-source models, we include advanced commercial systems such as Claude [1, 2], Gemini [12, 13, 14], and the GPT [36, 37] series. In total, we evaluate 20 high-performing MLLM, including 8 models with dedicated CoT reasoning optimization. A complete list of evaluated models is provided in the supplementary material.

Evaluation Protocols: We evaluate model performance using accuracy on our constructed VQA tasks. Each image is typically associated with two binary questions: a main question (answered "no") and an adversarial one (answered "yes"), corresponding to No-Acc and Yes-Acc metrics, respectively. To comprehensively assess model behavior, we follow NaturalBench [25] to use Q-Acc (either question correct), I-Acc (both questions per image correct), and G-Acc (all four questions in a group correct) metrics. Additionally, inspired by the reasoning steps evaluation of large models [22, 46], we propose a CoT-score to assess reasoning chain quality, based on annotated reasoning steps and GPT-4o scoring. The score is grounded on two key points: identifying detailed visual cues (1 point) and inferring the true reality (1 point). Multiple valid reasoning paths are allowed; direct yet logically sound answers receive the full 2 points as well. The final score is normalized to the [0,1] range. Most reasoning models produce directly evaluable chains, while typical MLLMs require step-by-step prompting to elicit such reasoning. Further evaluation details are in the supplementary materials.

4.2 Challenge to MLLMs

As shown in Table 1, BLINK-Twice poses a significant challenge to current multimodal models. Among open-source models, reasoning-enhanced approaches such as Qwen-2.5-VL-32B and QVQ achieve the strongest results. Notably, the smaller 32B model, after reinforcement learning-based reasoning enhancement, performs comparably to or even surpasses the earlier 72B version. Although QVQ is built on the previous Qwen2-VL architecture, it still achieves competitive results due to its self-criticism mechanism. In the InternVL series, performance consistently improves with larger language models despite using the same InternViT-6B vision encoder, indicating the positive impact of model scale on visual understanding. Among proprietary models, Gemini-2.5 Pro and OpenAI

Figure 3: BLINK-Twice qualitative evaluation visualization results. The upper part shows the results of a single difficult image. The lower part displays the results for a group of images.

o1 perform best, with the former slightly outperforming the latter. Nevertheless, performance on more challenging metrics such as I-Acc and G-Acc remains suboptimal, with I-Acc below 0.5 and G-Acc under 0.3, highlighting the ongoing limitations of current multimodal systems in complex visual reasoning tasks. Current models often remain at the level of visual perception, lacking genuine understanding and reasoning over images.

Compared to answer-level accuracy (No-Acc), the CoT-score tends to be lower, suggesting that some correct answers may result from guesses rather than genuinely sound reasoning. Among open-source models, QVQ demonstrates relatively high-quality reasoning. Similarly, GPT-4o exhibits strong reasoning performance. Figure 3 presents visualized results from model testing. Only a few models, such as GPT-4o and Gemini-2.5 Pro, answer certain questions correctly, while others consistently fail. Even correct answers may overlook key visual cues—for example, missing distant vehicles that imply the true ground in the first question—resulting in a CoT Score of 1/2 rather than full credit. The CoT Score focuses on evaluating the reasoning process rather than the final answer alone. In the adversarial QA examples below, only GPT-4o and Claude-3.7 answered two questions correctly for the same image, and no model succeeded on all four. This highlights the stringent demands that I-Acc and Q-Acc place on both visual perception and reasoning.

4.3 Reasoning Models Analysis

As shown in Figure 4 (a), we compare the performance of various base MLLMs and reasoning-augmented models on BLINK-Twice. The results indicate that incorporating chain-of-thought reasoning significantly enhances performance on this visual reasoning benchmark. For instance,

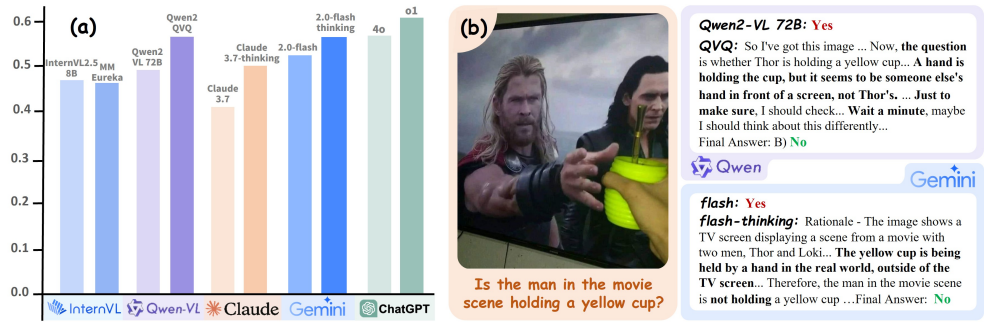

Figure 4: Comparison and Visualization of Reasoning Models and MLLMs.(a) Comparison of Q-Acc between reasoning models and based MLLMs. (b) The visualization illustrates the model's response process, highlighting step-by-step reasoning and self-critical chains of thought.

QVQ outperforms QwenVL2-72B by 15%, and the Claude-3.7 Thinking (16k) variant achieves a 20% improvement over its non-thinking counterpart. Similarly, Gemini-2.0-Thinking surpasses Gemini-2.0-Flash, and o1 also shows notable gains over GPT-4o. In contrast, MM-Eureka exhibits minimal improvement over InternVL2.5-8B, likely because its fine-tuning focuses primarily on mathematical reasoning, contributing less to general visual reasoning capabilities.

In Figure 4(b), we provide a qualitative visualization of reasoning processes. Compared to the base Qwen2-VL, the QVQ model performs step-by-step reasoning over the image and question, even without explicit prompts. For example, it first identifies the key visual cue: "A hand is holding the cup, but it seems to be someone else's hand in front of a screen, not Thor's.". Then, using phrases like "Just to make sure" or "Wait a minute," the model engages in self-refutation and reflective reasoning, ultimately providing a clear answer. Similarly, Gemini-2.0-Flash-Thinking also demonstrates a structured reasoning path to arrive at the correct answer.

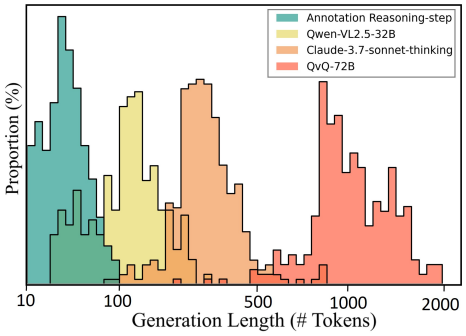

Figure 5: The distribution of generation length of Reasoning model.

While reasoning-augmented models achieve notable performance gains, such improvements often come at the cost of efficiency. As shown in Figure 5, we compare the length of annotated standard reasoning steps with the actual outputs from different models. The standard reasoning length generally remains below 100 tokens. Qwen-2.5-VL-32B, trained with chain-of-thought supervision, maintains efficient reasoning with an average output length of around 120 tokens. Claude-3.7-sonnet-thinking responds more slowly, with moderate efficiency. In contrast, QVQ consistently generates lengthy outputs—over 950 tokens on average—regardless of question difficulty or answer certainty, relying heavily on self-contradiction processes. While this improves accuracy, it also introduces significant redundancy. Future reasoning models should pursue more adaptive, selective strategies, akin to human reasoning, focusing effort only when necessary rather than indiscriminately extending the reasoning chain.

4.4 Multimodal Reasoning Paradigms

In prior reasoning model paradigms, the visual modality primarily serves for perception and feature extraction, while reasoning is predominantly driven by the language modality [26]. In such setups, the image is often encoded only once at the input stage—following a "see once" strategy, such as global feature extraction using CLIP [43]—and subsequent integration and logical inference rely entirely on the language model. However, for visual reasoning tasks, relying solely on language-based generation of intermediate reasoning steps may be limiting; repeatedly perceiving the image during reasoning can enhance the reliability of model decisions. To evaluate this, we design a multi-turn dialogue setup in which the model views the image twice before answering. As shown in Figure 6(a), models with weaker initial visual capabilities—such as Gemini-2.0-flash-thinking and Qwen2VL-72B exhibit notable performance improvements. In contrast, models with already strong visual grounding, such as

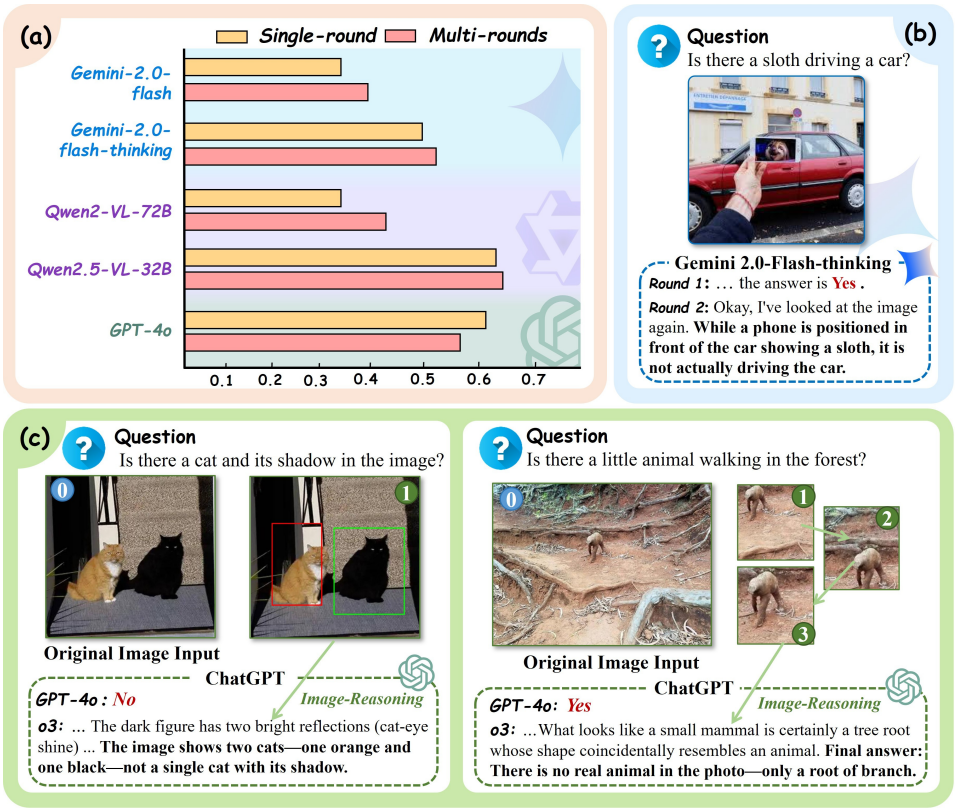

Figure 6: Multi-turn dialogue and multimodal reasoning. (a) shows the model's performance in single/multi-turn dialogue settings, (b) visualizes Gemini results in multi-turn dialogue, and (c) demonstrates the latest o3 model's ability to perform multimodal reasoning using image tools.

GPT-4o and QwenVL2.5-72B, show limited further gains. In Figure 6(b), Gemini-2.0-flash-thinking correctly identifies a raised phone only after a second observation.

As reasoning models continue to evolve, the visual modality should move beyond its passive role of perception and instead engage in collaborative reasoning with the language modality. In this paradigm, the visual module not only responds to language instructions but also initiates internal reasoning actions, such as invoking visual tools for image editing or transformation [62, 26], thereby forming explicit reasoning trajectories within the visual feature space. The recently released o3 model from OpenAI exhibits signs of such an evolution. As illustrated in Figure 6(c), the o3's reasoning process involves operations like "generating auxiliary bounding boxes" or "progressively zooming into image regions" as part of its internal deliberation. This suggests a promising trajectory for multimodal reasoning architectures, and underscores the value of our proposed BLINK-Twice dataset as a challenging and diagnostic benchmark for evaluating such capabilities. Further analysis and visualizations of multimodal reasoning are provided in the supplementary materials.

5 Conclusion

We introduce BLINK-Twice, a benchmark designed to systematically evaluate the performance of current multimodal large language models on complex visual perception and reasoning tasks. Experimental results indicate that current models exhibit notable limitations in visual perception and reasoning, often focusing on surface-level perception rather than truly understanding image content through reasoning. While Chain of Thought reasoning can partially improve performance on visual reasoning tasks, language-centric reasoning models remain dependent on the initial perception results and struggle with efficiency and redundancy issues. Future reasoning approaches are likely to evolve toward fully multimodal paradigms, such as repeatedly querying image perception or actively interacting with images for more robust visual reasoning. We hope BLINK-Twice will serve as a valuable benchmark for advancing research on perception and reasoning, and foster continued development in multimodal reasoning.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 62571560) and Shanghai Artificial Intelligence Laboratory.

References

[1] Anthropic: Claude-3.5. https://www.anthropic.com/news/claude-3-5-sonnet (2024)

[2] Anthropic: Claude-3.7. https://www.anthropic.com/news/claude-3-7-sonnet (2024)

[3] Cai, Z., Cao, M., Chen, H., Chen, K., Chen, K., Chen, X., Chen, X., Chen, Z., Chen, Z., Chu, P., Dong, X., Duan, H., Fan, Q., Fei, Z., Gao, Y., Ge, J., Gu, C., Gu, Y., Gui, T., Guo, A., Guo, Q., He, C., Hu, Y., Huang, T., Jiang, T., Jiao, P., Jin, Z., Lei, Z., Li, J., Li, J., Li, L., Li, S., Li, W., Li, Y., Liu, H., Liu, J., Hong, J., Liu, K., Liu, K., Liu, X., Lv, C., Lv, H., Lv, K., Ma, L., Ma, R., Ma, Z., Ning, W., Ouyang, L., Qiu, J., Qu, Y., Shang, F., Shao, Y., Song, D., Song, Z., Sui, Z., Sun, P., Sun, Y., Tang, H., Wang, B., Wang, G., Wang, J., Wang, J., Wang, R., Wang, Y., Wang, Z., Wei, X., Weng, Q., Wu, F., Xiong, Y., Xu, C., Xu, R., Yan, H., Yan, Y., Yang, X., Ye, H., Ying, H., Yu, J., Yu, J., Zang, Y., Zhang, C., Zhang, L., Zhang, P., Zhang, P., Zhang, R., Zhang, S., Zhang, S., Zhang, W., Zhang, W., Zhang, X., Zhang, X., Zhao, H., Zhao, Q., Zhao, X., Zhou, F., Zhou, Z., Zhuo, J., Zou, Y., Qiu, X., Qiao, Y., Lin, D.: Internlm2 technical report (2024)

[4] Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., Zhao, F.: Are we on the right way for evaluating large vision-language models? In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 27056–27087. Curran Associates, Inc. (2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/2f8ee6a3d766b426d2618e555b5aeb39-Paper-Conference.pdf

[5] Chen, Q., Qin, L., Zhang, J., Chen, Z., Xu, X., Che, W.: M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In: Proc. of ACL (2024)

[6] Chen, X., Zhang, R., Jiang, D., Zhou, A., Yan, S., Lin, W., Li, H.: Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. arXiv preprint arXiv:2506.05331 (2025)

[7] Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., Gu, L., Wang, X., Li, Q., Ren, Y., Chen, Z., Luo, J., Wang, J., Jiang, T., Wang, B., He, C., Shi, B., Zhang, X., Lv, H., Wang, Y., Shao, W., Chu, P., Tu, Z., He, T., Wu, Z., Deng, H., Ge, J., Chen, K., Zhang, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling (2025), https://arxiv.org/abs/2412.05271

[8] Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821 (2024)

[9] Feng, P., Lv, Z., Ye, J., Wang, X., Huo, X., Yu, J., Xu, W., Zhang, W., Bai, L., He, C., et al.: Earth-agent: Unlocking the full landscape of earth observation with agents. arXiv preprint arXiv:2509.23141 (2025)

[10] Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)

[11] Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)

[12] Gemini Team, G.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

[13] Gemini Team, G.: Gemini-2.0-flash-thinking (2025), https://deepmind.google/technologies/gemini/flash-thinking/

[14] Google, G.T.: Gemini 2.5 pro. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-pro (2025)

[15] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)

[16] Guo, Z., Lin, H., Yuan, Z., Zheng, C., Qiu, P., Jiang, D., Zhang, R., Feng, C.M., Li, Z.: Pisa: A self-augmented data engine and training strategy for 3d understanding with large models. arXiv preprint arXiv:2503.10529 (2025)

[17] Guo, Z., Zhang, R., Chen, H., Gao, J., Gao, P., Li, H., Heng, P.A.: Sciverse. https://sciverse-cuhk.github.io (2024), https://sciverse-cuhk.github.io/

[18] He, C., Luo, R., Bai, Y., Hu, S., Thai, Z., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., Liu, J., Qi, L., Liu, Z., Sun, M.: OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3828–3850. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.211, https://aclanthology.org/2024.acl-long.211/

[19] He, J., Lin, Y., Huang, Z., Yin, J., Ye, J., Zhou, Y., Li, W., Zhang, X.: Urbanfeel: A comprehensive benchmark for temporal and perceptual understanding of city scenes through human perspective. arXiv preprint arXiv:2509.22228 (2025)

[20] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020)

[21] Jiang, D., Guo, Z., Zhang, R., Zong, Z., Li, H., Zhuo, L., Yan, S., Heng, P.A., Li, H.: T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. arXiv preprint arXiv:2505.00703 (2025)

[22] Jiang, D., Zhang, R., Guo, Z., Li, Y., Qi, Y., Chen, X., Wang, L., Jin, J., Guo, C., Yan, S., et al.: Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In: Forty-Second International Conference on Machine Learning (2025)

[23] Jiang, D., Zhang, R., Guo, Z., Wu, Y., Lei, J., Qiu, P., Lu, P., Chen, Z., Song, G., Gao, P., et al.: Mmsearch: Benchmarking the potential of large models as multi-modal search engines. arXiv preprint arXiv:2409.12959 (2024)

[24] Kang, H., Wen, S., Wen, Z., Ye, J., Li, W., Feng, P., Zhou, B., Wang, B., Lin, D., Zhang, L., et al.: Legion: Learning to ground and explain for synthetic image detection. arXiv preprint arXiv:2503.15264 (2025)

[25] Li, B., Lin, Z., Peng, W., Nyandwi, J.d.D., Jiang, D., Ma, Z., Khanuja, S., Krishna, R., Neubig, G., Ramanan, D.: Naturalbench: Evaluating vision-language models on natural adversarial samples. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 17044–17068. Curran Associates, Inc. (2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/1e69ff56d0ebff0752ff29caaddc25dd-Paper-Datasets_and_Benchmarks_Track.pdf

[26] Lin, Z., Gao, Y., Zhao, X., Yang, Y., Sang, J.: Mind with eyes: from language reasoning to multimodal reasoning. arXiv preprint arXiv:2503.18071 (2025)

[27] Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al.: Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. ECCV 2024 (2023)

[28] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)

[29] Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)

[30] Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785 (2025)

[31] Lu, P., Bansal, H., Xia, T., Liu, J., yue Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. ArXiv **abs/2310.02255** (2023)

[32] Lu, Z., Ren, H., Yang, Y., Wang, K., Zong, Z., Pan, J., Zhan, M., Li, H.: Webgen-agent: Enhancing interactive website generation with multi-level feedback and step-level reinforcement learning (2025), https://arxiv.org/abs/2509.22644

[33] Lu, Z., Yang, Y., Ren, H., Hou, H., Xiao, H., Wang, K., Shi, W., Zhou, A., Zhan, M., Li, H.: Webgen-bench: Evaluating llms on generating interactive and functional websites from scratch (2025), https://arxiv.org/abs/2505.03733

[34] Meng, F., Du, L., Liu, Z., Zhou, Z., Lu, Q., Fu, D., Shi, B., Wang, W., He, J., Zhang, K., et al.: Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. arXiv preprint arXiv:2503.07365 (2025)

[35] OpenAI: GPT-4V(ision) system card (2023), https://openai.com/research/gpt-4v-system-card

[36] OpenAI: Hello gpt-4o. https://openai.com/index/hello-gpt-4o/ (2024)

[37] OpenAI: Introducing openai o1, 2024. (2024), https://openai.com/o1/

[38] OpenAI: introducing-o3-and-o4-mini, 2025. (2025), https://openai.com/index/introducing-o3-and-o4-mini/

[39] Peng, T., Li, M., Zhou, H., Xia, R., Zhang, R., Bai, L., Mao, S., Wang, B., He, C., Zhou, A., et al.: Chimera: Improving generalist model with domain-specific experts. arXiv preprint arXiv:2412.05983 (2024)

[40] Peng, Y., Wang, X., Wei, Y., Pei, J., Qiu, W., Jian, A., Hao, Y., Pan, J., Xie, T., Ge, L., et al.: Skywork r1v: pioneering multimodal reasoning with chain-of-thought. arXiv preprint arXiv:2504.05599 (2025)

[41] Qwen: Qwq-32b: Embracing the power of reinforcement learning. (2025), https://qwenlm.github.io/blog/qwq-32b/

[42] Qwen Team: Qwen2-vl (2024)

[43] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021), https://api.semanticscholar.org/CorpusID:231591445

[44] Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., Mariooryad, S., Ding, Y., Geng, X., Alcober, F., Frostig, R., Omernick, M., Walker, L., Paduraru, C., Sorokin, C., Tacchetti, A., Gaffney, C., Daruki, S., Sercinoglu, O., Gleicher, Z., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (2024), https://arxiv.org/abs/2403.05530

[45] Team, Q.: Qvq: To see the world with wisdom (December 2024), https://qwenlm.github.io/blog/qvq-72b-preview/

[46] Wang, J., Yuan, L., Zhang, Y., Sun, H.: Tarsier: Recipes for training and evaluating large video description models. arXiv preprint arXiv:2407.00634 (2024)

[47] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)

[48] Wen, S., Ye, J., Feng, P., Kang, H., Wen, Z., Chen, Y., Wu, J., Wu, W., He, C., Li, W.: Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. arXiv preprint arXiv:2503.14905 (2025)

[49] Xiao, H., Wang, G., Chai, Y., Lu, Z., Lin, W., He, H., Fan, L., Bian, L., Hu, R., Liu, L., et al.: Ui-genie: A self-improving approach for iteratively boosting mllm-based mobile gui agents. arXiv preprint arXiv:2505.21496 (2025)

[50] Yan, Z., Lin, K., Li, Z., Ye, J., Han, H., Wang, Z., Liu, H., Lin, B., Li, H., Xu, X., et al.: Can understanding and generation truly benefit together–or just coexist? arXiv preprint arXiv:2509.09666 (2025)

[51] Yan, Z., Ye, J., Li, W., Huang, Z., Yuan, S., He, X., Lin, K., He, J., He, C., Yuan, L.: Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. arXiv preprint arXiv:2504.02782 (2025)

[52] Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Fan, Z.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024)

[53] Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024)

[54] Ye, J., Jiang, D., Wang, Z., Zhu, L., Hu, Z., Huang, Z., He, J., Yan, Z., Yu, J., Li, H., et al.: Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. arXiv preprint arXiv:2508.09987 (2025)

[55] Ye, J., Lin, H., Ou, L., Chen, D., Wang, Z., Zhu, Q., He, C., Li, W.: Where am i? cross-view geo-localization with natural language descriptions. arXiv preprint arXiv:2412.17007 (2024)

[56] Ye, J., Zhou, B., Huang, Z., Zhang, J., Bai, T., Kang, H., He, J., Lin, H., Wang, Z., Wu, T., et al.: Loki: A comprehensive synthetic data detection benchmark using large multimodal models. arXiv preprint arXiv:2410.09732 (2024)

[57] Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., 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. pp. 9556–9567 (2024)

[58] Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.W., Gao, P., et al.: Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? ECCV 2024 (2024)

[59] Zhang, R., Wei, X., Jiang, D., Zhang, Y., Guo, Z., Tong, C., Liu, J., Zhou, A., Wei, B., Zhang, S., et al.: Mavis: Mathematical visual instruction tuning. arXiv preprint arXiv:2407.08739 (2024)

[60] Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)

[61] Zhou, B., Yang, H., Chen, D., Ye, J., Bai, T., Yu, J., Zhang, S., Lin, D., He, C., Li, W.: Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10707–10715 (2025)

[62] Zhou, Q., Zhou, R., Hu, Z., Lu, P., Gao, S., Zhang, Y.: Image-of-thought prompting for visual reasoning refinement in multimodal large language models. arXiv preprint arXiv:2405.13872 (2024)

[63] Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)

[64] Zong, Z., Ma, B., Shen, D., Song, G., Shao, H., Jiang, D., Li, H., Liu, Y.: Mova: Adapting mixture of vision experts to multimodal context. arXiv preprint arXiv:2404.13046 (2024)

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Please refer to the abstract in the main 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: Please refer to related section in the supplementary material.

   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 work do not involve assumptions.

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: Please refer to Section 4 in the main paper.

   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: The data we used are all from public datasets, and we commit to making our code publicly available. Please refer the abstract in the manuscript.

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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Section 4.1 in the main paper.

   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: This research performs adequate experiments without reporting error bars.

   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: Please refer to related section in the supplementary material.

   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: Please check our paper.

   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: [NA]

    Justification: There is no societal impact of the work performed.

    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: This 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: We ensure that all papers and datasets used or relevant to this work are properly cited.

    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: Please refer to Section 3 of the main paper and the link to the dataset presented in the abstract.

    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: Please refer to related section in the supplementary material.

    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: [NA]

    Justification: Our work does not involve research with human subjects.

    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: [NA]

Justification: LLMs are used solely for language polishing, checking grammar. They do not influence the core methodology, scientific rigor, or originality of the research.

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.