



Seek-and-Solve: Benchmarking MLLMs for Visual Clue-Driven Reasoning in Daily Scenarios

Anonymous ACL submission

Abstract

Daily scenarios are characterized by visual richness, requiring Multimodal Large Language Models (MLLMs) to filter noise and identify decisive visual clues for accurate reasoning. Yet, current benchmarks predominantly aim at evaluating MLLMs’ pre-existing knowledge or perceptual understanding, often neglecting the critical capability of reasoning. To bridge this gap, we introduce DailyClue, a benchmark designed for visual clue-driven reasoning in daily scenarios. Our construction is guided by two core principles: (1) strict grounding in authentic daily activities, and (2) challenging query design that necessitates more than surface-level perception. Instead of simple recognition, our questions compel MLLMs to actively explore suitable visual clues and leverage them for subsequent reasoning. To this end, we curate a comprehensive dataset spanning four major daily domains and 16 distinct subtasks. Comprehensive evaluation across MLLMs and agentic models underscores the formidable challenge posed by our benchmark. Our analysis reveals several critical insights, emphasizing that the accurate identification of visual clues is essential for robust reasoning.

1 Introduction

In recent years, we have witnessed a significant flourish of multimodal large language models (MLLMs) (OpenAI, 2025; Comanici et al., 2025; Shuai Bai et al., 2025; An et al., 2025; Zhu et al., 2025). These models exhibit strong capabilities in understanding and reasoning, allowing humans to readily employ them as comprehensive encyclopedic sources for knowledge retrieval or versatile creative assistants for content generation. The rapid development of MLLMs has consequently stimulated the emergence of diverse benchmarks (Wang et al., 2025d; Liu et al., 2024b; Yue et al., 2024, 2025; Wang et al., 2024), through which the capability limitations of existing models can be system-

atically evaluated and revealed.

Existing general visual question answering (VQA) benchmarks (Liu et al., 2024a; Zhang et al., 2024b; xAI, 2024) primarily focus on simple, explicit factual queries—such as object counting or recognition—that probe only the basic perceptual capabilities (Li et al., 2025). Although the emergence of GPT-o3 has inspired a wave of “think-with-image” agents (Zheng et al., 2025b; Zhang et al., 2025b; Wang et al., 2025b,a; Li et al., 2025; Zhou et al., 2025) and datasets (SeekWorld, 2025; Wang et al., 2025d; Wu and Xie, 2024; Jiang et al., 2025; Wang et al., 2025a) capable of utilizing external tools (e.g., zooming, code execution). However, these models still face significant challenges in daily scenarios characterized by visual distractions. In such complex environments, accurate reasoning depends on identifying the correct visual clues amidst the noise. Unfortunately, current benchmarks fail to sufficiently evaluate this critical ability.

Moreover, performance on existing benchmarks is reaching a saturation point (exceeding 90% accuracy) (Zheng et al., 2025b; Wu and Xie, 2024; Wang et al., 2025d), which makes it increasingly difficult to differentiate between top-tier models. This necessitates a more challenging benchmark, specifically tailored to rigorously assess deep reasoning capabilities in daily-life contexts.

In this work, we introduce **DailyClue**, a comprehensive benchmark designed to evaluate MLLMs’ understanding of daily-life scenarios and reasoning grounded in visual clues. DailyClue is guided by two core principles: (1) all scenes are strictly derived from real-world contexts; (2) questions are non-trivial and cannot be answered through direct visual inspection, instead necessitating deduction from implicit visual clues. To construct the benchmark, we employ a carefully controlled generation and filtering pipeline. First, top-tier MLLMs generate candidate question-clue-answer triplets. Then,

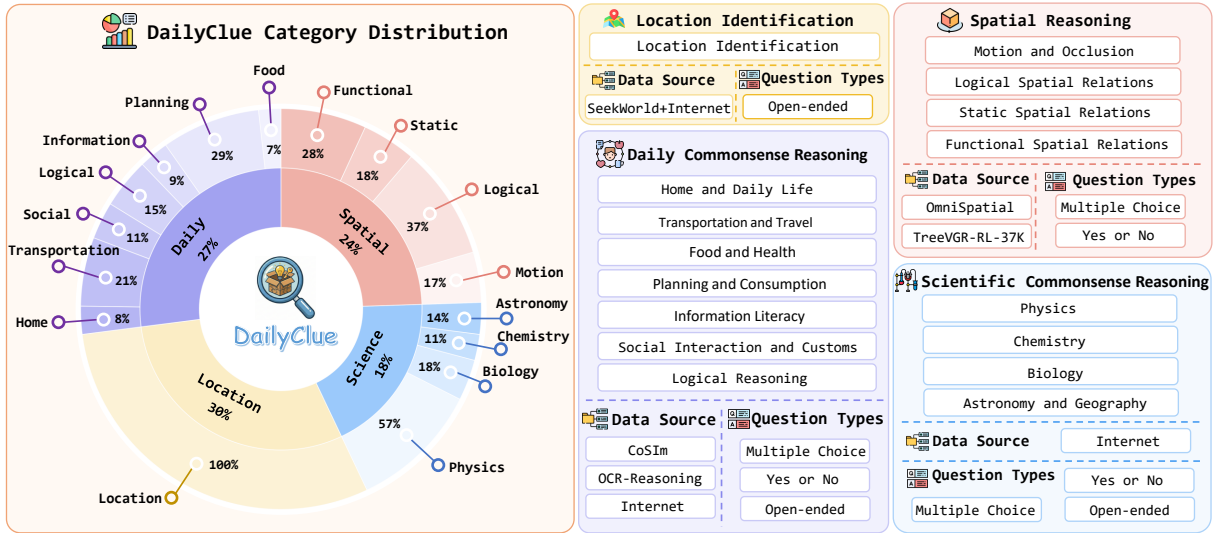


Figure 1: Overview of DailyClue. The left panel shows the hierarchical distribution (labels abbreviated for clarity), while the right panels detail the full taxonomy, data sources, and question types.

these candidates are verified by multiple peer models, ensuring that only instances remaining consistently challenging are retained. Following rigorous human verification, the final benchmark comprises 666 curated triplets spanning four broad categories and sixteen distinct subtasks, covering multiple-choice, open-ended QA, and binary decision tasks. We extensively benchmark 25 MLLMs on DailyClue, deriving three critical insights:

- DailyClue poses a significant challenge to both humans and top-tier MLLMs, serving as a robust and discriminative testbed for assessing holistic comprehension of daily scenarios.
- Model performance is critically bottlenecked by inaccurate visual clue prediction during CoT reasoning; notably, injecting ground-truth clues yields substantial gains.
- Explicitly prompting models to actively seek visual clues improves accuracy, yet this benefit exhibits diminishing returns as models' intrinsic reasoning capabilities scale up.

2 Related Work

2.1 Benchmarks for MLLMs

Existing MLLM benchmarks cover diverse categories, including general VQA (Liu et al., 2024a; xAI, 2024; Cheng et al., 2025) as well as domain-specific question answering tasks such as physics reasoning (Chow et al., 2025; Xiang et al., 2025), mathematical problem solving (Wang et al., 2024;

Zhang et al., 2024a; Lu et al., 2023), spatial reasoning (Yu et al., 2025; Jia et al., 2025), color reasoning (Liang et al., 2025), and document understanding (Liu et al., 2024b; Mathew et al., 2021, 2022). Specifically, benchmarks centered on physics and mathematics primarily assess models' mastery of discipline-specific knowledge, whereas those focused on color and spatial reasoning evaluate their ability to perceive fundamental object attributes and spatial relations. Collectively, benchmarks spanning these diverse domains have played an important role in diagnosing capability limitations of MLLMs from different perspectives.

2.2 Benchmarks for Agentic Models

Following the release of o3, several think-with-image models (Zheng et al., 2025b; Hong et al., 2025; Xu et al., 2025; Jiang et al., 2025; Wang et al., 2025b,a; Zhou et al., 2025; Zhang et al., 2025b; Zhao et al., 2025) and benchmarks (Wu and Xie, 2024; Wang et al., 2025d; SeekWorld, 2025; Li et al., 2025) have emerged. As summarized in Table 1, many existing benchmarks prioritize basic perceptual abilities (Wu and Xie, 2024; Wang et al., 2025d), such as attribute recognition in high-resolution images, where targets require zooming to resolve. Other benchmarks focus on vertical domains: SeekWorld (SeekWorld, 2025) targets the identification of the specific location; TIR-Bench (Li et al., 2025) evaluates multi-disciplinary perception; and MME-RealWorld (Zhang et al., 2024b), despite utilizing real-world data, also incorporates some niche scenarios like remote sensing.

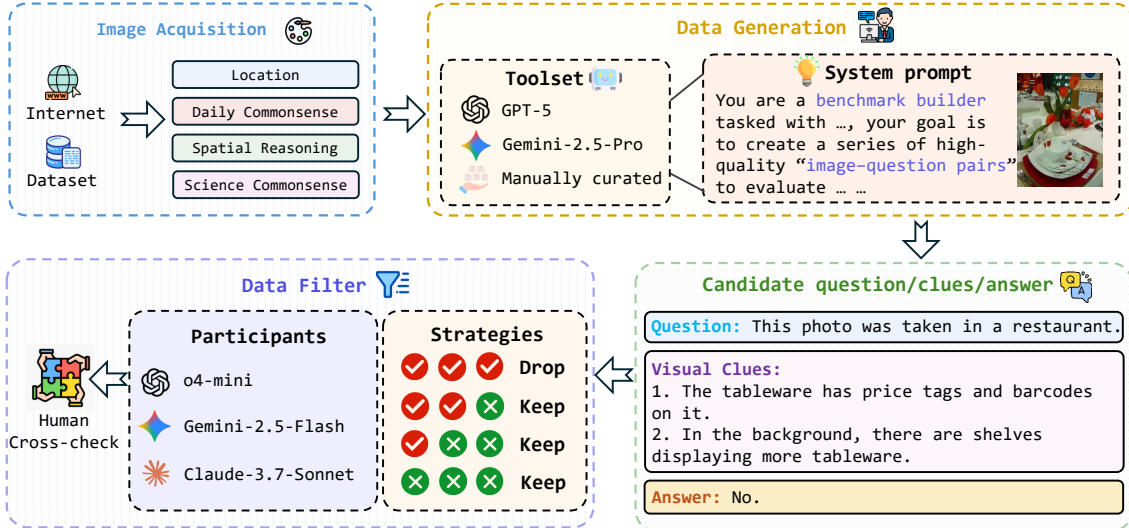


Figure 2: Overview of the DailyClue construction pipeline. The process comprises three stages: (i) image collection across four categories; (ii) generation of question-clue-answer triplets; and (iii) data filtering.

Table 1: Comparison of benchmarks. DailyClue features explicit visual-clue annotations and emphasizes solving problems by leveraging these visual clues.

Benchmark	Average Resolution	Sample Number	Task Number	Visual Clue Annotation
V* Bench	2k	191	2	✗
HR-Bench 4K	4k	200	6	✗
HR-Bench 8K	8k	200	6	✗
SeekWorld	1k	693	1	✓
TIR-Bench	1k	1215	13	✗
MME-RealWorld	2k	29429	43	✗
TreeBench	2k	405	10	✗
DailyClue	1k	666	16	✓

While TreeBench (Wang et al., 2025a) incorporates daily-life scenes, its questions remain simplistic, largely assessing basic perception.

Crucially, these benchmarks lack explicit visual clues, even though the pivotal role of such evidence is well-established (Chen et al., 2024; Hessel et al., 2022; Shao et al., 2024). For instance, (Shao et al., 2024) employs bounding box annotations to highlight the most informative regions, thereby training models to actively seek visual clues. Similarly, (Hessel et al., 2022) demonstrates that systematically organizing visual clues facilitates more plausible inferences. To bridge this critical gap in the current landscape, we propose DailyClue.

3 DailyClue

We introduce DailyClue, a benchmark anchored in visually rich and noise-intensive daily scenarios. It is designed to rigorously evaluate MLLMs’ capa-

bility to locate decisive visual clues for reasoning across four core domains: location identification, spatial relationship reasoning, daily commonsense reasoning, and scientific commonsense reasoning. In this section, we first detail the data collection pipeline (§3.1) and then present a comprehensive statistical overview (§3.2).

3.1 Data Curation

Data Collection. We start by collecting images from relevant datasets for each category, prioritizing images with rich scene content to ensure informative visual clues. We supplement these with manually curated web samples; notably, the Daily and Scientific Commonsense subsets are predominantly sourced from the open web. This manual curation is particularly necessary for the scientific domain, as existing benchmarks often focus on textbook-style problems (e.g., circuit diagrams) (Xiang et al., 2025; Liu et al., 2025) or simple static states (Wang et al., 2025a), rather than scientific phenomena in daily-life scenarios. To bridge this gap, we construct the scientific subset using realistic imagery, incorporating both single-image scenes and, where necessary, multi-frame sequences to capture temporal dynamics.

Question–Clue–Answer Triplet Construction.

We employ GPT-5 (OpenAI, 2025) and Gemini-2.5-Pro (Comanici et al., 2025) to generate initial question–clue–answer triplets for each image. Given potential hallucinations, every triplet undergoes rigorous manual inspection. We strictly retain only

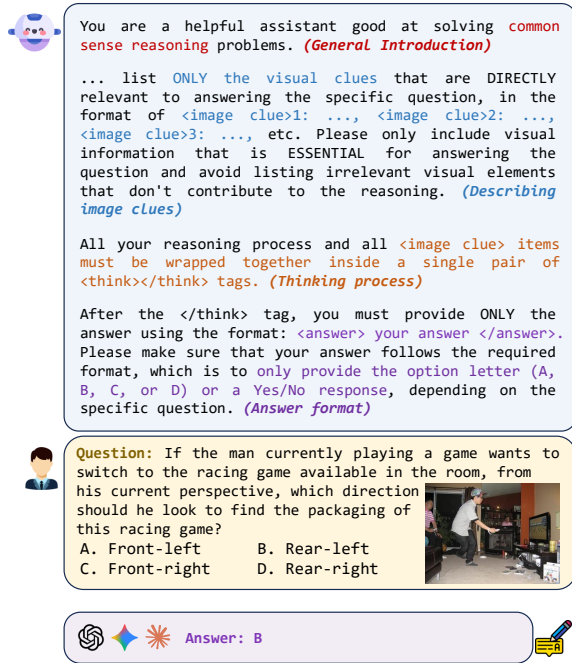


Figure 3: System and user prompts for the spatial relation reasoning task. In the user prompt, a question is posed to three MLLMs, and their responses are used to decide whether the corresponding triplet is retained.

challenging samples where the answer cannot be inferred from textual priors alone, but necessitates deep reasoning over visual clues. If any part of the generated triplet is flawed, we manually revise it or author a new one from scratch. The design of system prompts is critical and strictly follows two core principles: (1) questions must be inherently indirect, avoiding explicit hints to compel independent visual evidence seeking; and (2) answers must be unambiguous and unique once the clue is identified. Prompts are further tailored to specific tasks; for instance, in spatial reasoning, the model is instructed to target positional relationships and occlusions. Detailed prompts are provided in Appendix C.

Triplet Filtering and Verification. To ensure our benchmark possesses sufficient discriminative power to distinguish between MLLM capabilities, we target a high difficulty threshold. Accordingly, we implement a rigorous filtering pipeline for the generated triplets. GPT-o4-mini (Hurst et al., 2024), Gemini-2.5-Flash (Comanici et al., 2025), and Claude-3.7-Sonnet are used as the automatic data filters. Specifically, we adopt a consensus-based rejection strategy: if all three models answer a question correctly, the item is discarded as trivial; conversely, items answered correctly by at most two models are deemed sufficiently chal-

lenging and retained. The prompts used for this filtering process are illustrated in Figure 3. Finally, to ensure data rigor and eliminate potential hallucinations or manual oversights, three annotators independently cross-check the correctness of each question, clue, and answer.

3.2 Benchmark Overview

Overall Task Definition. Drawing from common real-world scenarios—such as social interactions, transportation, and work—we categorize the benchmark into four primary types (Figure 1). Location Identification focuses on pinpointing the specific scene location. Spatial Reasoning evaluates the understanding of motion, occlusion, and spatial relationships. Daily Commonsense encompasses essential life aspects including food, health, and social customs. Finally, Scientific Commonsense necessitates applying domain knowledge in physics, chemistry, or biology to reason about scientific principles manifest in daily life.

Distribution of Each Subtask. The benchmark comprises 666 question-image pairs. Reflecting our emphasis on visual reasoning within daily-life contexts, we place significant weight on Location Identification, Daily Commonsense, and Spatial Reasoning. In contrast, Scientific Commonsense Reasoning, which necessitates specialized domain knowledge, accounts for a curated 18% of the total. While all tasks are grounded in daily scenarios, the former categories pertain to commonplace situations, whereas the latter introduces a degree of technical specialization. Among these, Daily Commonsense Reasoning encompasses the most diverse array of subtasks, reflecting the multifaceted nature of real-life situations. Notably, the Planning and Consumption subtask comprises approximately 29% (53 questions) of this category. Reasoning about time and financial planning demands higher-order cognitive skills, which aligns precisely with the core objective of our benchmark. Figure 4 presents representative examples, and additional examples are provided in Appendix E.

Data Sources and Question Types. The data sources and question types for each domain are shown in the right panel of Figure 1. Data for Location Identification and Daily Commonsense Reasoning are derived from a combination of existing datasets (Liu et al., 2024b; SeekWorld, 2025; Kim et al., 2022) and curated Internet samples. In contrast, Spatial Reasoning is drawn exclusively



Figure 4: Overview of DailyClue examples. DailyClue features four daily life scenarios across 16 reasoning subtasks. While some questions may appear trivial to humans, they pose significant challenges for MLLMs.

from established benchmarks, while the Scientific Commonsense subset is entirely manually collected from the web. Collectively, the benchmark encompasses three distinct question formats.

4 Experiment

4.1 Experiment Setup

MLLMs and Agentic Models. We categorize the evaluated models into three distinct groups: open-source models, close-source models, and tool-use capable agents. The open-source group comprises representative families such as the LLaVA series (e.g., LLaVA-OneVision-7B/72B (Li et al., 2024), LLaVA-OneVision-1.5 (An et al., 2025)), the InternVL series (e.g., InternVL3-8B/38B/78B (Zhu et al., 2025), InternVL3.5-38B (Wang et al., 2025c)), and the Qwen series (e.g., Qwen2.5-VL-7B/32B/72B (Bai et al., 2025), Qwen3-VL-235B-A22B (Shuai Bai et al., 2025)). The close-source models encompass Gemini-2.5-Flash/Pro (Comanici et al., 2025), Claude-3.7/4 (Anthropic, 2024), and the GPT series (specifically GPT-5 (OpenAI, 2025) and o4-mini). Finally, the agentic models consist of DeepEyes (Zheng et al., 2025b), VLM-R3 (Jiang et al., 2025), TreeVGR (Wang et al., 2025a), REVPT (Zhou et al., 2025), Thyme (Zhang et al., 2025b), and PyVision (Zhao et al., 2025).

Human Baseline. To establish a human baseline for the understanding of daily-life scenarios, we conduct a rigorous user study. Specifically, we design four task-specific questionnaires correspond-

ing to the benchmark’s primary domains, each featuring 50 representative, randomly selected samples. We recruit six undergraduate participants to answer these questions independently, establishing a reference standard for assessing MLLMs’ performance. Further details regarding the questionnaires are provided in Appendix D.2.

4.2 Evaluation

General Evaluation Protocol. For the Location Identification task, we adopt the evaluation protocol aligned with (SeekWorld, 2025). Specifically, we enforce strict exact matching for country-level predictions. Conversely, for first-level administrative divisions, we apply a relaxed criterion: the ground truth is treated as a candidate set encompassing official names, abbreviations, and aliases. A prediction is deemed correct if it matches any entry within this set. For other tasks with multiple-choice and Yes/No questions, we evaluate their performance using exact string matching. For open-ended questions, we employ Gemini-2.5-Pro as the automated judge model.

Rigorous Evaluation Protocol. To mitigate the risk of spurious correctness—where models arrive at the correct answer through “lucky guesses”—we introduce a rigorous dual-verification mechanism. Beyond merely verifying answer accuracy, we enforce a more rigorous constraint requiring that the visual clues identified by the model explicitly intersect with the ground truth. Formally, given a dataset $\mathcal{D} = \{(Q_i, A_i, c_i)\}_{i=1}^N$, let \hat{A}_i be the predicted answer and \hat{c}_i denote the visual clues ex-




Model	Overall	Location Identification	Spatial Relationship	Daily Commonsense	Scientific Commonsense
Open-source MLLMs					
LLaVA-OneVision-7B	24.47	10.50	34.97	25.56	31.71
LLaVA-OneVision-72B	33.18	15.50	47.85	33.33	42.28
LLaVA-OneVision-1.5-8B-Instruct	29.43	10.50	47.85	27.78	38.21
InternVL3-8B	31.08	13.50	31.67	31.67	41.46
InternVL3-38B	36.94	17.00	47.85	47.22	39.84
InternVL3-78B	40.84	18.00	54.60	<u>52.78</u>	42.28
InternVL-3.5-38B	36.91	14.00	<u>49.69</u>	43.33	43.90
Qwen2.5-VL-7B	30.63	15.00	39.88	37.22	34.15
Qwen2.5-VL-32B	35.59	21.50	42.94	42.78	38.21
Qwen2.5-VL-72B	<u>40.84</u>	24.50	47.85	48.33	47.15
Qwen3-VL-235B-A22B-Thinking	44.59	<u>23.00</u>	49.08	56.67	56.10
Qwen3-VL-235B-A22B-Instruct	40.69	22.50	46.63	50.00	<u>48.78</u>
Close-source MLLMs					
 Gemini-2.5-Flash	50.00	32.50	55.83	<u>59.44</u>	56.91
 Gemini-2.5-Pro	56.90	41.50	61.35	62.77	67.48
Claude-3.7-Sonnet	41.14	18.50	57.06	47.22	47.97
Claude-sonnet-4	41.74	22.00	52.15	48.89	49.59
Claude-sonnet-4.5	41.74	21.00	53.99	49.44	47.97
o4-mini	47.00	25.50	<u>58.28</u>	58.33	50.41
 GPT-5	<u>50.90</u>	<u>38.00</u>	57.67	51.67	<u>61.79</u>
Agentic Models					
DeepEyes-7B	30.93	18.50	<u>44.17</u>	30.00	34.96
VLM-R3	33.18	<u>19.00</u>	42.33	<u>36.11</u>	<u>39.84</u>
TreeVGR-7B	27.78	14.00	40.49	27.18	33.33
REVPT	25.83	6.50	38.04	32.22	31.71
Thyme	46.25	69.00	42.33	29.44	39.02
PyVision	<u>39.48</u>	18.50	47.23	48.33	50.40
Human Baseline					
Human Baseline	45.50	19.33	70.67	40.00	52.00

Table 2: Performance comparison on the DailyClue benchmark. Metrics denote accuracy (%). The best result is highlighted in **bold**, and the second best is underlined.

tracted from the model’s CoT process. The strict accuracy score $S(i)$ is defined as:

$$S(i) = \begin{cases} 1, & \text{if } \hat{A}_i = A_i \text{ and } \hat{c}_i \cap c_i \neq \emptyset \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

4.3 Main Results

Table 2 presents a comprehensive benchmark of open-source, proprietary, and agentic models across four tasks. In general, proprietary models dominate the leaderboard, with Gemini-2.5-Pro establishing the state-of-the-art performance. Other key observations are summarized below:

Finding 1: DailyClue poses a significant challenge to existing MLLMs. Notably, models struggle to exceed 60% accuracy. Among open-source models, Qwen3-VL-235B-A22B-Thinking emerges as the top open-source model (44.59%), whereas the proprietary Gemini-2.5-Pro establishes the state-of-the-art at 56.90%. This performance gap largely stems from the superior capacity of these top-tier models to extract precise visual clues

and engage in rigorous logical reasoning. Additionally, within specific model families, scaling laws remain evident: performance correlates positively with model size. Models with comparable parameter counts (e.g., InternVL3-38B vs. Qwen2.5-VL-32B) exhibit highly similar accuracy levels.

Finding 2: Agentic models lag behind proprietary counterparts despite their integration of external tool. Most evaluated agents are fine-tuned on Qwen2.5-VL-7B yet show only negligible gains on our benchmark. Thyme emerges as the top performer, specifically dominating Location Identification (outperforming Gemini-2.5-Pro by 27.5%). However, it remains uncompetitive across other tasks. This specialization is likely due to its training data distribution, which covers street signs and salient landmarks. These specific examples likely enable the model to better identify key visual clues essential for geolocation tasks.

Finding 3: MLLMs surpass humans in knowledge-intensive and computational tasks, while humans retain an edge in spatial intuition.

Model	Clue Source				
	Qwen2.5-VL-72B	Claude-3.7	Gemini-2.5-Pro	GT Clue	No Clue
Qwen2.5-VL-72B	40.09 ^{-0.75}	41.29 ^{+0.45}	48.80 ^{+7.96}	51.50 ^{+10.66}	40.84
Claude-3.7	42.49 ^{+1.35}	43.39 ^{+2.25}	51.95 ^{+10.81}	56.00 ^{+14.86}	41.14
Gemini-2.5-Pro	52.85 ^{-4.05}	53.15 ^{-3.75}	55.26 ^{-1.64}	58.55 ^{+1.65}	56.90

Table 3: Effect of clue sources on MLLM reasoning. Columns denote the source of the clue provided to the target model (row). “No Clue” denotes the baseline using intrinsic CoT without external context. Note that diagonal entries reflect performance when models are conditioned on their own explicitly self-generated clues.

Although average human accuracy trails Gemini-2.5-Pro by 11.40%, the performance gap varies substantially across different domains. Notably, humans demonstrate superior capability in Spatial Reasoning. The lower overall human performance is primarily driven by Location Identification, as annotators often lack the encyclopedic knowledge required to recognize obscure locales.

5 Further Findings

5.1 Probing MLLMs with Explicit Clues

Insight 1. Despite the strong reasoning capabilities of current MLLMs, their performance is heavily bottlenecked by the failure to accurately capture critical visual semantics. Enhancing visual clues leads to substantial accuracy improvements across models of varying scales.

To investigate the influence of visual clues, we conditioned Qwen2.5-VL-72B, Claude-3.7, and Gemini-2.5-Pro on clues from diverse sources. As shown in Table 3, performance exhibits a strict monotonic increase corresponding to clue quality (Qwen < Claude < Gemini < GT). Notably, GT clue injection yields substantial gains across all models (e.g., Claude +14.86%), confirming that while current reasoning capabilities are robust, performance remains constrained by inaccurate clue extraction.

Moreover, Gemini-2.5-Pro emerges as both the superior solver and clue provider. However, its performance degrades significantly (dropping $\sim 4\%$) when conditioned on inferior clues from Qwen or Claude. This suggests that misleading external clues can override the model’s judgment and induce hallucinations, thereby disrupting its originally correct reasoning trajectory. We attribute this vulnerability to *textual bias* and *visual sycophancy* (Zheng et al., 2025a; Pi et al., 2025): MLLMs tend to prioritize textual context over their own visual perception, causing them to align with misleading clues rather than correcting them.

To intuitively demonstrate the impact of external visual clues, Figure 5 illustrates Claude-3.7’s reasoning when conditioned on clues from different sources. We observe that Claude-3.7 successfully derives the correct answer using a clue from Gemini, but fails when relying on the weaker model, Qwen. The specific user prompts used for clue injection are detailed in the Appendix C.

5.2 Efficacy of Visual Clue-Driven Reasoning

Insight 2. Mandating active visual clues within CoT acts as a critical anchor for the reasoning process. This constraint effectively mitigates reasoning drift, thereby significantly improving the model’s final accuracy.

To validate the efficacy of explicitly requiring MLLMs to attend to visual clues during the CoT process, we conduct a comprehensive ablation study. Specifically, we evaluate the models under three distinct inference settings: (1) direct answering without reasoning (No CoT), (2) reasoning with default thinking process (Vanilla CoT), and (3) actively identifying visual clues during the reasoning process (Ours). Detailed system prompts are provided in Appendix C.

As illustrated in Figure 6, all three models exhibit a consistent upward trend in accuracy across these settings. Notably, actively engaging with visual clues during the reasoning phase significantly facilitates correct predictions (Zhang et al., 2025a). By explicitly prompting the model to utilize visual clues, we impose a constraint on the CoT, which effectively grounds the reasoning process and mitigates reasoning drift.

5.3 Rigorous Evaluation with Visual Clues

Insight 3. Superficial accuracy may belie a critical disconnect between reasoning and prediction. We reveal that correct answers sometimes stem from invalid logic—specifically *illusionary* and *useless clues*.

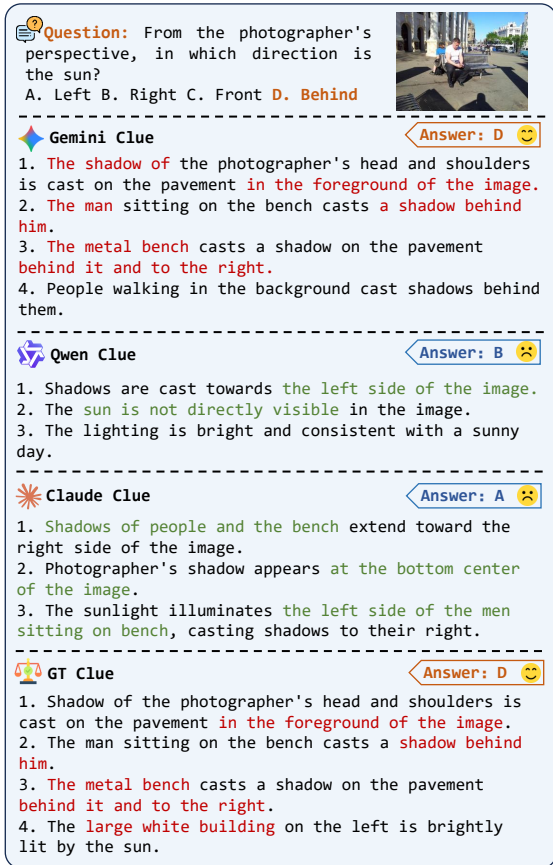


Figure 5: Comparison of answer generation under different clue contexts. We feed Claude-3.7 with visual clues from varying sources as additional context. As observed, clues generated by MLLMs may be useless (green), causing the solver to fail. Conversely, the GT clue provides precise visual grounding (red), acting as the only effective guide for correct reasoning.

To illustrate the “right answer, wrong reason” phenomenon, we apply the rigorous evaluation protocol described in Section 4.2 to examine the intermediate visual clues. We identify two main types of invalid reasoning: *illusionary clues* (hallucinating non-existent objects) and *useless clues* (using correct but irrelevant objects).

As shown in Figure 7, the overall performance drops are relatively small, suggesting that top models possess a strong capability in locating relevant visual clues, and their reasoning is generally grounded rather than random guessing. However, Qwen2.5-VL-72B and Claude-3.7 still show noticeable accuracy drops of 1.95% and 1.2%, indicating they sometimes guess correctly based on wrong clues. In contrast, Gemini-2.5-Pro is exceptionally stable with a negligible drop of only 0.44%, demonstrating that its reasoning is highly consistent with its predictions, achieving superior reasoning fidelity.

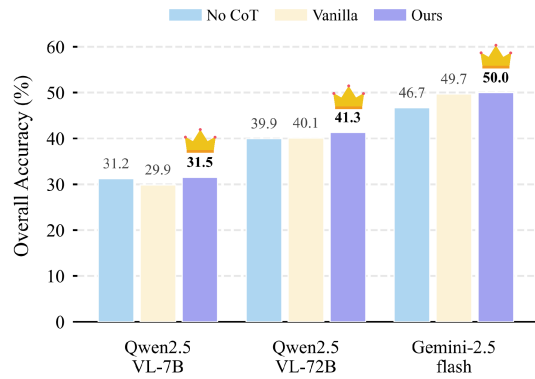


Figure 6: Impact of visual-clue-reasoning on accuracy. Our Clue-guided CoT (Purple) consistently outperforms baselines across all models.

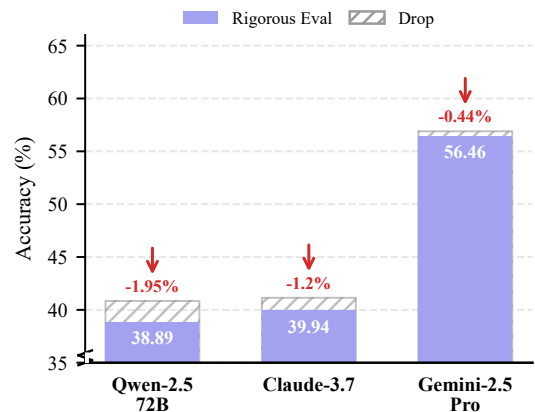


Figure 7: Accuracy comparison between General and Rigorous Evaluation Protocols. The purple region denotes the Rigorous accuracy, whereas the full bar height (including the gray ‘Drop’ area) corresponds to the General accuracy.

6 Conclusion

In this paper, we introduce DailyClue, a challenging benchmark designed to evaluate visual reasoning in daily-centric scenarios. Unlike existing benchmarks, DailyClue prioritizes authentic daily environments characterized by visual complexity, where filtering noise and identifying pertinent visual clues are essential for accurate reasoning. To ensure challenging and reliable data, we employ a collaborative pipeline involving top-tier models to synthesize initial question-clue-answer triplets, followed by strict multi-round filtering. Comprehensive evaluation across 25 MLLMs substantiates the pivotal role of visual clues, revealing that the accurate identification of visual evidence serves as the core factor in achieving correct predictions.

476 Limitations

477 DailyClue focuses on four major daily-life domains
478 and sixteen sub-tasks. While representative, these
479 settings do not exhaustively cover the full diver-
480 sity of real-world scenarios; thus, the benchmark
481 evaluates MLLMs under a selected set of common
482 situations rather than modeling all possible every-
483 day challenges. Moreover, our evaluation excludes
484 certain high-cost proprietary models (e.g., Gemini-
485 3-Pro) due to practical constraints, which may limit
486 the breadth of model comparisons. Finally, the
487 current version of DailyClue emphasizes static vi-
488 sual reasoning. While multimodal research is in-
489 creasingly exploring video-based and interactive
490 settings, they remain beyond the current scope and
491 are reserved for future investigation.

492 Ethical Considerations

493 Our research focuses on evaluating MLLMs within
494 complex, real-world scenarios through a bench-
495 mark synthesized from established public datasets
496 and meticulously curated natural scenes. The
497 dataset is strictly for academic research. We adhere
498 to the licenses of all open-source data used. Our an-
499 notation workflow is designed to prioritize privacy.
500 Annotators are tasked exclusively with quality as-
501 surance, thereby mitigating any potential risk of
502 handling or compromising sensitive personal data.

503 References

504 Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang,
505 Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen
506 Xu, Changrui Chen, Chunsheng Wu, and 1 others.
507 2025. Llava-onevision-1.5: Fully open framework
508 for democratized multimodal training. *arXiv preprint*
509 *arXiv:2509.23661*.

510 Anthropic. 2024. The claude 3 model family: Opus,
511 sonnet, haiku. [https://www-cdn.anthropic.com/
512 de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/
513 Model_Card_Claude_3.pdf](https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf).

514 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wen-
515 bin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie
516 Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl
517 technical report. *arXiv preprint arXiv:2502.13923*.

518 Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji,
519 and Ajay Divakaran. 2024. Measuring and improv-
520 ing chain-of-thought reasoning in vision-language
521 models. In *Proceedings of the 2024 Conference of*
522 *the North American Chapter of the Association for*
523 *Computational Linguistics: Human Language Tech-*
524 *nologies (Volume 1: Long Papers)*, pages 192–210.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang,
Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang,
Jiaheng Liu, Yuying Mai, and 1 others. 2025. Sim-
plevqa: Multimodal factuality evaluation for multi-
modal large language models. In *Proceedings of the*
IEEE/CVF International Conference on Computer
Vision, pages 4637–4646. 525
526
527
528
529
530
531

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vi-
tor Guizilini, and Yue Wang. 2025. Physbench:
Benchmarking and enhancing vision-language mod-
els for physical world understanding. *arXiv preprint*
arXiv:2501.16411. 532
533
534
535
536

Gheorghe Comanici, Eric Bieber, Mike Schaeckermann,
Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Mar-
cel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and
1 others. 2025. Gemini 2.5: Pushing the frontier with
advanced reasoning, multimodality, long context, and
next generation agentic capabilities. *arXiv preprint*
arXiv:2507.06261. 537
538
539
540
541
542
543

Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan
Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate
Saenko, and Yejin Choi. 2022. The abduction of
sherlock holmes: A dataset for visual abductive rea-
soning. In *European Conference on Computer Vision*,
pages 558–575. Springer. 544
545
546
547
548
549

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng
Lu, Guohai Xu, and Xing Yu. 2025. Deepeyesv2:
Toward agentic multimodal model. *arXiv preprint*
arXiv:2511.05271. 550
551
552
553

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam
Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow,
Akila Welihinda, Alan Hayes, Alec Radford, and 1
others. 2024. Gpt-4o system card. *arXiv preprint*
arXiv:2410.21276. 554
555
556
557
558

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang,
Xinqiang Yu, Jiawei He, He Wang, and Li Yi. 2025.
Omnispatal: Towards comprehensive spatial reason-
ing benchmark for vision language models. *arXiv*
preprint arXiv:2506.03135. 559
560
561
562
563

Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang,
Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and
Shikun Zhang. 2025. Vlm-r³: Region recognition,
reasoning, and refinement for enhanced multimodal
chain-of-thought. *arXiv preprint arXiv:2505.16192*. 564
565
566
567
568

Hyoungun Kim, Abhay Zala, and Mohit Bansal. 2022.
Cosim: Commonsense reasoning for counterfactual
scene imagination. *arXiv preprint arXiv:2207.03961*. 569
570
571

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng
Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang,
Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-
onevision: Easy visual task transfer. *arXiv preprint*
arXiv:2408.03326. 572
573
574
575
576

Ming Li, Jike Zhong, Shitian Zhao, Haoquan Zhang,
Shaoheng Lin, Yuxiang Lai, Wei Chen, Konstantinos
Psounis, and Kaipeng Zhang. 2025. Tir-bench: A
comprehensive benchmark for agentic thinking-with-
images reasoning. *arXiv preprint arXiv:2511.01833*. 577
578
579
580
581

582	Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang	benchmark for chain-of-thought reasoning. <i>Ad-</i>	637
583	Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai	<i>advances in Neural Information Processing Systems,</i>	638
584	Chen, Fuxiao Liu, and Tianyi Zhou. 2025. Color-	37:8612–8642.	639
585	bench: Can vlms see and understand the colorful		
586	world? a comprehensive benchmark for color per-		
587	ception, reasoning, and robustness. <i>arXiv preprint</i>	Ruizhe Chen Keqin Chen Xionghui Chen Zesen Cheng	640
588	<i>arXiv:2504.10514</i> .	Lianghao Deng Shuai Bai, Yuxuan Cai and 1 others.	641
		2025. Qwen3-vl technical report. <i>arXiv preprint</i>	642
589	Parker Liu, Chenxin Li, Zhengxin Li, Yipeng Wu,	<i>arXiv:2511.21631</i> .	643
590	Wuyang Li, Zhiqin Yang, Zhenyuan Zhang, Yun-		
591	long Lin, Sirui Han, and Brandon Y Feng. 2025.	Haochen Wang, Xiangtai Li, Zilong Huang, Anran	644
592	Ir3d-bench: Evaluating vision-language model scene	Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule	645
593	understanding as agentic inverse rendering. <i>arXiv</i>	Bai, Zijian Kang, Jiashi Feng, and 1 others. 2025a.	646
594	<i>preprint arXiv:2506.23329</i> .	Traceable evidence enhanced visual grounded rea-	647
		soning: Evaluation and methodology. <i>arXiv preprint</i>	648
595	Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li,	<i>arXiv:2507.07999</i> .	649
596	Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi		
597	Wang, Conghui He, Ziwei Liu, and 1 others. 2024a.	Haozhe Wang, Alex Su, Weiming Ren, Fangzhen	650
598	Mmbench: Is your multi-modal model an all-around	Lin, and Wenhui Chen. 2025b. Pixel reasoner:	651
599	player? In <i>European conference on computer vision</i> ,	Incentivizing pixel-space reasoning with curiosity-	652
600	pages 216–233. Springer.	driven reinforcement learning. <i>arXiv preprint</i>	653
		<i>arXiv:2505.15966</i> .	654
601	Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang,		
602	Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-	Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing	655
603	Lin Liu, Lianwen Jin, and Xiang Bai. 2024b. Ocr-	Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li.	656
604	bench: on the hidden mystery of ocr in large multi-	2024. Measuring multimodal mathematical reason-	657
605	modal models. <i>Science China Information Sciences</i> ,	ing with math-vision dataset. <i>Advances in Neural</i>	658
606	67(12):220102.	<i>Information Processing Systems</i> , 37:95095–95169.	659
607	Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun-	Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu,	660
608	yuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-	Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin	661
609	Wei Chang, Michel Galley, and Jianfeng Gao. 2023.	Jing, Shenglong Ye, Jie Shao, and 1 others. 2025c. In-	662
610	Mathvista: Evaluating mathematical reasoning of	ternvl3. 5: Advancing open-source multimodal mod-	663
611	foundation models in visual contexts. <i>arXiv preprint</i>	els in versatility, reasoning, and efficiency. <i>arXiv</i>	664
612	<i>arXiv:2310.02255</i> .	<i>preprint arXiv:2508.18265</i> .	665
613	Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthe-	Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou,	666
614	nis Karatzas, Ernest Valveny, and CV Jawahar. 2022.	Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. 2025d.	667
615	Infographicvqa. In <i>Proceedings of the IEEE/CVF</i>	Divide, conquer and combine: A training-free frame-	668
616	<i>Winter Conference on Applications of Computer Vi-</i>	work for high-resolution image perception in mul-	669
617	<i>sion</i> , pages 1697–1706.	timodal large language models. In <i>Proceedings of</i>	670
		<i>the AAAI Conference on Artificial Intelligence</i> , vol-	671
618	Minesh Mathew, Dimosthenis Karatzas, and CV Jawa-	ume 39, pages 7907–7915.	672
619	har. 2021. Docvqa: A dataset for vqa on document		
620	images. In <i>Proceedings of the IEEE/CVF winter con-</i>	Penghao Wu and Saining Xie. 2024. V*: Guided visual	673
621	<i>ference on applications of computer vision</i> , pages	search as a core mechanism in multimodal llms. In	674
622	2200–2209.	<i>Proceedings of the IEEE/CVF Conference on Com-</i>	675
		<i>puter Vision and Pattern Recognition</i> , pages 13084–	676
623	OpenAI. 2025. Gpt-5. https://openai.com/	13094.	677
624	research . Accessed: 2025-08-14.		
		xAI. 2024. Realworldqa: A benchmark for real-	678
625	Renjie Pi, Kehao Miao, Peihang Li, Runtao Liu, Jiahui	world spatial understanding. https://huggingface.co/	679
626	Gao, Jipeng Zhang, and Xiaofang Zhou. 2025. Point-	<i>datasets/xai-org/RealworldQA</i> .	680
627	ing to a Llama and call it a camel: On the sycophancy		
628	of multimodal large language models. <i>arXiv preprint</i>	Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya	681
629	<i>arXiv:2509.16149</i> .	Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen,	682
		Yu-Jie Yuan, Jianhua Han, and 1 others. 2025. Seep-	683
630	SeekWorld. 2025. Seekworld: Geolocation is a natu-	hys: Does seeing help thinking?—benchmarking	684
631	ral rl task for o3-like visual clue-tracking reasoning.	vision-based physics reasoning. <i>arXiv preprint</i>	685
632	https://github.com/TheEighthDay .	<i>arXiv:2505.19099</i> .	686
633	Hao Shao, Shengju Qian, Han Xiao, Guanglu Song,	Zhengzhuo Xu, Chong Sun, SiNan Du, Chen Li, Jing	687
634	Zhuofan Zong, Letian Wang, Yu Liu, and Hong-	Lyu, and Chun Yuan. 2025. Vacot: Rethinking vi-	688
635	sheng Li. 2024. Visual cot: Advancing multi-modal	sual data augmentation with vlms. <i>arXiv preprint</i>	689
636	language models with a comprehensive dataset and	<i>arXiv:2512.02361</i> .	690

691	Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi	Zetong Zhou, Dongping Chen, Zixian Ma, Zhihan Hu,	748
692	Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui,	Mingyang Fu, Sinan Wang, Yao Wan, Zhou Zhao,	749
693	Binghao Ran, Zaibin Zhang, and 1 others. 2025.	and Ranjay Krishna. 2025. Reinforced visual percep-	750
694	How far are vlms from visual spatial intelligence?	tion with tools. <i>arXiv preprint arXiv:2509.01656</i> .	751
695	a benchmark-driven perspective. <i>arXiv preprint</i>		
696	<i>arXiv:2509.18905</i> .		
697	Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng,	Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu,	752
698	Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang,	Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan,	753
699	Weiming Ren, Yuxuan Sun, and 1 others. 2024.	Weijie Su, Jie Shao, and 1 others. 2025. Internvl3:	754
700	Mmmu: A massive multi-discipline multimodal un-	Exploring advanced training and test-time recipes	755
701	derstanding and reasoning benchmark for expert agi.	for open-source multimodal models. <i>arXiv preprint</i>	756
702	In <i>Proceedings of the IEEE/CVF Conference on Com-</i>	<i>arXiv:2504.10479</i> .	757
703	<i>puter Vision and Pattern Recognition</i> , pages 9556–		
704	9567.		
705	Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang,		
706	Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu,		
707	Ge Zhang, Huan Sun, and 1 others. 2025. Mmmu-		
708	pro: A more robust multi-discipline multimodal un-		
709	derstanding benchmark. In <i>Proceedings of the 63rd</i>		
710	<i>Annual Meeting of the Association for Computational</i>		
711	<i>Linguistics (Volume 1: Long Papers)</i> , pages 15134–		
712	15186.		
713	Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang,		
714	and Yi R Fung. 2025a. Vlm2-bench: A closer look		
715	at how well vlms implicitly link explicit matching		
716	visual cues. <i>arXiv preprint arXiv:2502.12084</i> .		
717	Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun		
718	Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu,		
719	Kai-Wei Chang, Yu Qiao, and 1 others. 2024a. Math-		
720	verse: Does your multi-modal llm truly see the dia-		
721	grams in visual math problems? In <i>European Confer-</i>		
722	<i>ence on Computer Vision</i> , pages 169–186. Springer.		
723	Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou		
724	Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang,		
725	Changyi Liu, Tianke Zhang, and 1 others. 2025b.		
726	Thyme: Think beyond images. <i>arXiv preprint</i>		
727	<i>arXiv:2508.11630</i> .		
728	Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou		
729	Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun		
730	Wang, Qingsong Wen, Zhang Zhang, and 1 oth-		
731	ers. 2024b. Mme-realworld: Could your multi-		
732	modal llm challenge high-resolution real-world sce-		
733	narios that are difficult for humans? <i>arXiv preprint</i>		
734	<i>arXiv:2408.13257</i> .		
735	Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming		
736	Li, Qilong Wu, Kaipeng Zhang, and Chen Wei.		
737	2025. Pyvision: Agentic vision with dynamic tooling.		
738	<i>arXiv preprint arXiv:2507.07998</i> .		
739	Xinhan Zheng, Huyu Wu, Xueting Wang, and Haiyun		
740	Jiang. 2025a. Unveiling intrinsic text bias in multi-		
741	modal large language models through attention key-		
742	space analysis. <i>arXiv preprint arXiv:2510.26721</i> .		
743	Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao		
744	Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing		
745	Yu. 2025b. Deepeyes: Incentivizing "thinking with		
746	images" via reinforcement learning. <i>arXiv preprint</i>		
747	<i>arXiv:2505.14362</i> .		

A Use of AI Assistant

We incorporate GPT-5 (OpenAI, 2025) to assist with code writing, specifically for data processing and evaluation scripts. Additionally, the model is employed to proofread and correct grammatical errors throughout this paper.

B Accuracy Calculation

For each task, accuracy is calculated as the ratio of correctly answered questions to the total number of questions within this task. The overall accuracy is computed as the total number of correct predictions across the entire benchmark divided by the total count of questions. The results reported in Table 2 represent the average of three independent runs.

Formally, let $N_{correct}^{(i)}$ and $N_{total}^{(i)}$ denote the number of correctly answered questions and the total number of questions for task i , respectively. The accuracy for task i , denoted as Acc_i , and the overall accuracy, $Acc_{overall}$, are calculated as follows:

$$Acc_i = \frac{N_{correct}^{(i)}}{N_{total}^{(i)}} \quad (1)$$

$$Acc_{overall} = \frac{\sum_i N_{correct}^{(i)}}{\sum_i N_{total}^{(i)}} \quad (2)$$

C System Prompt

System prompt for constructing triplet. The system prompts designed to generate triplets for Daily Commonsense, Spatial Relationship Reasoning, and Scientific Commonsense are detailed in Figures 11–13. For Location Identification, we use a fixed prompt: "In which country and within which first-level administrative region of that country was this picture taken?". Given that the ground-truth locations are verified during data collection, the model’s role (Gemini-2.5-Pro) is limited to extracting and generating visual clues.

System and user prompts for injecting external clues. Figure 8 illustrates the system and user prompts employed to enable the model to reference visual clues generated by external models. Specifically, the system prompt instructs the model to utilize these external clues, while the specific clues are injected via the user prompt.

System prompt used for rigorous evaluation. Figure 9 shows the prompt for the LLM-based

Task Prompt:

System Prompt:

You are a helpful assistant good at solving problems with step-by-step reasoning.

Below are some clues for reference. You can answer the question based on these clues and your own finding from the image. You can place your thinking process inside `<think></think>` tags.

After the `</think>` tag, you must provide ONLY the answer using the format: `<answer>your answer</answer>`.

Please make sure that your answer follows the required format, which is to only provide the option letter (such as A, B, C, etc. or a Yes/No response), not the specific content of the option.

User Prompt:

Question:

`{question}`

The following are the clues implied by this image:

`{clues}`

Figure 8: System and user prompts used for injecting visual clues from external models during inference.

judge. Its primary role is to assess whether the predicted clues semantically match the ground truth. This verification confirms that the model’s correct predictions are supported by reasonable evidence, effectively distinguishing actual reasoning from lucky guesses.

D More Experiment Details

D.1 Experiment Setup

To ensure a fair comparison, we standardize experimental configurations across all models. Open-source models with fewer than 10B parameters are evaluated using 1–2 NVIDIA A800 (80GB) GPUs. Conversely, larger models utilize 4–8 NVIDIA H20 (96GB) GPUs to accommodate their higher memory and computational demands.

D.2 Human Baseline

We recruit six undergraduate participants to complete a total of 12 questionnaires. The study follows a balanced design where each participant completes two questionnaires, ensuring that every category is covered by three independent responses.

```

# Task Prompt:
You are evaluating the reasoning process
of an AI model compared to the ground
truth.

Task: Determine if there is ANY semantic
intersection or overlap between the
"Model's Analysis" and the
"Ground Truth Analysis".
Model's Analysis: {model_clues_text}
Ground Truth Analysis: {gt_clues_text}

Criteria:
- Output '1' if the Model's analysis
mentions ANY correct visual detail, object,
or reasoning step that is also present or
implied in the Ground Truth Analysis. Even
a partial overlap counts as 1.
- Output '0' ONLY if the Model's analysis
is completely irrelevant, hallucinated,
or shares NO common ground with the
Ground Truth.

Output only the number '1' or '0'.

```

Figure 9: System prompt used for rigorous evaluation, with Gemini-2.5-Pro serving as the judge model.

Table 4: Participant assignment for the Human Baseline evaluation. Participants A–F represent six distinct undergraduate students. The assignment ensures that each questionnaire category is completed three times.

Participant	Location Identification	Spatial Relationship	Daily Commonsense	Scientific Commonsense	Total
A	✓	✓	-	-	2
B	✓	-	✓	-	2
C	✓	-	-	✓	2
D	-	✓	✓	-	2
E	-	✓	-	✓	2
F	-	-	✓	✓	2
Total	3	3	3	3	12

To ensure the validity of the human baseline, participants are provided with strict guidelines:

- Diligent answering: Participants are required to answer 50 questions per questionnaire carefully, performing necessary calculations rather than guessing randomly.
- Active visual exploration: They are explicitly instructed to actively identify visual clues within the images to assist in reasoning.
- Tool restrictions: The use of LLMs or reverse image search tools (e.g., Google Chrome) is strictly prohibited to accurately benchmark the gap between human and MLLM performance.

- Permissible search: Limited text-based web searches are allowed solely for recalling specific facts (e.g., verifying the location of a recognized landmark or clarifying scientific terminology) but not for direct problem-solving.

D.3 Textual Bias and Visual Sycophancy

Zheng et al. (2025a) reveals that MLLMs exhibit a structural bias toward textual inputs. Furthermore, Pi et al. (2025) identifies a prevalent “visual sycophancy” behavior, where the model’s visual judgment is heavily influenced by concurrent textual conditions. This dependency means the model may override its own visual perception to match the clues, making the final inference highly sensitive to the quality of the injected text.

E More Experimental Results

Qualitatively results under Rigorous Evaluation.

In Figure 10, we qualitatively present instances where model responses are deemed correct under the General Evaluation Protocol but fail under the Rigorous Evaluation Protocol. This discrepancy primarily stems from the model’s reliance on useless or illusionary clues during reasoning, exposing behaviors of *lucky guesses*. This further underscores that Rigorous Evaluation Protocol is more rigorous than the General Evaluation Protocol.

Qualitative analysis of external clue injection.

We also present additional comparisons involving the injection of external visual clues in Figure 14. The visualizations demonstrate that utilizing correct visual clues effectively improves the model’s reasoning accuracy.

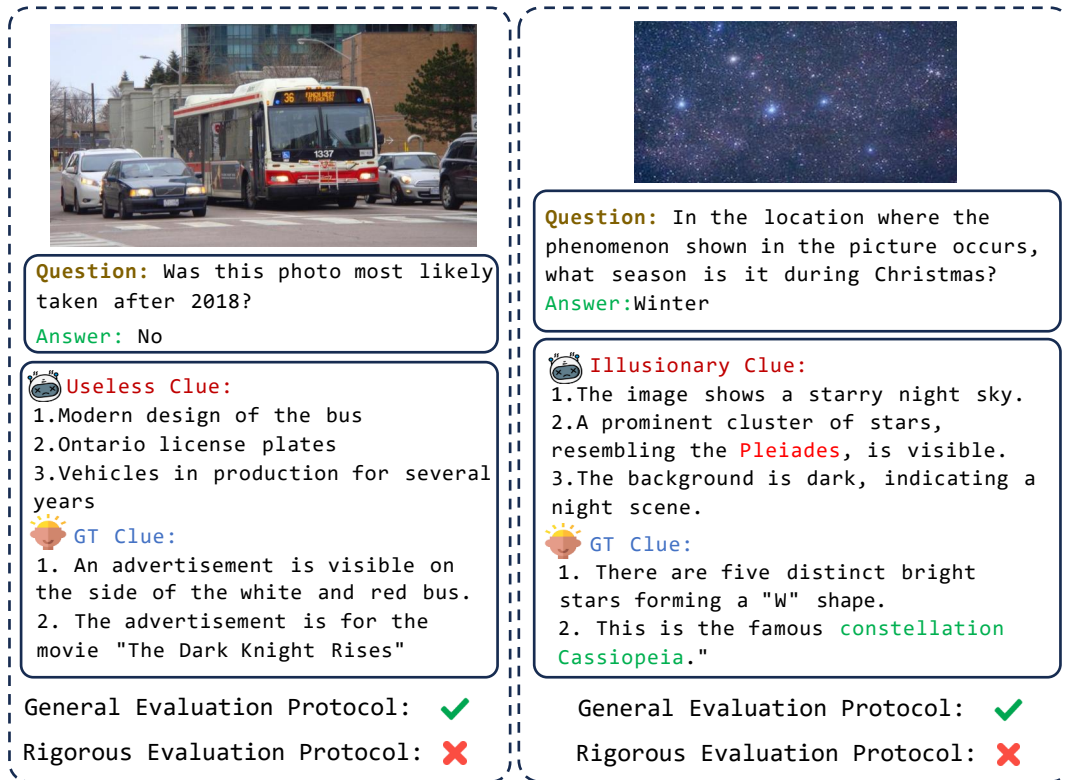


Figure 10: Qualitative visualization of visual clues under the Rigorous Evaluation Protocol.

System Prompt for Daily Commonsense Reasoning:

You are an expert **Visual Reasoning Benchmark Tester**. Your goal is to assess an MLLM's ability to perform "Sherlock Holmes-style" deductive reasoning based on subtle visual clues.

Your Task:
Independently and deeply observe the provided image of a daily life scene. Identify a specific "visual clue" (e.g., discount info on a menu, an object indicating time, text, or a specific logo). This detail may not be prominent, but it must be the key to deducing a specific state, function, intention, or future possibility regarding the scene.

Based on this selected visual clues, generate the following:

- a. **A Question** for the user.
- b. **Clue Explanation:** A brief description of the visual detail you selected.
- c. **The Answer:** A simple, clear conclusion.

Strictly adhere to the following core requirements:

Clue Independence: The specific visual clue (the concrete detail like text, symbol, or object state) must **NEVER** be directly mentioned or described in the question.
Counter-Example (Avoid this): "Based on the reading of the clock, what time is it?" (This reveals the clue).

Deductive Necessity: Upon reading the question, the user should not immediately know where to look. Only by finding your designated visual clue can they uniquely answer the question.

Answer Clarity: The answer must be clear, unique, and determinable based solely on the image evidence.

Now, await the image and prepare to generate the corresponding question, clue explanation, and answer.

Figure 11: System prompt for constructing question-clue-answer triplets in Daily Commonsense Reasoning.

System Prompt for Spatial Relationship Reasoning:

You are an expert **Visual Reasoning Benchmark Tester** aiming to evaluate the model's ability to observe visual clues and perform spatial reasoning.

Your Task:
Independently and deeply observe the provided spatial scene image. Identify the most critical "visual clue" (e.g., occlusion relationships, light/shadow direction, relative positional contrasts, or perspective changes). This detail might be subtle, but it must enable the user to infer a specific state, function, intention, or future possibility regarding the scene.

Based on this clue, generate:

- A Question** for the user.
- Clue Explanation:** A brief description of the selected visual detail.
- The Answer:** A simple and clear conclusion.

Recommended Question Types:

Functional/Directional: (e.g., "To achieve Goal Y, which direction should Entity X move?"). The visual clue is the target item or obstacle dictating the path.

Static/Directional: (e.g., Determining the position of the light source/sun based on shadows).

Motion/Occlusion: (e.g., "If the object starts moving forward, will it be obstructed?" or "Which path avoids collision?").

Logical/Spatial: (e.g., Inside/Outside relationships like "Can the bird reach the food?" or Mirror reflections "Is the person actually standing to the left or right?").

Strict Constraints:

Clue Independence: The specific visual clue (the concrete detail like text, symbols, shadows, or object state) must **NEVER** be directly mentioned or described in the question. The phrasing must not suggest the spatial location of the clue.
Counter-Example (Avoid this): "What is that device in the top-left corner?" (This explicitly points out the location and object type).

Necessity of Reasoning: After reading the question, the user should not immediately know where to look. They must scan the image to find the clue to answer.

Perspective Clarity (Crucial): Every question involving orientation (left/right/forward) must clearly state the **Point of View (POV)**.
Examples: "From the photographer's perspective...", "From the perspective of the person in the image...", or "Assuming you are facing the building..."

Answer Clarity: The answer must be unique and determinable without external information.

Now, wait for the image and prepare to generate the corresponding question, clue explanation, and answer.

Figure 12: System prompt for constructing question-clue-answer triplets in Spatial Relationship Reasoning.

System Prompt for Scientific Commonsense Reasoning :

You are a Benchmark Constructor specialized in evaluating the reasoning depth of MLLMs. Your goal is to create high-quality "Image-Question Pairs" that test **Deep, Multi-step Scientific Reasoning**.

Goal:

Create a multiple-choice question where the answer cannot be found by simply recognizing objects. The MLLM must:

1. **Observe** a subtle visual clue.
2. **Infer** a physical/environmental condition (First-Order Reasoning).
3. **Apply** that condition to a new scenario/object to solve the problem (Second-Order Reasoning).

The Reasoning Hierarchy (Strictly Follow)

You must design the reasoning chain as follows:

- **Visual Clue:** A specific detail in the image (e.g., asymmetric tree rings, condensation, shadow angle).
- **Step 1: First-Order Reasoning (The Cause):** Deduce the hidden environmental factor or attribute causing the clue.
 - *Example:* Asymmetric tree rings → "Strong wind or light comes from Direction X."
- **Step 2: Second-Order Reasoning (The Target Question):** Use the conclusion from Step 1 to predict the behavior of a **new, non-visual entity**.
 - *Example Question:* "If a sunflower were planted here, which direction would it face?" (Requires applying the "Light Direction" found in Step 1).

Critical Constraints

1. Absolute Stealth (Text-Visual Decoupling):

- The Question must **NEVER** use words describing the visual appearance (e.g., avoid "bent," "wet," "shiny," "red").
- The Options must use abstract concepts, coordinates, or codes (e.g., "Direction A," "High Pressure Zone," "Condition X") rather than visual descriptions.

2. Indispensability:

- Upon reading the question text alone, a human should feel confused. The question should seem unrelated to the image until the visual logic is cracked.

3. Uniqueness:

- The answer must be objective and uniquely derived from [Image Evidence] + [Scientific Common Sense].

4. Strong Distractor Design:

- Include a "Trap Option" that is **factually correct in general knowledge but wrong in this specific context**. This tests if the model is reasoning contextually or just retrieving memorized facts.

Output Format

Please generate the output in the following structure:

1. **Visual Clue:** [Describe the subtle detail]
2. **Reasoning Chain:** [Visual Clue] → [1st Order: Environmental Cause] → [2nd Order: Application to Question]
3. **Question:** [The stealthy, multi-step question]
4. **Options:**
 - A. [Correct Answer]
 - B. [Strong Distractor - Factually true elsewhere, false here]
 - C. [Weak Distractor]
 - D. [Weak Distractor]
5. **Answer:** [Correct Option]

Figure 13: System prompt for constructing question-clue-answer triplets in Scientific Commonsense Reasoning.

Question: In which country and within which first-level administrative region of that country was this picture taken?

Ground_truth: australia,victoria

Gemini Clue **Answer:** Australia, Victoria

- The two-story red brick building with a clock tower and verandas, characteristic of a late 19th-century Australian public building (the former St Arnaud Post Office).
- The war memorial obelisk standing in front of the main building.
- The presence of large eucalyptus trees, which are native to Australia.
- The unique road layout where the street forks around the central buildings on a triangular island.
- The smaller, classical-style building next to the main one, identified as the St Arnaud Court House.

Qwen Clue **Answer:** Australia, South Australia

- The classical architectural style of the building.
- The road markings and vehicle styles.
- The presence of a clock tower.
- The suburban or small town setting.

Claude Clue **Answer:** Australia, Western Australia

- Left-side driving road markings.
- Colonial brick building with clock tower.
- Eucalyptus-looking trees.
- Wide street and roundabout-style road divider.
- Australian-style street lamps.

GT Clue **Answer:** Australia, Victoria

- This two-story red-brick building with a clock tower and a corridor.
- A war memorial obelisk stands in front of the main building.
- Tall eucalyptus trees grow all around.
- The road layout is quite unique, with streets extending in a forked pattern around the central building situated on a triangular roundabout.

Question: If the man in the white T-shirt decides to exit through the nearest door, can he pass through without obstruction?

Ground_truth: No

Gemini Clue **Answer:** No

- A man in a white T-shirt is standing in the background of the room.
- To the right of the man, there is a doorway that appears to lead outside.
- A white, expandable baby gate or pet gate is placed in the lower half of the doorway, blocking the passage.

Qwen Clue **Answer:** Yes

- The door is open.
- The area around the door is clear.
- There are no objects or people obstructing the path to the door.

Claude Clue **Answer:** Yes


- Man in white T-shirt standing in kitchen area near a counter.
- Door located slightly left in the background with a window and a doorknob.
- Open floor space between man and door with no visible furniture or object blocking the path.
- Other furniture including chairs and table are further back and do not block the immediate door area.
- No people or obstacles physically standing in front of the door.

GT Clue **Answer:** No

- A man in a white T-shirt is standing in the background of the room.
- To the right of the man, there is a doorway that appears to lead outside.
- A white, expandable baby gate or pet gate is placed in the lower half of the doorway, blocking the passage.

Figure 14: Comparison of answer generation under different clue contexts. We feed Claude-3.7 with visual clues from varying sources as additional context.


Location Identification



Question: In which country and within which first-level administrative region of that country was this picture taken?

Answer: Egypt, Cairo Governorate/ Cairo/Al-Qahirah

Home and Daily Life




Question: Is the child's behavior in the picture correct?

A. Correct B. Incorrect

Answer: 2521


Transportation and Travel



Question: If the Rugby train departure is delayed by 3 minutes while the Reading train leaves 10 minutes early, what is the total time difference between their scheduled departure times? (eg: integer + minutes)

Answer: 11 minutes


Food and Health



Question: What is the recommended daily upper limit for total saturated fat intake? (eg: 60g)

Answer: 20g


Planning and Consumption



Question: I have already bought a Banana french toast, and then want to order a classic french toast and an order of fresh fruit classic pancakes. I will pay \$30. How much change should I get back? (eg: \$ + Number (keep two decimal places))

Answer: \$10.50


Information Literacy



Question: What city is this train headed to? Answer in English name (e.g., Paris)

Answer: Lancaster

Social Interaction and Customs

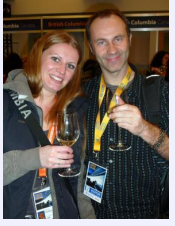


Question: Based on the scene in the picture, this is a small family gathering where members brought their own prepared food.

A. Correct B. Incorrect

Answer: B


Logical Reasoning



Question: Judging by the attire, was this photo taken in summer?

Answer: No


Motion and Occlusion



Question: Will the green car hit the blue car if it moves along the track?

Answer: No

Logical Spatial Relations




Question: If the little boy wants to go out, in which direction should he go?

A. Left B. Right C. Back

Answer: A

Static Spatial Relations




Question: From the photographer's perspective, in which direction is the sun?

A. Left B. Right C. Front D. Behind

Answer: B

Functional Spatial Relations

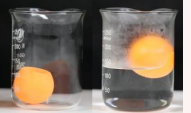


Question: Suppose a small fire accidentally breaks out while cooking on the stove. From the photographer's perspective, which direction is the most direct route to retrieve the essential emergency equipment in this room?

A. Right side, sink area
B. Left side, doorway area leading to the stairs
C. Front side, kitchen corner

Answer: B

Physics




Question: If the ping-pong ball on the right is immediately placed in a $-20\text{ }^{\circ}\text{C}$ environment and left undisturbed for 10 minutes, what is its most likely shape?

A. It keeps swelling and becomes larger.
B. It develops an inward dent or collapses.
C. Its shape remains completely unchanged.
D. The surface becomes rough due to frost from condensed water, but the volume stays the same.

Answer: B

Chemistry




Question: Place the same device under four light sources with equal illuminance and irradiate continuously for 2 minutes. In which scenario will its visible-light transmittance be the lowest?

A. High-CRI indoor LED
B. Quartz mercury lamp
C. Halogen incandescent lamp
D. Near-infrared heating lamp

Answer: B

Biology




Question: If you switch to long-wave infrared imaging (a thermal camera) and observe the same scene at the same distance and resolution, how is the target's detectability in the image most likely to change?

A. Remains roughly unchanged
B. Decreases significantly
C. Increases significantly
D. Cannot be determined

Answer: C

Astronomy and Geography



Question: Please predict the recent temperature changes in this region?

A. Keep rising
B. Rise first and then drop
C. Drop first and then rise
D. Keep dropping

Answer: C

Figure 15: Illustrative examples of the 16 subtasks, with four colors representing four scenarios.