

Can MLLMs Reason Beyond Language?

VisReason: A Comprehensive Benchmark for Vision-Centric Reasoning

Longteng Guo^{1,2*} Yifan Wang^{1,2*} Pengkang Huo^{1,2*} Tailai Chen^{1,2}
Yuze Wu^{1,2} Jing Liu^{1,2} Xinxin Zhu^{1,2†}

¹Institute of Automation, Chinese Academy of Sciences

²School of Artificial Intelligence, University of Chinese Academy of Sciences

{longteng.guo, jliu, xinxin.zhu}@nlpr.ia.ac.cn

{wangyifan2026, huopengkang2025}@ia.ac.cn

Abstract

Recent multimodal large language models (MLLMs) achieve strong performance on visual reasoning benchmarks, yet it remains unclear to what extent such performance reflects reasoning directly grounded in visual evidence. We introduce VisReason, a benchmark for vision-centric reasoning in everyday scenarios where perception and inference are tightly coupled. VisReason contains 1,505 questions across 10 categories spanning perceptual, structural, and conceptual reasoning. Our evaluation shows that VisReason poses a qualitatively different challenge from existing benchmarks, exposing substantial gaps between humans and current MLLMs and revealing limited benefits from test-time reasoning strategies. VisReason offers a focused diagnostic for evaluating vision-centric reasoning beyond language. We release our resources at <https://github.com/CASIA-IVA-Lab/VisReason>

1 Introduction

Recent advances in multimodal large language models (MLLMs) demonstrate strong performance on tasks such as scientific question answering, diagram understanding, and mathematical problem solving, suggesting an emerging ability to integrate visual and textual information for non-trivial reasoning.

However, many widely used visual reasoning benchmarks emphasize STEM-oriented or knowledge-intensive problems, where textual descriptions, symbolic representations, or domain priors play a dominant role. In these settings, models can often abstract visual inputs into language and perform most reasoning in the linguistic space, raising a fundamental question: *do current MLLMs truly reason from visual evidence, or do they primarily rely on language-mediated inference?*

To probe this question, we conduct a diagnostic comparison across existing benchmarks, as shown in Fig. 1. Using the same MLLM, we compare performance under direct vision-text inputs against a language-mediated setting where images are replaced by model-generated captions. On established benchmarks such as MMMU (Yue et al., 2024), MMMU-Pro (Yue et al., 2025), and MathVista (Lu et al., 2024), this substitution leads to only minor performance drops (< 6.35%), indicating that much of the reasoning can be preserved through textual descriptions alone.

These observations suggest that current evaluation protocols provide limited insight into vision-centric reasoning, as models can often succeed through language-mediated abstractions without robust visual grounding. By contrast, human visual reasoning routinely operates directly on visual input, enabling people to infer structure, relations, and implicit rules without relying on rich language descriptions or formal knowledge. This contrast raises a central question: *can MLLMs reason in a similarly vision-centric manner?*

To answer this question, we introduce **VisReason**, a benchmark designed to evaluate vision-centric reasoning beyond language mediation. VisReason targets everyday visual scenarios in which perception and reasoning are tightly coupled and textual cues alone are insufficient. As shown in Fig. 1, removing visual input leads to a substantial performance drop (48.12%), underscoring VisReason’s reliance on direct visual evidence.

VisReason consists of 1,505 carefully curated questions spanning 10 reasoning categories, covering perceptual, structural, and conceptual forms of visual reasoning, such as identifying visual differences, reasoning about 3D-spatial and game-board configurations, and inferring implicit rules from visual cues.

By isolating and evaluating vision-centric reasoning across these dimensions, VisReason of-

*Equal contribution.

†Corresponding authors.

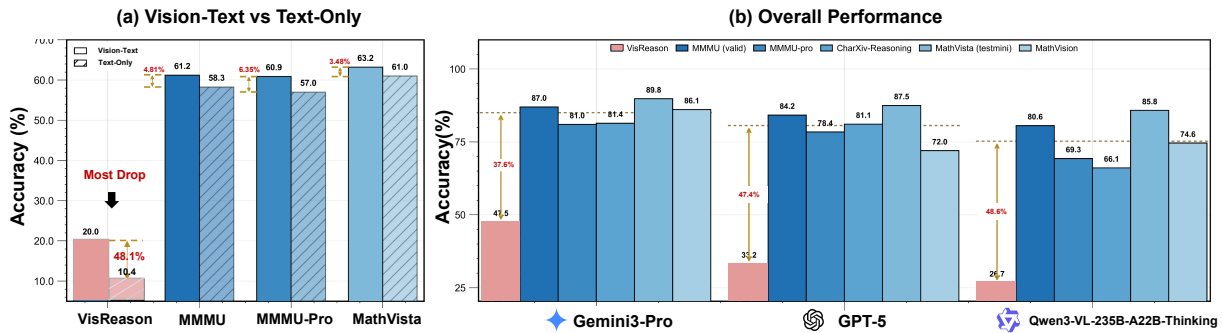


Figure 1: Diagnostic comparison of vision-centric reasoning across benchmarks and models. **Left:** Performance under direct vision–text inputs versus language-mediated text-only inputs, evaluated using the same MLLM (Qwen3-VL-32B), where images are replaced by captions generated by Qwen-VL-Max, highlighting the reliance on visual evidence. **Right:** Overall performance of representative proprietary and open-source MLLMs on VisReason and existing visual reasoning benchmarks, showing a larger performance gap on VisReason.

fers a diagnostic lens on the limitations of current MLLMs and a foundation for advancing multimodal models that reason more directly from visual evidence. Our experiments on VisReason lead to several key findings.

- **VisReason exposes a large and diagnostic gap between humans and leading MLLMs.** Humans solve VisReason reliably using the provided visual evidence, whereas even the strongest MLLMs remain far behind. Moreover, VisReason sharply differentiates model capabilities, with performance ranging from near-chance to substantially higher accuracy.
- **Increasing inference-time thinking budget yields consistent but slow and saturating gains.** Allocating more inference-time tokens generally improves accuracy, but the gains grow gradually and tend to plateau at higher budgets. As a result, additional test-time computation offers limited leverage in the absence of stronger visual grounding.
- **Explicit CoT prompting provides only marginal benefit for non-thinking models.** For non-thinking models, explicit CoT prompting yields only small average improvements (e.g., +1.1% for GPT-4o) and does not consistently benefit perception-heavy categories. Without accurate extraction of visual evidence, longer textual reasoning traces rarely translate into better answers.
- **Scaling model capacity consistently improves vision-centric reasoning performance.** Across model families, larger models achieve higher accuracy on VisReason, confirming that model ca-

capacity remains an important contributor to vision-centric reasoning. Nevertheless, even the largest models fall well short of human performance, leaving substantial room beyond scaling alone.

2 Related Work

2.1 Multimodal Large Language Models

Recent MLLMs have shown strong multimodal understanding and are increasingly tested on whether they can carry out multi-step reasoning grounded in visual evidence. Representative proprietary systems include GPT-4V (OpenAI, 2023), GPT-4o (OpenAI et al., 2024), Gemini (Team et al., 2023), and Claude, while competitive open-weight families include Qwen-VL (Wang et al., 2024) and InternVL (Chen et al., 2024; Zhu et al., 2025). In addition, thinking-style variants (Guo et al., 2025) allocate larger reasoning budgets at inference time, motivating closer examination of whether the gains come from stronger vision-centric reasoning with tighter visual grounding, rather than merely longer language-side deliberation.

2.2 Visual Reasoning Benchmarks

Recent multimodal evaluation has increasingly emphasized reasoning over visual evidence. Benchmarks such as ScienceQA (Lu et al., 2022), MathVista (Lu et al., 2024), and MMMU (Yue et al., 2024) introduce exam-style, visually grounded problems, but remain largely STEM- and knowledge-centered and do not always require fine-grained, directly verifiable visual evidence. In parallel, domain-specific benchmarks such as Sudoku-Bench (Seely et al., 2025) and puzzle-style datasets

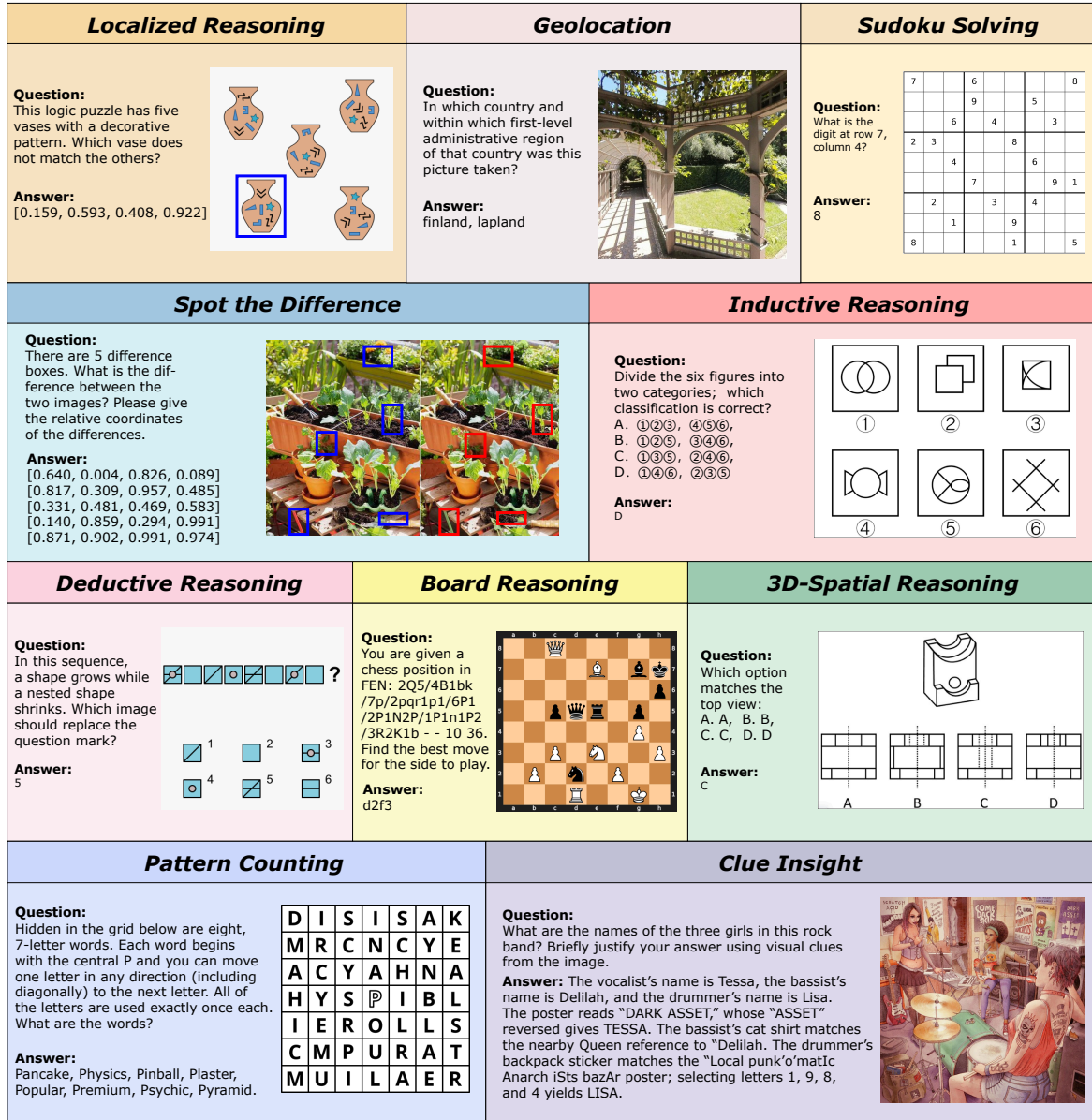


Figure 2: VisReason examples from different reasoning categories, covering everyday visual reasoning scenarios where perception and reasoning are tightly coupled.

like REBUS (Gritsevskiy et al., 2024) and VGRP-Bench (Ren et al., 2025) focus on narrower aspects of vision-centric reasoning, typically constrained to games, symbols, or rule-based settings. VisReason complements prior work by targeting common-context vision-centric reasoning across a broader spectrum of visual scenarios. It systematically covers perceptual, structural, and conceptual reasoning behaviors, organizing tasks into 10 categories and 36 subcategories to enable more comprehensive evaluation.

3 The VisReason Benchmark

3.1 Overview of VisReason

We propose VisReason, a multimodal benchmark designed to evaluate foundation models on vision-centric reasoning in everyday, perceptually grounded contexts. VisReason systematically spans perceptual, structural, and conceptual reasoning, covering abilities from low-level visual grounding to higher-level abstraction and inference. As illustrated in Figure 3, which summarizes the category structure, the benchmark comprises 10 reasoning categories organized into 36 fine-grained sub-

categories. Figure 2 further provides representative examples from each category. Together, these categories capture a broad spectrum of visual reasoning phenomena and emphasize tasks that require direct reasoning over visual evidence, rather than relying primarily on language priors or knowledge-intensive inference.

3.2 Data Curation Process

Data Collection. We first compile a candidate set of 10 visual reasoning categories. Guided by this candidate set, we curate questions from online sites and Chinese civil service exam repository. During data collection, we explicitly prioritized vision-centric items whose solutions hinge on extracting and integrating visual evidence, with a particular focus on graphic/diagrammatic reasoning problems that require interpreting shapes, layouts, and fine-grained visual cues rather than relying on text-only inference. During data collection, we manually verified that the selected samples do not contain personally identifying information.

Question Revision. After collecting the data, we recruited graduate students to convert the raw samples into a standardized visual question answering format. Firstly, we categorize questions by answer type into four forms: multiple-choice question, fill in the blank, open ended solution, and 2D bounding box selection. Each sample is independently judged by at least three annotators, and the final type is determined by majority vote. Secondly, we standardize questions and answers by format. For multiple-choice, we ensure the question contains explicit option indices and texts, and we restrict the answer to the option index only. For fill-in-the-blank question, we remove any extraneous content from the answer, keeping only the filled text. For open-ended, we ensure the question poses a clear, specific query. For bounding-box, annotators manually mark the target instance in the image and convert it to the $[x1, y1, x2, y2]$ format as the final answer.

Manual Validation. To ensure data quality, we conducted rigorous manual validation for all entries. Each image-question-answer triplet was independently reviewed by two annotators, who verified basic correctness and filtered out problematic cases. Disagreements were resolved through discussion and adjudication. During review, annotators evaluated multiple aspects of data quality. (1) Completeness and consistency: We checked

that the image, question, and answer are mutually consistent and that no essential information is missing. (2) Visual evidence sufficiency: Annotators assessed whether the key visual cues required to solve the question are clearly observable in the image and support a unique, well-defined answer. (3) Answer clarity: For multiple-choice questions, we verified that all candidate options are explicitly specified in the question text or visually present in the image, ensuring that each question is self-contained and the correct option is unambiguous. (4) Multimodal necessity and knowledge scope: We removed items that could be answered using only visual information or only textual information. In addition, we filtered out questions that depend on external knowledge beyond the provided context, so that solving each item relies primarily on reasoning over the given visual and textual evidence.

3.3 Reasoning Category

Each question in VisReason is manually assigned to a specific reasoning category. The 10 categories are organized under the three-level framework of *perceptual*, *structural*, and *conceptual* reasoning, spanning from low-level visual grounding to high-level abstraction and inference. Together, they provide a structured view of vision-centric reasoning behaviors grounded in visual evidence and commonly encountered in real-world scenarios.

- **Localized Reasoning:** Inferring the target instance and localizing it with a bounding box.
- **Spot the Difference:** Identifying visual differences between two scenes and localizing each difference with bounding boxes.
- **Pattern Counting:** Identifying all target patterns in complex scenes and counting their occurrences.
- **3D-Spatial Reasoning:** Reasoning about implicit 3D geometric relationships to infer the correct answer.
- **Board Reasoning:** Inferring the correct outcome given a board-game configuration.
- **Sudoku Solving:** Solving multiple variants of visually grounded Sudoku puzzles.
- **Geolocation:** Extracting geographic cues from the visual scene to infer its location.
- **Cue Insight:** Extracting and integrating textual and visual clues to derive the answer.
- **Inductive Reasoning:** Inferring underlying rules from observed instances and applying them to novel cases.
- **Deductive Reasoning:** Deriving logically en-

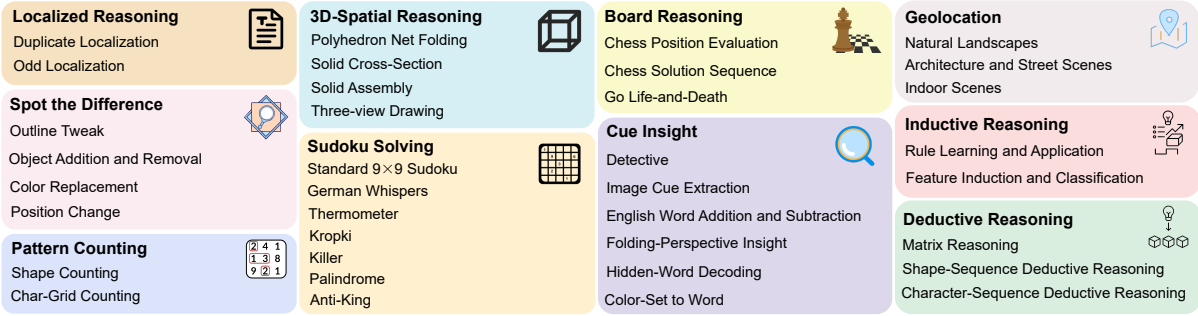


Figure 3: Domain diversity in VisReason. The benchmark covers 10 reasoning categories and 36 fine-grained subcategories, illustrating the breadth of vision-centric reasoning tasks.

Table 1: Key statistics of VisReason.

Statistic	Number
#Total questions	1505
- Multiple-choice	615 (40.9%)
- Fill-in-the-blank	441 (29.3%)
- Open-ended	309 (20.5%)
- Bounding-box	140 (9.3%)
#Reasoning categories	10
#Subcategories	36
Question length (min / avg / max)	5 / 45.9 / 282
Answer length (min / avg / max)	1 / 9.1 / 144

tailed conclusions from the given visual and contextual information.

3.4 Dataset Statistics

Table 1 summarizes the key statistics of VisReason. The benchmark contains 1,505 questions designed to support systematic evaluation of vision-centric reasoning across diverse settings.

VisReason is organized into 10 reasoning categories that span multiple levels of visual cognition, from perceptual grounding to abstract inference. Building on these core categories, the dataset further covers 36 fine-grained subcategories, capturing a broad spectrum of visual reasoning phenomena. Representative examples from each category and subcategory are provided in the Appendix.

VisReason includes four question formats that differ in how answers are specified and constrained. Multiple-choice, fill-in-the-blank, and open-ended questions impose different levels of answer openness, while all requiring reasoning over the visual input. In addition, VisReason contains a set of bounding-box questions arising from the Localized Reasoning and Spot-the-Difference categories, where models are required to infer the correct answer and identify its spatial location within the image. This format explicitly involves spatial lo-

calization as part of the reasoning process.

Overall, the dataset exhibits substantial linguistic and contextual diversity. Question lengths range from short prompts to complex multi-sentence descriptions, with an average length of 45.9 tokens.

4 Experiments

4.1 Experimental Setup

Metrics. We report accuracy (%) for all question types in VisReason, with evaluation protocols adapted to the corresponding answer formats. For multiple-choice and fill-in-the-blank questions, the final answer is extracted using regular expressions and directly compared with the ground truth. For open-ended questions, we employ GPT-5-mini as an automatic judge to assess correctness given the question and reference answer under a strict criterion; the full judging prompt is provided in the Appendix. For bounding-box questions, a single question may involve multiple ground-truth boxes. Predicted boxes are matched to ground truth using IoU-based Hungarian matching, and a match is considered correct if $\text{IoU} > 0.5$. This threshold was selected based on preliminary experiments, as it produces a more reasonable distribution of model performance and avoids settings that are either too loose or too strict. Accuracy is computed with partial credit and normalized to 1 per question.

Baseline Models. We evaluate VisReason on a diverse set of proprietary and open-source MLLMs, covering a wide range of model architectures, parameter scales, and reasoning capabilities.

Proprietary Models. We include (1) GPT-5.2, GPT-5, GPT-5-mini, o4-mini; (2) Gemini-2.5-Flash, Gemini-3-Pro. For the flagship GPT-5 and Gemini-3 models, their reasoning effort is set to *low* for cost control and evaluation consistency.

Open-source Models. We evaluate a broad

Table 2: Comparison of model performances across reasoning categories on VisReason. Gray rows indicate non-thinking models, while others employ explicit reasoning mechanisms.

Models	Avg.	Loc. Reas.	Spot Diff.	Pattern Count.	3D-Spat. Reas.	Board Reas.	Sudoku Solv.	Geolocation	Cue Insight	Ind. Reas.	Ded. Reas.
		Perceptual			Structural			Conceptual			
Human Performance	71.4	81.7	77.4	66.9	71.6	56.9	45.6	80.0	74.8	80.6	78.1
<i>Proprietary Models</i>											
GPT-4o	15.6	6.3	0.8	8.7	25.4	18.5	7.5	5.5	27.1	23.6	32.1
o4-mini	27.3	3.8	2.7	21.7	27.7	26.7	15.5	55.5	46.0	36.4	36.9
GPT-5-nano	19.0	3.8	0.1	17.4	23.9	22.2	9.0	29.0	29.7	29.8	24.6
GPT-5-mini (low)	28.1	12.5	2.2	15.2	30.8	30.4	15.0	55.0	45.1	37.5	37.7
GPT-5-mini (medium)	31.8	7.5	2.0	32.6	28.5	36.3	12.5	56.5	55.9	41.8	44.4
GPT-5-mini (high)	33.2	12.5	1.1	37.0	26.9	42.2	8.0	61.0	60.4	41.8	41.4
GPT-5	33.2	15.0	2.8	21.7	26.2	44.4	22.5	66.5	53.2	40.7	38.8
GPT-5.2	35.5	24.6	4.7	26.1	30.8	43.7	38.2	52.5	43.2	42.3	48.9
Gemini-2.5-Flash	16.0	0.0	0.5	8.7	31.5	15.6	6.0	12.5	19.8	31.6	34.0
Gemini-3-Pro	47.5	44.2	4.8	50.0	43.9	51.1	30.0	76.0	68.5	53.1	53.7
<i>Open-source Models</i>											
InternVL3-8B	10.9	0.0	0.1	2.2	31.5	6.7	8.5	11.0	9.9	19.3	20.2
InternVL3-14B	10.4	1.3	0.1	6.5	22.3	5.9	2.5	14.0	12.6	20.0	19.0
Qwen2.5-VL-7B-Instruct	7.2	0.0	0.3	0.0	18.5	1.5	1.5	0.0	9.0	22.2	19.0
Qwen2.5-VL-32B-Instruct	15.4	3.3	1.5	6.5	27.7	8.2	8.0	29.0	21.6	25.1	23.5
Qwen3-VL-8B-Instruct	18.8	1.3	1.1	8.6	22.8	13.1	9.2	29.5	34.3	33.7	34.1
Keye-VL-1.5-8B	10.9	0.0	0.8	10.9	19.2	6.7	4.5	5.0	16.6	24.9	20.5
MiMo-VL-7B-RL	15.9	2.1	0.7	10.6	25.4	17.0	2.5	37.5	12.6	24.7	25.4
Qwen3-VL-2B-Thinking	3.7	2.5	0.2	2.2	3.9	2.2	4.0	3.0	5.4	8.0	6.0
Qwen3-VL-8B-Thinking	19.0	9.6	0.7	6.5	33.1	14.1	6.0	25.3	25.2	35.0	34.0
Qwen3-VL-30B-A3B-Thinking	20.0	0.0	0.5	15.2	30.8	17.0	8.5	22.5	28.4	34.9	37.3
Qwen3-VL-32B-Thinking	22.1	1.3	1.3	17.4	30.8	27.4	7.0	31.0	27.0	38.6	39.2
Qwen3-VL-235B-A22B-Thinking	26.7	7.1	1.6	19.6	36.2	30.4	4.0	46.0	30.6	46.2	45.2

collection of leading open models, including (1) Qwen-VL series, including Qwen2.5-VL (Bai et al., 2025b) (7B-Instruct, 32B-Instruct), Qwen3-VL (Bai et al., 2025a) (8B-Instruct, 2B-Thinking, 8B-Thinking, 32B-Thinking, 30B-A3B-Thinking, 235B-A22B-Thinking); (2) InternVL3 (Zhu et al., 2025) (8B, 14B); (3) MiMo-VL-7B-RL (Team et al., 2025); (4) Keye-VL-1.5-8B (Yang et al., 2025).

All models are evaluated in a zero-shot setting. We design four prompt templates corresponding to the four question formats, all following a Chain-of-Thought (Wei et al., 2022) (CoT) prompting style. The complete prompt templates are provided in the Appendix.

Human Performance. We additionally include human performance to contextualize model results and to verify that VisReason tasks are reliably solvable given the provided visual evidence; the human evaluation protocol is detailed in the Appendix.

4.2 Overall Results

VisReason reveals a substantial gap between current MLLMs and human-level vision-centric reasoning. Table 2 reports model performance across all reasoning categories. Overall accuracy remains low across the board: the best-performing model achieves only 47.5% average accuracy, while many models fall below 20%. In contrast, human performance reaches 71.4%, indicating that VisReason is reliably solvable by humans under the provided visual evidence but remains highly challenging for current models. This gap confirms that VisReason effectively targets vision-centric reasoning abilities that are not yet well captured by existing MLLMs. **Proprietary vs. Open-Source Models.** State-of-the-art vision-centric reasoning is still dominated by proprietary models. A clear performance stratification emerges between proprietary and open-source systems. The strongest proprietary model, Gemini-3-Pro, achieves an overall accuracy of 47.5%, whereas the best open-source model, Qwen3-VL-235B-A22B-Thinking, reaches

only 26.7%. This gap suggests that large-scale pretraining and stronger cross-modal alignment remain critical for effective vision-centric reasoning. **VisReason Clearly Differentiates Model Capabilities.** Performance varies widely across models, with several weaker systems performing near chance level. For instance, Qwen3-VL-2B-Thinking attains an overall accuracy of only 3.7%, while a substantial gap separates such models from the strongest systems. This spread indicates that VisReason distinguishes model capabilities across a wide range of capacities.

4.3 Where Do MLLMs Struggle in Vision-Centric Reasoning?

Despite steady progress in multimodal modeling, VisReason reveals that current MLLMs struggle when reasoning must be tightly grounded in visual evidence. As shown in Table 2, across reasoning subjects, performance varies in a systematic but non-uniform manner: tasks that can be solved using coarse visual cues or scene-level abstraction are handled better, whereas tasks requiring fine-grained perceptual grounding and spatial precision remain challenging for most models.

These limitations are most evident in perception-grounded tasks that explicitly involve spatial localization. On Localized Reasoning and Spot-the-Difference, average model accuracy is only 6.9% and 1.3%, respectively, indicating substantial difficulty across all evaluated systems. In contrast, human performance on the same categories reaches 81.7% and 77.4%, confirming that these tasks are well-defined and reliably solvable given the visual evidence. The large human–model gap highlights persistent challenges in aligning inferred reasoning outcomes with localized visual evidence and in validating perceptual hypotheses against specific image regions.

4.4 Does CoT Help in Answering Vision-Centric Questions?

Figure 5 compares GPT-4o performance under two prompting settings: with and without CoT prompting. Overall, CoT yields only a marginal gain on VisReason, improving average accuracy by +1.1%. This improvement is smaller than the commonly assumed effectiveness of CoT for reasoning. A category-wise breakdown shows that CoT mainly helps higher-level reasoning tasks, such as Cue Insight and inductive or deductive reasoning, while providing little benefit or even degrading perfor-

mance on Spot-the-Difference, 3D-Spatial Reasoning, and Pattern Counting. These results suggest that longer reasoning traces alone are insufficient when task difficulty is dominated by visual perception and grounding.

4.5 How Much Does Compute Scaling Improve Vision-Centric Reasoning?

Scaling Model Capacity. Figure 4 (left) reveals a clear scaling trend within model families: accuracy generally improves as model parameters increase. This effect is most evident along the upward trajectories of the Qwen2.5 and Qwen3 series, where larger variants consistently outperform their smaller counterparts. Overall, increasing model capacity provides a reliable improvement for vision-centric reasoning on VisReason.

Increasing Inference-Time Thinking Budget. Figure 4 (right) shows that increasing the thinking budget leads to consistent but gradual accuracy improvements for the same model. This trend is most evident for GPT-5-mini, where moving from low to moderate token budgets yields noticeable gains, while further increases at higher budgets result in diminishing improvements, indicating that additional test-time computation offers limited leverage in the absence of stronger visual grounding.

4.6 Are Reasoning Abilities Correlated Across Categories?

Figure 6 shows the Pearson correlation of model performance across the 10 reasoning categories in VisReason. Overall, correlations are consistently positive and relatively high, ranging from 0.44 to 0.97, indicating that stronger models tend to perform well across multiple categories and reflecting a shared foundation of vision-centric reasoning ability. High-level reasoning categories are particularly aligned. Inductive and Deductive Reasoning exhibit the strongest correlation at 0.97, and Pattern Counting, Board Reasoning, and Cue Insight also form a tightly correlated group. In contrast, 3D-Spatial Reasoning, Geolocation, and Sudoku Solving show noticeably lower correlations with other categories, suggesting that these tasks rely on more specialized capabilities such as spatial imagination, geometric consistency, and strict constraint satisfaction that are not strongly predicted by overall model strength. These patterns indicate that VisReason captures both general reasoning competence and distinct category-specific challenges.

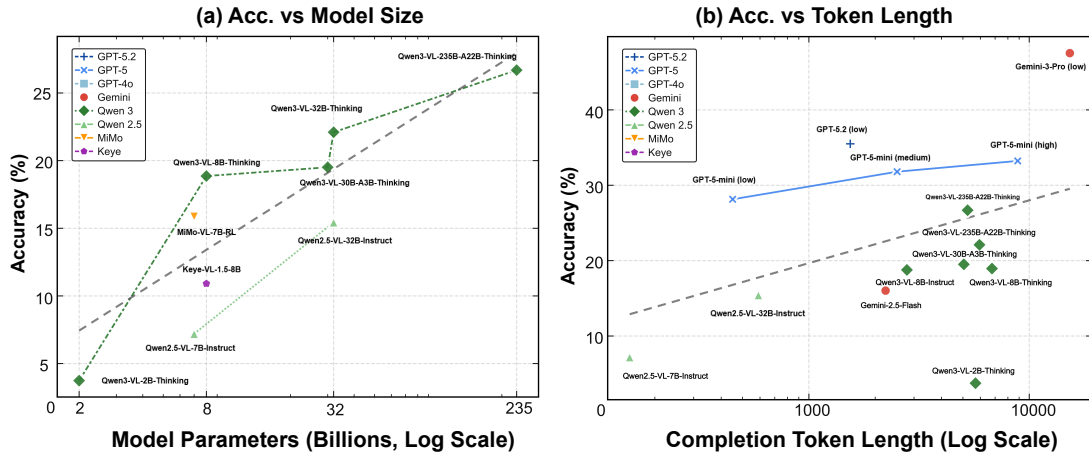


Figure 4: Scaling behavior of models on VisReason. **Left:** Accuracy versus model size (log scale). **Right:** Accuracy versus average inference-time tokens (log scale).

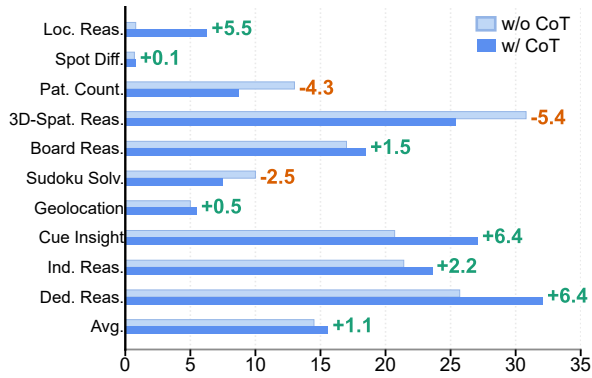


Figure 5: Effect of CoT prompting on GPT-4o performance across different reasoning categories.

4.7 Error Analysis

Figure 7 summarizes the primary error types identified through manual analysis of incorrect predictions on VisReason. Most mistakes fall into perceptual and reasoning errors, which together account for roughly 90% of all cases. Perceptual errors mainly arise from difficulties in scene parsing, visual grounding, and counting, while reasoning errors are largely due to selecting incorrect rules or producing hallucinated inferences, rather than simple calculation or step omissions. Format-related errors are less frequent but remain noticeable, especially for bounding-box questions with stricter output constraints.

5 Conclusion

We introduce VisReason, a benchmark designed to examine whether current MLLMs can reason in a vision-centric manner, beyond relying on language or domain knowledge. Experiments across a

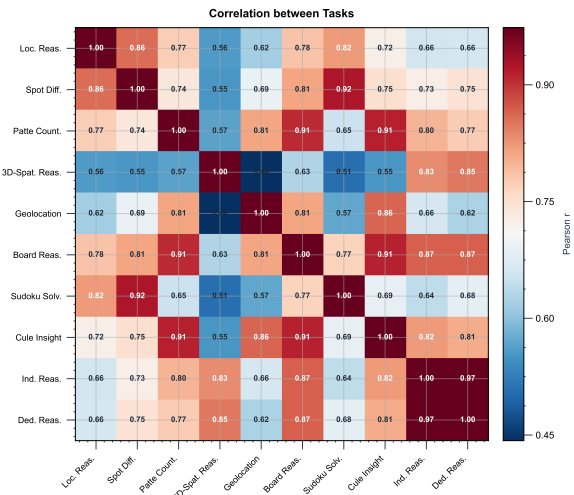


Figure 6: Pearson correlation of model accuracy across the 10 reasoning categories on VisReason.

broad set of models show that vision-centric reasoning in common visual contexts remains challenging, with a substantial gap to human performance even for strong contemporary systems. Our results suggest that gains from increased inference-time reasoning are meaningful but limited, and do not fully translate into robust reasoning from visual evidence. These findings indicate that advancing vision-centric reasoning will require deeper integration of visual perception and reasoning processes. We hope VisReason will help clarify this challenge and support progress toward multimodal models that genuinely reason beyond language.

6 Limitations

VisReason is designed as an evaluation benchmark and has several limitations. It focuses on static

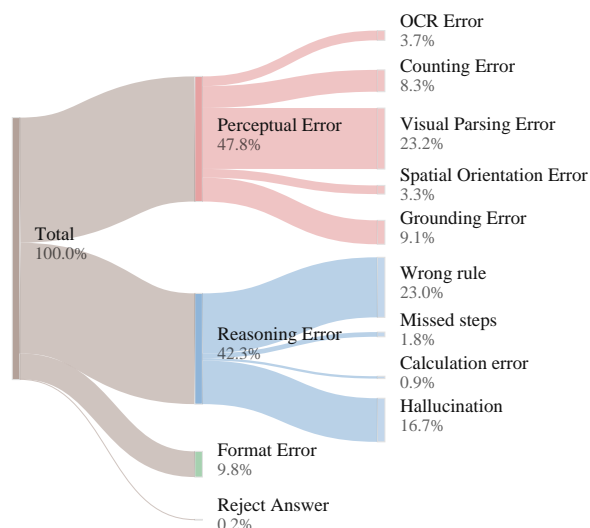


Figure 7: Distribution of error types for Qwen3-VL-235B-A22B-Thinking on VisReason.

images and therefore does not capture temporal dynamics, interactive perception, or embodied reasoning. While the benchmark spans a wide range of vision-centric reasoning types in common visual contexts, it cannot exhaustively cover the full diversity and complexity of real-world scenarios. We view these limitations as natural directions for future extensions, including dynamic settings and richer forms of vision-centric reasoning. A potential risk of benchmark-style resources such as VisReason is contamination or benchmark-specific overfitting as models and training corpora evolve, which may weaken the reliability of evaluation results over time. Although our automatic update framework can mitigate these issues to some extent, it cannot fully eliminate such risks.

7 Acknowledgment

This research is supported by Artificial Intelligence-National Science and Technology Major Project (2023ZD0121200), and the National Natural Science Foundation of China (62437001, 62436001, 62531026), the Key Research and Development Program of Jiangsu Province under Grant BE2023016-3, and the Natural Science Foundation of Jiangsu Province under Grant BK20243051.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng

Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv: 2511.21631*.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Andrew Gritsevskiy, Arjun Panickssery, Aaron Kirtland, Derik Kauffman, Hans Gundlach, Irina Gritsevskaya, Joe Cavanagh, Jonathan Chiang, Lydia La Roux, and Michelle Hung. 2024. *Rebus: A robust evaluation benchmark of understanding symbols*. *Preprint*, arXiv:2401.05604.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *arXiv preprint arXiv: 2410.21276*.

OpenAI. 2023. *GPT-4V(ision) System Card*. Accessed: 2025-04-18.

Yufan Ren, Konstantinos Tertikas, Shalini Maiti, Junlin Han, Tong Zhang, Sabine Süsstrunk, and Filippos Kokkinos. 2025. Vgrp-bench: Visual grid reasoning puzzle benchmark for large vision-language models. *arXiv preprint arXiv:2503.23064*.

Jeffrey Seely, Yuki Imajuku, Tianyu Zhao, Edoardo Cetin, and Llion Jones. 2025. *Sudoku-bench: Evaluating creative reasoning with sudoku variants*. *Preprint*, arXiv:2505.16135.

- Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, and 55 others. 2025. [Mimo-vl technical report](#). *Preprint*, arXiv:2506.03569.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, Fan Yang, Guorui Zhou, Guowang Zhang, Han Shen, Hao Peng, Haojie Ding, Hao Wang, Haonan Fan, Hengrui Ju, and 42 others. 2025. [Kwai keye-vl 1.5 technical report](#). *Preprint*, arXiv:2509.01563.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. 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 (CVPR)*, pages 9556–9567.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhao Chen, and Graham Neubig. 2025. [MMMU-pro: A more robust multi-discipline multimodal understanding benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, Vienna, Austria. Association for Computational Linguistics.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Additional Dataset Details

A.1 Category Distribution

Figure 8 shows the distribution of questions across the 10 reasoning categories in VisReason. The dataset is designed to span a wide range of vision-centric reasoning phenomena, covering perceptual grounding, structured spatial reasoning, and higher-level inference. Conceptual reasoning categories capture abstract inference based on visual evidence, while structural categories focus on spatial relations, rule-based reasoning, and state consistency. In addition, perception-grounded categories are included to explicitly evaluate fine-grained visual grounding and spatial verification. This composition reflects the diversity of reasoning types encountered in everyday visual contexts and supports complementary analyses across different reasoning dimensions.

A.2 Comparisons with Existing Benchmarks

Table 4 further distinguish the difference between VisReason and other existing ones. Regarding answer Forms, most existing datasets use a single format. In contrast, VisReason supports four forms: multiple-choice, fill-in-the-blank, open-ended, and bounding-box. Diverse answer forms enable a comprehensive evaluation of model capabilities. Regarding reasoning types, current benchmarks are predominantly STEM oriented and focus on structured data domains, while overlooking vision based reasoning in everyday contexts. VisReason advances beyond these limitations by focusing on human everyday visual reasoning. Finally, VisReason provides a richer set of categories and subcategories spanning diverse aspects of visual cognition,

enabling a more comprehensive analysis of models' visual reasoning capabilities.

B Additional Experimental Details

B.1 Human Performance Protocol

To provide a human reference for comparison, we recruited 10 graduate-student volunteers to answer questions from VisReason. Human evaluation serves to contextualize model performance and offers a reference point for understanding task difficulty under the same input conditions. To ensure adequate coverage and task familiarity, we selected volunteers with complementary strengths aligned with the domains covered by VisReason and assigned questions according to each participant's area of proficiency, such that each subject was answered by individuals familiar with the corresponding task style.

All participants completed the tasks under the same information constraints as the models: they were provided only with the image and question text, without access to external tools, AI assistants, search engines, or the internet. Volunteers followed the same answer-format requirements as specified in the benchmark using a lightweight annotation interface. We report the aggregated accuracy across all participants as the human performance reference.

B.2 Human Participants and Data Sources

Instructions to annotators and volunteers. For dataset construction, graduate-student annotators were instructed to convert raw samples into a standardized visual question answering format, assign each item to one of four answer types, standard-

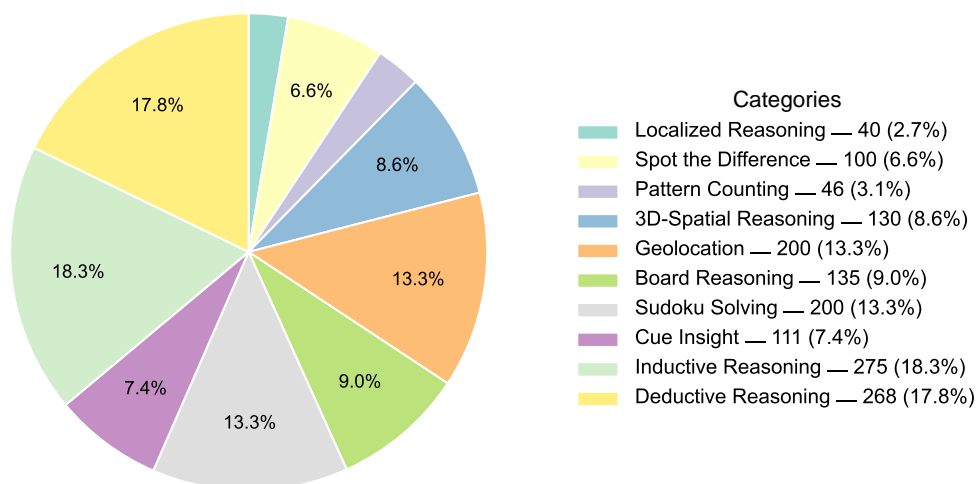


Figure 8: Category distribution of VisReason.

Table 3: Accuracy of different models on VisReason grouped by question form, including multiple-choice, fill-in-the-blank, open-ended, and bounding-box questions. Gray rows indicate non-thinking models, while others employ explicit reasoning mechanisms.

Models	Avg.	Question Forms			
		Multiple-choice	Fill-in-the-blank	Open-ended	Bounding-box
<i>Proprietary Models</i>					
GPT-4o	15.6	25.2	18.6	10.4	3.6
o4-mini	27.3	30.9	30.4	48.5	3.3
GPT-5-nano	19.0	23.6	20.2	29.8	2.0
GPT-5-mini (low)	28.1	32.7	28.3	50.5	7.4
GPT-5-mini (medium)	31.8	36.4	31.8	55.3	4.8
GPT-5-mini (high)	33.2	34.3	32.4	60.2	6.8
GPT-5	33.2	34.0	35.8	61.5	8.9
GPT-5.2	35.5	39.4	41.8	52.1	14.7
Gemini-2.5-Flash	16.0	31.5	16.3	12.0	0.3
Gemini-3-Pro	47.5	47.8	47.4	72.5	24.5
<i>Open-source Models</i>					
InternVL3-8B	10.9	22.1	10.2	8.4	0.1
InternVL3-14B	10.4	19.0	8.8	11.7	0.7
Qwen2.5-VL-7B-Instruct	7.2	20.0	4.1	2.9	0.2
Qwen2.5-VL-32B-Instruct	15.4	22.9	14.5	23.9	2.4
Qwen3-VL-8B-Instruct	18.8	27.8	20.8	28.6	1.2
Qwen3-VL-2B-Thinking	3.7	2.8	9.5	2.6	1.4
Qwen3-VL-8B-Thinking	19.0	30.6	17.5	25.1	5.2
Qwen3-VL-30B-A3B-Thinking	20.0	30.9	20.5	25.6	0.3
Qwen3-VL-32B-Thinking	22.1	34.5	22.2	29.8	1.3
Qwen3-VL-235B-A22B-Thinking	26.7	40.8	21.1	43.7	4.4

ize the question and answer format, and manually verify correctness, completeness, and multimodal necessity. For human evaluation, participants were instructed to answer each question using only the provided image and question text, without access to external tools, AI assistants, search engines, or the internet, and to follow the same answer-format

requirements as specified in the benchmark.

Recruitment and compensation. The annotators and human evaluators were graduate students recruited from our institution. Human evaluation involved 10 graduate-student volunteers. Each student participant was compensated 1,000 RMB. We consider this compensation appropriate for

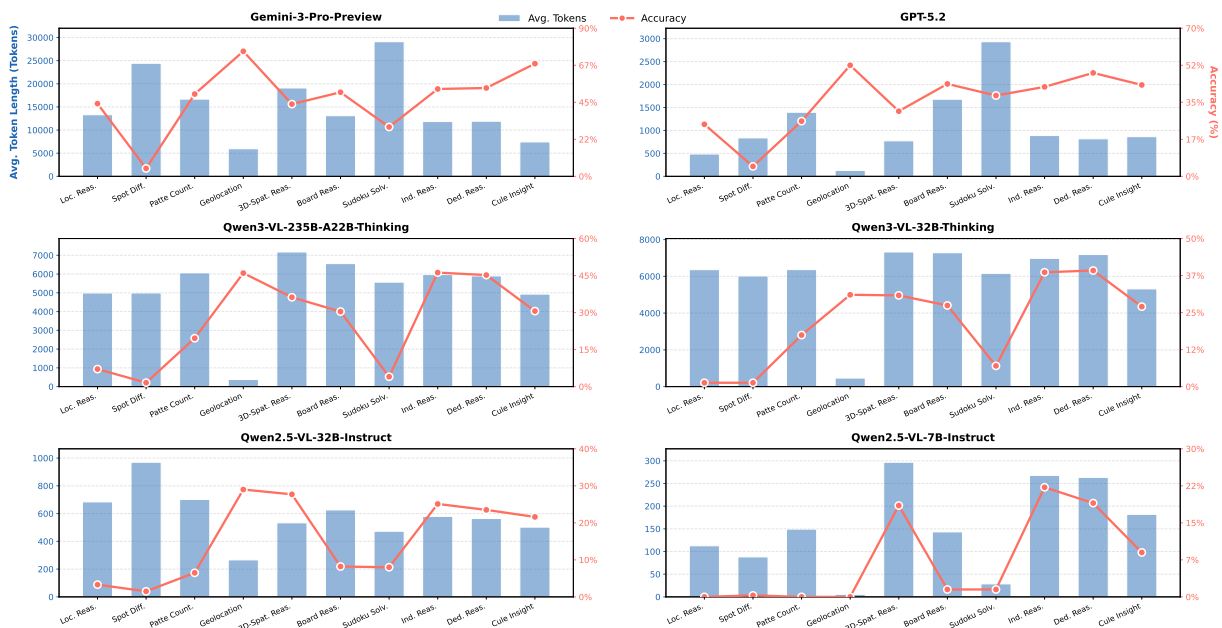


Figure 9: Token length and accuracy across categories for different models.

Table 4: Statistics for VisReason and comparisons with existing datasets. #Q: number of questions; Answer Forms: including multiple-choice(MC), fill-in-the-blank(FIB), open-ended(OE), and bounding-box(Bbox); Reasoning Types: the primary reasoning task categories covered by the dataset, SHAPE refers to Social Sciences, Humanities and the Arts for People and the Economy, and STEM refers to Science, Technology, Engineering, and Mathematics; Categories: the first-level taxonomy grouping related tasks; Subcategories: the second-level taxonomy that refines each category into more specific types.

Name	#Q	Answer Forms	Reasoning Types	Categories	Subcategories
WikiDiverse	8.0k	OE	Multimodal Entity Linking	10	7
MathVista	6.1k	MC	Mathematics	1	12
MATH-Vision	3.0k	MC	Mathematics	1	12
MMMU	11.5k	MC, OE	SHAPE, STEM	30	183
MMMU-pro	3.5k	MC	SHAPE, STEM	30	183
EMMA	2.7k	MC, OE	STEM	3	14
LogiQA	8.7k	MC	Deductive Logical Reasoning	1	5
Sudoku-Bench	2.6k	OE	Sudoku	3	3
LogicBench	2.0k	MC	Formal Logic	3	25
LogicVista	0.5k	MC	Multimodal Logical Reasoning	5	9
ARC-AGI-1	1.0k	OE	Abstract Reasoning	1	4
VisReason (Ours)	1.5k	MC, FIB, OE, Bbox	Everyday Visual Reasoning	10	36

graduate-student participants in our local context and for the time required to complete the assigned tasks.

Data sources and consent. The images and questions in VisReason were collected from publicly accessible websites and Chinese civil service exam repositories through URL-based access. We only used publicly available materials intended for open access and academic research. During data collection, we manually checked that the selected samples do not contain personally identifying information. Since the data were collected from public sources rather than through direct interaction with individuals, and no personally identifying information was retained in the released benchmark, we did not obtain individual consent from depicted persons.

B.3 Token Length Across Categories

Figure 9 visualizes how different models allocate generation budget across VisReason categories by reporting the average tokens generated per query, together with the corresponding accuracy. The results show substantial variation in token length across categories and models, reflecting different patterns of computation usage when solving vision-centric reasoning tasks.

B.4 Do MLLMs Already Know the Rules?

For several categories, the task requirements are largely specified in the question itself. By contrast, board reasoning and Sudoku solving rely on basic rules that cannot be fully restated in the prompt

without changing the nature of the task. To verify that this does not compromise the validity of our benchmark, we constructed two small entry-level sets with ten easy board reasoning problems and ten easy Sudoku problems in the same format as the main evaluation. Qwen3-VL-30B-A3B-Thinking, GPT-5-mini, and Gemini-2.5-Flash all achieved 100 percent accuracy on these problems. This result indicates that these models already know the basic rules required for such tasks. Therefore, VisReason does not primarily measure rule acquisition in these categories, and remains a valid benchmark for evaluating vision-centric reasoning.

C Analysis by Question Form

Table 3 reports model performance across different question forms. Overall, multiple-choice and fill-in-the-blank questions achieve higher accuracy than open-ended formats, reflecting the effect of answer constraints on reducing ambiguity. Open-ended questions show larger performance variance across models, with proprietary systems generally exhibiting stronger reasoning robustness. Bounding-box questions yield the lowest accuracy across models; however, this should not be interpreted as a lack of localization capability. Instead, these questions require models to perform vision-grounded reasoning while simultaneously maintaining and verifying spatial hypotheses during the reasoning process. The low accuracy therefore reflects the difficulty of integrating precise visual grounding into multi-step reasoning

D Continual Benchmark Updates

VisReason is intended not only as a benchmark dataset, but also as a scalable framework for continual evaluation. By supporting automatic or semi-automatic updates, this framework enables new benchmark versions to be generated as models and training corpora evolve. Such an update mechanism is valuable for three reasons: first, it helps keep the benchmark fresh and reduces the risk of contamination or saturation; second, it allows performance to be tracked under versioned and reproducible evaluation settings over time; third, it provides a practical foundation for community-driven extension, making VisReason a resource that can be maintained and expanded beyond the initial release.

The update pipeline for each VisReason category is briefly described as follows.

- **Localized Reasoning.** This category can be updated by tracking visual puzzle websites and adding 2D bounding boxes either manually or with grounding detectors. In addition, suitable multiple-choice items from other categories can be adapted into localization-style questions.
- **Spot the Difference.** This category supports a fully automatic pipeline based on instance detection, segmentation, controlled edits, and inpainting, which can be used to generate new image pairs and corresponding differences.
- **Pattern Counting.** This category can be generated automatically by rendering geometric shapes and constructing character grids under controlled generation rules.
- **3D Spatial Reasoning.** This category can be updated by collecting new items from annual Chinese civil service exams and engineering drawing exercises.
- **Board Reasoning.** This category can be updated from tournament game records and community puzzle sources such as online Go puzzles and Lichess.
- **Geolocation.** This category can be updated by mining image pairs from community sources, filtering leakage through OCR and model-based checks, and then applying manual verification.

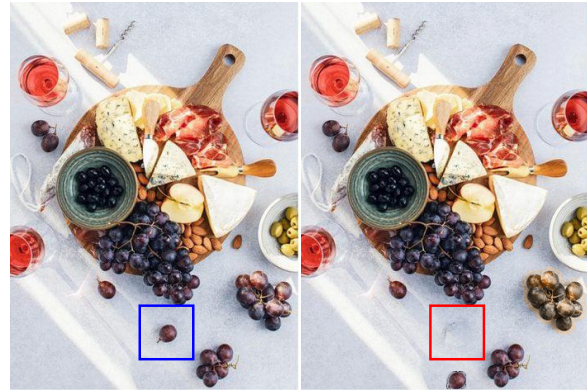


Figure 10: An example from the Spot the Difference category generated by our automatic update framework. **Left:** Original image. **Right:** Generated image with one grape removed.

- **Sudoku Solving.** This category can be updated with an automatic puzzle generator.
- **Cue Insight, Inductive Reasoning, and Deductive Reasoning.** These categories can be updated from civil service exams and visual puzzle websites with model-assisted categorization followed by human filtering. So far, we have collected several dozen new questions for each category.

Figure 10 shows an example from the Spot the Difference category generated by our automatic update framework.

E An Atomic-Capability Framework

To make the structure of vision-centric reasoning tasks more explicit, we introduce an atomic-capability framework that decomposes complex tasks into a small set of reusable visual reasoning primitives. The goal of this framework is not only to better organize VisReason itself, but also to provide a principled lens for analyzing shared requirements and common bottlenecks across different tasks. Such a perspective is valuable for future benchmark construction because it enables new tasks to be designed compositionally, supports capability-level diagnosis beyond category-level scores, and makes it easier to study how models generalize across related forms of visual reasoning. We hope this framework can serve as a lasting foundation for building, extending, and interpreting future benchmarks of vision-centric reasoning.

Table 5 summarizes the eight atomic visual reasoning capabilities in our framework. Therefore,

Table 5: Atomic visual reasoning capabilities used to organize and analyze VisReason.

Capability	Abbr.	Description
Layout Understanding	LU	Parsing multiple regions, spatial relations, and reading order.
Structured State Extraction	SSE	Converting grids, boards, or panels into structured representations.
Symbol and Entity Recognition	SER	Reading digits and letters, and identifying piece types, shapes.
Correspondence and Difference	CD	Aligning multiple views and localizing differences.
Relational Reasoning	RR	Modeling row or column constraints and move legality.
Rule Learning and Logical Inference	RLI	Inducing rules from examples or reasoning under them.
Search and Decision	SD	Performing constraint solving, backtracking, and decision making.
3D Hypothesis and Projection	3DHP	Forming 3D hypotheses and verifying them via projections.

Table 6: Capability-level results under the atomic visual reasoning framework. For each capability, we report the average accuracy over all VisReason categories that require it. Gray rows indicate the average performance of proprietary and open-source models, respectively.

Model	LU	SSE	SER	CD	RR	RLI	SD	3DHP
Human	71.36	64.17	70.69	77.40	68.28	74.42	62.80	71.60
o4-mini	27.29	21.30	30.02	2.70	25.08	35.72	21.10	27.70
GPT-4o	15.55	11.57	17.19	0.80	14.58	18.92	13.00	25.40
GPT-5-nano	18.95	16.20	21.04	0.10	19.60	23.38	15.60	23.90
GPT-5-mini-low	28.14	20.20	31.02	2.20	24.52	37.56	22.70	30.80
GPT-5-mini-medium	31.80	27.13	35.11	2.00	30.80	41.22	24.40	28.50
GPT-5-mini-high	33.23	29.73	36.80	1.10	32.00	44.14	25.10	26.90
GPT-5	33.18	29.53	36.89	2.80	32.08	42.04	33.45	26.15
GPT-5.2	35.50	36.67	38.81	4.70	39.05	42.12	40.95	30.80
Gemini-2.5-Flash	16.02	10.10	17.74	0.50	15.23	22.94	10.80	31.50
Gemini-3-Pro-Preview	47.53	43.70	52.06	4.80	49.78	59.10	40.55	43.90
Proprietary Models	28.72	24.48	31.67	2.17	27.89	36.60	24.77	29.56
Qwen2.5-VL-7B-Instruct	7.20	1.00	8.19	0.30	6.32	12.42	1.50	18.50
Qwen2.5-VL-32B-Instruct	15.44	7.57	17.65	1.50	12.70	20.50	8.10	27.70
Qwen3-VL-8B-Instruct	18.77	10.30	20.74	1.10	16.15	26.58	11.15	22.80
Qwen3-VL-2B-Thinking	3.74	2.80	4.13	0.20	3.08	5.18	3.10	3.90
Qwen3-VL-8B-Thinking	18.95	8.87	20.97	0.70	20.68	27.58	10.05	33.10
Qwen3-VL-32B-Thinking	22.10	17.27	22.19	1.30	25.10	27.42	17.20	30.80
Qwen3-VL-30B-A3B-Thinking	19.51	13.57	21.62	0.50	22.15	24.94	12.75	30.80
Qwen3-VL-235B-A22B-Thinking	26.69	26.87	28.92	1.60	32.90	35.02	17.20	36.20
InternVL3-8B	10.94	5.80	12.15	0.10	12.25	14.14	7.60	31.50
InternVL3-14B	10.42	6.10	11.58	0.10	13.68	14.86	4.20	22.30
MiMo-VL-7B-RL	15.85	10.03	17.53	0.70	16.18	20.44	9.75	25.40
Keye-VL-1.5-8B	10.91	7.37	12.03	0.80	10.68	13.08	5.60	19.20
Kimi-VL-A3B-Thinking-2506	4.21	1.47	4.66	0.20	3.48	4.10	2.95	10.80
Open-source Models	14.21	8.42	15.93	0.70	15.33	19.21	8.55	24.08

we represent each VisReason category as a composition of atomic visual reasoning capabilities:

- Localized Reasoning:** LU \oplus SER \oplus RLI
- Spot the Difference:** LU \oplus CD
- Pattern Counting:** LU \oplus SSE \oplus SER \oplus RR
- 3D Spatial Reasoning:** LU \oplus SER \oplus 3DHP
- Board Reas. :** LU \oplus SSE \oplus SER \oplus RR \oplus SD
- Sudoku Solving:** LU \oplus SSE \oplus SER \oplus RR \oplus SD
- Geolocation:** LU \oplus SER \oplus RLI
- Cue Insight:** LU \oplus SER \oplus RLI
- Inductive Reasoning:** LU \oplus SER \oplus RR \oplus RLI
- Deductive Reasoning:** LU \oplus SER \oplus RLI

As shown in Table 6, we add capability-level results to operationalize the atomic-capability view. For each capability, we average accuracies over

all categories that require it, and report per-model scores as well as proprietary and open-source averages. The results consistently identify Correspondence and Difference as the primary bottleneck: humans achieve 77.40, while proprietary and open-source models average 2.17 and 0.70. Structured State Extraction and Search and Decision are also challenging, whereas Rule Learning and Logical Inference is relatively stronger. We further report involvement analyses and capability-count trends, showing a large performance drop on categories that require Correspondence and Difference and a clear degradation as more capabilities are required, with Sudoku and Board Reasoning being the most affected.

F Prompt Templates

The prompt for response generation is as follows.

System Prompt with CoT

You are a highly intelligent question answering assistant.
{User Prompt}
{Questions} {Images}
You must think step by step.

System Prompt without CoT

You are a highly intelligent question answering assistant.
{User Prompt}
{Questions} {Images}
You must output only the final answer. Do not show any reasoning process or explanation.

User Prompt for Multiple-Choice Questions

Please answer the question from the given choices and put your final answer in one "\boxed{}".
There may be more than one correct option; please fill in all the options you consider correct in the \boxed{}

User Prompt for Fill-in-the-blank Questions

Please answer the question using a few words or phrases and put your final answer in one "\boxed{}".

User Prompt for Open-ended Questions

Please answer the question and summarize your answer concisely in one "\boxed{}".

User Prompt for Bounding-box Questions

First determine the required answer targets according to the task description, and then output bounding boxes only for these targets. Each bounding box must tightly cover exactly one answer target; do not include multiple objects or large regions in a single box. You must output exactly the number of bounding boxes specified in the question, no more and no fewer. Return a single array of bounding boxes in one "\boxed{}". Each bbox must be in the format [x1, y1, x2, y2], where (x1, y1) is the top-left corner and (x2, y2) is the bottom-right corner; different bboxes are separated by semicolons (";").

OCR Error

Question: Add lines to this grid and create five areas that each have 4 letters, to spell five 4-letter words. Please provide these five words.

Grounding Truth:
Trip, Trio, Open, Pale, and Palm

T	O	I	R
R	I	O	T
P	P	P	P
A	N	E	A
L	E	M	L

Qwen3-VL-235B-A22B-Thinking: To solve the puzzle, we are given a 5x4 grid of letters and must partition it into five connected regions, each containing exactly four letters, such that each region spells a valid 4-letter English word. --- ### ✓ Final Answer: The five 4-letter words formed by partitioning the grid into five connected regions are: 1. **PALE** - Cells: (3,1), (4,1), (5,1), (5,2) - Letters: P, A, L, E 2. **RIOT** - Cells: (2,1), (2,2), (2,3), (2,4) - Letters: R, I, O, T 3. **MEAL** - Cells: (4,3), (4,4), (5,3), (5,4) - Letters: M, E, A, L 4. **TOPE** - Cells: (1,1), (1,2), (3,2), (4,2) - Letters: T, O, P, N → Wait, this is not TOPE, but actually **TONE** - Correction: **TONE** - Cells: (1,1), (1,2), (4,2), (4,3) - Letters: T, O, N, E 5. **PILL** - Not valid. Reconsidering the final region. Instead, the correct fifth word is: 5. **PINE** - Cells: (3,2), (3,3), (4,2), (4,3) - Letters: P, I, N, E --- ### ✗

Summary: After analyzing all possible connected 4-letter groupings in the grid and verifying their validity as English words, the five valid 4-letter words that can be formed from the given grid using only contiguous cells are: - **PALE** - **RIOT** - **MEAL** - **TONE** - **PINE** Each word corresponds to a distinct, connected region of exactly four letters in the grid. - --- ### ✓ Final Answer in Box: **boxed{PALE, RIOT, MEAL, TONE, PINE}**

Error Reason: The model misreads the letter 'P' at grid position (3,3) as 'I' to form the word 'PINE'. The gold answer is 'Trip, Trio, Open, Pale, and Palm.', but the response depends on at least one misread token (letters/numbers/symbols) from the image, which changes the downstream reasoning.

Figure 11: OCR Error case of Qwen3-VL-235B-A22B-Thinking.

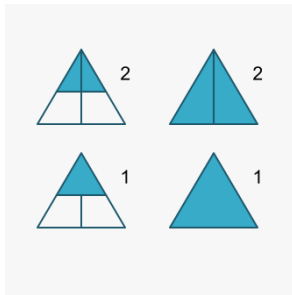
G Error Cases

We present several common types of errors made by different models during the reasoning process, as illustrated below.

Counting Error

Question: This simple riddle features a triangle subdivided by two intersecting lines. Count how many triangles appear in the picture. A triangle may be formed by multiple smaller triangles or by their outlines.

Grounding Truth: 6

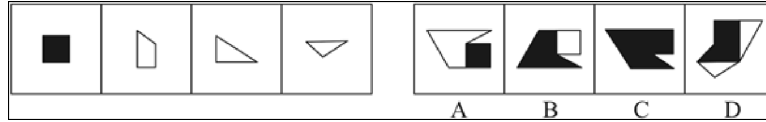


Qwen3-VL-235B-A22B-Thinking: To determine the number of triangles in the described figure, we analyze the structure formed by a large triangle subdivided by two intersecting lines: one vertical from the apex to the base and one horizontal cutting across the middle. --- ### Step-by-Step Count 1. ****The entire large triangle**** — this is the outermost triangle that encompasses all other regions. 2. ****Two medium-sized triangles**** — formed by the vertical line dividing the large triangle into left and right halves. 3. ****Two small triangles**** — formed at the top where the horizontal line intersects the sides and the vertical line. **The horizontal line divides the upper portion of the triangle into two smaller triangles (top-left and top-right), while the lower portions are quadrilaterals and do not form triangles on their own. No additional triangles can be formed from the quadrilateral regions because they lack the necessary triangular shape.** --- ### Final Count - 1 large triangle - 2 medium triangles - 2 small triangles **\boxed{5}**

Error Reason: The model undercounts by missing one specific triangle formed by the combination of the two small top triangles. The gold answer is '6', but the response's conclusion relies on an incorrect enumeration (missing, double-counting, or inconsistent counting rules). This is not just arithmetic: the failure occurs at the visual counting stage, so the derived total/choice is wrong even if later calculations are internally consistent.

Figure 12: Counting Error case of Qwen3-VL-235B-A22B-Thinking.

Visual Parsing Error



Question: Choose the most appropriate option: A. as shown above, B. as shown above, C. as shown above, D. as shown above

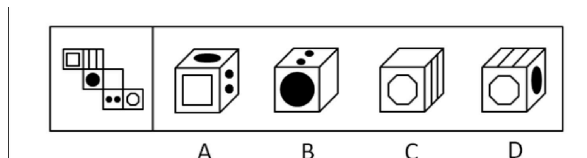
Grounding Truth: C

Qwen3-VL-235B-A22B-Thinking: To determine the most appropriate option, we analyze the sequence of shapes in the first row and how they relate to the options in the second row. --- ### **Understanding the Sequence** The top row contains four shapes: 1. A **solid black square** — this appears to be the base or starting shape. 2. A **white right trapezoid** — this could represent a portion of the square that has been cut out or transformed. 3. A **white right triangle** — possibly a further transformation or another part removed from the square. 4. A **white inverted triangle** — likely another geometric operation or removal. These white shapes appear to be **transformations or cuts** made on the original black square. The question seems to ask which of the four options (A–D) best represents the **resulting black area** after applying these transformations. --- ### **Analyzing the Options** Each option shows a black shape with some white regions. We are to find the one where the **black region corresponds to the remaining area of the square after subtracting the three white shapes**. **Let's examine Option B closely:** - It features a **black shape** that resembles a combination of a **trapezoid** and a **triangle**, with a **small white square** in the upper-right corner. - This white square may correspond to the **fourth shape** (inverted triangle), suggesting it's the last part to be removed. - The overall black shape aligns well with what remains if you **remove the right triangle (third shape)** and the **inverted triangle (fourth shape)** from the square, while preserving the **trapezoid-like structure (second shape)** as the main black area. This interpretation matches the logical progression: - Start with the square. - Remove the triangle (third shape). - Remove the inverted triangle (fourth shape). - The remaining **black area matches Option B**. --- ### **Why Other Options Don't Fit** - **Option A:** Contains a black square within a white trapezoid. This doesn't match the expected result of progressive subtraction from the square. - **Option C:** Has a large black shape with a notch on the right. While it might resemble a cut square, it lacks the specific structure of the trapezoid and triangle seen in the sequence. - **Option D:** Features a vertical black rectangle and an irregular white shape. It does not reflect the geometric transformations implied by the top row. --- ### **Conclusion** Only **Option B** accurately reflects the **remaining black area** after sequentially removing the three white shapes from the original black square. **boxed{B}**

Error Reason: "The model incorrectly claims a square notch in the option corresponds to an inverted triangle input, failing geometric composition. The gold answer is 'C', but the response builds a rule from an incorrect parse of the visual structure (local attributes, pattern changes, or composition relations). In effect, the model is reasoning over features that are not actually supported by the figure, so the final conclusion becomes systematically biased."

Figure 13: Visual Parsing Error case of Qwen3-VL-235B-A22B-Thinking.

Spatial Orientation Error



Question: Below is the net of the outer surface of a carton, which option on the right can be folded from it: A. A, B. B, C. C, D. D

Grounding Truth: A

Qwen3-VL-235B-A22B-Thinking: To determine which of the given options can be folded from the provided net, we must analyze the spatial relationships and adjacencies between the faces in the net. --- ### **Key Observations from the Net** The net consists of six distinct faces: 1. **Frame square** (a square with a smaller square inside) 2. **Three vertical stripes** 3. **Solid black circle** 4. **Empty square** 5. **Two small dots side by side** 6. **Octagon (white)** These faces are arranged in the following pattern: $[Frame][Stripes][Black][Empty][TwoDots][Octagon]$ This is a "zig-zag" type of net. When folded into a cube, the adjacency relationships can be deduced as follows: - **Frame** is adjacent to **Stripes** - **Stripes** is adjacent to **Frame**, **Black** - **Black** is adjacent to **Stripes**, **Empty** - **Empty** is adjacent to **Black**, **Two Dots** - **Two Dots** is adjacent to **Empty**, **Octagon** - **Octagon** is adjacent to **Two Dots** From this, we can infer the **opposite face pairs**: - **Frame** \leftrightarrow **Empty** - **Stripes** \leftrightarrow **Two Dots** - **Black** \leftrightarrow **Octagon** This means that **opposite faces cannot be adjacent** on the cube. --- ### **Evaluating Each Option** ### **Option A** - Front: Frame - Top: Black - Right: Two Dots - **Frame and Black** are adjacent in the net (via Stripes), so this is acceptable. - **Frame and Two Dots**: Since **Two Dots** is opposite **Stripes**, and **Frame** is not opposite **Two Dots**, they can be adjacent. - **Black and Two Dots**: In the net, **Black** is adjacent to **Empty**, which is adjacent to **Two Dots**, so they are also adjacent in the cube. However, **Two Dots** should be **bottom** if **Black** is front and **Empty** is right. The fact that it's shown on the **right** suggests an incorrect orientation. Thus, **this configuration is not consistent with the net's structure**. ✓ **Not valid** --- ### **Option B** - Front: Black - Top: Two Dots - **Black and Two Dots** are adjacent in the net (via Empty), so this is acceptable. - However, **Two Dots** should be **below Empty**, which would make it the **bottom face** when **Black** is front and **Empty** is right. - Showing **Two Dots** on the **top** contradicts the net's layout. ✓ **Not valid** --- ### **Option C** - Front: Octagon - Right: Stripes - **Octagon** is opposite **Black**, so **Black** is on the back. - **Stripes** is adjacent to **Black** (from the net), so it must also be adjacent to **Octagon** (since **Black** is behind). - **Octagon** and **Stripes** are **not opposites** (opposite of **Stripes** is **Two Dots**), so they can be adjacent. - **This configuration aligns with the net: Octagon** (front), **Stripes** (right), **Black** (back), **Two Dots** (bottom), **Empty** (left), **Frame** (top). All visible features match the expected spatial arrangement. ✓ **Valid** --- ### **Option D** - Front: Octagon - Right: Black - **Octagon** and **Black** are **opposites** (as per the net), so they **cannot** be adjacent. - This directly violates the cube folding rules. ✓ **Not valid** --- ### **Conclusion** Only **Option C** correctly represents a cube that can be folded from the given net without violating any spatial constraints or adjacency rules. **boxed{C}**

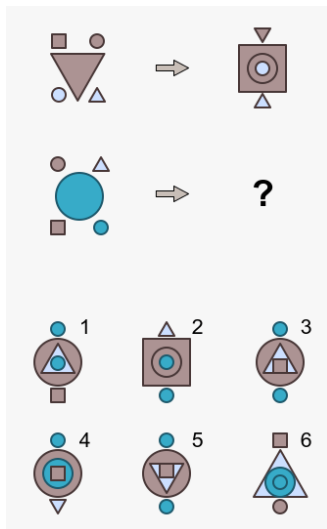
Error Reason: The model fails to correctly determine the relative spatial positions (top/right vs bottom) of adjacent faces after mentally folding the net. The gold answer is 'A', but the response applies the wrong spatial correspondence (rotation/mirroring/folding/view mapping). These tasks are reference-frame sensitive; a single wrong orientation assumption propagates and makes the final answer contradict the correct mapping.

Figure 14: Spatial Orientation Error case of Qwen3-VL-235B-A22B-Thinking.

Grounding Error

Question: This logic puzzle shows a square containing a triangle. After the arrow, the relationship between them changes due to a transformation. Which option is correct?

Grounding Truth: 3



Qwen3-VL-235B-A22B-Thinking: To solve this logic puzzle, we need to identify the transformation rule applied to the first example and then apply it consistently to the second example. --- ### **Step 1: Analyze the First Example** **Left Side (Before Transformation):** **Big Shape:** Brown triangle (pointing down) - **Top-left small shape:** Brown square - **Top-right small shape:** Brown circle - **Bottom-left small shape:** White circle - **Bottom-right small shape:** White triangle **Right Side (After Transformation):** **Big Shape:** Brown square - **Top small shape:** Brown inverted triangle (original big shape, now inverted) - **Center shape:** White circle (from bottom-left small shape) - **Bottom small shape:** White triangle (from bottom-right small shape) --- ### **Step 2: Identify the Transformation Rules** From the first example, we can deduce the following consistent rules: 1. **Big Shape \leftrightarrow Top-left Small Shape Swap:** - The **top-left small shape** becomes the **new big shape**. - The **original big shape** becomes the **top small shape**, **inverted** in orientation (e.g., a downward-pointing triangle becomes an upward-pointing one). 2. **Bottom-left Small Shape \rightarrow Center:** - The **bottom-left** small shape moves to the **center** of the new big shape. 3. **Bottom-right Small Shape \rightarrow Bottom:** - The **bottom-right small shape** moves to the **bottom** of the new big shape. 4. **Colors are preserved:** All shapes retain their original colors. --- ### **Step 3: Apply the Rules to the Second Example** **Left Side (Second Example):** **Big Shape:** Blue circle - **Top-left small shape:** Brown square - **Top-right small shape:** White triangle - **Bottom-left small shape:** Brown circle - **Bottom-right small shape:** Blue circle **Apply the Rules:** 1. **New Big Shape = Top-left small shape (brown square)** \rightarrow **New big shape is a brown square** 2. **Top small shape = Original big shape (blue circle), inverted** \rightarrow Since a circle has no direction, it remains a blue circle 3. **Center shape = Bottom-left small shape (brown circle)** \rightarrow **Center is a brown circle** 4. **Bottom small shape = Bottom-right small shape (blue circle)** \rightarrow **Bottom is a blue circle** --- ### **Step 4: Match with Answer Choices** Now, look for the option that matches: - **Big shape:** Brown square - **Top small shape:** Blue circle - **Center shape:** Brown circle - **Bottom small shape:** Blue circle Let's evaluate each option: - **Option 1:** Big brown circle \rightarrow **X** - **Option 2:** Big brown square , center blue circle **X** (should be brown), top white triangle **X**, bottom blue circle - **Option 3:** Big brown circle \rightarrow **X** - **Option 4:** Big brown circle \rightarrow **X** - **Option 5:** Big brown circle \rightarrow **X** - **Option 6:** Big white triangle \rightarrow **X** Only **Option 2** has the **correct big shape (brown square)** and **correct bottom small shape (blue circle)**. However, the **center shape** should be **brown circle**, but Option 2 shows a **blue circle**. This discrepancy suggests that the **top small shape** is also incorrect (white triangle instead of blue circle). But upon closer inspection, the **only option with a brown square as the big shape** is **Option 2**, and all other options fail to match even one key element (e.g., big shape not square, etc.). Therefore, **Option 2 is the best match** under the transformation rules derived from the first example. --- ### **Final Answer: \boxed{2}**

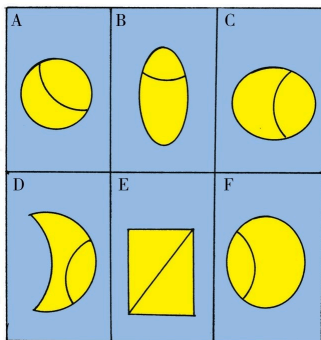
Error Reason: The model undercounts by missing one specific triangle formed by the combination of the two small top triangles. The gold answer is '6', but the response's conclusion relies on an incorrect enumeration (missing, double-counting, or inconsistent counting rules). This is not just arithmetic: the failure occurs at the visual counting stage, so the derived total/choice is wrong even if later calculations are internally consistent.

Figure 15: Grounding Error case of Qwen3-VL-235B-A22B-Thinking.

Wrong rule

Question: Carefully observe the rule of change, and infer which option should be chosen.

Grounding Truth: E



Qwen3-VL-235B-A22B-Thinking: To determine the correct option, we need to identify a consistent **rule of change** across the six options. The image shows a 2×3 grid with labeled shapes (A–F), each consisting of a yellow shape divided by a black line. --- ### Key Observations: 1. **Shape Type and Symmetry**: - All shapes except **E** have **curved outlines** and a **curved dividing line**. - **E** is the **only one** with a **straight-sided square** and a **diagonal straight line**. 2. **Dividing Line Orientation**: - In all other shapes (A, B, C, D, F), the dividing line is **curved**, often forming a **lens-like or crescent-like shape**. - This suggests a **consistent visual theme**: **curved outlines and curved internal lines**. 3. **Column-wise Pattern**: - **Column 3** contains: - **C**: A **horizontal oval** with a **curved dividing line** - **F**: A **circle** with a **curved dividing line** - These two are **visually related**—a horizontal oval can be seen as a stretched version of a circle, and both have curved internal lines. 4. **Row-wise Pattern**: - **Top row (A, B, C)**: All feature **curved outer shapes** with **curved internal lines**. - **Bottom row (D, E, F)**: D and F follow the same pattern; **E breaks it** with a **square and diagonal**. --- ### Rule of Change: The most consistent and observable **rule** is that **all shapes (except E) have a curved outer boundary and a curved internal dividing line**. This implies that **E is the odd one out**, and **the rest conform to a specific pattern**. However, the question asks: > \("Carefully observe the rule of change, and infer which option should be chosen.\)" This phrasing suggests **selecting the one that best fits the pattern**, not identifying the outlier. Given this, **E is not the correct choice** because it **does not fit the pattern**. Instead, we must look for the **most consistent continuation** of the pattern. --- ### Best Fit: - **F** is a **circle** with a **curved dividing line**, similar to **A**, but mirrored. - **C** is a **horizontal oval** with a **curved dividing line**, and **F** is its **circular counterpart**. - This supports a **column-based transformation** where **C → F** represents a return to a more symmetrical form (from an oval to a circle), maintaining the curved internal line. Thus, **F is the only option** that **completes the pattern** when viewed as a continuation of the top row's logic. --- ### Final Answer: \boxed{F}

Error Reason: The model correctly identified the outlier (E) but erroneously reasoned that the task required choosing an option that conforms to the majority pattern. The gold answer is 'E', but the response adopts an incorrect solving rule/assumption early on. Because the core rule is wrong, later steps may look coherent yet they solve a different problem than the one asked.

Figure 16: Wrong Rule case of Qwen3-VL-235B-A22B-Thinking.

Missed steps



Question: How did Slylock Fox get the rope up to the window so they could escape?

Grounding Truth:

He had the fisherman cast his line up to the window, tied the rope to the line, and pulled the rope up.

Qwen3-VL-235B-A22B-Thinking: Slylock Fox `[[used a fishing rod to cast the rope up to the window]].` `\boxed{He used a fishing rod to cast the rope up.}`

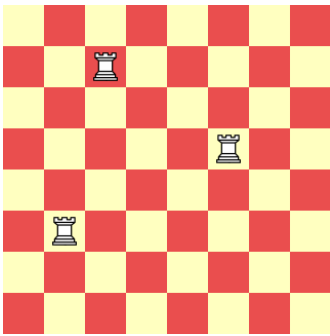
Error Reason: " The model points to the right tool (a fishing rod/line) but does not explain the essential intermediate steps required by the scene. In the comic, the rope cannot be directly 'cast' to the window; the workable procedure is to cast the fishing line upward, attach/tie the rope to the line, and then pull the rope up to the window so they can climb/escape. The response stops at Slylock Fox `[[used a fishing rod to cast the rope up to the window]].` `\boxed{He used a fishing rod to cast the rope up.}` and never states the necessary tie-and-pull step, so it omits a required condition/action that makes the escape feasible. As a result, the explanation is incomplete and does not match the ground-truth mechanism.

Figure 17: Missed Steps case of Qwen3-VL-235B-A22B-Thinking.

Calculation error

Question: How many squares of any size on this chessboard do not contain a rook?

Grounding Truth: 116



Qwen3-VL-235B-A22B-Thinking: To determine how many squares of any size on the chessboard do not contain a rook, we begin by calculating the total number of squares of all sizes on an 8x8 board, and then subtract the number of squares that contain at least one rook. --- ### ✓

Total Number of Squares on an 8x8 Chessboard The total number of squares of size $k \times k$ is: $(9 - k)^2$. Summing over all $k = 1$ to 8 : $\sum_{k=1}^8 (9 - k)^2 = 8^2 + 7^2 + \dots + 1^2 = 204$. --- ### ✓

Rook Positions From the image (and standard conventions), the rooks are placed at the following positions (in terms of column, row, where both are from 1 to 8): - Rook A: (3, 7) - Rook B: (6, 5) - Rook C: (2, 3) We compute for each square size $k \times k$, the number of $k \times k$ squares that contain at least one rook, using inclusion-exclusion to avoid double-counting overlapping squares. --- ### ✓

Count of Squares Containing At Least One Rook After computing for each k from 1 to 8, and applying inclusion-exclusion where necessary, we find the total number of squares that contain at least one rook is: $\text{Total rook-containing squares} = 3 + 12 + 21 + 22 + 16 + 8 + 3 + 1 = 86$. --- ### ✓

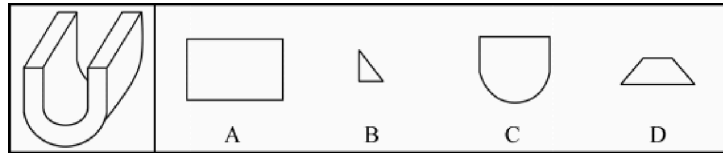
Final Calculation $\text{Squares without rooks} = \text{Total squares} - \text{Rook-containing squares} = 204 - 86 = 118$. --- ###

Final Answer **118**

Error Reason: The model correctly identifies rook positions but makes arithmetic errors in deriving the count of squares for sizes 6 and 7. The gold answer is '116 squares.'. The response broadly sets up the right quantities/logic, but an arithmetic/algebraic manipulation step derails the derivation. That small numerical error propagates to the final choice even if the intended rule is mostly correct.

Figure 18: Calculation Error case of Qwen3-VL-235B-A22B-Thinking.

Format Error



Question: The left figure is the given solid; when cut by a section from any face, which of the following may be its section: A. A, B. B, C. C, D. D

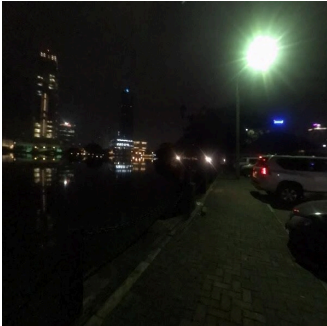
Grounding Truth: A

Qwen3-VL-235B-A22B-Thinking: To determine which of the given shapes (A, B, C, D) may be a cross-section of the solid when cut by a plane from any face, we must analyze the geometry of the solid and how it interacts with various cutting planes. --- ### **Understanding the Solid** The left figure depicts a **U-shaped solid** with: - A **flat rectangular top surface** (between the two vertical arms), - Two **vertical inner faces** (forming the sides of the U), - A **curved bottom** (semicircular in shape), and - An **outer curved surface** (on the right side). This solid can be visualized as a **prism** with a **U-shaped base**, where the base is a U with a straight top and a semicircular bottom. The solid extends uniformly along its length (into the page). --- ### **Analyzing Each Option** ##### **Option A: Rectangle** - If a **horizontal plane** cuts through the solid **parallel to the top surface**, it will intersect both vertical arms at the same height. - This results in a **rectangle**, matching the shape of the top surface. - **Possible** ##### **Option B: Triangle** - A triangle requires **three straight edges**, each from a different face. - However, the solid has **curved surfaces** (bottom and outer face), so any section that intersects these would have **curved edges**, not straight. - Even if the cut avoids the curved parts, the solid's geometry does **not allow for three straight-edge intersections** without involving curved faces. - **Not possible** ##### **Option C: U-shape (Rectangle on top, semicircular bottom)** - This matches the **front or back face** of the solid. - A **plane perpendicular to the length** of the solid (i.e., cutting across the U) will produce this exact shape. - **Possible** ##### **Option D: Trapezoid** - A **slanted vertical plane** that intersects the **two vertical inner faces** (which are parallel) and the **top surface** (a rectangle) will create a trapezoidal section. - The two vertical arms provide **parallel sides**, while the top surface provides **non-parallel sides** due to the angle of the cut. - **Possible** --- ### **Conclusion** **Only **A**, **C**, and **D** can be valid sections of the solid when cut by an appropriate plane.** **B** is impossible because the solid contains curved surfaces that would prevent forming a triangular section with only straight edges. --- ### **Final Answer:** **\boxed{A, C, D}**

Error Reason: This is a single-choice question that requires EXACTLY ONE option letter (A/B/C/D) as the final answer. The model provides a detailed attempt, but its final answer is formatted as multiple candidates.

Figure 19: Format Error case of Qwen3-VL-235B-A22B-Thinking.

Hallucination



Question: "In which country and within which first-level administrative region of that country was this picture taken? Please answer in the format of country, administrative_area_level?"

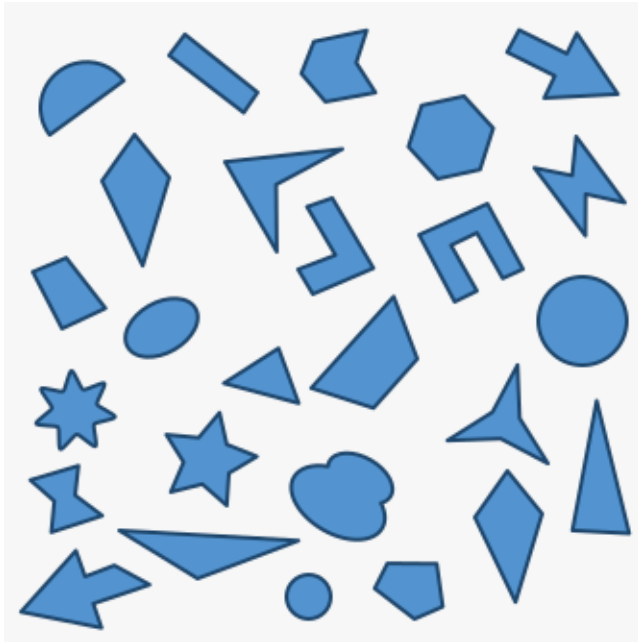
Grounding Truth:
sri lanka,western province

Qwen3-VL-235B-A22B-Thinking: The image depicts a nighttime urban waterfront scene with a promenade, parked cars, and illuminated high-rise buildings. Key clues include the **distinctive red-lit tower** (likely the TD Canada Trust Tower) and the waterfront layout, which align with **Toronto's Harbourfront** area along Lake Ontario. Toronto is located in the province of Ontario, Canada. **\boxed{Canada, Ontario}**

Error Reason: The model hallucinates a 'distinctive red-lit tower' which is not present in the image. The gold answer is 'sri lanka,western province'. The response commits to concrete intermediate facts or visual claims that are not justified by the provided evidence, and then reasons from those unsupported details. This is hallucination in the strict sense: the model fabricates or overconfidently asserts evidence and uses it to drive the answer.

Figure 20: Hallucination case of Qwen3-VL-235B-A22B-Thinking.

Localized Reasoning: Duplicate Localized



Q This puzzle has many different geometric shapes. Only two of them are identical. Can you find them? Return exactly 2 bounding boxes, no more and no fewer.

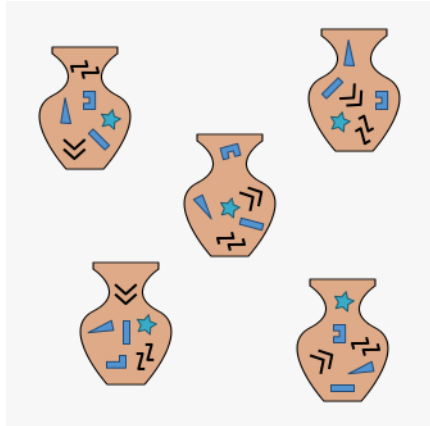
A [[43, 67, 96, 141], [237, 242, 290, 316]]

Table 7: Example of Duplicate Localized.

H Qualitative Subcategory Examples

We provide representative and canonical question–answer examples covering all 10 major categories and 36 subcategories to illustrate the intended task semantics and annotation standards.

Localized Reasoning: Odd Localization

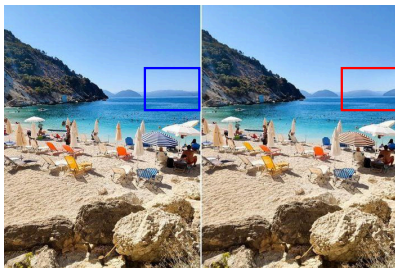


Q This logic puzzle has five vases with a decorative pattern. Which vase does not match the others? Return exactly 1 bounding boxes, no more and no fewer.

A [[52, 197, 136, 308]]

Table 8: Example of Odd Localization.

Spot the Difference: Outline Tweak

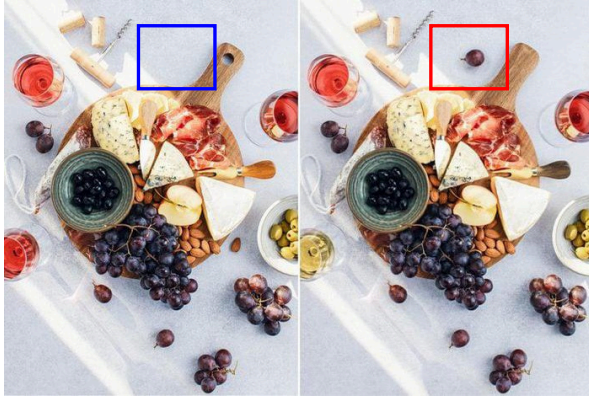


Q The two images constitute a spot-the-difference pair. Using the first image as reference, return the bounding boxes on the first image for each object that differs between the two images. Bounding boxes must be placed only on the first image: all bounding box coordinates must lie within the first image, and no bounding boxes should be output for the second image. The total number of differing objects is 5; exactly 5 bounding boxes must be returned.

A [[69, 235, 110, 273], [96, 200, 121, 223], [189, 202, 222, 250], [239, 210, 313, 270], [312, 140, 348, 175]]

Table 9: Example of Outline Tweak.

Spot the Difference: Object Addition and Removal



Q The two images constitute a spot-the-difference pair. Using the first image as reference, return the bounding boxes on the first image for each object that differs between the two images. Bounding boxes must be placed only on the first image: all bounding box coordinates must lie within the first image, and no bounding boxes should be output for the second image. The total number of differing objects is 5; exactly 5 bounding boxes must be returned.

A [[69, 235, 110, 273], [96, 200, 121, 223], [189, 202, 222, 250], [239, 210, 313, 270], [312, 140, 348, 175]]

Table 10: Example of Object Addition and Removal.

Spot the Difference: Color Replacement

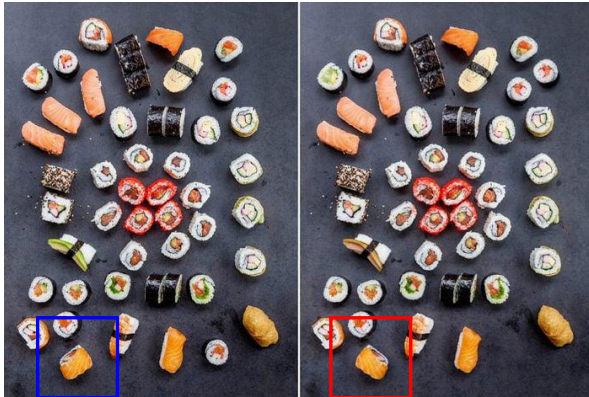


Q The two images constitute a spot-the-difference pair. Using the first image as reference, return the bounding boxes on the first image for each object that differs between the two images. Bounding boxes must be placed only on the first image: all bounding box coordinates must lie within the first image, and no bounding boxes should be output for the second image. The total number of differing objects is 6; exactly 6 bounding boxes must be returned.

A [[100, 0, 207, 39], [183, 48, 223, 128], [15, 185, 41, 219], [93, 251, 147, 342], [0, 303, 60, 377], [247, 118, 304, 181]]

Table 11: Example of Color Replacement.

Spot the Difference: Position Change

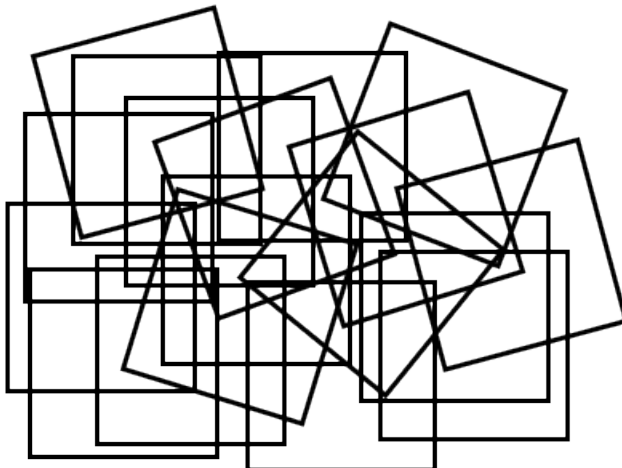


Q The two images constitute a spot-the-difference pair. Using the first image as reference, return the bounding boxes on the first image for each object that differs between the two images. Bounding boxes must be placed only on the first image: all bounding box coordinates must lie within the first image, and no bounding boxes should be output for the second image. The total number of differing objects is 5; exactly 5 bounding boxes must be returned.

A [[15, 62, 60, 111], [50, 266, 111, 323], [60, 403, 110, 453], [271, 67, 309, 104], [232, 391, 274, 435]]

Table 12: Example of Position Change.

Pattern Counting: Shape Counting



Q How many squares are in the drawing below?

A 18

Table 13: Example of Shape Counting.

Pattern Counting: Char-Grid Counting

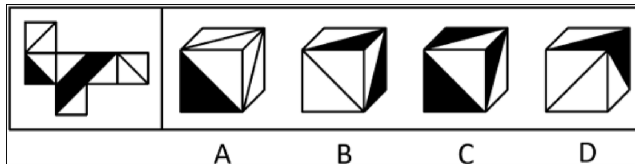
O T C O O C T O B E R O C R C
 C O T C O R E B O T C O O E O
 T O C T O B E R R C O C C B C
 O C T O B E R C R E T T T O T
 B R E B O T C O E O B O O T O
 E O C E O C T O B E R O C C B
 R C O R C C C E O T R O T O E
 R R R C O T R C T C E O O C R
 E E R E T C T T C O B C C R O
 B B R E B O T C O E O T T E C
 O O C O B O B O R O T O O B T
 T T T E C O T E B C C B B O O
 C C R O C T T C R E O E E T B
 O O R E B O T C O C R R R C E
 O C T O C O C T O B E R C O R

Q Can you find every occurrence of the word OCTOBER that appears in this grid (horizontally, vertically, or diagonally) ? The reasoning will reveal the number of times it occurs, but where are they?

A 29

Table 14: Example of Char-Grid Counting.

3D-Spatial Reasoning: Polyhedron Net Folding

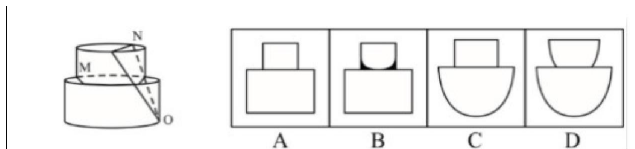


Q As shown is the planar net of a cube, the cube is most likely: A B C D ?

A D

Table 15: Example of Polyhedron Net Folding.

3D-Spatial Reasoning: Solid Cross-Section

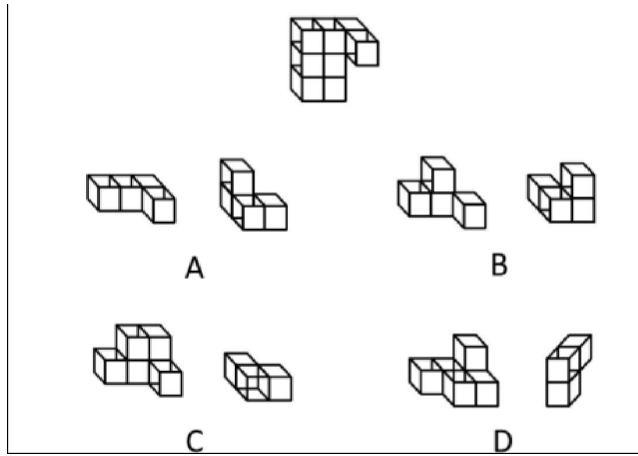


Q The cube below is obliquely cut along plane OMN; the section seen from the cutting plane may be: A B C D ?

A D

Table 16: Example of Solid Cross-Section.

3D-Spatial Reasoning: Solid Assembly

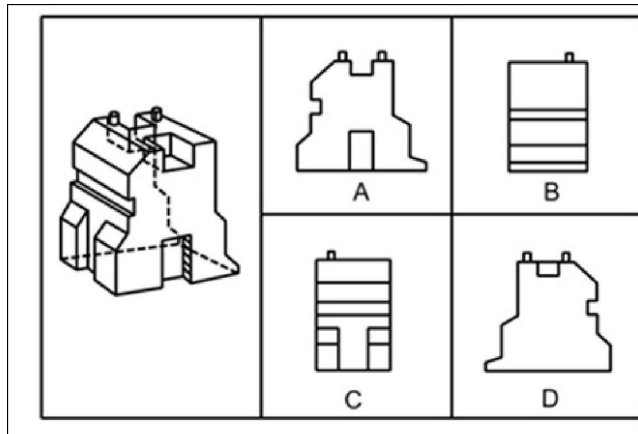


Q From the four given options, choose the most appropriate one to assemble the figure in the stem: A B C D ?

A A

Table 17: Example of Solid Assembly.

3D-Spatial Reasoning: Three-view Drawing



Q Among the four options on the right, which is not an elevation of the part on the left: A B C D ?

A A

Table 18: Example of Three-view Drawing.

Geolocation: Natural Landscapes



Q In which country and within which first-level administrative region of that country was this picture taken? Please answer in the format of country, administrative area level?

A india,tamil nadu

Table 19: Example of Natural Landscapes.

Geolocation: Architecture and Street Scenes



Q In which country and within which first-level administrative region of that country was this picture taken? Please answer in the format of country, administrative area level?

A france, nouvelle-aquitaine

Table 20: Example of Architecture and Street Scenes.

Geolocation: Indoor Scenes

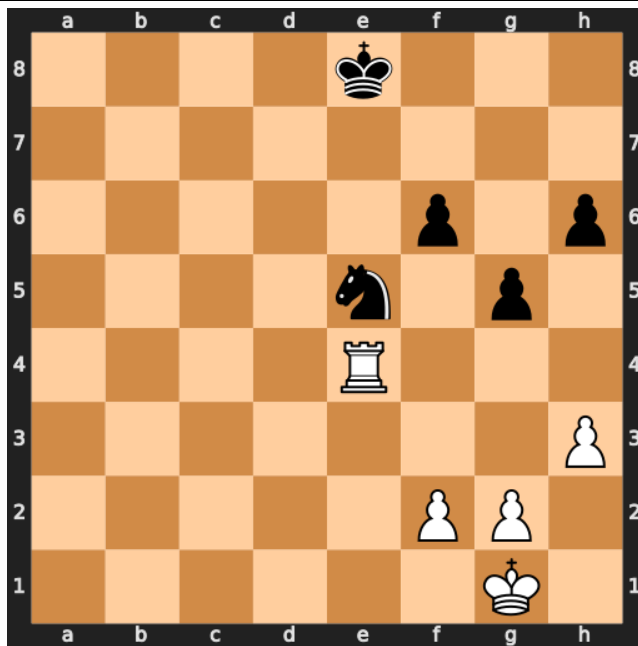


Q In which country and within which first-level administrative region of that country was this picture taken? Please answer in the format of country, administrative area level ?"

A argentina, buenos aires

Table 21: Example of Indoor Scenes.

Board Reasoning: Chess Position Evaluation

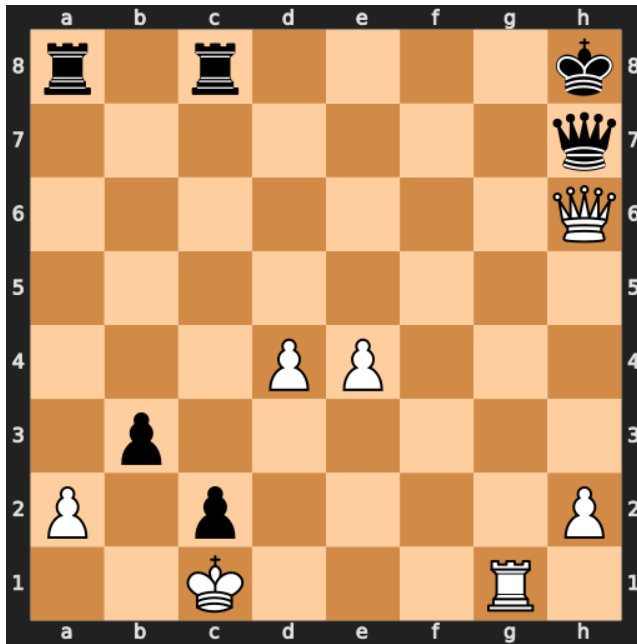


Q You are analyzing a chess position in FEN: 4k3/8/5p1p/4n1p1/4R3/7P/5PP1/6K1 w - -.the Stockfish evaluation in centipawns (from White's perspective). Think deeper about this position: Don't just evaluate the current board state. Consider what the most likely moves are for both sides and how the centipawn evaluation would change as the position develops. Analyze a moves ahead - what does the future of this position look like? How would a strong engine assess this position after calculating many moves deep? Analyze step by step and explain your reasoning.

A 400

Table 22: Example of Chess Position Evaluation.

Board Reasoning: Chess Solution Sequence

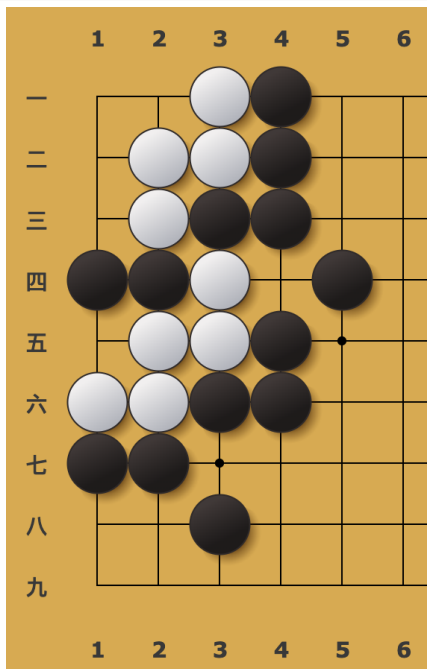


Q You are given a chess position in FEN: r1r4k/7q/7Q/8/3PP3/1p6/P1p4P/2K3R1 w - - 2 41. Find the best move for the side to play.

A h6f6

Table 23: Example of Chess Solution Sequence.

Board Reasoning: Go Life-and-Death



Q Go life & death. Black to play. What's the single best move? Please answer with the coordinate.

A 2

Table 24: Example of Go Life-and-Death.

Sudoku Solving: Standard 9 x 9 Sudoku

7			6					8
			9			5		
		6		4			3	
2	3				8			
		4				6		
			7				9	1
	2			3		4		
		1			9			
8					1			5

Q What is the digit at row 7, column 4 ?

A 8

Table 25: Example of Standard 9 x 9 Sudoku.

Cue Insight: Detective



Q Two gunshots were fired through the window of a coffee shop. When the police arrived, they successfully recognized which gunshot was fired first. Which was the first gunshot and how did they figure that out?

A The cracks of the left gunshot end up right at the cracks of the right gunshot. Therefore the first gunshot is the one on the right.

Table 26: Example of Detective.

Cue Insight: Image Cue Extraction

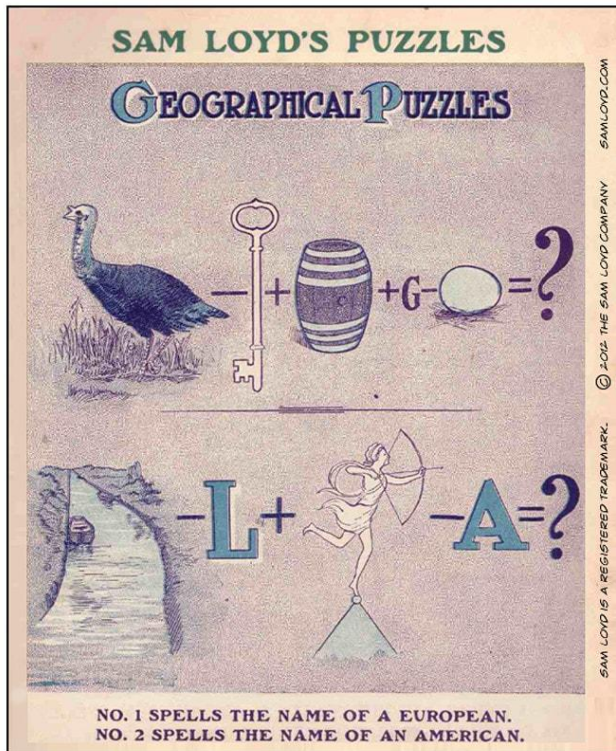


Q James sent this beautiful photo to a contest, hoping he will win first prize. The contest organizers, however, realized that the photo was fake and disqualified James. How did they figure out the picture was not real?

A The vocalist's name is Tessa. The bassist's name is Delilah. The drummer's name is Lisa.

Table 27: Example of Image Cue Extraction.

Cue Insight: English Word Addition and Subtraction



Q What nationalities do these sums spell?

A CANADIAN

Table 28: Example of English Word Addition and Subtraction.

Cue Insight: Folding-Perspective Insight

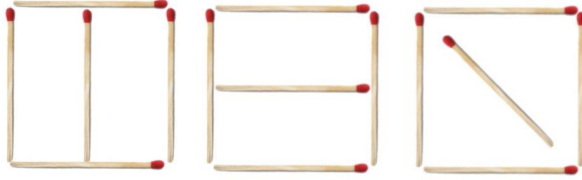


Q A boy receives a small note from a girl during his math class. He unfolds the note and sees the following expression. What does it mean?

A If you fold the note the other way around, the message "I LOVE YOU" appears.

Table 29: Example of Folding-Perspective Insight.

Cue Insight: Hidden-Word Decoding



Q You have fifteen matchsticks. If you must remove exactly six of them, what word should the remaining matchsticks spell?

A TEN

Table 30: Example of Hidden-Word Decoding.

Cue Insight: Color-Set to Word

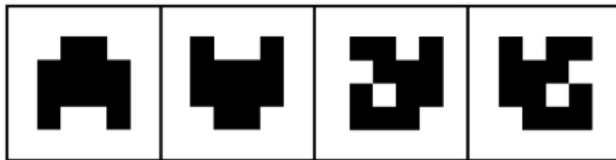
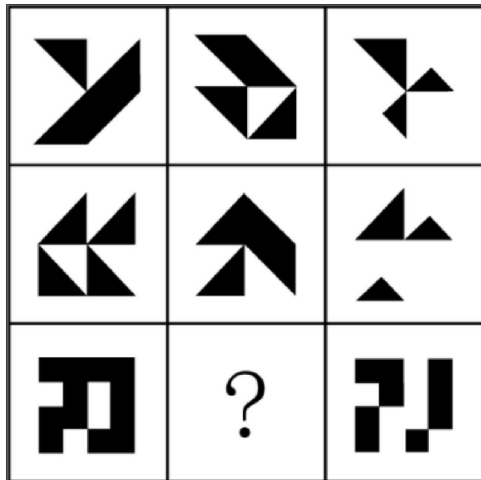


Q What does this Venn diagram depict?

A Each of the numbers corresponds to a letter from the alphabet: 1 – A, 4 – D, 5 – E, 7 – G, 14 – N, 15 – O, 18 – R. The three colors in the Venn diagram are GREEN, RED, ORANGE, and what is depicted is the letters they share.

Table 31: Example of Color-Set to Word.

Inductive Reasoning: Rule Learning and Application



A B C D

Q From the four options, choose the most appropriate one to fill in the question mark so that a pattern is presented: A B C D ?

A C

Table 32: Example of Rule Learning and Application.

Deductive Reasoning: Matrix Reasoning

3	3	39	6	17	4	22
27	4	80	5	49	7	61
2	42	5	93	2	54	3
98	33	6	21	4	1	28
7	19	65	8	48	31	6
56	5	2	77	?	23	12
43	8	37	16	51	30	1

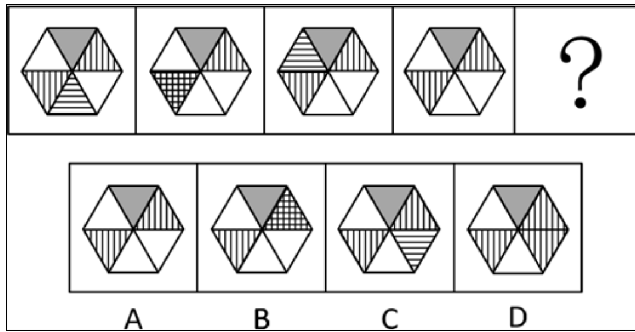
<input type="checkbox"/> 6	1	<input type="checkbox"/> 4	2	<input type="checkbox"/> 1	3
<input type="checkbox"/> 2	4	<input type="checkbox"/> 5	5	<input type="checkbox"/> 3	6

Q This difficult logic puzzle has 49 square cells with numbers, dark and light. The light numbers depend in some way on the dark numbers. You need to understand this relationship and choose the correct option for the light number in place of the question mark.

A 2

Table 33: Example of Matrix Reasoning.

Deductive Reasoning: Shape-Sequence Deductive Reasoning



Q From the four options, choose the most appropriate one to fill in the question mark so that a pattern is presented: A B C D ?

A B

Table 34: Example of Shape-Sequence Deductive Reasoning.