Diagnosing Visual Reasoning: Challenges, Insights, and a Path Forward

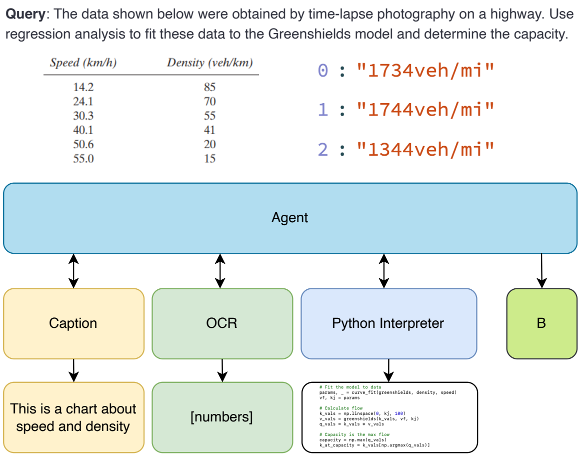
Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) that integrate visual and textual reasoning leverage chain-of-thought (CoT) prompting to tackle complex visual tasks, yet continue to exhibit visual hallucinations and an over-reliance on textual priors. We present a systematic diagnosis of state-of-the-art vision-language models using a three-stage evaluation framework, uncovering key failure modes. To address these, we propose an agent-based architecture that combines LLM reasoning with lightweight visual modules, enabling fine-grained analysis and iterative refinement of reasoning chains. Our results highlight future visual reasoning models should focus on integrating a broader set of specialized tools for analyzing visual content. Our system achieves significant gains (+10.3 on MMMU, +6.0 on MathVista over a 7B baseline), matching or surpassing much larger models. We will release our framework and evaluation suite to facilitate future research.

1 Introduction

The ability to perform coherent, structured reasoning is essential for solving complex visual understanding tasks. Unlike recognition, visual reasoning requires models to integrate perceptual cues with contextual knowledge, infer relationships between entities, track logical dependencies, and arrive at conclusions that are not immediately evident from raw pixel data. This cognitive process mirrors human problem-solving, where one sequentially interprets visual inputs and iteratively verifies conclusions (Liu et al., 2025b; Zhang et al., 2025; Yang et al., 2025; Fu et al., 2025).



Query: The data shown below were obtained by time-lapse photography on a highway. Use regression analysis to fit these data to the Greenshields model and determine the capacity.

Speed (km/h)	*Density (veh/km)*
14.2	85
24.1	70
30.3	55
40.1	41
50.6	20
55.0	15

```
0 : "1734veh/mi"
1 : "1744veh/mi"
2 : "1344veh/mi"
```

```
# Fit the model to data
params, _ = curve_fit(greenshields, density, speed)
vf, kj = params

# Calculate flow
k_vals = np.linspace(0, kj, 100)
v_vals = greenshields(k_vals, vf, kj)
q_vals = k_vals * v_vals

# Capacity is the max flow
capacity = np.max(q_vals)
k_at_capacity = k_vals[np.argmax(q_vals)]
```

Figure 1: This image showcases our agent system that leverages a pure LLM to solve a visual reasoning problem using external tools. It illustrates how complex tasks, such as fitting traffic speed-density data to the Greenshields model, can offload substantial token usage to a code interpreter, highlighting an efficient division of labor between perception, reasoning and computation.

Table 1: Comparison of our diagnostic agent with prior modular systems: MM-ReAct (Yang et al., 2023), MC-tree (Yao et al., 2024).

System	Math OCR	Iterative Diagnosis	Lightweight Backbone	Python Interpreter	Backtracing Thought
MM-ReAct	✓	✗	✓	✗	✗
MC-tree	✗	✓	✓	✗	✓
Ours	✓	✓	✓	✓	✓

Recent advancements in LLMs have accelerated progress in this direction with strong linguistic reasoning abilities. When extended into the multimodal domain, these capabilities enable models to interpret images, diagrams, and documents extending beyond recognition to include inference and abstraction. The emergence of Reasoning Multimodal LLMs (MLLMs), such as LLaVA-CoT (Xu et al., 2024), LlamaV-o1 (Thawakar et al., 2025), and Heima (hei, 2025), reflects this trend and demonstrates how the fusion of vision and language models can unlock new frontiers in visual intelligence. Central to these efforts is to encourage models to produce explicit intermediate steps. This structured reasoning is particularly impactful for visual tasks, where raw perceptual data must be transformed into high-level concepts through a series of inferential stages. For example, LlamaV-o1 combines Chain-of-thought (CoT) reasoning with curriculum learning and beam search to effectively solve multi-step visual tasks, while Heima accelerates inference by

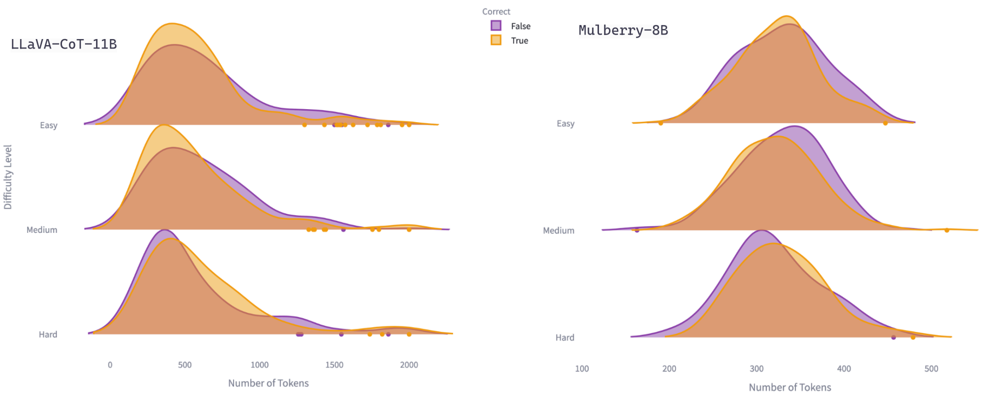


Figure 2: Comparison of token utilization and accuracy between Mulberry-7B and LLaVA-CoT-11B. Both models exhibit less adaptive reasoning, with token usage. We observe that LLaVA-CoT-11B frequently generates more tokens than Mulberry-7B, while Mulberry-7B tends to centralize its token usage around 300–350 tokens.

encoding CoT into compact representations (Shen et al., 2025)

Despite these advances, models still hallucinate, producing responses not grounded in the image, and often rely too heavily on textual priors. Most models reason in a single, unidirectional pass, lacking correction or self-reflection. To address these issues, we introduce a three-stage diagnostic framework and an agent-based architecture that tightly integrates stepwise textual reasoning with lightweight visual modules, enabling fine-grained analysis of reasoning failures (Author, s; Song et al., 2025; Kumar et al., 2025; Lu et al., 2025; Liu et al., 2025a). Unlike prior work (e.g., MC-tree (Yao et al., 2024), MM-ReAct (Yang et al., 2023)), our agent-based framework routes tool calls at each step for explicit intervention and diagnosis, similar in spirit to recent approaches like MMCTAgent (Kumar et al., 2024) and AgentRE (Shi et al., 2024). Our main contributions are:

Diagnosis: We present a diagnostic framework for math-centric visual reasoning, enabling granular identification and analysis of failures.

Agent-based Architecture: We propose an agent-based approach that decouples perception and reasoning, integrating LLMs with visual modules for iterative reasoning, yielding substantial empirical gains over strong 7B baselines.

Evaluation: We provide a comprehensive analysis of reasoning chains and release our diagnostic framework and evaluation suite to support future research in visual reasoning, enabling deeper understanding of model behaviors.

2 Diagnostic Methodology

Our analysis focuses on three representative visual reasoning models: QVQ (72B) (Team, 2024), Mulberry-7B (Yao et al., 2024)(Mulberry), and LLaVA-CoT-11B (Xu et al., 2024)(LLaVA-CoT). These models span a range of parameter sizes and are selected for their popularity and relevance. We exclude OpenAI's O-series reasoning models due to the unavailability of their reasoning paths, which prevents in-depth diagnostic analysis. This selection enables a comprehensive comparison across different model scales and reasoning strategies.

Dataset	Mulberry	LLaVA-CoT	QVQ
MMMU	52.8%	55.2%	60.9%
MathVista	63.1%	57.8%	65.4%
Base Model	QwenVL2-7B	Llama-3.2-7B	QwenVL2-72B

Table 2: Comparison of model accuracies (%) on MMMU and MathVista, and their base models.

2.1 Comparative Model Analysis

We begin our evaluation with the MMMU dataset (Yue et al., 2023) and MathVista (Lu et al., 2024), both selected for their comprehensive problem difficulty annotations and widespread use as benchmarks for MLLMs. For the MMMU dataset, we analyze token usage across varying difficulty levels to uncover patterns in reasoning efficiency and inefficiency, as shown in Figures 2,3.

As shown in Table 2, Mulberry is highly succinct, with token counts clustered between 200–400, but this brevity limits its accuracy (52.8%/63.1%). Incorrect Mulberry responses often occur at the upper end of its token range, suggesting that rigid, template-driven reasoning can be counterproductive when the model stretches beyond its typical

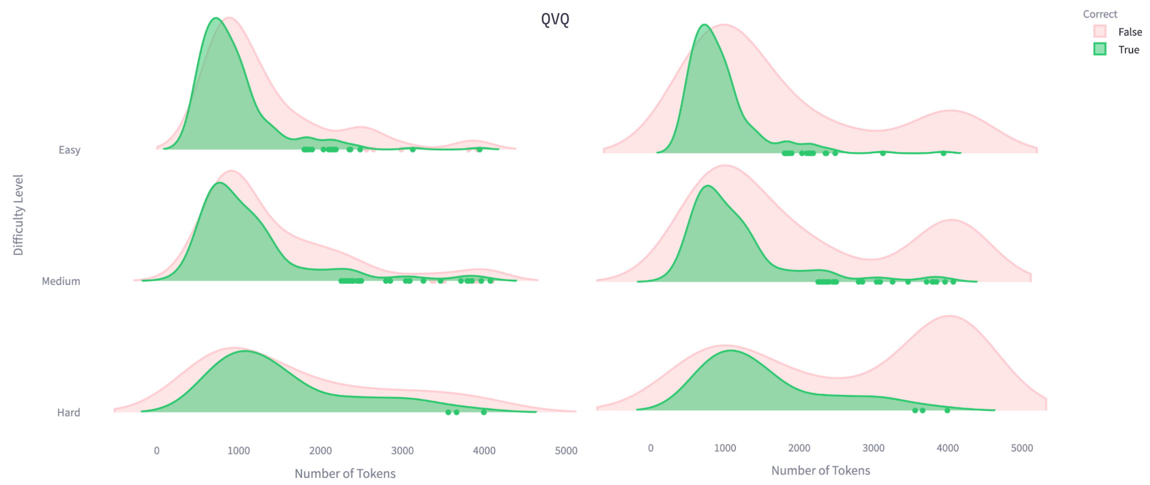


Figure 3: Token usage and accuracy trends for QVQ on MMMU. Left: Accuracy as a function of token count, showing diminishing returns and a decline beyond 2,000 tokens. Right: Distribution of token counts for correct and incorrect answers, including unfinished answers

patterns. LLaVA-CoT, with token counts typically between 800–1,200 for correct Easy/Medium answers, achieves intermediate accuracy. On harder tasks, longer responses often correspond to incorrect answers, suggesting that concise yet sufficiently detailed reasoning chains tend to be optimal, whereas excessive verbosity may signal confusion. Our analysis shows that while more verbose reasoning can indicate higher capability, excessive token usage (beyond 2,000 for QVQ) yields diminishing returns. The best models balance detail and brevity, providing enough reasoning steps without unnecessary verbosity. QVQ's larger size and flexible reasoning achieve the highest accuracy, but future work should aim to reduce verbosity while preserving reasoning quality.

2.2 In-depth Examination of QVQ

As shown in Table 2, QVQ achieves the highest accuracy—60.9% on MMMU and 65.4% on MathVista—outperforming both LLaVA-CoT and Mulberry. To better understand QVQ's strengths, we analyze its reasoning behavior in detail. Figure 4 illustrates the relationship between token count and accuracy, revealing that accuracy declines as token usage increases. This trend is partly due to our imposed hard threshold of 4,000 tokens, responses exceeding this limit are typically incomplete and considered incorrect. In Figure 3, the right panel displays the distribution of token counts for answers, while the left panel excludes unfinished responses. QVQ's superior performance comes with more tokens: it often generates 1,000–2,000 tokens for Easy/Medium tasks and over 3,000 tokens for Hard cases (see Figure 3). Incorrect answers are typically even longer, suggesting that excessive reasoning does not guarantee correctness. Notably, QVQ exhibits adaptive reasoning: for more difficult questions, it generates longer and more detailed reasoning chains, reflecting increased effort to address complexity. However, our analysis indicates that once token usage exceeds approximately 2,000 tokens, further reasoning does not improve accuracy and may even reduce it. We set a 4,000-token cutoff to avoid excessive computation and latency, as well as to align with practical deployment constraints. Upon manually examining QVQ's incorrect cases, we find that the reasoning steps themselves are often logically sound. However, errors frequently arise during visual readout operations—when revisiting the image, they sometimes produces statements that do not align with the visual content. These failures are commonly due to mistakes in reading numbers, misidentifying

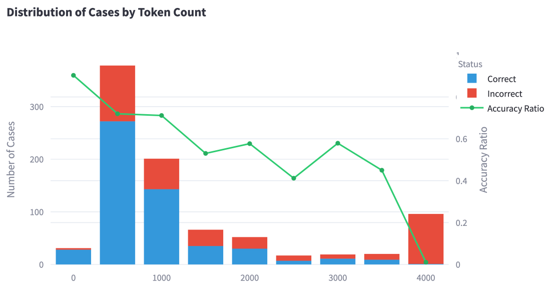


Figure 4: This chart illustrates the number of correct and incorrect cases across different token count ranges, with a green line as accuracy ratio. As token count increases, the number of cases generally decreases, and the accuracy ratio tends to decline.

Table 3: Quantitative results on the MMMU and MathVista datasets. All results are averaged over 3 random seeds; 95% confidence intervals are computed via bootstrap resampling.

Dataset	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Qwen2.5-VL-32B	Qwen2.5-VL-72B	Gemini-2 Flash	GPT-4o	Claude3.5 Sonnet	Qwen2-VL-72B	QVQ	Ours 3B	Ours 11B	Ours 7B
MMMU	53.1	58.6	70.0	70.2	70.7	70.3	70.4	64.5	60.9	60.2	66.7	68.9
MathVista	62.3	68.2	74.7	74.8	73.1	63.8	65.4	70.5	65.4	67.1	72.0	74.2

details, or other perceptual inaccuracies. Moreover, some hard problems require intensive computation, which consumes a large number of tokens and increases the likelihood of errors.

2.3 Agent-based Diagnostic and Intervention

To overcome the limitations of purely LLM-driven visual reasoning, we propose an agent-based architecture that seamlessly combines LLM reasoning with lightweight visual modules. This modular design enables precise analysis and iterative refinement of reasoning chains, allowing us to pinpoint whether failures stem from perceptual errors or reasoning weaknesses. We evaluate our approach across a suite of multimodal tasks, leveraging specialized tools such as *OCR*, *Image Captioning*, *Image Question and answering* and a *Python interpreter*, with `Qwen2.5-VL-3b,7B` (Bai et al., 2025) serving as the backbone. Our agent's reasoning mode is denoted as `qwq` (Yang et al., 2024). Based on the hypothesis that visual grounding errors are a major source of failure, we experiment with agent-based systems using three backbone sizes—3B, 7B, and 11B, with the latter matching LLaVA-CoT.

Our results yield several key insights: 1. *Strong Performance Without Large-Scale Models:* Our agent-based system ("Ours 7B") achieves 68.9% on MMMU and 74.2% on MathVista, rivaling top-tier models such as Qwen2.5-VL-72B, Gemini-2 Flash, and GPT-4o, despite using a much smaller backbone. This demonstrates that modularizing perception and reasoning can yield substantial gains without increasing model size. Notably, the 7B backbone consistently outperforms the larger 11B variant, highlighting the effectiveness of our modular approach. 2. *Dedicated Visual Tools Enhance Reasoning:* On MathVista, our system matches the performance of much larger models, underscoring that perceptual grounding (e.g., accurate text and layout extraction) is a key bottleneck. Specialized tools such as OCR are essential for these tasks. 3. *Task-Specific Gains:* On MMMU, our system outperforms the base Qwen2.5-VL-7B by 10.3 points; on MathVista, where perceptual accuracy is critical, the improvement is even greater (+6.0 points). This supports the view that many visual reasoning failures stem from perceptual errors, which modular pipelines can address. Unlike monolithic VL models, our agent architecture enables multi-step reasoning, such as re-querying OCR or cross-checking visual entities with logical constraints, providing greater flexibility and effectiveness without increasing model size. We performed analysis on 100 incorrect responses from both baseline and our models, categorizing errors as OCR, spatial, math. Baseline errors: OCR (38%), spatial (22%), math (19%). With our agent, these dropped to OCR (19%), spatial (15%), math (13%).

2.4 Ablation Study

Table 4: Ablation study of our agent-based system (7B backbone) on MMMU and MathVista. Each column disables a specific module.

Dataset	**Full**	**- OCR**	**- Python**	**- Caption**	**- QA**	**- Backtrace**
MMMU	68.9	62.1	65.4	66.2	67.0	60.8
MathVista	74.2	66.7	70.3	71.1	72.0	69.5

Table 4 summarizes the effect of removing each module. Disabling OCR causes the largest drop, especially on MathVista, confirming its critical role. The Python interpreter and captioning modules also yield notable gains, while the QA tool has a smaller effect. Removing backtracing significantly reduces performance, underscoring its importance for error correction for iterative reasoning.

3 Conclusion and Limitations

Our diagnostic framework demonstrates that targeting common failure modes enables strong performance even with smaller backbones. Looking forward, **future visual reasoning models should focus on integrating a broader set of specialized tools for analyzing visual content**. Beyond simply calling external tools, models should natively incorporate these capabilities. This direction will help models better adapt to diverse and complex real-world scenarios. Our analysis is limited to math-centric visual reasoning, and findings may not generalize to other domains such as document understanding or natural scene understanding.

References

2025. Efficient reasoning with hidden thinking. *Preprint*, arXiv:2501.19201.

Author(s). 2025. Rethinking reflection in pre-training. *Journal Name*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. 2025. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*.

Adarsh Kumar, Hwiyoon Kim, Jawahar Sai Nathani, and Neil Roy. 2025. Improving the reliability of llms: Combining cot, rag, self-consistency, and self-verification. *arXiv preprint arXiv:2505.09031*.

Somnath Kumar, Yash Gadhia, Tanuja Ganu, and Akshay Nambi. 2024. Mmctagent: Multi-modal critical thinking agent framework for complex visual reasoning. *arXiv preprint arXiv:2405.18358*.

Qiang Liu, Xinlong Chen, Yue Ding, Shizhen Xu, Shu Wu, and Liang Wang. 2025a. Attention-guided self-reflection for zero-shot hallucination detection in large language models. *arXiv preprint arXiv:2501.09997*.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. 2025b. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*.

Haolang Lu, Yilian Liu, Jingxin Xu, Guoshun Nan, Yuanlong Yu, Zhican Chen, and Kun Wang. 2025. Auditing meta-cognitive hallucinations in reasoning large language models. *arXiv preprint arXiv:2505.13143*.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xuan Shen, Yizhou Wang, Xiangxi Shi, Yanzhi Wang, Pu Zhao, and Jiuxiang Gu. 2025. Efficient reasoning with hidden thinking. *arXiv preprint arXiv:2501.19201*.

Yuchen Shi, Guochao Jiang, Tian Qiu, and Deqing Yang. 2024. Agentre: An agent-based framework for navigating complex information landscapes in relation extraction. *arXiv preprint arXiv:2409.01854*.

Xiaoshuai Song, Yanan Wu, Weixun Wang, Jiaheng Liu, Wenbo Su, and Bo Zheng. 2025. Progco: Program helps self-correction of large language models. *arXiv preprint arXiv:2501.01264*.

Qwen Team. 2024. Qvq: To see the world with wisdom. *Preprint*, arXiv:2409.12191.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *Preprint*, arXiv:2501.06186.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-cot: Let vision language models reason step-by-step. *Preprint*, arXiv:2411.10440.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Diyi Yang, Junnan Li, Xiangru Li, Wayne Zhao, Ming Tan, Jing Du, Zhou Yu, Kai-Wei Chang, Zichao Wu, and Mohit Bansal. 2023. Mm-react: Prompting multi-modal chain-of-thought reasoning in language-image models. *arXiv preprint arXiv:2303.11381*.

Shuo Yang, Siwen Luo, Soyeon Caren Han, and Eduard Hovy. 2025. Magic-vqa: Multimodal and grounded inference with commonsense knowledge for visual question answering. *arXiv preprint arXiv:2503.18491*.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and Dacheng Tao. 2024. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *Preprint*, arXiv:2412.18319.

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. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Yi Zhang, Qiang Zhang, Xiaozhu Ju, Zhaoyang Liu, Jilei Mao, Jingkai Sun, Jintao Wu, Shixiong Gao, Shihan Cai, Zhiyuan Qin, Linkai Liang, Jiaxu Wang, Yiqun Duan, Jiahang Cao, Renjing Xu, and Jian Tang. 2025. Embodiedvsr: Dynamic scene graph-guided chain-of-thought reasoning for visual spatial tasks. *arXiv preprint arXiv:2503.11089*.