EVA: Evidence-seeking Visual Agent for Hallucination-Resistant Multimodal Reasoning

Ting Lei **Yu Chen** **Yang Liu***
Wangxuan Institute of Computer Technology, Peking University
ting_lei@pku.edu.cn yu_chen@stu.pku.edu.cn yangliu@pku.edu.cn

Abstract

Hallucination remains a fundamental challenge in multimodal large language models (MLLMs), often stemming from unverified visual assumptions and reasoning processes that are weakly grounded in observable evidence. Existing approaches attempt to mitigate hallucination through self-reflection, multi-agent debate, or counterfactual verification. However, these methods remain largely opinion-driven: they lack explicit verification of the visual understanding process and may fail when multiple agents share the same incorrect perceptual assumptions, leading to consistent yet hallucinated conclusions. A promising direction is to augment MLLMs with external visual tools (e.g., OCR and grounding models) to acquire verifiable evidence. However, effective tool-augmented reasoning remains challenging due to two key failure modes: unreliable tool invocation, where models select inappropriate tools or produce incorrect parameters, and evidence–reasoning inconsistency, where models ignore or contradict the acquired evidence when forming final predictions. In this work, we propose EVA, an Evidence-seeking Visual Agent that reframes multimodal reasoning as an explicit process of evidence acquisition and verification. At the core of EVA is a unified evidence-grounded execution pipeline designed to address the above challenges through two tightly coupled mechanisms. First, we introduce *Iterative Tool Replanning*, which dynamically refines tool selection and parameters based on intermediate observations, improving the reliability of tool usage. Second, we develop *Evidence Consistency Verification*, an answer validation mechanism that improves alignment between the final prediction and the collected evidence, mitigating evidence–reasoning inconsistencies. In addition, to balance robustness and efficiency, EVA incorporates *Selective Evidence Routing*, an adaptive fast–slow reasoning strategy that determines when external evidence is necessary. Visually straightforward queries are handled via direct reasoning, while complex or high-risk cases trigger tool-based evidence acquisition. By jointly addressing how to reliably acquire evidence and how to use it consistently, EVA transforms multimodal reasoning from passive answer validation into explicit evidence-backed decision making. Experimental results demonstrate that EVA achieves significantly improved reasoning reliability over direct VLM inference and multi-agent methods, while maintaining efficiency through selective tool invocation. These findings highlight structured and adaptive evidence utilization as a practical paradigm for hallucination-resistant multimodal reasoning. Code is available at https://github.com/oceanflowlab/EVA.

1 Introduction

Multimodal large language models (MLLMs) have recently demonstrated strong capabilities in visual understanding and reasoning. However, hallucination remains a fundamental obstacle to their

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

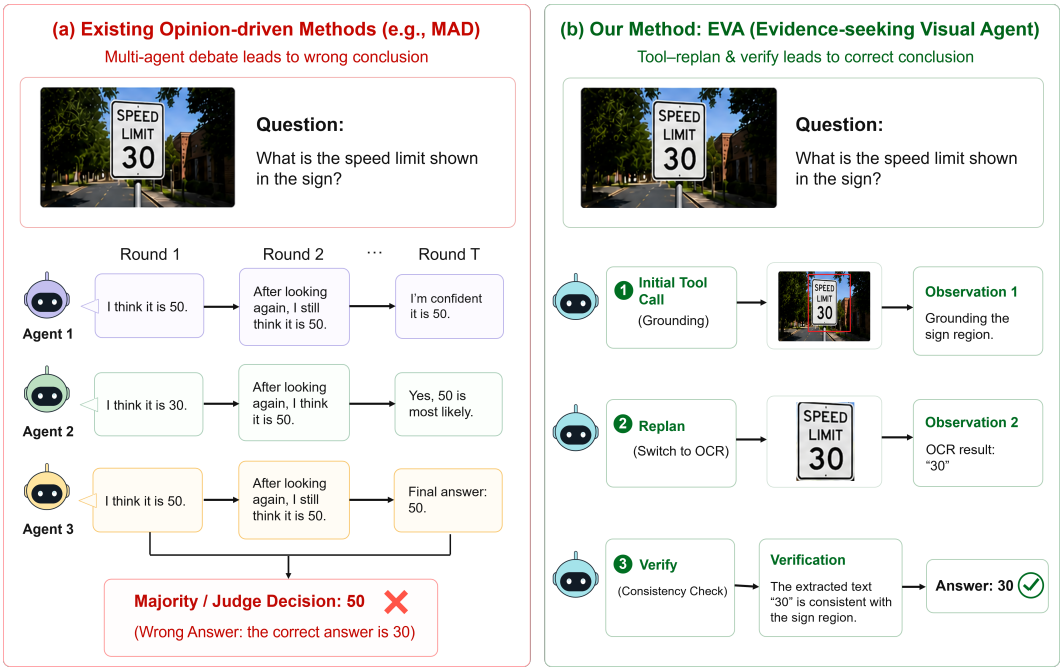


Figure 1: **Opinion-driven vs. evidence-driven multimodal reasoning.** (a) Existing methods rely on multi-agent debate or iterative reasoning, where agents may converge to an incorrect answer due to shared perceptual errors. (b) EVA actively seeks and verifies visual evidence through tool use and replanning, ensuring that the final answer is grounded in reliable observations.

reliability, especially in multimodal settings where reasoning depends on both visual perception and language generation. In many cases, models produce plausible but incorrect answers grounded in unsupported visual assumptions, or fail to verify whether their reasoning is consistent with observable evidence. This issue is particularly critical for tasks requiring fine-grained perception, such as reading text in images, identifying subtle attributes, or interpreting structured visual content.

A growing body of work attempts to mitigate hallucination by improving reasoning robustness. Representative approaches include self-reflection, multi-agent debate (MAD), and counterfactual verification. These methods aim to identify unreliable outputs through iterative reasoning, cross-agent interaction, or hypothetical testing. While promising, they remain largely *opinion-driven*: reasoning is conducted primarily in the language space, and correctness is often inferred from agreement, consistency, or confidence rather than explicit grounding in visual evidence. As illustrated in Figure 1(a), such approaches may lead to consistent yet incorrect conclusions, even when multiple agents participate in deliberation. In particular, when different reasoning paths share the same incorrect perceptual assumptions, agreement among agents does not guarantee correctness.

We argue that hallucination in multimodal reasoning is not merely a failure of reasoning, but more fundamentally a failure of *evidence acquisition and verification*. Reliable reasoning should be grounded in explicitly acquired, task-relevant visual evidence, rather than inferred from internal deliberation or inter-agent consensus.

A natural solution is to augment MLLMs with external visual tools (e.g., OCR and grounding models) to acquire verifiable evidence. However, effective tool-augmented reasoning introduces new challenges. First, models often suffer from *tool invocation unreliability*, where inappropriate tools are selected or incorrect parameters are produced, leading to noisy or irrelevant observations. Second, even when correct evidence is obtained, models may exhibit *evidence–reasoning inconsistency*, ignoring or contradicting the collected evidence when forming final predictions. These issues reveal that simply providing tools is insufficient; reliable multimodal reasoning requires both accurate evidence acquisition and consistent evidence utilization.

To address these challenges, we propose EVA, an Evidence-seeking Visual Agent that reframes multimodal reasoning as an explicit evidence-driven process. At the core of EVA is a unified evidence-grounded execution pipeline composed of two tightly coupled mechanisms. First, we

introduce *Iterative Tool Replanning*, which dynamically refines tool selection and parameters based on intermediate observations, improving the reliability of tool invocation. Second, we develop *Evidence Consistency Verification*, an answer validation mechanism that enforces alignment between the final prediction and the collected evidence, mitigating evidence–reasoning inconsistencies. As shown in Figure 1(b), this replan-and-verify process enables the model to correct earlier decisions and produce evidence-consistent answers.

In addition, to balance robustness and efficiency, EVA incorporates *Selective Evidence Routing*, an adaptive fast–slow reasoning strategy that determines when external evidence is necessary. Visually straightforward queries are handled via direct reasoning, while complex or high-risk cases trigger tool-based evidence acquisition (e.g., OCR, object grounding, and chart parsing). This design avoids unnecessary tool usage while preserving robustness on challenging inputs.

By jointly addressing how to reliably acquire evidence and how to use it consistently, EVA transforms multimodal reasoning from passive validation of generated answers into explicit evidence-backed decision making.

Our contributions are threefold: (1) We reformulate hallucination-resistant multimodal reasoning as an evidence acquisition and verification problem, highlighting the limitations of opinion-driven validation in multimodal settings. (2) We propose EVA, a unified evidence-seeking visual agent built upon an evidence-grounded execution pipeline, featuring *Iterative Tool Replanning*, *Evidence Consistency Verification*, and *Selective Evidence Routing*. (3) Experiments demonstrate that EVA improves reasoning reliability while maintaining efficiency through selective evidence utilization.

2 Related Work

2.1 Multimodal Reasoning for Hallucination Removal

Early efforts in multimodal reasoning focused on aligning visual and textual representations, leading to the development of multimodal large language models (MLLMs) Liu et al. [2023a], Chen et al. [2024b] and the adoption of in-context learning (ICL) Baldassini et al. [2024] and chain-of-thought (CoT) Wei et al. [2022] reasoning. These techniques enable models to decompose complex tasks into interpretable intermediate steps, improving both performance and transparency. The extension of CoT to multimodal settings, often referred to as multimodal chain-of-thought (MCoT), has led to diverse reasoning paradigms, including linear Zhang et al. [2023] and graph-based Besta et al. [2024] structures, as well as modality-specific designs such as MVoT Li et al. [2025], and Video-of-Thought Fei et al. [2024]. To further enhance reasoning robustness, multi-agent systems (MASs) Wooldridge [2009] have been introduced to leverage collective intelligence across multiple agents. In particular, multi-agent debate (MAD) frameworks Zheng et al. [2024], Wang et al. [2025c], Lin et al. [2024] aim to mitigate hallucination by aggregating diverse reasoning paths through interaction and consensus. Recent work Liang et al. [2025] further reframes MAD as a process of detecting undercover agents by employing multimodal counterfactual tests.

Despite these advances, existing approaches remain largely confined to language-space reasoning and implicit verification signals, such as agreement, consistency, or counterfactual plausibility. As a result, they may still fail when multiple agents share the same incorrect visual assumptions, leading to consistent but unsupported conclusions. Moreover, these methods lack mechanisms for selectively acquiring task-relevant visual evidence, and therefore cannot adaptively determine when additional perceptual grounding is required. In contrast, our approach introduces an adaptive fast–slow reasoning mechanism that explicitly triggers evidence acquisition when necessary, enabling more reliable multimodal reasoning grounded in observable evidence.

2.2 Tool-Augmented Multimodal Reasoning

Recent work has explored augmenting MLLMs with external tools to improve visual reasoning capabilities. Early efforts focused on building infrastructure and datasets for tool integration. For instance, LLaVA-Plus Liu et al. [2024b] introduced a dedicated tool server, while CogCoM Qi et al. [2024] and TACO Liu et al. [2024c] constructed large-scale training data with structured manipulation traces and tool-augmented reasoning demonstrations. These approaches aim to teach models how to interact with tools through supervised learning. Subsequent research has investigated different paradigms for tool use. One prominent direction Zhao et al. [2025], Zhang et al. [2025] trains models

to generate executable code for complex reasoning tasks, enabling flexible and expressive interactions with external environments. Another line of work Wang et al. [2025b,d,a], Zhu et al. [2025b], Zheng et al. [2025], Song et al. [2026] focuses on lightweight, atomic visual tools, such as zoom-in or region selection, to enhance perception in a more structured and controllable manner.

However, existing tool-augmented approaches typically rely on extensive data curation and model fine-tuning, resulting in high training costs and limited scalability. Furthermore, their tool-use policies are often brittle, exhibiting poor generalization to unseen task distributions beyond the training scope. In contrast, our approach treats tool use as an adaptive inference-time process, guided by explicit evidence acquisition and verification. By combining selective fast–slow reasoning with a replan-and-verify execution pipeline, EVA enables flexible and reliable tool use without requiring exhaustive supervision or fixed tool policies.

3 Method

3.1 Problem Formulation

We study hallucination-resistant multimodal reasoning, where a model answers a question grounded in an input image using verifiable visual evidence. Given an image I and a question q, the goal is to predict an answer a based on task-relevant evidence E extracted from the image:

$$a = f(q, I, E), \quad \text{s.t. } E \subseteq I \text{ and } E \Rightarrow a. \tag{1}$$

Here, E denotes observable evidence such as text regions, objects, attributes, or spatial relations. The condition $E \Rightarrow a$ requires the answer to be supported by explicit visual evidence rather than implicit visual assumptions.

However, reliable evidence-grounded reasoning remains challenging for tool-augmented MLLMs. We identify two key issues: (1) **Tool Invocation Reliability:** The model may select inappropriate tools or generate suboptimal tool parameters. (2) **Evidence–Reasoning Inconsistency:** The model may ignore or contradict the acquired evidence during answer generation. To address these challenges, we formulate multimodal reasoning as an evidence-grounded decision process that iteratively acquires evidence and verifies answer consistency.

3.2 Overview of EVA

We propose EVA, an Evidence-seeking Visual Agent that reframes multimodal reasoning as explicit evidence acquisition and verification. EVA is built around an evidence-grounded execution pipeline designed to address the two challenges above. The pipeline contains two core mechanisms. First, *Iterative Tool Replanning* improves tool invocation reliability by refining tool selection and parameters based on intermediate observations. Rather than relying on a single-pass tool plan, EVA allows the model to inspect obtained evidence and replan subsequent tool usage. Second, *Evidence Consistency Verification* addresses evidence–reasoning inconsistency by checking whether the synthesized answer is fully supported by the collected evidence and revising unsupported conclusions. In addition, EVA incorporates *Selective Evidence Routing*, an adaptive fast–slow reasoning mechanism that determines whether external evidence is necessary for a given query. Since many visually simple questions can be answered directly, this routing mechanism avoids unnecessary tool invocation while preserving robustness for complex or high-risk cases.

3.3 Evidence-Grounded Execution Pipeline

EVA's core execution pipeline targets reliable evidence acquisition and evidence-consistent answer generation. Given an image-question pair (I, q), EVA first produces an initial tool-use plan:

$$(\mathcal{P}^{(0)}, \mathcal{T}^{(0)}) = \pi_\theta(I, q), \tag{2}$$

where $\mathcal{P}^{(0)}$ denotes the initial reasoning plan and $\mathcal{T}^{(0)} = t_1^{(0)}, \ldots, t_k^{(0)}$ denotes the corresponding tool calls, including tool types and parameters.

Each tool call is executed on the image to obtain visual evidence:

$$E_i^{(0)} = t_i^{(0)}(I), \quad E^{(0)} = E_1^{(0)}, \ldots, E_k^{(0)}. \tag{3}$$

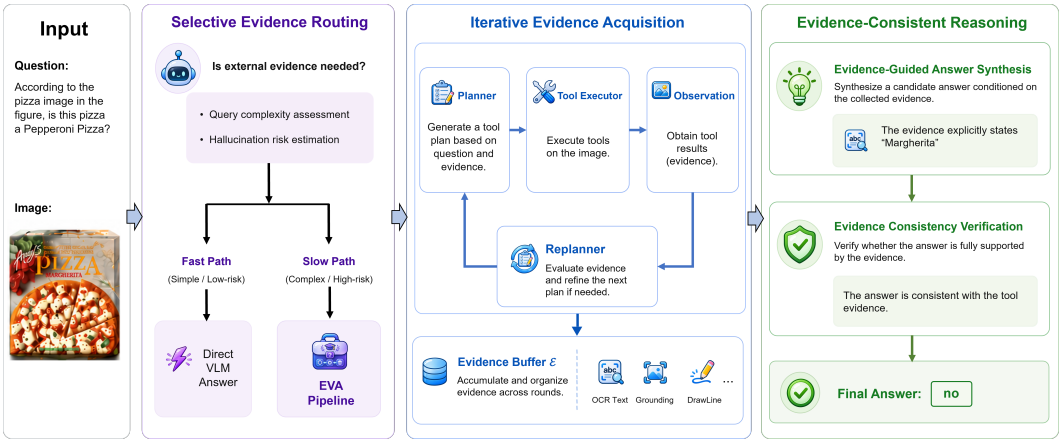


Figure 2: Overview of EVA. EVA first performs Selective Evidence Routing to determine whether external evidence is required. For complex queries, EVA activates an evidence-grounded execution pipeline with iterative tool replanning, evidence-guided answer synthesis, and evidence consistency verification.

However, a single tool invocation round can be unreliable. The initial plan may use an unsuitable tool, localize the wrong image region, apply incorrect OCR settings, or extract evidence that is insufficient for answering the question. EVA therefore introduces an iterative replanning process.

Iterative Tool Replanning. At each round r, EVA evaluates the accumulated evidence $E^{(r)}$ with respect to the question q and updates the tool plan when the evidence is incomplete, irrelevant, or uncertain:

$$(\mathcal{P}^{(r+1)}, \mathcal{T}^{(r+1)}) = \pi_\theta(I, q, E^{(r)}). \tag{4}$$

The new tool calls are then executed to acquire additional evidence:

$$\Delta E^{(r+1)} = \mathcal{T}^{(r+1)}(I), \quad E^{(r+1)} = E^{(r)} \cup \Delta E^{(r+1)}. \tag{5}$$

This mechanism allows EVA to dynamically adjust its evidence-seeking behavior. For example, if object grounding fails to resolve a queried entity, EVA may switch to OCR, or refine the target description depending on the intermediate observation. In this way, Iterative Tool Replanning directly improves tool invocation reliability by correcting suboptimal tool choices and parameters during execution.

Evidence-Guided Answer Synthesis. After evidence acquisition, EVA synthesizes a candidate answer from the collected evidence:

$$a_{\text{syn}} = g_\theta(q, E). \tag{6}$$

Unlike direct MLLM inference, the synthesis step conditions the answer primarily on explicit evidence rather than implicit visual assumptions. The model is encouraged to aggregate relevant observations and produce an answer that is traceable to the acquired evidence.

Evidence Consistency Verification. Even when the evidence is sufficient, the synthesized answer may still be inconsistent with the tool results. To address this, EVA performs a final verification step:

$$a = h_\theta(q, E, a_{\text{syn}}). \tag{7}$$

The verifier checks whether the candidate answer is entailed by the collected evidence. If the answer ignores relevant observations, contradicts tool outputs, or relies on unsupported assumptions, the verifier revises the prediction or rejects the unsupported reasoning path. This stage explicitly enforces evidence–reasoning consistency and reduces hallucinated conclusions caused by ungrounded answer synthesis.

3.4 Selective Evidence Routing

Although external tools improve reliability, invoking them for every query is inefficient and sometimes unnecessary. Many visually straightforward questions can be answered accurately through direct visual reasoning, while complex or high-risk queries require explicit evidence acquisition.

To balance efficiency and robustness, EVA introduces *Selective Evidence Routing*. Before activating the full evidence-grounded execution pipeline, EVA predicts whether external evidence is needed:

$$z = \mathbb{I}[\text{need_evidence}(I, q)]. \tag{8}$$

If $z = 0$, EVA follows a fast path and directly answers the query:

$$a = f_\theta(I, q). \tag{9}$$

If $z = 1$, EVA follows the slow path and activates the full evidence-grounded pipeline:

$$a = h_\theta(q, E, g_\theta(q, E)). \tag{10}$$

This adaptive strategy treats tool use as a targeted intervention rather than a default operation. It avoids unnecessary tool calls for simple cases, while routing complex, ambiguous, or hallucination-prone queries to the more reliable replan-and-verify pipeline.

4 Experiments

4.1 Experimental Setup

Evaluated Benchmarks. We adopt the VLMEvalKit Duan et al. [2024] framework for fair evaluation. We evaluate EVA on four benchmark datasets that cover diverse aspects of multimodal reasoning:

General Reasoning Benchmarks. We adopt MMStar Chen et al. [2024a] and MMMU Yue et al. [2024] as comprehensive benchmarks for evaluating overall multimodal reasoning ability. MMStar consists of 1,500 challenging visual questions spanning multiple domains, while MMMU contains 11,500 college-level multimodal problems that require advanced cross-disciplinary reasoning.

Hallucination Detection Benchmarks. We further evaluate hallucination robustness using HallusionBench Guan et al. [2024] and POPE Li et al. [2023]. HallusionBench includes 346 carefully designed samples that induce hallucinations through misleading visual cues. POPE contains 3,000 binary (yes/no) questions aimed at systematically evaluating object hallucination in image descriptions.

Baselines. We compare EVA against a diverse set of baseline methods to comprehensively benchmark its performance: **(1) Single-Agent Models.** These include both open-source and closed-source multimodal models. Open-source baselines include DeepSeek-VL-7B Lu et al. [2024], LLaVA-OneVision-7B Li et al. [2024], LLaVA-NEXT-Llama3-8B Liu et al. [2024a], LLaVA-v1.5-13B Liu et al. [2023b], InternVL2-8B Chen et al. [2024b], InternVL3-14B Zhu et al. [2025a], and Qwen2.5-VL Bai et al. [2025]. Closed-source baselines include Gemini-1.5-Pro Team et al. [2024], GPT-4o-mini Hurst et al. [2024], GPT-4V OpenAI [2023], and Claude-3.5-Sonnet Anthropic [2024]. These models integrate visual and textual inputs using different architectural and training strategies to enhance multimodal understanding. **(2) Self-Refine Madaan et al. [2023].** Self-Refine is an approach that improves initial outputs of large language models through iterative self-feedback and refinement. **(3) Multi-Agent Debate (MAD) Liang et al. [2024].** MAD represents a representative multi-agent paradigm, where multiple agents engage in iterative debate and reach final decisions via majority voting or a centralized judge. **(4) Multi-agent Undercover Gaming (MUG) Liang et al. [2025].** MUG reframes multi-agent debate as a process of identifying "undercover" agents (i.e., hallucinating agents) via multimodal counterfactual testing, where modified visual evidence is introduced to verify agents' consistency against ground truth.

Controlled Tool-Use Baselines. We introduce four controlled baselines using the same backbone, tools, and evaluation protocol as EVA to isolate the effects of tool access and inference computation. **OneShot** performs a single round of tool use without replanning or verification. **ReAct**

Model	MMMU	MMStar	HallusionBench				POPE			
	Acc.	Acc.	aAcc.	fAcc.	qAcc.	Avg.	Acc.	Pre.	Rec.	F1
DeepSeek-VL-7B	38.3	40.5	53.9	24.9	24.6	34.5	86.8	94.3	78.4	85.6
LLaVA-OneVision-7B	47.9	61.9	48.4	21.4	25.1	31.6	87.9	96.7	78.5	86.6
LLaVA-Next-Llama3-8B	43.1	43.9	52.3	25.7	21.3	33.1	87.1	88.1	95.1	87.1
LLaVA-v1.5-13B	37.0	34.3	45.3	34.9	11.0	24.5	87.5	91.4	82.3	85.9
InternVL2-8B	50.2	61.5	63.9	35.0	36.0	45.0	86.0	96.6	74.7	84.2
InternVL2-26B	50.7	61.0	67.9	44.8	41.8	51.5	87.7	96.4	78.3	86.4
Gemini-1.5-Pro	60.6	59.1	63.0	36.1	37.6	45.6	88.6	93.0	83.9	88.2
GPT4o mini (detail-high)	57.3	46.5	61.9	41.3	34.9	46.1	84.2	95.5	71.7	81.9
GPT-4v (detail-high)	53.8	43.9	65.8	38.4	35.2	46.5	83.9	93.9	72.5	81.8
Claude3.5-Sonnet	65.9	62.2	66.4	41.6	41.8	49.9	78.7	97.1	59.2	73.6
Qwen2.5VL-7B	45.0	61.2	64.8	34.9	39.6	46.4	87.4	96.8	77.3	85.9
Qwen2.5VL-7B (Self-Refine)	45.8	61.5	67.3	38.2	40.9	48.8	85.9	97.4	73.8	84.0
Qwen2.5VL-7B (MAD-Vote)	44.7	57.4	56.4	26.9	30.1	37.8	80.0	80.4	68.7	74.1
Qwen2.5VL-7B (MAD-Judge)	47.4	62.3	64.5	43.2	42.9	50.2	85.2	94.4	74.9	83.5
Qwen2.5VL-7B (MUG)	50.3	63.8	69.4	43.9	**47.9**	53.8	**88.4**	**95.6**	80.5	**87.4**
Qwen2.5VL-7B (EVA(Ours))	**51.4**	**64.3**	**70.4**	**45.4**	47.3	**54.4**	88.1	94.1	**81.3**	87.3
InternVL3-14B	59.8	68.7	69.8	47.7	47.7	55.1	89.3	92.1	86.5	89.5
InternVL3-14B (Self-Refine)	45.9	61.2	70.3	46.5	45.5	54.1	88.6	91.9	86.7	89.2
InternVL3-14B (MAD-Vote)	55.2	62.9	70.5	48.6	46.4	55.2	89.4	94.2	83.9	88.8
InternVL3-14B (MAD-Judge)	60.2	68.9	72.5	47.9	48.9	56.4	89.6	94.1	87.3	90.6
InternVL3-14B (MUG)	60.7	**69.1**	**73.3**	51.2	49.5	58.0	**90.1**	94.1	**88.2**	**91.1**
InternVL3-14B (EVA(Ours))	**62.7**	**69.1**	72.2	**52.3**	**51.9**	**58.8**	89.5	**94.2**	84.2	88.9

Table 1: Performance comparison across MMMU, MMStar, HallusionBench, and POPE benchmarks. Acc.: accuracy, aAcc.: average accuracy, fAcc.: unique figure accuracy, qAcc.: unique question accuracy, Avg.: average score of the three metrics. Pre.: precision, Rec.: recall.

iteratively selects tools based on intermediate observations without explicit evidence verification. **FixedSecondRound** always executes a second tool-use round regardless of evidence sufficiency. **SameEvidenceSelfRefine** refines its answer using exactly the same evidence acquired by EVA, isolating the effect of structured reasoning from evidence access.

4.2 Main Results

Table 1 summarizes the comparison between EVA and existing multimodal reasoning approaches across MMMU, MMStar, HallusionBench, and POPE. Overall, EVA consistently improves both reasoning capability and hallucination resistance across different model backbones, demonstrating the effectiveness of evidence-grounded reasoning.

EVA consistently improves both small and large MLLMs. Integrating EVA leads to stable gains on both Qwen2.5VL-7B and InternVL3-14B across nearly all benchmarks. On Qwen2.5VL-7B, EVA improves the baseline from 61.2% to 64.3% on MMStar, and from 46.4% to 54.4% on HallusionBench. Similar improvements are observed on InternVL3-14B, where EVA achieves 69.1%, and 58.8%, respectively. Notably, EVA built on Qwen2.5VL-7B surpasses several proprietary models such as GPT-4v and Claude3.5-Sonnet on HallusionBench, suggesting that structured evidence-grounded reasoning can effectively compensate for limitations in model scale.

EVA achieves stronger hallucination resistance than self-reflection and multi-agent methods. Compared with Self-Refine, MAD-Vote, MAD-Judge, and MUG, EVA achieves the best or near-best performance across most hallucination-related metrics. On HallusionBench, EVA obtains the highest average scores of 54.4% and 58.8% on the two backbones respectively, while also achieving strong figure and question accuracy. Compared with self-reflection and multi-agent approaches, EVA directly improves the reliability of evidence acquisition and evidence utilization instead of relying solely on answer refinement or opinion aggregation. While multiple agents may still converge to the same incorrect perceptual assumption, EVA explicitly verifies visual evidence through Iterative Tool Replanning and Evidence Consistency Verification, leading to more reliable multimodal reasoning.

4.3 Behavior Analysis of EVA

To better understand EVA beyond aggregate benchmark performance, we analyze its tool usage behavior, iterative replanning dynamics, and performance–efficiency trade-off on HallusionBench Guan et al. [2024].

4.3.1 Tool Usage Behavior

We first analyze EVA's tool invocation behavior. EVA activates external tools on only 19.87% of samples, with an average of 0.323 tool calls per sample, indicating that *Selective Evidence Routing* invokes tools conservatively rather than relying on external evidence by default. Most queries are resolved through direct multimodal reasoning, while tool usage is primarily triggered for perception-intensive cases.

We further examine EVA's multi-round execution dynamics, where the tool invocation pattern also reveals the adaptive behavior of EVA. The tool-call rate increases from 20.34% in the first round to 56.61% in the second round, indicating that EVA progressively refines earlier observations through targeted tool usage once additional evidence becomes necessary. This behavior aligns with the design objective of *Iterative Tool Replanning*, where intermediate observations guide subsequent tool selection and parameter refinement.

Interestingly, tool-triggered samples achieve higher accuracy (76.72%) than no-tool samples (68.37%), despite generally corresponding to more difficult perception-intensive queries. This suggests that EVA effectively identifies cases where external evidence is beneficial and improves reasoning reliability through targeted evidence acquisition.

Figure 3 further illustrates the distribution of external tool usage in EVA. OCR is the most frequently invoked tool, followed by chart-understanding and grounding-related tools. This distribution suggests that many challenging multimodal reasoning cases require fine-grained perception capabilities, particularly for text recognition and structured visual understanding.

Finally, EVA demonstrates highly reliable tool execution in practice, with failures (undefined tool name output) observed in only 0.11% of evaluated samples. Together, these results validate EVA's evidence-grounded reasoning paradigm: tools are used sparingly on simple queries while iterative evidence acquisition is selectively applied when the initial evidence is insufficient.

4.3.2 Performance–Latency Trade-off

Figure 4 compares EVA with direct inference and representative multi-agent reasoning methods under the performance–latency trade-off. All latency measurements are conducted on a single NVIDIA GeForce RTX 4090 GPU to ensure a fair and consistent comparison across methods.

Although multi-agent approaches such as MAD-Judge and MUG improve reasoning accuracy through iterative deliberation, they incur substantially higher inference costs due to repeated model interactions and unconditional reasoning expansion. In contrast, EVA achieves the best overall balance between reasoning performance and inference efficiency. Several important observations can be drawn from the results: (1) Direct inference with Qwen2.5-VL-7B achieves 46.4% performance with low latency (4.35s), but suffers from severe hallucination and insufficient evidence verification. (2) Multi-agent methods improve accuracy through iterative deliberation, with MAD-Judge reaching 50.2% and MUG achieving 53.8%. However, these gains come at significantly increased latency, requiring 22.83s and 26.35s per sample, respectively. (3) EVA without routing already achieves strong reasoning performance (53.5%) while reducing latency to 17.50s, demonstrating that explicit evidence acquisition and verification are more efficient than purely language-space deliberation. Full EVA further improves performance to 54.4% while reducing latency to only 9.84s through *Selective Evidence Routing*, which adaptively invokes external tools only when additional evidence is necessary. These results suggest that explicit evidence-grounded reasoning provides a more efficient and scalable path toward hallucination-resistant multimodal reasoning than simply increasing language-space deliberation depth.

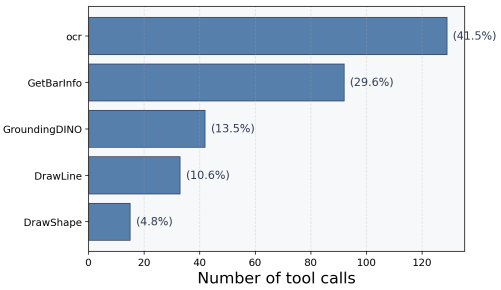


Figure 3: Distribution of external tool usage in EVA. OCR and chart-understanding tools dominate the invocation distribution, indicating that fine-grained perception remains a major source of multimodal hallucination.

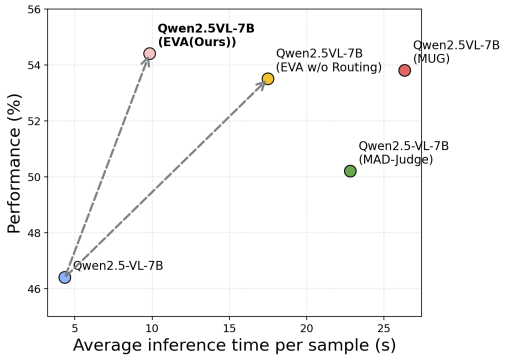


Figure 4: Performance–latency trade-off comparison. EVA achieves the best overall performance while maintaining significantly lower inference latency than multi-agent reasoning methods. w/o Routing: without Selective Evidence Routing.

Table 2: Ablation study on HallusionBench. aAcc.: average accuracy, fAcc.: unique figure accuracy, qAcc.: unique question accuracy, Avg.: average score of the three metrics.

Method	aAcc.	fAcc.	qAcc.	Avg.
EVA	**70.4**	**45.4**	**47.3**	**54.4**
w/o Iterative Tool Replanning	68.1	41.7	44.2	51.3
w/o Evidence Consistency Verification	67.4	40.9	43.5	50.6
w/o Selective Evidence Routing	69.5	44.3	46.7	53.5
w/o Replanning & Verification	65.8	38.6	41.4	48.6

4.3.3 Failure Source Analysis

To better understand EVA's remaining limitations, we manually examine 50 failure cases and categorize each by the earliest decisive error in its reasoning process. Poor tool selection or incorrect tool parameters account for 46% (23/50) of the failures, while incorrect or insufficient tool outputs account for another 22% (11/50). The remaining 32% (16/50) arise from incorrect interpretation or utilization of otherwise correct and sufficient evidence. In total, 68% of the failures originate during evidence acquisition, whereas 32% occur during evidence utilization. These results indicate that improving tool planning and evidence quality offers substantial headroom, while evidence-consistent answer generation remains important even when reliable observations have been obtained.

4.4 Ablation Study

Component Ablation. To better understand the contribution of each component in EVA, we conduct ablation studies on HallusionBench. We progressively remove the proposed mechanisms from the full model, including *Iterative Tool Replanning*, *Evidence Consistency Verification*, and *Selective Evidence Routing*. As shown in Table 2, removing either *Iterative Tool Replanning* or *Evidence Consistency Verification* leads to clear performance degradation, demonstrating that both reliable evidence acquisition and evidence-consistent answer generation are critical for hallucination-resistant reasoning. In particular, removing *Evidence Consistency Verification* causes a notable drop (4.5% in fAcc. and 3.8% in qAcc.), suggesting that models may still generate unsupported conclusions even when external tool evidence is available. Similarly, removing *Iterative Tool Replanning* weakens the model's ability to recover from incorrect tool choices or suboptimal tool parameters, resulting in lower overall accuracy.

When both replanning and verification are removed, the performance drops further to 48.6% average accuracy, indicating that these two mechanisms are complementary. In contrast, removing *Selective Evidence Routing* causes a smaller but consistent decrease. This suggests that routing mainly

Table 3: Evaluation of EVA with a frontier MLLM. The same GPT-5.5 backbone is used for both settings. Tool calls are reported as per-sample averages.

Method	MMMU	MMStar	HallusionBench	POPE	Tool calls
GPT-5.5 (Direct)	78.2	75.4	62.3	87.0	0
GPT-5.5 + EVA	**79.3**	**76.2**	**64.4**	**89.6**	0.14

Table 4: Controlled comparison on HallusionBench using Qwen2.5-VL-7B. All methods are training-free and share the same tool set and evaluator. Resource statistics are per-sample averages.

Method	HallusionBench Avg.	Model calls	Generated tokens	Tool calls	Latency (s)
OneShot	52.1	1.92	175.80	1.06	7.98
ReAct	51.5	2.18	183.54	1.18	8.72
FixedSecondRound	52.5	2.84	242.80	1.98	13.53
SameEvidenceSelfRefine	53.1	2.62	221.38	0.32	10.20
EVA	**54.4**	2.62	218.26	0.32	9.84

improves the efficiency–robustness trade-off by avoiding unnecessary tool usage for simple cases, while the main reliability gains come from the replan-and-verify execution pipeline. Overall, the ablation results validate the effectiveness of EVA's design: *Iterative Tool Replanning* improves tool invocation reliability, *Evidence Consistency Verification* mitigates evidence–reasoning inconsistency, and *Selective Evidence Routing* provides adaptive and efficient evidence utilization.

Generalization to Frontier MLLMs. We further evaluate EVA on a stronger frontier MLLM by comparing GPT-5.5 with and without EVA under the same benchmark settings. As shown in Table 3, EVA consistently improves performance across all four benchmarks. The largest improvement on HallusionBench suggests that explicit evidence acquisition and verification remain beneficial even when the backbone has stronger built-in visual understanding capabilities. A category-level analysis further indicates that these gains are particularly pronounced on visual-semantic reasoning problems involving charts and text-rich images, including a 4.2% improvement on OCR-related questions. Notably, these improvements require only 0.14 external tool calls per sample on average, demonstrating that EVA can complement frontier MLLMs with limited additional tool usage.

Controlled Tool-Use and Compute Comparison. We further investigate whether EVA's improvement can be explained solely by access to external tools or increased inference-time computation. Table 4 compares EVA against four training-free baselines with the same backbone, tool set, and evaluator. Despite having access to the same tools, OneShot and ReAct underperform EVA by 2.3 and 2.9 points, respectively, indicating that tool availability alone is insufficient. FixedSecondRound also remains 1.9 points below EVA, despite requiring more model calls, generated tokens, tool calls, and latency. More importantly, SameEvidenceSelfRefine receives exactly the same acquired evidence as EVA and uses the same average number of model and tool calls, yet scores 1.3 points lower while generating more tokens and incurring higher latency. Together, these controlled comparisons suggest that EVA's improvements stem from its coordinated selective routing, conditional replanning, and evidence consistency verification, rather than simply from additional tool access, evidence availability, or inference computation.

5 Conclusion

We presented EVA, an evidence-seeking visual agent for reliable multimodal reasoning. EVA improves reasoning reliability through three key mechanisms: *Selective Evidence Routing*, *Iterative Tool Replanning*, and *Evidence Consistency Verification*. Together, these components enable EVA to adaptively acquire external evidence, refine tool usage, and ensure that final predictions remain grounded in verifiable observations. Extensive experiments show that EVA consistently reduces hallucination and improves multimodal reasoning performance across challenging benchmarks, outperforming direct MLLM inference and existing multi-agent reasoning methods while avoiding unnecessary tool usage. Our results highlight the importance of explicit evidence acquisition and verification for building reliable multimodal agents.

Acknowledgement

This work was supported by the grants from the National Natural Science Foundation of China 62372014, Beijing Nova Program and Beijing Natural Science Foundation 4252040.

References

Anthropic. Claude 3.5 sonnet, 2024.

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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1539–1550, 2024.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690, 2024.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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, 2024b.

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

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Dayong Liang, Xiao-Yong Wei, and Changmeng Zheng. Multi-agent undercover gaming: Hallucination removal via counterfactual test for multimodal reasoning. *arXiv preprint arXiv:2511.11182*, 2025.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 17889–17904, 2024.

Zheng Lin, Zhenxing Niu, Zhibin Wang, and Yinghui Xu. Interpreting and mitigating hallucination in mllms through multi-agent debate. *arXiv preprint arXiv:2407.20505*, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. In *European conference on computer vision*, pages 126–142. Springer, 2024b.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21740–21751, 2024c.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

OpenAI. Gpt-4v(ision) system card, 2023.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. Cogcom: Train large vision-language models diving into details through chain of manipulations. 2024.

Mingyang Song, Haoyu Sun, Jiawei Gu, Linjie Li, Luxin Xu, Ranjay Krishna, and Yu Cheng. Adareasoner: Dynamic tool orchestration for iterative visual reasoning. *arXiv preprint arXiv:2601.18631*, 2026.

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

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025b.

Mengzhen Wang, Xunbin Huang, Jiayuan Xie, Shukai Ma, Jiale Men, Dayong Liang, and Yi Cai. From model diagram to code: A benchmark dataset and multi-agent framework. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 1754–1763, 2025c.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7907–7915, 2025d.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Michael Wooldridge. *An introduction to multiagent systems*. John wiley & sons, 2009.

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, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. Pyvision: Agentic vision with dynamic tooling. *arxiv preprint arxiv:2507.07998*, 2025.

Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, and Qing Li. A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 419–428, 2024.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

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

Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Zheng Huang, Mingyu Liu, Hao Chen, Cheng Zou, Jingdong Chen, Ming Yang, et al. Active-o3: Empowering multimodal large language models with active perception via grpo. *arXiv preprint arXiv:2505.21457*, 2025b.

A Benchmark Details

We evaluate EVA on four representative multimodal reasoning and hallucination benchmarks, covering general multimodal reasoning, fine-grained visual perception, and hallucination-resistant reasoning.

MMMU. MMMU Yue et al. [2024] is a comprehensive multimodal benchmark designed to evaluate multimodal large language models on massive multi-discipline tasks requiring college-level domain knowledge and deliberate reasoning. The benchmark spans a wide range of subjects, including science, engineering, medicine, humanities, and social sciences. Following prior work, we report results on the validation set.

MMStar. MMStar Chen et al. [2024a] is a high-quality vision-centric benchmark consisting of 1,500 carefully curated samples selected by human experts. The benchmark emphasizes visually grounded reasoning and fine-grained perception, requiring models to accurately interpret visual details rather than relying on language priors or commonsense biases.

POPE. POPE Li et al. [2023] is a benchmark specifically designed to evaluate object hallucination in multimodal models. It contains three evaluation settings: *random*, *popular*, and *adversarial*, each targeting different hallucination scenarios. Following standard evaluation protocols, we report the average F1 score across all three settings as the overall metric. We additionally report accuracy, precision, and recall:

$$\mathrm{F1} = \frac{2 \times (\mathrm{Precision} \times \mathrm{Recall})}{\mathrm{Precision} + \mathrm{Recall}}. \tag{11}$$

HallusionBench. HallusionBench Guan et al. [2024] is a benchmark designed to evaluate hallucination behaviors in vision-language models. For each question, the benchmark provides an original image and a modified image, where the correct answers may differ depending on the visual content. This setup explicitly tests whether the model grounds its reasoning in observable evidence rather than relying on memorized priors or language biases.

We report three metrics following the official evaluation protocol:

- **aAcc**: the overall accuracy across all atomic question-image pairs;
- **qAcc**: the mean accuracy over unique questions. A question may correspond to multiple image variants, and a question is considered correct only if the model answers all associated question-image pairs correctly;
- **fAcc**: the mean accuracy over figures. Since each figure is associated with multiple questions, a figure is considered correct only if the model correctly answers all questions related to that figure.

B Tool Library

EVA relies on a lightweight yet diverse set of external visual tools for explicit evidence acquisition. These tools are designed to complement the perception limitations of general MLLMs by providing structured visual observations that can be iteratively queried and verified during reasoning.

Each tool is invoked through structured arguments generated by the model. Table 5 summarizes the tool set used in EVA. The tools in EVA serve complementary roles during evidence acquisition. OCR provides textual evidence, GroundingDINO enables object-level localization, and visualization tools such as `DrawLine` and `DrawShape` assist spatial reasoning and region inspection. In addition, `GetBarInfo` provides structured parsing for chart understanding tasks. Importantly, EVA does not assume that a single tool invocation is always sufficient. Instead, tools are dynamically selected and refined through Iterative Tool Replanning. For instance, the model may first localize an object using `GroundingDINO`, then invoke OCR to extract the potential text information, or add visual guides via `DrawLine` to better estimate spatial relations. This adaptive tool usage enables EVA to progressively acquire evidence that is both task-relevant and verifiable.

To support flexible tool-augmented reasoning, EVA interfaces with external tools through a unified Tool Server. The Tool Server acts as an MCP-like orchestration layer that standardizes communication between the language model and heterogeneous visual tools.

Table 5: External visual tools used in EVA.

Tool	Function	Input Arguments
`ocr`	Extract text regions and corresponding bounding boxes from the image.	No additional arguments required.
`GroundingDINO`	Locate image regions corresponding to a textual description. Used for object grounding and region-level evidence acquisition.	`description`: textual query describing the target object or region.
`DrawLine`	Overlay horizontal or vertical guide lines on the image to facilitate spatial reasoning, alignment estimation, or chart interpretation.	`line_type`: horizontal or vertical; `coordinates`: absolute pixel coordinates.
`DrawShape`	Overlay geometric shapes (e.g., rectangles, ellipses, circles) on the image for region highlighting and evidence annotation.	`bboxes`: list of shapes and corresponding coordinates.
`GetBarInfo`	Parse structural information from bar charts or tabular visualizations, including bar positions and relative values.	No additional arguments required.

C Prompt Templates

We provide the prompt templates used in EVA in Fig. 5–Fig. 7. The (re)planning prompt serves two purposes. First, it decides whether external tools are necessary for the current image-question pair. Second, when tools are needed, it generates structured tool calls according to the predefined tool descriptions and argument specifications. We do not introduce a separate `need_evidence` classifier. Instead, the evidence-routing decision is implicitly represented by the generated tool list: if `tool_calls=[]`, EVA follows the fast path and directly returns an answer; otherwise, EVA enters the slow path and invokes the evidence-grounded execution pipeline. After tool execution, the answer synthesis prompt aggregates the collected evidence and produces a candidate answer. This stage is instructed to rely on tool results as explicit evidence while avoiding over-trusting noisy or inconclusive outputs. Finally, the consistency check prompt verifies whether the candidate answer is fully supported by the tool results. It only revises the answer when there is a clear contradiction or missing support, thereby reducing unnecessary changes while enforcing evidence–reasoning consistency.

D Impact of Replanning Rounds

We analyze the effect of the number of replanning rounds, which corresponds to the iterative evidence refinement steps performed after each round of tool execution. The performance metrics across different replanning rounds are shown in Table 6. The results indicate that the initial replanning rounds yield higher performance, likely because early-stage replanning effectively corrects incorrect tool choices and refines evidence acquisition strategies based on fresh intermediate observations. Specifically, the performance peaks at round 1, with a notable improvement over the no-replanning setting, reflecting the effectiveness of a single corrective replanning step for reliable tool invocation. As the number of replanning rounds increases, the performance gradually stabilizes or slightly declines, suggesting diminishing returns from additional evidence refinement iterations.

(Re)Planning Prompt

You are a careful visual reasoning agent. First decide whether tools are actually needed.
Only propose tool calls when they are likely to improve accuracy for reading tiny text, locating an object, zooming in, or extracting structured evidence.
Only output valid JSON with schema: {“plan”: [string], “tool_calls”: [{“name”: string, “args”: object}], “final_answer”: string}.
If tools are unnecessary, set tool_calls=[].
Tool descriptions: {TOOL_DESCRIPTIONS}
Tool argument specs: {TOOL_ARG_SPECS}
Previous evidence and tool history: {payload}
Question: {question}

Figure 5: Prompt template for (re)planning.

Answer Synthesis Prompt

You are a strict evidence synthesizer. Read the question, plan, and tool results carefully, then produce the best supported answer.
Treat tool_results as evidence, but do not let noisy tool outputs override obvious direct visual evidence.
Do not invent measurements or conclusions that are not supported by the tool outputs.
If the tool results are inconclusive, return the most conservative answer.
Only output valid JSON with schema: {"explanation": string, “final_answer ": string}.
Previous evidence and tool history: {payload}
Question: {question}

Figure 6: Prompt template for evidence-guided answer synthesis.

E Qualitative Analysis

To better understand the behavior of EVA, we present three qualitative examples covering different multimodal reasoning challenges, including visual illusion, chart reasoning, and text-based visual understanding.

Figure 8 demonstrates EVA’s ability to refine tool execution during reasoning. The model first applies a horizontal measurement line to compare the orange circles, but the initial alignment is insufficient for accurate comparison. Through Iterative Tool Replanning, EVA adjusts the measurement line to intersect more representative regions of the circles, resulting in more reliable evidence extraction. This example highlights the importance of adaptive tool parameter refinement in perception-sensitive scenarios.

Figure 9 illustrates the role of Evidence Consistency Verification. EVA first extracts structured values from the bar chart and correctly identifies the import values of multiple countries. However, the answer synthesis stage still produces an incorrect conclusion that contradicts the extracted evidence. The verification stage subsequently detects this inconsistency and revises the final answer. This example demonstrates that reliable evidence acquisition alone is insufficient; the reasoning process must also explicitly enforce consistency between evidence and final predictions.

Figure 10 shows how EVA improves tool invocation reliability. The model initially selects an inappropriate chart analysis tool for a text-based image, resulting in an invalid tool response. EVA then replans the execution strategy and switches to OCR to extract textual evidence from the image. This example demonstrates EVA’s ability to dynamically recover from incorrect tool selection and adapt its evidence acquisition strategy based on intermediate observations.

Consistency Check Prompt

You are a strict evidence-reasoning consistency checker.
Step 1 produced tool calls and step 2 produced a candidate answer from tool evidence. Your job is to judge whether the candidate answer is fully supported by the tool results.
Only change the answer when there is a clear contradiction or missing support.
Otherwise preserve the candidate answer.
Only output valid JSON with schema: {"explanation": string, " final_answer ": string}
Previous evidence and tool history: {payload}
Question: {question}

Figure 7: Prompt template for Evidence Consistency Verification.

Table 6: Impact of replanning rounds on HallusionBench.

# Replanning Rounds	aAcc.	fAcc.	qAcc.	Avg.
0	68.1	41.7	44.2	51.3
1	**70.5**	**45.4**	**47.3**	**54.4**
2	70.2	44.8	47.1	54.0
3	69.6	44.1	46.4	53.4

Overall, these qualitative examples demonstrate the complementary roles of EVA's core mechanisms. Iterative Tool Replanning improves the reliability of evidence acquisition by refining tool usage during execution, while Evidence Consistency Verification ensures that the final prediction remains grounded in the collected evidence.

F Limitations and Broader Impact

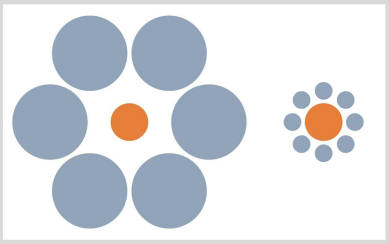
Limitations. Although EVA improves multimodal reasoning reliability through explicit evidence acquisition and verification, several limitations remain. First, EVA still depends on the quality of external visual tools. When tool outputs are inaccurate or incomplete (e.g., failed OCR on blurry text or incorrect grounding under heavy occlusion), the downstream reasoning process may still produce incorrect conclusions despite replanning and verification. Second, EVA introduces additional inference cost compared with direct MLLM inference due to iterative tool execution and consistency checking. While Selective Evidence Routing reduces unnecessary tool usage for simple cases, complex queries may still require multiple rounds of evidence acquisition and replanning, leading to increased latency.

Broader Impact. This work aims to improve the reliability and factual grounding of multimodal large language models by introducing explicit evidence acquisition and verification mechanisms. We believe such evidence-driven reasoning frameworks can positively impact real-world applications that require trustworthy multimodal understanding, including document analysis, educational assistants, scientific reasoning, accessibility tools, and human–AI collaboration systems. By reducing hallucination and encouraging verifiable reasoning processes, EVA may help improve transparency and user trust in multimodal AI systems.

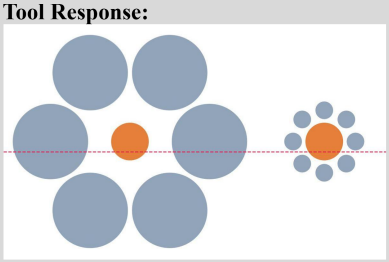
At the same time, our approach may also introduce potential risks. Although EVA improves evidence consistency, it still relies on external visual tools and underlying MLLMs, which may inherit biases, recognition errors, or dataset limitations from pretrained components. Incorrect tool outputs or biased visual evidence could still propagate through the reasoning pipeline and lead to misleading conclusions. Furthermore, EVA introduces additional computational overhead due to iterative evidence acquisition and verification, which may increase energy consumption and resource requirements compared with direct inference methods. Future work should therefore explore more efficient evidence selection and lightweight verification mechanisms. We hope this work encourages future research on

EVA Example

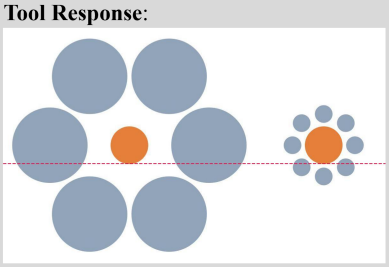
Question: Is the right orange circle the same size as the left orange circle?
Question Image:



Response by EVA (Planning):
{“plan”: “Use the DrawLine tool to measure the size of the orange circles", 'tool_calls': [{'name': DrawLine', 'args': {"line_type":"horizontal”, "coordinates":"y1=400"}}]}
Tool Response:



Response by EVA (Re-Planning):
{“plan”: “Adjust the horizontal measurement line to measure the size of the orange circles", 'tool_calls': [{'name': DrawLine', 'args': {"line_type":"horizontal”, "coordinates":"y1=425"}}]}
Tool Response:



Response by EVA (Answer Synthesis):
{'explanation': 'After adjusting the measurement line to pass through the bottom of both orange circles, the measured horizontal position are nearly identical, indicating that the two orange circles have the same size.','final_answer': 'yes'}
Response by EVA (Consistency Verification):
{'explanation': 'The synthesized answer is consistent with the measurement evidence. The horizontal line support the conclusion that the circles are the same size.','final_answer': 'yes'}

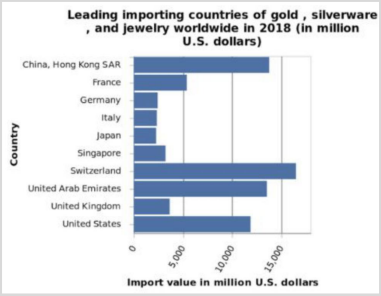
Figure 8: EVA resolves a visual illusion through iterative evidence acquisition. The initial measurement line is suboptimal for comparing the circle diameters, and EVA replans the tool invocation to adjust the measurement position. After refinement, EVA correctly concludes that the two orange circles have the same size.

trustworthy, transparent, and verifiable multimodal reasoning systems, while also promoting careful consideration of their societal impacts and responsible deployment.

EVA Example

Question: Is China, Hongkong SAR, the leading importing country of gold, silverware, and jewelry with the highest import value in 2018?

Question Image:



Response by EVA (Planning):
{'plan': 'Extract the bar info using tool and check the value of China, Hong Kong SAR.', 'tool_calls': [{'name': 'GetBarInfo', 'args': {}}]}

Tool Response:
Country | Import value in million U.S. dollars
Switzerland | 16000
China, Hong Kong SAR | 13500
United Arab Emirates | 13500
United States | 12000
…

Response by EVA (Re-Planning):
{'plan': 'Compare the extracted values to identify the country with the maximum import value and verify whether China, Hong Kong SAR ranks first.','tool_calls': []}

Response by EVA (Answer Synthesis):
{'explanation': 'The bar chart shows that China, Hong Kong SAR, had the highest import value of gold, silverware, and jewelry in 2018, with an import value of approximately 15,000 million U.S. dollars.', 'final_answer': 'yes'}

Response by EVA (Consistency Verification):
{'explanation': 'The synthesized answer is inconsistent with the extracted evidence. The tool response shows that Switzerland has the highest import value at 16,000 million U.S. dollars, while China, Hong Kong SAR has an import value of 13,500 million U.S. dollars. Therefore, China, Hong Kong SAR is not the highest.', 'final_answer': 'no'}

Figure 9: EVA corrects a hallucinated conclusion through Evidence Consistency Verification. Although the answer synthesis stage incorrectly predicts that China, Hong Kong SAR has the highest import value, the verification stage identifies that the extracted chart evidence actually supports Switzerland as the maximum.

EVA Example

Question: According to the text given in the image, is this a Washington style cheesecake?
Question Image:



Response by EVA (Planning):
{'plan': 'Determine whether the cheesecake is labeled as Washington style by analyzing the information in the image.','tool_calls': [{'name': 'GetBarInfo', 'args': {}}]}
Tool Response:
{'error': 'No bar chart detected in the image.'}
Response by EVA (Re-Planning):
{'plan': 'The image primarily contains textual information rather than chart elements. Use OCR to extract the text describing the cheesecake style. The text should be analyzed to determine if it's a New York style cheesecake or not.', 'tool_calls': [{'name': 'OCR', 'args': {}}]}
Tool Response:
{'label': 'Style', 'confidence': 0.97, 'pixel_bbox': {'x_min': 350, 'y_min': 6, 'x_max': 431, 'y_max': 52}}, {'label': 'Cheesecake', 'confidence': 0.98, 'pixel_bbox': {'x_min': 103, 'y_min': 26, 'x_max': 438, 'y_max': 133}}, {'label': 'Washington', 'confidence': 1.0, …
Response by EVA (Answer Synthesis):
{'explanation': "The text in the image explicitly mentions 'New York Style Cheesecake'.", 'final_answer': 'no'}
Response by EVA (Consistency Verification):
{'explanation': "The answer is consistent with the tool result'.", 'final_answer': 'no'}

Figure 10: EVA recovers from incorrect tool invocation through Iterative Tool Replanning. The model initially applies an unsuitable chart-parsing tool to a text-centric image. After observing the failed tool response, EVA replans the execution and switches to OCR-based evidence extraction.

NeurIPS Paper Checklist

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

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

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

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

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

IMPORTANT, please:

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

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly describe the motivation, proposed method (EVA), key components (*Iterative Tool Replanning*, *Evidence Consistency Verification*, and *Selective Evidence Routing*), and the main experimental findings, which are consistent with the paper's scope and contributions.

   Guidelines:

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

2. **Limitations**

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

Answer: [Yes]

Justification: The paper discusses limitations, including dependence on external tool quality, and potential latency introduced by replanning and verification. Computational efficiency and trade-offs are also analyzed.

Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: The paper does not present formal theoretical theorems or proofs, focusing instead on empirical system design and experimental evaluation.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

Justification: The paper describes the EVA framework, tool execution pipeline, evaluation benchmarks, and implementation details necessary to reproduce the main experimental results.

Guidelines:

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

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

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

Answer: [No]

Justification: The current submission does not yet provide public access to the implementation code or released evaluation scripts, although reproducibility details are described in the paper.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper specifies datasets, evaluation settings, external tools, prompting strategies, and other implementation details required to understand the experiments and reproduce the reported results.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: The paper reports evaluation metrics across benchmarks but does not include statistical significance tests or error bars, as experiments primarily evaluate deterministic inference pipelines on standard benchmarks.

   Guidelines:

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

8. **Experiments compute resources**

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

Answer: [Yes]

Justification: The paper provides information about the computational environment, including the use of GPU-based inference and external visual tools, along with discussions of efficiency and selective tool invocation.

Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: The research complies with the NeurIPS Code of Ethics and does not involve unethical data collection, harmful experimentation, or violations of responsible AI research practices.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: The paper discusses both positive impacts, such as improving reliability and trustworthiness of multimodal AI systems, and potential risks including misuse of automated reasoning systems.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release high-risk generative models or sensitive datasets that require additional safeguards beyond standard responsible research practices.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Existing datasets, models, and external tools used in the experiments are properly cited, and their original sources and usage conditions are acknowledged in the paper.

    Guidelines:

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

13. **New assets**

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

Answer: [No]

Justification: The paper does not currently release new datasets, benchmarks, or public code artifacts alongside the submission

Guidelines:

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

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

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

Answer: [No]

Justification: The work does not involve crowdsourcing experiments or studies with human participants.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The paper does not involve human subject research requiring IRB approval or equivalent ethical review.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: Large language models (LLMs) were used as general-purpose tools in this work. Specifically, LLMs assisted in polishing the writing to improve clarity and readability.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.