How Much Vision Does Multimodal Reasoning Need? Vision-Stripping for Multimodal Benchmarks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multimodal reasoning benchmarks are usually reported as a single accuracy number, but that number can mix language priors, visual summaries, OCR, structured visual evidence, VLM captions, question-guided perception, tool use, and end-to-end multimodal fusion. We ask a complementary system-side question: on the original benchmark input and with a fixed final text-only reasoner, how does answerability change as image-derived evidence reaches the solver through different evidence paths? We introduce the **Vision-Stripping Test** (**VST**), a framework built on a same-input protocol that keeps the benchmark image, question, options, grader, and final text-only reasoner fixed while varying only the evidence path. The resulting visual-evidence answerability profile shows that benchmarks differ not only in difficulty, but in the visual evidence paths they reward: some retain strong text-only answerability, some become largely answerable from task-agnostic VLM captions, and others depend on structured visual text. Across six evaluated subsets, task-agnostic VLM captions and structured deterministic tools yield large average gains over the no-image anchor, plain OCR alone adds no average gain, and question guidance helps selectively rather than monotonically. As the profile endpoint, VST-Full is the highest-scoring VST evidence path on all six evaluated subsets. We further report a first-success regime partition u_p that, by construction, isolates the strict-marginal contribution of image-derived evidence: agentic acquisition contributes $u_{\text{Full}} = 16.0$ on OCR-BenchV2 and 10.0 on MMMU-Pro Vision, while $u_{\text{Full}} = 0$ on MMMU indicates the cumulative gap there is fully overlap with simpler paths. VST therefore reframes multimodal reasoning evaluation from a single endpoint score into a profile of which image-derived evidence paths make benchmark items answerable.

1 Introduction

A multimodal reasoning system cannot reason from evidence that never enters its context, yet this simple point is hidden when multimodal benchmarks are reported as a single accuracy number. The same item may be answered from the question and options alone; from a task-agnostic VLM caption; from OCR, table, formula, chart, or layout extraction; from question-guided VLM evidence; from agentic evidence acquisition; or from direct multimodal fusion in a monolithic VLM. Even when these routes yield the same final answer, they imply different conclusions about what kind of "vision" the benchmark rewards. A single score tells us whether a system answered correctly, but not which evidence path made the item answerable, nor what visual evidence was missing when the system failed.

This concern is not hypothetical. MMMU evaluates expert-level multimodal understanding Yue et al. [2024a], yet in our pilot runs a text-only reasoner with no image access reaches 46.2% on the MMMU validation split. A mirage-mode evaluation reports a related image-omitted effect on MMMU-Pro:

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

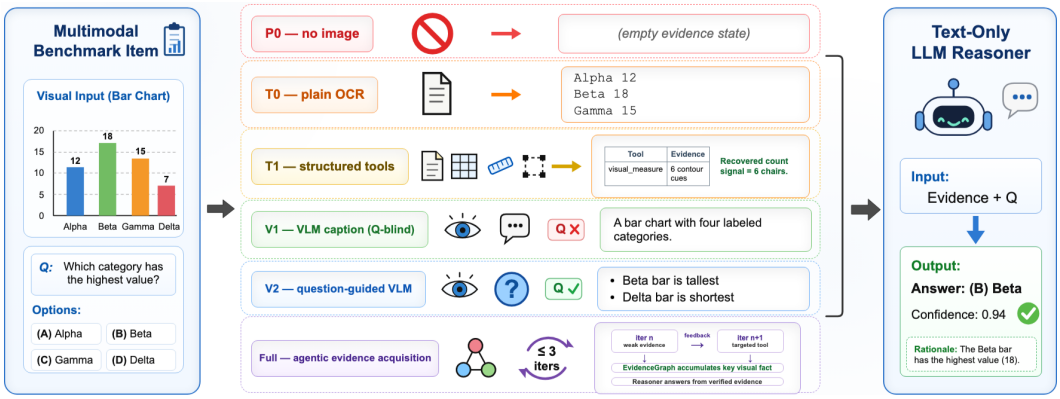


Figure 1: Overview of the **Vision-Stripping Test** (VST).

Gemini-2.5-Pro answers 858 of 1,730 questions correctly without image access Asadi et al. [2026]. MMMU-Pro filters text-only-solvable questions and introduces a Vision setting to reduce shortcut answerability Yue et al. [2024b]. Other diagnostic benchmarks sharpen this concern. VLind-Bench separates language priors, commonsense knowledge, and visual perception, showing that language-prior reliance remains a substantial failure mode in LVLMs Lee et al. [2024]. CrossMath constructs text-only, image-only, and image+text versions with matched task-relevant information, exposing that VLMs may reason better from text than from equivalent visual inputs Xu et al. [2026]. These works show that visual access and language priors cannot be taken for granted. We therefore turn to the evidence path inside the solver: for the original benchmark item and a fixed final text-only reasoner, how does answerability change as image-derived evidence is made available through different paths before final reasoning?

We introduce the **Vision-Stripping Test** (**VST**), a same-input protocol that constructs a visual-evidence answerability profile by holding the benchmark item, grader, and final text-only reasoner fixed and varying only the evidence path. The six VST paths span a no-image anchor (P0), deterministic tools (T0/T1), one-shot VLM evidence with and without question guidance (V1/V2), and an agentic tool-feedback acquisition endpoint (VST-Full); a direct VLM reference (D1) is reported alongside but is not part of the profile. Together they separate the major interfaces through which an image can become usable evidence for the same text-only solver (Section 3, Table 1).

Our experiments show that visual-evidence answerability profiles are benchmark-specific and non-monotonic, and that **VST-Full** is the highest-scoring VST evidence path on all six evaluated subsets. Because cumulative scores conflate evidence access with language-model capacity, we additionally report a first-success regime partition u_p, the share of items *first* answered correctly under each path; since P0 already uses the final Qwen3-8B reasoner without image access, u_p for paths beyond P0 isolates the contribution of image-derived evidence. The partition cleanly separates evidence-side and reasoning-side bottlenecks: $u_{\text{Full}} = 16.0$ on OCR-BenchV2 and 10.0 on MMMU-Pro Vision identify items only agentic acquisition exposes, while $u_{\text{Full}} = 0$ on MMMU indicates the cumulative gap there is fully overlap with simpler paths (Section 4).

Our contributions are:

- We introduce the **Vision-Stripping Test** (**VST**), a framework that defines a same-input, fixed-reasoner protocol for measuring benchmark answerability under different image-derived evidence paths.
- We instantiate VST with controlled paths spanning no-image answerability (P0), plain OCR (T0), structured deterministic tools (T1), task-agnostic VLM captions (V1), question-guided VLM evidence (V2), agentic tool-feedback evidence acquisition (VST-Full), and direct VLM answering (D1). The paired and endpoint contrasts measure the effects of adding structured deterministic evidence, question guidance, and agentic evidence acquisition.
- We define a first-success regime partition u_p that, by construction, controls for language-model capacity and reports the strict marginal answerability exposed by each path. The

resulting benchmark-specific profiles, summarized in Section 5, show where each evidence interface contributes and where it does not.

2 Related Work

Expert multimodal reasoning benchmarks. We evaluate on benchmarks that target expert-level multimodal reasoning over diagrams, documents, charts, tables, and visual text: MMMU, MMMU-Pro, MIA-Bench, ChartQA, MathVision, MathVerse, and OCR-BenchV2 Yue et al. [2024a,b], Qian et al. [2024], Masry et al. [2022], Wang et al. [2024], Zhang et al. [2024], Fu et al. [2024]. MMMU-Pro additionally filters text-only shortcuts and adds a Vision setting to stress visual dependence. We do not introduce a new benchmark; we ask which image-derived evidence paths make items from these benchmarks answerable at inference time.

Multimodal evaluation and profiling. Most evaluation suites report scalar endpoint performance Duan et al. [2024]. Diagnostic datasets such as MMStar, MathVerse, and CrossMath show that endpoint scores can overstate visual dependence by exposing leakage, text-dominant variants, or matched text/image versions Chen et al. [2024], Zhang et al. [2024], Xu et al. [2026], and MIRAGE finds that frontier models retain surprisingly high accuracy under image-omission Asadi et al. [2026]. VST uses this no-image setting as an anchor (P0) but keeps the original item fixed and intervenes on the solver's inference-time evidence path to profile how answerability changes as specific forms of visual evidence become available.

Textualization and structured visual evidence. Images can be converted into language for a text-only reasoner through captions Alayrac et al. [2022], Li et al. [2023], Khademi et al. [2023], OCR and scene-text reading Singh et al. [2019], Hegde et al. [2023], or structured summaries such as tables, key-value fields, and evidence graphs Cheng et al. [2024], Sun et al. [2023]. VST turns these representation choices into profiling controls: OCR-derived, caption-derived, structured, and local-perception evidence are measured as separate paths rather than treated as one visual-to-text bottleneck.

Tool-using multimodal systems. Tool-using and visual chain-of-thought systems call external tools, search over regions, and refine evidence over multiple steps Yao et al. [2023], Shinn et al. [2023], Wei et al. [2022], Wu et al. [2023], Wu and Xie [2024], while monolithic VLMs process image and text jointly Liu et al. [2024], Bai et al. [2023], Achiam et al. [2023]. Endpoint accuracy alone does not reveal whether success came from language priors, OCR, localization, tool routing, or fusion. VST is complementary to both paradigms: it reports direct VLM baselines alongside controlled decomposed evidence paths, making the workflow itself a measurement instrument for which image-derived evidence becomes sufficient.

3 Method

We describe the Vision-Stripping Test (VST), a measurement framework built on a same-input protocol for studying how image-derived evidence becomes available to a fixed text-only reasoner. VST keeps the benchmark image, question, answer options, grader, and final reasoner fixed, and varies the evidence path through P0/T0/T1/V1/V2/VST-Full. The resulting accuracies form a visual-evidence answerability profile spanning no-image answerability, plain OCR, structured deterministic extraction, one-shot VLM evidence, and agentic evidence acquisition. D1 is reported separately as a direct monolithic VLM reference. The V2 and VST-Full paths form a combined endpoint contrast from question-guided VLM evidence to tool-feedback evidence acquisition before final text-only reasoning.

3.1 Problem Setup

Each benchmark example consists of an image I, a question Q, answer options O, and a gold answer $A^\star$. A system returns an answer A, which is evaluated by the original benchmark grader. Standard VLM evaluation reports accuracy for a complete multimodal system. VST instead measures a family of accuracies indexed by evidence path:

$$\mathrm{Acc}(p) = \mathbb{E}_{(I,Q,O)}[\mathbf{1}\{f_p(I,Q,O) = A^\star\}], \tag{1}$$

Table 1: Evidence paths used in VST, with D1 included as a direct monolithic answering reference.

Path	Evidence source	VLM sees image?	VLM sees Q/O?	Tool use?	Reasoner input	Profiling role
P0	no image	no	no	no	empty evidence state	question/options answerability
T0	plain OCR tools	no	no	yes	reading-order OCR text	visual text without layout
T1	structured deterministic tools	no	no	yes	OCR + layout/table/formula/CV facts	value of structure over OCR
V1	task-agnostic VLM caption	yes	no	no	global caption	generic visual summary
V2	question-guided VLM evidence	yes	yes	no	targeted visual facts/caption	effect of question conditioning
VST-Full	agentic evidence acquisition	yes	yes	yes	compact evidence graph	adaptive tool-feedback evidence
D1	direct VLM	yes	yes	no	direct answer	monolithic reference

where p denotes the VST evidence path used inside the inference system. We interpret $\mathrm{Acc}(p)$ as the answerability exposed by path p for the fixed final reasoner and benchmark grader. All VST paths receive the same benchmark input and are scored by the same grader. They differ only in how image-derived evidence enters the reasoning context.

3.2 VST Evidence Paths

Table 1 defines the evidence paths reported by VST. Each path specifies how image-derived information may enter the inference system before the same text-only reasoner answers, ranging from the no-image anchor P0, through deterministic tool translators (T0 plain OCR; T1 adding layout, table, formula, key-value, region OCR, text grounding, and visual measurement) and one-shot VLM evidence with and without question guidance (V1/V2), to the agentic evidence-acquisition loop VST-Full, which combines the deterministic tool family with question-guided VLM evidence under provenance and quality metadata. D1 is reported alongside as a direct monolithic VLM reference. The **Tool use?** column refers to explicit external evidence tools or multi-tool acquisition; V1, V2, and direct answering use the VLM itself as the visual interface.

Why these paths? These paths instantiate a compact set of visual-to-text handoffs for a language reasoner. P0 matches the image-omitted setting used by recent mirage-style diagnostics Asadi et al. [2026] and anchors the profile at zero image-derived evidence. T0/T1 and V1/V2 then separate two common ways that systems convert an image into text usable by a language reasoner: deterministic specialist tools and VLM-mediated evidence. T0 versus T1 compares plain OCR text with structured deterministic evidence, holding the final reasoner and budget bookkeeping fixed. V1 versus V2 tests question guidance under a matched one-pass VLM budget: the same 3B VLM either produces a task-agnostic caption without seeing the question or produces an answer-forbidden question-guided caption after seeing the question and options. Intermediate prompts or budgets would form a continuum; these endpoints give reproducible paired contrasts. V2 versus VST-Full then moves from a flat question-guided evidence string to agentic evidence acquisition: the Translator VLM can use tool-result feedback to update an evidence graph before committing the final visual evidence state.

Evidence-path contrasts. Let $a_p = \mathrm{Acc}(p)$ be the accuracy under evidence path p. We report endpoint accuracy, paired contrasts for one-step evidence paths, and the VST-Full endpoint contrast:

$$\Delta_{\mathrm{VLMcap}} = a_{\mathrm{V1}} - a_{\mathrm{P0}}, \tag{2}$$

$$\Delta_{\mathrm{ocr}} = a_{\mathrm{T0}} - a_{\mathrm{P0}}, \tag{3}$$

$$\Delta_{\mathrm{struct}} = a_{\mathrm{T1}} - a_{\mathrm{T0}}, \tag{4}$$

$$\Delta_{\mathrm{qVLM}} = a_{\mathrm{V2}} - a_{\mathrm{V1}}, \tag{5}$$

$$\Delta_{\mathrm{agent}} = a_{\text{VST-Full}} - a_{\mathrm{V2}}, \tag{6}$$

$$\Delta_{\mathrm{endpoint}} = a_{\text{VST-Full}} - \max(a_{\mathrm{T0}}, a_{\mathrm{T1}}, a_{\mathrm{V1}}, a_{\mathrm{V2}}). \tag{7}$$

Cumulative path accuracies a_p overlap: an item solved by both T1 and V1 contributes to both columns. To measure the *strict marginal* answerability exposed by each path, we additionally partition each benchmark by the first path in the chain $\mathrm{P0} \to \mathrm{T0} \to \mathrm{T1} \to \mathrm{V1} \to \mathrm{V2} \to$ VST-Full that answers the item correctly. For path p at chain position k,

$$u_p = \Pr\big[\, f_p(I,Q,O) = A^\star \ \wedge\ \forall\, q{<}k:\ f_q(I,Q,O) \neq A^\star \big], \tag{8}$$

i.e., the share of items *first* answered correctly under path p. The residual $u_{\mathrm{unsolved}} = 1 - \sum_p u_p$ is the share answered by no path. Because P0 already uses the same fixed text-only reasoner as VST-Full, u_p for $p \succ \mathrm{P0}$ isolates the contribution of image-derived evidence from language-model capacity. The

contrasts Δ are signed and need not be monotonic: adding a particular evidence interface can help, leave answerability unchanged, or introduce evidence that is incomplete, noisy, or poorly matched to the item. The T0 $\rightarrow$ T1 contrast measures the value of adding layout, table, formula, and region structure on top of plain OCR. The V1 $\rightarrow$ V2 contrast measures the effect of question conditioning in a single VLM pass. The V2 $\rightarrow$ VST-Full endpoint contrast measures the combined effect of replacing flat question-guided VLM evidence with a provenance- and quality-annotated evidence graph plus tool-feedback acquisition before final text-only reasoning. The endpoint contrast summarizes the margin between VST-Full and the strongest non-agentic VST evidence path.

3.3 Measurement Instrument: Controllable Evidence Workflow

To instantiate these paths, we use a two-stage evidence workflow with path-specific access rules. A visual *evidence extractor* gathers evidence and writes it into a visual evidence state; a text-only *reasoner* consumes that state together with the question and options to produce the answer.

Visual evidence state. The visual evidence state is a textual, provenance-tracked representation containing tool outputs, extracted claims, locations, and confidence scores. The primary runs use a *neutral* visual evidence state: upstream visual components may extract text, tables, formulas, regions, or local captions, but they do not add option-level support or contradiction labels before final reasoning. The text-only reasoner receives the rendered visual evidence state, the question, and the options, but no raw image tokens. This makes the evidence interface explicit: each VST path is represented by the textual evidence state available to the final solver.

Tool families and algorithm. The workflow uses deterministic non-VLM tools, one-shot VLM evidence, and agentic evidence acquisition under different access rules. T0/T1 enable deterministic tool subsets, V1 uses a single Qwen2.5-VL-3B caption pass, V2 uses one question-guided Qwen2.5-VL-3B evidence pass, and VST-Full enables the agentic evidence-acquisition loop. Appendix A.1 gives the full tool-family inventory and the VST-Full evidence-construction algorithm.

VST-Full implementation. VST-Full is the adaptive VST path. It uses Qwen2.5-VL-3B-Instruct for the Translator/evidence tools and Qwen3-8B for final text-only reasoning. The Translator maintains a compact evidence graph and, for up to three Translator/tool steps ($N_T = 3$), either routes one visual tool, observes its result, or commits a sufficient visual evidence summary. Only after this acquisition loop terminates is the final text-only reasoner invoked on the question, options, and compact evidence graph; the reasoner does not open the next visual acquisition pass. Appendix A.1 gives the full evidence-construction algorithm, and Appendix A.2 gives the evidence schema, prompts, and quality guards.

Budget matching. For T0/T1, we match tool-call limits, retained-evidence limits, serialization, final reasoner, and answer extraction. T0 versus T1 changes the deterministic tool set. V1 versus V2 matches the single small-VLM pass and output budget, while V2 additionally exposes the question/options. V2 versus VST-Full adds the evidence-graph/tool interface and an agentic acquisition budget.

Reasoning. Across T0/T1/V1/V2/VST-Full, the final answer is produced by the same Qwen3-8B text-only reasoner with the same extraction and grading pipeline. P0 uses the same reasoner with an empty visual evidence state, keeping the reasoning substrate fixed while only the evidence path changes.

4 Experiments

We evaluate VST through three questions:

1. How much answerability is exposed when the fixed final reasoner receives no image-derived evidence?
2. How does answerability change as image-derived evidence enters through plain OCR, structured deterministic tools, one-shot VLM textualization, question-guided VLM evidence, and tool-feedback evidence acquisition (VST-Full)?

Table 2: Main Vision-Stripping matrix. Scores are reported on a 0–100 scale under the benchmark grader, with bracketed uncertainty intervals. MIA-Bench and OCR-BenchV2 use benchmark scores scaled by 100; the other entries are accuracies. D1 is the direct Qwen2.5-VL-3B reference baseline, not part of the VST evidence-path profile. VST-Full is the tool-feedback evidence-acquisition path ($N_T \leq 3$). Subset sizes and sampling details are reported in Appendix A.3.

Benchmark	P0-text	T0-OCR	T1-structured tools	V1-VLM caption	V2-qVLM caption	VST-Full	D1
MMMU-Pro Vision	21.0 [14.17,29.98]	21.0 [14.17,29.98]	31.0 [22.78,40.63]	34.0 [25.46,43.72]	29.0 [21.02,38.54]	45.0 [35.61,54.76]	24.61 [19.16,31.02]
MMMU-Pro Standard	31.0 [22.78,40.63]	32.0 [23.67,41.66]	39.0 [30.02,48.80]	40.0 [30.94,49.80]	42.0 [32.80,51.79]	47.0 [37.51,56.71]	25.82 [20.25,32.30]
MMMU	46.2 [41.88,50.58]	44.0 [37.98,50.20]	50.2 [45.83,54.56]	55.4 [51.02,59.70]	51.11 [43.86,58.31]	63.0 [53.22,71.82]	48.33 [43.98,52.71]
MIA-Bench (score)	51.9 [45.27,58.40]	52.4 [46.07,58.93]	59.9 [54.13,65.53]	86.2 [82.80,89.40]	87.1 [83.67,90.20]	90.73 [87.47,93.60]	76.90 [72.53,80.76]
OCR-BenchV2 (score)	6.0 [2.78,12.48]	2.0 [0.55,7.00]	35.0 [28.79,41.81]	34.5 [28.32,41.30]	50.0 [43.20,56.91]	52.5 [45.65,59.37]	52.0 [45.87,58.45]
ChartQA	1.0 [0.18,5.45]	0.0 [0.00,3.70]	15.0 [9.31,23.28]	21.0 [14.17,29.98]	29.0 [21.02,38.54]	33.0 [24.56,42.70]	40.0 [30.94,49.80]

3. How do these evidence-availability profiles differ across benchmarks and relative to direct monolithic VLM references?

4.1 Benchmarks and Baselines

Benchmarks. The evaluation set contains six benchmarks. **MMMU-Pro Vision** is our primary benchmark because it is explicitly designed to reduce text-only shortcuts and emphasize visual dependence **?**. We also evaluate **MMMU-Pro Standard** and the **MMMU validation** split for historical comparability and as a no-image text-only anchor Yue et al. [2024a]. **MIA-Bench** provides an out-of-family multi-image and instruction-following benchmark for testing how evidence-availability profiles differ beyond the MMMU family Qian et al. [2024]. **ChartQA val** contributes chart and table reasoning examples where structured evidence should be especially useful Masry et al. [2022]. **OCR-BenchV2** contributes OCR-centric examples that stress visual text extraction and grounding **?**. We report all entries on a 0–100 scale: MIA-Bench and OCR-BenchV2 use benchmark scores scaled by 100, while the remaining benchmarks report accuracy.

Systems. All VST evidence paths share the same Qwen3-8B text-only reasoner; V1, V2, and VST-Full use Qwen2.5-VL-3B for VLM-based evidence components, keeping the small visual backbone fixed while tool access and evidence budgets vary by path. Path definitions are given in Table 1 and Section 3.3.

Direct references. D1 is direct Qwen2.5-VL-3B image-question answering under the same benchmark grader. It is included as a monolithic VLM reference, not as part of the VST evidence-path profile. The comparison to D1 is therefore an evidence-interface comparison against the same small visual backbone, not a compute-matched claim against larger multimodal systems. Table 4 additionally reports larger direct-answer references where available, including Qwen2.5-VL-7B, Qwen2.5-VL-32B, LLaVA-1.5-7B, and GPT-4o-mini Bai et al. [2023], Liu et al. [2024], Achiam et al. [2023].

4.2 Main Vision-Stripping Matrix

Table 2 is the main Vision-Stripping result matrix. The columns correspond to the evidence path: P0 measures no-image text-only answerability, T0 isolates plain OCR text without layout, T1 tests structured deterministic tools, V1/V2 measure one-shot VLM evidence paths with and without question conditioning, VST-Full reports tool-feedback evidence acquisition with an evidence graph and tool-result feedback, and D1 is a direct-answer reference baseline.

Macro-average evidence-path contrasts (Δ_{ocr}, Δ_{struct}, Δ_{qVLM}, Δ_{agent}, etc.) across the six benchmarks are reported in Appendix A.5 (Table 7).

Legible evidence makes many items answerable. On MMMU-Pro Vision, P0 already solves 21% of the slice, showing that some items remain answerable from the question and options alone. Plain OCR (T0) does not improve this score, while structured deterministic evidence (T1) raises accuracy to 31% and a task-agnostic VLM caption (V1) reaches 34%. Across the matrix, the same pattern recurs in sharper form: plain OCR can add little or even hurt, while structured deterministic evidence creates large gains on OCR-BenchV2 (2.0 $\rightarrow$ 35.0) and ChartQA (0% $\rightarrow$ 15%). These contrasts show why VST separates plain text extraction from structured visual evidence.

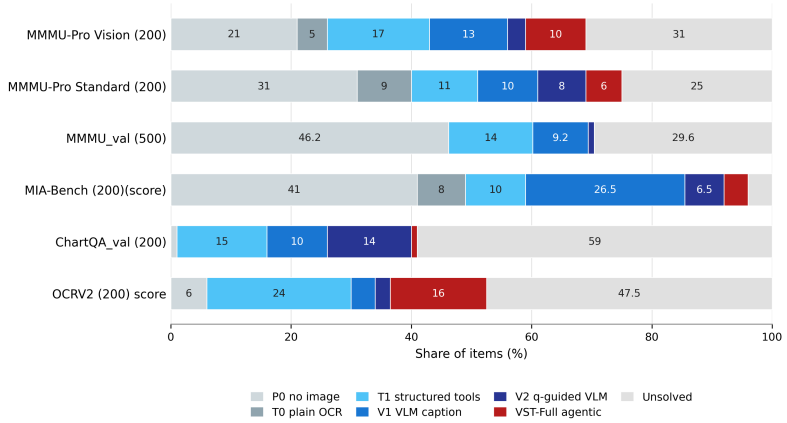


Figure 2: First-success regime partition.

Table 3: First-success regime partition u_p (%). For each benchmark the row sums to 100.

Benchmark	P0	T0	T1	V1	V2	VST-Full	Unsolved
MMMU-Pro Vision	21.0	5.0	17.0	13.0	3.0	10.0	31.0
MMMU-Pro Standard	31.0	9.0	11.0	10.0	8.0	6.0	25.0
MMMU	46.2	0.0	14.0	9.2	1.0	0.0	29.6
MIA-Bench (score)	41.0	8.0	10.0	26.5	6.5	4.0	4.0
OCR-BenchV2 (score)	6.0	0.0	24.0	4.0	2.5	16.0	47.5
ChartQA	1.0	0.0	15.0	10.0	14.0	1.0	59.0

Cumulative scores overlap; the strict marginal view follows in Section 4.3. Cumulative path accuracies a_p count the same item every time it is solved, so they do not distinguish "new evidence solved this item" from "a simpler path already had it." Section 4.3 reports the first-success regime partition u_p defined in Eq. equation 8, which assigns each item to the earliest path in the chain that answers it correctly and is therefore comparable across paths and benchmarks. Appendix A.4 provides qualitative adjacent-path examples illustrating how individual items change when the evidence interface changes.

4.3 First-Success Regime Partition

Figure 2 and Table 3 report the visual-evidence answerability profile through the first-success regime partition. Each item is assigned to the earliest path $\mathrm{P0} \to \mathrm{T0} \to \mathrm{T1} \to \mathrm{V1} \to \mathrm{V2} \to \text{VST-Full}$ that answers it correctly under the same fixed Qwen3-8B reasoner; items not solved by any path are reported as Unsolved. The partition is exclusive: the columns sum to 100. The non-Unsolved share, $1 - u_{\mathrm{unsolved}}$, is the *evidence ceiling* reachable by the protocol on this benchmark.

The partition reshapes the VST-Full claim. In Table 2, VST-Full has the highest cumulative score on five of six subsets. The partition shows the share of items for which VST-Full is *strictly* required, in the sense that no earlier path in the chain answers them: $u_{\mathrm{Full}} = 16.0$ on OCR-BenchV2 and $u_{\mathrm{Full}} = 10.0$ on MMMU-Pro Vision are the largest, while $u_{\mathrm{Full}} = 0.0$ on MMMU, 1.0 on ChartQA, 4.0 on MIA-Bench, and 6.0 on MMMU-Pro Standard are smaller. Because P0 uses the same Qwen3-8B reasoner, these strict-marginal shares cannot be explained by language-model capacity alone: they isolate items for which agentic evidence acquisition exposes answerability that the same reasoner cannot reach through any of the five simpler paths.

Question guidance and agentic acquisition contribute selectively. On MMMU, $u_{\mathrm{V2}} + u_{\mathrm{Full}} = 1.0$: the question-guided VLM path and the agentic loop together first-solve only 1% of items, and the 63.0 cumulative score of VST-Full in Table 2 is almost entirely overlap with simpler paths. On OCR-

Table 4: Endpoint and direct-answer reference results. Scores are on a 0–100 scale; MIA-Bench and OCR-BenchV2 use benchmark scores scaled by 100, while the other entries are accuracies. VST-Full uses Qwen2.5-VL-3B visual evidence components and Qwen3-8B final text-only reasoning.

Benchmark	VST-Full	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Qwen2.5-VL-32B	LLaVA-1.5-7B	GPT-4o-mini
MMMU-Pro Vision (200)	**45.00**	24.61	23.26	28.77	12.13	31.72
MMMU-Pro Standard (10 option, 200)	**47.00**	25.82	23.60	32.93	14.21	38.99
MMMU (500)	**63.00**	48.33	51.11	51.56	32.47	59.40
OCR-BenchV2 (score, 200)	52.50	52.00	53.96	52.22	19.21	**55.08**
MIA-Bench (score, 200)	**90.73**	76.90	79.90	89.60	69.80	88.58

BenchV2 and MMMU-Pro Vision, by contrast, $u_{\mathrm{T1}} + u_{\mathrm{Full}} = 40.0$ and 27.0 respectively, locating most of the additional answerability in structured deterministic tools and the agentic endpoint. On MIA-Bench, $u_{\mathrm{P0}} + u_{\mathrm{V1}} = 67.5$, consistent with a benchmark that is largely answerable from question/options or a generic visual summary.

Unsolved residuals mark the protocol ceiling. ChartQA leaves 59.0% of items Unsolved and OCR-BenchV2 leaves 47.5%; these are the benchmarks where the current VST workflow reaches its ceiling furthest below 100. MMMU and MMMU-Pro Standard leave 29.6% and 25.0%, while MIA-Bench leaves only 4.0%. Unsolved residuals are not a failure mode of the partition; they are the measurement: they identify items for which none of the six interfaces—no-image priors, plain OCR, structured deterministic tools, generic captions, question-guided VLM captions, or agentic evidence acquisition—deliver evidence sufficient for the fixed text-only reasoner under the configured budgets.

4.4 Agentic Evidence Endpoint Results

Table 4 places the VST-Full evidence path alongside direct-answer VLM references at several model scales. Table 2 shows how answerability changes across VST evidence paths, while Table 4 summarizes the endpoint comparison to common monolithic references.

The VST-Full endpoint stress-tests the evidence handoff exposed by VST. With the same fixed Qwen3-8B text reasoner and a small Qwen2.5-VL-3B visual evidence backbone, the cumulative VST-Full score in Table 4 exceeds the direct Qwen2.5-VL-3B reference on five of six subsets, with the largest gaps on MMMU-Pro Standard (47.00 vs. 25.82), MMMU-Pro Vision (45.00 vs. 24.61), and MMMU validation (63.00 vs. 48.33). The cumulative comparison is confounded with reasoner capacity, since the Qwen2.5-VL-3B reference uses a smaller language backbone than the Qwen3-8B reasoner used by every VST path. We therefore read this table together with the partition in Table 3: the strict-marginal share u_{Full} controls for reasoner capacity by construction, because $\mathrm{P0}$ already uses Qwen3-8B without image access. Under that lens, VST-Full's largest evidence-side contributions are on OCR-BenchV2 ($u_{\mathrm{Full}} = 16.0$) and MMMU-Pro Vision ($u_{\mathrm{Full}} = 10.0$); on MMMU $u_{\mathrm{Full}} = 0.0$, indicating that the cumulative gap on this benchmark is fully attributable to overlap with simpler paths, not to items that only agentic acquisition exposes. Cumulative endpoint scores and the regime partition are therefore complementary: the matrix shows which simpler interface already approaches the endpoint, while the partition shows where adaptive evidence accumulation exposes *strictly* new answerability.

5 Discussion

Findings. The VST profile separates four observations that an endpoint score collapses. (i) Substantial no-image answerability persists on MMMU validation and MMMU-Pro Standard ($a_{\mathrm{P0}} = 46.2$ and 31.0). (ii) Plain OCR alone contributes essentially nothing on average ($\Delta_{\mathrm{ocr}} = -1.0$ macro), while structured deterministic tools deliver $+13.1$ macro and large per-benchmark gains on layout-heavy items (T0→T1: $+33.0$ on OCR-BenchV2, $+15.0$ on ChartQA). (iii) Global VLM captions already approach the protocol ceiling on instruction-following ($u_{\mathrm{P0}} + u_{\mathrm{V1}} = 67.5$ on MIA-Bench) but not on visual-text benchmarks. (iv) Question-guided VLM captioning is non-monotonic with respect to the task-agnostic caption (V2 below V1 on MMMU-Pro Vision; V2 above V1 on OCR-BenchV2 and ChartQA), confirming that the relevant variable is the evidence delivered to the reasoner, not whether the upstream VLM saw the question. Each is a reproducible paired contrast under matched evidence budgets.

When agentic acquisition adds no strict-marginal answerability. The cleanest finding from the regime partition is that agentic acquisition does not always help. On MMMU, $u_{\text{Full}} = 0$: every item VST-Full answers is already first-solved by a simpler path. Combined with $u_{\text{V2}} = 1.0$ on the same benchmark, the implication is that for items beyond the reach of simpler paths—the 29.6% Unsolved on MMMU and the 25.0% Unsolved on MMMU-Pro Standard, where u_{Full} is also small (6.0)—the bottleneck is not "more visual evidence" but something the protocol does not address: domain knowledge, multi-step reasoning, or numeric/symbolic computation. The asymmetry suggests a useful diagnostic. Where agentic acquisition does help in the strict-marginal sense, the contributions concentrate on layout- and perception-heavy benchmarks ($u_{\text{Full}} = 16.0$ on OCR-BenchV2, 10.0 on MMMU-Pro Vision); where $u_{\text{Full}} \approx 0$, additional tool budget yields no language-model-controlled gain, and compute should be redirected toward reasoning capacity rather than evidence acquisition.

Reading the Unsolved residual. The Unsolved column $1 - \sum_p u_p$ marks items that no VST path answers under the configured budgets. ChartQA leaves 59.0% Unsolved and OCR-BenchV2 leaves 47.5%, the largest residuals in Table 3. These residuals are not artifacts of weak tools—the cumulative VST-Full scores on the same items reach 33.0 and 52.5, at or near the per-benchmark ceiling (41.0 and 52.5)—but a measurement: items that resist every interface tested. On chart and OCR-heavy benchmarks the residual likely concentrates around fine-grained numerical reading and dense layout dependencies that exceed the 3B visual backbone's per-item evidence quality; on the MMMU family the residual is plausibly dominated by domain reasoning rather than evidence access (cf. the $u_{\text{Full}} = 0$ result on MMMU). The Unsolved column lets the protocol report what it cannot answer instead of confounding it with what it can.

Workflow dependence. The VST profile is measured through one controllable evidence workflow, not through all possible decomposed systems. The profile therefore characterizes benchmark dependence as exposed by this measurement instrument. If the deterministic tools, acquisition policy, evidence serialization, or final reasoner change, the absolute scores may change. The more stable claim is the pattern of paired contrasts under matched budgets, especially whether deterministic tool translation, one-shot VLM evidence, and tool-feedback evidence acquisition remain separated.

Scope. Our framework is aimed at reasoning-centric, document-like, chart-like, diagram-like, and evidence-heavy multimodal tasks. It does not claim that all visual tasks can be reduced to structured text. Low-level perception, dense correspondence, fine-grained recognition, medical imaging, and tasks requiring precise continuous geometry may exhibit low answerability under T0/T1/V1/V2 and large residual gaps even under **VST-Full**. The protocol is designed so that those residuals are observable, not hidden.

Cost and fairness. u_{Full} controls for language-model capacity but not for token cost. On the four benchmarks for which we logged per-path tokens (Appendix A.7), VST-Full uses 3.8–$22\times$ more total tokens than the direct Qwen2.5-VL-3B reference. Paired with the partition, this gives a cost diagnostic: on MMMU ($u_{\text{Full}} = 0$) the extra tokens are overhead, while on OCR-BenchV2 ($u_{\text{Full}} = 16.0$) they buy answerability no simpler path reaches. Wall-clock and USD depend on hardware and provider; tokens are the reproducible cross-system unit.

6 Conclusion

We introduced the Vision-Stripping Test (VST), a same-input, fixed-reasoner framework for measuring benchmark answerability under explicit image-derived evidence paths. Beyond cumulative path accuracy, we defined a first-success regime partition u_p that, by construction, controls for language-model capacity (the no-image anchor uses the same final reasoner) and reports the strict marginal answerability exposed by each path. Read together, the cumulative matrix and the partition turn "how visual is this benchmark?" into separable questions: which simpler evidence interfaces already approach the protocol ceiling, where the agentic endpoint contributes strict-marginal answerability ($u_{\text{Full}} = 16$ on OCR-BenchV2 and 10 on MMMU-Pro Vision), where it does not ($u_{\text{Full}} = 0$ on MMMU validation), and which items resist every interface tested (the Unsolved residual, 59% on ChartQA and 47.5% on OCR-BenchV2). VST therefore shifts evaluation from a single endpoint score to a profile of which image-derived evidence paths make benchmark items answerable, and at what cost.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Fardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, 2024.

Kewei Cheng, Nesreen K Ahmed, Theodore Willke, and Yizhou Sun. Structure guided prompt: Instructing large language model in multi-step reasoning by exploring graph structure of the text. *arXiv preprint arXiv:2402.13415*, 2024.

Haodong Duan, Xinyu Fang, Junming Yang, Xiangyu Zhao, Yuxuan Qiao, Mo Li, Amit Agarwal, Zhe Chen, Lin Chen, Yuan Liu, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *arXiv preprint arXiv:2407.11691*, 2024.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2024.

Shamanthak Hegde, Soumya Jahagirdar, and Shankar Gangisetty. Making the v in text-vqa matter. *arXiv preprint arXiv:2308.00295*, 2023.

Mahmoud Khademi, Ziyi Yang, Felipe Frujeri, and Chenguang Zhu. Mm-reasoner: A multi-modal knowledge-aware framework for knowledge-based visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6571–6581, 2023.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. Vlind-bench: Measuring language priors in large vision-language models. *arXiv preprint arXiv:2406.08702*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. Mia-bench: Towards better instruction following evaluation of multimodal llms. *arXiv preprint arXiv:2407.01509*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *arXiv preprint arXiv:2307.07697*, 2023.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024.

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.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.

Yige Xu, Yongjie Wang, Zizhuo Wu, Kaisong Song, Jun Lin, and Zhiqi Shen. Do vision-language models truly perform vision reasoning? a rigorous study of the modality gap. *arXiv preprint arXiv:2604.16256*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024a.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024b.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, 2024.

A Appendix

A.1 VST Run Configuration

The deterministic-evidence subfamily contrast is T0 versus T1. Both paths use bounded evidence serialization and the same final reasoner. T0 restricts the action space to one plain-OCR call, while T1 enables up to three structured deterministic tool calls, including layout, table, formula, region, and measurement evidence. V1 and V2 both use one Qwen2.5-VL-3B pass: V1 is task-agnostic VLM captioning, while V2 is question-guided VLM caption/evidence generation; neither path calls external visual tools. VST-Full combines deterministic and VLM-based visual tools inside an agentic evidence-acquisition loop. In the reported runs, VST-Full allows up to three Translator/tool steps ($N_T = 3$) and stops earlier when the Translator commits a compact visual evidence summary. It uses Qwen2.5-VL-3B-Instruct for VLM-based evidence tools and Qwen3-8B for final reasoning.

Evidence interfaces by VST path. Table 5 lists the evidence interfaces exposed by each VST path.

Table 5: Visual evidence interfaces used by each VST path.

VST path	Evidence interface	Enabled tools
P0	no visual evidence	none
T0	plain OCR	`ocr`
T1	structured deterministic tools	OCR, layout OCR, page OCR, table extraction, region OCR, key-value extraction, formula OCR, visual measurement
V1	one-shot task-agnostic VLM textualization	none
V2	one-shot question-guided VLM textualization	none
VST-Full	agentic evidence acquisition	deterministic tools, VLM-based local evidence tools, append-only EvidenceGraph, terminate/evidence-summary action

Algorithm 1 VST-Full agentic evidence acquisition with up to $N_T = 3$ Translator/tool steps and adaptive early termination.

Require: image I, question Q, options O
Require: Translator/tool budget $N_T = 3$
$G_0 \leftarrow \emptyset$ ▷ append-only visual evidence graph
$G_\star \leftarrow G_0$
for $t = 1$ to N_T **do**
  $a_t \leftarrow \mathrm{TranslatorVLM}(I, Q, O, \mathrm{CompactView}(G_{t-1}))$ ▷ select a tool or commit the evidence state
  if $a_t = \mathrm{ToolCall}(u, \theta)$ **then**
    $y_t \leftarrow u(I, \theta)$
    $G_t \leftarrow \mathrm{AppendToolResult}(G_{t-1}, u, y_t)$
  else if $a_t = \mathrm{TerminateEvidence}(E)$ **then**
    $G_t \leftarrow \mathrm{AppendEvidenceSummary}(G_{t-1}, E)$
    $G_\star \leftarrow G_t$
    break
  else
    $G_t \leftarrow G_{t-1}$
  $G_\star \leftarrow G_t$
return $\mathrm{TextReasoner}(Q, O, \mathrm{CompactView}(G_\star))$

A.2 Visual Evidence State Schema and Prompt Templates

The visual evidence state is the only image-derived input visible to the text-only reasoner in T0, T1, V1, V2, and VST-Full. It is rendered as plain text before final reasoning, but VST-Full stores it internally as an append-only EvidenceGraph with provenance fields so that evidence can be audited and assigned to an evidence path.

Evidence-quality guards. VST-Full uses deterministic evidence-quality guards before rendering the compact evidence graph to the reasoner. These guards do not call an LLM, do not remove evidence, and do not assign option-level support or contradiction labels. They only attach quality metadata to each evidence node. Evidence is marked `failed` when a tool returns an execution error or no parseable output; `low` when an output is structurally noisy, duplicated, empty, or based on a visibly poor crop; `uncertain` when partial signal exists but should be corroborated; and `ok` otherwise. The compact evidence view keeps these labels so the text-only reasoner can treat noisy evidence cautiously without hiding it from the final decision.

A.3 Benchmark Subsets

The VST profiles are run on fixed subsets so that all paths and baselines are compared on identical examples. Table 6 lists the subset sizes used in the VST matrix.

A.4 Qualitative Pair Examples

Figures 8–10 show representative adjacent-path examples. In each example, the left VST path fails on the benchmark item while the right VST path answers correctly, illustrating the additional answerability exposed by the next evidence interface.

Visual Evidence State Schema

```
{
  "nodes": [
    {
      "id": "e1",
      "iteration": 1,
      "tool": "ocr | layout_text_ocr | table_extract_compact | ...",
      "claim": "bounded reasoner-facing evidence claim",
      "raw": "stored raw tool output when retained",
      "confidence": "unknown | low | mid | high",
      "region_id": "optional region/crop reference",
      "bbox": [x1, y1, x2, y2],
      "crop_path": "optional crop artifact path",
      "tags": ["ocr", "table", "formula", "region", "..."],
      "quality": "ok | uncertain | low",
      "quality_score": 1.0,
      "quality_reasons": ["short deterministic notes, e.g., empty_ocr,
    blurry_crop, duplicate"],
      "inner_vlm_usage": {"input_tokens": 0, "output_tokens": 0, "
    total_tokens": 0},
      "failed": false
    }
  ],
  "rendering": {
    "compact_view": "newest and most relevant nodes, max_nodes=8, max_chars
    =1800",
    "omitted_in_primary_runs": "option-level support/contradiction labels"
  },
  "metadata": {
    "vst_path": "P0 | T0 | T1 | V1 | V2 | VST-Full",
    "tool_counts": {},
    "vlm_tool_calls": 0,
    "non_vlm_tool_calls": 0
  }
}
```

Figure 3: Visual evidence state used to pass image-derived information to the text-only reasoner. VST-Full stores tool outputs as append-only EvidenceGraph nodes and renders a bounded compact view before final reasoning. The primary runs use neutral evidence items and omit option-level support or contradiction labels from the reasoner-visible rendering. The rendered text contains only the fields available in the current VST path.

Table 6: Benchmark subset sizes.

Benchmark	Size	Role
MMMU-Pro Vision	200	primary visual-dependence benchmark
MMMU-Pro Standard	200	ten-option robustness setting
MMMU validation	500	historical comparison and text-only diagnostic
MIA-Bench (score)	200	out-of-family multi-image / instruction-following check
ChartQA validation	200	chart and table structured-evidence benchmark
OCR-BenchV2	200	OCR-centric boundary case

Text-Only Final-Answer Prompt Template

```
You are answering a multiple-choice multimodal reasoning question.

You do not have access to the raw image. You may use only:
1. the question,
2. the answer options,
3. the visual evidence state rendered below.

If the evidence is insufficient, choose the best-supported option from the
available information. Do not invent visual facts that are absent from the
visual evidence state.

Question:
{question}

Options:
{options}

Visual evidence state:
{rendered_evidence_state}

Return only the final option letter.
```

Figure 4: Final-answer prompt used after rendering a fixed evidence state in P0, T0, T1, V1, V2, and VST-Full. P0 uses an empty visual evidence state; VST-Full renders the compact evidence graph produced by the acquisition steps. The text-only reasoner receives no raw image tokens.

VST-Full Acquisition Policy

```
At each acquisition step, inspect:
1. the image,
2. the question and answer options,
3. the current compact EvidenceGraph/SIR,
4. recent tool-result feedback when available.

If the current evidence appears incomplete, uncertain, or mismatched with the
question, call exactly one useful visual tool. Otherwise call
terminate_and_output_caption with a compact visual evidence summary.
Do not choose the final answer or add option-level support labels.
```

Figure 5: Translator-side acquisition policy used by VST-Full. The Translator VLM can inspect the question, image, current evidence graph, and tool-result feedback, then route another tool call when the existing evidence appears incomplete or mismatched.

Task-Agnostic VLM Caption Prompt Template

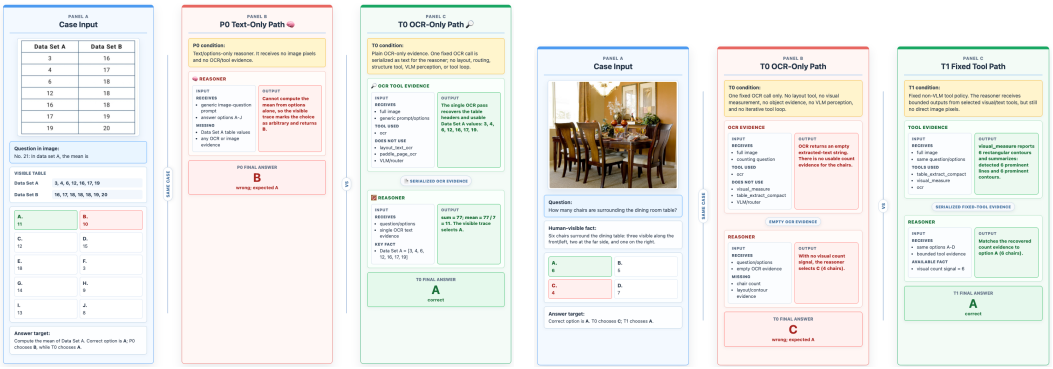
```
You are a visual captioner, not the final solver.

Produce a generic one-shot caption of the image as observable visual evidence.

Include visible text, objects, spatial layout, charts/tables/diagrams if
obvious, and major numbers/symbols.

Hard constraints:
- Do not answer the question or solve the task.
- Do not compare answer options, rank options, eliminate options, or say an
  option is supported.
- Do not mention option letters or answer choices unless they are visibly
  printed inside the image.
- Do not use domain knowledge or infer hidden facts beyond visible evidence.
- If a detail is unclear, mark it uncertain instead of guessing.

Keep it concise but complete; do not omit relevant visible evidence just to
    be brief.
```

Figure 6: V1 prompt. The 3B VLM receives the image and this task-agnostic caption prompt only; it does not receive the benchmark question or answer options.

Question-Guided VLM Caption Prompt Template

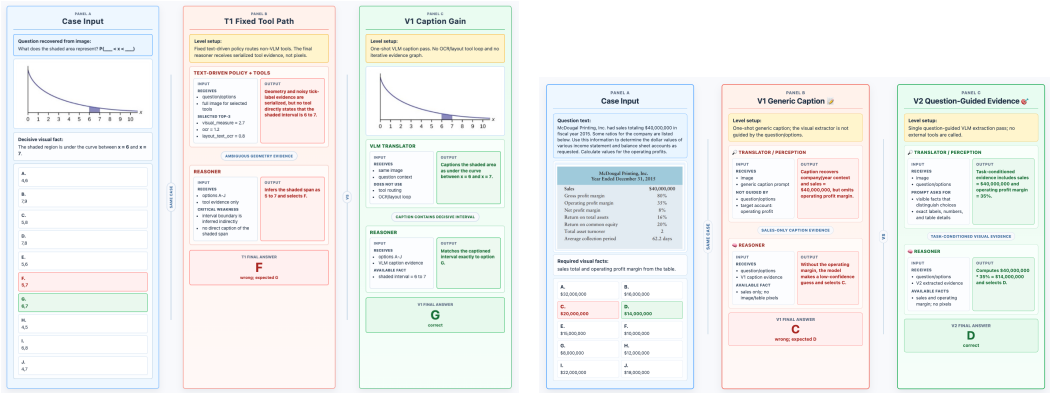
```
You are a compact visual evidence captioner, not the final solver.

Question:
{question}

Options:
{options}

Extract only visual facts needed to answer:
- exact visible labels, numbers, symbols, relations, positions, counts,
  table/chart/diagram details;
- mention uncertainty when a detail is unclear;
- do not choose an option, compare options, or state the final answer;
- do not use domain knowledge or infer hidden facts beyond visible evidence.

Keep it concise but complete.
```

Figure 7: V2 prompt. The same 3B VLM used for V1 receives the question, options, and image, then returns an answer-forbidden question-guided caption.



P0 → T0 **T0 → T1**

Figure 8: Qualitative adjacent-path examples where the right-hand evidence path succeeds and the left-hand path fails.



T1 → V1 **V1 → V2**

Figure 9: Qualitative adjacent-path examples where the right-hand evidence path succeeds and the left-hand path fails.

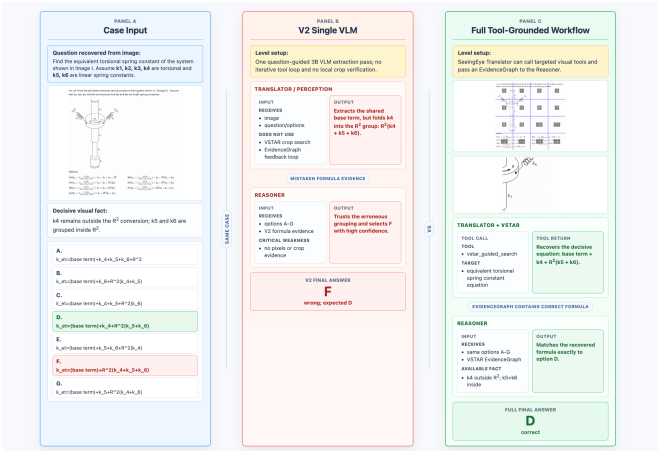


Figure 10: Qualitative adjacent-path example: V2 fails while VST-Full succeeds.

Table 7: Descriptive macro-average evidence-path contrasts across the six VST benchmarks in Table 2. MIA-Bench and OCR-BenchV2 contribute benchmark scores scaled to 0–100. T0 is the plain-OCR path, while T1 adds structured deterministic tools.

Comparison	**Avg. gain**
V1 VLM-caption − P0 text-only	+19.0
T1 structured deterministic evidence − P0 text-only	+12.2
T0 plain-OCR evidence − P0 text-only	-1.0
T1 structured deterministic evidence − T0 plain-OCR evidence	+13.1
V2 question-guided VLM evidence − V1 VLM-caption	+2.9
V2 question-guided VLM evidence − T1 structured deterministic evidence	+9.7
VST-Full tool-feedback acquisition − V2 question-guided VLM (Δ_{agent})	+7.2
VST-Full − best non-agentic VST path	+5.6

Table 8: Per-item average token cost (input + output, summed across all VLM and reasoner calls in the path). All numbers in tokens. The input/output split shows where the budget is spent: P0/T0/T1 spend most output on Qwen3-8B reasoning, while VST-Full spends a larger share of input on image-conditioned tool calls.

Benchmark	**P0**	**T0**	**T1**	**V1**	**V2**	**VST-Full**	**D1**
Total tokens (input + output)							
MMMU-Pro Vision (200)	1,828	3,026	3,630	2,279	3,408	17,086	1,265
MMMU-Pro Standard (200)	4,418	3,499	3,950	4,719	4,328	14,307	650
MMMU (500)	3,298	1,845	3,422	3,043	2,506	5,784	825
OCR-BenchV2 (200)	2,750	961	2,929	2,756	2,571	4,743	1,241
Input / output split (tokens)							
MMMU-Pro Vision	262 / 1566	530 / 2496	1553 / 2077	295 / 1984	325 / 3083	12458 / 4628	1260 / 5
MMMU-Pro Standard	303 / 4115	368 / 3131	1225 / 2725	330 / 4389	357 / 3971	9773 / 4534	645 / 5
MMMU	232 / 3066	219 / 1626	1155 / 2267	260 / 2783	250 / 2256	3952 / 1832	820 / 5
OCR-BenchV2	158 / 2592	330 / 631	1533 / 1396	180 / 2576	161 / 2410	4256 / 487	1236 / 5

A.5 Macro-Average Evidence-Path Contrasts

Table 7 reports the descriptive macro-average evidence-path contrasts across the six VST benchmarks in Table 2. MIA-Bench and OCR-BenchV2 contribute benchmark scores scaled to 0–100; the other entries are accuracies. The same per-benchmark numbers used to compute these macro averages are in Table 2.

A.6 Uncertainty

For each benchmark and VST path, the primary statistic is accuracy on the fixed subset, except for MIA-Bench, where we use the official GPT-4o-as-judge score scaled by 100. The bracketed intervals reported in Table 2 are 95% Wilson score intervals computed from the per-subset binary correctness counts. For paired diagnostic contrasts, we report descriptive example-level differences whenever the same examples are evaluated under both paths. We do not make statistical significance claims in the main tables.

A.7 Token Cost per Path

Table 8 reports per-item average token consumption for each VST path, summed across all VLM and reasoner calls. Tokens are the reproducible cross-system unit; wall-clock and USD costs depend on hardware and provider and are not reported. Costs were logged for the four benchmarks listed; MIA-Bench and ChartQA logs are not included. The direct-answer reference D1 (Qwen2.5-VL-3B image-question answering) is included for comparison.

The cost-to-strict-marginal ratio differs sharply across benchmarks. On MMMU, $u_{\text{Full}} = 0$, so every VST-Full token beyond the simpler paths is overhead. On OCR-BenchV2, $u_{\text{Full}} = 16.0$ at a per-item budget of $4{,}743$ tokens against $1{,}241$ for D1, i.e., $3.8\times$ the direct-answer budget purchases 16pp of strict-marginal answerability. On MMMU-Pro Vision and Standard, u_{Full} of 10.0 and 6.0 are

purchased at $13.5\times$ and $22\times$ the direct-answer budget, the least favorable ratios in the four-benchmark subset.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and Section 1 state the paper's scoped claims: the Vision-Stripping Test (VST) decomposes benchmark accuracy into controlled evidence paths; the visual-evidence answerability profiles show benchmark-specific behavior; and VST-Full is the most permissive VST evidence path whose endpoint contrast is interpreted through the full profile. Sections 3, 4, and 5 define the protocol, report the corresponding profiles, and state the limits of the interpretation.

   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: Section 5 discusses what a VST visual-evidence answerability profile can and cannot show, the task families for which the protocol is intended, and why cost-aware comparisons are needed when a decomposed system is compared with monolithic VLMs. The paper does not claim that all visual tasks can be reduced to structured textual evidence.

   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 include theoretical theorems or proofs. The equations in Section 3 define the profile accuracies and paired evidence-path contrasts used by the protocol rather than new theoretical results.

   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: Section 3 defines the P0/T0/T1/V1/V2/VST-Full evidence paths, Algorithm 1, the visual evidence workflow, and the tool families enabled in each VST path. Section 4 names the benchmark splits, subset sizes, reasoner and VLM backbones, the D1 direct Qwen2.5-VL-3B baseline, additional direct-answer references, and grading setup; Appendix A.2 provides prompt, tool-policy, quality-guard, and visual-evidence representation details for the workflow.

   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: [Yes]

   Justification: The evaluation uses public benchmarks cited in Section 4: MMMU-Pro, MMMU, MIA-Bench, ChartQA, and OCR-BenchV2. The implementation, prompts, tool definitions, and run scripts are provided as anonymized supplementary material for review and will be released publicly upon acceptance.

   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 protocol is training-free and uses pre-trained models. Section 4 specifies the benchmark splits and subset sizes, the text-only reasoner (Qwen3-8B), the 3B VLM captioner/evidence extractor (Qwen2.5-VL-3B), the D1 direct Qwen2.5-VL-3B baseline, additional direct-answer references, and the evidence-path constraints enabled in each VST path.

   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: [Yes]

   Justification: Table 2 reports 95% Wilson score intervals for the main VST matrix, computed from per-subset binary correctness counts. Tables 7 and 4 are descriptive summaries of benchmark scores on fixed evaluation subsets: accuracies for multiple-choice benchmarks and benchmark scores for MIA-Bench and OCR-BenchV2. Appendix A.6 states the interval definition and that the paper does not make paired statistical significance claims.

   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: [No]

   Justification: All inference is run through hosted model APIs (Qwen2.5-VL-3B, Qwen3-8B, and the direct-answer reference VLMs), so no local GPU type, memory, or wall-clock time is reported. The paper notes in Section 5 that tool-call counts, model-role logs, token usage, wall-clock time, and API or local-inference cost are needed for fair endpoint comparisons. The experiments use public pretrained models and fixed evaluation subsets; implementation logs record model roles and tool calls for the stripping paths.

   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 work uses only publicly released models and public academic VQA benchmarks; no human subject data was collected and no personally identifying information is used. The authors have reviewed the NeurIPS Code of Ethics and conform to it in every respect.

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: [No]

Justification: The paper focuses on a benchmark-profiling protocol and does not include a dedicated broader-impacts section. Potential positive impacts include more transparent evaluation of multimodal reasoning benchmarks; potential negative impacts include over-trusting textualized visual evidence or tool-routed systems in high-stakes settings outside the paper's intended scope.

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 a new pre-trained model, image generator, scraped dataset, or benchmark containing newly collected sensitive content. The released artifact is an inference and evaluation workflow over existing public benchmarks and models.

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: The models and benchmarks used in the paper are credited through citations in Sections 2 and 4, including Qwen, MMMU-Pro, MMMU, MIA-Bench, ChartQA, and OCR-BenchV2. The work evaluates these existing assets without redistributing third-party model weights or benchmark data.

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: [Yes]

Justification: The new asset is the VST evaluation workflow, not a new model or dataset. The supplementary material documents the VST paths, prompts, evidence-state format, tool definitions, and run scripts; Appendix A.2 reproduces the structured-representation and prompt details.

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: [N/A]

Justification: The paper does not involve any crowdsourcing or human-subject research. All evaluation is conducted on existing public VQA benchmarks.

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: No human subjects are involved in this research, so IRB approval is not applicable.

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: [Yes]

Justification: LLMs and VLMs are central to the method and baselines. Sections 3 and 4 identify the text-only reasoner (Qwen3-8B), the captioner/question-guided VLM evidence setting (Qwen2.5-VL-3B), the VST-Full VLM-based evidence tools (Qwen2.5-VL-3B), the D1 direct Qwen2.5-VL-3B baseline, and the additional direct-answer VLM references.

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.