MM-RAGChecker: Validated Claim-Level Diagnosis for Visual Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Evaluating visual retrieval-augmented generation (RAG) requires more than answer-level accuracy. A system may produce a correct short answer while grounding its long-form justification in the wrong modality, ignoring retrieved evidence, or hallucinating unsupported claims. We present **MM-RAGChecker**, a multimodal claim-level diagnostic framework for visual RAG. MM-RAGChecker decomposes long-form answers into atomic claims, verifies each claim against retrieved text and image evidence, and aggregates these judgments into interpretable metrics for hallucination, claim recall, context precision, evidence utilization, and cross-modal grounding balance.

To validate the framework, we evaluate both its individual components and its end-to-end diagnostic reliability on model-generated answers. We further deploy MM-RAGChecker on **MRAG-Suite**, a cross-domain visual RAG testbed spanning web photos, charts, scanned documents, slides, and scholarly PDFs under controlled ambiguity and distractor settings. Across evaluated vision-language models (**VLMs**), we find that (1) domains with similar EM can differ sharply in grounding quality, (2) retrieved text is often underused relative to images, and (3) ambiguity and distractors substantially worsen both answer accuracy and claim-level faithfulness. These results show that MM-RAGChecker surfaces failure modes that remain invisible under conventional answer-level metrics alone.

1 Introduction

Retrieval-augmented generation (RAG) grounds large language models (LLMs) in external evidence to improve factuality and reliability. With the rise of multimodal models, *visual* RAG (jointly leveraging text and images) has gained traction for complex questions involving charts, plots, figures, photos, and document pages. Evaluation has not kept pace. Most multimodal benchmarks either assume pre-selected visuals, bypassing retrieval, or remain text-only. This limits our ability to measure retrieval accuracy, multimodal grounding, and robustness in realistic settings. Existing multimodal QA datasets such as ScienceQA (Lu et al., 2022), MMMU (Yue et al., 2023), and MMMU-Pro (Yue et al., 2024) supply the necessary visual context in the input, but they sidestep retrieval. Retrieval-centric efforts like Visual Haystacks (Wu et al., 2024) emphasize hard visual look-up but focus on a single domain and lack fine-grained diagnos-

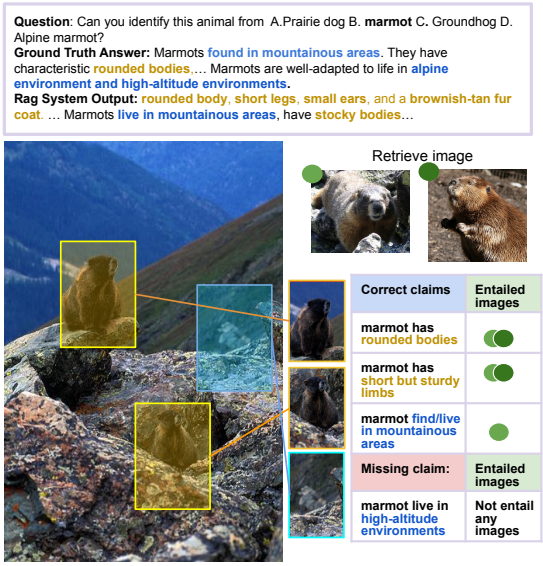

Figure 1: Illustration of our multimodal claim verification process. Given a long-form answer from a RAG system, we extract atomic claims and assess support from retrieved visual evidence. Claims are labeled **Entailment** if at least one image supports them. In this example, several traits are correctly described, but the gold claim about "high-altitude environments" remains unsupported.

Code, prompts, and diagnostics available at: https://anonymous.4open.science/r/MRAGChecker-8DCD/

tics. Text-only tools such as RAGChecker and RefChecker (Ru et al., 2024; Hu et al., 2024b) offer claim-level analysis, yet cannot assess how claims are grounded in *visual* evidence.

Three systematic evaluation gaps. First, most multimodal benchmarks supply the relevant image directly to the model, bypassing retrieval entirely (Lu et al., 2022; Yue et al., 2023). This inflates scores by removing the retrieval challenge that real deployments face. Second, answer-level metrics cannot attribute failures to specific modalities: a model grounding all claims in text while ignoring images receives the same EM score as one that genuinely uses visual evidence. Third, query ambiguity is a significant but uncontrolled confound: our experiments show it inflates Hallucination Rates by 9.5 points, yet existing benchmarks provide no mechanism to isolate its effect.

To address these gaps, we introduce **MM-RAGChecker**, a multimodal claim-level diagnostic framework for visual RAG. Rather than evaluating only final answers, MM-RAGChecker verifies whether individual generated claims are supported by retrieved text and image evidence, and attributes errors to specific modalities and retrieved items. We further deploy it on **MRAG-Suite**, a cross-domain visual RAG testbed with controlled ambiguity and distractor settings.

Contributions.

1. We present **MM-RAGChecker**, a multimodal claim-level diagnostic framework for visual RAG that attributes support and errors to retrieved text and image evidence at the claim-level.

2. We provide **multi-level validation** of MM-RAGChecker, including component-level evaluation for claim segmentation and claim verification, as well as end-to-end human validation on model-generated answers.

3. We introduce a set of **diagnostic metrics** for visual RAG (including hallucination, claim recall, context precision, evidence utilization, and cross-modal grounding balance) and show that they capture failure modes that are not revealed by EM or ROUGE alone, while remaining complementary to answer-level accuracy and correlated with human judgments of groundedness.

4. We deploy MM-RAGChecker on **MRAG-Suite**, a cross-domain evaluation testbed that we assemble from eight existing multimodal QA sources under unified schema and controlled retrieval conditions, and use it to uncover systematic modality skew and evidence under-utilization in current VLMs.

2 Related Work

Claim-level evaluation for RAG. Evaluation of retrieval-augmented generation has moved beyond answer-level metrics toward finer-grained attribution and factuality analysis. ALCE evaluates attributed long-form generation through citation quality (Gao et al., 2023); ARGUE uses nugget-style evaluation for information coverage (Mayfield et al., 2024); and RAGAS provides text-based faithfulness and relevance metrics for RAG systems (Es et al., 2024). Most closely related on the text side, RAGChecker and RefChecker diagnose claim-level consistency and support in textual RAG outputs (Ru et al., 2024; Hu et al., 2024b). More broadly, FActScore established the paradigm of decomposing long-form outputs into atomic claims and verifying them against external evidence (Min et al., 2023), while WICE motivates minimal verifiable units for grounded evaluation (Kamoi et al., 2023). We build on this line of work, but focus on *retrieval-conditioned multimodal grounding* rather than text-only factuality.

Multimodal factuality and claim verification. Prior work on multimodal factuality has studied hallucination and faithfulness in settings such as visual question answering and image captioning. HallusionBench evaluates visual hallucination in multimodal generation, but does not model retrieval-conditioned evidence use (Guan et al., 2024). FaithScore measures multimodal factuality for image-conditioned generation rather than RAG (Jing et al., 2024). Most closely related to our setting, MiRAGE extends claim checking to multimodal RAG and adapts text-based RAG metrics to multimodal evidence (Martin et al., 2025). MM-RAGChecker differs in emphasizing (i) retrieval-conditioned diagnosis, (ii) explicit cross-modality attribution, (iii) Self-Knowledge separated from retrieved support, and (iv) controlled ambiguity and distractor analysis across multiple domains.

Visual RAG benchmarks. Existing multimodal benchmarks such as ScienceQA, MMMU, and

MMMU-Pro evaluate reasoning with provided image-text context, thereby bypassing retrieval (Lu et al., 2022; Yue et al., 2023, 2024). Visual Haystacks stresses multimodal retrieval difficulty, but is single-domain and does not provide claim-level grounding diagnostics (Wu et al., 2024). Recent work has also highlighted modality and positional biases in multimodal systems (Yao et al., 2025). In contrast, MRAG-Suite serves as a *cross-domain evaluation testbed* for stress-testing visual RAG under controlled ambiguity and distractor conditions, while MM-RAGChecker provides the claim-level diagnostic layer needed to analyze grounding failures.

3 MRAG-Suite: Cross-Domain Evaluation Testbed

3.1 Task Definition

We evaluate multimodal Retrieval-Augmented Generation (RAG) for open-ended QA. Given a query q, the system retrieves textual (E^{txt}) and visual (E^{img}) evidence and must produce (i) a concise short answer and (ii) a grounded long answer: $(a^{\text{short}}, a^{\text{long}}) = f(q, E^{\text{txt}}, E^{\text{img}})$.

3.2 Corpora Overview

MRAG-Suite consolidates eight heterogeneous sources: **WebQA** (Chang et al., 2022) (open-web multi-hop QA with images and short passages), **Chart-MRAG** (Yang et al., 2025; Masry et al., 2022) (chart understanding with OCR/metadata), **Visual-RAG** (Wu et al., 2025) (fine-grained visual knowledge requiring object and attribute recognition), **MRAG-Bench** (Hu et al., 2024a) (1,353 vision-centric multiple-choice items; we retain the original answer choices and treat the correct option text as a^{short}, making this a constrained-choice task rather than free-form generation), and four **VisRAG** subsets (Yu et al., 2024; Li et al., 2024; Methani et al., 2019; Tanaka et al., 2023; Tito et al., 2022) covering scholarly PDFs (ArXiv), technical plots, slide decks, and scanned documents. Each source contributes distinct modalities and reasoning demands. Full descriptions and dataset statistics are reported in Appendix A.

3.3 Normalization and Indexing

We normalize all samples to `{question, short_answer, long_answer, evidence_imgs, evidence_txts}`. Images use CLIP (ViT-L/14) features plus captions (Radford et al., 2021); text passages use dense embeddings. FAISS indexes the visual items; a dense passage retrieval (DPR) / ElasticSearch store holds the text.

Pipeline Stage	Model Used
Stage 1: Claim extraction	Claude 3.5 Sonnet
Stage 2: Per-evidence judging	Claude 3.5 Sonnet
Gold salient claim extraction	Claude 3.5 Sonnet
Reference correctness check	Claude 3.5 Sonnet
Step 1 filter (closed-book)	Claude 3.5 Sonnet
Step 2 filter (VLM panel)	6 VLMs (Table 2)

Table 1: Models used at each stage of the MM-RAGChecker pipeline. All judge stages use the same model and prompt configuration to avoid confounding.

3.4 End-to-End Pipeline

Table 1 summarizes the model components used in the pipeline. Given a normalized sample, we (i) retrieve shared text and image evidence with a fixed hybrid retriever, (ii) run each target VLM once to produce short and long answers, and (iii) apply MM-RAGChecker to decompose the long answer into claims, judge each claim against retrieved evidence, and aggregate the results into diagnostic metrics. We use a single retriever, a single generation pass, and a single judge configuration throughout.

3.5 Two-Step Filtering for Non-Triviality

We remove trivial and guessable items in two steps.

Step 1: Retrieval-independent filter. We first remove questions that are likely solvable without retrieval. We discard items meeting either condition: (i) a strong closed-book model (Claude 3.5) answers the short gold answer correctly with high confidence; or (ii) the gold answer appears verbatim in the question. This step excludes items solvable from parametric memory or surface answer leakage.

Step 2: Difficulty-based filter with a model panel. For the remaining items, we estimate difficulty via a 6-model VLM panel (Table 2) in the `gt_only` (GO) setting, computing per-question success rate and removing the easiest $\sim$10% per dataset. This design ensures the final difficulty ranking is not driven by a single judge (Appendix H reports per-dataset statistics). We refer to the post-filtering subset as the **Filtered split** and the original data as the **Full split** throughout this paper.

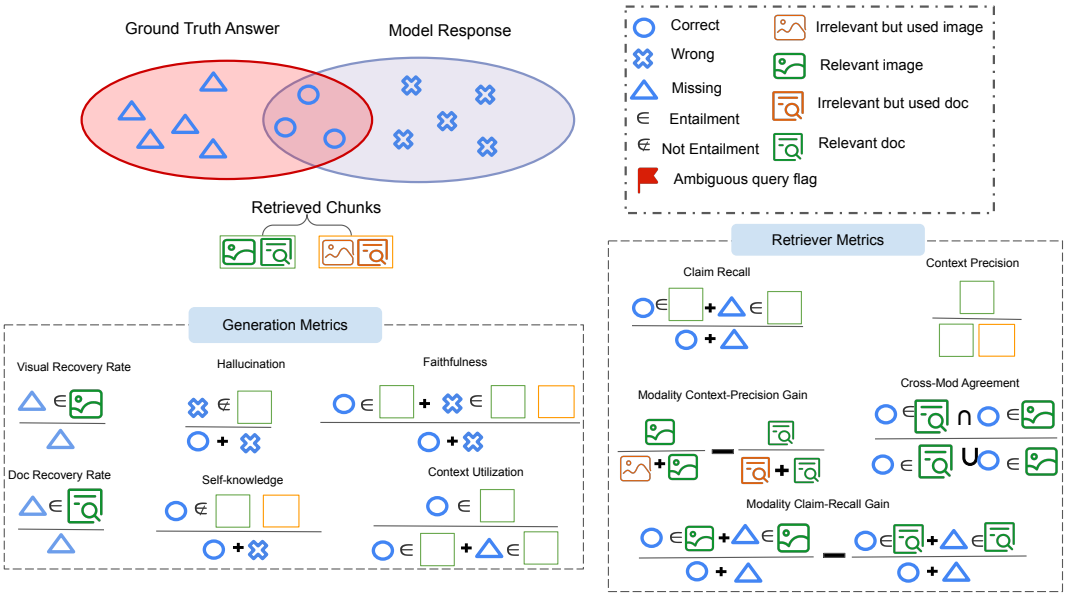

Figure 2: MM-RAGChecker pipeline overview. **(Left)** A long-form answer is decomposed into atomic claims via an LLM-based splitter (Stage 1). **(Center)** Each claim is independently judged against each retrieved text passage and image using three-way natural language inference (NLI): ENTAILMENT, NEUTRAL, or CONTRADICTION (Stage 2, capped at K_{img}=K_{txt}=3). **(Right)** Claim verdicts are aggregated via precedence rule (ENTAILMENT > CONTRADICTION > NEUTRAL) and used to compute the full diagnostic suite: generation metrics (Hallucination, Faithfulness, Self-Knowledge; top group) and retriever metrics (Claim Recall, Context Precision, Cross-Modal Agreement; bottom group). Cross-modality gaps ΔCR and ΔCP compare image-attributed vs. text-attributed claim sets.

3.6 Output Specification

Systems must return both short and long answers. We report Exact Match (EM) on short answers and ROUGE-L plus claim-level diagnostics on long answers (MM-RAGChecker, Sec. 4). Given a model answer $\hat{a}$ and the gold short answer a^{short},

$$\text{EM}(\hat{a}, a^{\text{short}}) = \mathbf{1}\big[\text{norm}(\hat{a}) = \text{norm}(a^{\text{short}})\big],$$

where $\text{norm}(\cdot)$ lowercases, strips punctuation, and normalizes whitespace. For multiple-choice items (e.g., MRAG-Bench), we report Accuracy over the label set instead.

3.7 Retrieval Modes and Distractors

To probe robustness, we evaluate three settings: `gt_only` (**GO**), providing only gold evidence; `gt_plus_distractors` (**GPD**), mixing gold with distractors selected by CLIP cosine similarity ≥ 0.8 to the gold image (CLIP ViT-L/14 embeddings, cosine distance); and `distractors_only` (**DO**), providing only distractors to test spurious reasoning.

4 MM-RAGChecker: Multimodal Claim-Level Diagnosis

We focus here on the claim-level decision rules and metrics; Figure 2 overviews the three-stage pipeline (claim extraction, per-evidence judging, aggregation).

MM-RAGChecker is deliberately minimal: it does not train new models or perform multi-round self-consistency. Instead, it applies a single off-the-shelf LLM/VLM judge in three fixed stages with deterministic decoding, reused for all datasets and systems.

Stage 1 — Claim extraction. We split the long answer a^{long} into minimal verifiable units $\mathcal{C} = \{c_i\}$ using an LLM-based splitter (Claude 3.5 Sonnet) plus light heuristics for compound clauses. To avoid anaphoric ambiguity (e.g., "they" referring to a prior noun), the splitter is prompted to resolve pronouns to their explicit antecedents within the same answer before producing claim units.

Stage 2 — Per-evidence judging. Given retrieved images $\{I_k\}$ and text passages $\{t_k\}$, we directly judge each pair (c_i, e), $e \in \{I_k\} \cup \{t_k\}$, with a three-way decision $\{$ENTAILMENT, NEUTRAL, CONTRADICTION$\}$.

We cap to top K_{img}=3 and K_{txt}=3 and use temperature 0.

Stage 3 — Aggregation. Each claim receives a final label L_i via a simple precedence rule: ENTAILMENT if any judgment is ENTAILMENT; else CONTRADICTION if any is CONTRADICTION; otherwise NEUTRAL. For modality analysis, we record whether a claim is supported by images or text and mark an evidence item as "used" if it entails some claim.

Implementation. Table 1 summarizes the models used at each stage. The claim splitter, entailment judge, and gold-claim extractor all use the same model (Claude 3.5 Sonnet) with deterministic decoding (temperature 0, fixed seeds) to ensure reproducibility. Using a single judge model across all stages simplifies the pipeline but means that systematic biases of this judge propagate throughout; we discuss this trade-off in the Limitations section.

We validate MM-RAGChecker at both component and end-to-end levels in Section 5.2.

Notation. L_i is the aggregated full-context label for claim c_i, while $j(c_i, e)$ is the single-evidence label for an individual evidence item $e \in \mathcal{E}$. A claim c_i receives L_i=ENTAILMENT if any single-evidence judgment is ENTAILMENT, even if others are NEUTRAL. **Gold salient claims** $\mathcal{G}$ are extracted from reference long answers using the same LLM splitter as Stage 1, applied to ground-truth answers rather than model outputs.

4.1 Metrics

Core metrics. Let $\mathcal{C}$ be the set of extracted claims, L_i the final label of c_i, and $\mathcal{E}$ the set of retrieved items. Then

$$\text{Ungrounded Rate} = \frac{\#\{\, i : L_i = \text{NEUTRAL}\,\}}{|\mathcal{C}|}, \quad (1)$$

$$\text{Faithfulness} = \frac{\#\{\, i : L_i = \text{ENTAILMENT}\,\}}{|\mathcal{C}|}. \quad (2)$$

Hallucination notation. Ungrounded Rate (Eq. 1) counts NEUTRAL claims and is our primary generation metric. The modality-attributed V-HR/D-HR (Appendix F) instead use the broader non-Entailment set (NEUTRAL∪CONTRADICTION), since a wrong-modality grounding is still a failure. "Hallucination Rate" in tables denotes Ungrounded Rate unless noted.

Let $\mathcal{G}$ be gold salient claims mined from a reference long answer.

$$\text{Claim Recall (CR)} = \frac{\#\{\text{gold } g \in \mathcal{G} \text{ that appear in } a^{\text{long}}\}}{|\mathcal{G}|}. \quad (3)$$

Let $\mathcal{E}_{\text{used}} = \{e \in \mathcal{E} : \exists i,\ j(c_i, e) = \text{ENTAILMENT}\}$.

$$\text{Context Precision (CP)} = \frac{|\mathcal{E}_{\text{used}}|}{|\mathcal{E}|}. \quad (4)$$

CP rewards all evidence-supported claims, including tangential ones; restricting to gold claims in a^{long} would yield stricter estimates, left as future work.

Self-Knowledge. We define Self-Knowledge as the fraction of claims that are *correct* with respect to the reference answer but *not supported* by any retrieved evidence item. Let $\mathcal{C}_{\text{correct}}$ be the set of generated claims that match a gold salient claim in $\mathcal{G}$, with the match determined by the same judge via pairwise entailment. Let $\mathcal{C}_{\text{unsup}} = \{c_i : \forall e \in \mathcal{E},\ j(c_i, e) \neq \text{ENTAILMENT}\}$ be the set of claims unsupported by any retrieved item. Then:

$$\text{Self-Knowledge} = \frac{|\mathcal{C}_{\text{correct}} \cap \mathcal{C}_{\text{unsup}}|}{|\mathcal{C}_{\text{correct}}|}. \quad (5)$$

Note that $\mathcal{C}_{\text{unsup}}$ corresponds to claims with $L_i \in \{\text{NEUTRAL}, \text{CONTRADICTION}\}$ under our aggregation rule. Self-Knowledge thus measures the proportion of correct answer content that originates from the model's parametric memory rather than from retrieved evidence.

Cross-modality metrics. For each claim c_i, we record binary indicators s_i^{img} and s_i^{txt} denoting whether at least one image or text item entails it. We then compute modality-specific claim recall ($\text{CR}_{\text{img}}, \text{CR}_{\text{txt}}$) and context precision ($\text{CP}_{\text{img}}, \text{CP}_{\text{txt}}$), along with their gaps ΔCR and ΔCP. We also report Visual Hit@k, Text Miss Rate, Cross-Modal Agreement (CMA), and modality-attributed Hallucination Rates (V-HR, D-HR). We report representative cross-modality diagnostics in Table 3; per-modality Recall, Precision, and Hallucination diagnostics for all model–dataset pairs under all retrieval modes are in Appendix G, from which ΔCR and ΔCP can be derived. Key patterns are consistent: ΔCR is positive on chart-heavy domains (image preference), TXT-MissRate exceeds 80% in most settings, and CMA remains below 15%. Formal definitions of all cross-modality metrics are in Appendix F.

5 Experiments and Results

We evaluate MRAG-Suite along three axes: (1) end-task accuracy on short answers, (2) factuality

Model	Chart		MRAG		ArXiv		Doc		Plot		Slide		Visual-RAG		WebQA		Macro	
	F	*U*	*F*	*U*	*F*	*U*	*F*	*U*	*F*	*U*	*F*	*U*	*F*	*U*	*F*	*U*	*F*	*U*
Qwen2.5-VL-72B	**22.4**	19.1	32.6	33.8	55.3	48.7	31.1	**35.4**	36.7	45.8	18.5	**25.3**	**13.8**	13.1	16.2	14.9	28.3	29.5
LLaVA-v1.6-34B	15.3	25.7	19.5	24.6	34.6	34.5	20.2	21.9	12.3	13.6	8.7	10.4	9.6	11.5	12.7	13.4	16.6	19.5
Claude 3.5	10.8	11.9	28.3	26.1	35.5	35.1	28.5	29.2	35.9	36.8	17.6	22.1	11.7	10.6	14.8	17.3	22.9	23.6
Phi-4	11.2	13.5	32.4	30.7	38.6	36.9	27.6	28.2	6.4	9.2	**24.3**	22.6	8.5	9.2	**20.6**	**22.1**	21.2	21.6
Pixtral-12B-2409	14.7	**29.4**	**38.8**	**37.5**	43.1	43.7	26.9	26.6	**51.4**	**56.5**	18.2	15.5	12.1	**13.3**	**26.4**	**27.9**	**29.0**	**31.3**
InternVL3-8B	19.2	27.6	21.9	21.4	**63.9**	**55.4**	**39.7**	34.6	33.5	36.1	20.8	22.9	11.2	12.6	17.8	28.7	28.5	29.9
Macro Avg	15.6	21.2	28.9	29.7	44.8	42.4	28.9	29.5	29.4	33.0	18.0	19.8	11.1	11.8	18.1	20.5	24.4	26.0

Table 2: EM/Accuracy (%) on **Filtered split** (*F*) and **Full** (*U*) splits. Cell color encodes performance intensity (darker green = higher). **Bold** marks per-column best. The bottom row reports macro averages across all eight datasets.

Dataset	Model	ΔCR	ΔCP	VIS-Hit@k	TXT-MissRate	CMA	V-HR	D-HR
Chart-MRAG	Phi-4 (GPD)	54.54	59.50	**81.38**	83.00	**13.02**	**66.87**	**45.2**
WebQA	Phi-4 (GPD)	6.59	1.50	**14.79**	87.00	**10.86**	**12.41**	**0.33**
WebQA	Claude 3.5 (GPD)	3.09	-0.35	13.00	**83.33**	6.52	32.44	3.70

Table 3: Cross-modality diagnostics for selected pairs spanning the largest $|\Delta\mathrm{CR}|$ range (see Figure 3 for visualisation). Per-modality diagnostics for all 48 model–dataset pairs are in Appendix G. ΔCR/ΔCP: image vs. text evidence gaps; positive = image preference.

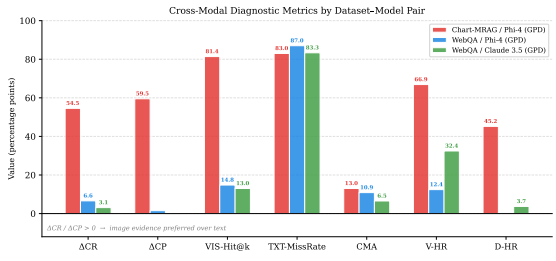

Figure 3: Cross-modal diagnostic metrics for selected dataset–model pairs. Chart-MRAG/Phi-4 (red) shows strong image preference (ΔCR=54.5, V-HR=66.9), while WebQA pairs show near-zero ΔCR. Full numeric values in Table 3.

and grounding quality of long answers via MM-RAGChecker, and (3) robustness to query difficulty, ambiguity, and distractors.

5.1 Experimental Setup

Models. We evaluate six VLMs: **Qwen2.5-VL-72B** (primary open-source baseline), **LLaVA-v1.6-34B** (for comparison with prior work (Martin et al., 2025)), **Claude 3.5** (API, proprietary upper bound), **Phi-4**, **Pixtral-12B-2409**, and **InternVL3-8B**. All share identical retrieval conditions: a CLIP-based dual encoder for images and DPR for text, fetching up to 5 passages and 5 images per query.

Implementation Details. We use the pipeline of Section 3.4 with Claude 3.5 as the fixed judge for all systems and datasets. Decoding is greedy (temperature 0); we do no sampling, voting, or self-refinement, so each sample yields a single answer and a single set of diagnostic labels.

Metrics. For short answers we report EM/Accuracy. For long answers we compute ROUGE-L against our reference explanations and all claim-level diagnostics using MM-RAGChecker (Hallucination Rate, Faithfulness, Claim Recall, Context Precision, Self-Knowledge, Modality Bias). We further report the fraction of retrieved items actually used (*evidence utilization*). Retrieval robustness is probed under GO, GPD, and DO (Section 3.7). Unless noted, we report macro averages on the Filtered split with decoding temperature $= 0$ and identical retriever outputs across models. We acknowledge that candidate pool sizes (e.g., ∼300 in Visual-RAG) are smaller than real-world corpora; results under larger, noisier pools may differ (see Limitations).

5.2 Validation of MM-RAGChecker

We validate MM-RAGChecker at two levels.

Component-Level. For claim segmentation, human annotators achieve Boundary F1 of 86.3% against the automatic splitter (Table 15 in Appendix E). For claim–evidence entailment, agreement on N=200 pairs from Chart-MRAG and MRAG-Bench reaches 80.4% (Cohen's κ=0.73; Table 16).

Validation target	Metric	Score
Claim–evidence support	Accuracy / κ	80.4 / 0.73
Question-level verdict	Accuracy / κ	82.0 / 0.71

Table 4: Human validation of MM-RAGChecker at component and end-to-end levels.

Feature	**Ours**	**MiRAGE**	**FActSc.**	**RAGAS**
Visual evidence	✓	✓	×	×
Modality attribution	✓	×	×	×
Self-Knowledge	✓	×	×	×
ΔCR / ΔCP	✓	×	×	×
Distractor eval	✓	×	×	×
Ambiguity subset	✓	×	×	×
Human validated	κ=.73	–	varies	–

Table 5: Comparison of MM-RAGChecker with related evaluation frameworks. ✓= supported; × = not supported; – = not reported.

End-to-End. To test whether component errors compound, we annotate 50 complete questions and compare MM-RAGChecker's question-level hallucination verdict against human judgments: agreement is 82.0% (κ=0.71), indicating that aggregation does not amplify errors (Table 4).

Complementarity with EM/ROUGE. On the Filtered split, Hallucination Rate correlates with EM at ρ=−0.31 and ROUGE-L at ρ=−0.28, while Context Precision is nearly uncorrelated with EM (ρ=0.09). This confirms that claim-level diagnostics capture grounding failures not visible from answer accuracy alone.

5.3 Comparison with Existing Evaluation Metrics

Table 5 situates MM-RAGChecker relative to the most closely related evaluation tools. MiRAGE (Martin et al., 2025) is the closest prior work: it extends claim checking to multimodal RAG but does not perform cross-modality attribution or distinguish parametric from retrieved knowledge. FActScore (Min et al., 2023) and RAGAS (Es et al., 2024) operate on text only. The key differentiators of MM-RAGChecker are: (i) cross-modal attribution via modality-disaggregated recall and precision (ΔCR, ΔCP), which expose whether a model preferentially grounds claims in images vs. text; (ii) Self-Knowledge measurement, isolating parametric support from retrieval; and (iii) validation across three controlled retrieval conditions spanning eight heterogeneous domains.

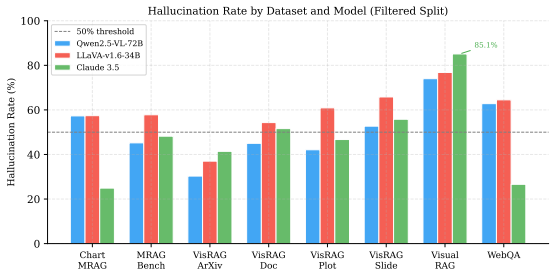

Figure 4: Hallucination Rates by dataset and model (Filtered split). Visual-RAG consistently exceeds 74%, while VisRAG-ArXiv stays below 42%, confirming that hallucination varies substantially by domain independent of EM. Full diagnostics (Faithfulness, Claim Recall, Context Precision) are in Table 14 (Appendix E).

5.4 End-Task Accuracy (short-answer EM/Acc)

Table 2 shows that accuracies vary widely by domain: VisRAG-ArXiv and VisRAG-Plot are comparatively easier, while Visual-RAG and VisRAG-Slide remain challenging. Moving from the Filtered to Full split generally increases scores (e.g., Pixtral-12B on Chart-MRAG: 14.7% → 29.4%), confirming that filtering removes easier, less retrieval-dependent items. Figure 4 confirms that Visual-RAG occupies the high-hallucination region regardless of EM.

5.5 Claim-Level Diagnostics (MM-RAGChecker)

Table 14 (Appendix E) summarizes claim-level diagnostics on the Filtered split. Hallucination varies sharply by domain: it is lowest on structured documents (VisRAG-ArXiv: 30–41%) and highest on open-world visual recognition (Visual-RAG: 74–85%). On WebQA, image Context Precision is high (89–91%) while text precision is much lower (13–31%), indicating systematic text under-utilization. Across datasets, models commit to one modality rather than reconciling both. Macro-averaged hallucination remains 48–59% even for the strongest models.

5.6 Qualitative Error Analysis

Case 1: EM-invisible grounding failure. On a WebQA question about a historical building, Qwen2.5-VL-72B produces the correct short answer (EM=1) but grounds its long answer entirely in parametric knowledge: MM-RAGChecker labels 4/5 claims as NEUTRAL, yielding Faithfulness=20% despite perfect EM. This exemplifies the

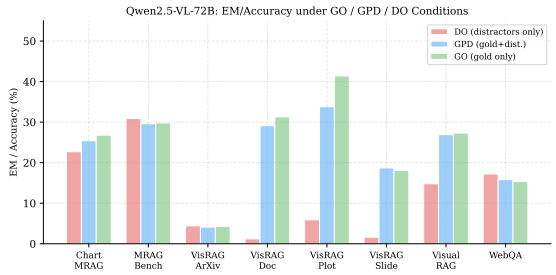

Figure 5: EM/Accuracy of Qwen2.5-VL-72B under gold-only (GO), gold+distractors (GPD), and distractors-only (DO) conditions. The GO–GPD gap quantifies distractor susceptibility; VisRAG-Doc and VisRAG-Plot exhibit the largest drops, revealing brittle evidence selection. Full per-model results in Table 17 (Appendix G).

EM–grounding disconnect that motivates claim-level evaluation.

Case 2: Modality skew. On a Chart-MRAG question requiring both the chart image and its caption, Claude 3.5 extracts all numeric claims from the image (s_i^{img}=1) but ignores the text legend (s_i^{txt}=0 for all claims), resulting in ΔCR= + 62 points. When the chart is replaced by a distractor (DO setting), the model hallucinates plausible but incorrect values, confirming brittle single-modality dependence.

Case 3: Pipeline failure mode. On a VisRAG-Slide question, the claim splitter fails to decompose a compound sentence ("The slide shows quarterly revenue of $2.1M and a 15% YoY growth"), leading to a single claim that receives ENTAILMENT based on the revenue figure alone, masking the unsupported growth claim. This under-segmentation pattern accounts for $\sim$40% of end-to-end errors (Appendix E).

5.7 Effect of Retrieval Modes

Figure 5 summarizes EM/Accuracy under DO, GPD, and GO for our primary baseline. DO severely degrades performance, while GPD lies between DO and GO. The GO–GPD gap quantifies distractor susceptibility; Chart-MRAG and VisRAG-Doc exhibit the largest drops. Full per-model breakdowns are in Table 17.

5.8 Prompt Sensitivity Study

We sweep 12 prompt designs varying init style, exemplar count, and reasoning pattern (Table 6; Figure 6). `retrieve_then_reason` dominates on text-heavy tasks (MRAG-Bench, WebQA), while `direct` is better on focused visual QA (VisRAG-* splits). Per-dataset EM standard deviation across the 12 designs averages 2.9 points, with the reasoning pattern contributing most variance. Full results are in Appendix D.

Dataset	Best Prompt	Mean Acc. (%)
MRAG-Bench	expert3 retrieve & reason	31.5
VisRAG-ArXiv	expert6 direct	5.0
VisRAG-Doc	plain6 direct	34.3
VisRAG-Plot	expert3 direct	36.5
VisRAG-Slide	plain3 direct	35.2
WebQA	expert6 retrieve&reason	23.0

Table 6: Best-performing prompt (by mean EM/Acc across models) per dataset on the Filtered split. The patterns highlight when "retrieve-then-reason" helps (text-heavy tasks) versus when direct prompting is sufficient (focused visual tasks).

5.9 Impact of Distractors and Filtering

From GO $\rightarrow$ GPD/DO, we observe higher hallucination, lower context precision, and stronger modality skew ($|\Delta\text{CR}|$ increases, CMA decreases), indicating that distractors degrade both *what* is answered and *how* it is justified (Table 17, Appendix G). Filtering also raises mean long-answer length from 45 to 60 tokens but flags more clauses as unsupported—harder questions trigger verbosity, not better grounding.

6 Conclusion

We presented **MM-RAGChecker**, a multimodal claim-level diagnostic framework for visual retrieval-augmented generation. By decomposing long-form answers into atomic claims and verifying them against retrieved text and image evidence, MM-RAGChecker exposes grounding failures that remain invisible under answer-level metrics alone. We validated the framework at both the component and end-to-end levels, and deployed it on MRAG-Suite, a cross-domain evaluation testbed, to study ambiguity, distractors, evidence under-utilization, and modality skew. Our findings suggest that improving visual RAG requires not only stronger answer accuracy, but also better evidence selection, tighter grounding, and more balanced cross-modal evidence use.

Limitations

This work targets visual RAG for images, charts, slides, scanned pages, and scholarly PDFs. Results may not transfer to videos, embodied settings, or low-resource languages. Hybrid retrieval depends

on CLIP-like encoders and OCR text. Known dataset and model biases, plus OCR noise, can affect measured recall/precision and inflate hallucination labels. Our diagnostics rely on LLM/VLM components for claim extraction and entailment; despite prompt controls and spot-checks, such judges can diverge from expert assessments in domain-specific cases. More rigorous extractors such as Claimify (Metropolitansky and Larson, 2025) could be substituted to strengthen the extraction step in future work. Our pipeline uses a single, fixed judge configuration; systematic biases of this judge therefore affect all reported diagnostics. In particular, because Claude 3.5 serves as both judge and evaluated system, its diagnostic scores may benefit from self-evaluation bias; future work should assess with an independent judge. Scores are also sensitive to prompt format, retrieval k, and distractor sampling; we report settings but do not exhaust all combinations.

Retrieval Pool Scale. Candidate pool sizes in MRAG-Suite range from $\sim$300 (Visual-RAG) to $\sim$19,000 (WebQA). The smaller pools may underestimate retrieval noise relative to real-world deployments where corpora contain millions of items. We expect that scaling to larger pools would increase distractor effects and lower Context Precision further. However, the *relative* ranking of models and the diagnostic patterns (e.g., modality skew, text under-utilization) likely persist, since these arise from generation-side behavior rather than retrieval noise alone. Validating this hypothesis with million-scale pools is an important direction for future work.

Model Recency. We use Claude 3.5 as our proprietary baseline and evaluate open-source models up to 72B parameters. Newer models (e.g., Qwen3-VL, GPT-4o) have since become available and may perform differently on MRAG-Suite. MRAG-Suite and MM-RAGChecker are model-agnostic by design; we encourage the community to benchmark newer models using the released code and data.

Ethics Statement

MM-RAGChecker is designed as a diagnostic and debugging aid for multimodal RAG systems, not as a safety filter for deployment. Automatic claim-level verification can be wrong; false negatives may miss harmful or misleading claims, and false positives may over-flag well-grounded content. We recommend using MM-RAGChecker as a screening tool with human oversight for high-stakes settings (e.g., medical, legal, or financial visual QA).

All datasets used in MRAG-Suite are publicly released for research use under their respective licenses (WebQA, ScienceQA, ChartQA, MRAG-Bench, VisRAG, Visual-RAG), and we comply with the original intended use of each source. Retrieved images and text passages are drawn from publicly available web pages, scholarly PDFs, and benchmark corpora; no private or personally identifiable content is processed.

For our human validation (Section 5.2), annotators are graduate students who consented to the annotation task as part of standard research duties; no crowdsourcing platform was used and no separate per-task compensation was provided. Annotators were not asked to provide medical, legal, or other professional advice.

We will release the MM-RAGChecker checker code, the MRAG-Suite schema, and evaluation scripts under CC BY 4.0 for non-commercial research use, to support transparency and reproducibility.

Acknowledgments

Use of Generative AI. Claude 3.5 Sonnet is used as a methodological component of this work, serving as the claim splitter, per-evidence entailment judge, gold-claim extractor, and closed-book non-triviality filter; the full mapping of models to pipeline stages is documented in Section 3.4 and Table 1. In addition, the authors used general-purpose LLM assistants for light editing and polishing of prose; all technical content, experimental design, results, and analyses are the authors' own.

References

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of EACL (Demo)*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of EMNLP*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of CVPR*.

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. 2024a. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. *arXiv preprint arXiv:2410.08182*.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024b. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *ArXiv*, abs/2405.14486.

Liqiang Jing and 1 others. 2024. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of EMNLP*.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in wikipedia. In *Proceedings of EMNLP*.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Annual Meeting of the Association for Computational Linguistics*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and A. Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *ArXiv*, abs/2209.09513.

Alexander Martin and 1 others. 2025. MiRAGE: A multimodal retrieval-augmented generation evaluation framework. *Preprint*, arXiv:2510.24870.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

James Mayfield, Eugene Yang, Dawn Lawrie, Sean MacAvaney, Samuel Huston, and Andrew Yates. 2024. ARGUE: Argument-based retrieval-augmented generation evaluation. In *Proceedings of SIGIR*.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2019. Plotqa: Reasoning over scientific plots. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1516–1525.

Dasha Metropolitansky and Jonathan Larson. 2025. Towards effective extraction and evaluation of factual claims. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6995–7027.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of EMNLP*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, and 1 others. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999–22027.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: a dataset for document visual question answering on multiple images. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Rubèn Pérez Tito, Dimosthenis Karatzas, and Ernest Valveny. 2022. Hierarchical multimodal transformers for multi-page docvqa. *ArXiv*, abs/2212.05935.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph Gonzalez, Trevor Darrell, and David M. Chan. 2024. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *International Conference on Learning Representations*.

Yin Wu, Quanyu Long, Jing Li, Jianfei Yu, and Wenya Wang. 2025. Visual-rag: Benchmarking text-to-image retrieval augmented generation for visual knowledge intensive queries. *ArXiv*, abs/2502.16636.

Yuming Yang, Jiang Zhong, Li Jin, Jingwang Huang, Jingpeng Gao, Qing Liu, Yang Bai, Jingyuan Zhang, Rui Jiang, and Kaiwen Wei. 2025. Benchmarking multimodal rag through a chart-based document question-answering generation framework. *arXiv preprint arXiv:2502.14864*.

Jiayu Yao, Shenghua Liu, Yiwei Wang, Lingrui Mei, Baolong Bi, Yuyao Ge, Zhecheng Li, and Xueqi Cheng. 2025. Who is in the spotlight: The hidden bias undermining multimodal retrieval-augmented generation. *ArXiv*, abs/2506.11063.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *ArXiv*, abs/2410.10594.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Annual Meeting of the Association for Computational Linguistics*.

A Dataset Statistics and Implementation Details

Dataset statistics. Table 7 summarizes domains, filtering counts, image resolutions, and text length statistics.

Computing Infrastructure. All open-source VLM inference is run on NVIDIA H100 (80GB) GPUs provided by an academic compute cluster. Qwen2.5-VL-72B uses 4 H100s per inference pass via tensor parallelism; LLaVA-v1.6-34B uses 2 H100s; Phi-4, Pixtral-12B-2409, and InternVL3-8B each use a single H100. Claude 3.5 Sonnet is accessed via API for both the evaluated baseline and the MM-RAGChecker judge stages. The total local inference budget across the eight datasets, six models, three retrieval modes (GO/GPD/DO), and 12 prompt designs is approximately **320** H100 GPU-hours.

B Ambiguity Process

B.1 Ambiguity Definition and Annotation

We curate a 200-question subset to study ambiguity in multimodal RAG.

Definition. A question is **AMBIGUOUS** if a reasonable reader cannot determine a *single* correct answer without additional constraints (time, location, unit, object, visibility in the image, etc.); otherwise it is **CLEAR**. For image-conditioned queries, we additionally mark as ambiguous when key visual evidence is occluded or blurred.

Detection Pipeline. We adopt a two-stage procedure:

1. **LLM-based pre-filter.** We prompt an assessor model to output `CLEAR` or `AMBIGUOUS` together with a short rationale (30 words). We use separate prompts for text-only and image-conditioned questions (see Appendix B.3).

2. **Human adjudication.** Two annotators independently review the flagged items and resolve disagreements. We achieve substantial agreement (Cohen's $\kappa = 0.74$).

Rewriting and Answer Normalization. For each ambiguous item we:

- *Evidence-grounded answer rewrite*: given retrieved evidence and the original answer, an LLM produces a concise answer or returns "Evidence inconclusive." (Appendix B.5).

- *Query disambiguation*: another prompt rewrites the question into a precise version that can be answered solely from the evidence (Appendix B.6).

The final ambiguity split (200 queries) are used in Appendix. B.7 to isolate the impact of ambiguity on accuracy, claim recall, and Hallucination Rate. To construct the matched clear set, we randomly sample CLEAR questions from the same datasets while roughly matching their answer type and length distribution.

B.2 Error Patterns Under Ambiguity

On the 200-query ambiguity subset, two dominant patterns emerge:

1. **Single-interpretation bias**: the model commits to one reasonable sense, lowering Claim Recall and short-answer accuracy.

2. **Interpretation conflation**: retrieval returns evidence for multiple senses; the generator merges them, producing composite neutral claims and higher Hallucination Rates.

These results motivate an *ambiguity detection & clarification* step before final answer generation.

Table 7: Overview of the eight sources in MRAG-Suite, covering domains, filtering counts, image resolution, and question/answer lengths. The suite spans open-web images, charts/plots, scanned documents, slides, and scholarly PDFs to enable cross-domain visual RAG evaluation.Max Q / Max A / Avg Q / Avg A are max length average length of the question and answer.

Dataset	Domain	Two-way filter	Avg. Img (px)	Max Q	Max A	Avg Q	Avg A	Ambig.#
MRAG-Bench	Natural photos (animals, objects)	845/1353	803×658	20	9	8	2	753/1353
Chart-MRAG	Business / science charts	2720/4738	466×576	74	81	7	20.7	4093/4738
Visual-RAG	Fine-grained natural images	298/378	429×430	39	34	19	5.8	309/378
WebQA	Open-web images & captions	19480/21382	792×641	70	99	17.5	12.5	9730/21382
VisRAG-ArXiv	Scholarly PDF pages	709/816	1790×1278	40	21	22	12	46/816
VisRAG-Plot	Technical plots	794/863	1098×684	45	4	15	8	9/863
VisRAG-Slide	Slide decks (mixed media)	525/556	1026×734	39	15	14	10	101/556
VisRAG-Doc	Scanned manuals / forms	562/591	1845×2244	24	37	20	4	6/591

B.3 Ambiguity Detection (Image-Conditioned)

DETECT AMBIGUOUS QUERY WITH IMAGE

You are an expert multimodal question-quality assessor.
Your goal is to decide **from the question plus the attached image**(which may be blurred or obscured) whether it unambiguously identifies exactly one correct answer.

Ignore any answer or multiple-choice options you see below; pretend you do **not know whether they are correct.**

Special considerations when an image is provided:
• If the image is too blurred or key details are not visible → "AMBIGUOUS".
• If the question refers to visual features not visible in the image → "AMBIGUOUS".
• Otherwise, apply the rule that the question lacks a needed constraint (time,
location, unit, object, etc.) →"AMBIGUOUS".
If reasonable readers would arrive at one and only one answer → "CLEAR".
Return **only** this JSON object:
{
"status": "CLEAR" | "AMBIGUOUS",
"explanation": "<≤30 words why>"
}

B.4 Ambiguity Detection (Text-Only)

DETECT AMBIGUOUS QUERY WITHOUT IMAGE

You are an expert question-quality assessor.
Your goal is to decide **from the question alone** whether it unambiguously identifies exactly one correct answer.

Ignore any answer or multiple-choice options you see below; pretend you do **not know whether they are correct.**

If the question lacks a needed constraint (time, location, unit, object, etc.) →
"AMBIGUOUS".
If reasonable readers would arrive at one and only one answer → "CLEAR".

Return **only** this JSON object:
{
"status": "CLEAR" | "AMBIGUOUS",
"explanation": "<≤30 words why>"
}
Question: "{q_esc}"
The following answer is provided for reference only.
Do NOT rely on it when deciding ambiguity.
Answer: "{a_esc}"

B.5 Evidence-Based Answer Rewrite

REWRITE ANSWER

You are a meticulous answer-generation expert.
You will first examine the "Retrieved Evidence" below, then produce a concise,unambiguous answer strictly supported by that evidence.
Follow these steps:
1. **Identify Evidence**
• Quote or paraphrase the specific sentence(s)/data that directly relate to the question.
• If evidence pieces contradict or only partially address the question, label them clearly(e.g., "Conflicting Evidence A / B") and note the conflict.

2. **Answer the Question**
• **If the evidence is clear and consistent**, provide the answer in ≤ 2 sentences.
• **If the evidence is related but insufficient or conflicting**, output exactly: Evidence inconclusive.
Then add one short sentence (< 20 words) explaining why it is inconclusive(e.g., "figures cover different years").
• **If no relevant evidence exists**, output exactly: Insufficient evidence to answer.

3. **Do NOT speculate** beyond the given evidence. Never invent facts.

Question:
{question}
Original Answer (for reference):
{orig_answer}
Retrieved Evidence (bullet-list each item):
{evidence}

Now, produce your evidence-based answer following the instructions above.

B.6 Ambiguous Query Rewrite

REWRITE QUERY

You are an expert at rewriting ambiguous queries into precise, unambiguous questions THAT CAN BE ANSWERED *SOLELY* FROM THE EVIDENCE BELOW.
The evidence includes:
- An explanation of why the query was flagged as ambiguous.
- All relevant text or image-based evidence needed for answering.
EVIDENCE:
{evidence}
REQUIREMENTS (must all be satisfied):
• Embed ANY explicit time-stamp, location, unit and source that the evidence uses.
• If evidence gives a percentage, ask for that percentage (not an absolute count) and vice-versa.
• For image items, phrase the question so that the answer is inferable from THE PROVIDED IMAGE alone.
• Avoid superlatives/comparatives unless the evidence contains a unique maximum/minimum.
• Produce ONLY the rewritten question text on one line – no bullets, no "Rewritten Question:" label.

1. The new question must NOT contain phrases like "according to the evidence / image / chart / passage".
2. Do NOT refer to "the following image" or "the study" unless you embed the study's full name, year, and author in your question text.
3. If the answer comes from quantitative data, embed the dataset's year OR the exact figure you expect, so only one answer is possible.
4. Avoid vague temporal words ("recent", "modern", "in the 2000s"); give exact years.
5. The rewritten text must be **just the question**, one sentence if possible.

ORIGINAL Q → "{raw_question}"
ORIGINAL A → "{raw_answer}"
REWRITE:

Table 8: Ambiguity subset construction statistics. We report candidate counts, items flagged as ambiguous by the two-stage pipeline, and the final curated set (200).

	Candidates	Marked Ambiguous	Final Set
Text-only	26,120	13,823	184
Image-conditioned	4,557	1,224	16
Total	30,677	15,047	200

B.7 Ambiguity Subset Results

On the 200-query ambiguous subset, short-answer EM drops by **5.9** points compared to clear queries, while Hallucination Rate increases by **9.5** points. Claim Recall and Context Precision also decrease, indicating single-interpretation bias or interpretation conflation; see Appendix. B.2 for qualitative patterns. Table 9 reports the aggregate metrics.

Table 9: Results on the 200-query ambiguity subset versus a matched clear set for **Claude 3.5** on all datasets. We report EM/Accuracy, Hallucination Rate (lower is better), Claim Recall, and Context Precision, isolating the effect of query underspecification on retrieval and grounding.

	EM/Acc ↑	Halluc. ↓	Claim Recall ↑	Ctx.Prec. ↑
Clear (200 random)	23.5	49.2	12.4	14.1
Ambiguous (200)	17.6	58.7	9.1	12.5
Δ (Ambig − Clear)	−5.9	+9.5	−3.3	−1.6

C Retrieval System Settings

Dataset	Text Retriever	Image Retriever
WebQA	SBERT (mpnet)	CLIP ViT-B/32
Chart-MRAG	DPR (ctx-enc)	CLIP ViT-B/32
MRAG	—	CLIP ViT-B/32 (img+txt avg)
VisRAG	VisRAG-Ret	—
Visual-RAG	—	CLIP ViT-B/32

Table 10: Backbones used in the hybrid retriever for each sub-corpus. Images are indexed with CLIP variants; text uses DPR/SBERT; VisRAG documents use a document-image retriever.

D 12 Prompt Experiment Settings

Nickname	init_style	ex_style	Example template
expert1	expert	ex1	Q: {q} A: {ans}
expert2	expert	ex3	Question + brief Explanation + Final Answer
expert3	expert	ex6	Dialog style: User: {q} / Assistant: {ans}
plain1	plain	ex1	Q: {q} A: {ans}
plain2	plain	ex3	Question + brief Explanation + Final Answer
plain3	plain	ex6	Dialog style: User: {q} / Assistant: {ans}

Table 12: Mapping from prompt nicknames to template families used in the 12-way prompt sweep. Variations differ in initialization style and exemplar formatting.

E Human Validation of MM-RAGChecker

To assess the reliability of our automatic diagnostics, we ran a small human study on a subset of MRAG-Suite. Annotators are graduate students in our lab; no separate compensation was provided. To verify that our automatic claim segmentation and entailment pipeline is reliable, we conduct a human study on 200 QA instances randomly sampled from the test split of MRAG-Bench

Layer	Arg name	Values in this paper	Template / Semantics
System Prompt	`init_style`	`plain`	`plain`: "Please read the following question and retrieve the relevant document(s)/image(s) to produce the best possible answer."
		`expert`	`expert`: "You are an expert in <DOMAIN>. Analyze the question carefully, think step-by-step about which sources to use, and present your answer clearly at the end."
Few-shot Demo	`example_style`	`ex1`, `ex3`, `ex6`	Three fixed 0-shot style exemplars (chosen from the pool). Content differs in wording/format hints.
	`max_examples`	0 or 5	Whether to insert 5 illustrative QA pairs (kept at 0 in our sweep).
Reasoning Instruction		`direct`	`direct`: "Question: ... Answer:" (no mandated reasoning steps).
		`reasoning_process`	`retrieve_then_reason`: "First outline retrieval/reasoning, then give final answer."
Answer Format	`answer_format`	`free`	Free-form text (no forced JSON or letter).
Context Packaging	`context_order`	`img_first`	Images precede text in the prompt (fixed in our runs).
	`include_doc_ids`	0 / 1	Whether to show doc/image IDs (0 in our runs).

Table 11: Prompt dimensions and concrete templates used in our 12-way sweep.

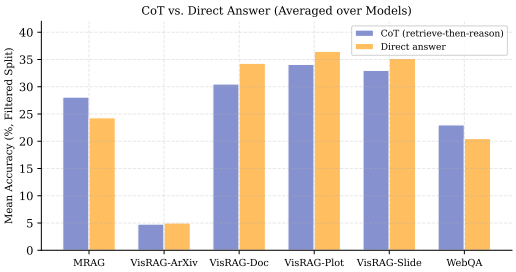

Figure 6: CoT vs. direct answer accuracy averaged over all evaluated models on each MRAG-Suite domain. Overall, CoT provides only modest gains and can even hurt performance on some domains, suggesting that adding long chains of thought does not automatically resolve multimodal reasoning and retrieval noise.

and Chart-MRAG. Two graduate-level annotators with background in NLP and data science independently annotate (1) atomic claim boundaries in the model answers and (2) entailment labels (ENTAILMENT / CONTRADICTION / NEUTRAL) for each claim–evidence pair, given the retrieved multimodal context

For **Stage 1 (claim segmentation)**, annotators were given the model answer and instructed to split it into a small number of minimal statements that can each be judged as true or false. They were asked to split only when a sentence clearly contained multiple independent facts (e.g., joined by "and", "or", ";") or discussed clearly different entities, and to avoid over-segmenting into tiny fragments. We then compared the human segments against the output of our automatic splitter and measured over-splitting and under-splitting rates.

For **Stage 2 (entailment judgement)**, the same annotators received individual (claim, evidence) pairs, where the evidence was either a text passage or an image/slide from the retrieved context. They labeled each pair with one of three categories: ENTAILMENT (the evidence alone is sufficient to support the claim), CONTRADICTION (the evidence clearly conflicts with the claim), or NEUTRAL (the evidence is insufficient or irrelevant).

Figure 7 provides a heatmap overview.

End-to-end error propagation. To assess whether Stage 1 errors compound through Stage 3, we additionally annotated 50 complete questions (not just claim–evidence pairs) from MRAG-Bench and Chart-MRAG. For each question we independently labeled the final question-level hallucination verdict and compared it with MM-RAGChecker's aggregated output. Question-level agreement is 82.0% (κ=0.71), slightly above the claim-level figure, indicating that aggregation does not amplify individual claim errors. When MM-RAGChecker

p_reason	Reasoning block (excerpt)	COT line?
direct	Answer directly without explaining your reasoning.	No
retrieve_then_reason	First list which retrieved snippets/images you will use (by Doc/Image IDs). Then reason step by step using them, and finally give the answer.	Yes
structured	Follow this structure: 1) Question understanding 2) Relevant sources (IDs or short quotes) 3) Reasoning 4) Final Answer	Yes
plan_execute	Outline a short plan first (bullet points). After the plan, execute it and derive the final answer.	Yes
verify	Propose an initial answer. Then self-check it briefly. If it passes, output the final answer prefixed with "FINAL:".	Yes
none	(no additional block)	No

Table 13: Reasoning patterns injected into prompts. Variants differ in whether to expose chain-of-thought–like structure and explicit self-verification before the final answer.

errs at the question level, two patterns dominate: (1) claims requiring world knowledge beyond retrieved evidence ($\sim$60% of errors), where the judge over-predicts NEUTRAL; (2) compound claims incompletely decomposed in Stage 1 ($\sim$40%), where the aggregation propagates the segmentation error.

Dataset	Model	Recall (img/txt)	Precision (img/txt)	F1 (img/txt)	Halluc. (img/txt)	Faith. (img/txt)	Self-know. (img/txt)	Claim R. (img/txt)	Ctx.Prec. (img/txt)
Chart-MRAG	Qwen2.5-VL-72B	12.1/12.9	7.9/8.1	**6.3**/7.2	57.3/**58.9**	54.9/7.8	8.3/7.7	8.9/14.7	85.2/**31.8**
	LLaVA-v1.6-34B	7.6/10.1	0.6/18.9	1.1/**13.2**	57.4/66.8	50.5/8.8	3.6/10.6	2.0/12.8	86.1/22.3
	Claude 3.5	13.3/13.2	0.8/2.1	1.4/3.4	**24.9**/87.3	65.1/10.6	0.0/2.1	75.0/10.0	**90.0**/15.0
MRAG-Bench	Qwen2.5-VL-72B	16.3	1.2	2.2	**45.2**	62.6	12.4	11.3	72.1
	LLaVA-v1.6-34B	10.9	1.0	1.8	57.8	54.8	3.6	8.5	**87.4**
	Claude 3.5	53.3	4.1	**7.6**	48.2	66.3	5.7	11.0	79.6
VisRAG-ArXiv	Qwen2.5-VL-72B	23.7	1.9	**3.5**	**30.3**	67.2	17.2	23.2	61.0
	LLaVA-v1.6-34B	19.7	1.7	3.1	37.0	65.5	12.3	19.4	67.4
	Claude 3.5	17.0	1.6	2.9	41.4	64.4	9.1	16.9	**71.6**
VisRAG-Doc	Qwen2.5-VL-72B	16.8	1.5	**2.8**	**45.0**	57.5	8.0	11.1	74.3
	LLaVA-v1.6-34B	11.0	0.9	1.6	54.3	56.3	4.2	5.6	**82.3**
	Claude 3.5	12.1	1.2	2.2	51.6	57.7	4.7	7.3	80.4
VisRAG-Plot	Qwen2.5-VL-72B	17.7	1.9	**3.4**	**42.1**	65.5	9.6	16.4	75.1
	LLaVA-v1.6-34B	9.5	0.8	1.5	60.9	50.1	3.2	1.9	**88.6**
	Claude 3.5	17.1	1.5	2.8	46.7	62.8	8.3	13.5	75.9
VisRAG-Slide	Qwen2.5-VL-72B	8.7	0.6	**1.1**	**52.7**	47.7	4.4	3.2	85.9
	LLaVA-v1.6-34B	6.3	0.5	0.9	65.8	46.3	1.2	1.3	**92.2**
	Claude 3.5	8.2	0.5	0.9	55.8	51.3	2.1	1.0	89.6
Visual-RAG	Qwen2.5-VL-72B	6.5	0.5	1.0	**74.0**	42.3	2.7	5.2	11.0
	LLaVA-v1.6-34B	5.3	0.4	0.8	76.8	44.1	3.2	4.5	12.0
	Claude 3.5	6.0	0.6	**1.1**	85.1	39.5	1.8	6.1	**14.5**
WebQA	Qwen2.5-VL-72B	8.8/8.1	0.8/16.1	1.5/**10.8**	62.8/**70.4**	52.5/7.9	6.5/3.6	5.9/13.5	89.7/**30.4**
	LLaVA-v1.6-34B	6.5/7.1	0.7/15.8	1.3/9.8	64.5/70.8	50.2/6.2	2.5/4.1	1.1/12.2	**91.4**/28.6
	Claude 3.5	9.0/7.9	1.1/1.8	**1.7**/3.7	**26.6**/98.8	62.7/10.7	0.3/0.6	72.0/8.8	87.4/13.6

Table 14: Diagnostic metrics on the Filtered split (image/text pairs where applicable). **Best** and Second-best are highlighted per row and split. All metrics are percentage points except F1. Macro-averaged hallucination across the six single-modality datasets: Qwen2.5-VL-72B 48.2, LLaVA-v1.6-34B 58.8, Claude 3.5 54.8, confirming consistently high hallucination rates even for the strongest model.

Metric	MRAG-Bench	Chart-MRAG	Overall
Boundary Precision	88.2	86.5	87.4
Boundary Recall	85.7	84.9	85.3
Boundary F1	86.9	85.7	86.3
Claim Recall	92.1	90.4	91.3
Over-seg Rate (%)	14.0	16.5	15.2
Under-seg Rate (%)	8.5	9.2	8.9

Table 15: Human evaluation of claim segmentation quality.

Setting	Accuracy (%)	Macro-F1	Cohen's κ
MRAG-Bench	81.5	0.78	0.74
Chart-MRAG	79.2	0.76	0.71
Overall	80.4	0.77	0.73

Table 16: Agreement between our automatic entailment checker and human annotators.

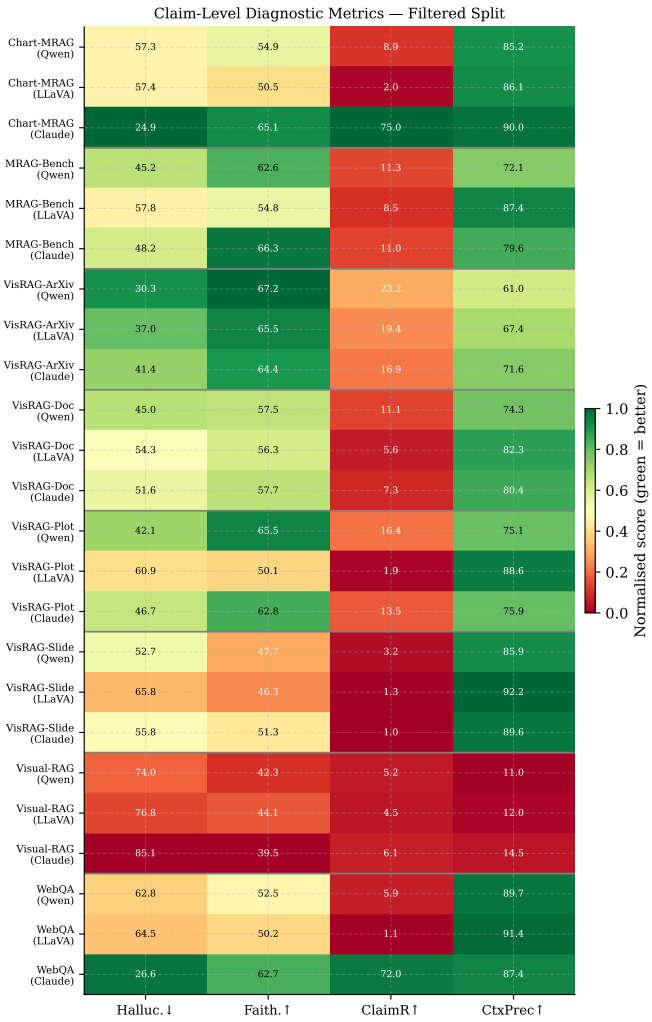

Figure 7: Claim-level diagnostic heatmap on the Filtered split. Rows = Dataset × Model; columns = Hallucination (lower is better), Faithfulness, Claim Recall, Context Precision. Green = better performance. Macro-averaged hallucination: Qwen2.5-VL-72B 48.2%, LLaVA-v1.6-34B 58.8%, Claude 3.5 54.8%.

F Cross-Modality Metric Definitions

Let $\mathcal{C}$ be the set of generated atomic claims and L_i the verdict from MM-RAGChecker for claim $i \in \mathcal{C}$. Let $\mathcal{I}$ and $\mathcal{T}$ denote the retrieved image and text sets; we write $v \models i$ if evidence v entails claim i.

$$s_i^{\text{img}} = \mathbf{1}\{\, \exists v \in \mathcal{I} : v \models i \,\}, \tag{6}$$

$$s_i^{\text{txt}} = \mathbf{1}\{\, \exists t \in \mathcal{T} : t \models i \,\}. \tag{7}$$

Claim-level recall/precision for $m \in \{\text{img}, \text{txt}\}$:

$$\text{CR}_m = \frac{\sum_{i \in \mathcal{C}} \mathbf{1}[L_i = \textsc{Entailment}]\, \mathbf{1}[s_i^m = 1]}{\sum_{i \in \mathcal{C}} \mathbf{1}[L_i = \textsc{Entailment}]}, \tag{8}$$

$$\text{CP}_m = \frac{\sum_{i \in \mathcal{C}} \mathbf{1}[s_i^m = 1]}{\sum_{i \in \mathcal{C}} \mathbf{1}[\, s_i^{\text{img}} = 1 \vee s_i^{\text{txt}} = 1 \,]}. \tag{9}$$

Cross-modality gaps:

$$\Delta\text{CR} = \text{CR}_{\text{img}} - \text{CR}_{\text{txt}}, \tag{10}$$

$$\Delta\text{CP} = \text{CP}_{\text{img}} - \text{CP}_{\text{txt}}. \tag{11}$$

Coverage (restricting to top-k per modality):

$$\text{VIS}-\text{Hit@}k = \frac{1}{k} \sum_{i=1}^{k} \mathbb{I}\big[\, I_i \vDash \exists c \,\big], \tag{12}$$

$$\text{TXT}-\text{MissRate} = 1 - \frac{1}{k} \sum_{i=1}^{k} \mathbb{I}\big[\, t_i \vDash \exists c \,\big]. \tag{13}$$

Cross-modal agreement:

$$\text{CMA} = \frac{\sum_{i \in \mathcal{C}} \mathbf{1}[\, s_i^{\text{img}} = 1 \wedge s_i^{\text{txt}} = 1 \,]}{\sum_{i \in \mathcal{C}} \mathbf{1}[\, s_i^{\text{img}} = 1 \vee s_i^{\text{txt}} = 1 \,]}. \tag{14}$$

Hallucination attribution:

$$\text{V}-\text{HR} = \frac{\#\{\, i : \; s_i^{\text{img}} = 1 \wedge L_i \neq \textsc{Entailment} \,\}}{\#\{\, i : \; L_i \neq \textsc{Entailment} \,\}}, \tag{15}$$

$$\text{D}-\text{HR} = \frac{\#\{\, i : \; s_i^{\text{txt}} = 1 \wedge L_i \neq \textsc{Entailment} \,\}}{\#\{\, i : \; L_i \neq \textsc{Entailment} \,\}}. \tag{16}$$

Unless noted, all values are percentage points on the Filtered split, and we use $k = 5$ per modality at test time.

G Distractor Results Overview

In this section, we provide the full per-dataset and per-model scores under the three retrieval modes `distractors_only` (DO), `gt_plus_distractors` (GPD), and `gt_only` (GO) discussed in Section 5. The main paper only reports summary trends; Table 17 below contains the complete EM/Accuracy results.

H Difficulty-Based Filter Scoring Rubrics

We score each item on three binary axes using GPT-4o:

Dataset	Qwen2.5-VL-72B			LLaVA-v1.6-34B			Claude 3.5			Phi-4			InternVL3-8B			Pixtral-12B-2409		
	DO	GPD	GO	DO	GPD	GO	DO	GPD	GO	DO	GPD	GO	DO	GPD	GO	DO	GPD	GO
Chart-MRAG	22.7	**25.4**‡	**26.8**†	10.3	18.1	13.9	8.2	15.5	9.7	15.9	19.8	17.4	23.3	19.5	21.7	14.8	14.1	16.6
MRAG-Bench	30.9	29.6	**29.8**†	17.8	16.4	20.7	26.9	24.1	29.7	2.1	2.4	27.9	26.2	21.3	28.5	**33.6**	**29.8**‡	28.2
VisRAG-ArXiv	4.4	4.1	4.3	4.3	**11.7**‡	**8.5**†	**7.3**	4.1	5.0	1.7	4.4	5.5	6.8	5.1	5.2	3.2	1.4	1.8
VisRAG-Doc	1.2	29.1	31.3	12.9	20.5	20.2	2.4	26.6	25.9	2.3	21.6	28.3	4.7	**38.4**‡	**38.1**†	0.6	26.5	27.1
VisRAG-Plot	5.9	33.8	41.4	8.4	12.9	12.6	12.7	25.7	29.3	6.2	6.7	32.1	4.9	28.3	28.4	0.8	**51.2**‡	**52.0**†
VisRAG-Slide	1.6	18.7	18.1	1.9	8.4	8.2	4.3	**18.9**‡	**18.8**†	1.2	24.6	15.3	0.9	17.5	15.7	0.6	15.3	15.6
Visual-RAG	14.8	26.9	**27.3**†	13.6	18.7	18.1	15.5	25.4	26.5	15.1	21.4	20.2	17.6	**23.1**‡	23.2	16.5	23.6	18.9
WebQA	17.2	**15.8**	15.4	12.3	12.6	13.4	9.9	14.3	**16.4**†	17.5	7.3	12.8	20.1	20.8	17.4	**25.6**	25.3	0.9

Table 17: Scores (EM–exact match preferred; Accuracy used when EM is unavailable) under the three retrieval modes: DO = *distractors_only*, GPD = *gt_plus_distractors*, GO = *gt_only*. † marks the best in GO; ‡ marks the best in GPD; * marks the smallest $|\mathrm{GO} - \mathrm{GPD}|$ (most robust to distractors).

Dataset	Model	Distractor	Recall (img/txt)	Precision (img/txt)	F1 (img/txt)	Halluc. (img/txt)	Faith. (img/txt)	Self-know. (img/txt)	Claim R. (img/txt)	Ctx.Prec. (img/txt)
VisRAG-ArXiv	Qwen2.5-VL-72B	distractors only	21.8	6.7	7.9	66.2	23.5	5.9	20.4	31.0
		gt_plus_distractors	29.6	8.9	10.8	60.1	28.9	7.1	24.1	33.5
		gt_only	33.4	10.8	12.7	55.3	32.1	8.2	25.9	35.8
	LLaVA-v1.6-34B	distractors only	18.4	5.9	6.8	69.4	20.1	4.8	18.0	28.2
		gt_plus_distractors	24.7	7.6	9.1	63.5	24.9	5.9	21.6	30.4
		gt_only	28.9	9.2	10.9	58.2	29.6	6.8	23.7	33.1
	Claude 3.5	distractors only	30.4	9.2	10.1	61.4	28.0	7.6	29.9	40.0
		gt_plus_distractors	40.7	10.6	14.3	56.0	37.0	7.0	25.6	33.0
		gt_only	35.9	8.6	11.3	57.7	36.5	4.7	27.6	39.0
	Phi-4	distractors only	22.5	15.3	13.7	50.8	20.6	12.9	22.2	33.8
		gt_plus_distractors	34.4	24.0	23.5	45.3	22.3	16.3	26.0	36.0
		gt_only	37.8	27.2	26.6	41.1	24.7	18.5	27.9	37.9
	Pixtral-12B-2409	distractors only	17.6	7.1	6.3	70.5	18.7	5.1	16.0	27.6
		gt_plus_distractors	25.9	10.4	9.7	63.0	23.3	7.4	18.9	29.5
		gt_only	30.5	12.8	12.0	58.9	25.6	8.3	21.5	31.2
	InternVL3-8B	distractors only	26.5	14.6	12.6	59.3	24.0	10.8	28.8	37.0
		gt_plus_distractors	39.2	15.7	16.5	54.4	30.0	10.6	29.0	38.0
		gt_only	40.3	14.6	17.6	48.2	27.4	9.5	24.4	35.0
VisRAG-Doc	Qwen2.5-VL-72B	distractors only	10.8	4.0	3.6	57.9	13.4	2.6	21.0	23.5
		gt_plus_distractors	46.7	31.9	27.2	28.1	22.7	20.4	25.9	27.0
		gt_only	52.9	36.2	31.0	24.4	24.5	22.6	24.1	25.1
	LLaVA-v1.6-34B	distractors only	9.4	3.7	3.2	60.1	12.1	2.3	19.5	22.0
		gt_plus_distractors	39.8	26.1	22.0	33.9	20.1	16.4	23.8	25.6
		gt_only	44.0	28.9	24.7	30.2	22.4	18.5	22.9	24.2
	Claude 3.5	distractors only	12.0	4.6	4.2	38.2	16.9	3.0	24.5	26.0
		gt_plus_distractors	75.5	35.2	38.2	41.4	32.2	25.4	27.7	31.0
		gt_only	68.8	33.6	33.7	41.5	30.7	23.8	26.5	28.0
	Phi-4	distractors only	18.1	12.2	11.0	27.6	7.5	10.5	20.8	22.7
		gt_plus_distractors	45.3	37.8	37.0	15.2	4.7	32.3	24.5	26.0
		gt_only	50.6	41.9	41.1	13.0	3.9	35.2	23.6	24.8
	Pixtral-12B-2409	distractors only	7.8	3.5	3.0	61.3	10.6	2.1	18.8	21.3
		gt_plus_distractors	34.4	25.6	20.4	36.8	17.7	15.1	22.1	23.9
		gt_only	39.1	29.7	24.1	32.6	19.8	17.2	21.5	22.8
	InternVL3-8B	distractors only	8.0	4.1	3.4	55.2	12.8	3.1	22.5	25.0
		gt_plus_distractors	68.3	65.2	26.5	15.8	25.6	45.5	28.0	26.5
		gt_only	60.3	77.8	57.8	17.2	27.3	54.5	21.5	23.0
MRAG-Bench	Qwen2.5-VL-72B	distractors only	22.6	6.2	5.1	79.4	14.3	4.0	7.3	4.5
		gt_plus_distractors	28.4	7.9	6.7	75.8	16.2	4.7	8.9	5.2
		gt_only	33.7	9.5	8.4	71.1	18.6	5.3	10.4	6.0
	LLaVA-v1.6-34B	distractors only	21.1	5.3	4.1	82.4	13.2	3.5	8.2	4.2
		gt_plus_distractors	27.8	4.8	3.1	77.1	17.6	4.3	10.1	6.0
		gt_only	31.9	6.7	4.6	73.4	19.1	4.8	9.8	5.0
	Claude 3.5	distractors only	42.2	3.9	4.6	73.5	19.8	3.8	9.2	5.5
		gt_plus_distractors	47.2	2.8	3.9	80.5	16.1	2.4	7.7	6.0
		gt_only	47.7	4.2	5.6	78.8	17.1	4.1	5.5	3.0
	Phi-4	distractors only	14.5	12.6	10.8	36.4	8.8	11.9	7.1	4.6
		gt_plus_distractors	18.7	18.3	16.2	31.4	9.3	16.3	8.6	5.5
		gt_only	22.6	21.4	18.7	27.8	10.4	18.5	9.5	6.2
	Pixtral-12B-2409	distractors only	12.8	6.9	5.7	84.6	11.0	5.6	6.4	3.9
		gt_plus_distractors	17.2	8.8	7.2	80.3	12.5	6.4	7.3	4.2
		gt_only	21.0	10.6	8.6	76.1	14.3	7.2	8.4	4.9
	InternVL3-8B	distractors only	35.3	37.6	31.0	54.5	9.3	35.1	5.3	3.0
		gt_plus_distractors	29.0	33.1	25.1	46.7	8.2	32.1	5.3	3.5
		gt_only	31.8	31.6	25.6	62.0	7.5	28.6	6.2	4.0
VisRAG-Plot	Qwen2.5-VL-72B	distractors only	10.1	3.9	2.6	61.7	12.2	2.7	9.4	10.2
		gt_plus_distractors	17.6	7.2	5.8	58.5	11.6	5.4	9.7	9.6
		gt_only	23.8	10.7	8.2	56.0	12.9	7.7	8.6	8.8
	LLaVA-v1.6-34B	distractors only	7.9	3.1	1.9	64.9	10.7	2.1	9.1	10.4
		gt_plus_distractors	13.4	6.1	4.0	61.2	11.5	3.9	9.5	9.8
		gt_only	19.6	8.3	5.7	58.9	12.1	5.5	8.9	9.3
	Claude 3.5	distractors only	8.0	4.1	2.1	55.1	10.1	2.8	10.5	11.0
		gt_plus_distractors	19.8	9.4	6.6	77.8	13.4	8.9	9.0	9.0
		gt_only	26.5	11.6	9.4	75.2	12.7	11.1	7.5	8.0
	Phi-4	distractors only	11.3	4.7	1.8	83.7	10.3	4.0	10.0	10.0
		gt_plus_distractors	16.8	8.9	3.7	79.5	11.2	6.8	9.6	9.7
		gt_only	22.4	11.5	5.2	76.4	12.0	9.4	8.4	8.7
	Pixtral-12B-2409	distractors only	6.5	2.8	1.1	66.0	9.8	1.7	8.8	9.5
		gt_plus_distractors	12.9	6.4	2.9	62.1	10.5	3.7	9.2	9.6
		gt_only	18.7	8.9	4.6	59.6	11.2	5.2	8.7	9.0
	InternVL3-8B	distractors only	2.3	1.8	0.4	52.3	11.0	1.7	8.5	9.0
		gt_plus_distractors	25.5	15.1	7.9	72.9	9.3	13.8	10.0	10.0
		gt_only	21.5	11.2	3.2	76.5	8.2	10.3	11.1	12.0
VisRAG-Slide	Qwen2.5-VL-72B	distractors only	12.6	5.5	4.8	63.3	14.7	3.8	51.1	56.8
		gt_plus_distractors	41.3	27.2	26.1	36.5	42.2	21.1	46.9	50.6
		gt_only	49.7	34.9	33.5	31.2	48.4	23.2	43.8	48.0
	LLaVA-v1.6-34B	distractors only	10.9	4.9	4.1	66.0	13.3	3.3	49.6	54.3
		gt_plus_distractors	35.6	22.7	21.0	40.2	39.1	18.8	45.1	49.2
		gt_only	42.8	29.6	27.4	35.5	43.7	20.6	41.9	46.6
	Claude 3.5	distractors only	23.1	12.0	11.2	53.3	36.7	9.1	53.3	59.1
		gt_plus_distractors	78.8	30.6	37.3	38.7	45.8	15.4	55.7	61.1
		gt_only	64.2	39.1	36.0	30.1	49.2	18.9	49.2	56.0
	Phi-4	distractors only	19.9	11.7	10.3	41.2	31.6	8.6	50.1	55.6
		gt_plus_distractors	71.8	42.7	44.5	33.7	40.1	23.4	45.9	48.6
		gt_only	60.4	47.9	45.0	29.0	37.7	26.1	43.6	46.2
	Pixtral-12B-2409	distractors only	9.6	5.1	4.2	67.9	12.0	3.4	48.5	53.7
		gt_plus_distractors	29.8	18.9	16.7	45.6	34.7	16.3	44.3	47.5
		gt_only	36.7	24.5	22.1	39.8	38.9	17.9	41.1	45.3
	InternVL3-8B	distractors only	17.2	4.4	5.1	58.7	34.0	3.2	54.1	59.7
		gt_plus_distractors	67.1	42.6	41.9	23.4	51.4	22.5	47.3	51.4
		gt_only	57.8	41.1	38.7	28.8	44.6	23.8	40.8	45.8
WebQA	Qwen2.5-VL-72B	distractors only	14.0/13.1	6.5/5.8	5.0/4.3	69.1/71.2	9.4/8.6	5.0/5.4	12.8/12.0	14.0/13.2
		gt_plus_distractors	18.3/17.2	8.9/8.1	7.3/6.5	66.5/68.7	11.2/10.3	6.3/6.8	12.0/11.3	13.5/12.6
		gt_only	22.4/21.0	11.3/10.2	9.1/8.0	62.8/64.9	13.0/12.0	7.4/8.0	11.2/10.5	12.7/11.8
	LLaVA-v1.6-34B	distractors only	12.6/11.4	5.7/5.0	4.5/3.8	71.6/73.9	8.0/7.2	4.8/5.1	12.1/11.3	13.1/12.1
		gt_plus_distractors	16.8/15.6	8.2/7.4	6.6/5.7	68.7/70.8	10.4/9.5	5.9/6.3	11.3/10.6	12.7/11.8
		gt_only	21.3/19.9	10.5/9.4	8.4/7.3	65.2/67.3	12.6/11.6	7.0/7.6	10.7/10.0	12.0/11.2
	Claude 3.5	distractors only	23.0/21.5	4.7/4.3	5.2/4.4	63.4/86.4	31.9/4.1	3.7/3.5	12.6/4.8	15.0/12.0
		gt_plus_distractors	22.4/20.7	5.9/4.3	5.4/4.7	61.1/87.2	30.9/4.4	4.9/3.3	10.1/7.0	13.0/16.5
		gt_only	26.8/24.1	6.8/5.0	6.6/5.4	57.6/84.3	33.5/5.2	5.5/3.9	11.2/8.1	14.2/17.8
	Phi-4	distractors only	14.2/12.7	22.0/19.1	8.7/7.5	53.4/60.2	9.1/0.7	20.1/18.5	10.9/4.7	13.6/12.3
		gt_plus_distractors	10.4/12.5	25.6/22.3	9.5/11.1	49.7/56.5	8.4/0.5	23.9/22.0	11.8/5.2	14.5/13.0
		gt_only	16.1/14.8	29.0/24.7	12.1/12.4	46.2/53.0	10.2/0.9	26.7/24.1	12.3/5.8	15.9/13.9
	Pixtral-12B-2409	distractors only	10.5/ 9.6	15.3/14.1	6.4/5.7	58.9/65.2	6.9/1.0	13.8/12.6	9.6/3.9	12.0/10.9
		gt_plus_distractors	13.4/12.6	19.6/17.4	8.7/7.5	55.2/61.8	8.0/1.2	17.3/15.5	10.5/4.4	13.2/11.9
		gt_only	17.4/16.0	23.5/20.7	11.3/9.8	51.7/58.1	9.6/1.6	20.4/18.0	11.8/5.0	14.7/13.0
	InternVL3-8B	distractors only	27.9/27.3	13.8/14.6	14.7/12.8	68.1/77.2	13.8/8.8	13.1/11.9	13.0/4.5	14.0/10.5
		gt_plus_distractors	32.0/31.0	15.4/15.3	17.0/15.1	69.0/74.4	14.2/8.4	14.8/12.2	11.2/6.3	14.0/14.0
		gt_only	16.3/16.9	7.0/6.9	5.6/5.1	64.0/63.6	8.9/12.0	6.1/5.3	8.5/6.4	11.5/13.5

Table 18: Full MM-RAGChecker diagnostics across datasets, models, and retrieval modes. For each setting we report per-modality (image/text) Recall, Precision, F1, Hallucination, Faithfulness, Self-Knowledge, Claim Recall, and Context Precision.

Dataset	Model	Distractor	Recall (img/txt)	Precision (img/txt)	F1 (img/txt)	Halluc. (img/txt)	Faith. (img/txt)	Self-know. (img/txt)	Claim R. (img/txt)	Ctx.Prec. (img/txt)
Chart-MRAG	Qwen2.5-VL-72B	distractors only	11.0/10.0	4.2/3.7	2.9/2.4	60.6/62.8	12.8/11.9	2.6/2.9	9.8/9.1	10.6/10.0
		gt_plus_distractors	18.4/17.1	7.6/6.9	6.2/5.4	57.1/59.4	12.2/11.1	5.1/5.6	10.1/9.4	10.0/9.7
		gt_only	24.6/23.1	11.1/10.3	8.6/7.7	54.6/56.9	13.4/12.6	7.4/8.1	9.0/8.3	9.2/8.9
	LLaVA-v1.6-34B	distractors only	8.6/7.6	3.4/2.9	2.2/1.7	63.7/65.8	11.2/10.5	2.0/2.3	9.5/8.8	10.7/10.2
		gt_plus_distractors	14.2/13.0	6.5/5.8	4.3/3.5	60.1/62.3	12.0/11.0	3.8/4.2	9.9/9.2	10.2/9.8
		gt_only	20.7/19.2	8.9/8.1	6.1/5.3	57.8/60.1	12.6/11.8	5.4/6.0	9.2/8.6	9.6/9.1
	Claude 3.5	distractors only	8.6/7.5	4.3/3.9	2.3/1.9	54.0/56.0	10.4/9.8	2.7/3.0	10.8/10.2	11.2/10.7
		gt_plus_distractors	20.7/18.9	9.8/8.9	6.9/6.2	76.6/78.9	13.8/12.7	9.2/9.8	9.2/8.9	9.2/8.8
		gt_only	27.5/25.4	12.2/11.0	9.7/8.9	74.3/76.6	13.1/12.2	11.2/11.9	7.7/7.3	8.2/7.8
	Phi-4	distractors only	12.1/10.6	4.9/4.4	2.0/1.7	82.4/84.7	10.9/9.8	3.7/4.0	10.2/9.7	10.2/9.9
		gt_plus_distractors	17.6/16.0	9.4/8.5	4.1/3.4	78.3/80.6	11.7/10.8	6.5/7.0	9.9/9.3	9.9/9.5
		gt_only	23.3/21.6	12.0/10.9	5.5/4.9	75.1/77.4	12.5/11.6	9.1/9.7	8.7/8.1	8.9/8.5
	Pixtral-12B-2409	distractors only	7.2/6.0	3.1/2.6	1.3/1.0	64.8/66.9	10.3/9.4	1.6/1.9	9.1/8.5	9.8/9.3
		gt_plus_distractors	13.7/12.2	6.7/5.9	3.1/2.7	60.9/63.1	11.0/10.1	3.6/3.9	9.5/8.9	9.9/9.4
		gt_only	19.6/18.0	9.3/8.2	4.9/4.3	58.6/60.8	11.7/10.8	5.1/5.6	9.1/8.5	9.3/8.9
	InternVL3-8B	distractors only	2.6/2.1	1.9/1.6	0.5/0.4	51.1/53.5	11.3/10.7	1.6/1.8	8.7/8.3	9.2/8.8
		gt_plus_distractors	26.5/24.6	15.7/14.4	8.4/7.5	71.8/74.0	9.7/8.9	14.3/15.0	10.1/9.7	10.1/9.8
		gt_only	22.3/20.8	11.6/10.7	3.5/2.9	75.4/77.5	8.5/7.8	10.6/11.1	11.4/10.8	12.1/11.4
Visual-RAG	Qwen2.5-VL-72B	distractors only	18.4	6.3	6.8	42.1	27.2	48.3	4.9	18.2
		gt_plus_distractors	44.5	14.7	16.8	36.3	34.5	53.9	6.8	25.0
		gt_only	41.3	13.2	15.4	37.5	32.8	51.1	6.5	23.3
	LLaVa-v1.6-34B	distractors only	16.1	5.2	5.7	43.7	25.4	45.0	4.4	16.1
		gt_plus_distractors	42.0	12.9	14.7	38.8	30.1	51.6	6.0	22.7
		gt_only	38.8	11.8	13.2	39.5	28.0	50.2	5.7	21.4
	Claude 3.5	distractors only	23.3	7.0	8.4	40.2	28.6	50.3	6.2	20.0
		gt_plus_distractors	52.7	16.5	19.7	35.0	36.7	55.5	7.5	28.8
		gt_only	48.1	15.2	17.4	36.1	34.9	53.2	7.1	26.5
	Phi-4	distractors only	20.7	6.5	7.3	39.0	26.0	46.4	5.3	18.0
		gt_plus_distractors	47.2	15.0	17.6	34.1	32.5	52.9	6.7	25.2
		gt_only	43.6	13.9	16.0	35.3	31.0	50.7	6.4	23.7
	Pixtral-12B-2409	distractors only	15.3	5.8	6.2	44.9	23.5	43.2	4.1	15.3
		gt_plus_distractors	40.8	12.0	13.9	38.2	29.8	49.8	5.8	21.9
		gt_only	37.9	11.1	12.8	39.0	27.7	48.1	5.5	20.4
	InternVL3-8B	distractors only	22.6	8.1	9.2	41.3	30.0	49.0	6.0	19.5
		gt_plus_distractors	50.1	17.2	20.3	33.7	37.8	54.4	7.8	27.1
		gt_only	45.7	15.5	18.1	34.9	35.3	52.0	7.3	25.0

Table 19: Full MM-RAGChecker diagnostics for Chart-MRAG and Visual-RAG across models and retrieval modes.

Criterion	Scoring Rule
Multi-hop requirement	Score 1 if the question requires combining information from $\geq$2 distinct evidence pieces; else 0.
Modality dependency	Score 1 if removing all images would change the correct answer; else 0.
Baseline accuracy	Score 0 if answered correctly by Qwen2.5-VL-72B with no retrieval; else 1.

Table 20: Scoring criteria for the difficulty-based filter. Items with weighted sum below the 10th percentile per domain are removed.