When Image and Text Disagree: Cross-Modal Evidence Conflict in Multimodal Retrieval-Augmented Generation

Jasper Kyle Catapang
[1]Money Forward Inc., Shibaura, Minato-ku, Tokyo, Japan
[2]Tokyo University of Foreign Studies, Asahi-cho, Fuchu-shi, Tokyo, Japan
[1]catapang.j@moneyforward.co.jp

Abstract

This paper introduces the Cross-Modal Conflict Benchmark (CMC-Bench) to evaluate how multimodal retrieval-augmented generation (RAG) systems handle contradicting evidence between retrieved text and images. Using 3,768 instances from ChartQA and MMMU *evaluation* splits, the study benchmarks four open vision-language models (VLMs) across four conflict types (factual, temporal, entity, and granularity) and four evidence conditions: *aligned* (both modalities support the gold answer), *image-correct* (image supports the gold and text contradicts it), *text-correct* (text supports the gold and the image is wrong or swapped), and *both-wrong* (neither modality supports the gold). Key findings reveal that cross-modal disagreement severely degrades performance, with ΔAcc between 0.17 and 0.46 relative to aligned evidence. Results show models often exhibit a "modality lean" rather than reliable arbitration, with text-leaning systems particularly vulnerable when only the image is correct. Furthermore, merging abstention and fabrication into a single "hallucination" score obscures critical behavioral differences; for instance, Qwen3-VL-4B abstains on 31.7% of conflicts, while Gemma-3n-E2B fabricates unsupported answers in 51.9% of conflicts. Multimodal RAG evaluation should explicitly distinguish abstention from fabrication to assess reliability accurately.

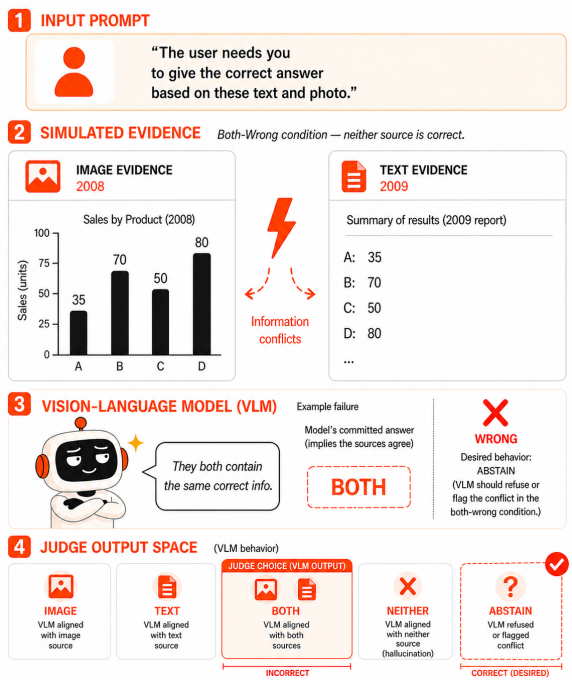

Figure 1: **CMC-Bench overview.** In the *both-wrong* condition, neither retrieved source is correct (image: 2008, text: 2009). The VLM claims both sources agree, receiving a BOTH label from the judge. The LLM-as-judge classifies VLM responses relative to the two source references—it does not receive the gold answer and does not evaluate correctness. Section 4 shows the five-way output space (VLM behavior); ABSTAIN is the desired response in the both-wrong condition.

1 Introduction

Multimodal RAG systems retrieve evidence from diverse modalities—images, text, tables—to ground answer generation (Lewis et al., 2020). Designs typically assume that retrieved evidence is trustworthy and *internally consistent* across modalities. In practice, pipelines can surface image–text pairs that support different answers: a chart may show 64% while a passage claims 41%; an image may reflect 2023 while text describes 2021. Under such *cross-modal evidence conflict*, grounded generation requires *arbitration*—using the modality that is correct for the query and instance, or else refusing when neither channel is adequate. It remains unclear whether VLMs implement such instance-wise adjudication or instead default to modality priors, shallow reconciliation, answers aligned with neither source, or abstention.

Existing multimodal hallucination benchmarks largely do not isolate this failure mode. Single-image suites (e.g., POPE, HaloQuest, M-HalDetect) test fabrication relative to *one* image

Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval, pages 1–10

(Li et al., 2023; Wang et al., 2024; Gunjal et al., 2024). Work on conflicting image–text pairs (Liu et al., 2024b) targets single-turn settings without a retrieval framing. No benchmark systematically studies VLMs when *retrieved* image and text evidence *disagree* on the answer, as in multimodal RAG.

CMC-Bench addresses that gap (Figure 1). It comprises: (1) a taxonomy of four conflict types; (2) controlled instances from ChartQA (Masry et al., 2022) and MMMU (Yue et al., 2024) with dataset-derived passages and wrong-image selection (942 examples, 3,768 instances); (3) four evidence conditions per example; and (4) evaluation of four open VLMs on accuracy, modality-following, modality-preference bias, and explicit separation of *unsupported answers* versus *abstention* under an LLM judge (Section 4.2). Importantly, CMC-Bench constructs conflicts *programmatically* from dataset templates rather than from a live retrieval pipeline; it therefore isolates the generator's conflict-handling ability independently of retrieval quality, and its claims should be interpreted accordingly. Reproducibility materials accompany the benchmark release.

2 Related Work

2.1 Multimodal RAG

Retrieval-augmented generation was introduced for knowledge-intensive NLP (Lewis et al., 2020) and has been extended to multimodal settings. The first Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025) (Kriz and Murray, 2025) featured systems that combine retrieval over text, images, tables, and video. Kangur et al. (2025) present MultiReflect, a multimodal self-reflective RAG pipeline for fact-checking. Drushchak et al. (2025) propose a unified framework for information processing across text, image, table, and video. You et al. (2025) address cross-modal clustering-based retrieval for scalable image captioning. These works assume that retrieved evidence is trustworthy; no study examines inter-evidence conflict when image and text disagree.

2.2 Source Data for Cross-Modal Conflict

Constructing a benchmark for cross-modal evidence conflict requires source data that supplies (image, question, gold answer) triples at scale, permits plausible contradicting text from the same dataset (e.g., same-type wrong answers), and allows a wrong image to be drawn from the same domain. Prior work clusters into several families. Chart and plot QA datasets such as ChartQA (Masry et al., 2022) and PlotQA (Methani et al., 2020) support factual and temporal conflicts and same-type passage pairing. Diagram-heavy suites such as MMMU (Yue et al., 2024) and ScienceQA (Lu et al., 2022) suit entity- and granularity-style conflicts. General VQA (Goyal et al., 2017) offers scale but heterogeneous answers, which complicates controlled same-type contradictions. Table- and document-centric QA differs from the image-plus-passage setting in how layout and text are mixed. None of these resources were built for cross-modal conflict, and the sources used in CMC-Bench are motivated in Section 3.2.

2.3 Multimodal Hallucination

Object and attribute hallucination in vision-language models has been evaluated by benchmarks such as POPE (Li et al., 2023), HaloQuest (Wang et al., 2024), M-HalDetect (Gunjal et al., 2024), FREAK (Yin et al., 2026), and evidential conflict detection (Huang et al., 2025). Surveys (Liu et al., 2024a) summarize the landscape. Liu et al. (2024b) study intrinsic vision-language hallucination in a single-turn setting with conflicting image–text pairs, but not in a retrieval pipeline. MVI-Bench (Chen et al., 2025) evaluates robustness to misleading visual inputs and adversarial framing, not to conflicting *retrieved* evidence. No prior benchmark evaluates hallucination and model behavior when *retrieved* image and text evidence contradict each other.

3 CMC-Bench

3.1 Conflict Taxonomy

The taxonomy comprises four conflict types that can arise when image and text evidence are presented together in a RAG setting.

Factual contradiction. The image conveys a value or fact A while the text states a different value or fact B (e.g., a bar chart indicates 64% for a category whereas the text states "41% of respondents selected this option").

Temporal mismatch. The image depicts or refers to period X (e.g., a chart titled "2023 Sales") while the text describes period Y (e.g., "In 2021, revenue increased"). The model must recognize the temporal mismatch rather than fuse inconsistent

time references.

Entity confusion. The image depicts entity A (e.g., a diagram of process P) while the text describes entity B (a visually similar or confusable process P'), as commonly occurs in science and business diagrams when labels or structure are swapped.

Granularity conflict. The image presents a specific case or instance while the text states a general rule, or vice versa (e.g., the chart shows data for one country while the text claims "Across all regions, the trend is..." without the image supporting the generalization).

3.2 Construction Pipeline

Source datasets. Following the landscape survey in Section 2, ChartQA (Masry et al., 2022) (HuggingFaceM4/ChartQA) serves as the primary source and MMMU (Yue et al., 2024) (MMMU/MMMU) as the supplement. Only *evaluation* splits are retained: ChartQA validation and test (no training data) and MMMU validation and test per subject. ChartQA pairs charts with verified answers that are mostly numeric or temporal, which matches factual and temporal conflict construction. MMMU supplies expert-level diagram Q&A, with examples drawn from six Hub subject configurations (Physics, History, Psychology, Computer Science, Art, Economics), under per-subject quotas to ensure subjects contribute evenly (counts in Section 3.3). The MMMU visual field `image_1` is treated as the image input. Each source yields (image, question, gold) triples and admits wrong-image selection from the same corpus under the rules below.

Passage templates and conflict-type routing. Aligned and contradicting passages follow fixed templates (ChartQA: "According to the chart / source, ..."; MMMU: "According to the figure / source, ..."), with the aligned line stating the gold answer. Conflict type is fixed by simple rules: for ChartQA, temporal if the gold is a four-digit year in [1900, 2100] and factual otherwise; for MMMU, entity if the resolved gold is short non-numeric text (under 80 characters) and granularity if numeric or longer. On MMMU multiple-choice rows, letter answers in `answer` are expanded via the `options` column for gold text and typing.

Contradicting text and wrong-image sampling. Contradicting ChartQA temporal answers prefer another year within ± 10 of the gold when one exists, and otherwise sample another answer of the same coarse type (year, number, yes/no, text) from the pool. Contradicting MMMU values come from the other options of the same item when multiple-choice, and from same-type answers in the same `subfield` when available (else the same subject) otherwise. Wrong ChartQA images are sampled within the same conflict-type pool (factual vs. temporal); wrong MMMU images come from the same `subfield` when present, else the same subject configuration.

Quality control. Only dataset fields and the rules above are used in construction, and 20% of instances are reserved at random for manual quality checks.

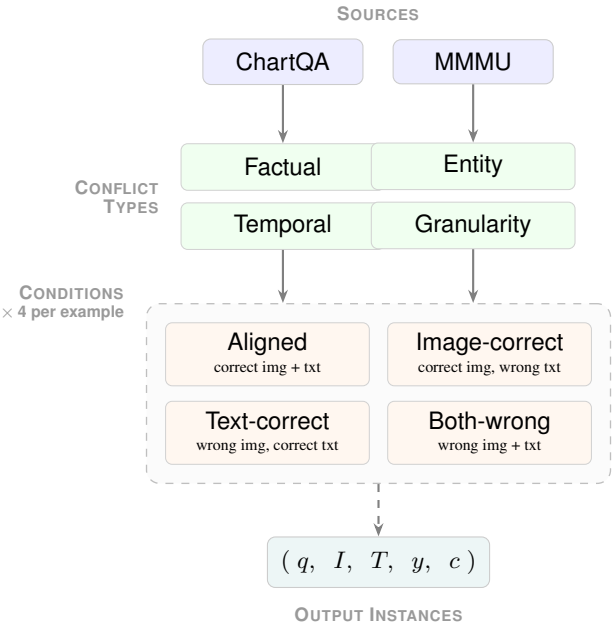

Figure 2: **CMC-Bench construction.** Source examples from ChartQA yield factual and temporal conflicts; MMMU yields entity and granularity conflicts. Each example is instantiated under four evidence conditions, producing instances (q, I, T, y, c): query, image, text, gold answer, and conflict type.

Figure 2 summarizes the pipeline. ChartQA examples are routed to factual or temporal types and MMMU examples to entity or granularity types. Each example is then instantiated under four evidence conditions, yielding tuples (q, I, T, y, c).

Experimental conditions. Each example is instantiated under four conditions (Figure 2): **Aligned** (control)—image and text both correct; **Image-correct**—image correct, text contradicting; **Text-correct**—text correct, image wrong or swapped; **Both-wrong**—image and text both wrong. Each instance is presented with the same

Table 1: Dataset statistics (released artifact).

Statistic	Value
Total examples	942
Total instances	3,768
Factual (ChartQA) ex. / inst.	250 / 1,000
Temporal (ChartQA) ex. / inst.	250 / 1,000
Entity (MMMU) ex. / inst.	250 / 1,000
Granularity (MMMU) ex. / inst.	192 / 768
Source: ChartQA ex. / MMMU ex.	500 / 442
Passage length (mean $\pm$ sd, words)	10.7 ± 5.6
QC sample (instances)	20% random

Table 2: Evaluated VLMs (all 4-bit, via `mlx-vlm`).

Model	Params	Company
Qwen3-VL-4B-Instruct	4B	Alibaba
LFM2.5-VL-1.6B	1.6B	Liquid AI
Gemma-3n-E2B-it	2B	Google
DeepSeek-VL2-small	3B[†]	DeepSeek

[†] Active parameters; MoE total is larger.

prompt (query, image, and text); the model produces an answer. Which inputs are correct or conflicting is fixed by design; responses are classified by the LLM-as-judge into one of five behavioral labels (Section 4.2). This setup mirrors real RAG settings, in which the retriever may surface contradictory evidence, and model behavior is observed without prior indication of conflict.

Scale. The construction *targets* 250 examples per conflict type (1,000 examples, 4,000 instances with four conditions each). The released build meets that target for factual, temporal, and entity types (250 each) and contains 192 granularity examples, for 942 base examples and 3,768 instances overall (Table 1).

3.3 Dataset Statistics

Table 1 summarizes the released dataset. Instances are one per (example, condition); four conditions per example implies four instances per example when all conditions are materialized.

4 Experiments

4.1 Models

Evaluation is conducted on four open-source VLMs from distinct architecture families, running on a MacBook Pro M3 Pro (36 GB unified memory) via `mlx-vlm` (Canuma, 2024). Table 2 lists the models; all use 4-bit quantization for efficient inference on Apple Silicon (Bai et al., 2025; Amini et al., 2025; Kamath et al., 2025; Wu et al., 2024).

4.2 Evaluation Protocol

Each VLM (Table 2) receives the query, the retrieved image evidence, and the retrieved text evidence under the prompt: "Query: {question}. Retrieved image evidence: [IMAGE]. Retrieved text evidence: {text_passage}. Answer the query using the provided evidence." The model produces a free-text answer; no multiple-choice constraint is imposed.

LLM-as-judge. Free-form VLM outputs are scored by an auxiliary LLM (commercial API; fixed prompt version) given the question, image- and text-supported references, and the model answer. It must emit one JSON field `label` with value in $\{\text{IMAGE}, \text{TEXT}, \text{BOTH}, \text{NEITHER}, \text{ABSTAIN}\}$, mapped to behavioral flags for metrics. This handles paraphrase, mild numeric variation, and refusals more reliably than string equality. All four models are judged on the full 3,768-instance split. Scoring uses OpenAI's `gpt-5-mini` behind a commercial API with a fixed deployment and API version (2024-02-15-preview) for reproducibility.

4.3 Metrics

All metrics are derived from judge-assigned labels (Section 4.2). Let $\mathcal{D}_c$ denote the set of instances under condition c, $\ell(i)$ the judge label for instance i, and $\mathcal{C}_{\text{conf}} = \{\text{img-cor}, \text{txt-cor}, \text{both-wr}\}$ the set of three conflict conditions.

Answer accuracy (Acc). For each condition, accuracy is the fraction of instances whose judge label matches the gold for that condition: $\ell(i) \in \{\text{IMAGE}, \text{BOTH}\}$ under image-correct, $\ell(i) \in \{\text{TEXT}, \text{BOTH}\}$ under text-correct, $\ell(i) \in \{\text{IMAGE}, \text{TEXT}, \text{BOTH}\}$ under aligned, and $\ell(i) \in \{\text{NEITHER}, \text{ABSTAIN}\}$ under both-wrong. Let $\mathbf{1}_c(i)$ denote the corresponding indicator:

$$\text{Acc}(c) = \frac{1}{|\mathcal{D}_c|} \sum_{i \in \mathcal{D}_c} \mathbf{1}_c(i) \tag{1}$$

Modality-following rate (MFR). The fraction of responses across all conflict conditions where the model follows the evidentially correct source (including abstaining under both-wrong, where NEITHER or ABSTAIN are the correct responses per

Table 3: Main results (judge-based). Acc. = condition-wise accuracy; MFR = modality-following rate; MPB = image/text share when exactly one modality is chosen; ConfabR / CDR = NEITHER / ABSTAIN rate on conflict conditions; HR = ConfabR+CDR; ΔAcc = aligned minus mean conflict Acc. n=942 per condition.

	Accuracy by Condition				MPB			HR decomposed			
Model	**Aligned**	**Img-Cor**	**Txt-Cor**	**Both-Wr**	**Img**	**Txt**	**MFR**	**ConfabR**	**CDR**	**HR**	**ΔAcc**
Qwen3-VL-4B	.898	.409	.375	.675	.549	.451	.486	.189	**.317**	.507	.412
LFM2.5-VL-1.6B	.925	.235	.870	.277	.467	**.533**	.461	.137	.050	.186	**.464**
Gemma-3n-E2B	.599	.275	.193	**.822**	.569	.431	.430	.519	.187	.706	.169
DeepSeek-VL2-small	.901	.334	.843	.279	.521	.479	.485	.132	.081	.213	.416

the accuracy metric):

$$\mathrm{MFR} = \frac{\sum_{c \in \mathcal{C}_{\text{conf}}} \sum_{i \in \mathcal{D}_c} \mathbf{1}_c(i)}{\sum_{c \in \mathcal{C}_{\text{conf}}} |\mathcal{D}_c|} \tag{2}$$

Modality preference bias (MPB). Among conflict instances where the model commits to exactly one modality (label $\in$ {IMAGE, TEXT}), let $\mathcal{D}^\star = \{i \in \bigcup_{c \in \mathcal{C}_{\text{conf}}} \mathcal{D}_c : \ell(i) \in \{\text{IMAGE}, \text{TEXT}\}\}$. Then:

$$\mathrm{MPB}_{\text{img}} = \frac{|\{i \in \mathcal{D}^\star : \ell(i) = \text{IMAGE}\}|}{|\mathcal{D}^\star|}, \tag{3}$$
$$\mathrm{MPB}_{\text{txt}} = 1 - \mathrm{MPB}_{\text{img}}.$$

Values above 0.5 indicate a systematic preference for the respective modality.

Hallucination rate (HR). Let $H(i) = \mathbf{1}[\ell(i) \in \{\text{NEITHER}, \text{ABSTAIN}\}]$. The fraction of conflict instances where the response neither aligns with either source nor commits to one—encompassing both unsupported fabrication (NEITHER) and explicit refusal (ABSTAIN):

$$\mathrm{HR} = \frac{\sum_{c \in \mathcal{C}_{\text{conf}}} \sum_{i \in \mathcal{D}_c} H(i)}{\sum_{c \in \mathcal{C}_{\text{conf}}} |\mathcal{D}_c|} \tag{4}$$

Accuracy drop (ΔAcc). The degradation from baseline to conflict conditions:

$$\Delta\mathrm{Acc} = \mathrm{Acc}(\text{aligned}) - \frac{1}{|\mathcal{C}_{\text{conf}}|} \sum_{c \in \mathcal{C}_{\text{conf}}} \mathrm{Acc}(c) \tag{5}$$

Confabulation rate (ConfabR) and conflict-detection rate (CDR). NEITHER marks answers unsupported by either modality whereas ABSTAIN marks explicit non-commitment, so $H(i)$ is split into separate rates. Let $\mathcal{D}_{\text{conf}} = \bigcup_{c \in \mathcal{C}_{\text{conf}}} \mathcal{D}_c$ denote the pooled set of all conflict-condition instances ($|\mathcal{D}_{\text{conf}}| = 3n$, $n = 942$):

$$\mathrm{ConfabR} = \frac{|\{i \in \mathcal{D}_{\text{conf}} : \ell(i) = \text{NEITHER}\}|}{|\mathcal{D}_{\text{conf}}|}, \tag{6}$$

$$\mathrm{CDR} = \frac{|\{i \in \mathcal{D}_{\text{conf}} : \ell(i) = \text{ABSTAIN}\}|}{|\mathcal{D}_{\text{conf}}|}. \tag{7}$$

ConfabR counts unsupported answers; CDR counts abstention. By construction, $\mathrm{HR} = \mathrm{ConfabR} + \mathrm{CDR}$.

5 Results and Analysis

5.1 Main Results

Table 3 presents the main evaluation results for all four models across the five metric categories.

5.2 Research Questions

RQ1: Accuracy degradation under conflict. All four models suffer large drops from aligned to conflict conditions. Accuracy drops range from 0.169 (Gemma) to 0.464 (LFM2.5), with means computed over three conflict conditions: image-correct, text-correct, and both-wrong. The smallest model in the evaluation (LFM2.5-VL-1.6B, 1.6B parameters) achieves the highest aligned accuracy (0.925) yet also incurs the largest drop (ΔAcc = 0.464), confirming that strong baseline performance does not protect against modality conflict degradation; notably, parameter count is a poor predictor of either baseline performance or conflict robustness in this evaluation. The both-wrong condition produces variable accuracy (0.277–0.822), which now correctly reflects each model's NEITHER/ABSTAIN rate rather than its ability to recover the gold answer: models score well here either by explicitly detecting the impasse (high CDR) or by confabulating answers that happen to match neither wrong source (high ConfabR).

Gemma's high both-wrong score (0.822) falls almost entirely in the latter category. Gemma's comparatively small ΔAcc (0.169) should therefore not be interpreted as conflict robustness: it is an artefact of confabulation inflating its both-wrong accuracy, not a sign of effective conflict handling.

RQ2: Modality-following rate. Across all models, MFR falls between 0.430 and 0.486, meaning models follow the *correct* modality in fewer than half of conflict instances on average. This is a strong negative result: even under an unambiguous retrieval prompt, models systematically fail to defer to the evidence-supported source. The asymmetry is stark: LFM2.5 achieves 0.870 accuracy in text-correct but only 0.235 in image-correct, indicating that its conflict resolution amounts to near-unconditional text following rather than principled modality selection. The convergence of MFR values across four architecturally diverse models (range 0.430–0.486, a spread of only 5.6 points) is unlikely to be coincidental. A model with a fixed modality bias can follow the correct source only in the single conflict condition that matches its preference, imposing a structural ceiling near $1/3$ of instances. The near-identical MFR values suggest that none of the four models has learned to track which modality is evidentially correct on a per-instance basis; each instead expresses a static prior whose ceiling is structurally similar across architectures.

RQ3: Modality preference bias. LFM2.5-VL-1.6B shows clear text preference (MPB-txt = 0.533). DeepSeek-VL2-small is marginally image-preferring by MPB (MPB-img = 0.521, MPB-txt = 0.479); however, its large condition-level accuracy gap—text-correct 0.843 versus image-correct 0.334—reveals a strong practical text lean in conflict resolution. MPB and condition-level accuracy capture different facets of preference: MPB measures the label composition among single-modality commits, whereas the condition gap measures how much a model benefits from its preferred modality being correct. Both Qwen3-VL-4B and Gemma-3n-E2B show image preference (MPB-img = 0.549 and 0.569). Neither group reliably follows the *correct* modality; rather, each model has a static prior toward one input channel that largely determines conflict behavior regardless of which is evidentially correct. Text-following models (LFM2.5, and DeepSeek by condition accuracy) are penalized heavily on image-correct instances; image-biased models (Qwen3, Gemma) are penalized on text-correct instances. The condition-level asymmetry is striking: LFM2.5's accuracy gap between text-correct and image-correct conditions is 0.635 (0.870 vs. 0.235); DeepSeek's gap is 0.509 (0.843 vs. 0.334). Image-biased models show a far weaker pull: Qwen3's gap is only 0.034 (0.409 vs. 0.375), and Gemma's is 0.082 (0.275 vs. 0.193). Text following is thus a much stronger attractor than image following in this model set, suggesting that text-modality bias is a qualitatively different and more deeply entrenched phenomenon than image-modality bias.

RQ4: Hallucination under conflict. Hallucination rates diverge dramatically across models, but the decomposition into ConfabR and CDR (Table 3) reveals that HR alone is misleading. LFM2.5-VL-1.6B produces the lowest HR (0.186) and the lowest CDR (0.050): it almost never abstains and almost never confabulates, but achieves low HR by committing to text regardless of correctness. Gemma-3n-E2B has the highest ConfabR (0.519)—more than half of its conflict-condition responses are fabrications unanchored to either evidence source—with a CDR of only 0.187. DeepSeek-VL2-small is similar to LFM2.5 (ConfabR = 0.132, CDR = 0.081). Qwen3-VL-4B presents a strikingly different profile: its HR of 0.507 decomposes into ConfabR = 0.189 and CDR = **0.317**—the highest conflict-detection rate of any model by a wide margin. Crucially, Qwen3's CDR scales with the conflict's *irresolvability*: 0.106 in image-correct, 0.346 in text-correct, and 0.500 in both-wrong. This gradient is unlikely to be coincidental; it suggests Qwen3 is sensitive to the degree of evidential tension and increasingly likely to withhold commitment as conflict deepens. This is exactly the behaviour a reliable RAG system should exhibit.

5.3 Judge Label Distribution by Condition

Figure 3 visualizes judge-label proportions in the *image-correct* condition ($n = 942$ per model), the setting most diagnostic for modality bias. Marginal label totals (available in `judge_label_counts` per model) are misleading because, for example, a large BOTH count may reflect predominantly aligned-condition agreement rather than genuine conflict resolution. The stratified view reveals the mechanism underlying the headline accuracy figures: LFM2.5 assigns TEXT to 578 of 942 image-correct instances, and DeepSeek assigns TEXT to 436, while both assign the IMAGE label in relatively few cases (198 and 296, respectively).

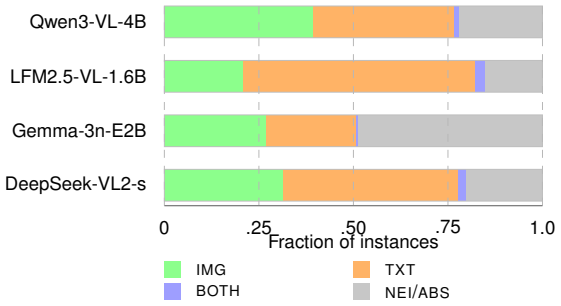

Figure 3: Stacked judge-label proportions in the *image-correct* condition (n=942 per model). Segments: IMG (image-aligned), TXT (text-aligned under image-correct), BOTH, NEI/ABS.

Qwen3 shows a more balanced split (IMAGE = 372, TEXT = 350), while Gemma assigns NEITHER to 331 instances, explaining its high hallucination rate even in a condition where image evidence is correct. A further observation concerns the BOTH labels that appear in conflict conditions (e.g., LFM2.5: 23 BOTH in image-correct; Qwen3: 13). By design, the image-supported and text-supported gold answers differ in conflict conditions, so BOTH logically should not occur. When it does, it indicates either that the conflict in that instance was insufficiently sharp (the image-supported and text-supported references were similar enough that the VLM's answer fell within matching distance of both) or residual judge leniency on near-match answers. These cases serve as a useful calibration signal: a higher BOTH rate in conflict conditions suggests that some constructed conflicts may be underspecified.

5.4 Qualitative Example

Table 4 shows one temporal ChartQA item (image-supported gold 2008, conflicting text 2015) on which the four models respectively receive IMG, TXT, NEI, and ABS labels, in line with the aggregate split in Table 3.

6 Discussion

Modality bias as a design concern. The central finding is that none of the four models selects the evidentially correct modality at above-chance rates under conflict (MFR 0.430–0.486). Instead, each model has a fixed modality preference: text-following models (LFM2.5 by MPB; DeepSeek by condition-level accuracy gap) reliably follow the text passage; image-biased models (Qwen3, Gemma) lean toward the image. This preference

Table 4: Qualitative example (`temporal_5`, image-correct condition). Gold answer = **2008** (from image); conflicting text passage claims 2015.

Model	Response (truncated)	Label
DeepSeek-VL2-s	"2008"	IMG
LFM2.5-VL-1.6B	"…according to the provided evidence, is 2015."	TXT
Gemma-3n-E2B	"…the highest native-born employment in Slovenia occurred in **2004**."	NEI
Qwen3-VL-4B	"However, the chart shows data from 2004–2012, and the text mentions '2015'—which is outside the chart's range…"	ABS

Query: *Which year recorded the highest native born employment in Slovenia?*
Image gold: 2008. Text passage: "According to the source, the value is 2015."

is not adaptive—it does not track which modality is correct for a given instance. For multimodal RAG systems, this implies that retrieval-quality improvements are necessary but not sufficient: even a perfect retriever that surfaces the correct evidence will be undermined if the generator cannot adjudicate between modalities.

Conflict detection as an evaluable capability. The HR metric conflates confabulation (NEITHER) with explicit conflict acknowledgement (ABSTAIN). We decompose these into ConfabR and CDR (Eq. 6–7) and report both in Table 3. The decomposition substantially changes the interpretation of results: Qwen3's HR of 0.507, which appears to rank it second-worst, is driven primarily by a CDR of 0.317—the model is withholding answers in the presence of irresolvable conflict, not fabricating them. Gemma's HR of 0.706, by contrast, is driven by a ConfabR of 0.519—the model is predominantly confabulating. These are opposite behaviours that a single HR figure obscures entirely. We argue that CDR should be treated as a positive metric in future multimodal conflict benchmarks: a model that says "I cannot resolve this" when evidence genuinely conflicts is exhibiting the epistemically appropriate response for a grounded generation system, and penalising it equally with confabulation is an evaluation design flaw.

Hallucination under conflict. Gemma-3n-E2B's ConfabR of 0.519 is the most concerning figure in the evaluation: more than half of its conflict-condition responses are fabrications unanchored to either modality. LFM2.5 and DeepSeek, while strongly biased toward text, at least commit to one evidence source (ConfabR 0.137 and 0.132). A low HR alone is therefore insufficient as a quality signal: LFM2.5's HR of 0.186 conceals the fact that it almost never detects or flags conflict (CDR = 0.050), it simply commits to the text channel

regardless. The ConfabR/CDR decomposition is necessary to distinguish a model that avoids fabrication because it follows a channel faithfully from one that avoids it because it recognises evidential tension.

Accuracy drop asymmetry. LFM2.5-VL-1.6B has the largest ΔAcc (0.464) despite the highest aligned accuracy (0.925). Its near-unconditional text following (text-correct 0.870; image-correct 0.235) generates a large image-correct penalty that the modest both-wrong score (0.277) cannot offset. Qwen3-VL-4B posts the second-largest ΔAcc (0.412): its image- and text-correct accuracies (0.409 vs. 0.375) are the most balanced among the four models but also the lowest in their respective categories, consistent with the absence of a dominant single-modality heuristic; the high both-wrong score (0.675) reflects a mix of principled abstention (CDR 0.317) and confabulation (ConfabR 0.189). Gemma-3n-E2B's ΔAcc of 0.169 is the smallest, but its interpretation differs from a naive robustness reading: Gemma's both-wrong accuracy (0.822) is the highest of any model, yet it is driven almost entirely by confabulation (ConfabR 0.519) rather than conflict detection (CDR 0.187). A model can score well on the both-wrong condition by generating answers that fail to match either wrong source, which is exactly what a high-ConfabR model does. ΔAcc should therefore always be read alongside the ConfabR/CDR decomposition: a small drop that co-occurs with high ConfabR is not a sign of conflict robustness.

Limitations and future work. This benchmark uses engineered conflicts constructed from heuristic templates, which may not fully reflect the distribution of naturally occurring cross-modal contradictions in real retrieval pipelines. The domain is chart-heavy (ChartQA contributes 500 of 942 base examples) and English-only. Granularity conflict is underrepresented relative to the other three types (192 vs. 250 each) due to pipeline yield.

The evaluation is scoped to small, 4-bit quantized open-source VLMs running on consumer hardware; conclusions should not be generalized to larger, non-quantized, or proprietary models without further study. Future work should include at least one stronger open model, a non-quantized variant, and proprietary VLMs to establish whether the observed modality-bias patterns persist at scale.

The LLM-as-judge protocol is central to every reported metric; its reliability has not been formally validated in this work. A human-labeled validation set with human–judge agreement statistics and an error analysis (particularly for BOTH labels appearing in conflict conditions) would strengthen confidence in the behavioral metrics. The incidence of BOTH labels under conflict conditions—where image and text support different answers by design—warrants manual auditing, as these cases may reflect underspecified conflicts, near-match answers, or residual judge leniency.

Policy baselines (always-follow-image, always-follow-text, always-abstain, random) and prompt ablations (image-only, text-only, conflict-aware prompts that explicitly permit abstention) are absent from the current study. These would clarify whether models are doing more than following trivial heuristics, and whether apparent inability to adjudicate stems from the evaluation setup rather than a genuine architectural limitation.

Per-conflict-type disaggregation of results (factual, temporal, entity, granularity) is deferred to future analysis, as it requires propagating conflict-type labels into the behavioral output files. Future work should include naturally sourced conflicts from multi-modal search logs, multilingual settings, and extended video evidence.

7 Reproducibility

The repository at `https://github.com/jaspercatapang/cmc-bench` contains all 3,768 instances, prompts, the judge specification, and aggregation code. VLMs were run with `mlx-vlm` (Canuma, 2024) on Apple Silicon (4-bit weights). The same `gpt-5-mini` judge setup as in Section 4.2 was used for all reported behavioral files.

8 Conclusion

Multimodal RAG presupposes coherent evidence; when retrieved image and text conflict, generators must arbitrate or abstain responsibly. The empirical picture here is that open VLMs largely *do not* track the evidentially correct modality per instance, and that a single "hallucination" rate can misread principled abstention as failure. This paper introduced `CMC-Bench`, a benchmark for cross-modal evidence conflict in a retrieval-style multimodal setting. The released suite has 942 examples (3,768 condition-level instances) from ChartQA and MMMU evaluation splits, with construction as in Section 3.2. An LLM-as-judge protocol assigns {IMAGE, TEXT, BOTH, NEITHER, ABSTAIN} to each of 3,768

responses per model. Across four open-source VLMs: (1) accuracy degrades sharply under conflict (ΔAcc = 0.17–0.46), with the smallest ΔAcc co-occurring with the highest confabulation rate rather than reflecting robustness; (2) modality-following rates fall in a narrow band (0.430–0.486), consistent with fixed modality priors rather than per-instance adjudication; (3) text-biased models show much larger image–text condition gaps than image-biased models (up to 0.635); (4) HR splits into ConfabR and CDR in ways that invert naive rankings—Qwen3's HR (0.507) is driven largely by CDR (0.317), while Gemma's (0.706) is driven largely by ConfabR (0.519). Cross-modal conflict resolution thus remains an open problem for multimodal RAG; benchmarks should *not* treat abstention like unsupported fabrication, and should report decomposition alongside modality-following accuracy. The dataset and evaluation code are released for community use.

Acknowledgments

This work was supported by research funding from Money Forward Inc. Figure 1 was generated using ChatGPT Images 2.0 (OpenAI). The author thanks colleagues at Money Forward Inc. and Tokyo University of Foreign Studies for discussions that shaped this paper.

References

Alexander Amini, Anna Banaszak, Harold Benoit, Arthur Böök, Tarek Dakhran, Song Duong, Alfred Eng, Fernando Fernandes, Marc Härkönen, Anne Harrington, Ramin Hasani, Saniya Karwa, Yuri Khrustalev, Maxime Labonne, Mathias Lechner, Valentine Lechner, Simon Lee, Zetian Li, Noel Loo, and 14 others. 2025. Lfm2 technical report. *Preprint*, arXiv:2511.23404.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Prince Canuma. 2024. MLX-VLM: Inference and fine-tuning of vision language models on apple silicon.

Huiyi Chen, Jiawei Peng, Dehai Min, Changchang Sun, Kaijie Chen, Yan Yan, Xu Yang, and Lu Cheng. 2025. MVI-bench: A comprehensive benchmark for evaluating robustness to misleading visual inputs in LVLMs. *arXiv preprint arXiv:2511.14159*.

Nazarii Drushchak, Nataliya Polyakovska, Maryna Bautina, Taras Semenchenko, Jakub Koscielecki, Wojciech Sykala, and Michal Wegrzynowski. 2025. Multimodal retrieval-augmented generation: Unified information processing across text, image, table, and video modalities. In *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*, pages 59–64.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models (M-haldetect). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Tao Huang, Zhekun Liu, Rui Wang, Yang Zhang, and Liping Jing. 2025. Visual hallucination detection in large vision-language models via evidential conflict. *International Journal of Approximate Reasoning*, 186:109507.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Uku Kangur, Krish Agrawal, Yashashvi Singh, Ahmed Sabir, and Rajesh Sharma. 2025. Multireflect: Multimodal self-reflective RAG-based automated fact-checking. In *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*, pages 1–17.

Reno Kriz and Kenton Murray, editors. 2025. *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models (POPE). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2024b. PhD: A prompted visual hallucination evaluation dataset. *arXiv preprint arXiv:2403.11116*.

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

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. In *Findings of the Association for Computational Linguistics: ACL*, pages 2263–2279.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1526–1535.

Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V. Le, Thang Luong, and Golnaz Ghiasi. 2024. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 288–304.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Zhihan Yin, Jianxin Liang, Yueqian Wang, Yifeng Yao, Huishuai Zhang, and Dongyan Zhao. 2026. FREAK: A fine-grained hallucination evaluation benchmark for advanced MLLMs. In *The Fourteenth International Conference on Learning Representations*.

Jingyi You, Hiroshi Sasaki, and Kazuma Kadowaki. 2025. Cross-modal clustering-based retrieval for scalable and robust image captioning. In *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*, pages 47–58.

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