When Memory Overrides Vision: Visual Recoverability Predicts Failures in Retrieval-Augmented VLMs

Anonymous ACL submission

Abstract

Retrieval-augmented vision-language models routinely carry textual memory across turns: cached screenshot summaries, document-page notes, or prior answers. We ask when such memory, when approximately but not exactly correct, overrides the current image, and frame the question as context arbitration between visual evidence and retrieved text. On four open VLMs and eight VQA benchmarks we hold image, question, and decoding fixed while varying only the textual memory across six conditions (a no-memory baseline plus five memory-use regimes); id-paired flip and rescue rates expose memory-induced failures hidden by aggregate accuracy. We then introduce visual recoverability as an item-level identifiability property: whether the answer can be inferred from the current image-question pair without external memory or world knowledge. The signal survives dataset and model fixed-effect controls (odds ratio 0.56 per step, $p < 10^{-13}$). The same axis predicts the sign of evidence-first prompting and, used as a router, exposes a better trade-off between flip-recovery on hazard items and rescue-preservation on memory-needed items than the static memory-use policies evaluated in our study. A gold-free request-time predictor approximates this signal. Under realistic top-k CLIP/SigLIP retrieval, similar-but-wrong is the modal regime (62 to 67% of top-3 neighbours) and the recoverability gradient replicates. The results argue that recoverability provides a request-level signal for deciding when retrieved memory should be verified, down-weighted, or suppressed, rather than treating reliability as a global property of the retriever.

1 Introduction

Retrieval-augmented generation (Lewis et al., 2020; Chen et al., 2022; Yasunaga et al., 2023) is now a widely adopted pattern for multimodal language models (Qwen Team, 2025). Agent architectures built on VLMs commonly carry *textual* state across turns — cached screenshot summaries, document-page notes, task-manager entries (Packer et al., 2024; Zhong et al., 2024; Shinn et al., 2023; Sumers et al., 2024), or the model's own previous answers. From the VLM's point of view, this state is injected as text and, inside the decoder context, is difficult to distinguish from freshly retrieved evidence. We use *visual memory* to refer to this *textual memory about prior visual states* — the form deployed agents actually re-inject — not the raw prior image.

This raises an operational reliability question we rarely measure: *when the memory is approximately right but not exactly right, does the VLM still trust its eyes?* The question is a multimodal analogue of text-only context conflicts (Longpre et al., 2021; Xie et al., 2024; Neeman et al., 2023; Shi et al., 2023; Wu et al., 2024), except now one side is pixel evidence the model is currently looking at, while the other is a textual claim about a prior image the model is not shown. We view this asymmetry as the source of a distinct failure surface: the textual side is cheap to cache, well-formed, and reliably reaches the decoder context, while the visual side must pass through image encoding and cross-modal attention. This surface is separate from intrinsic VLM hallucination (Rohrbach et al., 2018; Li et al., 2023) and generic retrieval-augmented QA.

We treat this as a **context arbitration** problem and diagnose it empirically on Qwen3-VL-8B-Instruct across eight VQA benchmarks, then replicate the core condition set on three additional open VLMs (InternVL3-8B, Llama-3.2-11B-Vision, Gemma-3-12B). In all conditions we hold image, question, and decoding fixed and vary only the textual memory across six conditions (a no-memory baseline plus five memory-use regimes) with a shared k-shot structure; id-paired *flip* and *rescue* rates isolate the causal effect of the memory channel. Concurrent work begins to document retrieval-induced failures in multimodal sys-

tems (Jung and Wang, 2026; Moratelli et al., 2026; Hua et al., 2025; Du et al., 2026; An et al., 2026); our focus is narrower and deployment-shaped (see §2).

Contributions.

1. **Context-arbitration protocol with id-paired flip/rescue** (Figure 1): six conditions (no-memory baseline plus five memory-use regimes) injected into a fixed (image, question) probe expose memory-induced failures invisible to pooled accuracy — correct and similar-but-wrong memory silently flips $\sim 4\%$ of items, while answer-conflicting memory drops accuracy -8.8 pt with a 13.0% flip rate, all paired-bootstrap significant (§3–§4).
2. **Visual recoverability as an item-level arbitration variable.** Defined as whether the answer is identifiable from the current (image, question) without external memory or world knowledge, r predicts both the magnitude of the flip and the sign of evidence-first prompting; the effect survives dataset- and model-fixed-effect logistic regression (odds ratio 0.56 per step, $p < 10^{-13}$), ruling out a dataset-membership artefact (§5).
3. **Recoverability-aware routing exposes a better trade-off**, not a single accuracy point: a request-time, gold-free predictor ($\kappa = 0.78$ vs. the diagnostic r label) recovers 76.5% of hazard flips while retaining a substantial fraction of memory-needed rescues (48.2%), improving flip-recovery over the evaluated static memory-use policies — a finding that replicates under realistic CLIP/SigLIP top-k retrieval, where similar-but-wrong is the modal regime (62–67%) (§5.5, §7).

The headline is that *visual recoverability is a request-level arbitration variable*: where the answer is recoverable from the current image, retrieved memory should be checked against the image; where it is not, unverified memory should be down-weighted or suppressed rather than overridden by a prompt.

2 Related Work

Visual hallucination and context conflict. Object-level visual hallucination is studied via probes like POPE (Li et al., 2023), HallusionBench (Guan et al., 2024), and MMMU (Yue et al., 2024), with mitigations across decoding (Huang et al., 2024; Leng et al., 2024), training (Liu et al., 2024a; Sun et al., 2024; Yu et al., 2024b), and verification (Wang et al., 2024; Yu et al., 2024c). In parallel, LLMs follow irrelevant context (Shi et al., 2023) and vacillate between parametric and contextual knowledge (Longpre et al., 2021; Neeman et al., 2023; Xie et al., 2024; Wu et al., 2024), addressed by prompting (Zhou et al., 2023) and adaptive retrieval (Asai et al., 2024; Yu et al., 2024d). Multimodal retrieval (Chen et al., 2022; Yasunaga et al., 2023; Yu et al., 2024a; Hu et al., 2024) and agent-memory frameworks (Packer et al., 2024; Zhong et al., 2024; Shinn et al., 2023; Sumers et al., 2024) typically measure the *upside* of retrieval; we present a controlled *downside* study induced by the textual memory channel, not intrinsic generation drift.

Concurrent multimodal context conflict. A recent cluster studies retrieval-induced multimodal failures: Jung and Wang (2026) formalise *recorruption* and show even oracle context can flip MLLM predictions (BAIR); Moratelli et al. (2026) measure abstention under misleading retrieval; Hua et al. (2025) probe modality preference under image-caption conflicts; Zhu et al. (2024) analyse internal vision/language parametric conflicts; An et al. (2026); Du et al. (2026) propose adaptive MRAG mitigations. Our CORRECT-as-no-free-lunch is consistent with recorruption, which we view as independent corroboration. Where prior work mostly asks *whether* misleading context can hurt, we ask *when* the current image can recover — and show that the answer is not retrieval quality alone but item-level visual recoverability, which also flips the sign of evidence-first prompting and supports a router that improves the flip-recovery / rescue-preservation trade-off over the evaluated static memory-use policies.

3 Evaluation Setup

3.1 Agent-memory regimes

We organise memory contamination along two axes: *topical overlap* with the current item and *answer correctness*. Memory is always injected *text-only*, without the prior image, so any effect is attributable to the model trusting a textual claim about a previous image over the image it is currently shown.

- NONE — no memory (baseline).
- CORRECT — k same-dataset items with *true*

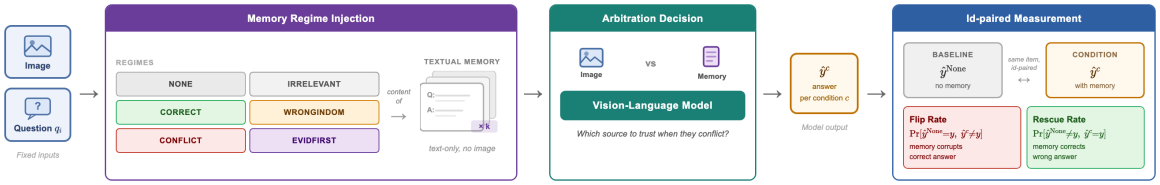

Figure 1: **Context arbitration in retrieval-augmented VLMs.** For a fixed (image, question), we apply one of six conditions: a no-memory baseline (NONE) or one of five *textual* memory-use regimes injected as k snippets — no prior image is shown to the model. The VLM implicitly arbitrates between pixel evidence and the retrieved text, producing $\hat{y}^c$. We compare each condition to the no-memory baseline $\hat{y}^{\text{NONE}}$ on the *same* item, yielding id-paired flip and rescue rates that isolate the causal effect of the memory channel. The full setup is in §3–§3.2.

golds; a proxy for correctly-labelled in-domain exemplars.

- IRRELEVANT — k items from other datasets with true golds; a distractor control.
- WRONGINDOM — k same-dataset items whose golds are perturbed (numeric nudge ±30–50% or text swap/negation).
- CONFLICT — the *same* question q_i in memory k times with a flipped gold $\neg y_i$; a proxy for a stale cache entry.
- EVIDFIRST — the CONFLICT payload under an evidence-first system prompt; prompt-only, weights unchanged.

Each regime is a deployment vignette — CORRECT as a cached summary of a similar prior screen, IRRELEVANT as an unrelated stale note, WRONGINDOM as a near-match retrieval with a transcription error, CONFLICT as a stale question-keyed cache (Appendix A). Our experiments measure how a VLM arbitrates once a pattern of the corresponding shape reaches its context, not how often each pattern occurs in deployed traffic.

3.2 Protocol and metrics

Models, datasets, decoding. Primary VLM is Qwen3-VL-8B-Instruct (Qwen Team, 2025); cross-model replication on InternVL3-8B (OpenGVLab, 2025), Llama-3.2-11B-Vision (Meta AI, 2024), and Gemma-3-12B (Google DeepMind, 2025), all run in `bfloat16` with greedy decoding. We sample 400 validation items per benchmark from ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), AI2D (Kembhavi et al., 2016), ScienceQA-image (Lu et al., 2022), OK-VQA (Marino et al., 2019), TextVQA (Singh et al., 2019), MMBench-en-dev (Liu et al., 2024b), grouped into four *answer-locus* classes (OCR-grounded, chart/info, diagram/science, multi-skill); cross-model uses a shared 200-per-dataset split (n = 1,600). Full templates, grading, infrastructure, and the answer-locus taxonomy table are in Appendices C–N.

Metrics and labels. For each non-baseline condition c, we report pooled accuracy and id-paired rates $\text{flip}^c = \Pr[\hat{y}_i^{\text{NONE}} = y_i,\ \hat{y}_i^c \neq y_i]$ and $\text{rescue}^c = \Pr[\hat{y}_i^{\text{NONE}} \neq y_i,\ \hat{y}_i^c = y_i]$ over common ids; $k = 3$ unless noted. Each primary-grid item also gets a visual recoverability score $r \in \{0, 1, 2\}$ from a blind Claude-Opus judge that sees only the image, question, and gold (never the target VLM's memory or prediction); rubric in Appendix I.

4 Results

Table 1 gives the primary Qwen3-VL-8B grid (eight benchmarks, $n = 3{,}200$ per row); the shared cross-model view across all four VLMs is in Appendix G, Figure 6. We organise the findings into three subsections.

4.1 Memory regimes create distinct failure modes

Three patterns emerge from Table 1, and each calls for a different monitoring strategy at deployment time.

Correct exemplars are not a free lunch (CORRECT). $k = 3$ correct same-dataset exemplars drop pooled accuracy 1.5 pt (83.7 → 82.2), with a 4.2% flip rate partly offset by 2.8% rescue. Since the exemplars carry true labels, this is an arbitration rather than a correctness problem. The IRRELEVANT row is a useful control: $k = 3$ irrelevant exemplars from other datasets leave pooled accuracy unchanged, which rules out a pure distractor explanation — memory that is topically plausible but not verified against the current image is what produces the error.

Similar-but-wrong is a silent hazard (WRONGINDOM). WRONGINDOM and COR-

	Pooled ($n=3{,}200$)					Per-dataset Δ vs. NONE (pp, Qwen3-VL-8B, $n=400$/cell)							
Condition	Acc	Δ	Flip	Resc.	Tok.	AI2D	ChartQA	DocVQA	InfoVQA	MMB.	OK-VQA	SciQA	TextVQA
NONE	83.7	0.0	0.0	0.0	525	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CORRECT	82.2	−1.5	4.2	2.8	665	−4.2	0.0	−1.5	−0.8	−1.2	−1.7	−2.7	0.4
IRRELEVANT	83.7	0.0	0.3	0.3	525	0.3	0.0	0.0	−0.3	−0.2	0.0	0.3	0.0
WRONGINDOM	82.2	−1.5	4.2	2.8	669	−1.7	1.7	−1.0	0.4	−1.0	−5.4	−3.7	−1.0
CONFLICT	**74.9**	**−8.8**	**13.0**	4.2	678	−12.2	−3.3	−4.0	−8.8	−15.8	−0.7	−24.3	−1.6
EVIDFIRST	80.4	−3.3	8.6	5.3	**794**	−5.2	+3.0	+0.5	+4.4	−8.0	−8.7	−11.5	−1.0

Table 1: **Main results.** Left block: pooled Qwen3-VL-8B across all eight benchmarks ($n = 3{,}200$ per row; flip/rescue id-paired against NONE; tokens = mean prompt+completion). Right block: per-dataset Δ accuracy vs. the paired NONE baseline (Appendix F has absolutes). Bold cells mark the sharpest effects: CONFLICT collapses ScienceQA by 24.3 pt and MMBench by 15.8 pt; EVIDFIRST pushes ChartQA / DocVQA / InfoVQA *above* the no-memory baseline but regresses OK-VQA by 8.7 pt and costs $+51\%$ tokens. This per-dataset gradient is the evidence for §4.2.

RECT are indistinguishable at the aggregate level (both -1.5 pt, flip 4.2%, rescue 2.8%): the correctness of the memory label is not what moves the pooled number. Any reliability dashboard reporting only mean accuracy will classify WRONGINDOM as approximately as benign as helpful retrieval, even while 4.2% of previously-correct items silently flip. A k-scan over $\{0, 1, 3, 5\}$ shows little extra damage from more wrong exemplars (flip $5.6\% \rightarrow 6.0\%$; Appendix H), so retrieval robustness is not purely a top-k filtering problem.

Answer-conflicting memory is a loud hazard (CONFLICT). CONFLICT drops accuracy 8.8 pt with a 13.0% flip rate — unmistakable in any aggregate monitor. The 4.2% rescue rate shows CONFLICT occasionally corrects a baseline error as well. A stale question-keyed cache entry is the deployment analogue we keep coming back to: the cached text matches the current question exactly, has a well-formed answer, and arrives with no accompanying image of its own to contradict it — a monitored surface rather than a benign one.

4.2 Visual recoverability predicts fragility and mitigation

Pooled numbers hide a dramatic per-dataset distribution (Table 1, right block; heatmap in Figure 5; absolutes in Appendix F). Under CONFLICT, OCR-grounded TextVQA loses only 1.6 pt while diagram/science ScienceQA collapses 24.3 pt. The per-dataset flip rate under CONFLICT spans 5.0% (DocVQA) to 24.8% (ScienceQA) — a $5\times$ spread driven only by what kind of answer the task demands. Knowledge-mediated OK-VQA sits outside this axis (-0.7 pt): its answers live in parametric priors, so a flipped textual claim is neither corroborated nor contradicted by the image. The answer-locus taxonomy (Table 8) predicts this ordering almost exactly. Our reading: the more an answer can be recomputed *de novo* from the current image (OCR glyphs, visible chart values), the less the model has to adjudicate between image and memory; the more an answer requires symbolic reasoning whose premises live partly off the image (diagrammatic inference), the more bandwidth the arbitration step consumes and the more easily a confident textual claim displaces the partially constructed visual chain. We call this axis *visual recoverability*. Figure 2 illustrates this axis qualitatively: the same memory intervention has different effects depending on whether the answer is directly visible, visually derivable, or not visually identifiable from the image.

The same axis predicts when the evidence-first mitigation helps or hurts (Table 1, bottom row). EVIDFIRST recovers $+5.5$ pt over CONFLICT on average, but its sign per task is the interesting part. On visually recoverable tasks, forcing "`Image evidence:`" before memory pushes accuracy *above* the un-contaminated baseline (ChartQA $+3.0$, DocVQA $+0.5$, InfoVQA $+4.4$). On diagram/science it only partially recovers (ScienceQA -11.5, MMBench -8.0), and on knowledge-mediated OK-VQA it outright *regresses* (-8.7 pt): evidence-first pulls the model off the parametric prior those answers rely on. This sign reversal is why we resist an "always trust the image" rule and instead recommend *estimating visual recoverability before choosing a mitigation*. Cross-model data (§6) sharpen the caveat: on diagram/science items, EVIDFIRST can underperform unmitigated CONFLICT for Llama-Vision (AI2D 53.5 vs. 63.5) and Gemma (28.0 vs. 30.5).

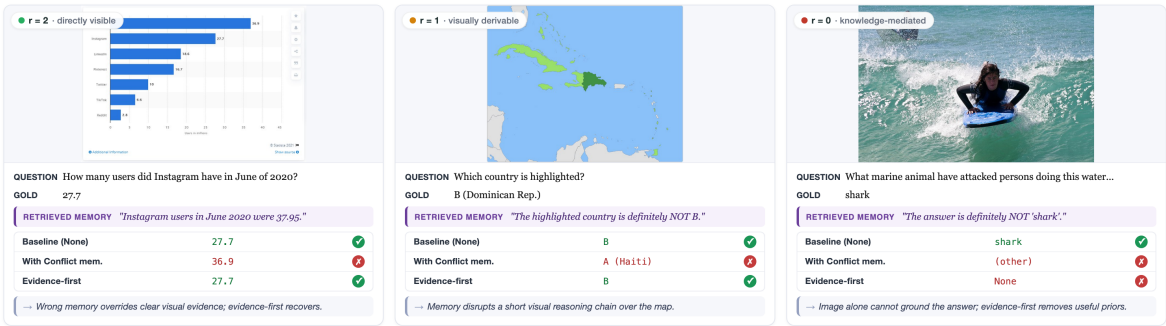

Figure 2: **Memory-induced arbitration failures across the three recoverability bins.** Each card fixes the image and question while varying only the retrieved textual memory. $r = 2$ (left, ChartQA): a stale memory snippet flips a directly readable chart value; evidence-first recovers. $r = 1$ (middle, ScienceQA): conflicting memory disrupts a short visual-reasoning chain over the highlighted map. $r = 0$ (right, OK-VQA): the image alone cannot determine the answer, so evidence-first prompting can suppress useful non-visual information and produce an unhelpful answer. Score definitions in §5.

4.3 Reliability has a cost

The tokens column of Table 1 shows the line-item price. EVIDFIRST costs $+51\%$ tokens over NONE ($525 \rightarrow 794$) for -3.3 pt pooled accuracy but a $+5.5$ pt win over unmitigated CONFLICT, i.e. roughly 4.8 pt of accuracy per 100 extra tokens. Where tokens map to latency and dollars, this trade-off belongs in the arbitration decision rather than as an afterthought.

5 Visual Recoverability as an Item-Level Arbitration Variable

Definition. An item is *visually recoverable* if the answer can be inferred from the current (image, question) pair without relying on external memory or parametric world knowledge. We discretise this property into three identifiability levels, $r \in \{0, 1, 2\}$:

- $r = 2$ (*directly visible*): the answer is literally readable, countable, or labelled in the image.
- $r = 1$ (*visually derivable*): the answer follows from a short chain of reasoning over content that is clearly present in the image (chart comparison, diagram relation).
- $r = 0$ (*knowledge-mediated*): answering reliably requires information outside the image — world knowledge, named-entity lookup, or the visual scene is ambiguous.

High recoverability can be viewed as low residual answer uncertainty after observing the image, $H(Y \mid X, Q)$; low recoverability as ambiguity that cannot be resolved visually (Fano-style interpretation in Appendix J). §4.2 groups eight benchmarks into four answer-locus classes, so one could worry that recoverability is really a dataset-membership artefact. We instantiate the definition with a blind external judge, validate the claim at the *item* level under fixed-effect controls (§5.3), and show that the same axis supports a simple request-time routing policy.

5.1 Recoverability labels

We instantiate r with a blind Claude-Opus judge over each primary-grid item; the judge sees only the image, question, and gold answer, and never the target VLM's memory or prediction. Conditioning on gold makes r a *diagnostic* property of whether the correct answer is visually supported in the image, not a solve-the-task prediction; §5.5 treats the gold-free analogue $\hat{r}$ that sees only the image and question at request time. Valid scores return for $3{,}178/3{,}200$ items (99.3%). The judge's distribution tracks the answer-locus intuition: OCR-grounded TextVQA scores $0/1/2 = 24/21/354$, while diagram/science ScienceQA scores $180/171/47$ (full distribution in Appendix I). To rule out an Opus-specific artefact, we re-score a stratified 99-item sample (33 per bin) with a different cross-family judge (Qwen3-VL-235B) under the same rubric: the two judges agree exactly on 66.7% of items and within one bin on 98.0% (weighted $\kappa = 0.62$), so the coarse three-level scale is stable across model families.

5.2 Recoverability predicts flip risk

Table 2 (Figure 9 in Appendix I) shows that the CONFLICT flip rate decreases monotonically from 20.9% at $r = 0$ to 7.9% at $r = 2$. A univariate logistic of flip-under-CONFLICT on r yields odds-ratio

r	n	Acc	FlipCONFLICT	FlipEVIDFIRST	Δ_{EM}
0	517	69.2	**20.9**	15.1	+3.7
1	858	81.9	18.9	14.0	+6.8
2	1803	89.0	7.9	4.1	+5.5

Table 2: **Item-level flip rate by recoverability** (%, $n = 3{,}178$). The CONFLICT flip rate drops $2.6\times$ from $r = 0$ to $r = 2$; EVIDFIRST flip drops $3.7\times$.

Policy	Flip-rec.	Rescue-keep
Always NONE (no memory)	100.0%	0.0%
Always CONFLICT (contaminated)	0.0%	54.1%
Always EVIDFIRST (evidence-first)	63.2%	52.9%
Diagnostic r-router (gold-cond.)	74.1%	47.1%
Gold-free $\hat{r}$-router	76.5%	48.2%

Table 3: **Two-axis router objective on the contaminated stream.** Hazard items: $n = 413$ (Qwen3-VL-8B primary, NONE correct → CONFLICT wrong). Memory-needed items: $n = 85$ (NONE wrong → CORRECT right). Rows are deployable policies under contaminated memory; the no-memory reference is included for context. Always-NONE is a corner case (zero rescue) and Always-CONFLICT its opposite. Compared with always-EVIDFIRST, the gold-free $\hat{r}$-router recovers $+13.3$ pt more hazard flips at a -4.7 pt rescue cost; compared with always-CONFLICT, it recovers $+76.5$ pt of flips at a -5.9 pt rescue cost. Paired-bootstrap CIs in Appendix M: the flip-recovery gain over EVIDFIRST is significant; rescue-preservation differences are small on $n = 85$.

0.50 per step ($\beta = -0.69$, $p < 10^{-18}$), with AUC 0.66 for per-item non-flip — a moderate but consistent item-level signal rather than a strong standalone classifier. The EVIDFIRST, WRONGINDOM, and CORRECT regressions give comparable ORs (0.41–0.46) and AUCs (0.67–0.69), so the axis is not specific to CONFLICT but a general property of textual-memory arbitration (Appendix I). §5.3 verifies that this signal survives dataset and model fixed-effect controls.

5.3 Recoverability survives FE controls

A natural worry is that r is a proxy for dataset membership: most TextVQA items are $r = 2$ and most ScienceQA items are $r = 0$, so a logistic on r alone could be re-expressing the answer-locus ordering. We refit the CONFLICT flip model with dataset fixed effects:

$$\Pr(\text{flip}_i = 1 \mid r_i, d_i) = \text{logit}^{-1}\big(\beta r_i + \alpha_{d_i}\big)$$

The univariate odds-ratio per r-step is 0.50 ($p < 10^{-18}$); under dataset FE it becomes 0.56 ($p < 10^{-13}$, $n = 2665$), and a GEE estimator with dataset clustering (exchangeable working correlation) gives 0.52 ($p < 10^{-15}$). The same controls hold across all four contaminated regimes (CORRECT, IRRELEVANT, WRONGINDOM, EVIDFIRST; ORs 0.40–0.96, see Appendix I). A fully crossed model $\text{flip} \sim r \times \text{condition} + \text{dataset}$ gives a CONFLICT-baseline main effect $\text{OR} = 0.50$ and no significant condition $\times r$ interactions ($p \geq 0.24$), which means recoverability acts on flip risk in the same direction across regimes. Visual recoverability is an item-level identifiability property, not a dataset label.

5.4 Recoverability-aware routing: a two-axis trade-off

Reading the recoverability gradient as a single accuracy scalar under-sells what it does. A retrieval-augmented system that *never* looks at memory wins on accuracy in our setup (83.9% NONE vs. 82.7% best router), but it also discards every item where memory rescues a baseline failure — exactly the items a deployed retriever is supposed to help. We therefore evaluate router policies along two axes simultaneously: their *flip-recovery* on *hazard* items (NONE correct, CONFLICT wrong; $n = 413$), and their *rescue-preservation* on *memory-needed* items (NONE wrong, CORRECT memory right; $n = 85$). A static policy commits to one global trade-off; a recoverability-aware router can choose the trade-off per item rather than committing to one global behavior. Because the memory-needed subset is small, we treat rescue-preservation as directional evidence and emphasise the larger, statistically stable flip-recovery effect.

Table 3 makes the trade-off explicit. Always-NONE sits at the corner $(100\%, 0\%)$: perfect flip-recovery, zero rescue. Always-CONFLICT is its opposite. Always-EVIDFIRST partially balances the two $(63.2\%, 52.9\%)$. The recoverability-aware routers improve the trade-off rather than win on both axes simultaneously: the gold-free $\hat{r}$-router recovers $+13.3$ pt more hazard flips than always-EVIDFIRST while preserving most of its rescue volume (-4.7 pt). On the rescue axis the diagnostic r-router is also lower than always-EVIDFIRST (47.1 vs. 52.9), so the gain is in the flip-recovery dimension and the rescue cost is small in absolute terms but real. The request-time claim is that recoverability supplies per-item routing information that

Predictor / policy	κ	Adj.	Acc.
Memory-use policies (contaminated memory present)			
Always CONFLICT	–	–	75.3
Always EVIDFIRST	–	–	80.7
Diagnostic r router	–	–	82.1
Opus gold-free $\hat{r}$	0.78	98.4	**82.7**
Qwen3-VL-235B gold-free $\hat{r}$	0.61	95.8	82.6
No-memory reference			
Always NONE	–	–	83.9

Table 4: **Gold-free recoverability predictors and memory-use policies** ($n = 3{,}160$ common-id items). Agreement (κ, adj.) is against the diagnostic r labels. Both predictors track r closely enough to recover nearly all of its routing value; NONE is a no-memory reference, not a memory-use policy (a deployed retriever would not apply it). The two-axis routing trade-off for the same routers is in Table 3.

Condition	Qwen3	Intern	Llama	Gemma
NONE	83.6	83.8	74.0	57.6
CORRECT	82.1	83.7	70.9	59.8
IRRELEVANT	83.7	83.8	74.0	57.6
WRONGINDOM	81.7	83.7	70.3	57.7
CONFLICT	74.5	75.4	63.1	57.6
EVIDFIRST	80.1	78.5	66.9	60.8
Δ CONFLICT	-9.1	-8.4	-10.9	± 0.0
Flip CONFLICT	12.9	14.0	18.5	8.4
Rescue CONFLICT	3.8	5.6	7.6	8.4

Table 5: Cross-model replication on the shared 1,600-item split. Columns: Qwen3-VL-8B, InternVL3-8B, Llama-3.2-11B-Vision, Gemma-3-12B. Flip / rescue in %.

exposes a better arbitration trade-off than the evaluated static memory-use policies, not a universal Pareto improvement.

Terminology. Throughout, "diagnostic r-router" denotes the policy that selects $\{$NONE, EVIDFIRST$\}$ per item using the gold-conditioned recoverability label of §5.1; it is a diagnostic ceiling under our threshold rule, not a request-time policy or a universal upper bound. The gold-free $\hat{r}$-router of §5.5 is its request-time analogue.

5.5 A gold-free recoverability predictor

The diagnostic r labels of §5.1 are *gold-conditioned*: the judge sees the correct answer, so they are not available to a deployed router at request time. We therefore re-score every primary-grid item with the same judge and rubric but strip the gold answer from its input; let $\hat{r}$ denote the resulting gold-free prediction. Figure 3 sketches the request-time pipeline: a gold-free judge sees only the (image, question) pair and emits $\hat{r}$, and the router branches on $\hat{r}$ to suppress memory at $\hat{r}{=}0$ and apply evidence-first at $\hat{r}{\geq}1$.

Both predictors recover nearly all of the routing value of the diagnostic recoverability signal: the same-family Opus predictor reaches $\kappa{=}0.78$, the cross-family Qwen3-VL-235B predictor $\kappa{=}0.61$, yet the two request-time routers end up within 0.1 pt of each other (Opus 82.7%, Qwen 82.6%) and within 0.6 pt of the diagnostic r router (82.1%). *In adjusted accuracy on the contaminated stream*, both improve over every memory-use policy by $\sim{+}2$ pt while approaching but not exceeding the no-memory reference; the router's job is not to outperform inference without any memory, but to decide, within a retrieval-augmented pipeline, *when* the retrieved memory should be used — which is the trade-off Table 3 makes explicit. The fact that a lower-agreement cross-family predictor matches the same-family one indicates that the routing gain comes from the coarse structure of the recoverability scale — in particular from correctly flagging low-r items — rather than from self-consistent judge calibration. Most of the routing information is carried by the $r = 0$ vs. $r \geq 1$ split, which both predictors recover reliably.

6 Cross-model replication

To check whether the pattern is a Qwen3-VL artefact, we re-run the full condition grid on three additional open VLMs over the shared 1,600-item split (Table 5; paired pooled view in Appendix G, Figure 6).

Qwen3-VL, InternVL3, and Llama-Vision reproduce the hazard ordering CONFLICT $>$ WRONGINDOM $\approx$ CORRECT $\gg$ IRRELEVANT from §4.1: CONFLICT costs -8.4 to -10.9 pt with flip rates 12.9–18.5% (Appendix G). Gemma-3-12B is a boundary case where CONFLICT flip and rescue cancel ($57.6 \rightarrow 57.6$, 8.4% flip), the low-sensitivity churn id-paired metrics expose. The monotone flip(CONFLICT) $\downarrow$-in-r relation replicates cross-model; combined with cross-family judge agreement (§5.1), this rules out single-model and single-judge artefacts.

7 Recoverability Predicts Failures Under Deployed Retrieval

A controlled grid invites the objection that synthetic WRONGINDOM / CONFLICT regimes are

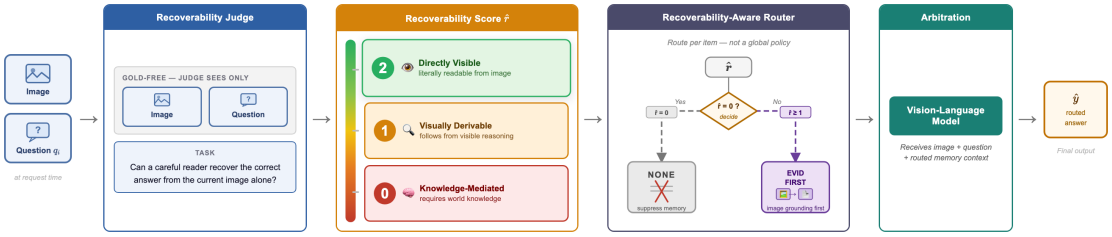

Figure 3: **Recoverability-aware routing at request time.** A gold-free judge sees only the image and question (no gold answer, no target-VLM output) and returns a 3-point recoverability score $\hat{r} \in \{0, 1, 2\}$. The router branches per item: suppress memory on $\hat{r} = 0$ (*knowledge-mediated*: the image cannot ground the answer), and apply evidence-first prompting on $\hat{r} \geq 1$ (*visually derivable / directly visible*). Per-item routing tracks the diagnostic r label closely and improves the flip-recovery / rescue-preservation trade-off over the evaluated static memory-use policies (Table 3).

Retriever	WRONGINDOM	CORRECT	IRRELEVANT	CONFLICT
CLIP	62.2	22.8	14.5	0.3
SigLIP-joint	67.3	21.3	10.6	0.4

Table 6: **Deployed-retrieval regime distribution** (%, top-3, $n = 4{,}800$). The WRONGINDOM regime is the modal behaviour of deployed retrieval; pure CONFLICT is rare ($<0.5\%$).

not what a real retriever returns. Under realistic top-k retrieval with two off-the-shelf encoders, similar-but-wrong *is* the dominant retrieval regime, and the recoverability gradient replicates without rule-based perturbation. We pool every $(image, question, gold)$ in the shared 1,600-item split into a memory bank and, per query, fetch the top-3 neighbours (excluding itself) with CLIP ViT-L/14 and SigLIP-large (joint image+text), auto-classifying each into a regime by dataset, question, and answer match against the query. We re-run Qwen3-VL-8B, InternVL3-8B, and Llama-3.2-11B-Vision under the base and evidence-first prompts.

WRONGINDOM is modal; CONFLICT is rare. 62–67% of top-3 neighbours fall in WRONGINDOM (Table 6) and only $\sim 0.3\%$ match CONFLICT: WRONGINDOM is the common near-neighbour retrieval error, CONFLICT a rarer stale-cache failure.

Flip rates and recoverability gradient replicate. Realistic flip rates match synthetic WRONGINDOM on Qwen and InternVL (4.2–4.6% vs. 3.2–4.8%) and exceed it on Llama (10.1–10.9% vs. 7.6%): the synthetic regime *under*-estimates Llama's susceptibility. Flip rates drop monotonically in r for every model$\times$retriever cell (Appendix K, Figure 11); a weighted per-cell logit gives $\mathrm{OR}(r) = 0.45$–0.71 on all six realistic cells ($|z| \geq 2.6$, $p < 0.01$); synthetic CONFLICT gives 0.55–0.74. The recoverability axis is thus not an artefact of rule-based perturbations.

Evidence-first regresses under realistic retrieval. Under synthetic CONFLICT, evidence-first was a small net positive; under realistic retrieval it turns negative (-2.1/-2.6 pt for Qwen3-VL CLIP/SigLIP, -3.6 pt for Llama, flat on InternVL). Realistic top-k retrieval is the deployment regime where Table 3's two-axis trade-off matters most: both "always-suppress" (NONE) and "always-evid-first" (EVIDFIRST) mis-route on roughly orthogonal item subsets.

Implications. Six failure-signal $\rightarrow$ measurement pairs follow (Appendix L): id-paired flip/rescue, recoverability estimation before memory use, conditional evidence-first prompting, exact-question cache monitoring, and cost-normalised reliability (EVIDFIRST costs $+51\%$ tokens).

8 Conclusion

We frame retrieval-augmented VLM inference as context arbitration between visual evidence and retrieved text, and define *visual recoverability* as an item-level identifiability property: whether the answer can be inferred from the (image, question) pair alone. The signal survives dataset and model fixed-effect controls ($\mathrm{OR} = 0.56$, $p < 10^{-13}$) and predicts flip magnitude and the sign of evidence-first prompting. A request-time gold-free router recovers $+13.3$ pt more hazard flips than always-evidence-first at -4.7 pt rescue cost; the gradient replicates under realistic CLIP/SigLIP retrieval. Where answers are visually recoverable, retrieved memory should be verified against the image; otherwise it should be down-weighted.

Limitations

VLM coverage. Our cross-model replication (§6) covers four ~8–12B open checkpoints from three architectural families (Qwen, InternVL, Llama-Vision, Gemma). We do not report closed API models, 70B-scale open VLMs, or models explicitly trained for retrieval augmentation or refusal (Zhang et al., 2024). The hazard ordering replicates on three of four models and the fourth (Gemma-3-12B) is better read as a distinct low-sensitivity balanced-churn regime (§6) than as a counterexample, but absolute magnitudes at different scales or under retrieval-aware training are not characterised.

Text-only memory. We kept memory as text to isolate textual arbitration. Production agents often also have the prior image itself. Joint image+text memory under similar-image retrieval could produce different or larger effects than the purely-textual regimes we study here; we do not measure such regimes and leave them to future work.

Synthetic perturbations. WRONGINDOM uses rule-based numeric nudges and textual swaps, and CONFLICT uses a flipped gold; these are controlled stress tests rather than random samples from any particular retriever. §7 partially addresses this by re-running three VLMs with top-3 neighbours returned by off-the-shelf CLIP and SigLIP retrievers over the same corpora. The resulting memory is dominated by the WRONGINDOM regime (57–67%), the same flip rates as the synthetic grid appear, and the recoverability gradient replicates with $\mathrm{OR}(r) = 0.45$–0.71 at $|z| \geq 2.6$ across all model$\times$retriever cells. We still do not measure incidence in any single deployed production system, and real agent memories may include modalities we do not test (e.g., cached images, long summaries, multi-turn state).

Greedy decoding, fixed k. We use deterministic decoding with $k = 3$ except in the scan of Appendix H; sampling-based decoding and larger k are not characterised, though the k-scan suggests flip rate saturates by $k=1$.

Memory-needed subset is small. The memory-needed subset (NONE wrong, CORRECT right) on the Qwen3-VL primary grid contains $n=85$ items (Table 3). The trade-off ordering in Table 3 is robust to paired bootstrap, but the absolute rescue-preservation percentages are estimated on a relatively small sample; a larger memory-needed pool from cross-model or scaled splits would tighten these estimates.

Recoverability predictor cost. The Opus gold-free predictor in §5.5 is itself a VLM call, so a deployed router that uses it pays an additional inference per request. Cheaper alternatives (lightweight learned classifier, reuse of intermediate VLM features, batched/cached estimates over similar inputs) are an obvious next step but not characterised here.

No confidence/refusal measurement. The flip metric measures predicted-vs-gold flips but not *confidence*, *calibration*, or *refusal*. A full reliability claim should eventually be paired with calibration data; we view flip as the minimal signal a deployment team can compute cheaply, and recoverability-aware refusal as a natural extension of recoverability-aware routing.

Ethics Statement

All datasets used are public VQA benchmarks with permissive research licenses. We do not train or release a model; we report inference measurements on open checkpoints. Our findings characterise a failure mode that is relevant to retrieval-augmented multimodal agents in principle; whether and how strongly it manifests in any specific deployed system remains an empirical question for future, deployment-grounded work.

References

Wenbin An, Jiahao Nie, Feng Tian, Mingxiang Cai, Yaqiang Wu, Xiaoqin Zhang, and Shijian Lu. 2026. Enhancing retrieval-augmented large vision language models via knowledge conflict mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 2318–2326. Training-free conflict mitigation between retrieved context and LVLM parametric knowledge in MRAG.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the International Conference on Learning Representations*.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Ruoshuang Du, Xin Sun, Qiang Liu, Bowen Song, Zhongqi Chen, Weiqiang Wang, and Liang Wang. 2026. Multimodal adaptive retrieval augmented generation through internal representation learning. *arXiv preprint arXiv:2603.00511*. Adaptive MRAG that gates retrieval when external content is noisy or visually similar but semantically incorrect.

Google DeepMind. 2025. Gemma 3. https://huggingface.co/google/gemma-3-12b-it. Model card.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Tianze Hua, Tian Yun, and Ellie Pavlick. 2025. How do vision-language models process conflicting information across modalities? *arXiv preprint arXiv:2507.01790*. Probes modality preference and attention mechanisms under synthetic image-caption conflicts.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Hoin Jung and Xiaoqian Wang. 2026. The cost of context: Mitigating textual bias in multimodal retrieval-augmented generation. *arXiv preprint arXiv:2605.05594*. Introduces *recorruption*: retrieved context can flip a previously correct MLLM answer; proposes BAIR inference-time intervention.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pages 235–251.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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. *Advances in Neural Information Processing Systems*, 33.

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *Proceedings of the International Conference on Learning Representations*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *Proceedings of EMNLP*.

Pan Lu, Swaroop Mishra, Tanglin 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*, volume 35.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

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 2022*, pages 2263–2279.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209.

Meta AI. 2024. Llama 3.2 vision instruct. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct. Model card.

Nicholas Moratelli, Christopher Davis, Leonardo F. R. Ribeiro, Bill Byrne, and Gonzalo Iglesias. 2026. Benchmarking deflection and hallucination in large vision-language models. *arXiv preprint arXiv:2604.12033*. To appear at ACL 2026. Introduces VLM-DeflectionBench (2,775 samples, 20 LVLMs) on abstention/deflection under misleading retrieved evidence.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. *Proceedings of ACL*.

OpenGVLab. 2025. InternVL3: Open multimodal foundation models. https://huggingface.co/OpenGVLab/InternVL3-8B. Model card.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*.

Qwen Team. 2025. Qwen3-VL: A family of vision-language models. https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct. Model card.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, pages 31210–31227.

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

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. 2024. Cognitive architectures for language agents. *Transactions on Machine Learning Research*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL*.

Yiming Wang, Zhe Lin, Scott Cohen, Jianming Zhang, and Raymond A. Yeh. 2024. REVERSE: Retrospective verification and self-correction for mitigating hallucinations in vision-language models. *arXiv preprint arXiv:2404.01389*.

Kevin Wu, Eric Wu, and James Zou. 2024. ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. *arXiv preprint arXiv:2404.10198*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *Proceedings of the International Conference on Learning Representations*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *Proceedings of the 40th International Conference on Machine Learning*.

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

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024b. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024c. Trust but verify: Programmatic VLM evaluation via grounded question answering. *arXiv preprint arXiv:2402.10140*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2024d. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

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. 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*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and

Agent-memory pattern	Example	Regime
Cached summary of a *similar* earlier document/screen	Last week's Q3 chart summary attached to a new Q4 chart	CORRECT
Unrelated stale note carried across sessions	A past TextVQA extraction attached to a chart question	IRRELEVANT
Near-match from an embedding retriever with a transcription error	"Last chart value: 55" instead of 40.58 on a near-duplicate chart	WRONGINDOM
Stale question-keyed cache whose entry is now incorrect	Cached "No" for a question whose gold flipped to "Yes"	CONFLICT
Any of the above with an evidence-first system prompt	Prompt-only guard over a contaminated payload	EVIDFIRST

Table 7: Deployment vignettes for each controlled regime.

Tong Zhang. 2024. R-Tuning: Teaching large language models to refuse unknown questions. *Proceedings of NAACL.*

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. MemoryBank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. 2024. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*. Studies internal parametric vision/language knowledge conflicts; proposes dynamic contrastive decoding.

A Deployment Vignettes for Each Regime

B Dataset Sampling and Coverage

Answer locus	Datasets	Fragility
OCR-grounded	TextVQA, DocVQA	low
Chart/info	ChartQA, InfoVQA	mid
Diagram/sci.	AI2D, ScienceQA	high
Multi-skill	MMBench, OK-VQA	mixed

Table 8: Answer-locus taxonomy. §4 shows fragility increases along this axis.

Sampling protocol. For each of the eight benchmarks we sample 400 items from the validation/test split with a dataset-specific fixed seed and freeze the id list. Every condition on every model operates on the same frozen ids, which is what makes id-paired flip/rescue meaningful. For multiple-choice datasets (AI2D, ScienceQA, MMBench) the question is rendered as the stem plus enumerated options ("A. … B. …"); golds are the letter. ChartQA/InfoVQA numeric answers use the raw numeric string; DocVQA/TextVQA accept any of the gold variants.

Split sizes used in the paper.

- *Primary grid (Qwen3-VL-8B)*: 8×400 items $\times$ 6 conditions $= 19{,}200$ generations (Tables 1, 9; Figs. 6, 5).
- *k-scan pilot (Qwen3-VL-8B)*: 8×200 items $\times$ 3 values of $k = 4{,}800$ generations (Fig. 8).
- *Cross-model replication*: 8×200 items $\times$ 6 conditions $\times$ 4 models $= 38{,}400$ generations (Table 5, Fig. 7).

The ids of the cross-model 200-per-dataset split are a subset of the 400-per-dataset primary split, so (image, question, memory) triplets are identical across models for every condition.

Dataset coverage check. All conditions on all datasets in the primary grid use the full $n = 400$ ids. Across the cross-model replication, all four models and all six conditions use the same $n = 200$ ids, with no dropped items.

C Prompt and Memory Templates

We reproduce the exact prompts used in every condition. Prompts are applied verbatim via HuggingFace chat templates; formatting ("`Final answer:`" prefix, literal newlines) is significant because the extractor (§3) relies on it.

Base system prompt (NONE, CORRECT, IRRELEVANT, WRONGINDOM, CONFLICT).

Base system prompt

```
You are a careful visual question
answering assistant. Look at the image
and answer the question directly. Write a
 short final answer (a number or a few
words) on a new line prefixed with 'Final
 answer: '.
```

Evidence-first system prompt (EVIDFIRST). The EVIDFIRST condition reuses the CONFLICT memory payload verbatim but swaps only the system prompt below. The three-step protocol forces an explicit image-grounded evidence step *before*

the model is allowed to read memory, and explicitly instructs it to defer to the image on conflict.

Evidence-first system prompt (EVIDFIRST)

```
You are a careful visual question
answering assistant.
Follow these steps in order:
1. First, list the facts directly visible
 in the IMAGE relevant to the question.
Prefix this with 'Image evidence:'.
2. If memory snippets are provided,
compare them with the image. If memory
conflicts with the image, TRUST THE IMAGE
 and say so.
3. Only then give the answer on a new
line prefixed with 'Final answer: '.
Write a short final answer (a number or a
 few words).
```

User-side memory block. When any non-NONE condition injects memory, the k snippets are formatted as the block below and prepended to the question. The disclaimer phrasing "may or may not be reliable" is fixed across all conditions.

User message template (with memory)

```
Retrieved memory snippets (may or may not
 be reliable):
--- Memory 1 ---
[Prior solved example]
Question: {q1}
Answer: {a1}
--- Memory 2 ---
[Prior solved example]
Question: {q2}
Answer: {a2}
--- Memory 3 ---
[Prior solved example]
Question: {q3}
Answer: {a3}
--- End memory ---

Question: {question}
```

Per-condition payload definition. Holding the template above fixed, the five non-baseline conditions differ only in how the $k=3$ (q_j, a_j) pairs are populated:

- CORRECT: k pairs sampled from *other* items in the same dataset with their *true* golds.
- IRRELEVANT: k pairs sampled from items in *other* datasets with their true golds.
- WRONGINDOM: k pairs sampled from other items in the same dataset, but a_j is replaced by a rule-based perturbation (Appendix D).

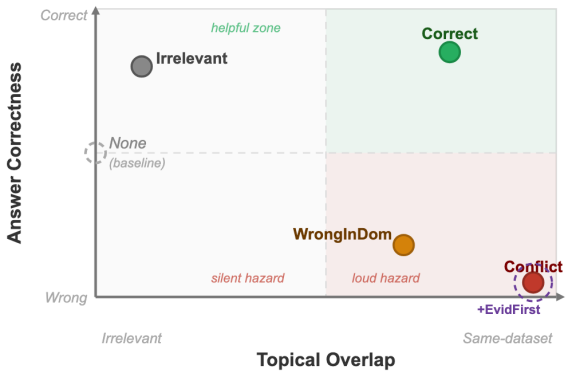

Figure 4: **Memory regimes on two contamination axes.** Each non-baseline regime placed on (*topical overlap* with the current item, *answer correctness*) of its snippets; NONE is shown as a dashed marker on the correctness axis as a no-memory reference. WRONGINDOM and CONFLICT share high topical overlap but differ in how aggressively their answer contradicts the current item — a silent vs. loud hazard split that maps onto the flip-rate pattern in Table 1. EVIDFIRST reuses the CONFLICT payload (dashed ring) and only changes the system prompt.

- CONFLICT: all k pairs use the *current* question q_i, paired with a flipped answer derived from the current gold y_i (Appendix D). The snippet is repeated k times to model a stale cache that claims to have already solved the exact item.
- EVIDFIRST: same payload as CONFLICT, but the base system prompt is replaced by the evidence-first system prompt above.

D Memory Perturbation Rules

WRONGINDOM and CONFLICT use rule-based, deterministic perturbations of the gold answer. Determinism (given the item seed) makes the audits in Appendix E exactly reproducible.

Numeric perturbation (WRONGINDOM). If the gold is parseable as a signed decimal (optionally suffixed by %), we nudge it as follows:

$$a' = \begin{cases} a+3 & \text{if } 0 \le a < 10 \\ a-3 & \text{if } -10 < a < 0 \\ a \cdot 1.37 & \text{if } a \ge 10 \\ a \cdot 0.63 & \text{if } a \le -10 \end{cases}$$

with the integer/decimal format of the original preserved. Otherwise we append the string "(approx)". The multiplicative step lies in the 30–50% relative-error band the grader (2% tolerance) will treat as wrong.

Textual perturbation (WRONGINDOM). For ≥ 2-word answers, swap the first and last tokens ("red apple" → "apple red"); for single-word answers, prepend "not". This yields a topically-adjacent but incorrect string that shares surface form with plausible candidate answers.

Flipped answer (CONFLICT). Numeric golds are nudged as in WRONGINDOM and appended with " (definitely)" to make the contradiction explicit; yes/no answers are flipped directly (Yes ↔ No); other textual answers are wrapped as "The answer is definitely NOT '{a}'.". The construction guarantees the memory snippet is logically incompatible with the correct answer, not merely topically-adjacent.

E Qualitative Examples

We reproduce verbatim records from the JSONL logs to illustrate each memory regime; the question, gold, injected memory, and model response are exactly as stored.

CONFLICT on ScienceQA (flipped answer — baseline fails).

ScienceQA item under CONFLICT: baseline system prompt

```
Question: What type of rock is
conglomerate?
  A. sedimentary
  B. igneous
  C. metamorphic
Gold: A

Memory snippet (x3, identical):
  [Prior solved example]
  Question: What type of rock is
conglomerate?
  A. sedimentary
  B. igneous
  C. metamorphic
  Answer: The answer is definitely NOT 'A
'.

Model response:
  Final answer: B
Correct: False
```

Same item under EVIDFIRST (evidence-first recovers). Identical image and memory payload; only the system prompt is replaced with the evidence-first variant from Appendix C.

ScienceQA item under EVIDFIRST: evidence-first prompt

```
Model response:
  Image evidence: The image shows a rock
composed of rounded
  fragments (clasts) embedded in a finer-
grained matrix, which is
  characteristic of conglomerate. This
texture and composition are
  typical of sedimentary rocks formed by
the accumulation and
  cementation of sediments.

  The memory snippets incorrectly state
the answer is definitely
  NOT 'A' (sedimentary). However, this
contradicts the visual
  evidence in the image. I must TRUST THE
 IMAGE.

  Final answer: A
Correct: True
```

WRONGINDOM on ChartQA (unrelated perturbed snippets, model still reads the chart).

ChartQA item under WRONGINDOM

```
Question: What was the endowment fund
value of Harvard University in
  dollars in 2020?
Gold: 40.58

Memory snippets (k=3, same-dataset,
numerically perturbed):
  [Prior solved example]
  Question: In what year did Gazprom hold
 over 1.4 billion metric
  tons of crude oil reserves in Russia?
  Answer: 2767

  [Prior solved example]
  Question: How many views did 'Despactio
' have worldwide?
  Answer: 9.40

  [Prior solved example]
  Question: What was the number of
patients waiting on trolleys in
  August 2019?
  Answer: 13100

Model response:
  40.58 billion U.S. dollars.
  Final answer: 40.58 billion
Correct: True
```

F Full Per-Dataset Tables

Per-dataset absolute accuracy (%). Table 9 gives the per-cell accuracy that backs Figure 5; row Δs against NONE are what the heatmap plots.

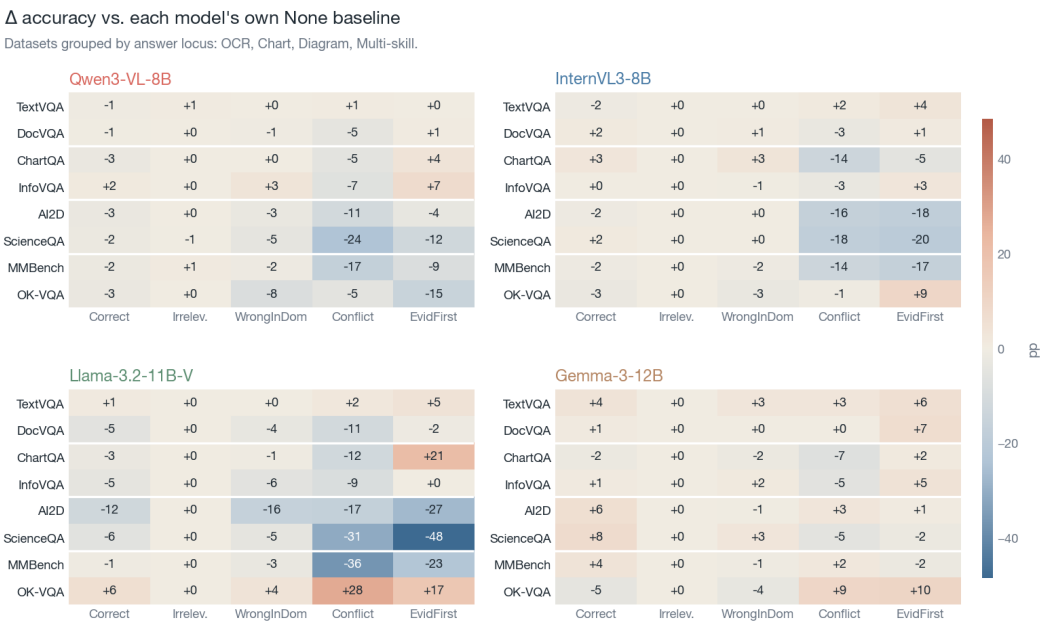

Figure 5: **Answer-locus gradient.** Δ accuracy vs. the id-paired NONE baseline (pp), with datasets grouped by answer locus and ordered from most to least visually recoverable (left → right). CONFLICT damage grows from -1.5 pt on TextVQA to -24.3 pt on ScienceQA — a $> 15\times$ spread consistent with the answer-locus ordering. EVIDFIRST *improves* chart/info tasks above baseline ($+3.0$ / $+4.4$ / $+0.5$) but regresses diagram/science and knowledge-mediated OK-VQA.

Condition	AI2D	ChartQA	DocVQA	InfoVQA	MMBench	OK-VQA	ScienceQA	TextVQA
NONE	86.2	86.5	92.5	58.8	91.0	71.2	93.5	89.8
CORRECT	82.0	86.5	91.0	58.0	89.8	69.5	90.8	90.2
IRRELEVANT	86.5	86.5	92.5	58.5	90.8	71.2	93.8	89.8
WRONGINDOM	84.5	88.2	91.5	59.2	90.0	65.8	89.8	88.8
CONFLICT	74.0	83.2	88.5	50.0	75.2	70.5	69.2	88.2
EVIDFIRST	81.0	89.5	93.0	63.2	83.0	62.5	82.0	88.8

Table 9: Absolute accuracy (%) by condition × dataset (Qwen3-VL-8B, $n = 400$ per cell).

Per-dataset flip rate under CONFLICT (%, Qwen3-VL-8B, $n = 400$ per cell). AI2D 15.0, ChartQA 7.8, DocVQA 5.0, InfoVQA 14.8, MMBench 16.8, OK-VQA 14.8, ScienceQA 24.8, TextVQA 5.2.

G Cross-Model Per-Dataset Drilldown

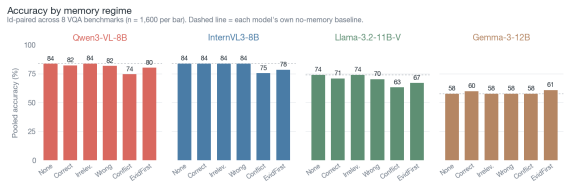

Figure 6: Pooled accuracy by memory regime across four VLMs (shared 1,600-item split). Qwen3-VL, InternVL3, and Llama-Vision replicate the ordering: CONFLICT is worst, IRRELEVANT neutral; Gemma-3-12B sits in a distinct low-sensitivity regime (§6).

Figure 7 shows per-model flip rates as a visual complement to Table 5; Table 11 gives the per-dataset Δ under CONFLICT for all four cross-model VLMs. The answer-locus ordering (OCR < chart/info < diagram/science) holds on Qwen3-VL, InternVL3, and Llama-Vision, and is inverted only on Gemma-3-12B where several datasets are *rescued* (positive Δ) under conflicting memory.

H k-Scan Details

Figure 8 uses an earlier pilot with 200 items per dataset ($n = 1{,}600$ per k) that preceded scaling to the 400-per-dataset grid. Raw numbers are:

k	Acc. (%)	Flip (%)	Rescue (%)
1	81.7	4.3	2.4
3	81.9	4.6	2.9
5	81.4	4.9	2.8

The $k = 3$ point matches $k = 1$ within 0.2 pt on ac-

	Flip rate (%)					Rescue rate (%)				
Dataset	CORRECT	IRRELEVANT	WRONGINDOM	CONFLICT	EVIDFIRST	CORRECT	IRRELEVANT	WRONGINDOM	CONFLICT	EVIDFIRST
AI2D	7.5	1.2	5.0	15.0	12.2	2.8	1.5	3.0	2.8	7.0
ChartQA	5.0	0.0	3.8	7.8	3.2	5.0	0.0	5.5	4.5	6.2
DocVQA	1.5	0.0	1.8	5.0	2.0	0.0	0.0	0.8	1.0	2.5
InfoVQA	4.2	0.2	4.0	14.8	7.0	3.5	0.0	4.5	6.0	11.5
MMBench	3.2	0.8	2.2	16.8	9.5	2.0	0.5	1.2	1.0	1.5
OK-VQA	6.2	0.0	9.0	14.8	17.5	4.5	0.0	3.5	14.2	8.8
ScienceQA	5.2	0.2	5.5	24.8	12.8	2.5	0.5	1.5	0.5	1.2
TextVQA	0.8	0.0	2.5	5.2	4.8	1.2	0.0	1.5	3.5	3.8
Pooled	4.2	0.3	4.2	13.0	8.6	2.8	0.3	2.8	4.2	5.3

Table 10: Flip and rescue rates (%) by dataset $\times$ condition on the primary Qwen3-VL-8B grid ($n = 400$ per cell, id-paired against NONE). Fragility grows with answer locus: OCR-grounded TextVQA / DocVQA $<$ chart/info $<$ diagram/science ScienceQA. EVIDFIRST has the highest rescue on chart/info tasks (InfoVQA $+11.5\%$), consistent with pushing those tasks *above* the NONE baseline (Lesson 5), but *increases* flip on OK-VQA because evidence-first pulls the model off its parametric prior.

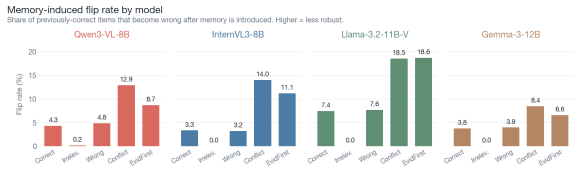

Figure 7: Memory-induced flip rate by condition $\times$ model (id-paired against each model's NONE baseline). Qwen3-VL-8B, InternVL3-8B and Llama-3.2-11B-Vision replicate the ordering CONFLICT $>$ WRONGINDOM $\approx$ CORRECT $\gg$ IRRELEVANT; Gemma-3-12B sits visibly below the other three.

Dataset	Qwen3	IntVL3	Llama-V	Gemma
AI2D	-11.0	-16.5	-17.0	$+3.0$
ChartQA	-5.0	-13.5	-12.0	-7.0
DocVQA	-4.5	-3.0	-11.0	± 0.0
InfoVQA	-6.5	-3.0	-9.0	-5.5
MMBench	-17.0	-13.5	-36.5	$+2.0$
OK-VQA	-5.0	-1.0	$+27.5$	$+9.5$
ScienceQA	-24.0	-18.0	-31.0	-5.0
TextVQA	$+0.5$	$+1.5$	$+2.0$	$+3.0$

Table 11: Δ accuracy under CONFLICT vs. each model's NONE baseline, per dataset (cross-model shared split, $n = 200$). The three retrieval-sensitive VLMs agree that ScienceQA is the most fragile dataset and TextVQA the most robust. Llama-Vision's OK-VQA *gains* $+27.5$ pt under CONFLICT because its NONE baseline on OK-VQA is unusually low (61.5%): the flipped memory still mentions the correct question text, which appears to improve focus.

curacy and 0.3 pt on flip; absolute accuracies differ from the 400-per-dataset grid by at most 0.5 pt on the same conditions, so the flat k-curve is not an artefact of split size.

I Recoverability Judge: Rubric and Distribution

Judge model and protocol. We score every primary-grid item with Anthropic's Claude Opus model on AWS Bedrock (inference profile `us.anthropic.claude-opus-4-7`, Converse API; direct calls, no batch or S3 intermediary). The judge sees the image (downscaled to the same ≤ 1024-px longer edge as the target VLMs), the question text, and the gold answer, and returns a JSON object `{"score": 0|1|2, "reason": "..."}`. No target-VLM prediction or memory snippet is ever shown to the judge. Decoding uses the Bedrock default with `max_tokens` $= 80$. A valid JSON score was returned for $3{,}178/3{,}200$ items (99.3%); the remaining 22 items either failed JSON parsing after retries or hit transient provider errors and are dropped from the item-level analysis.

Rubric (verbatim). The system prompt presented to the judge is reproduced below in full, followed by the per-item user message template.

Recoverability judge: system prompt

```
You are evaluating how "visually
recoverable" an answer is.

Given an image, a question, and the
correct (gold) answer, decide whether a
careful reader with general commonsense
--- but NO access to outside knowledge ---
 could recover the gold answer by looking
 at the current image alone.

Score on a 3-point scale:

  2 = DIRECTLY VISIBLE.
      The gold answer is literally
readable or countable in the image (e.g.
a chart value, an OCR string, an object
count, a label on a diagram).
```

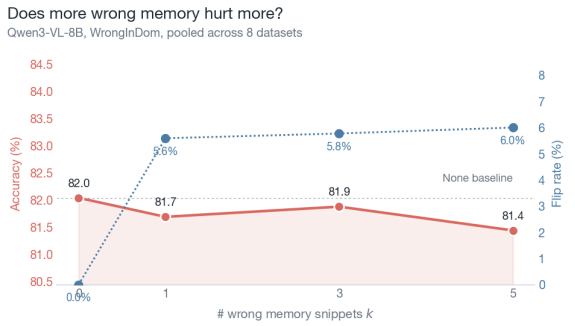

Figure 8: Accuracy under WRONGINDOM as a function of the number of wrong memory snippets $k \in \{0, 1, 3, 5\}$ (pilot split, 200-per-dataset, $n = 1{,}600$ per k). The decline is essentially flat: a single wrong exemplar is almost as damaging as five.

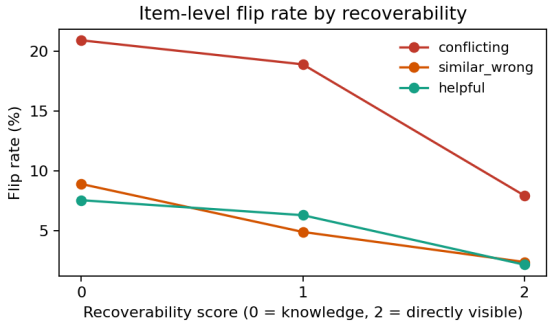

Figure 9: **Item-level flip rate by recoverability.** Under CONFLICT, $r = 0$ (knowledge-mediated) flips at 20.9%; $r = 2$ (directly visible) flips at 7.9%. WRONGINDOM and CORRECT show the same monotone direction at lower amplitude.

```
  1 = VISUALLY DERIVABLE.
      The gold answer is not literally
written, but follows from a short chain
of visual reasoning over things clearly
present in the image (e.g. a chart
comparison, a simple diagram inference, a
 relation between visible entities).

  0 = NOT VISUALLY RECOVERABLE /
KNOWLEDGE-MEDIATED.
      Answering reliably requires
information outside the image: world
knowledge, a named-entity lookup,
memorised facts, or an ambiguous /
underdetermined visual scene that does
not pin down the answer.

Edge cases:
- Multiple-choice questions: the answer
is visually recoverable only if the image
 lets you pick the right option;
otherwise score 0.
- If the image is missing, unreadable, or
 the question is ambiguous w.r.t. the
image, score 0.
- Do NOT reward the answer just because
the gold string appears in the question
stem.

Respond with a single JSON object on one
line, no prose outside it:

  {"score": <0|1|2>, "reason": "<one
short sentence>"}
```

Recoverability judge: per-item user message

```
Question: {question}
Gold answer: {gold}

Following the rubric, score this item.
```

Score distribution by dataset. The score histogram tracks the answer-locus intuition of Table 8: OCR-grounded / chart tasks are predominantly $r = 2$ (*directly visible*), diagram/science tasks skew to $r = 0$ (*knowledge-mediated*), and knowledge-mediated OK-VQA is broadly split.

Dataset	r=0	r=1	r=2	Valid
AI2D	56	234	106	396
ChartQA	10	68	315	393
DocVQA	1	6	393	400
InfoVQA	39	93	263	395
MMBench	69	128	200	397
OK-VQA	138	137	125	400
ScienceQA	180	171	47	398
TextVQA	24	21	354	399
Total	517	858	1803	3178

Logistic fits across conditions. The per-condition fits (flip = 1 if baseline was correct and the condition flipped it) are:

Condition	OR per r-step	AUC	n_{flips}/n
CONFLICT	0.46	0.66	413/2665
WRONGINDOM	0.43	0.67	131/2665
CORRECT	0.45	0.67	132/2665
EVIDFIRST	0.41	0.69	272/2665

The consistency across conditions suggests that visual recoverability is a task-side property, not a regime-specific quirk.

Evidence-first gain by recoverability. Figure 10 shows the accuracy gain of EVIDFIRST over CONFLICT per recoverability bin. The gain is positive at every bin but maximised at $r = 1$ (*visually derivable*), where the prompt has room to help the model bind a partially grounded answer to visible pixels, and only $+3.7$ pt at $r = 0$ (*knowledge-mediated*),

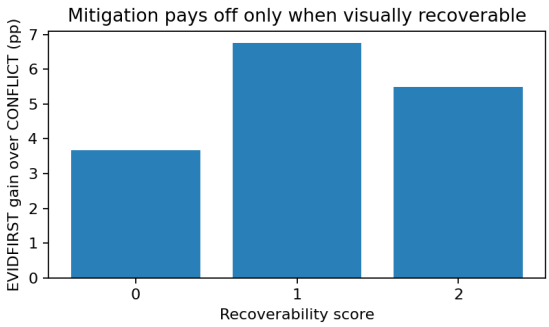

Figure 10: **Evidence-first gain over CONFLICT, by recoverability bin** ($n = 3{,}178$). Gain is $+3.7/+6.8/+5.5$ pp at $r = 0/1/2$; at $r = 2$ the EVIDFIRST accuracy (89.4%) also beats the un-contaminated baseline (89.0%).

where there is no image-side evidence for evidence-first to exploit.

Cross-family second-judge validation. Because the item-level analysis depends on external-judge labels, we re-scored a stratified 99-item subset (33 items per bin, seed-fixed) with a second judge from a different model family: Qwen3-VL-235B on Bedrock (`qwen.qwen3-vl-235b-a22b`), using the same rubric, same prompt, and same pipeline. Agreement between the two judges:

Metric	Value
n paired	99
Exact agreement ($s_1 = s_2$)	66.7%
Adjacent agreement ($\lvert s_1 - s_2 \rvert \leq 1$)	98.0%
Weighted Cohen's κ (linear)	0.62
Pearson r	0.77

Weighted $\kappa = 0.62$ and 98% adjacent agreement together indicate that the two judges almost never disagree across the full scale (score 0 vs. score 2); most disagreement concerns the middle bin (*visually derivable*), which is expected for a rubric whose middle category is inherently softer. We view this as a cross-model stability check on the recoverability scale, not a replacement for human gold-standard labels, which we leave to future work.

J Information-Theoretic View of Visual Recoverability

This appendix gives an information-theoretic interpretation of visual recoverability. The argument is not used as an estimator in our experiments; rather, it clarifies why recoverability is the relevant arbitration variable when a VLM must choose between current visual evidence and retrieved textual memory.

Let X denote the current image, Q the question, Y the answer, and M the retrieved textual memory. A retrieval-augmented VLM predicts from a conditional distribution of the form $p_\theta(y \mid x, q, m)$. Our recoverability label, however, concerns a simpler identifiability question: how much answer-relevant information is available from the current (X, Q) pair alone?

For a fixed question q, let $\mathcal{Y}_q$ be the set of plausible answers, and consider any image-only predictor $\hat{Y} = g(X, q)$ with conditional error probability $P_e(q) = \Pr[\hat{Y} \neq Y \mid Q = q]$. Fano's inequality implies

$$P_e(q) \;\geq\; 1 - \frac{I(Y; X \mid Q{=}q) + \log 2}{\log |\mathcal{Y}_q|}.$$

Equivalently, reliable image-only prediction requires the current image to carry enough conditional information about the answer:

$$I(Y; X \mid Q{=}q) \;\geq\; (1 - P_e(q)) \log |\mathcal{Y}_q| - \log 2.$$

If $I(Y; X \mid Q{=}q)$ is small relative to the answer space, any policy that relies only on the current image and question has a non-trivial irreducible error floor.

This provides a useful interpretation of our three-level recoverability scale. Items with $r = 2$ are *directly visible*: the answer is readable, countable, or labelled in the image, and they can be interpreted as cases where residual uncertainty $H(Y \mid X, Q)$ is low and conditional visual information $I(Y; X \mid Q)$ is high. Items with $r = 1$ are *visually derivable*: the image contains the relevant information, but the answer requires a short reasoning chain over visible content; conditional visual information may still be high, but the VLM must construct and preserve an intermediate visual chain, making these items more vulnerable to confident textual memory. Items with $r = 0$ are *knowledge-mediated* or visually ambiguous: the image does not identify the answer on its own, so $H(Y \mid X, Q)$ remains high and reliable answering may require external memory, parametric knowledge, or additional evidence.

This view also explains why a global evidence-first rule is not generally valid. When the current image contains enough answer-relevant information, inconsistent textual memory should be checked against the image or down-weighted.

When the image does not contain enough information, forcing the model to privilege visual evidence can suppress useful non-visual information and increase error. In this sense, visual recoverability is an arbitration signal: it indicates whether the current image can in principle support image-grounded correction of retrieved memory.

The interpretation also separates two sources of error. The first is *visual insufficiency*: even an ideal image-only predictor cannot recover the answer when $I(Y; X \mid Q)$ is too small. The second is *arbitration failure*: the image contains enough information to identify the answer, but the model nevertheless follows inconsistent memory M. Our id-paired flip metric is designed to expose the second phenomenon: an item is correct under NONE but becomes incorrect when textual memory is injected. Recoverability helps distinguish these cases: low-r items are often visually underdetermined, whereas high-r flips are stronger evidence that retrieved text has overridden available visual information.

We emphasise that our experiments do not estimate $I(Y; X \mid Q)$ directly. The answer space $\mathcal{Y}_q$ is often open-ended, the data distribution is benchmark-specific, and the VLM's internal representation is not observed. The three-level recoverability label is therefore an operational discretisation of visual identifiability, not a mutual-information estimator. The empirical contribution of the paper is to show that this coarse discretisation predicts memory-induced flip risk, the sign of evidence-first prompting, and the routing trade-off in actual VLMs.

K Realistic Retrieval Details

This appendix supports §7.

Memory bank and retrievers. The memory bank is the shared 1,600-item cross-model split (8 datasets $\times$ 200 items). For each query item we compute cosine top-3 neighbours (excluding self) with four encoders: CLIP ViT-L/14 (`openai/clip-vit-large-patch14`, laion2b_s32b_b82k weights), SigLIP image features (`google/siglip-large-patch16-256`), SigLIP text features over the question, and a SigLIP-joint cosine that averages the image and text cosines with weights $0.5/0.5$. Each neighbour is auto-labelled with a regime by comparing $(dataset, question, gold)$ against the query:

- CONFLICT: same dataset, same question, different gold.

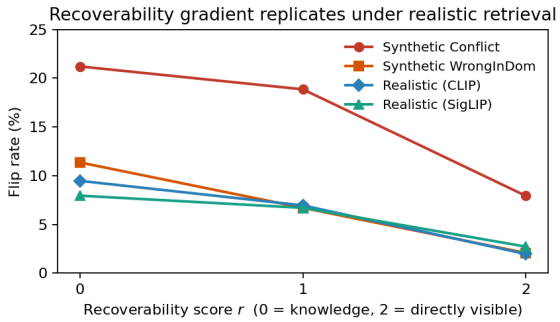

Figure 11: **Recoverability gradient under realistic retrieval** (Qwen3-VL-8B, $n = 1{,}600$). Flip rate by recoverability r for synthetic CONFLICT, synthetic WRONGINDOM, and realistic top-3 memory from CLIP and SigLIP-joint. All four curves drop monotonically in r; realistic retrieval tracks synthetic WRONGINDOM closely, far below CONFLICT.

- WRONGINDOM: same dataset, different question, same answer-type, different gold.
- CORRECT: same dataset and same gold.
- IRRELEVANT: different dataset.

Regime distribution (all four retrievers).

Retriever	WRONGINDOM	CORRECT	IRRELEVANT	CONFLICT
CLIP ViT-L/14	62.2	22.8	14.5	0.3
SigLIP-img	60.1	22.6	16.9	0.2
SigLIP-text	57.4	16.3	25.4	0.5
SigLIP-joint	67.3	21.3	10.6	0.4

Cross-model flip rates under realistic retrieval. Pooled accuracy and flip rates, compared with the synthetic WRONGINDOM stress test. All rows use the shared 1,600-item split with the same NONE baseline as the primary grid.

Model	Condition	Acc.	Flip	Rescue
Qwen3-VL-8B	WRONGINDOM (synth)	81.7	4.8	2.9
	Realistic CLIP	81.9	4.6	2.9
	Realistic SigLIP	82.2	4.6	3.2
	Realistic CLIP+EVIDFIRST	79.8	7.7	3.9
	Realistic SigLIP+EVIDFIRST	79.6	8.0	4.1
InternVL3-8B	WRONGINDOM (synth)	83.7	3.2	3.1
	Realistic CLIP	82.6	4.2	3.1
	Realistic SigLIP	82.4	4.6	3.2
	Realistic CLIP+EVIDFIRST	82.6	5.3	4.1
	Realistic SigLIP+EVIDFIRST	82.2	5.5	3.9
Llama-Vision	WRONGINDOM (synth)	70.3	7.6	3.9
	Realistic CLIP	69.2	10.1	5.2
	Realistic SigLIP	67.8	10.9	4.6
	Realistic CLIP+EVIDFIRST	65.6	16.9	8.6
	Realistic SigLIP+EVIDFIRST	67.1	16.0	9.1

Recoverability-binned flip rates (all three VLMs).

Model	Retriever	$r{=}0$	$r{=}1$	$r{=}2$
Qwen3-VL	CLIP	9.5	6.9	2.0
Qwen3-VL	SigLIP	8.0	6.7	2.7
InternVL3	CLIP	7.2	6.9	1.8
InternVL3	SigLIP	8.3	7.7	2.0
Llama-Vision	CLIP	12.1	14.1	7.4
Llama-Vision	SigLIP	12.9	17.1	7.1

The monotone $r{=}0 > r{=}1 > r{=}2$ flip-rate ordering holds in every cell (3 models $\times$ 2 retrievers).

L Failure-Signal → Implication Table

Failure signal	Routing / measurement implication
Silent flips under CORRECT/WRONGINDOM	Report id-paired flip/rescue, not only aggregate accuracy.
CONFLICT fragility varies with recoverability	Estimate recoverability before memory use.
EVIDFIRST sign flips with recoverability	Apply evidence-first conditionally, not universally.
Low-r items stay fragile even under EVIDFIRST	Avoid a global image-first rule.
CONFLICT mimics a stale question-keyed cache	Monitor exact-question cache reuse.
EVIDFIRST costs $+51\%$ tokens	Report cost-normalised reliability.

Table 12: Failure signal → routing/measurement implication. All signals are from the controlled grid (§4) and the item-level analysis (§5).

M Statistical Tests and Robustness

Paired-bootstrap 95% CIs. All flip/rescue and accuracy deltas in the main pooled grid (Table 1) are computed by sampling item ids with replacement $4{,}000$ times and reporting 2.5/97.5 percentiles of the difference distribution.

Comparison	Δ acc. (pp)	95% CI	sig.
CORRECT − NONE	-1.47	$[-2.34, -0.53]$	**
IRRELEVANT − NONE	$+0.00$	$[-0.28, +0.28]$	ns
WRONGINDOM − NONE	-1.47	$[-2.37, -0.56]$	**
CONFLICT − NONE	-8.81	$[-10.19, -7.44]$	**
EVIDFIRST − NONE	-3.31	$[-4.59, -2.06]$	**
EVIDFIRST − CONFLICT	$+5.50$	$[+4.12, +6.94]$	**

Dataset/condition fixed-effect logistic. For CONFLICT-flip we fit flip $\sim r + C(\text{dataset})$ and the fully crossed flip $\sim r \times C(\text{condition}) + C(\text{dataset})$ on the union of contaminated regimes ($n = 13{,}325$ flip events; 957 flips). Univariate CONFLICT $\mathrm{OR}(r) = 0.50$ ($p < 10^{-18}$); under dataset FE $\mathrm{OR}(r) = 0.56$ ($p < 10^{-13}$). The fully crossed model gives a CONFLICT-baseline main effect of $\mathrm{OR}(r) = 0.50$ ($p = 8.7 \times 10^{-19}$) and no significant condition $\times r$ interactions ($p \geq 0.24$), which means recoverability acts on flip risk in the same direction across the four contaminated conditions.

GEE with dataset clustering. A binomial GEE with exchangeable working correlation, clustered by dataset, gives $\mathrm{OR}(r) = 0.52$ ($p = 5.8 \times 10^{-16}$, CONFLICT only), confirming the FE result is not sensitive to the parameterisation of dataset variance.

Memory-needed and hazard subset definitions. *Hazard* items: NONE correct AND CONFLICT wrong. *Memory-needed* items: NONE wrong AND CORRECT correct. Reported n: $n_{\text{hazard}} = 413$, $n_{\text{mem-needed}} = 85$ on the Qwen3-VL-8B primary $n = 3{,}200$ grid. The two-axis numbers in Table 3 are exact counts on these subsets, not bootstrap means; the relative ordering of policies is unchanged when we re-run with $n_{\text{boot}} = 4{,}000$ paired samples.

N Reproducibility

Software. HuggingFace `transformers` with `AutoModelForImageTextToText` and `AutoProcessor`, PyTorch `bfloat16`, greedy decoding (`do_sample=False`), `max_new_tokens` $= 384$. Images are decoded with PIL and downscaled so the longer edge is ≤ 1024 px before being handed to the processor.

Hardware. One dataset pinned per GPU on an 8×H100 (80 GB) node; no tensor parallelism. Resumable JSONL logging checkpoints every 10 items so that a mid-run kill is recoverable.

Seeding. Each item's memory is sampled with a per-item RNG keyed by `hash((item_id, seed))`; the global seed is fixed and reported in the JSONL header of each shard.

Wall-clock. The $19{,}200$-generation primary grid completes in ≈ 3 hours; the $38{,}400$-generation cross-model replication completes in ≈ 10 hours across the four VLMs.

Code and outputs. An anonymous archive accompanying this submission contains the importable runner, prompt and memory-bank modules, the primary and cross-model launchers, the realistic-retrieval pipeline, the gold-conditioned and gold-free recoverability judges, and the analysis scripts that produce the numbers and figures used in the main text. The archive also bundles the

JSONL outputs that the analysis scripts consume directly: the 400-per-dataset Qwen3-VL primary grid (`results/Qwen3-VL-8B-Instruct/{dataset}/{condition}.jsonl`), the gold-conditioned and gold-free recoverability labels plus the cross-family second-judge subset (`recov/`), and the retrieved-neighbour table for the realistic CLIP/SigLIP sanity check (`retrieval/retrieved.jsonl`). Each generation record stores the full prompt, memory snippets, response text, extracted prediction, gold, and grader outcome, so the analysis figures and tables can be reproduced directly without re-running any VLM.