HARD TO UNSEE: SOURCE ADDRESSABILITY AFTER MULTIMODAL INTEGRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A system that integrates several sources often learns which one to trust only afterwards, when the original evidence can no longer be re-read. We ask what a vision-language model can still do with a late, source-conditioned instruction once only an intermediate representation remains. An image and one or two texts disagree about one attribute; an inserted summary records or omits which source said what; access to the evidence is kept or removed by a verified attention mask; only then is a source queried by name or declared unreliable. After a source-neutral summary, Qwen3-VL-8B and Pixtral-12B do not retrieve what a named non-default source reported above a shortcut ceiling, and a late reliability statement is followed only when it agrees with a persistent, image-favoring default. Carrying arbitrary source handles through integration, without binding any value to them, restores most named-source retrieval; handles introduced after the bottleneck do not; answers follow a handle when its source assignment is permuted; the effect survives eviction of the evidence from the cache; and blocking attention to the carried handle at question time removes it while a matched sham does not. The hierarchy replicates in Pixtral-12B and on controlled natural images, is weaker for visual than for text targets, and reliability-conditioned selection lags retrieval only in Qwen3-VL-8B. In these settings, later source-conditioned access depends strongly on a referent carried through integration, while explicit source-value binding provides additional support for reliability-conditioned selection.

1 INTRODUCTION

Systems that combine several sources rarely know in advance which deserves trust: a photograph is later shown to be manipulated after the evidence has been read and often after it has been folded into a summary, a memory entry or a compressed context. We ask what a vision-language model (VLM) does with source-conditioned information arriving at that point.

Late correction requires source-value addressability (finding the content that came from the named source), reliability revision and answer selection from the revised state rather than a default; if the evidence cannot be re-read, everything depends on what the intermediate representation preserved and whether later computation can address it. Existing multimodal conflict and source-monitoring studies generally keep the original evidence accessible or supply source information before or during evidence processing (Section 2). We instead ask what remains addressable when source-conditioned information arrives only after multimodal evidence has been integrated and made inaccessible.

Our paradigm (Sections 3 and 4) separates evidence availability from provenance in the intermediate representation, with a text-only reader of the same summary as reference. The failure we report is not that a source-neutral summary omits provenance: the hidden states computed with the evidence in view still carry evidence-derived information (an evidence swap changes the logits, and a persistent default is produced through them), and a summary that merely names the sources makes that information addressable while a text-only reader of the same summary cannot use it.

In the two-source design, with only a source-neutral summary readable, Qwen3-VL-8B follows a late statement that the caption is unreliable at 1.00 and one that the image is unreliable at 0.22-0.23, against 0.90-0.94 when the evidence can be re-read or the summary records provenance: a default-override failure, in which the statement is "followed" exactly when following it coincides with a persistent default that the readable text does not specify.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Is this a failure of trust updating or of source-conditioned access more broadly? With three named sources reporting three different values we ask, under the same bottleneck, what X reported, which value if X is reliable, and which value if the other two are unreliable. Retrieval would survive if only trust revision had failed; it does not: in every cell for both models, retrieval of a non-default source's value falls at or below a text-only reader of the summary, and so do trust and distrust; explicit provenance restores retrieval to 1.00 under the same mask.

The phenomenon is therefore a loss of source-conditioned addressability around a persistent default, not an isolated failure of trust updating: after a source-neutral bottleneck a default remains behaviorally influential while non-default source-to-value associations become unusable when the intermediate text names no source; a persistent source referent, whether a native name or an arbitrary handle assigned before integration, restores most retrospective retrieval and part of reliability-conditioned selection, and explicit source-value binding restores the rest. A selective trust-failure reading, on which source tracking survives while trust updating fails, is not supported.

2 Related Work

Conflict and credibility (Appendix A). Image-text conflict studies measure which source prevails: synthetic shape-and-caption stimuli close to ours (Nguyen et al., 2025), legibility-degraded sources (Ghosh et al., 2026), conflicting descriptions inside multimodal evidence (MMKC-Bench, Jia et al., 2025). Credibility-aware work (Pan et al., 2024; Schuster et al., 2026; Pradhan & Goley, 2026; Ashkinaze et al., 2026) supplies source or credibility information with or before the evidence and keeps the evidence in context. None declares a source unreliable afterwards or removes the evidence. Two lines of work are closest. Source-Modality Monitoring (Hua et al., 2026) shows that, with the evidence in view, VLMs can report by arbitrary labels and follow a reassignment of labels to modalities; Binding-ID work (Feng & Steinhardt, 2024) permutes entity-attribute bindings by intervening on activations while the context remains readable. Our accessible-evidence conditions replicate the first; our handle permutation is a stimulus-level analogue of the second. What neither examines is the combination studied here: source identity fixed before integration, the evidence then unavailable, a readable intermediate text that does not restate any source-value mapping, and a later instruction that names a source. These prior lines do not jointly test post-integration source addressability under evidence removal together with later named-source retrieval and use and reliability-conditioned selection.

Compression and late correction. The structural precedent is text-only: "Reclaim Evaluation" (Kwon, 2026a) delivers a correction after a session has been compressed into a note with manipulated content (a point that applies to our factorial too, Section 5.1), and "Manufactured Confidence" (Kwon, 2026b) applies a distrust instruction to source-stripped agent memory. Neither has conflicting sources, a named-source statement, multimodal input or a bottleneck that hidden states can cross; that what an intermediate keeps bounds what a late signal can achieve is their idea.

Mechanisms of arbitration. "Instruction Anchor" (Zhang et al., 2026) localizes modality arbitration to instruction-token states, with the instruction after the evidence in the same prompt; router-head work (Hua et al., 2025) shows causal control over which modality is reported; text-only analogues (Prakash et al., 2025; Leong, 2026) are in Appendix A.

Table A1 (Appendix A) tabulates these papers against the dimensions of our setting.

3 Problem Formulation

Sources and conflict. An episode contains sources $S = \{s_1, ..., s_k\}$, each reporting a different value $v(s)$ of the same attribute of the same scene: $k = 2$ (an image and a caption) or $k = 3$ (Section 4).

Intermediate representation. After the model processes the evidence E, an intermediate representation Z is placed in the context. Its textual part z is provenance-explicit (z_E: "The image shows a. The caption says b.") or provenance-neutral (z_N: "Reported <attribute>: a and b.", order counterbalanced). Its non-textual part is the hidden states at the positions of z, computed with E in view.

Evidence access and reference. In the accessible condition (A) later positions can attend to E and Z; in the blocked condition (B), to Z but not E. The text-only reference (T) never received E and reads only z under the same mask: at chance by construction for z_N (1/2 with two sources, 1/3 with three) and at ceiling for z_E. Such a reader succeeds in every cell but (B, z_N), so a positive access x provenance interaction is expected by construction and is not the contribution (Section 5.1).

Three capacities the target behavior conflates. (A) Source retrieval, or source-value addressability: given a source name, produce the value it reported. (B) Reliability updating: change the standing of a source's content when a statement changes the source's standing. (C) Answer selection: output the value the updated state supports rather than a default. Two-source experiments cannot separate the three, nor "selects the reliable source" from "avoids the default". The three-source experiment separates A from B by asking the retrieval question without any change of trust, and with three values a policy that knows only the default and picks among the other two cannot exceed 0.50 on non-default targets: the shortcut ceiling.

4 Experimental Setup

Models and inference. Qwen3-VL-8B (Instruct), Pixtral-12B and Idefics3-8B-Llama3 run every condition (baselines included) through the same attention-hook code path (inference settings, Appendix D, E).

Stimuli. Three synthetic families, COUNT, COLOR and CHART, with labels correct by construction and fresh seeds for every experiment (renderer and item counts, Appendix C).

Conversation structure. Two-source: an image and a caption disagree about exactly one attribute (the image's value is true of the rendered scene; image first); an inserted assistant summary z_E or z_N follows, then the reliability statement (both directions per item) and the question (prompts, Appendix B); 360 confirmatory items (120 per family) and a 90-item fresh replication anchor run with the original design and cue. Three-source: three sources report three distinct values; source sets ITT (image, caption, field note), TTT (three texts) and IIT (two images, caption); label schemes NATIVE and NEUTRAL (Source A, B, C; no modality word in any turn); order, letters, summary value order and value-to-source assignment counterbalanced (Appendix B, F). The final turn is NONE (the bare question: the default), QUERY-X (what X reported; no trust statement), TRUST-X, DISTRUST-YZ (the correct source never named) or COMBINED-X (the two-source wording), X rotated within item; 540 items plus the anchor (counts, Appendix C).

Removing evidence access. In blocked conditions every position from the final user turn onward, including generated tokens, cannot attend to the turn-1 positions (Appendix E). Completeness is verified in every run by an evidence-swap test: with evidence and summary both blocked, replacing the sources with another item's under identical forced continuations must change the logits by exactly 0.0, and did in every check of every run; with only the evidence blocked the same swap changes logits substantially, so direct access is removed while evidence-derived information still propagates through the intermediate states (sham mask and cache eviction, Section 5.4).

Cells, gates and estimands. AN, AE: evidence accessible, neutral or explicit summary; BN, BE: evidence blocked; TN, TE: text-only readers under the same mask; C0: single-source capability checks. A three-source item enters the confirmatory analysis only if all its C0 checks are correct (every verdict is unchanged without the exclusions); a condition whose accessible-neutral accuracy is below 0.60 fails the prerequisite and its blocked result is not interpreted. The two-source criteria P1-P4 and the factorial estimand I = [BE - BN] - [AE - AN] (item-bootstrap 95% intervals, exact sign test) are defined in Appendix D. In the three-source experiment the default is the item's BN-NONE answer, accuracies are over non-default targets, and blocked-neutral QUERY is compared with the 0.50 ceiling and the text reader (remaining definitions and interval rules, Appendix D, F).

Status of each analysis. Analysis rules were fixed before model outputs; the two-source Qwen experiment is confirmatory, the referent and modality controls are pre-specified follow-ups, and the sufficiency patching is exploratory. Full design and reproducibility details are in the appendix.

5 RESULTS

Unless stated, two-source numbers are accuracies for the direction in which the image is later declared unreliable, cells are written AN / AE / BN / BE, and three-source numbers are accuracies over non-default targets.

5.1 DOES THE ORIGINAL PHENOMENON REPLICATE?

Confirmatory two-source experiment (Qwen3-VL-8B). On 360 items all four pre-specified criteria hold: under BN, accuracy is 1.00 when the caption is declared unreliable and 0.23 when the image is, and the factorial interaction is I = +0.64 [+0.59, +0.70] (cells, discordant counts and tests, Appendix G).

Fresh replication anchor. Ninety fresh two-source items with the original design and cue replicate the result in Qwen (BN 0.22 when the image is declared unreliable, 1.00 in the other direction; I = +0.60 [+0.47, +0.72]) and, more weakly, in Pixtral (0.50 and 0.99; I = +0.36 [+0.24, +0.47]; Appendix G).

The directional asymmetry. Following the caption-unreliable statement coincides with keeping the image's value (Qwen 1.00 in both experiments); following the image-unreliable statement requires giving it up (0.22-0.23 for Qwen, 0.34-0.50 for Pixtral). This is the default-override failure of Section 1: near-ceiling accuracy in one direction is what a persistent default alone would produce. Generic instruction-following failure and summary order are ruled out for Qwen (Appendix G). COUNT exhibits a distinct pattern from COLOR and CHART; we leave the source of this difference unresolved.

5.2 WHAT SURVIVES WHEN PROVENANCE IS ABSENT?

Under BN the answers sit on the image's value. In the confirmatory experiment 638 of 720 BN answers equal the image's value (0.89), against 0.49 for the text-only reader, and with no reliability statement 115 of 120 do. The same holds for Pixtral, Idefics3 and all six cue wordings (Appendix G).

In the three-source design's bare-question cell, Qwen's blocked-neutral answers sit on the image's value at 0.87 (native labels) and 0.95 (neutral labels) against a text-only reader with no source information (image share above the reader by +0.52 [+0.40, +0.63] and +0.60 [+0.49, +0.71]); Pixtral's default is weaker and label-dependent (0.39 native, not above the reader; 0.71 neutral, +0.35 [+0.23, +0.48]); with two conflicting images the image dominance disappears in both models (Appendix F).

A persistent default that the readable text does not specify thus reaches the answer through the intermediate states. We call this a persistent non-textual default: one value keeps being produced, which does not mean the model can say where it came from (Section 5.3). Presentation order does not explain it: the ITT default sits on the image in every slot, with only a mild first-slot preference in TTT (Appendix F). Since the image's value is also the true value in all completed stimuli, the default could track the visual source or the value that was true or confidently perceived; Section 5.6 weakens a certainty-only account. Two agreeing textual sources do not displace it (pre-specified diagnostic, Appendix G).

5.3 DOES USABLE SOURCE-VALUE ADDRESSABILITY SURVIVE?

The three-source experiment asks, under the same bottleneck and with no trust at issue, what a named source reported (QUERY, "According to X, ...").

In no cell of either model does blocked-neutral QUERY exceed the shortcut ceiling; the highest value is Qwen IIT native at 0.48 [0.40, 0.56]. The verdict, in every cell, both models, both label schemes, with and without the capability exclusions, is that blocked-neutral retrieval does not exceed the shortcut ceiling. In Qwen's ITT cells retrieval of a non-default source's value is below the text-only reader, which has no source information at all (QUERY BN minus TN, pooled over label schemes, -0.18 [-0.24, -0.13]; TTT and IIT, Appendix F). Pixtral is at the text reader in TTT and IIT and above it only in ITT, still below the ceiling (Appendix F).

cell	Qwen3-VL-8B BN [TN] (AN)	Pixtral-12B BN [TN] (AN)
ITT native	0.09 [0.32] (0.98)	0.20 [0.29] (0.84)
ITT neutral labels	0.16 [0.31] (0.98)	0.27 [0.29] (0.91)
TTT native	0.16 [0.19] (1.00)	0.28 [0.26] (1.00)
TTT neutral labels	0.25 [0.29] (1.00)	0.33 [0.32] (0.99)
IIT native	0.48 [0.29] (0.92)	0.29 [0.34] (0.94)
IIT neutral labels	0.24 [0.30] (0.94)	0.30 [0.29] (0.83)

Table 1: QUERY accuracy for non-default targets, evidence blocked with a neutral summary; the text-only reader of the same summary in brackets, the evidence-accessible cell in parentheses.

When the queried source is the default, QUERY is high: Qwen ITT 0.98 (native) and 0.80 (neutral), Pixtral 0.79 and 0.55: under the bottleneck the models return the default value whichever source is named.

TRUST and DISTRUST are at floor with QUERY: blocked-neutral TRUST / DISTRUST accuracies for non-default targets are 0.10 / 0.04 in Qwen ITT native, and no cell of either model exceeds 0.41 / 0.28 (per-cell values, Appendix F). By the pre-specified gate these trust results are not interpreted as source-specific rebinding failures, because QUERY fails first; the pre-fixed cross-cell readings are all false in every source set, so the failure is not attributable to priors attached to the native source words (Appendix F).

This does not support the selective trust-failure account, which predicts retrieval above the ceiling with trust at floor. Explicit provenance in the summary preserves addressability: 1.00 for retrieval everywhere and high for TRUST / DISTRUST (Qwen ITT neutral 0.93 / 0.65, Qwen TTT about 1.00, Pixtral 0.96 / 0.93 or higher), with Qwen's ITT native cell lower (0.62 / 0.27; Appendix F).

5.4 Is the result a masking artifact?

Four checks address off-distribution computation under the mask. The evidence-swap difference is exactly 0.0 in every check of every run, all source sets and models. A sham mask changes pooled accuracy by +0.03 [+0.01, +0.05] (Qwen) and +0.01 [-0.01, +0.04] (Pixtral), within the pre-set equivalence band. Cache eviction agrees with the mask on 0.987 (Qwen) and 0.972 (Pixtral) of answers (Appendix E). The explicit summary is read correctly under the mask (BE QUERY 1.00 in every cell), so the mask does not remove the ability to follow a late source-conditioned instruction.

5.5 Source referents: a three-level hierarchy

Is the failure a failure to resolve the source names? In the blocked source-neutral condition the readable text contains no antecedent for "the image" or "the caption". Pre-specified controls (Appendix I) insert instead a summary that names the sources without binding any value to them ("Sources consulted: the field note, the caption and the image. Reported counts of orange triangles: 8, 6 and 4."). On the 90 image-text-text items, evidence blocked, non-default targets, Qwen3-VL-8B answers the named-source QUERY at 0.09 [0.06, 0.13] with the ordinary neutral summary, 0.88 [0.83, 0.93] with the anchor and 1.00 with the explicit summary; the combined reliability cue goes from 0.14 [0.09, 0.20] to 0.70 [0.62, 0.77] and 0.98. The anchor's text-only reader is at chance (0.33), the accessible-evidence cell is unchanged (0.98), and so is the default (image share 0.87 vs 0.94). Referential continuity matters. Pixtral-12B gives the same result with a smaller effect (QUERY 0.20 to 0.71 to 1.00; combined cue 0.33 to 0.65 to 1.00; color and chart only, Appendix I).

Is it the words "image" and "caption"? A second control uses 90 independently generated scenes whose sources are introduced at turn 1 only as arbitrary handles, "Source A/B/C", with letter-to-modality and letter-to-slot assignment counterbalanced, and an anchor summary that names the handles ("Sources consulted: Source B, Source A and Source C. Reported ..."), with the value order in the summary counterbalanced independently of the handle order. Figure 1 gives the result for both models. With the evidence blocked, Qwen3-VL-8B retrieves a named non-default handle at 0.80 [0.74, 0.86] against 0.16 [0.11, 0.22] for the ordinary neutral summary (text-only reader 0.36; accessible-evidence prerequisite 0.97), and the combined reliability cue rises from 0.14 to 0.61 [0.53, 0.68]. The handle must be the one carried through integration: fresh handles ("Source P/Q/R") introduced only after the bottleneck recover far less (0.37; continuity contrast +0.44 [+0.35, +0.52]); with stable handles the answers land on the queried source's value (0.79) rather than on the third value (0.08), whereas with late handles the three values are near-equiprobable, so the handle raises

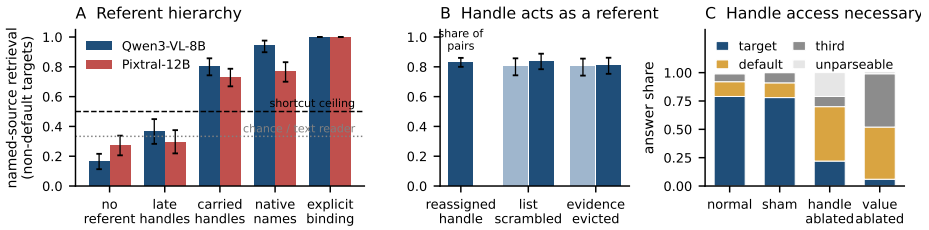

Figure 1: Source-conditioned retrieval after a source-neutral bottleneck (evidence blocked; non-default targets; item-bootstrap 95% intervals). (A) Retrieval by the intermediate text's treatment of sources, 90 held-out three-source scenes; dashed: the 0.50 ceiling for a policy that only avoids the default; dotted: chance and the text-only reader. (B) Qwen3-VL-8B: answers follow a handle when its source assignment is permuted with content fixed (share of pairs), are unchanged when the handle list is scrambled, and persist when the evidence is evicted from the cache (light bars: baseline). (C) Qwen3-VL-8B: blocking attention to the two carried-handle tokens at question time removes retrieval, a matched sham does not, and blocking the value token moves answers to the third value.

the intended source rather than merely suppressing the default. The same scenes relabeled with native names replicate the first hierarchy on held-out items, and Pixtral-12B replicates the hierarchy and the continuity effect with the same magnitudes (stable handles 0.27 to 0.73 and 0.25 to 0.71; continuity contrast +0.44 [+0.36, +0.52]; queried source 0.73 vs third value 0.07), although its two cues are restored equally, so the retrieval-over-reliability gap is Qwen-specific. The result holds in every family.

Is the handle a source referent or a fixed default? A third pre-specified control (Appendix I), the post-integration counterpart of label reassignment with the evidence in view (Hua et al., 2026), keeps every scene, its evidence, order and values fixed and permutes only the letter-to-source assignment. Under blocked evidence, Qwen3-VL-8B retrieves the queried letter's current source at 0.80 / 0.81 / 0.80 in the three assignments (text readers 0.33-0.36), and when the same letter is reassigned to a different source within a scene the answer follows it in 0.83 [0.80, 0.86] of 635 pairs (per answer: 0.80 on the queried source, 0.10 on the default, 0.10 on the other non-default source). Tracking is uniform across letters, assignments and families but weaker for the third-presented source (0.56 vs 0.96 and 0.90 by slot), and the benefit is largest where the image default is strongest (Appendix I).

The rescue is not a list-position cue and does not depend on familiar letters: scrambling the handle list leaves retrieval unchanged (0.84 vs 0.80), an ordinal-only anchor ("the second source") gives far less (0.46), and nonce handles ("Source Zeb/Kip/Vox") restore retrieval equally (0.78 vs 0.80) while reliability-conditioned answering falls (0.41 vs 0.61); a decomposition locates the nonce cost in mapping reliability statements to source selection, with retrieval, use and a direct "Source Zeb is reliable" instruction equal (0.78 / 0.79 / 0.78) and selection from statements about the other sources at 0.43 (Appendix I). Pixtral-12B shows the same list-order pattern (0.76 vs 0.73; ordinal 0.31).

The same control decomposes the remaining failure: retrieval and use ("Use Source A") are equally available after the bottleneck (0.80 and 0.83) while reliability-conditioned selection lags (0.61; +0.22 [+0.16, +0.29]), so in Qwen3-VL-8B addressability is intact and reliability-to-source control is the bottleneck; and the rescue persists when the turn-1 cache entries are physically evicted instead of masked (0.81 vs 0.80; agreement 0.99), so it is carried through the intermediate computation rather than by retained access to the handle tokens. Pixtral-12B replicates the reassignment (the answer follows the letter in 0.74 [0.70, 0.78] of pairs; color clean, count and chart partial) and the eviction persistence (0.74 vs 0.74), but shows no reliability bottleneck once the handle is addressable (0.74 / 0.76 / 0.74), so that bottleneck is Qwen-specific.

Three levels of intermediate text are therefore distinguished: no source referent, a persistent referent without value binding, and explicit source-value provenance. A referent alone, whether a native name or an arbitrary handle assigned before integration, recovers much of retrospective source retrieval and a smaller part of reliability-conditioned behavior, while explicit source-value binding adds substantial benefit beyond identity continuity, most for the reliability cue. We do not claim that exact lexical token matching is the mechanism: the referent must be one the model can map

(a renaming of the sources is not resolved even with the evidence accessible, Appendix G), not a modality word.

5.6 Modality versus veridicality

A sensor-corruption control (Appendix I) separates the visual percept from its certainty: the target shape's color is mixed toward the color the caption reports on the 60 three-source color items, keeping only items the model still reads as the shown color. The manipulation reduces the model's confidence (first-token margin 2.13 to 1.50) but leaves the blocked-neutral default essentially unchanged (image share 1.00 to 0.98), which weakens a simple certainty-only explanation of the default without ruling out every confidence or veridicality mechanism. The stronger corruption level is underpowered (19/60 readable items) and Pixtral-12B is directionally similar (margin reduced, default 0.83 to 0.83).

5.7 Causal interventions

Sufficiency (exploratory). Residual-stream patching could redirect answers from a late question-token state carrying the reliability statement, but source-specific patching controls were invalid, so we draw no mechanistic conclusion from that analysis; the behavioral handle-permutation control (Section 5.5) is where source-directedness is established.

Handle necessity (pre-specified). Is later access to the preserved handle tokens necessary? On the 90 stable-handle scenes with the evidence blocked, additionally blocking attention from the final turn to the two summary tokens of the queried handle drops non-default retrieval from 0.79 to 0.22 and use from 0.82 to 0.17, while a matched sham on the two adjacent tokens ("consulted:") leaves both unchanged (0.78 and 0.82; specificity +0.56 [+0.48, +0.64] and +0.65 [+0.58, +0.72]; meeting the pre-specified necessity criterion). The handle remains readable in the final question; the intervention removes only its earlier occurrence in the intermediate summary, which links the later query to the integrated source information. Under handle ablation answers revert to the default or become unparseable rather than moving to the third value, whereas blocking the value token sends them to the third value (Figure 1C), so each arm removes what it targets. The combined reliability cue falls below its 0.70 prerequisite at 0.62 and is reported descriptively (0.14 ablated against 0.50 sham; its drop is not larger than retrieval's, -0.09 [-0.18, -0.01]). Pixtral-12B fails the prerequisite (0.68) and shows the same pattern descriptively (0.32 ablated against 0.69 sham). This establishes causal necessity of access to the preserved handle-token positions at question time, not the identification of a provenance circuit (Appendix I).

5.8 Cross-model and naturalistic boundaries

Pixtral-12B reproduces the referent hierarchy, the counterfactual tracking and the eviction persistence (Sections 5.3, 5.5) despite under-updating with the evidence in view and failing the original replication criterion's no-bottleneck threshold (Appendix G).

A natural-image validation (100 VQAv2 photographs with one conflicting color attribute; Appendix I) reproduces the ordering in Qwen3-VL-8B with the evidence blocked: a neutral summary 0.015, stable handles 0.87 (retrieval) and 0.61 (reliability-conditioned), explicit binding 1.00 and 0.93, text readers at chance. It is a small controlled check on one attribute and one model, not evidence of broad generalization.

A modality stress test over all 270 neutral-label three-source items (Appendix I) shows that stable-handle rescue of retrieval extends to non-default visual targets (+0.41 [+0.29, +0.52]) and to both image and text defaults, but is stronger for text targets (+0.65 [+0.59, +0.70]; difference +0.24 [+0.12, +0.37]); the visual-target evidence comes almost entirely from the two-image set, and default modality is observed rather than assigned, so the contrast is descriptive.

The two-source effect is robust in sign and wording-dependent in size over six cue wordings, and an emphatic wording lowers Qwen's blocked-explicit cell to 0.63-0.68, so explicit provenance is not unconditionally sufficient (Appendix G).

Explanation	Assessment	Evidence
Trust-update failure with association intact	Not supported	predicts retrieval above ceiling with trust at floor; both at floor
Late source-label binding failure	Remains possible	explains BN, BE and QUERY; needs a default
Persistent non-textual default	Consistent with results	Qwen ITT default above the text reader; default-source QUERY 0.98 / 0.80
Visual-modality prior (default = image's value)	Remains possible	ITT default follows the image across slots; disappears in IIT
Certainty / truth tag (default = perceived value)	Weakened by control	sensor-corruption control: certainty reduced, default unchanged (5.6)
Primacy (default = first-presented source)	Weakened by control	ITT default on the image at every slot; mild first-slot effect in TTT
Lexical antecedent / referential continuity	Consistent with results	anchor summaries restore blocked QUERY 0.09 -> 0.88 (Qwen), 0.20 -> 0.71 (Pixtral); text reader at chance
Lossy by construction (text omits provenance)	Remains possible for non-default sources	explains interaction and non-default retrieval failure, not the default or the rescue

Idefics3-8B and InternVL3-8B fail the accessible-evidence prerequisite (they do not follow a late source statement even with the evidence in view; Appendix H, I) and therefore support no conclusion about the mechanism.

6 Alternative Explanations

Table 3 lists the explanations the paper argues about; with three candidates the models keep the default rather than avoid it, so a two-choice shortcut does not explain the results, and answer leakage, an incomplete mask and generic instruction failure are ruled out by the controls (Sections 5.3, 5.4; Appendix G).

7 Limitations

Stimuli are synthetic apart from a 100-item natural-image color validation, with one conflicting attribute, and the image's value is always the true value of the scene; summaries are inserted and templated, and model-written summaries were not tested. The mask (or equivalent cache eviction) removes direct access rather than information and is not claimed to resemble a deployed system. The stable-handle effect is established mainly for addressing a text source against an image default; visual-target rescue is smaller and its evidence comes almost entirely from the two-image set, and tracking is weaker for the third-presented source. Models differ: reliability-conditioned selection lags retrieval in Qwen3-VL-8B but not in Pixtral-12B, and two further models fail the accessible-evidence prerequisite, so nothing here supports claims about moderators. The necessity test is token-level, confined to one summary wording and to a sham on adjacent function tokens; it does not identify a circuit or say at what stage of computation the association becomes unusable. Effect sizes swing with cue wording, and a renaming of the sources is not resolved even with the evidence accessible.

8 Discussion

In terms of Section 3, a carried referent restores capacity (A) and, in Qwen3-VL-8B, less of (B, C); explicit source-value binding restores both. The carried signal itself remains constrained rather than settled: primacy is weakened and a certainty-only account is unsupported but not excluded (Section 5.6). In a continuous-context or key-value-cache system the evidence's hidden states persist and the failure we measure can arise; with serialized text memory a summary that omits provenance loses source identity by construction. We do not model a deployed system or estimate how often the failure occurs.

9 Conclusion

We asked what a vision-language model can do with a late, source-conditioned instruction once the evidence is gone and only an intermediate representation remains. After a source-neutral bottleneck, later source-conditioned access largely collapses unless a source referent is carried through

integration. Arbitrary handles assigned before integration restore most named-source retrieval, handles introduced afterwards do not, answers follow a handle when its source assignment is permuted, and the effect persists when the evidence is physically evicted from the cache. Causally, access to the carried handle's tokens at question time is necessary for that retrieval; whether the handle addresses provenance held in hidden states or re-derives it we do not know, and we identify no circuit. Two things remain model- or modality-dependent: the rescue is weaker for visual than for text targets, and reliability-conditioned selection lags retrieval in Qwen3-VL-8B but not in Pixtral-12B. For intermediate representations meant to support later correction, the implication is concrete: source-conditioned access after integration is determined not only by which values survive but by whether a usable source referent survives with them, and explicit source-value binding further supports reliability-conditioned selection.

AI USE STATEMENT

Large language models (LLMs) were used as supporting tools during this work. The authors made the scientific decisions and take full responsibility for the manuscript and results.

Writing support. LLM assistants were used to help revise and polish portions of the manuscript. All text was reviewed and edited by the authors.

Literature discovery. LLM assistants were used to help identify potentially relevant related work. The closest papers were checked against their primary sources.

Research and coding assistance. LLM assistants were used to assist with research workflows, coding, debugging, and analysis. The authors determined the research questions, experimental designs, analysis plans, and interpretation of the results, and reviewed the relevant code and outputs.

REPRODUCIBILITY STATEMENT

Exact prompts and conversation structure (Appendix B), stimulus construction and item counts (Appendix C), model revisions, inference settings, statistical procedures and decision rules (Appendix D), the evidence-removal implementation with its completeness checks (Appendix E), and the full central-result tables (Appendices F to I) are provided in the appendix. Analysis rules were fixed before model outputs. Code, stimuli and generation records will be released with the camera-ready version; the natural-image items are public VQAv2 validation questions identified by question id.

REFERENCES

Mihir Shriniwas Arya. RECON: Benchmarking Agent Memory for Compositional Reasoning over Long Contexts. *arXiv preprint arXiv:2607.16716*, 2026.

Joshua Ashkinaze, Laura Kurek, Alina Faisal, Tongyuan Miao, Mariam Joseph, Ceren Budak, and Eric Gilbert. Information Discernment in Large Language Models. *arXiv preprint arXiv:2607.19355*, 2026.

Ivo Brink, Alexander Boer, and Dennis Ulmer. Probing for Knowledge Attribution in Large Language Models. *arXiv preprint arXiv:2602.22787*, 2026.

Wenxiao Fan, Jingling Fu, Fang Li, Luohang Liu, Yu He, Lichen Ma, Zhiyang Yu, Weishan Bi, Junshi Huang, Yan Li, Gu Simiu, and Kan Li. Recompute or Reuse? Diagnosing and Mitigating Textual Shortcuts in VLM Self-Reflection. *arXiv preprint arXiv:2608.01930*, 2026.

Jiahai Feng and Jacob Steinhardt. How do Language Models Bind Entities in Context? In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.17191.

Rodela Ghosh, Aviral Gupta, and Guangjing Wang. Which Source Wins? Task-Dependent Reliance in Vision-Language Models. *arXiv preprint arXiv:2608.17205*, 2026.

Etha Tianze Hua, Tian Yun, and Ellie Pavlick. Source-Modality Monitoring in Vision-Language Models. *arXiv preprint arXiv:2604.22038*, 2026.

Tianze Hua, Tian Yun, and Ellie Pavlick. How Do Vision-Language Models Process Conflicting Information Across Modalities? *arXiv preprint arXiv:2507.01790*, 2025.

Yifan Jia, Kailin Jiang, Yuyang Liang, Qihan Ren, Yi Xin, Rui Yang, Fenze Feng, Mingcai Chen, Hengyang Lu, Haozhe Wang, Xiaoye Qu, Dongrui Liu, Lizhen Cui, and Yuntao Du. Benchmarking Multimodal Knowledge Conflict for Large Multimodal Models. *arXiv preprint arXiv:2505.19509*, 2025.

Alex Kwon. Reclaim Evaluation: A Lossy Memory Is Worse Than an Empty One. *arXiv preprint arXiv:2606.25449*, 2026a.

Alex Kwon. Manufactured Confidence: How Memory Consolidation Turns Hearsay into Confident Facts. *arXiv preprint arXiv:2606.29279*, 2026b.

Jun Wen Leong. Recognition Without Enforcement: Configuration-Dependent Failures in LLM Agent Instruction Arbitration and External Control. *arXiv preprint arXiv:2608.28502*, 2026.

Trang Nguyen, Jackson Michaels, Madalina Fiterau, and David Jensen. Challenges in Understanding Modality Conflict in Vision-Language Models. *arXiv preprint arXiv:2509.02805*, 2025.

Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. Not All Contexts Are Equal: Teaching LLMs Credibility-aware Generation. In *Proceedings of EMNLP*, 2024. arXiv:2404.06809.

Athulith Paraselli, Etha Tianze Hua, and Ellie Pavlick. Slow to See, Slow to Suppress: Understanding the Effects of Modality in Context-Memory Conflicts. *arXiv preprint arXiv:2609.00293*, 2026.

Rohan N. Pradhan and Steve Goley. Trust, but Don't Verify: Epistemic Blind Spots in LLM Source Evaluation. *arXiv preprint arXiv:2606.05403*, 2026.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language Models use Lookbacks to Track Beliefs. *arXiv preprint arXiv:2505.14685*, 2025.

Saurabh Ranjan, Konstantina Sokratous, and Brian Odegaard. Reality Monitoring in Large Language Models: Self-Knowledge That Transforms with Conversation Memory. *arXiv preprint arXiv:2607.23927*, 2026.

Jakob Schuster, Vagrant Gautam, and Katja Markert. Whose Facts Win? LLM Source Preferences under Knowledge Conflicts. In *Proceedings of ACL*, 2026. arXiv:2601.03746.

Alessandro Pietro Serra, Francesco Ortu, Emanuele Panizon, Lucrezia Valeriani, Lorenzo Basile, Alessio Ansuini, Diego Doimo, and Alberto Cazzaniga. The Narrow Gate: Localized Image-Text Communication in Native Multimodal Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2412.06646.

Zineddine Tighidet, Andrea Mogini, Jiali Mei, Benjamin Piwowarski, and Patrick Gallinari. Probing Language Models on Their Knowledge Source. In *Proceedings of the BlackboxNLP Workshop (EMNLP)*, 2024. arXiv:2410.05817.

Jinghan Xu, Yiyong Xiao, Wanru Shao, Hankai Liu, and Xinjin Li. Memory Provenance Laundering in LLM Agents: A Non-Amplification Firewall for Persistent Memory. *arXiv preprint arXiv:2607.29167*, 2026.

Yu Zhang, Mufan Xu, Xuefeng Bai, Kehai Chen, Pengfei Zhang, Yang Xiang, and Min Zhang. Instruction Anchor: Dissecting the Mechanistic Dynamics of Modality Arbitration. *arXiv preprint arXiv:2602.03677*, 2026.

Paper (arXiv)	late reliability	conflicting sources	multi-modal	evidence inaccessible	intermediate provenance manipulated	retrieval vs reliability	causal intervention
Source-Modality Monitoring (2604.22038)	N	Y	Y	N	N	retrieval only	Y
Instruction Anchor (2602.03677)	N (late modality instruction)	Y	Y	N (windowed knockout)	N	N	Y
Router heads (2507.01790)	N	Y	Y	N	N	N	Y
MMKC-Bench (2505.19509)	N	Y	Y	N	N	N	N
Whose Facts Win? (2601.03746)	N	Y	N	N	N	N	N
Trust, but Don't Verify (2606.05403)	N	Y (four)	N	N	N	N (detect vs discount)	Y (probes, tracing)
Reclaim Evaluation (2606.25449)	Y (late correction, no source)	N	N	Y (fresh session)	Y (retained content, not source)	N	N
Manufactured Confidence (2606.29279)	partial (distrust at read)	mostly N	N	Y	Y (attribution framing)	reliability only	N
THIS WORK	Y	Y	Y	Y (verified mask; cache eviction)	Y (neutral / referent-only / explicit)	Y (named-source QUERY vs TRUST / DISTRUST)	Y (one model, exploratory)

A Additional Related Work

Table A1. Closest work against the dimensions of our setting. Y = present; N = absent; (.) = qualification.

A.1 Multimodal knowledge conflict and modality arbitration

"Challenges in Understanding Modality Conflict in VLMs" (Nguyen et al., 2025) uses synthetic colored shapes with captions asserting another color, close to our color stimuli, and studies conflict detection and resolution with linear probes; the authors note the absence of causal tests. Synthetic shape/caption conflict stimuli are therefore not new. Our setting differs in that reliability is stated, stated late, and evaluated with and without access to the evidence.

"Which Source Wins? Task-dependent reliance in VLMs" (Ghosh et al., 2026) studies image–text conflicts in which one source is degraded in legibility, and stresses that legibility is not factual reliability; reliability is never stated to the model. We adopt its lesson that source naming changes effects. Our setting states factual reliability explicitly and late.

MMKC-Bench (Jia et al., 2025) evaluates multimodal knowledge conflicts, including an inter-context condition in which two pieces of evidence (each an image with a description) conflict through their descriptions, with four-option questions. It contains text-versus-text conflict inside multimodal evidence and more than two answer options; it has no reliability cue, no evidence removal, and does not use same-modality conflict as a control for a cross-modal effect.

"Slow to See, Slow to Suppress" (Paraselli et al., 2026) studies context-versus-parametric conflicts while varying whether the entity is shown as an image or as text; the conflicting context is always text. It is a modality-of-entity control, not a source-conflict control.

A.2 Source monitoring, attribution and provenance

"Source-Modality Monitoring in VLMs" (Hua et al., 2026) is the closest behavioral neighbor. It presents inconsistent image–caption pairs to eleven VLMs and asks what the image, or the caption, contains; it studies marker tokens and source-cue wording, shows with "freeze-remove" patching that source identity propagates into content-token states, and learns vectors that induce source misattribution. It establishes that VLMs track source modality, that this can be manipulated, and that cue wording matters. Our failing source-name paraphrase echoes its cue findings and is not a new observation. The differences are that in its setting the target source is named at query time with the evidence in view, the task is retrieval rather than a change of trust, and no intermediate representation or evidence removal is involved.

Text-only knowledge-source probing (Brink et al., 2026; Tighidet et al., 2024) shows that whether a statement derives from context or parametric memory is linearly decodable. "Reality Monitoring in LLMs" (Ranjan et al., 2026) tests whether models distinguish self-generated from user-provided items as a function of how conversational memory is structured; it is a retrieval-type source test without a reliability manipulation. "Memory Provenance Laundering" (Xu et al., 2026) observes that memory consolidation can rewrite a low-trust observation as apparent user history; the observation that summaries lose provenance is thus established at the systems level.

A.3 Credibility-aware reasoning and source reliability

Credibility-aware generation (Pan et al., 2024) attaches a credibility label to each retrieved document at input. "Whose Facts Win?" (Schuster et al., 2026) places source type in the evidence itself (table headers) and measures which source wins; its three-table condition still has two candidate values. "Trust, but Don't Verify: Epistemic Blind Spots in LLM Source Evaluation" (Pradhan & Goley, 2026) presents four named sources with credibility cues inside their messages and conflicting numeric estimates, and includes probes and causal tracing; it reports a dissociation between detecting invalid statistics in isolation and discounting them in synthesis. "Information Discernment" (Ashkinaze et al., 2026) pits a named source's claim against the model's prior. In all four, read at the level of methods and prompts, credibility information arrives with or before the evidence, the evidence remains in context, no intermediate representation stands between evidence and answer, and no source is declared unreliable afterwards. They establish that models use (and misuse) concurrent credibility cues; they do not address late updates.

A.4 Memory compression, retraction and late correction

This is where the structural precedent lies. "Reclaim Evaluation: A Lossy Memory Is Worse Than an Empty One" (Kwon, 2026a) runs a text-only pipeline of problem, planted error, compression into a note whose retained content is the manipulated variable (with a length-matched control), a fresh session in which the original is gone, and then a late correction. Its authors state plainly that "the information loss is definitional; the finding is behavioral". That sentence applies to our interaction as well. What it lacks relative to our setting: several sources in conflict, a reliability statement about a named source, any multimodal content, internals, and a condition in which hidden states (rather than only text) cross the bottleneck. "Manufactured Confidence" (Kwon, 2026b) applies a distrust instruction to consolidated, source-stripped agent memory and finds blanket escalation unless a second source is available, and that attribution wording in the memory text does not protect; the distrust text accompanies the memory rather than arriving later about one of several conflicting sources. RECON (Arya, 2026) includes formal invalidations that arrive later in long case files and are evaluated through lossy memory systems, but reliability of a named source is never asserted to the model. Our contribution is instead to isolate post-integration source addressability under evidence removal and separate it from reliability-conditioned selection.

Two text-only mechanistic studies are structural analogues. "Lookbacks" (Prakash et al., 2025) shows a later visibility statement changing which previously bound state applies, via pointer-like mechanisms; "Recognition Without Enforcement" (Leong, 2026) shows that the channel of an instruction is decodable at high accuracy yet not enforced. "Decodable but not used" is thus an established motif; our setting extends this distinction to multimodal evidence after source access has been removed.

A.5 Mechanistic studies of multimodal arbitration

"Instruction Anchor" (Zhang et al., 2026) is the closest mechanistic precedent. With image and text contexts supporting different answers and an instruction (placed after the evidence, in the same prompt) to follow one modality, it uses pathway attention knockout, logit-lens analysis on instruction tokens, and head blocking/amplification to localize arbitration to instruction-token states in Qwen2.5-VL-7B, InternVL3 and DeepSeek-VL2. "How do VLMs process conflicting information" (Hua et al., 2025) recovers both labels with probes (rejecting loss of provenance in its setting) and identifies modality-agnostic "router" heads and modality-specific heads by last-token head scaling. Given these, an instruction-token locus for arbitration cannot be claimed as new. Our mechanistic question is narrower: does that locus remain sufficient when the evidence itself can no longer be

attended and only a provenance-neutral intermediate is readable, and is the effect specific to the content of the instruction? "The Narrow Gate" (Serra et al., 2024) uses attention knockout on a natural one-token bottleneck; attention knockout as a method is not new either. "Recompute or Reuse" (Fan et al., 2026) masks attention to a stale reasoning trace at all layers; our masks block evidence rather than reasoning.

B EXACT PROMPTS

All prompts are rendered with each model's native chat template; no system prompt. {q} is the item question in lower case (e.g. "how many orange triangles are there?"), {Q} the same with its first letter capitalized, {kind} is "number" or "colour" from the item's answer_kind. Every answer turn ends with the tail "Answer with the {kind} only."

B.1 TWO-SOURCE DESIGN (PILOT, THE CONFIRMATORY TWO-SOURCE EXPERIMENT,,,, AND THE THE THREE-SOURCE EXPERIMENT IT2 ANCHOR)

Capability checks (C0): C0i (image only): [image] "{Q} Answer with the {kind} only." C0c (caption only): "Caption: {caption}\{}n\{}nAccording to the caption, {q} Answer with the {kind} only."

Turn 1 (user), evidence arms (AN, AE, BN, BE): [image] then "Caption: {caption}\{}n\{}nYou are given an image and a caption. Note what each one says about {attribute_phrase}." Turn 1 (user), text-only readers (TE, TN): "You are given an image and a caption. Note what each one says about {attribute_phrase}." with no image and no caption.

Turn 2 (assistant, inserted), templated summaries : COUNT: s_img = "The image shows {v_img} {type}." s_cap = "The caption says {v_cap} {type}." COLOR: s_img = "The image shows a {v_img} {type}." s_cap = "The caption says the {type} is {v_cap}." CHART: s_img = "The image shows a value of {v_img} for bar {X}." s_cap = "The caption says a value of {v_cap} for bar {X}." explicit = "{s_img} {s_cap}" when the image value is listed first, else "{s_cap} {s_img}". neutral = "Reported {attribute_noun}: {v_1} and {v_2}." (v_1, v_2 in the same counterbalanced order; e.g. "Reported counts of red squares: 5 and 3.")

Turn 3 (user), original reliability cue c1 : image unreliable: "The image was corrupted and is unreliable; the caption was reliable." caption unreliable: "The caption was corrupted and is unreliable; the image was reliable." followed on the same line by " Based on the reliable source, {q} Answer with the {kind} only." (The pilot's reliability-BEFORE arms used "Note: the {bad} is corrupted and unreliable; the {good} is reliable." at the top of turn 1 and the final turn "Based on the reliable source, {q} Answer with the {kind} only."; these arms were dropped from the confirmatory two-source experiment onward.)

No-summary baseline NS : [image] then "Caption: {caption}\{}n\{}n{reliability sentence} Based on the reliable source, {q} Answer with the {kind} only."

emphatic cue and the six cues. Each replaces ONLY the late reliability statement; the final turn becomes "{cue}\{}n{Q} Answer with the {kind} only." (for NS the same two lines follow "Caption: {caption}\{}n\{}n"). c1 keeps the exact the confirmatory two-source experiment/ wording above. The mirrored (caption-unreliable) wording swaps the source names. c1_original (unreliable-first; the confirmatory two-source experiment/ wording, past tense): as above. c2_concise (unreliable-first): image unreliable "The image is unreliable. Use the caption." / caption unreliable "The caption is unreliable. Use the image." c3_disregard (reliable-first): "The caption is trustworthy; disregard any information that originated from the image." / "The image is trustworthy; disregard any information that originated from the caption." c4_modality (unreliable-first; modality paraphrase): "The visual source is corrupted; base your answer on the textual source." / "The textual source is corrupted; base your answer on the visual source." c5_verbose (reliable-first; long, narrative): "An update since you looked at the material: the caption has been checked and verified as accurate, whereas the image you were given turned out to have been altered and can no longer be trusted. Please rely only on the trustworthy one of the two when you answer." / "An update since you looked at the material: the image has been checked and verified as accurate, whereas the caption you were given turned out to have been altered and can no longer be trusted. Please rely only on the trustworthy one of the two when you answer." c6_strong_v5 (unreliable-first; the emphatic wording, one realization among six,

not canonical): “IMPORTANT: The IMAGE source has been determined to be corrupted and must not be used. The CAPTION source is reliable. Answer using only the information that came from the reliable source.” / “IMPORTANT: The CAPTION source has been determined to be corrupted and must not be used. The IMAGE source is reliable. Answer using only the information that came from the reliable source.” note: “Six realizations of the SAME late reliability information (which source is unreliable, which is reliable). No cue names or hints at any value. Fixed before any output.”

the confirmatory two-source experiment pre-specified three-source diagnostic (audit record agreeing with the caption; first 40 items per family): Turn 1: [image] “Caption: {caption}\{}n{audit}\{}n\{}nYou are given an image, a caption and an audit record. Note what each one says about {attribute_phrase}.” where {audit} = “Audit record: {caption sentence with v_cap}.” D0 / B0: “{Q} Answer with the {kind} only.” D1 / D1n: “The image was corrupted and is unreliable; the caption and the audit record were reliable. Based on the reliable sources, {q} Answer with the {kind} only.” D2 / D2n: “The caption and the audit record were corrupted and are unreliable; the image was reliable. Based on the reliable source, {q} Answer with the {kind} only.”

B.2 Model-written summaries Turn 1 instruction (replaces the last sentence of the ordinary turn 1): “You are given an image and a caption. “ followed by exactly one of explicit: “Reply with one sentence stating what the image shows and what the caption says about {attribute_phrase}.” neutral: “Reply with one sentence that states both reported values for {attribute_phrase} in ascending order (alphabetical order for colours) and does not say where either value came from.” The model generates the summary with no mask, greedy, at most 48 new tokens; the text is inserted as the assistant turn and the identical turn 3, mask and scoring are applied (ME, MN; TMN = text-only control for MN with the same summary text). Compliance (decided from the summary text alone, before any turn-3 generation, never from outcomes): neutral compliant iff finish = eos AND both v_img and v_cap appear (digits or number words normalized; color words) AND none of {image, picture, photo, figure, chart, graph, plot, caption, text, written, shows, shown, says, said, states, according, visual} appears as a whole word. explicit compliant iff finish = eos AND a clause containing “image|picture|chart|figure” contains v_img AND a clause containing “caption” contains v_cap (clauses split on “.”, “;” or “, and / while / whereas / but”). ME/MN/TMN are analyzed on compliant items only; non-compliant items are still run and reported separately.

B.2 Three-source design

Turn 1 (user): the three sources in presentation order, each introduced by its header, then the instruction “You are given three sources about the same scene. Keep track of what each one reports about {attribute_phrase}.” (Amendment the three-source experiment.4: the verb “Note” was removed because it collided with the source name “field note” and with the NEUTRAL-scheme word ban.) Headers, NATIVE scheme: “Image:” (image follows), “Caption: {text}”, “Field note: {text}”, “Catalogue entry: {text}”, “Image 1:”, “Image 2:”. Labels used in later turns: “the image”, “the caption”, “the field note”, “the catalogue entry”, “image 1”, “image 2”. Headers and labels, NEUTRAL scheme: “Source A:”, “Source B:”, “Source C:” (letters assigned by the balance matrix, not always alphabetical in presentation). Text-source sentences use the two-source caption templates: “There are {n} {colour} {shape}s.” / “The {shape} is {colour}.” / “Bar {X} has value {v}.” SHAM arm only: turn 1 begins with “Background (unrelated to the task): {filler}\{}n\{}n” (filler sentences drawn item-seeded from, 72 sentences, to at least 0.95 x the item’s blocked-evidence token count).

Turn 2 (assistant, inserted): neutral: “Reported {attribute_noun}: a, b and c.” (values in the counterbalanced summary order) explicit: one sentence per source in presentation order. NATIVE: “The image shows...”, “Image 1 shows...”, “The caption says...”, “The field note says...”, “The catalogue entry says...” with the per-family phrasing of (article corrected to “an orange...”). NEUTRAL: “Source B reports...” with the image phrasing for every source, so the sentence form carries no modality cue.

Final turn (user) (X = target source label, Y and Z = the other two in presentation order; cap capitalizes the first letter): NONE: “{Q} Answer with the {kind} only.” QUERY-X: “According to {X}, {q} Answer with the {kind} only.” TRUST-X w1: “{cap(X)} was found to be reliable.\{}n{Q} Answer with the {kind} only.” TRUST-X w2: “Trust only {X}.\{}n{Q} Answer with the {kind} only.” DISTRUST-YZ w1: “{cap(Y)} and {Z} were found to be unreliable.\{}n{Q} Answer with the {kind} only.” DISTRUST-YZ w2: “Do not trust {Y} or {Z}.\{}n{Q} Answer with the {kind}

only." COMBINED-X: "{cap(Y)} and {Z} were found to be unreliable; {X} was reliable. Based on the reliable source, {q} Answer with the {kind} only." (bridge to the confirmatory two-source experiment–; ITT-NATIVE, AN and BN only)

Capability checks, one per source: image source: [image] "{Q} Answer with the {kind} only." text source: "{header} {text}\{}n\{}nAccording to {label}, {q} Answer with the {kind} only."

One item per family, verbatim from (ITT, NATIVE, main; design row 0): (seed 83053017). Presentation order: Field note "There are 4 orange triangles." (4) · Caption "There are 8 orange triangles." (8) · Image (6 orange triangles among 3 blue circles). Question "how many orange triangles are there?" neutral_summary: "Reported counts of orange triangles: 8, 6 and 4." explicit_summary: "The field note says 4 orange triangles. The caption says 8 orange triangles. The image shows 6 orange triangles." (seed 83054017). Field note "The circle is blue." (blue) · Caption "The circle is purple." (purple) · Image (green circle, orange triangle, red square). Question "what colour is the circle?" neutral_summary: "Reported colours of the circle: purple, green and blue." explicit_summary: "The field note says the circle is blue. The caption says the circle is purple. The image shows a green circle." (seed 83055017). Field note "Bar C has value 1." (1) · Caption "Bar C has value 6." (6) · Image (bars A 9, B 5, C 4, D 2). Question "what is the value of bar C?" neutral_summary: "Reported values of bar C: 6, 4 and 1." explicit_summary: "The field note says a value of 1 for bar C. The caption says a value of 6 for bar C. The image shows a value of 4 for bar C." Other schemes and source sets (count family, row 0): (NEUTRAL): Source B "There are 3 green triangles." (field note) · Source A "There are 8 green triangles." (caption) · Source C (image, 6). neutral "Reported counts of green triangles: 8, 6 and 3." explicit "Source B reports 3 green triangles. Source A reports 8 green triangles. Source C reports 6 green triangles." : Catalogue entry "There are 2 blue circles." · Field note "There are 8 blue circles." · Caption "There are 4 blue circles." neutral "Reported counts of blue circles: 8, 4 and 2." explicit "The catalogue entry says 2 blue circles. The field note says 8 blue circles. The caption says 4 blue circles." : Caption "There are 2 red squares." · Image 1 (8) · Image 2 (5). neutral "Reported counts of red squares: 8, 5 and 2." explicit "The caption says 2 red squares. Image 1 shows 8 red squares. Image 2 shows 5 red squares."

B.4 Lexical-anchor summary L = "Sources consulted: {label1}, {label2} and {label3}. Reported {attribute_noun}: a, b and c." Labels in presentation order with native names ("the image", "the caption", "the field note"); values in the the three-source experiment counterbalanced summary order; no value bound to any source. For this gives "Sources consulted: the field note, the caption and the image. Reported counts of orange triangles: 8, 6 and 4." Final turns: QUERY-X and COMBINED-X exactly as in B.3; NONE for the default.

C Stimulus Generation

C.1 Renderer (all families, all experiments; PIL only, seeded, deterministic) CANVAS = 448 x 448 px; BACKGROUND = (230, 230, 230) (#e6e6e6); OUTLINE = (40, 40, 40), OUTLINE_W = 2; supersampling SS = 4 with LANCZOS downsampling (shapes); GRID = 4 x 4 cells of 112 px (at most 16 objects); MARGIN = 8 px from any cell border (so >= 8 px from the canvas edge and >= 16 px between bounding boxes); SIZE_RANGE = (56, 72) px. SHAPES = ("circle", "square", "triangle"). COLORS (palette, 5 named colors): red (220, 40, 40) · blue (40, 90, 220) · green (40, 160, 70) · orange (245, 140, 20) · purple (140, 60, 190). COUNT_RANGE = range(2, 9), i.e. 2..8. Chart: 4 bars labeled A–D on the x-axis, integer heights 1..9, y-axis ticks 0..10 with gridlines, no value labels on bars; the fixed chart renderer uses ImageFont.load_default(size=...) and therefore needs Pillow >= 10.1 (the three-source experiment and images were generated with Pillow 12.3.0; the manifest is a pure function of the code and the design matrix). Every rendered image is pixel-verified at generation time (object-center / bar pixels equal the palette color; background pixel check).

C.2 Two-source families (–, the three-source experiment IT2 anchor) COUNT: n_img, n_cap in 2..8 with |n_img − n_cap| >= 2; one distractor type (a different shape AND a different color); distractor count 2..5, never equal to either claimed value, and n_img + n_d never equals the caption's value. Caption "There are {n_cap} {colour} {shape}s."; question "how many {colour} {shape}s are there?". COLOR: three objects, one per shape; the target shape has color c_img; c_cap is absent from the image; the two other shapes take two colors drawn from the palette colors outside {c_img, c_cap}. Caption "The {shape} is {c_cap}."; question "what colour is the {shape}?". CHART:

target bar X has height v_img; caption “Bar X has value {v_cap}.” with |v_img − v_cap| >= 2 and v_cap not equal to any bar’s height (no bar may show the caption’s value); from on the 4-height combination of every chart is unique across sets (seeded redraw, ‘avoid‘). Question “what is the value of bar X?”. In every item the image and the caption disagree on exactly one attribute; the image’s value is the true value of the rendered scene; the image precedes the caption in turn 1; the visual fact is never printed as its value. Counterbalancing (exact within family): summary order (image value first / caption value first) x which source has the larger (alphabetically first) value; – use with 4 cells of equal size where divisible (IT2: 30 per family -> 8/7/7/8; dev 2/1/1/2). Seeds: BASE_SEED = 20260918 (pilot : count seeds BASE + 1000 + i, color BASE + 2000 + i, N_PER_FAMILY = 24); BASE_SEED_V2 = 73450019 (the confirmatory two-source experiment: 120 per family, “all the confirmatory two-source experiment seeds are disjoint from every seed”); – use fresh base seeds with the identical specification (“Identical specification to, new base seed; - are frozen (hash-checked in tests)”); the three-source experiment BASE_SEED_V7 = 83052017. Every manifest ships with a SHA256 guard file („ ; no exists in the repo). Manifest sizes (lines): 48; 360; 270 and 270 (gate + main parts); 180; 120; 108 (90 main + 18 dev); 60.

C.3 Three-source items Families count, color, chart; source sets ITT (image, caption, field note), TTT (caption, field note, catalogue entry; no image), IIT (image X, image Y, caption); label schemes NATIVE / NEUTRAL. 30 main items per family x source set x scheme (18 cells, 540 items) + 3 dev items per cell (never analyzed) + IT2 anchor 30 per family (90) + 6 dev per family. has 702 rows; 909 PNGs (ITT 198, IIT 396 as ‘_x‘/‘_y‘ renderings, IT2 108, blurred copies 207 = (60 SUBSET-S + 9 dev) x 3 Gaussian-blur radii {2, 4, 6} px, files “). SUBSET-S = design rows 0..19 of ITT-native main (60 items). Values: three distinct values per item. count 2..8 and chart 1..9 keep the rule that any two claimed values differ by >= 2 (count: the 10 such triples x 3 per cell; chart: 30 of the 35 triples, each at most once per cell); color: the 10 triples of the 5-color palette x 3 per cell. Targets spread evenly (count: 15 shape-color types x 2; color: 3 shapes x 10; chart: bars 7/8). Distractors: no distractor carries, counts or totals a claimed value. count — one distractor type (other shape AND color), n_d not in the three values and n_shown + n_d not in them; n_d in 2..5, with 6 used only when 2..5 is impossible (15 of 198 count images; at most 8 + 6 = 14 objects). color — the two other shapes take the two palette colors outside the triple. chart — non-target bars avoid all three values and every 4-height combination is unused in the confirmatory two-source experiment– and within the three-source experiment (234 unique). IIT pairs differ only in the target attribute (count: same distractors and positions, the smaller target set is a prefix of the larger; color: only the target’s fill; chart: only the target bar’s height). Counterbalancing is GIVEN by the analytic matrix (sha256 3f32648419e8c6dc58c3ff6aebed3621263ec1615120ae81479a212f50503e4b), 30 rows, each a quadruple of permutations of the three presentation slots: P (kind order), L (letter A/B/C, neutral scheme), S (position of each slot’s value in the neutral summary), V (rank of each slot’s value). Row j is applied to the j-th item of every cell; no search, solver or rejection loop over designs is run. Realized balance: every 3 x 3 table (slot x kind, slot x value rank, slot x summary position, kind x value rank, kind x summary position, summary position x value rank, and in neutral cells slot x letter, kind x letter, letter x summary position, letter x value rank) is exactly 10/10/10 in every row and column in each of the 18 main cells; each of the 6 orders of P, S, V (and L) occurs exactly 5 times; max pair multiplicity PL 2, PS 1, PV 2, LS 2, LV 2, SV 2. Readable consequences: the image is first in 10/30 ITT items; letters are alphabetical in presentation in exactly 5 of 30 items per neutral cell. SUBSET-S (first 20): slot x kind 6–7 per cell (image first/second/third = 7/6/7), kind x value rank 4–8 (image smallest/middle/largest = 4/8/8; caption 8/4/8; field note 8/8/4). Dev rows 0..2 are not balanced. Value usage over main three-source items: color 108 each; count 2:108 3:54 4:72 5:72 6:72 7:54 8:108; chart 1:81 2:51 3:59 4:55 5:55 6:55 7:58 8:53 9:73 (value RANK, which the matrix balances, is exact). Seeds: cell k (1..42) uses item seeds BASE_SEED_V7 + 1000k + j and design-rng seed BASE_SEED_V7 + 1000k + 999; item seeds 83053017 . . . { } 83094022, asserted disjoint from every – seed and > 200000 away from every earlier base seed. Neutral-scheme wording check: the source sentences, headers, summaries and every final turn contain none of image/picture/photo/caption/note/catalogue/entry (tested through the consumer’s builders).

C.4 Item counts per experiment pilot : 48 items (count, color) x 2 directions · the confirmatory two-source experiment: 360 (120 per family) x 2 = 7,200 answers + 720 model-written summaries · (Pixtral): 160 (80 color + 80 chart) x 2 = 1,920 answers, after a 30-item gate set · : Qwen 360 (reused cells + AE, NS), Idefics3 240 fresh x 2 = 3,840 answers (30-item gate), Pixtral 160 (cells + AE, NS) · : 180 (60 per family), identical across the three models · : 120 (40 per family) x 2 directions x 6 cues = 10,320 answers per model · the three-source experiment: 540 three-source

items + 90 IT2 anchor items; Qwen 42,840 answers, Pixtral 42,840, Idefics3 7,560 · M1 patching: 72 failure-selected items (24 per family; 38 localization / 34 test) · M2: the same 72 (patching) and all 360 items (probes).

C.5 Color-sensor corruption stimuli COLOR family only (counts and charts have no equally principled corruption). Latent target color C* = the value the caption reports; C' = the palette color the existing the three-source experiment image shows. The rendered target shape passes through a linear color-mixing sensor fault of strength s: RGB_shown = s · RGB(C') + (1 − s) · RGB(C*) s = 1.0 is the existing image; s = 0.75 and s = 0.55 are partial corruptions (gen_v8b LEVELS = {75: 0.75, 55: 0.55}; each channel rounded to the nearest integer). Only the target shape's RGB changes; distractor shapes, positions, sizes, background, caption, field note, summaries and all turn wording are held constant. Items: the 60 the three-source experiment ITT COLOR main items (30 native + 30 neutral labels); files {item_id} and (120 images, generated with the unchanged renderer, asserting that exactly one object has the target shape and that its color equals the image value); manifest (60 rows, fields subset_s, degraded_image_paths, v8b_shown = C', v8b_latent = C*).

D Models, Checkpoints, Inference, Gates And Exclusions

D.1 Checkpoints (HF ids @ pinned revisions) Qwen/Qwen3-VL-8B-Instruct @ 0c351dd01ed87e9c1b53cbc748cba10e6187ff3b (36 decoder layers; all experiments) mistral-community/pixtral-12b @ c2756cbbb9422eba9f6c5c439a214b0392dfc998 (Mistral-NeMo LM + Pixtral vision encoder;,,,, the three-source experiment, Exp 1–2) HuggingFaceM4/Idefics3-8B-Llama3 @ fddb4ff79181e55a994674777e06cd5456ce3dc3 (Llama-3.1 LM + SigLIP;,, the three-source experiment prerequisite cells) OpenGVLab/InternVL3-8B-hf @ 259a3b64a14623c0ec91a045cb43f7c5af5fa6af (Exp 4, option B; InternViT vision tower + Qwen2.5-based language model; "distinct from Qwen3-VL, Pixtral and Idefics3 as a released VLM, though the LM backbone family overlaps Qwen") Third pre-declared candidate, never run: llava-hf/llava-.6-mistral-7b-hf @ 2424fdd47412fccc66d91719126b420e9fbd7065. Gemma-3 and Llama-3.2-Vision were excluded as license-gated. Model choice in was fixed before any second-model output ("the FIRST candidate that passes the GATE").

D.2 Inference settings (identical in every behavioral run) HuggingFace transformers 4.57.1 (logs/setup_env.log), attn_implementation = "sdpa", dtype bfloat16 (image tensors cast to bfloat16 for Pixtral, an engineering fix made before any output), device_map cuda, greedy decoding, max_new_tokens = 16 for answers (48 for model-written summaries), every arm including baselines through the same attention-hook code path (a no-op hook when nothing is blocked). Decoder layers are located generically (the single ModuleList named "...layers" under language_model/text_model whose members have self_attn); turn boundaries are found model-agnostically from prefix lengths of the rendered prompt (prefix property asserted per item x context). Scoring: the first number (digits or number word) or color word in the reply; categories reliable / unreliable / other / unparseable; labels correct by construction, no human annotation. Truncations are reported, never rerun; 0 errors and 0 truncated answers in every reported run.

D.3 Gates, thresholds and exclusions per experiment the confirmatory two-source experiment (Qwen, confirmatory): stage 0 leakage checks first; abort if L1 != 0.0 on any of >= 12 pairs (4 per family incl. one diagnostic context) or no-block swap not > 0. P4 controls: TN within [0.40, 0.60] overall and per direction; TE >= 0.95. Result: L1 (the evidence-swap difference with evidence and summary both blocked) = 0.0 in 15/15; TN 0.50 (0.49 / 0.51); TE 720/720. (Pixtral): gate on a separate 30-item set, only arms that cannot reveal the effect (never BE/BN): hook reaches every decoder layer in prefill and decode; L1 = 0.0 on 6 evidence-swap checks; no-block swap > 0; C0 image-only >= 0.80 and caption-only >= 0.95; N >= 0.75 in each direction; TE >= 0.90; TN within [0.35, 0.65]. Family rule: image-only capability < 0.70 on the gate set drops the family (>= 2 must remain). Result: passed; L1 6/6; image-only 26/30; caption-only 30/30; N 25/30 and 26/30; TE 60/60; TN 29/60; COUNT dropped (image-only 6/10). Main: L1 = 0.0 on 8/8. (factorial): Q1 capability C0i < 0.70 on the gate set drops the family; gates L1 = 0.0 exactly, TE >= 0.90, TN within [0.35, 0.65]. Qwen: reproducibility of the model-agnostic runner against the saved records (arm AN, 180 trials) 180/180; L1 12/12. Idefics3 gate (30 separate items): L1 6/6; image-only 28/30 (count 8/10, color 10/10, chart 10/10); caption-only 30/30; TE 60/60; TN 27/60; no family dropped; main L1 12/12. Pixtral follow-up: L1 8/8. : family retained iff C0i >= 0.70 on the 180 items; gates L1 =

0.0, TE >= 0.90, TN within [0.35, 0.65]. All gates passed for all three models: L1 12/12 each; TE 1.00 / 1.00 / 0.99; TN 0.48 / 0.51 / 0.51 (Qwen / Pixtral / Idefics3); Pixtral drops COUNT (36/60). : family retention C0i >= 0.70 (Pixtral drops COUNT, 23/40); cue VALIDITY per model decided from controls that cannot see the effect: TE >= 0.90 in each direction under that cue and TN within [0.35, 0.65]; at least 4 valid cues required. L1 = 0.0 on 12/12 per model. c4_modality INVALID for Qwen (text-only explicit reader 0.82 < 0.90); valid for Pixtral. the three-source experiment : L1 re-verified per model and source set (3 evidence-swap pairs per set); any L1 != 0.0 aborts that model x source set. Item-level: an item enters the confirmatory analysis only if ALL its C0 checks (one per source) are correct; every primary estimate is also reported without the exclusion. Exclusions (native / neutral): Qwen ITT 8 / 7, IIT 11 / 11; Pixtral ITT 8 / 10, IIT 20 / 21; Idefics3 ITT 3 / 7, IIT 8 / 15; none in TTT; every verdict unchanged without them. Condition-level label prerequisite failure (never a deletion): if accessible-neutral (AN) accuracy for a model x source set x scheme x condition is < 0.60, its blocked result is shown but not interpreted as a compression effect. Unparseable/other answers count as incorrect. V0 INVALID-MASK if L1 != 0, SHAM not EQUAL (|ACC(SH1) − ACC(SH0)| CI outside ± 0.10) or BE w1 QUERY < 0.80. DEGRADED arm engineering rule: blur level = the highest of {2, 4, 6} px at which every dev ITT-native image C0 check is still correct (none -> arm not run; outcome: not run, Qwen 8/9, 8/9, 5/9; Pixtral 8/9, 5/9, 3/9). EVICT validity: the manual decode loop without deletion must reproduce hook-path answers on dev items at >= 0.98 (71/72 Qwen, 72/72 Pixtral). Idefics3: C0, all AN cells and BN-NONE only; never part of a verdict. Exp 4 gate : PASS requires ALL of C0i >= 0.80 and C0c >= 0.90; TE >= 0.90 and TN within [0.35, 0.65]; AN >= 0.60 in BOTH reliability directions; AN QUERY (all targets) >= 0.70; unparseable <= 10%; FAIL on any -> prerequisite regime only. Engineering check: a one-item development run (‘–part dev –limit 1 –cells prereq‘) must complete with L1 = 0.0, else the fourth-model screen is abandoned (logged), not debugged.

E Evidence-Access Mask: Definition And Verification

E.1 Definition (sink-spared mask, the only mask used from .1 on) Queries Q = every position from the first token of the final user turn (the ‘<|im_start|>‘ of turn 3 in Qwen; found by prefix length in other models) onward, including every generated token. Blocked keys K = positions 1 . . . {} (first summary content token − 1): everything before the summary — the turn-1 evidence, its headers, its end-of-turn marker and the delimiter tokens up to the summary — except position 0. All decoder layers (36 in Qwen); prefill and every decode step. Implementation: forward pre-hooks on every decoder layer replace the attention mask with an explicit boolean mask built from the cache position; the K/V entries of blocked positions remain in the cache and are masked at read time. Accessible conditions run through the same hook with an empty key set (BlockSpec name “noop”), so baseline and intervention share the attention kernel path. Text-only readers (TN, TE) run under the same mask with no evidence in turn 1. Summary start s0 and question start q0 are asserted per item x context from prefix rendering (onward: “prefix property (text/tokens) failed” assertions). Three pilot findings fixed the definition (pilot findings): (i) blocking position 0, the attention sink, destroys instruction following — a text-only reader of the explicit summary fell from 96/96 to 58/96, so the pilot’s original bottleneck arms were invalid and position 0 is spared; (ii) a single delimiter token after the end-of-turn marker relays blocked content in one hop unless the blocked key span includes the closing delimiters (invariance error 2.26 before the fix, 0.0 after); (iii) masks applied only in prefill or only in decoding leak (positive controls move logits by 5–30).

E.2 Verification: evidence-swap invariance, every run L1 (evidence-swap invariance): with evidence AND summary blocked (keys 1 . . . {} q0 − 1), swapping the image and caption (all sources in the three-source experiment) for another item’s with identical token layout, under identical forced continuations (the first 8 non-eos tokens the blocked context generates), the max absolute logit difference must be exactly 0.0. L2: the same swap with only the evidence blocked (> 0 shows that summary-span states carry evidence beyond the text). no_block: the same swap with nothing blocked (must be > 0). Values below are read from the leakage JSON of each run (max | logit| per pair; ranges over pairs, one decimal):

An earlier version quoted the values as 9.0–18.4 / 3.2–9.6 (L2) and 5.8–32.3 / 2.7–16.5 (no block); the differences in the last digit are rounding of the same numbers.

run (file)	n pairs	L1	L2 (evidence only blocked)	no block
the confirmatory two-source experiment Qwen	15	0.0 in all	5.7–22.1	4.5–30.8
Qwen follow-up (...{}v4_qwen_followup_...{})	12	0.0	5.7–22.1	4.5–30.8
Idefics3 gate / main (...{}v4_idefics3_...{})	6 / 12	0.0	2.1–5.2 / 2.6–6.3	7.6–12.1 / 7.2–13.6
Pixtral follow-up (...{}v4_pixtral_followup_...{})	8	0.0	3.2–7.6	6.2–13.8
Qwen (...{}v6_cueinv_...{})	12	0.0	9.0–18.4	5.8–32.2
Pixtral (...{}v6_cueinv_...{})	12	0.0	3.1–9.6	2.7–16.5
the three-source experiment Qwen three-source (...{}v7_...{}; 3 pairs per source set ITT/TTT/IIT)	9	0.0	4.8–17.2	10.1–22.9
the three-source experiment Pixtral three-source (...{})	9	0.0	2.4–5.9	3.5–7.0
the three-source experiment Idefics3 three-source (...{})	9	0.0	2.5–5.6	3.2–9.0
the three-source experiment Qwen IT2 anchor (...{})	12	0.0	11.6–18.3	17.9–29.6
the three-source experiment Pixtral IT2 anchor (...{})	12	0.0	3.4–11.0	4.1–11.7

E.3 Sham mask and cache eviction SHAM: a filler passage "Background (unrelated to the task): ...{}" of token length >= 0.95 x the item's blocked-evidence span is placed at the start of turn 1; SH0 = filler present, no mask; SH1 = the filler span (positions 1 through the last filler token, with its delimiters) blocked while the evidence stays accessible; neutral summary; all w1 final-turn conditions. ACC(SH1) − ACC(SH0) pooled over QUERY/TRUST/DISTRUST: Qwen +0.030 [+0.007, +0.052], Pixtral +0.013 [−0.013, +0.039] -> EQUAL by the pre-set rule (CI within ±0.10); answer agreement 0.92 / 0.87. EVICT: turns 1–2 are prefilled with NO mask; then the K/V entries of exactly positions 1 ...{} s0 − 1 are deleted from every layer's cache (position 0 kept, original position ids kept); the final turn is processed and decoded greedily. Validity check (same manual loop without deletion must reproduce hook-path AN answers on dev items at >= 0.98): 71/72 Qwen, 72/72 Pixtral. Agreement of EVICT with the mask (600 answers each): 0.987 (Qwen), 0.972 (Pixtral); accuracy 0.361 vs 0.365 and 0.350 vs 0.346. Implementation robustness only; never described as a natural bottleneck. V0 INVALID-MASK was not triggered anywhere; the explicit summary is readable under the mask (BE QUERY = 1.00 in every cell of both models).

F Three-Source Experiment : Design And Full Results

F.1 Design summary Stimuli, contexts and final-turn conditions as in B.3 and C.3. Contexts: AN, AE (accessible), BN, BE (blocked), TN, TE (text-only readers; chance 1/3). Cells per three-source item : C0 x 3 sources; for each of the six contexts NONE + QUERY/TRUST/DISTRUST x 3 targets (w1); w2 TRUST/DISTRUST x 3 targets in AN and BN; COMBINED x 3 targets in AN and BN for ITT-native; on SUBSET-S additionally SHAM (SH0, SH1) x all w1 conditions and EVICT (BN, all w1 conditions); DEGRADED not run (rule above). Models: Qwen and Pixtral full design; Idefics3 C0, all AN cells and BN-NONE only (a prerequisite regime, never part of a verdict). Runs: Qwen 42,840 answers, Pixtral 42,840, Idefics3 7,560; 0 errors each. Estimands (section 6): d_i = item i's BN-NONE answer (its default). A target is NON-DEFAULT if its value != d_i; all accuracies below are over non-default targets (default-target accuracies tabulated separately). K1 = default vs text reader (ITT: image share BN − TN; IIT: summed image share; TTT: total-variation distance between BN and TN source-share distributions plus share by presentation slot). ACCq/ACCt/ACCd = accuracy on non-default targets for QUERY/TRUST/DISTRUST; SHORTCUT CEILING = 0.50 (a policy that knows only the default and picks among the other two values). K2 = ACCq(BN) vs 0.50 and vs ACCq(TN). K3t = [ACCq − ACCt]_BN − [ACCq − ACCt]_AN; K3d likewise with DISTRUST; K3x = [ACCt − ACCd]_BN − [ACCt − ACCd]_AN. K5 = SHAM and EVICT contrasts. Item-cluster bootstrap 95% CI, 10,000 resamples, seed 0. Conventions: ABOVE = point estimate >= 0.15 and CI excludes 0 (for K2: ACCq >= 0.65 and CI lower bound > 0.50); EQUAL = CI within ±0.10; otherwise INDETERMINATE.

F.2 Verdict rules V0 INVALID-MASK L1 != 0, or SHAM not EQUAL (|ACC(SH1) − ACC(SH0)| CI outside ±0.10), or BE w1 QUERY accuracy < 0.80 (explicit summary unreadable under the

cell	Qwen BN [TN] (AN)	Pixtral BN [TN] (AN)
ITT native	0.09 [0.32] (0.98)	0.20 [0.29] (0.84)
ITT neutral labels	0.16 [0.31] (0.98)	0.27 [0.29] (0.91)
TTT native	0.16 [0.19] (1.00)	0.28 [0.26] (1.00)
TTT neutral labels	0.25 [0.29] (1.00)	0.33 [0.32] (0.99)
IIT native	0.48 [0.29] (0.92)	0.29 [0.34] (0.94)
IIT neutral labels	0.24 [0.30] (0.94)	0.30 [0.29] (0.83)

mask). -> F-b, pivot criterion P-1 is examined. V1 PREMISE-FAIL K1 not ABOVE and ACCq(BN) EQUAL to ACCq(TN): nothing source-linked is carried. -> F-a. V2 retrieval below the ceiling ACCq(BN) not above the shortcut ceiling. Sub-label DEFAULT-ONLY if K1 is ABOVE. -> the source-rebinding hypothesis NARROWS (DEFAULT-ONLY) or FAILS (with V1) for that cell; it is not re-read through another statistic. V3 DISSOCIATION-TRUST ACCq(BN) above the ceiling AND [ACCq − ACCt]_BN ABOVE AND K3t ABOVE. V4 DISSOCIATION-DISTRUST same with ACCd and K3d. (V3 and V4 can both hold. V4 without V3 = the failure is distrust/exclusion, not trust as such; the framing narrows to that.) V5 NO-DISSOCIATION ACCq(BN) above the ceiling AND K3t and K3d EQUAL: late trust updates work after compression once retrieval works. V6 INDETERMINATE anything else. prerequisite failure (section 5) overrides V3–V5 for the affected condition. w2 robustness: a V3/V4 verdict is "wording-robust" only if the corresponding K3 has the same sign with CI excluding 0 under w2; otherwise it is reported as w1-specific. Cross-cell readings: GENERIC late source-reliability binding = V3 or V4 in ITT and in at least one same-modality-containing comparison where retrieval succeeds (TTT; IIT for image-vs-image targets). VISUAL-SPECIFICITY SUPPORTED = V3 or V4 in image-containing sets while TTT is V5 (if TTT is V2 or V1 the comparison is "not informative"). LABEL-DEPENDENT = a verdict V3–V5 under NATIVE labels but V2 under NEUTRAL labels in the same source set; the NEUTRAL scheme is the stricter test and bounds the claim. Two models: a reading "strengthens" only if Qwen and Pixtral receive the same verdict in ITT. Anchor: if the IT2 P1 contrast or interaction fails to replicate in Qwen on fresh items, pivot criterion P-1 is examined first. Amendment the three-source experiment.1: V1 is also evaluated pooled over label schemes within each source set (n = 180) and over all cells (n = 540). Guardrails the three-source experiment.3: (a) where NATIVE and NEUTRAL disagree, NEUTRAL arms are the basis for any modality/source-specificity claim; (b) QUERY is a hard gate — if blocked-neutral QUERY does not exceed the ceiling, TRUST/DISTRUST failures are not reinterpreted as source-specific rebinding through any other statistic; (c) EVICT is secondary and is dropped, not debugged, if its validity check is not clean.

F.3 Results Validity controls: see E.2–E.3. Two-source replication anchor (fresh IT2 items, original design and cue): Qwen AN 0.90 / AE 1.00 / BN 0.22 / BE 0.92 (image unreliable); BN caption-unreliable 1.00; I = +0.60 [+0.47, +0.72]; discordant 70 vs 0 (p = 2e-21) -> REPLICATES. Pixtral 0.90 / 0.99 / 0.50 / 0.94; BN caption-unreliable 0.78; I = +0.36 [+0.24, +0.47]; discordant 37 vs 12 (p = 5e-4) -> REPLICATES by the coded rule (the anchor script applies no capability exclusion, so COUNT items are included for Pixtral here).

Primary result — named-source retrieval after compression (the hard gate). QUERY accuracy for non-default targets, blocked + neutral summary [text-only reader] (evidence accessible):

In NO cell of either model does blocked-neutral QUERY exceed the shortcut ceiling (highest: Qwen IIT native 0.48 [0.40, 0.56]). Verdict in every cell, both models, both label schemes, with and without exclusions: V2 retrieval below the ceiling. Sub-labels: DEFAULT-ONLY (K1 above reference) in Qwen ITT native, ITT neutral, TTT neutral and Pixtral ITT neutral, TTT native; V1 PREMISE-FAIL also in Qwen TTT native and Pixtral TTT neutral, IIT neutral. Pooled premise test (the three-source experiment.1): Qwen not premise-fail in any set (pooled QUERY BN − TN: ITT −0.18 [−0.24, −0.13], i.e. BELOW the text reader; TTT −0.03; IIT +0.07 [+0.01, +0.13]); Pixtral pooled premise-fail in TTT and IIT, not in ITT. When the queried source IS the default, QUERY is high (Qwen ITT 0.98 / 0.80; Pixtral 0.79 / 0.55): under the bottleneck the models return the default value whichever source is named. Per family (Qwen ITT native QUERY BN): count 0.30, color 0.00, chart 0.03. No V3, V4 or V5 anywhere; GENERIC = false, VISUAL-SPECIFICITY SUPPORTED = false (TTT "not informative"), LABEL-DEPENDENT = false in all sets.

The no-instruction default (K1). ITT: blocked-neutral answers sit on the image's value — Qwen 0.87 (native) / 0.95 (neutral labels), K1 = +0.52 [+0.40, +0.63] / +0.60 [+0.49, +0.71]; by presentation slot of the image 0.89 / 0.75 / 0.96 (native) and 1.00 / 0.90 / 0.96 (neutral): the default follows the image, not its position. Pixtral: 0.39 (native; K1 +0.07, not above; by image slot 0.57 / 0.31 / 0.28)

criterion	result
P4 controls	L1 = 0.0 in 15/15 checks; TN 359/720 = 0.50 (0.49 / 0.51 by direction); TE 720/720
P3 no bottleneck (N = AN)	caption unreliable 341/360 = 0.95; image unreliable 337/360 = 0.94
P1 BN	caption unreliable 360/360 = 1.00; image unreliable 82/360 = 0.23; items correct only when the caption is unreliable 278 vs only when the image is unreliable 0 (both 82, neither 0); exact McNemar p = 4e-84. By family (image-unreliable): count 0.49, color 0.05, chart 0.14 — all three pass
P2 BE	image unreliable 337/360 = 0.94 (caption unreliable 360/360); BE − BN on image-unreliable trials = +0.71, p = 3e-77

and 0.71 (neutral labels; K1 +0.35 [+0.23, +0.48]; 0.73 / 0.68 / 0.72). TTT: no source is favored by name; the first-presented slot is over-represented (Qwen 0.46 / 0.42, Pixtral 0.46 / 0.42 vs ≈ 0.33 for the text reader), and "other" answers are frequent (Qwen 0.13 / 0.34). IIT: K1 ≈ 0 in both models and both schemes (-0.01, +0.04; +0.01, +0.01): with two conflicting images the image values are NOT favored over the caption's beyond the text reader (Qwen native 0.34 / 0.32 / 0.34 for image 1 / image 2 / caption).

TRUST and DISTRUST (observations only; by the gate they are NOT interpreted as source-specific rebinding). Blocked-neutral TRUST / DISTRUST (non-default targets): Qwen ITT 0.10 / 0.04 (native), 0.14 / 0.05 (neutral); TTT 0.17 / 0.06, 0.26 / 0.08; IIT 0.41 / 0.28, 0.22 / 0.13. Pixtral ITT 0.18 / 0.20, 0.30 / 0.09; TTT 0.24 / 0.09, 0.27 / 0.20; IIT 0.22 / 0.21, 0.26 / 0.21. With the evidence ACCESSIBLE the separated cues are already weak whenever an image is among the sources: Qwen ITT native TRUST 0.59, DISTRUST 0.18 (-> prerequisite failure: TRUST, DISTRUST); ITT neutral 0.78 / 0.55; IIT native 0.72 / 0.44; TTT at ceiling (0.99–1.00). Pixtral: ITT native 0.52 / 0.32, IIT native 0.45 / 0.46 (prerequisite failure), TTT 0.92–0.99. Bridge cell with the original COMBINED cue (ITT native): Qwen AN 0.89 / BN 0.14, Pixtral AN 0.76 / BN 0.33. K3t / K3d are negative or null in image-containing sets and small positive for DISTRUST in TTT (Qwen +0.10 / +0.17; Pixtral +0.15 / +0.05). With explicit summaries under the mask (BE), TRUST / DISTRUST are 0.62 / 0.27 for Qwen ITT native but 0.93 / 0.65 with neutral labels and ≈ 1.00 in TTT; Pixtral 0.96 / 0.93 and higher.

Idefics3 (prerequisite cells only). Evidence accessible, QUERY / TRUST / DISTRUST: ITT native 0.51 / 0.29 / 0.24; ITT neutral 0.69 / 0.52 / 0.26; IIT native 0.58 / 0.33 / 0.17; IIT neutral 0.49 / 0.39 / 0.31; TTT 1.00 / 0.99 / 0.99 in both schemes. Blocked-neutral default on the image 0.51 / 0.66 (ITT).

Deviations and notes: amendments the three-source experiment.1–the three-source experiment.4 all committed before stimuli or outputs; DEGRADED arm not run for either model (rule above) — the certainty-tag vs image-prior question is untouched by the three-source experiment; stimuli generated with Pillow >= 10.1, exact analytic counterbalancing matrix, 15/198 count images use 6 distractors; the text-only reader's DISTRUST accuracy is below 1/3 (0.11–0.20), not examined further; K1 for TTT is a total-variation distance whose CI lower bound exceeds 0 almost by construction (noted in the script before data; permutation p-values in).

G Two-Source Cell Tables With Intervals And The Post Hoc Answer-Direction Table

Accuracy = the answer equals the RELIABLE source's value. Cells AN / AE / BN / BE = accessible or blocked evidence x neutral or explicit summary; NS = no summary; TN / TE = text-only readers under the mask. I = [BE − BN] − [AE − AN], item bootstrap 95% CI (10,000 resamples, seed 0), exact sign test.

G.1 Confirmatory two-source experiment

ALL HOLD -> REPLICATED. BN answers equal the image's value in 638/720 trials (first-listed value 342/720 = no order effect); text-only readers pick the first-listed value 591/720 and the image value 355/720. Restricted to items with correct image-only perception (n = 342): 342/342 vs 66/342. Pre-registered secondary diagnostic (120 items; audit record agrees with the caption): B0 image value 115/120 · D0 image value 104/120 (caption 16) · D1 caption value 26/120 · D2 image value

arm	caption declared unreliable	image declared unreliable
no bottleneck (N)	146/160 = 0.91	122/160 = 0.76
BE	160/160 = 1.00	140/160 = 0.88
BN	140/160 = 0.88	55/160 = 0.34
TN	0.53	0.49

model	AN	AE	BN	BE	NS	I [95% CI]	label
Qwen3-VL-8B (360; reused cells + AE, NS)	337/360 = 0.94	360/360 = 1.00	82/360 = 0.23	337/360 = 0.94	0.93	+0.644 [+0.589, +0.697], sign test p = 4e-58	INTERACTION, symmetric baseline
Pixtral-12B (160; cells + AE, NS)	0.76	0.91	0.34	0.88	0.71	+0.381 [+0.294, +0.469], p = 4e-14	INTERACTION amplifying a baseline asymmetry
Idefics3-8B (240 fresh)	42/240 = 0.17	170/240 = 0.71	68/240 = 0.28	164/240 = 0.68	0.35	−0.133 [−0.208, −0.062], p = 1e-3	NO EFFECT (by the pre-set labels)

117/120 · D1n 97/120, D2n 105/120 (13 other). Joint outcome: BOTH correct 23 | D2-only 94 | D1-only 3 | neither 0 => 94/120 = 0.78 >= 0.75 in the H-IMG cell; D1n = 0.81 and D2n = 0.875 meet the >= 0.80 precondition. PRE-REGISTERED CALL: H-IMG (image preference / perceptual-grounding tag). H-SRC holds for a minority: 23/120 items, 22 of them COUNT (count: both 22, D2-only 15, D1-only 3; color 0/40; chart 1/40).

G.2 Pixtral-12B original test

TE 320/320; image-only and caption-only 160/160 each.

BN = image value in 245/320. By family, BN image-unreliable: chart 0.20 (caption-unreliable 0.80), color 0.49 (0.95). Criteria: P1 caption-unreliable 0.875 MISSES 0.90; image-unreliable 0.344 meets; discordant 94 vs 9, p = 5e-19; color passes (0.95 / 0.49), chart fails the 0.90 bar (0.80 / 0.20). P2 PASSES (0.875 vs 0.344, +0.53, p = 1e-22). **P3 FAILS** (0.91 / 0.76; image-unreliable direction below 0.80). P4 PASSES. CODED VERDICT: "MEASUREMENT PROBLEM (P3/P4) — no claim"; not "replicated" and not "qualitatively replicated" (P3 is a precondition of the graded labels). The conditional causal test was therefore not run.

G.3 Factorial experiment

By family (image-unreliable I): Qwen chart +0.80 [0.72, 0.87], color +0.95 [0.91, 0.98], count +0.18 [0.08, 0.28]; Pixtral chart +0.49 [0.35, 0.62], color +0.28 [0.17, 0.39]; Idefics3 chart −0.39 [−0.53, −0.26], color +0.04 [−0.07, +0.15], count −0.05 [−0.16, +0.05]. Caption declared unreliable (AN / AE / BN / BE / NS; I): Qwen 0.95 / 1.00 / 1.00 / 1.00 / 0.88; I = −0.053 [−0.078, −0.031] (driven by counts: AN 0.84). Pixtral 0.91 / 1.00 / 0.88 / 1.00 / 0.98; I = +0.037 [−0.019, +0.094]. Idefics3 0.87 / 0.96 / 0.76 / 1.00 / 0.84; I = +0.146 [+0.087, +0.204] (chart +0.34).

G.4 Emphatic cue

Caption declared unreliable: Qwen 0.93 / 1.00 / 0.99 / 1.00; Pixtral 1.00 / 1.00 / 0.94 / 1.00; Idefics3 0.84 / 0.91 / 0.78 / 0.97. By family (image-unreliable I): Qwen chart +0.55, color +0.68, count −0.17 [−0.32, 0.00]; Pixtral chart +0.52, color +0.25; Idefics3 chart −0.28, color 0.00, count −0.30. Cost of this wording: Qwen BE falls to 0.68 (0.94), driven by counts (0.38) and colors (0.68); charts 0.98.

G.5 Six cue wordings

Cue as a random factor (valid cues): Qwen $\bar{I}$ = +0.585, item-bootstrap CI [+0.505, +0.662], items x cues CI [+0.400, +0.753], 5/5 valid cues positive, between-cue SD 0.18 -> CUE-ROBUST. Pixtral $\bar{I}$ = +0.356, [+0.302, +0.412], items x cues [+0.156, +0.583], 5/6 valid cues positive, none significantly reversed, between-cue SD 0.26 -> CUE-ROBUST (COUNT dropped, 23/40). Caption-unreliable direction: per-cue interactions between −0.06 and +0.13 in both models. NS swings with wording (Qwen 0.15–0.97; Pixtral 0.15–0.74). Qwen COUNT is inconsistent across cues (−0.38 . . . { } +0.60).

model	AN	AE	BN	BE	NS	I [95% CI]	label rule	pre-registered reading
Qwen3-VL-8B (180)	0.80	1.00	0.13	0.68	0.93	+0.356 [+0.256, +0.450]	interaction, symmetric baseline	R1-type pattern
Pixtral-12B (120; color, chart)	0.65	1.00	0.26	0.99	0.46	+0.383 [+0.292, +0.475]	interaction amplifying a baseline asymmetry	INTERMEDIATE (AN 0.65 < 0.75)
Idefics3-8B (180)	0.18	0.31	0.19	0.13	0.13	−0.194 [−0.272, −0.117]	no effect by the pre-set labels	R3 PREREQUISITE FAILS

cue	Qwen cells	Qwen I_c [95% CI]	Pixtral cells (color, chart)	Pixtral I_c [95% CI]
c1 original (the confirmatory two-source experiment/)	0.92 / 1.00 / 0.25 / 0.93	+0.592 [+0.483, +0.692]	0.82 / 0.95 / 0.44 / 0.90	+0.338 [+0.225, +0.450]
c2 concise	1.00 / 1.00 / 0.12 / 0.97	+0.850 [+0.783, +0.908]	0.96 / 1.00 / 0.11 / 1.00	+0.850 [+0.762, +0.925]
c3 disregard/trustworthy	0.67 / 0.68 / 0.12 / 0.70	+0.567 [+0.442, +0.683]	0.64 / 1.00 / 0.38 / 1.00	+0.263 [+0.138, +0.375]
c4 modality paraphrase	0.03 / 0.01 / 0.14 / 0.00	−0.125 [−0.192, −0.058] (INVALID for Qwen: TE 0.82 < 0.90)	0.30 / 0.94 / 0.36 / 0.97	−0.025 [−0.150, +0.100] (valid; null)
c5 verbose narrative	0.81 / 1.00 / 0.14 / 0.96	+0.625 [+0.525, +0.725]	0.53 / 1.00 / 0.17 / 1.00	+0.350 [+0.237, +0.475]
c6 emphatic	0.79 / 1.00 / 0.13 / 0.63	+0.292 [+0.167, +0.417]	0.62 / 1.00 / 0.26 / 1.00	+0.362 [+0.263, +0.475]

G.7 POST HOC answer-direction table. DA = the contrast of the two reliability directions under BN minus the same contrast for the text-only reader.

H IDEFICS3-8B-LLAMA3: PREREQUISITE FAILURE

Selection: chosen before any output (ungated; Llama-3.1 LM + SigLIP, distinct from Qwen and Pixtral); gate passed (Appendix D.3). (240 fresh items; original cue c1), image declared unreliable: AN 0.17 / AE 0.71 / BN 0.28 / BE 0.68 / NS 0.35; I = −0.133 [−0.208, −0.062]; label "NO EFFECT (by the pre-set labels)" — the interaction is not >= +0.20 and the scaffold label required the interaction CI to include 0, which it does not (it is NEGATIVE). Descriptively: a very large provenance-label effect that does NOT depend on evidence access (explicit − neutral = +0.53 accessible, +0.40 blocked) on top of a severe baseline failure to discount the image even with full access (0.17; 0.35 with no summary). Caption unreliable: 0.87 / 0.96 / 0.76 / 1.00 / 0.84; I = +0.146 [+0.087, +0.204]. (180 fresh items; emphatic cue): 0.18 / 0.31 / 0.19 / 0.13 / NS 0.13; I = −0.194 [−0.272, −0.117]; of the three readings fixed in advance (R1 hidden interaction, R2 binding survives, R3 prerequisite fails: accessible-neutral remains < 0.50) the data select R3 PREREQUISITE FAILS. A text-only reader of the same explicit summary follows the same instruction on 0.98, so instruction following is intact when provenance is literally in the text; the multimodal binding operation is not performed. The stronger cue made its explicit-summary cells WORSE than in (0.71 -> 0.31 accessible, 0.68 -> 0.13 blocked). : not run ("counts neither for nor against the interaction"). the three-source experiment (prerequisite cells only; 7,560 answers): evidence accessible QUERY / TRUST / DISTRUST — ITT native 0.51 / 0.29 / 0.24; ITT neutral 0.69 / 0.52 / 0.26; IIT native 0.58 / 0.33 / 0.17; IIT neutral 0.49 / 0.39 / 0.31; TTT 1.00 / 0.99 / 0.99 in both schemes. Blocked-neutral default on the image 0.51 / 0.66 (ITT). C0 exclusions ITT 3 / 7, IIT 8 / 15. Treatment: a prerequisite regime, never part of a verdict; it neither supports nor contradicts the pattern in the other two models. Possible readings (not adjudicated): a generic failure to act on late instructions when image-derived evidence is in play, a strong preference for visual input, format sensitivity. The fourth-model gate of Appendix I.3 (AN >= 0.60 in both directions) is the same prerequisite made explicit for a fourth model.

I CONTROLS ON THE SOURCE-REFERENT RESULT: DESIGNS AND RESULTS

For each experiment the wording, items, n, exclusions, primary statistic and thresholds are fixed before it is run; its analysis script is committed before its run finishes; no partial outcome is inspected. Engineering decisions are logged (e1–e5 below).

run	n	BN acc (caption unrel / image unrel)	TN acc (caption / image)	DA [95% CI]	P(image-backed answer) BN	same for TN
Qwen the confirmatory two-source experiment confirmatory (360; 3 families)	360	1.00 / 0.23	0.49 / 0.51	+0.786 [+0.689, +0.883]	0.89	0.49
Pixtral (160; color, chart)	160	0.88 / 0.34	0.53 / 0.49	+0.494 [+0.338, +0.650]	0.77	0.52
Idefics3 (240; 3 families)	240	0.76 / 0.28	0.49 / 0.53	+0.517 [+0.408, +0.625]	0.74	0.48
qwen c1_original	120	1.00 / 0.25	0.42 / 0.51	+0.833 [+0.667, +1.008]	0.88	0.46
qwen c2_concise	120	0.99 / 0.12	0.47 / 0.49	+0.892 [+0.708, +1.083]	0.93	0.49
qwen c3_disregard	120	1.00 / 0.12	0.48 / 0.50	+0.900 [+0.708, +1.092]	0.94	0.49
qwen c4_modality	120	1.00 / 0.14	0.51 / 0.52	+0.867 [+0.692, +1.042]	0.93	0.50
qwen c5_verbose	120	1.00 / 0.14	0.47 / 0.53	+0.917 [+0.725, +1.108]	0.93	0.47
qwen c6_strong_v5	120	1.00 / 0.13	0.45 / 0.53	+0.942 [+0.767, +1.117]	0.93	0.46
pixtral c1_original	80	0.88 / 0.44	0.53 / 0.49	+0.400 [+0.175, +0.625]	0.72	0.52
pixtral c2_concise	80	0.99 / 0.11	0.45 / 0.47	+0.900 [+0.688, +1.113]	0.94	0.49
pixtral c3_disregard	80	0.84 / 0.38	0.50 / 0.50	+0.463 [+0.312, +0.625]	0.73	0.49
pixtral c4_modality	80	0.88 / 0.36	0.46 / 0.47	+0.525 [+0.362, +0.675]	0.76	0.49
pixtral c5_verbose	80	0.89 / 0.17	0.50 / 0.45	+0.662 [+0.475, +0.850]	0.86	0.53
pixtral c6_strong_v5	80	0.89 / 0.26	0.49 / 0.51	+0.650 [+0.475, +0.825]	0.81	0.49

I.1 Lexical-anchor neutral control (Experiment 1) Question. Does blocked-neutral failure reflect the absence of readable antecedents for the source names? Stimuli. the three-source experiment ITT-native MAIN items (90; existing, not fresh — chosen because QUERY and the original combined cue are already implemented for them; stated as such). Sources in presentation order; three distinct values. Turn-2 summaries: N ordinary neutral "Reported {noun}: a, b and c." · L lexical anchor (new wording) "Sources consulted: {label1}, {label2} and {label3}. Reported {noun}: a, b and c." (labels in presentation order with native names; values in the the three-source experiment counterbalanced summary order; no value bound to any source) · E explicit "The image shows a. The caption says b. The field note says c." (presentation order). Final turn. QUERY-X and COMBINED-X (the original combined reliability cue), X over all three sources; NONE for the default. Contexts: B (evidence blocked, the three-source experiment mask) primary; A (accessible) control; TL text-only reader of summary L (QUERY, COMBINED) as a validity reference. Cells per item: {N,L,E} x {B,A} x (NONE + 3 QUERY + 3 COMBINED) + TL x (3 QUERY + 3 COMBINED) = 48 -> 4,320 answers per model. Models: Qwen (primary), Pixtral (run second). Exclusions. Items whose the three-source experiment C0 checks all passed for that model (Qwen 82, Pixtral 82); estimates also reported without exclusion. Default per item = the N-B-NONE answer in THIS run; non-default targets = sources whose value != default. Primary statistics (non-default targets; item-cluster bootstrap 10,000, seed 0): R_q = ACC_QUERY(L-B) − ACC_QUERY(N-B); R_c = ACC_COMBINED(L-B) − ACC_COMBINED(N-B). Reference: ACC(E-B) − ACC(N-B). Validity gates: TL QUERY <= 0.50 (expected ≈ 1/3); ACC_QUERY(L-A) >= 0.80; L1 = 0.0. Pre-committed verdicts (per model). PREDICTION-A (antecedents do not explain the failure): neither R_q nor R_c is ABOVE (>= 0.15 with CI excluding 0). PREDICTION-B (referential continuity matters): both ABOVE. PREDICTION-C (mixed): exactly one ABOVE — reported as mixed, no mechanism invented. Secondary: fraction of the explicit-summary effect recovered, R/(E−N). Runtime estimate Qwen ≈ 10 min, Pixtral ≈ 15 min incl. load. Engineering log: e1 existing the three-source experiment ITT-native main items rather than fresh items (QUERY/COMBINED already implemented; 90 items; time). e2 ordinary-neutral and explicit cells are re-run inside the same job (same-run paired contrasts) rather than borrowed from records. Results.

Qwen3-VL-8B:

default image share N-B 0.87, L-B 0.94; fraction recovered {'QUERY': 0.87248322147651, 'COMBINED': 0.6642335766423357}; no-exclusion verdict: PREDICTION-B (referential conti-

cond	N-B	L-B	E-B	N-A	L-A	E-A	TL	R = L−N (B)	E−N (B)	verdict
QUERY	+0.091 [+0.055, +0.134]	+0.884 [+0.829, +0.933]	+1.000 [+1.000, +1.000]	+0.976 [+0.945, +1.000]	+0.976 [+0.951, +0.994]	+1.000 [+1.000, +1.000]	+0.329 [+0.264, +0.394]	+0.793 [+0.720, +0.860]	+0.909 [+0.866, +0.945]	PREDICTION-B (referential continuity matters)
COMBINED	+0.140 [+0.085, +0.201]	+0.695 [+0.622, +0.768]	+0.976 [+0.951, +0.994]	+0.890 [+0.835, +0.939]	+0.902 [+0.848, +0.951]	+0.994 [+0.982, +1.000]	+0.309 [+0.240, +0.382]	+0.555 [+0.470, +0.640]	+0.835 [+0.768, +0.896]	PREDICTION-B (referential continuity matters)

cond	N-B	L-B	E-B	N-A	L-A	E-A	TL	R = L−N (B)	E−N (B)	verdict
QUERY	+0.195 [+0.140, +0.250]	+0.713 [+0.634, +0.787]	+1.000 [+1.000, +1.000]	+0.835 [+0.780, +0.890]	+0.878 [+0.829, +0.921]	+0.994 [+0.982, +1.000]	+0.321 [+0.260, +0.386]	+0.518 [+0.427, +0.610]	+0.805 [+0.750, +0.860]	PREDICTION-B (referential continuity matters)
COMBINED	+0.329 [+0.256, +0.402]	+0.652 [+0.573, +0.726]	+1.000 [+1.000, +1.000]	+0.762 [+0.701, +0.823]	+0.768 [+0.701, +0.829]	+1.000 [+1.000, +1.000]	+0.313 [+0.248, +0.382]	+0.323 [+0.226, +0.415]	+0.671 [+0.598, +0.744]	PREDICTION-B (referential continuity matters)

nuity matters) per family verdicts: {"count": "PREDICTION-B (referential continuity matters)", "colour": "PREDICTION-B (referential continuity matters)", "chart": "PREDICTION-B (referential continuity matters)"}

Pixtral-12B:

default image share N-B 0.39, L-B 0.59; fraction recovered {'QUERY': 0.6439393939393939, 'COMBINED': 0.48181818181818187}; no-exclusion verdict: PREDICTION-B (referential continuity matters) per family verdicts: {"count": "GATE-FAIL: PREDICTION-B (referential continuity matters)", "colour": "PREDICTION-B (referential continuity matters)", "chart": "PREDICTION-C (mixed)"}

I.2 Visual modality vs perceptual certainty under a principled sensor corruption (Experiment 2; design fixed before stimulus generation, code completion or any run) Why this design. Veridicality is not observable to the model: an image that "truly" shows C and a caption that says C* is, from the model's side, identical to an image corrupted to show C while the caption reports the latent C*. What the model CAN register is how certain its percept is. The unresolved question is therefore modality (H1: the default follows the visual percept whatever its certainty) vs perceptual certainty (H3: the default weakens as the visual percept becomes uncertain). A corrupted visual source, with the text carrying the latent truth, is implemented with a known generative sensor model rather than a verbal declaration. Generative setup (COLOR family only). RGB_shown = s · RGB(C') + (1 − s) · RGB(C*), C' = the palette color the the three-source experiment image shows, C* = the caption's value; s = 1.0 (clean, existing image), 0.75 and 0.55 -> 120 new images on the 60 the three-source experiment ITT COLOR main items (30 native + 30 neutral labels). Only the target shape's RGB changes (Appendix C.5). Cells per item x level: image-only C0 with first-token logit margin; AN-NONE; BN-NONE; BN-QUERY/TRUST/DISTRUST for the caption target (plus the three C0 checks for exclusion). ≈ 14 answers/item/level -> ≈ 1,700 per model. Models: Qwen, Pixtral. Inclusion (per item, per level): the image C0 answer at that level equals C' (if it flips to C* or another color the item no longer contains a conflict and is excluded at that level; counts reported; estimates also given without exclusion). Target statistic. Image share of BN-NONE answers by level (paired within item), first-token margin by level; secondary: Spearman correlation across items between the clean C0 margin and the BN-NONE image indicator. Predictions. H3 (certainty): margin falls with corruption and the BN-NONE image share falls with it; correlation positive. H1 (modality): margin falls but the image share is unchanged among included items; no correlation. Rule (per model). MARGIN-REDUCED if the paired-bootstrap CI of (margin_clean − margin_0.55) excludes 0. Then DEFAULT-TRACKS-CERTAINTY (compatible) if image share clean − 0.55 is ABOVE; DEFAULT-UNCHANGED (counts against H3) if EQUAL and MARGIN-REDUCED; UNINFORMATIVE if not MARGIN-REDUCED; else INDETERMINATE. The 0.75

run (inclusion)	n	margin clean / corrupted (median)	margin diff [CI]	BN-NONE image share clean / corrupted	share diff [CI]	Spearman(clean margin, default)	verdict
Qwen 55 (primary)	19	2.75 / 1.25	+1.07 [+0.38, +1.74]	1.00 / 0.95	+0.05 [+0.00, +0.16]	-0.49	INDETERMINATE
Qwen 55 (no_exclusion)	60	1.88 / 1.25	+0.56 [+0.09, +1.04]	1.00 / 0.75	+0.25 [+0.15, +0.37]	-0.04	DEFAULT-TRACKS-CERTAINTY (compatible with H3)
Qwen 75 (primary)	54	2.12 / 1.50	+0.50 [+0.21, +0.80]	1.00 / 0.98	+0.02 [+0.00, +0.06]	0.01	DEFAULT-UNCHANGED (counts against H3)
Qwen 75 (no_exclusion)	60	1.88 / 1.50	+0.53 [+0.23, +0.83]	1.00 / 0.98	+0.02 [+0.00, +0.05]	-0.04	DEFAULT-UNCHANGED (counts against H3)
Pixtral 55 (primary)	18	7.19 / 7.03	+1.47 [+0.49, +2.53]	0.83 / 0.83	+0.00 [+0.00, +0.00]	-0.12	DEFAULT-UNCHANGED (counts against H3)
Pixtral 55 (no_exclusion)	60	7.00 / 5.62	+1.91 [+1.13, +2.68]	0.83 / 0.67	+0.17 [+0.07, +0.27]	-0.01	DEFAULT-TRACKS-CERTAINTY (compatible with H3)
Pixtral 75 (primary)	50	7.00 / 7.25	+0.40 [-0.11, +0.95]	0.86 / 0.86	+0.00 [-0.06, +0.06]	-0.02	UNINFORMATIVE (margin not reduced)
Pixtral 75 (no_exclusion)	60	7.00 / 7.00	+0.85 [+0.33, +1.41]	0.83 / 0.82	+0.02 [-0.03, +0.08]	-0.01	DEFAULT-UNCHANGED (counts against H3)

level is reported as a dose check. Runtime $\approx$ 4 min per level per model plus load. Engineering log: e3 gains a '–cells degraded' mode (the three-source experiment itself is complete and archived; the added mode only selects cells). e4 corrupted images are generated with the unchanged renderer, files named * / * so the existing suffix rule applies. Results. Levels 55 = s 0.55 (decisive), 75 = s 0.75 (dose check).

I.3 Fourth model with a pre-committed prerequisite gate (Experiment 4, option B; design fixed before any output) Option A (natural-image validation on MMKC-Bench) was not attempted: adapting an external conflict dataset to the three-turn design with deterministic late queries and a comparable restricted-evidence implementation could not be done cleanly within the available time (decision e5). Model. OpenGVLab/InternVL3-8B-hf @ 259a3b64a14623c0ec91a045cb43f7c5af5fa6af (native transformers 4.57 InternVL implementation; InternViT vision tower + Qwen2.5-based language model; distinct from Qwen3-VL, Pixtral and Idefics3 as a released VLM, though the LM backbone family overlaps Qwen). Chosen because it is contemporary, ungated, natively supported and ~16 GB. Engineering check: a one-item development run ('–part dev –limit 1 –cells prereq') must complete with L1 = 0.0; if the model does not load or the mask/prefix machinery fails, the experiment is abandoned (logged), not debugged. Gate. Two-source anchor items plus named-source QUERY with evidence accessible. PASS requires ALL of: C0i >= 0.80 and C0c >= 0.90; TE >= 0.90 and TN within [0.35, 0.65]; AN >= 0.60 in BOTH reliability directions (the accessible late-instruction prerequisite that Idefics3 failed); AN QUERY (all targets) >= 0.70; unparseable <= 10%. FAIL on any -> reported as a prerequisite regime only (like Idefics3), nothing else run. If PASS, the same 1,440 + 420 answers ARE the core subset: report the four cells both directions, I with item-bootstrap CI, the P1 discordant contrast, BN-NONE image share vs 1/3, and AN/BN QUERY; label the model by its gate/result only. Runtime $\approx$ 15 min including load. Nothing from this model enters a mechanistic claim. Results. Two-source anchor (90 fresh items, original cue, both directions; 1,440 answers): C0 image 0.956, caption 1.000; text-only explicit 0.928, neutral 0.494; cells (image unreliable / caption unreliable): NS 0.078 / 0.844, AN 0.133 / 0.978, AE 0.522 / 0.989, BN 0.189 / 0.944, BE 0.300 / 0.956; interaction I = -0.278 [-0.422, -0.133]; P1 discordant 73 vs 5; BN image-backed share 0.878; unparseable 0.031. Three-source prerequisite cells (30 ITT-native items, 420 answers): image C0 0.733, text C0 1.000; with evidence accessible QUERY 0.889 (all targets), TRUST 0.844, DISTRUST 0.178; blocked-neutral no-instruction image share 0.533; unparseable 0.173. GATE: FAIL on 'AN >= 0.60 in both directions' (image-unreliable 0.133) and 'unparseable

cond	N_h-B	H-B	E_h-B	H_new-B	N_n-B	L_n-B	E_n-B	TH	TH_new	R_H	R_L	G_E	C_cont
QUERY	+0.162 [+0.113, +0.216]	+0.802 [+0.743, +0.857]	+1.000 [+1.000, +1.000]	+0.365 [+0.283, +0.449]	+0.156 [+0.108, +0.206]	+0.940 [+0.898, +0.976]	+1.000 [+1.000, +1.000]	+0.361 [+0.289, +0.434]	+0.349 [+0.277, +0.426]	+0.641 [+0.566, +0.710]	+0.784 [+0.717, +0.845]	+0.198 [+0.143, +0.257]	+0.437 [+0.353, +0.524]
COMBINED	+0.138 [+0.084, +0.192]	+0.611 [+0.533, +0.683]	+0.964 [+0.934, +0.988]	+0.305 [+0.232, +0.381]	+0.210 [+0.151, +0.269]	+0.725 [+0.657, +0.789]	+0.958 [+0.928, +0.982]	+0.373 [+0.301, +0.450]	+0.349 [+0.277, +0.426]	+0.473 [+0.399, +0.548]	+0.515 [+0.443, +0.584]	+0.353 [+0.281, +0.429]	+0.305 [+0.228, +0.385]

summary / cue	target	default	third	none	n
N_h/QUERY	0.17	0.63	0.20	0.00	181
N_h/COMBINED	0.15	0.79	0.06	0.00	181
H/QUERY	0.79	0.13	0.08	0.00	181
H/COMBINED	0.61	0.35	0.04	0.00	181
H_new/QUERY	0.34	0.35	0.31	0.00	181
H_new/COMBINED	0.29	0.49	0.22	0.00	181
E_h/QUERY	1.00	0.00	0.00	0.00	181
E_h/COMBINED	0.96	0.04	0.00	0.00	181
N_n/QUERY	0.18	0.66	0.17	0.00	181
N_n/COMBINED	0.21	0.66	0.13	0.00	181
L_n/QUERY	0.92	0.05	0.03	0.00	181
L_n/COMBINED	0.72	0.24	0.04	0.00	181
E_n/QUERY	1.00	0.00	0.00	0.00	181
E_n/COMBINED	0.95	0.05	0.00	0.00	181

<= 0.10' (0.173); PASS on C0i, C0c, TE, TN, AN QUERY. Verdict: prerequisite regime only; no further cells run. Engineering deviations e7 (the engineering check was repeated with a 3-item development manifest because a 1-item manifest gave the completeness check no pair) and e8 (the gate manifest was recreated identically and the prerequisite stage re-run) are logged in the design document

I.4 Arbitrary stable source handles, late-handle continuity and a held-out replication Items: the 90 the three-source experiment ITT main scenes with the neutral (Source A/B/C) label scheme, never used in the lexical-anchor control (I.1), plus native-label twins of the same scenes. Summaries: N_h ordinary neutral; H handle anchor "Sources consulted: Source B, Source A and Source C. Reported..."; E_h explicit with handles; H_new handles P/Q/R introduced only at turns 2-3; N_n, L_n, E_n the native twins. 67 cells per item; 6,030 answers per model.

Qwen3-VL-8B:

accessible QUERY: H-A +0.972 [+0.952, +0.988], L_n-A +0.988 [+0.972, +1.000]; D_QR +0.168 [+0.090, +0.244] VERDICTS: HANDLES-RESCUE + QUERY-MORE-THAN-RELIABILITY + BINDING-ADDS(QUERY,COMBINED) + CONTINUITY-MATTERS + replication: PREDICTION-B no-exclusion: HANDLES-RESCUE + QUERY-MORE-THAN-RELIABILITY + BINDING-ADDS(QUERY,COMBINED) + CONTINUITY-MATTERS + replication: PREDICTION-B per family: {"count": "HANDLES-RESCUE + BINDING-ADDS(QUERY,COMBINED) + CONTINUITY-MATTERS + replication: PREDICTION-B", "colour": "HANDLES-RESCUE + QUERY-MORE-THAN-RELIABILITY + BINDING-ADDS(COMBINED) + CONTINUITY-MATTERS + replication: PREDICTION-B", "chart": "HANDLES-RESCUE + QUERY-MORE-THAN-RELIABILITY + BINDING-ADDS(QUERY,COMBINED) + CONTINUITY-MATTERS + replication: PREDICTION-B"}

Candidate-level addendum (Qwen; blocked evidence; non-default targets; share of answers on the queried source's value, the default (image) value, the third value, or none):

With stable handles the queried source's value is raised (0.79) while the third value stays low (0.08), so the handle raises the intended source rather than merely suppressing the default; with handles introduced only after the bottleneck the three values are near-equiprobable (0.34 / 0.35 / 0.31). Pixtral-12B:

accessible QUERY: H-A +0.917 [+0.875, +0.954], L_n-A +0.921 [+0.879, +0.954]; D_QR -0.000 [-0.094, +0.094] VERDICTS: HANDLES-RESCUE + BINDING-ADDS(QUERY,COMBINED) + CONTINUITY-MATTERS + replication: PREDICTION-B no-exclusion: HANDLES-RESCUE + BINDING-ADDS(QUERY,COMBINED) + CONTINUITY-MATTERS + replication: PREDICTION-B per family: {"count": "HANDLES-RESCUE + BINDING-ADDS(QUERY,COMBINED) + CONTINUITY-MATTERS + replication: PREDICTION-B",

cond	N_h-B	H-B	E_h-B	H_new-B	N_n-B	L_n-B	E_n-B	TH	TH_new	R_H	R_L	G_E	C_cont
QUERY	+0.269 [+0.206, +0.338]	+0.731 [+0.669, +0.787]	+1.000 [+1.000, +1.000]	+0.294 [+0.219, +0.375]	+0.281 [+0.212, +0.350]	+0.769 [+0.700, +0.831]	+1.000 [+1.000, +1.000]	+0.338 [+0.267, +0.412]	+0.317 [+0.246, +0.388]	+0.462 [+0.381, +0.544]	+0.488 [+0.394, +0.581]	+0.269 [+0.213, +0.331]	+0.437 [+0.356, +0.519]
COMBINED	+0.250 [+0.188, +0.319]	+0.713 [+0.644, +0.775]	+1.000 [+1.000, +1.000]	+0.269 [+0.194, +0.344]	+0.331 [+0.256, +0.406]	+0.656 [+0.575, +0.731]	+1.000 [+1.000, +1.000]	+0.346 [+0.275, +0.421]	+0.317 [+0.246, +0.388]	+0.463 [+0.363, +0.556]	+0.325 [+0.225, +0.419]	+0.287 [+0.225, +0.356]	+0.444 [+0.363, +0.525]

summary / cue	target	default	third	none	n
N_h/QUERY	0.28	0.46	0.26	0.00	180
N_h/COMBINED	0.27	0.62	0.11	0.00	180
H/QUERY	0.73	0.21	0.07	0.00	180
H/COMBINED	0.72	0.23	0.04	0.00	180
H_new/QUERY	0.32	0.36	0.32	0.00	180
H_new/COMBINED	0.29	0.43	0.28	0.00	180
E_h/QUERY	1.00	0.00	0.00	0.00	180
E_h/COMBINED	1.00	0.00	0.00	0.00	180

“colour”: “HANDLES-RESCUE + BINDING-ADDS(QUERY,COMBINED) + CONTINUITY-MATTERS + replication: PREDICTION-B”, “chart”: “HANDLES-RESCUE + BINDING-ADDS(QUERY,COMBINED) + CONTINUITY-MATTERS + replication: PREDICTION-B”}

Pixtral candidate-level addendum:

Handle audit: in both tokenizers every handle is two tokens (’Source’, ’ X’) for A/B/C and P/Q/R alike, so stable and late handles are token-length matched; native names are two to three tokens (’the’, ’ image’; ’the’, ’ field’, ’ note’).

I.5 Counterfactual handle reassignment, capacity decomposition and cache-eviction replication Items: the 90 stable-handle scenes of I.4; three letter-to-source assignments per scene (cyclic shifts) with evidence, order, values and template fixed; USE-X = “Use Source X.” + question; RELIABILITY-X = the combined cue; eviction as in the three-source experiment. 33 cells per scene.

Qwen3-VL-8B: P1 handle tracking (blocked, non-default targets): v0 +0.801 [+0.741, +0.855] [text +0.361 [+0.289, +0.434]], +0.807 [+0.747, +0.861] [text +0.337 [+0.269, +0.410]], the confirmatory two-source experiment +0.795 [+0.741, +0.849] [text +0.333 [+0.261, +0.406]]; all +0.801 [+0.755, +0.847] P1 predicted switch (0.8330708661417323, 0.8020808205342808, 0.8600551455643578) (n pairs 635); stayed on old source (0.07559055118110236, 0.05748576771475312, 0.09879960881676894); target margin median 8.25, positive share 0.8433734939759037 P1 confusion (queried kind -> answered kind): {’image->image’: 232, ’image->text’: 17, ’text->image’: 42, ’text->text’: 456} P1 VERDICT: HANDLE-TRACKING

P2 QUERY +0.801 [+0.741, +0.855] USE +0.831 [+0.777, +0.880] RELIABILITY +0.608 [+0.530, +0.681]; D_QU -0.030 [-0.066, +0.006]; D_UR +0.223 [+0.157, +0.289]; VERDICT ADDRESSABILITY-INTACT/CONTROL-BOTTLENECK

P3 QUERY mask +0.801 [+0.741, +0.855] evict +0.807 [+0.753, +0.861] (diff -0.006 [-0.018, +0.000]); RELIABILITY mask +0.608 [+0.530, +0.681] evict +0.620 [+0.542, +0.693] (diff -0.012 [-0.030, +0.000]); answer agreement 0.9919678714859438; VERDICT PERSISTS

no-exclusion verdicts: {”P1_verdict”: “HANDLE-TRACKING”, “P2_verdict”: “ADDRESSABILITY-INTACT/CONTROL-BOTTLENECK”, “P3_verdict”: “PERSISTS”} per family: {”count”: {”P1_verdict”: “HANDLE-TRACKING”, “P2_verdict”: “MIXED”, “P3_verdict”: “PERSISTS”}, “colour”: {”P1_verdict”: “HANDLE-TRACKING”, “P2_verdict”: “ADDRESSABILITY-INTACT/CONTROL-BOTTLENECK”, “P3_verdict”: “PERSISTS”}, “chart”: {”P1_verdict”: “HANDLE-TRACKING”, “P2_verdict”: “MIXED”, “P3_verdict”: “PERSISTS”}}

Pixtral-12B: P1 handle tracking (blocked, non-default targets): v0 +0.736 [+0.676, +0.793] [text +0.338 [+0.267, +0.412]], +0.764 [+0.696, +0.829] [text +0.296 [+0.229, +0.362]], the confirmatory two-source experiment +0.741 [+0.674, +0.808] [text +0.308 [+0.246, +0.371]]; all +0.747 [+0.697, +0.796] P1 predicted switch (0.7425925925925926, 0.7040933970029265, 0.7776645266277776) (n pairs 540); stayed on old source (0.11481481481481481, 0.09060400044336862,

cond	N_h	H (presentation order)	H_scr (scrambled list)	H_ord (ordinal only)	T_scr	T_ord	D_scr = H − H_scr	R_H	R_scr
QUERY	+0.162 [+0.113, +0.216]	+0.802 [+0.743, +0.857]	+0.838 [+0.784, +0.888]	+0.461 [+0.389, +0.533]	+0.349 [+0.281, +0.422]	+0.349 [+0.277, +0.426]	-0.036 [-0.108, +0.036]	+0.641 [+0.566, +0.710]	+0.677 [+0.608, +0.743]
COMBINED	+0.138 [+0.084, +0.192]	+0.611 [+0.533, +0.683]	+0.623 [+0.548, +0.695]	+0.281 [+0.211, +0.351]	+0.337 [+0.269, +0.410]	+0.349 [+0.277, +0.426]	-0.012 [-0.084, +0.059]	+0.473 [+0.399, +0.548]	+0.485 [+0.413, +0.560]

cond	N_h	H (presentation order)	H_scr (scrambled list)	H_ord (ordinal only)	T_scr	T_ord	D_scr = H − H_scr	R_H	R_scr
QUERY	+0.269 [+0.206, +0.338]	+0.731 [+0.669, +0.787]	+0.762 [+0.700, +0.819]	+0.306 [+0.231, +0.388]	+0.358 [+0.287, +0.433]	+0.317 [+0.246, +0.388]	-0.031 [-0.106, +0.044]	+0.462 [+0.381, +0.544]	+0.494 [+0.419, +0.569]
COMBINED	+0.250 [+0.188, +0.319]	+0.713 [+0.644, +0.775]	+0.706 [+0.644, +0.769]	+0.269 [+0.194, +0.344]	+0.375 [+0.308, +0.446]	+0.317 [+0.246, +0.388]	+0.006 [-0.081, +0.094]	+0.463 [+0.363, +0.556]	+0.456 [+0.362, +0.550]

0.14446738799032235); target margin median 3.5625, positive share 0.7375 P1 confusion (queried kind -> answered kind): {'image->image': 173, 'image->text': 67, 'text->image': 77, 'text->text': 403} P1 VERDICT: HANDLE-TRACKING

P2 QUERY +0.736 [+0.676, +0.793] USE +0.759 [+0.698, +0.817] RELIABILITY +0.741 [+0.676, +0.803]; D_QU -0.023 [-0.064, +0.017]; D_UR +0.017 [-0.029, +0.064]; VERDICT ADDRESSABILITY-DOMINATES

P3 QUERY mask +0.736 [+0.676, +0.793] evict +0.736 [+0.676, +0.793] (diff +0.000 [+0.000, +0.000]); RELIABILITY mask +0.741 [+0.676, +0.803] evict +0.741 [+0.676, +0.803] (diff +0.000 [+0.000, +0.000]); answer agreement 0.99375; VERDICT PERSISTS

no-exclusion verdicts: {"P1_verdict": "HANDLE-TRACKING", "P2_verdict": "ADDRESSABILITY-DOMINATES", "P3_verdict": "PERSISTS"} per family: {"count": {"P1_verdict": "PARTIAL", "P2_verdict": "MIXED", "P3_verdict": "PERSISTS"}, "colour": {"P1_verdict": "HANDLE-TRACKING", "P2_verdict": "ADDRESSABILITY-DOMINATES", "P3_verdict": "PERSISTS"}, "chart": {"P1_verdict": "PARTIAL", "P2_verdict": "ADDRESSABILITY-DOMINATES", "P3_verdict": "PARTIAL"}}

I.6 List-position control Summaries: H handles in presentation order; H_scr the same handles in a scrambled list order; H_ord ordinal-only anchor; N_h baseline; text readers for H_scr and H_ord.

Qwen3-VL-8B:

VERDICTS: handles PARTIAL; ordinal ORDINAL-NOT no-exclusion: PARTIAL / ORDINAL-NOT per family: {"count": ["PARTIAL", "INDETERMINATE"], "colour": ["PARTIAL", "ORDINAL-NOT"], "chart": ["PARTIAL", "ORDINAL-NOT"]}

Pixtral-12B:

VERDICTS: handles PARTIAL; ordinal ORDINAL-NOT no-exclusion: PARTIAL / ORDINAL-NOT per family: {"count": ["PARTIAL", "ORDINAL-NOT"], "colour": ["PARTIAL", "ORDINAL-NOT"], "chart": ["PARTIAL", "ORDINAL-NOT"]}

I.7 Nonce handles Handles Source Zeb / Kip / Vox replace Source A / B / C at turn 1 (assignment inherited from the the three-source experiment matrix); tokenization 2–3 tokens per handle in both models (e13).

Qwen3-VL-8B:

VERDICT: NONCE-RESCUE; D_nonce EQUAL no-exclusion: NONCE-RESCUE; D_nonce EQUAL per family: {"count": "NONCE-RESCUE; D_nonce INDETERMINATE", "colour": "NONCE-RESCUE; D_nonce EQUAL", "chart": "NONCE-RESCUE; D_nonce INDETERMINATE"}

I.8 Compact nonce reliability-control decomposition Design: the 90 ITT-neutral items with the nonce-handle summary of I.7, evidence blocked, every target slot; final turns QUERY ("Accord-

cond	N_h	H_nonce	H (Exp 5, same scenes)	T_nonce	R_nonce	D_nonce = H − H_nonce
QUERY	+0.162 [+0.113, +0.216]	+0.784 [+0.726, +0.839]	+0.802 [+0.743, +0.857]	+0.349 [+0.277, +0.426]	+0.623 [+0.545, +0.695]	+0.018 [-0.024, +0.060]
COMBINED	+0.138 [+0.084, +0.192]	+0.413 [+0.331, +0.491]	+0.611 [+0.533, +0.683]	+0.349 [+0.277, +0.426]	+0.275 [+0.205, +0.347]	+0.198 [+0.120, +0.277]

cue	N_h blocked	H blocked	E_h blocked	H accessible	text reader of H	rescue H - N_h	binding gain E_h - H
QUERY	0.015 [0.000, 0.035]	0.865 [0.820, 0.905]	1.000	1.000	0.363 [0.307, 0.420]	+0.850 [+0.800, +0.895]	+0.135 [+0.095, +0.180]
COMBINED	0.015 [0.000, 0.040]	0.610 [0.550, 0.670]	0.930 [0.885, 0.965]	0.993	0.350 [0.300, 0.403]	+0.595 [+0.535, +0.655]	+0.320 [+0.265, +0.375]

ing to Source X,..."), USE ("Use Source X."), DIRECT ("Source X is reliable. Use Source X.") and INFERRED (the the three-source experiment combined cue: the other two sources declared unreliable, so X must be selected from the statements). Non-default targets defined by the item's blocked no-instruction answer; leakage check L1 = 0.0 on 2/2, no_block > 0. Pre-committed reading (item-bootstrap 10,000; BELOW = difference <= -0.15 with CI excluding 0): ACCESS-USE impairment if USE below QUERY; STATEMENT impairment if DIRECT below USE; LANGUAGE-MAPPING impairment if INFERRED below DIRECT with neither of the former; otherwise MIXED. Result (180 non-default targets per condition): QUERY 0.78, USE 0.79, DIRECT 0.78, INFERRED 0.43; USE - QUERY +0.006 [-0.028, +0.039] (EQUAL), DIRECT - USE -0.006 [-0.039, +0.028] (EQUAL), INFERRED - DIRECT -0.350 [-0.439, -0.261] (BELOW). PRE-COMMITTED LABEL: LANGUAGE-MAPPING impairment. With nonce handles, access, use and a direct reliability statement about the named source are equally available; the cost is specific to selecting the source from reliability statements about the other sources. The nonce-handle INFERRED value of I.7 (0.41) is re-measured here at 0.43 on the same items and wording. Secondary, pre-registered; no verdict of I.7 changes.

I.9 Natural-image validation Items: 100 VQAv2 validation photographs whose unanimous human answer is a color; the image is one source and two templated text sources report counterfactual colors (three distinct values; handles Source A/B/C assigned at turn 1, letters counterbalanced); conditions as in the stable-handle experiment (I.4): neutral summary N_h, stable-handle summary H, explicit binding E_h, evidence blocked (sink-spared mask) or accessible, text-only readers, QUERY and COMBINED cues; C0 single-source checks (image 1.00, text 1.00; all 100 items kept, verdict identical without exclusions). Gates: G1 evidence-swap L1 = 0.0 on 3/3 (no_block > 0); G2 C0 image >= 0.70; G3/G3b text readers <= 0.50 (0.36, 0.30); G4 handle summary with evidence accessible >= 0.80 (1.00); G5 unparseable rate <= 0.10 (0.000). The pre-committed analysis first reported G5 as failed because its implementation tested '(rate or 1) <= 0.10', which treats a rate of exactly 0.0 as missing; the expression was corrected in the analysis code only (thresholds, prompts, items and model outputs untouched), the analysis re-run on the saved generations, and the original output preserved.

Rescue difference QUERY minus COMBINED +0.255 [+0.190, +0.325]; blocked no-instruction answers on the image's value 1.00 (0.97 accessible); per-answer decomposition H/QUERY target 0.865, default 0.13, third 0.005; H/COMBINED 0.61 / 0.385 / 0.005. PRE-COMMITTED VERDICT: NATURAL-REPLICATES + HANDLES-RESCUE + QUERY-MORE-THAN-RELIABILITY + BINDING-ADDS. A small controlled validation on one attribute and one model; not evidence of broad natural-world generalization.

I.10 Handle necessity via selective attention ablation Design: the 90 stable-handle scenes of I.4 (ITT, neutral labels, summary "Sources consulted: Source A, Source B and Source C. Reported <noun>: a, b and c."), evidence blocked by the sink-spared mask; final turns NONE, QUERY, USE ("Use Source X.") and COMBINED for every target. Arms add positions to the blocked keys for all final-turn queries and all layers: HANDLE blocks the two turn-2 tokens of the queried handle (" Source", " X"); SHAM blocks the same number of adjacent non-handle tokens (" consulted", ":"); VALUE blocks the queried value's token in the "Reported..." clause (diagnostic). Blocked positions are recorded per answer. Gates fixed in advance: normal accuracy >= 0.70 (prerequisite), unparseable <= 0.10, sham drop <= 0.15 (else NONSPECIFIC). handle necessary if the handle drop

cue	NORMAL (target/default/third/unparseable)	HANDLE ablated	SHAM	VALUE ablated	drop NORMAL-HANDLE	drop NORMAL-SHAM	specificity	label
QUERY	0.79 [0.73, 0.85] (0.79/0.13/0.07/0.00)	0.22 [0.16, 0.28] (0.22/0.48/0.09/0.21)	0.78 [0.72, 0.84] (0.78/0.13/0.09/0.00)	0.06 [0.03, 0.09] (0.06/0.46/0.47/0.01)	+0.57 [+0.49, +0.64]	+0.01 [-0.03, +0.05]	+0.56 [+0.48, +0.64]	handle necessary
USE	0.82 [0.77, 0.87] (0.82/0.13/0.05/0.00)	0.17 [0.11, 0.22] (0.17/0.54/0.07/0.22)	0.82 [0.76, 0.87] (0.82/0.13/0.06/0.00)	0.07 [0.03, 0.11] (0.07/0.52/0.41/0.01)	+0.66 [+0.58, +0.72]	+0.01 [-0.03, +0.04]	+0.65 [+0.58, +0.72]	handle necessary
COMBINED	0.62 [0.54, 0.69] (0.62/0.35/0.03/0.00)	0.14 [0.09, 0.19] (0.14/0.79/0.07/0.00)	0.50 [0.43, 0.57] (0.50/0.43/0.07/0.00)	0.03 [0.01, 0.06] (0.03/0.78/0.18/0.01)	+0.48 [+0.41, +0.55]	+0.12 [+0.07, +0.17]	+0.36 [+0.29, +0.43]	prerequisite failure (0.62 < 0.70)

set	target	default	n	N_h	H	E_h	rescue H - N_h	H target/default/third
ITT	image	text	5	0.2	0.4	1.0	0.2	0.4/0.4/0.2
ITT	text	image	168	0.161	0.792	1.0	0.631	0.792/0.125/0.083
ITT	text	text	5	0.2	0.8	1.0	0.6	0.8/0.2/0.0
IIT	image	image	64	0.266	0.656	1.0	0.39	0.656/0.219/0.125
IIT	image	text	52	0.192	0.635	1.0	0.443	0.635/0.154/0.192
IIT	text	image	64	0.203	0.797	1.0	0.594	0.797/0.109/0.094
TTT	text	text	116	0.112	0.81	1.0	0.698	0.81/0.086/0.103

and the handle-minus-sham specificity are both >= 0.15 with CIs excluding 0; NOT-NECESSARY-AT-TOKEN if the handle drop's CI lies within +-0.10; otherwise PARTIAL. Qwen3-VL-8B: 3,330 answers, 0 errors, L1 = 0.0 on 3/3, unparseable 0.042; 180 non-default targets per cell.

Secondary (reliability more dependent than retrieval): not supported; COMBINED handle drop minus QUERY handle drop -0.09 [-0.18, -0.01]. Reading: a handle-specific decrement with a null sham is causal necessity of access to the preserved handle-token positions at question time; it does not identify a provenance circuit. Pixtral-12B, minimal QUERY-only replication (arms NORMAL / HANDLE / SHAM; 900 answers, 0 errors, L1 = 0.0, unparseable 0.069, 152 non-default targets), reported independently: NORMAL 0.68 [0.62, 0.75] (0.68/0.24/0.08/0.00), HANDLE 0.32 [0.25, 0.40] (0.32/0.32/0.19/0.16), SHAM 0.69 [0.63, 0.75]; drop +0.36 [+0.28, +0.44] vs sham -0.01 [-0.05, +0.04]; specificity +0.37 [+0.28, +0.45]. PRE-COMMITTED LABEL: prerequisite failure (0.68 < 0.70); the same direction and specificity as Qwen, descriptively only.

I.11 Modality / default-symmetry stress test Design: all 270 neutral-label three-source main items (image-text-text, image-image-text, text-text-text), evidence blocked; final turns NONE, QUERY and COMBINED for every target; stable-handle summary with evidence accessible and its text-only reader as prerequisites. No item selection; the pre-existing capability exclusions are a sensitivity analysis only. Strata recorded per answer and never used for selection: target modality (kind of the queried slot), default modality (kind of the item's blocked no-instruction answer), slot, source set, family. Pre-committed labels on the QUERY rescue difference D = R_text - R_visual: RESCUE-VISUAL if R_visual >= 0.15 with CI excluding 0; MODALITY-SYMMETRIC if the CI of D lies within +-0.15; TEXT-FAVORED if D >= 0.15 with CI excluding 0; VISUAL-FAVORED if D <= -0.15; otherwise INDETERMINATE. Gates: L1 = 0.0 (3/3), unparseable <= 0.10 (0.011). Prerequisites (stable handles, evidence accessible / text reader): image-text-text 0.96 / 0.32, two-image 0.89 / 0.34, three-text 1.00 / 0.32. Default modality: image-text-text image 84 / text 5 / unparseable 1; two-image image 64 / text 26; three-text text 58 / unparseable 32. Primary result (7,290 answers, 0 errors). Retrieval (QUERY): visual targets n = 121, 0.231 -> 0.636 -> 1.000, rescue +0.405 [+0.289, +0.517], binding gain +0.364 [+0.271, +0.460]; text targets n = 353, 0.153 -> 0.799 -> 1.000, rescue +0.646 [+0.592, +0.698], gain +0.201 [+0.161, +0.243]; D = +0.241 [+0.115, +0.368]: RESCUE-VISUAL + TEXT-FAVORED. Combined cue: visual 0.455 -> 0.653 -> 1.000, rescue +0.198 [+0.092, +0.310]; text 0.193 -> 0.731 -> 0.980, rescue +0.538 [+0.482, +0.595]; D = +0.340 [+0.211, +0.463]: RESCUE-VISUAL + TEXT-FAVORED. By default modality (QUERY rescue): image default n = 296, +0.571 [+0.506, +0.636]; text default n = 178, +0.607 [+0.523, +0.688]; combined cue +0.436 and +0.478. By slot (QUERY rescue): +0.54 / +0.71 / +0.50 (n = 114 / 176 / 184); by source set: image-text-text +0.62, two-image +0.48, three-text +0.70; by family: color +0.81, chart +0.55, count +0.36. Candidate shares with stable handles (target / default / third): QUERY visual 0.64 / 0.20 / 0.16, text 0.80 / 0.11 / 0.09; combined visual 0.65 / 0.21 / 0.14, text 0.73 / 0.22 / 0.05. Exact N per target modality x default modality x source set (QUERY):

The same strata for the combined cue:

set	target	default	n	N_h	H	E_h	rescue H - N_h	H target/default/third
ITT	image	text	5	0.4	0.6	1.0	0.2	0.6/0.4/0.0
ITT	text	image	168	0.131	0.601	0.958	0.47	0.601/0.357/0.042
ITT	text	text	5	0.0	0.8	1.0	0.8	0.8/0.2/0.0
IIT	image	image	64	0.453	0.688	1.0	0.235	0.688/0.234/0.078
IIT	image	text	52	0.462	0.615	1.0	0.153	0.615/0.154/0.231
IIT	text	image	64	0.188	0.734	1.0	0.546	0.734/0.156/0.109
TTT	text	text	116	0.293	0.914	1.0	0.621	0.914/0.069/0.017

Sensitivity (pre-existing capability filter from the three-source experiment; 18 count items removed; no outcome-based exclusion): labels unchanged; visual rescue +0.404 [+0.276, +0.529] (n = 104), text +0.665 [+0.611, +0.717] (n = 334), D = +0.261 [+0.127, +0.398]. Limitations fixed in advance: default modality is observed rather than assigned, so the modality contrast is descriptive, not causal; the visual-target evidence is concentrated in the two-image set (116 of 121); MODALITY-SYMMETRIC did not hold, so no symmetry claim is made.

I.1 Analyses of existing data without new model runs : COUNT diagnostic; patching alignment audit (M1/M2 code); probe control analysis. Recorded in design document