Life of Research · Published 3 November 2026 · DOI 10.5281/zenodo.23919442

A Rendered Passage Is Not a Picture: Four Things Called Modality Preference in Multimodal Knowledge Conflict, Why the Reported Reversal of Text Dominance Changes the Channel but Not the Content, and the Reliability-Matched Test That Would Settle It

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. A recurring finding about multimodal language models is that they trust text over images when the two disagree. A 2026 study of thirteen models reports the reverse for one kind of evidence: a passage that contradicts the model's memory is adopted more often when it is shown as a rendered image than when it is typed, and when a rendered passage and a typed passage contradict each other the winning channel changes with order, model and dataset. In this analysis the author reviews and synthesises published results to show that the reversal and the earlier finding are not about the same quantity. "Modality preference" names at least four measurements: how attention is shared across token types, which source a model follows when two sources disagree, how far a named target modality is disturbed by the other one, and whether the same content gives the same answer in two channels. The text-dominance studies mostly vary content type, pitting a written proposition against a perceptual scene; the 2026 study varies the channel while holding the proposition fixed. Typographic attacks, in which words inside an image override what the image shows, indicate that what has been called text dominance travels with propositional content rather than with the token channel. Re-reading the 2026 study's own tables, the author shows that its order effect is structured rather than arbitrary (in 11 of 13 models on one dataset the image gains when it is placed second), that most single-evidence gaps are no larger in magnitude than the study's same-modality controls (though the controls lean negative, and on a signed comparison seven of thirteen NQ-Swap gaps exceed all of them), that its normalized gap hides that, in the cells where the rates can be recovered, two proprietary models are not counted as following either passage on most items, and that for the five models, from three families, it shares with an earlier benchmark the sign of the cross-modal preference agrees on both datasets. Human cue-combination research resolved a similar dispute by showing that dominance follows relative reliability. The paper states a reliability-matched protocol as an algorithm and specifies what a claim of modality bias should report.

Keywords: multimodal large language models; knowledge conflict; modality preference; text dominance; text-as-image; cue combination; position bias; evaluation methodology

1. Introduction

Multimodal large language models now receive evidence in more than one form. A retrieval system may hand them a typed passage, a screenshot of a web page, or a scanned document page, and vision-based retrieval pipelines index pages as images precisely so that layout and figures survive (Faysse et al., 2025; Yu et al., 2025; Cho et al., 2024). Rendering text as an image has also been proposed as a way to compress long contexts, on the grounds that a page of text costs fewer visual tokens than text tokens (Cheng et al., 2025; Wei, Sun and Li, 2025; Li, Lan and Zhou, 2025). In all of these settings the model may meet evidence that disagrees with what it already believes, or with other evidence in the same prompt. Which source wins is a question with practical weight, and the literature on it has produced a slogan.

The slogan is that these models have "blind faith in text". The phrase comes from Deng et al. (2025), who added textual descriptions to vision-centric question answering tasks, sometimes accurate and sometimes corrupted, and found that ten models frequently answered from the text when it disagreed with the image. Related results point the same way. A benchmark built from images that contradict common sense found that almost all tested models leaned on language priors (Lee et al., 2025). A controlled evidence-conflict benchmark found that most of twenty open models favoured the textual context, with LLaVA-1.5-7B choosing the visual answer in only 13.4 percent of decided cases (Zhang et al., 2025a). An attention analysis across five non-text modalities concluded that text dominance is pervasive (Wu et al., 2025). The older visual question answering literature had already documented models answering from question statistics rather than from the image (Goyal et al., 2017; Agrawal et al., 2018).

A 2026 study by Lee, Yoon and Kim, accepted to Findings of EMNLP 2026, reports a result that it presents as contrary to this belief (Lee, Yoon and Kim, 2026). It takes passages from two knowledge-conflict datasets that contradict what each model answers closed-book, and shows each passage either as text or as an image of the same text rendered with a standard font. Across thirteen

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

models, the rendered version is adopted somewhat more often for most models, and much more often for Claude Sonnet 4.5. When a typed passage and a rendered passage are given together and support different wrong answers, which one wins depends on which comes first, on the model, and on the dataset. The authors conclude that modality preference "is thus better understood as an artifact of the evaluation setup than as a fixed property of the model."

Both bodies of evidence are careful, and they appear to contradict each other. This paper argues that they mostly do not, because they measure different things, and that the part of the disagreement that remains is about reliability rather than about modality. The argument has four parts.

First, "modality preference" is used for at least four distinct quantities: a ratio of attention per token across modalities; the share of disagreeing items on which the model's answer follows one source rather than the other; the loss in accuracy when a model asked about one modality is shown a conflicting other modality; and the agreement of answers when identical content is delivered in two channels. Section 4 separates them and shows which studies measure which.

Second, the studies that report text dominance and the 2026 study that reverses it change different variables. In the former, the text carries a proposition and the image carries a scene; the conflict is between kinds of content. In the latter, both carry the same proposition and only the channel changes, tokens against pixels. Typographic attacks, in which a word written into an image overrides the object the image shows (Cheng et al., 2024; Qraitem et al., 2024; Westerhoff et al., 2025), fill in the remaining cell: propositional content can win from inside the pixel channel. Section 5 argues that the evidence fits content-type dominance better than channel dominance, and that on this reading the 2026 result does not contradict the earlier one.

Third, preference is not bias until it is measured against reliability. Human perception research went through the same argument about "visual dominance" and settled it by showing that the sense with lower variance dominates, so that blurring the visual stimulus reverses the effect (Ernst and Banks, 2002; Alais and Burr, 2004). One multimodal-model study has already found the analogous law, with the probability of following a modality falling steadily as its relative uncertainty rises (Zhang et al., 2025b). Section 6 develops what this implies for each construction.

Fourth, the 2026 study's own tables support a more specific reading than its conclusion. Section 3 shows that its order effect has a direction, that most of its single-evidence gaps are no larger in magnitude than its own controls, that its normalized metric conceals how often proprietary models are not counted as following either passage, and that two cells are each reported with opposite signs in two of its tables. Section 8 shows that for the five models, from three families, it shares with an earlier benchmark built on natural images, the sign of each family's cross-modal preference agrees (Figure 1b).

The contribution is a separation and a set of consequences, stated as six claims.

Claim 1. The four quantities called modality preference are not interchangeable, and the 2026 reversal and the earlier text-dominance results sit in different cells of a channel-by-content table (Table 2).

Claim 2. The order effect in the 2026 study is structured: for most open-weight models, whichever channel is placed second gains, and in the image-first order the familiar text dominance reappears for 9 of 12 models on one dataset and 8 of 12 on the other, with one disputed cell set aside on each (Section 3.4).

Claim 3. Most single-evidence gaps in the 2026 study are smaller in magnitude than the largest of its same-modality controls on the same dataset, and those controls are the only noise yardstick it reports; because the controls lean negative, a signed comparison puts three ConflictQA gaps and seven of thirteen NQ-Swap gaps beyond every control (Section 3.2).

Claim 4. The model-family ordering, with Qwen2.5-VL and InternVL3 leaning toward the image and LLaVA-1.6 leaning toward text, replicates across three studies with different constructions; the rates do not (Section 8).

Claim 5. A preference measured without per-item reliability cannot be called a bias, because a reliability-weighted observer also prefers one source; a balance-point design separates the two (Section 6).

Claim 6. A reliability-matched protocol, given as Algorithm 1, specifies the controls and reports a claim of modality bias should carry (Section 10).

Table 1 consolidates the measured values the argument rests on.

2. Scope, sources and prior records

The paper concerns vision-language and omni-modal models that take images and text as input and produce text. Its subject is behaviour under conflicting inputs: an image and a text that support different answers, or an input of either kind that contradicts the model's parametric knowledge. Audio-visual conflict enters at one point, as evidence that the pattern is not peculiar to vision (Sudheer et al., 2026). Hallucination under consistent inputs, and the decoding methods proposed against it (Leng et al., 2024), are cited only where they bear on preference.

The evidence base is 2016 to 2026 work, verified against arXiv or Crossref records. Several of the most relevant results are 2026 preprints without peer review; where a claim depends on one, the text says so. The anchor study (Lee, Yoon and Kim, 2026) was read in full, including its

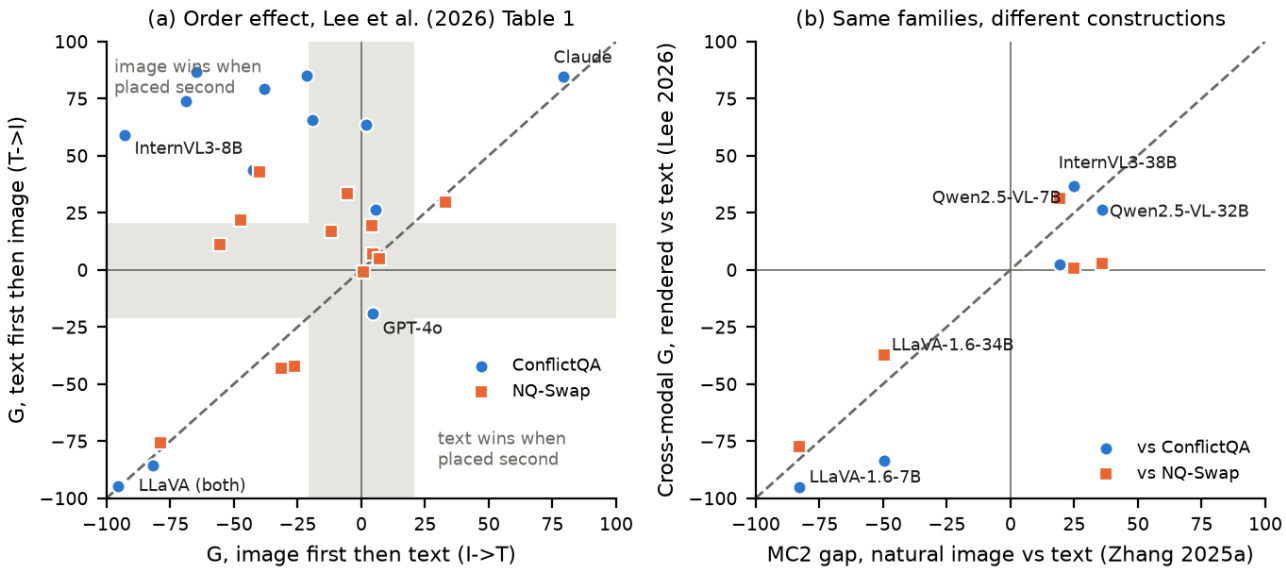


Figure 1. (a) Normalized image-over-text gap G in the anchor study's multi-evidence setting, image-first order (x) against text-first order (y), 13 models on two datasets; values from Lee, Yoon and Kim (2026), Table 1. Shaded bands mark the magnitude of the same study's same-modality controls (Table 10, |G| up to about 20); the controls lean negative (largest positive value +5.84 on ConflictQA, +4.63 on NQ-Swap), so the positive half of each band is wider than any control. Two plotted cells are disputed between the study's tables (Section 3.5): Qwen2.5-VL-7B on NQ-Swap at (33.10, 29.51) and MiniCPM-o2.6 on ConflictQA at (5.42, 26.36); both are shown at their Table 1 values and left out of the counts in the text. (b) For the five model families and sizes reported in both studies, the cross-modal gap in the anchor study (Table 11, both orders pooled) against the MC2 benchmark's vision-versus-text accuracies (Zhang et al., 2025a, Table 13), converted by the author to the same scale as (V - T)/(V + T) x 100. The Qwen2.5-VL-7B NQ-Swap point averages in the disputed cell and is not counted in Section 8's sign comparison.

appendix tables, and every value attributed to it below was taken from those tables. Where the paper combines two published values into a third, as in Section 3.3, the arithmetic is the author's and is labelled as such.

Three of the author's prior records border this one and are cited rather than repeated. One argued that context-versus-memory adoption rates in text-only knowledge-conflict studies are set by how the conflict is constructed, so that moderators replicate across studies while point estimates do not (Mahendrakar, 2026a). This paper does not re-argue that thesis for text; it addresses a question that record excluded, which modality wins when two channels disagree, and where the construction argument applies it is cited. A second argued, for audio-language benchmarks, that removing a modality measures answerability rather than reliance (Mahendrakar, 2026b). A third analysed why degrading a rendered harmful request raises jailbreak success in one study and lowers it in another (Mahendrakar, 2026c). The present question, which source a model follows and whether that choice should be called a bias, is not addressed in any of them.

3. What the anchor study measured, read at table resolution

3.1 The design

Lee, Yoon and Kim (2026) use two datasets. ConflictQA pairs questions from PopQA with passages generated by a language model to support either the model's remembered answer or a counter-answer (Xie et al., 2024; Mallen et al., 2023). NQ-Swap takes Natural Questions passages and substitutes the answer entity (Longpre et al., 2021). For each of thirteen models, the authors first ask each question closed-book five times and keep only items on which all five answers agree; that answer is taken as the model's memory. This filter is a strength. A conflict that is not checked against what the model actually believes is not a conflict with its memory, and the authors point out that earlier multimodal work did not always check.

Each passage is then shown either as text or as a rendered image: black DejaVu Sans at 24 points on white, wrapped at 800 pixels. The authors report that font sizes from 24 to 32 and widths from 600 to 1,000 pixels made little difference, and that three open models transcribe the rendered passages almost perfectly, with ROUGE-L between 98.37 and 99.97 (their Appendix F). The outcome, the external-source following rate, counts an item only if all three sampled answers match the passage's answer. The headline metric is a normalized gap, the difference between the image and text following rates divided by their sum, scaled to lie between -100 and +100.

In the single-evidence setting, one passage is shown, in one channel or the other. In the multi-evidence setting, the model's memory agrees with the true answer, and two different counterfactual passages are shown together, one typed and one rendered, in both orders,

with the assignment of passages to channels swapped and averaged. A second appendix table reports two same-modality controls, both passages typed or both rendered.

3.2 Single-evidence gaps, and the yardstick for them

The single-evidence result is described by its authors as a weak but pervasive image preference. The description is accurate as to sign: 9 of 13 models have a positive normalized gap on each dataset (their Table 1). It needs a qualification as to size. On ConflictQA, 11 of 13 gaps are smaller than 10 in magnitude, and on NQ-Swap 9 of 13 are. In percentage points (their Table 9), 7 of 13 single-evidence differences on ConflictQA are under 5 points.

The study reports no confidence intervals. The nearest thing to a noise estimate in the paper is the pair of same-modality controls in its Table 10. When both passages are typed, the normalized gap on ConflictQA ranges from -19.51 to +5.84 across models; when both are rendered, from -20.16 to 0.00. On NQ-Swap the ranges are -4.26 to +4.63 and -11.57 to +3.96. These controls come from the multi-evidence design rather than the single-evidence one, so they are not a confidence interval for the single-evidence gaps. They are, however, the only quantity the study offers for how large a gap can be when the channel does not differ. The yardstick used here is a magnitude band: |G| up to 20.16 on ConflictQA and up to 11.57 on NQ-Swap, the largest control magnitude on each dataset. Most single-evidence gaps fall inside that band. The single-evidence gaps whose magnitude exceeds it are Claude Sonnet 4.5 on both datasets (+42.86 on ConflictQA, +22.63 on NQ-Swap) and, on NQ-Swap only, Qwen3-Omni (+12.98), GPT-4o (-14.74) and LLaVA-v1.6-7B (-34.17). GPT-5.4's +15.29 on ConflictQA is smaller in magnitude than the largest ConflictQA control, but it is not inside the signed control range.

The magnitude convention has a limitation that matters here. The controls are not symmetric about zero. On ConflictQA, 23 of the 26 control values are negative, one is 0.00, and the largest positive one is +5.84; on NQ-Swap the controls run from -11.57 to no higher than +4.63. A band of |G| therefore treats as noise positive gaps that no control produced. Compared with the signed ranges instead, three ConflictQA gaps lie beyond every control, Claude Sonnet 4.5 (+42.86), GPT-5.4 (+15.29) and GPT-4o (+9.75), and so do seven of the thirteen NQ-Swap gaps: Claude Sonnet 4.5 (+22.63), Qwen3-Omni (+12.98), InternVL3-38B (+8.53), Qwen2.5-Omni (+6.74), Qwen2.5-VL-32B (+5.49), GPT-4o (-14.74) and LLaVA-v1.6-7B (-34.17). On NQ-Swap, the statement that most gaps fall inside the controls holds only under the magnitude reading.

The paper's per-model sample sizes are large, from 589 to 4,049 items on ConflictQA and from 2,169 to 3,310 on NQ-Swap (its Table 6), so a gap of a few points may well be statistically real for a given model. What the absence of intervals prevents is a claim about which small gaps are real. The direction-level summary, that most models adopt the rendered passage slightly more, is therefore best read as a hypothesis the data are consistent with rather than as an established property.

3.3 What the normalized gap does not show

Normalizing by the sum of the two following rates lets the authors compare models whose overall willingness to follow context differs. It also hides that willingness. Because the paper reports both the normalized gap (Table 1) and the absolute difference in percentage points (Table 9), the sum of the two following rates can be recovered by dividing the second by the first, and from the sum and the difference each rate follows. The arithmetic below is the author's, from the published rounded values; it is unstable when the normalized gap is small, so it is reported only where the gap exceeds 5 in magnitude.

For Claude Sonnet 4.5 in the single-evidence setting on ConflictQA, the implied rates are about 53 percent for the rendered passage and about 21 percent for the typed one. The large image preference therefore arises in a model that mostly declines a typed counter-memory passage and accepts a rendered one about half the time. For LLaVA-v1.6-34B on the same dataset the implied rates are about 80 and 89 percent, so its small text preference sits near a ceiling at which both channels are followed most of the time. A gap measured near a ceiling is compressed; the same underlying channel effect would produce a larger normalized gap in a model that resists context more.

In the multi-evidence setting the recovered sums have a clearer meaning, because both rates come from the same inputs and cannot exceed 100 percent together. The remainder is the share of items not counted as consistently following either passage. Because the anchor credits an item "only when all three responses match the external answer candidate" (Lee, Yoon and Kim, 2026, Section 3.2), this remainder includes answering from memory, answers that split across the three samples, and other answers; the anchor does not report how it divides. For Claude Sonnet 4.5 on ConflictQA, the implied share not counted as following either passage is about 57 percent in the image-first order and about 67 percent in the text-first order; on NQ-Swap in the text-first order it is about 78 percent. For GPT-5.4 on ConflictQA in the text-first order it is about 59 percent. The normalized gaps for the three ConflictQA cells, between +63 and +85, and for the NQ-Swap cell, +19.38, describe the minority of items on which the model was persuaded. The headline "image preference" of these two models in the multi-evidence setting is real for that minority; on most items, these models did not consistently adopt either wrong passage. The most common outcome for the study's strongest models is thus one its headline

metric does not describe at all.

3.4 The order effect has a direction

The anchor's second claim is that when both channels are present the preferred one is essentially arbitrary. Its own numbers show structure. Figure 1a plots, for every model and dataset, the normalized gap in the image-first order against the gap in the text-first order. A point above the diagonal means the image did better when it came second.

On ConflictQA, the text-first gap exceeds the image-first gap for 11 of 13 models, by more than 10 points for 9 of them; the exceptions are GPT-4o, for which the rendered passage does better when it comes first (a shift of about 23 points), and LLaVA-v1.6-34B, which is strongly text-preferring in both orders. On NQ-Swap, where the two passages differ only in the substituted entity, the pattern is weaker: 8 of 12 models above the diagonal, leaving out the disputed Qwen2.5-VL-7B cell (Section 3.5), 6 of them by more than 10 points, and two (MiniCPM-o2.6 and LLaVA-v1.6-34B) shifting more than 10 points the other way. For the Qwen and InternVL families on ConflictQA the shift is between 85 and 152 points: a model that prefers the typed passage by a wide margin when it is placed second prefers the rendered passage by a wide margin when that is placed second.

For most open models this is a recency effect with respect to channel, and it is consistent with what is known about position in language models: performance depends on where relevant information sits in the context (Liu et al., 2024a), primacy and recency effects are widespread across tasks (Guo and Vosoughi, 2024), and reordering multiple-choice options moves accuracy by large margins (Pezeshkpour and Hruschka, 2023). For multi-image inputs, open-source models have been reported to reason best about images placed later (Tian et al., 2025). The anchor's same-modality controls, whose gaps are at most about 20 in magnitude, suggest that order matters much less when the channel does not change, so that it is the interaction of position with channel that is large. The anchor does not state how the image and text labels are assigned in its same-modality controls, so the controls cannot be read as a pure position effect, and this paper does not read them that way.

Two further observations follow from the same table. First, in the image-first order, 9 of 12 models on ConflictQA and 8 of 12 on NQ-Swap have negative gaps, that is, they follow the typed passage, with one disputed cell set aside on each dataset (Section 3.5). The text dominance the anchor sets out to contradict is reproduced in its own data, in one of its two orders. Second, the direction of the order effect is not the same across studies. Deng et al. (2025) found, for Phi-3.5, that placing text tokens before image tokens increased text preference, which is a primacy effect for text; the anchor finds, for most of its models, that the later channel gains. Position matters in both studies and in opposite directions, which is consistent with an effect that depends on each model's training format rather than on a general law.

"Arbitrary" is thus the wrong word for the multi-evidence result. The preferred channel is predictable from order for most open models, from family for LLaVA (text in both orders) and Claude (image in both orders), and less predictable for GPT-4o, GPT-5.4 and MiniCPM. What is true is that no single channel is preferred, and that a study run in only one order would report a confident preference of either sign.

3.5 Two cells reported with two signs

For Qwen2.5-VL-7B on NQ-Swap in the image-first order, the anchor's Table 1 and its Table 10 give a normalized gap of +33.10, an image preference, while its Table 9 gives an absolute difference of -20.18 points, a text preference. A positive normalized gap and a negative difference cannot both describe the same pair of rates. For MiniCPM-o2.6 on ConflictQA in the image-first order, the conflict runs the other way: Table 1 gives a normalized gap of +5.42 and Table 9 a difference of +4.44 points, while Table 10 gives -26.18, which is the same value as the model's NQ-Swap entry for that order and may be a copying error. In each case one of the entries is in error, and this paper treats both cells as unknown in its counts. The discrepancies do not affect the anchor's qualitative conclusions. They are noted here because both cells bear on the counts in Section 3.4: with them set aside, the image-first text preference is 9 of 12 on ConflictQA and 8 of 12 on NQ-Swap. Figure 1 still plots both cells at their Table 1 values, so that the panel reproduces that table as published; the caption marks them. The Qwen2.5-VL-7B cell also enters the order-pooled NQ-Swap value used in Section 8, which is therefore set aside there too.

3.6 What survives

Four findings in the anchor study are well supported by its own data and are not in dispute here. Rendering a passage changes how often it is adopted against memory, most clearly for Claude Sonnet 4.5. When channels compete, order changes the outcome by large margins for most open models. Model family predicts the direction where order does not. And transcription accuracy is not the bottleneck, since three open models transcribe the renderings nearly perfectly yet treat them differently from typed text. The last point matches two independent results: a model that transcribed 87.6 percent of rendered questions exactly answered only 48.4 percent correctly (Wang et al., 2026b), and GPT-4o's answers about documents agreed with its own OCR of those documents in only 75.26 percent of cases (Shao et al., 2024).

What is not supported is the step from these findings to the claim that they contradict text dominance. That step

requires that the anchor and the text-dominance studies measure the same quantity, and Section 4 argues they do not.

4. Four quantities under one name

Table 1 collects the measured values that the rest of the argument uses, one row per cited result. The rows are grouped by the quantity each one measures, which is the point of this section.

4.1 Attention share

The broadest use of "text dominance" is a claim about computation: the model attends more to text tokens than to non-text tokens. Wu et al. (2025) define a modality dominance index as the ratio of average attention per text token to average attention per non-text token, and report that it is high across images, video, audio, time series and graphs, except graphs under standard token counts. They attribute much of it to redundancy, since non-text inputs arrive as many tokens carrying little information each, and show that compressing the non-text tokens brings LLaVA-7B's index from 10.23 to 0.86. An earlier inference-efficiency study made the same observation from the other side: in the deeper layers of several vision-language models, attention to visual tokens is so inefficient that half of them can be pruned after the second layer with little loss (Chen et al., 2024).

An attention ratio per token is a fact about how a network distributes computation over its input. It is not a fact about which source the answer follows. A model can spread little attention over each of a thousand image tokens and still answer from the image, and a high per-token text share is what one would expect if a few text tokens carry the question. Wu et al.'s own remedy, compressing image tokens until the ratio is balanced, illustrates the gap: the ratio responds to token count by construction. Attention share is useful for diagnosing where an imbalance might come from; it cannot settle which source wins a conflict.

4.2 Following under disagreement

The second quantity is behavioural. Present inputs that support different answers and count which answer the model gives. Every study in this family computes a ratio, but the ratios differ in what they condition on.

Deng et al. (2025) compute a text preference ratio over items on which the model's image-only answer and its text-only answer differ, among the answers that match one of those two. They do this for three kinds of text: a description written by GPT-4o to match the image, one written to contradict it, and an irrelevant Wikipedia passage. The prompt introduces the text as a description based on the image and warns that it may be inaccurate. Two features of this design matter. The first is that in the match condition an item enters the ratio only when the model's own image-only answer disagrees with a description written to support the correct answer; on such items, following the text usually means following the more accurate source. Deng et al. report high text preference in the match condition, 87 percent for Claude Haiku on VQAv2, and treat it alongside the corruption condition as evidence of text bias. They do not report how many of those items had a correct image-only answer, so the share of "text bias" that is reliance on a correct description over a wrong visual reading is not known. The second feature is that the text is declared to be derived from the image. Under that framing, the image is the primary source and the description a secondary report about it, and preferring the report in a conflict is a defect in a way it would not be if the two were independent witnesses. The corruption condition is diagnostic of over-trust in text; the match condition is not.

Zhang et al. (2025a) build their MC2 benchmark from MS-COCO images with generated textual contexts that agree with the scene except in the attribute the question asks about, and verify that models answer correctly from either context alone (all above 95 percent). They ask the question without saying which context to use, discard items whose answer changes when the options are swapped, and compute the share of decided items answered from the image. This is closer to a pure default: both sources are reliable in isolation and the instruction does not favour either. Most of the twenty open models favour text; Qwen2.5-VL and InternVL3 lean toward the image, increasingly so with size.

Zhang et al. (2025b) start from the observation that such ratios still mix two things, how hard each source is for the model on a given item and how much the model favours one source when difficulty is equal. They construct items whose visual and textual difficulty are varied in steps, measure each model's entropy on each source alone, and plot the probability of following text against the relative uncertainty. For all six models the curve falls steadily as the text becomes relatively harder. The point at which the model follows each source half the time, the balance point, is the paper's measure of inherent preference. On their data, LLaVA models have a balance point near zero or on the text side and Qwen models a balance point on the vision side, and they show that a raw following ratio can rank two models the wrong way round when one is simply better at the visual task.

The anchor's multi-evidence gap is a following ratio of this kind, computed over items where both passages contradict memory and each other and the model answered with one of them. Its single-evidence gap is not; it is a comparison of adoption rates across two prompts, each containing one source and the model's memory.

Table 1. Measured values the argument rests on, one row per cited result, with the source named on each row. "pp" is percentage points; G is the anchor study's normalized image-over-text gap. Values are as reported; counts across models ("9/13") are the author's tallies of the source's table.

Source	Construction	Models	Reported value	Quantity
Wu et al. 2025	attention per token, five modalities	LLaVA-7B	MDI 10.23, 0.86 after token compression	attention share
Deng et al. 2025	natural image vs written description	Claude Haiku, VQAv2	TPR 87% (match), 83% (corrupt)	following, conditioned on disagreement
Deng et al. 2025	same, corrupted text	Qwen2-VL-7B	accuracy falls to 59, 63, 52% of base on VQAv2, DocVQA, MathVista (96% on brand detection)	following
Deng et al. 2025	text tokens moved before image	Phi-3.5	text preference rises	position
Zhang et al. 2025a	image vs generated text, both solvable	LLaVA-1.5-7B	vision ratio 13.4%	following, default
Zhang et al. 2025a	same	Qwen2.5-VL-32B; InternVL3-38B	vision/text acc 62.9/29.7; 51.1/30.8	following, default
Zhang et al. 2025b	graded unimodal difficulty	six models	following falls monotonically with relative entropy	following vs reliability
Hua et al. 2025	image vs caption, target named	InstructBLIP, LLaVA-OneVision	larger loss when asked for the caption	target interference
van Sprang et al. 2026	same content in text, image, mixed	15 models	answer agreement 6.6-90.7%; text best ($t = 17.7$)	render equivalence
Sun et al. 2026	rendered vs typed benchmarks	7 models	outputs 5-19x shorter; font moves accuracy up to 47 pp	render equivalence
Wang et al. 2026b	question moved into the image	6 models, 4 benchmarks	-17.8 pp mean; 87.6% transcribed, 48.4% correct	render equivalence
Liu et al. 2026	visualized vs pure text	Qwen3-VL-8B, MMLU	75.99 to 68.46%	render equivalence
Lee, Yoon and Kim 2026	rendered vs typed passage vs memory	13 models	single G: 9/13 positive, 11/13 below 10 (ConflictQA)	adoption by channel
Lee, Yoon and Kim 2026	same	Claude Sonnet 4.5	single G $+42.86$ / $+22.63$	adoption by channel
Lee, Yoon and Kim 2026	rendered vs typed, both counterfactual	13 models	sign flips with order: 8/13, 6/13	following, position
Lee, Yoon and Kim 2026	same-modality controls	13 models	abs G up to 20.16 / 11.57	noise yardstick
Lee, Yoon and Kim 2026	harmful instruction rendered	3 families	ASR $+6.99$ pp mean; 126 of 156 positive	safety
Cheng et al. 2024	words written into images	LVLMs	degradation 42.07%, 13.90% after mitigation	content in pixel channel
Gong et al. 2025	harmful request as typography	6 open LVLMs	ASR 82.50%	content in pixel channel
Tian et al. 2025	image position in multi-image input	open vs proprietary	open models best on later images	position
Zhu et al. 2024	visual vs textual route to same fact	LLaVA 7B/13B/34B	conflict rate at least 21.36, 28.10, 20.53%	cross-route consistency
Shao et al. 2024	document answer vs own OCR	GPT-4o	75.26% consistent	cross-route consistency
Liu et al. 2025	images that defy common sense	9 models	memory over image on about 20% of queries	image vs memory
Lietzow et al. 2026	image vs memory, head ablation	3 families	ablation flips 68-96% to visual; 0.8-7.5% reverse	image vs memory

4.3 Interference with a named target

Hua, Yun and Pavlick (2025) ask a different question. They give a model an image and a caption that disagree (an image of one class, a caption naming another) and instruct it to report what the image shows, or what the caption says. A perfect model answers each instruction correctly regardless of the conflict. The measured quantity is how much the irrelevant modality pulls the answer away from the named one. They find that InstructBLIP and LLaVA-OneVision lose much more accuracy when asked about the caption than when asked about the image, a pull toward the image, while Qwen2.5-VL mostly reports whichever modality it is asked for, except on CIFAR images where text pulls it. They describe the image-biased models as showing blind faith in images rather than in text.

This is a test of selective attention under instruction, not of preference between sources of evidence. Its result can differ from a default-preference result for the same model, and it does: LLaVA-OneVision-7B in MC2 answers from the textual context far more often than from the image, with vision and text accuracies of 20.9 and 58.6 percent (Zhang et al., 2025a, Table 13), while in Hua et al.'s task it is pulled toward the image. These are not

contradictory findings. One measures what the model does when it must choose without guidance, the other what it fails to suppress when told which input matters.

4.4 Render equivalence

The fourth quantity holds content fixed and changes the channel. A growing set of benchmarks renders text-only tasks as images and compares accuracy. Li et al. (2023) found large drops for open models when instructions were given only in pixels. Van Sprang et al. (2026) presented the same questions as text, as an image, and as a mixture to fifteen models, analysed both all items and those each model transcribed correctly, and found agreement across the three forms ranging from 6.6 to 90.7 percent across models, with text the best-performing form overall. Liu et al. (2026) report the same direction across more than thirty models, with Qwen3-VL-8B falling from 75.99 to 68.46 percent on MMLU questions when they are rendered. Wang et al. (2026b) moved the question itself into the image and found lower accuracy in all 24 model-task pairs they tested, by 17.8 points on average. Sun et al. (2026) found the gap on synthetic renderings to be large for multi-step mathematics and modest for knowledge questions, smaller or reversed on real document images, and driven mainly by much shorter outputs in image mode, which they describe as a collapse of reasoning effort rather than of perception.

The anchor's single-evidence comparison belongs in this family with one difference: the outcome is not accuracy on a task but adoption of a passage against memory. In the render-equivalence literature the typed channel usually does better. In the anchor, the rendered channel is usually adopted more. These are not in tension, because doing better at a task that requires reasoning over a passage is not the same as being more persuaded by it. Section 6.3 returns to a possible link between them.

4.5 What the four quantities share, and what they do not

The four quantities share a vocabulary and almost nothing else. An attention ratio can be high while answers follow the image. A following ratio conditioned on disagreement can be high because the image was misread. A target-interference loss can point toward the image in a model whose default preference is text. And a render-equivalence gap can favour text in accuracy while the rendered passage is adopted more often against memory. A claim that a model "prefers text" is underspecified until it says which of the four it means, and a claim that a new result "contradicts" text dominance is underspecified until it says that the two results measure the same one.

5. Channel and content

5.1 Two variables, not one

The text-dominance studies and the anchor differ in a variable that the word "modality" covers. In Deng et al. (2025) and MC2 (Zhang et al., 2025a), the text is a sentence asserting something and the image is a scene from which an answer must be perceived. Two things change at once between the sources: the channel (tokens or pixels) and the kind of content (a stated proposition or a perceptual array). In the anchor, both sources are the same sentence; only the channel changes. Table 2 places the main studies in a grid formed by these two variables, and adds the reference point against which preference is measured.

5.2 The missing cell: words inside images

If text dominance were a preference for the token channel, a word written into an image should behave like the image, not like text. Typographic attacks show the opposite. When misleading text is written onto an image, vision-language models misclassify the object toward the written word; Cheng et al. (2024) report a performance degradation of 42.07 percent from such attacks on large vision-language models, which their mitigation reduced to 13.90 percent; Qraitem et al. (2024) report reductions in classification performance of up to 60 percent when a model is attacked with words it generated itself; and a real-world collection of handwritten attack images, SCAM, confirms that the effect is not an artefact of synthetic text (Westerhoff et al., 2025). In this cell both inputs arrive through the pixel channel, and the propositional one wins.

The same holds for instructions. FigStep rendered harmful requests as typographic images and reported an average attack success rate of 82.50 percent on six open models, attributing the success to safety training that covers text embeddings but not visual ones (Gong et al., 2025). MM-SafetyBench found that query-relevant images, including typographic ones, compromise models whose language backbones are safety-aligned (Liu et al., 2024b). The anchor's own safety experiment extends this by moving the entire instruction into the image and reports a mean rise of 6.99 points in attack success over the benchmark's baselines, positive in 126 of 156 comparisons (Lee, Yoon and Kim, 2026). A request delivered as pixels is still read and acted on as a request.

The pattern across the three cells is therefore this. Proposition against scene, with the proposition typed: the proposition tends to win (Deng et al., 2025; Zhang et al., 2025a). Proposition against scene, with the proposition in pixels: the proposition tends to win (typographic attacks). Proposition against the same proposition, in different channels: no consistent channel wins, and with one source against memory the pixel version is somewhat more persuasive (Lee, Yoon and Kim, 2026). Read together, these suggest that the

Table 2. What each line of work holds fixed. "Propositional" means the input asserts the answer in words; "perceptual" means the answer must be read off a scene. Direction is as each source reports it.

Study	Source A	Source B	Same content?	Reported direction
Deng et al. 2025	typed proposition	perceptual image	no	toward A (text)
Zhang et al. 2025a	typed proposition	perceptual image	no	toward A, except Qwen2.5-VL, InternVL3
Hua et al. 2025	typed caption	perceptual image	no	varies; some models toward B
Typographic attacks	words in pixels	perceptual scene in pixels	no	toward A
van Sprang et al. 2026	typed question	rendered question	yes	typed answered better
Lee, Yoon and Kim 2026, single	typed passage vs memory	rendered passage vs memory	yes	rendered adopted more, mostly
Lee, Yoon and Kim 2026, multi	typed passage	rendered passage	same kind	follows order and family

dominance documented by the earlier literature attaches to propositional content, not to the token channel. On that reading the anchor's result is not a reversal of text dominance. It is a test of a different variable, and its result, that the channel matters but not in a fixed direction, is compatible with the earlier finding.

This is an inference from cross-study comparison, with different models and tasks in each cell, and it should be held at that level. No single study has varied channel and content in a factorial design on the same items and models. Section 10 specifies one.

5.3 Role: evidence, instruction and question

A third variable is the role the text plays. A passage offered as evidence, an instruction telling the model what to do, and a question to be answered all arrive as text, and the rendered-text results suggest they are not treated alike when rendered. A rendered question exerts less control over the answer than a typed one even when transcribed correctly (Wang et al., 2026b). A rendered harmful instruction is less often refused than a typed one (Gong et al., 2025; Lee, Yoon and Kim, 2026). A rendered counter-memory passage is somewhat more often adopted (Lee, Yoon and Kim, 2026). One coherent description of these three results is that rendering weakens the treatment of text as an instruction to be parsed and evaluated, which lowers both the control a rendered question exerts and the scrutiny a rendered request receives, while leaving the content available as material to be copied into an answer. That description is a conjecture. It is consistent with Sun et al.'s observation that image mode shortens reasoning (Sun et al., 2026) and with FigStep's account of uneven safety coverage, but no study has measured it directly, and Section 12 lists it among the open questions.

6. Preference relative to reliability

6.1 The human precedent

Perception research spent decades on a claim of the same shape as text dominance. When vision and another sense disagree, people's judgments often follow vision; the Colavita effect, in which people sometimes fail to respond to a sound presented with a light, was long treated as a paradigm case of visual dominance. Two lines of work changed the reading. Ernst and Banks (2002) measured the variance of visual and haptic estimates of an object's height and showed that combined judgments behave like a maximum-likelihood integrator that weights each sense by its reliability, so that vision dominates when its variance is the lower of the two. Alais and Burr (2004) showed the same for the ventriloquist effect: with a sharp visual stimulus, vision captures the location of a sound; with a heavily blurred one, sound captures vision; in between, neither dominates. A later causal-inference model added that observers also infer whether the two signals come from one cause, and integrate them only to the extent that they do (Kording et al., 2007). And even the Colavita effect, long resistant to manipulation, was reversed by a change of task (Ngo et al., 2011).

The lesson is not that people are optimal. It is that a measurement of which sense wins, taken without measuring how reliable each sense was on the trials in question, does not identify a bias. An ideal observer with no preference at all will follow vision on most trials of a task in which vision is more precise, and will appear "visually dominant". The bias, if there is one, is the departure from the reliability-weighted prediction.

6.2 The same law in multimodal models

Zhang et al. (2025b) found the machine counterpart. On their graded datasets, each model's probability of following a modality decreases as that modality's uncertainty rises relative to the other's, and the curve's crossing point gives a measure of preference with reliability held equal. Deng et al. (2025) report a coarser version: grouping items by the model's certainty on each source alone, models follow the image when image certainty is high and text certainty low, follow the text in the reverse case, and give neither answer more often when both are low. A test-time adaptation study of a different architecture describes the failure mode from the other side: when one branch becomes unreliable under shift but stays confident, it dominates the fused prediction (Chen et al., 2026). Each of these says that which source wins depends on each source's reliability as the model registers it, and that apparent preferences can

move with the item mix.

This has a direct consequence for the macro-level ratios in Section 4.2. Zhang et al. (2025b) show that a following ratio can rank models in the wrong order of inherent preference when the models differ in unimodal skill: on their data, Qwen2-VL looks more vision-following than Qwen2.5-VL in raw ratio because it finds their visual items easier, while Qwen2.5-VL has the more vision-leaning balance point. MC2 reports Qwen2-VL-7B as strongly text-preferring on its own items, with vision and text accuracies of 15.3 and 78.3 percent (Zhang et al., 2025a). The two studies disagree about the same family, and the balance-point argument predicts that they should be able to, because the items differ in relative difficulty.

6.3 What reliability means for each construction

The normative question, which source a model should follow, has a different answer in each construction, and a preference is a defect only relative to that answer.

In the anchor's single-evidence setting, the two prompts contain the same passage, so the evidence is equally reliable by construction, and so is the memory it contradicts. Any difference in adoption is a presentation effect and is a defect under any reasonable standard. This is the cleanest test in the literature, and on it the result is that rendering raises adoption by a small amount for most open models and a large amount for Claude Sonnet 4.5. Sun et al.'s finding offers one candidate mechanism, which is a conjecture: if image input shortens deliberation (Sun et al., 2026), a rendered counter-memory passage may be checked against memory less often and copied more often. The account predicts that adoption gaps should shrink when the model is required to reason before answering, and should be largest in models that resist typed counter-evidence most, as Claude Sonnet 4.5 does in Section 3.3. Neither prediction has been tested.

In the anchor's multi-evidence setting, both passages are counterfactual and contradict each other as well as memory. There is no basis for following either channel, and a reliability-weighted observer that registers the contradiction should fall back on its memory or report the conflict. Section 3.3 found that Claude Sonnet 4.5 and GPT-5.4 are not counted as following either passage on most items, in the cells where the rates can be recovered; that is compatible with such behaviour but does not show it, because the anchor does not report how often those answers matched memory. The order effects among open models are defects, but they are defects of conflict handling generally rather than of modality: text-only work reports that models side with evidence that appears more often or agrees with memory (Jin et al., 2024; Xie et al., 2024), and that they recognise that a conflict exists more reliably than they locate it or give answers that respect both sides (Wang et al., 2024).

In Deng et al.'s design the image is the primary source and the description derived from it, so following a corrupted description is a defect; following a correct description over a misread image is not. In MC2 both sources are reliable in isolation and the instruction is neutral, so the vision ratio is a default preference, but without per-item uncertainty it confounds preference with relative ease. In Hua et al.'s design the instruction names the target, so any pull from the other modality is an error. Only in Zhang et al. (2025b) is the preference estimated at matched reliability; Deng et al. measure per-item certainty but report it only as binned proportions.

6.4 Two kinds of reliability

One further distinction matters for the rendered-passage case. The reliability of a source to the model has two components: how well the model can extract its content (perceptual reliability), and how far the model should trust the content once extracted (epistemic reliability). For a rendered passage that the model transcribes at a ROUGE-L of 99, perceptual reliability is close to that of typed text, so the reliability-weighting account predicts no channel effect. The anchor finds one. This is therefore the one setting in which the reliability account, as usually stated, does not explain the observed preference, and the departure is informative: it implies that something other than how well the content is read differs between channels. Whether that something is reasoning effort, as Sun et al. (2026) suggest for task accuracy, positional handling of image tokens (Wang et al., 2026a), or a learned association between images and authoritative documents, is not known.

7. Position as a model-by-format interaction

Section 3.4 showed that in the anchor's multi-evidence setting the channel placed second usually gains, and that Deng et al. (2025) found the opposite direction for Phi-3.5. Three further results bear on why position and channel interact.

First, positional encodings were designed for one-dimensional text. When rotary position embeddings are applied to a sequence that interleaves image patches and text tokens, the relative positions of text and image tokens enter attention as if they were distances in a sentence, which Wang et al. (2026a) describe as a spurious cross-modal relative-position bias and propose to remove by placing image tokens equidistant from every text token. If such a bias exists in the tested models, moving a passage of several hundred image tokens from before a typed passage to after it changes not only its order but its distance from the question, and the effect would depend on each model's encoding scheme.

Second, the number of tokens differs by channel. A rendered passage may occupy many more positions than the same passage typed, or fewer, depending on the

vision encoder's resolution handling; the anchor's own preprocessing experiment, which gave Qwen2.5-VL the image tiling used by LLaVA and moved its single-evidence gap on NQ-Swap from +9.71 to -37.23, close to LLaVA's own -34.92 (values from its Table 5, which differ from Table 1, e.g. +9.71 against +3.80 for Qwen2.5-VL-7B, for reasons the anchor does not state), shows that how an image is cut into tokens can by itself reproduce one family's preference in another (Lee, Yoon and Kim, 2026, Table 5). The reverse intervention, giving LLaVA Qwen's resizing, moved its gap only from -34.92 to -30.91, so the effect is not symmetric, and the authors are explicit that preprocessing is one factor rather than the mechanism.

Third, multi-image work finds that open models reason best about images placed late in the input and proprietary models about images at the start or end (Tian et al., 2025), the same split the anchor shows, where the open Qwen and InternVL models move strongly with order and Claude does not.

None of this establishes a mechanism for the anchor's order effect. It does establish that order is not a nuisance variable to be averaged away. The anchor averages over orders in its Table 11 and reports the orders separately in Table 1; a study that ran only one order would have reported a firm preference of whichever sign that order produced. Of the other preference studies reviewed here, only Deng et al. (2025) report a manipulation of modality order, and they tested it on one model.

8. Families replicate; rates do not

Figure 1b compares the anchor with MC2 for the five models, from three families, that both report: Qwen2.5-VL at 7B and 32B, InternVL3 at 38B, and LLaVA-1.6 (LLaVA-NeXT) at 7B and 34B. The two constructions could hardly differ more. MC2 sets a natural photograph against a generated sentence that contradicts it on one attribute, with both sources verified to be easy; the anchor sets a rendered passage against a typed passage, both counterfactual. The author converted MC2's reported vision and text accuracies (its Table 13) to the anchor's scale. The converted values are +19.2 and +35.9 for the two Qwen2.5-VL models, +24.8 for InternVL3-38B, -83.0 for LLaVA-1.6-7B and -49.7 for LLaVA-1.6-34B. The anchor's cross-modal gaps, pooled over order (its Table 11), have the same sign for all five on ConflictQA and for the four that can be compared on NQ-Swap; the Qwen2.5-VL-7B NQ-Swap value (+31.23) is set aside because it averages in the disputed image-first cell of Section 3.5. Its single-evidence gaps have the same sign for all five on both datasets.

The magnitudes do not agree. Qwen2.5-VL-32B is at +26.21 on ConflictQA and +2.61 on NQ-Swap; Qwen2.5-VL-7B is at +2.27 and, taking the disputed cell at its Table 1 value, +31.23; on that reading the ordering of the two Qwen models reverses between datasets. MC2 does not say whether its 7B LLaVA-NeXT checkpoint uses the same language backbone as the anchor's vicuna-7B model, so that pair may not be the same model. With five models from three families and two datasets, sign agreement in nine of nine comparisons is suggestive but not a test; the comparisons are not independent, the families were not sampled at random, and the LLaVA pair contributes four of the nine agreements from strongly negative values.

The same family ordering appears in two other studies. Zhang et al. (2025b) find LLaVA models with a balance point near zero or toward text and Qwen models toward vision, using coloured shapes and CLEVR scenes. Hua et al. (2025) find Qwen2.5-VL the most selective of four models, though their task measures interference rather than preference. The pattern, LLaVA toward text and recent Qwen and InternVL models toward images, holds across rendered passages, natural photographs and synthetic scenes. What does not hold is any rate. This mirrors a finding the author reported for text-only context-memory conflicts, that moderators replicate across constructions while adoption rates do not (Mahendrakar, 2026a), and it suggests the same remedy: report the ordering and the conditions under which it holds, not a rate that will not transfer.

The family pattern also has a partial explanation in the literature. Deng et al. (2025) give a theoretical argument that text preference can arise from an imbalance between text-only and multimodal training data, and report that fine-tuning on text-augmented data lowers it; MC2 finds that fine-tuning with a larger share of conflict-context data shifts preference further in whichever direction the data favour, and that vision preference rises with size within families (Zhang et al., 2025a). Families trained with more multimodal and document data would on this account lean toward images. That is consistent with the data and has not been tested by training matched models, which the anchor names as the experiment it could not run.

9. Images against memory: the other half of the literature

The anchor's single-evidence setting is a context-memory conflict in which the context is an image. A separate literature studies the same conflict with natural images rather than rendered passages, and it too is divided. MMKC-Bench, which presents factual knowledge conflicts with multimodal context, reports that current models tend to favour their internal knowledge over the external evidence (Jia et al., 2025). A benchmark of images that contradict common sense finds over-reliance on parametric knowledge for roughly a fifth of queries (Liu et al., 2025). SegSub, which edits image regions to create conflicts, reports adherence rates of about 20 percent in parametric conflicts, which its authors read as robustness, alongside much weaker performance at detecting counterfactual conditions and resolving

conflicts between sources (Carragher et al., 2025). On the other side, a mechanistic study across three families finds that answering from the image is the default and answering from memory against the image depends on a small set of attention heads, between 2.5 and 4.8 percent of heads, whose ablation flips 68 to 96 percent of memory-based answers to visual ones while changing only 0.8 to 7.5 percent of visual answers (Lietzow et al., 2026). A dataset of counterfactual images built to contradict commonsense knowledge was used to locate heads that mediate the conflict and that can steer the model toward either source (Ortu et al., 2025).

Two further results show that conflicts can arise inside one model across routes to the same fact. Zhu et al. (2024) asked models the same question about an entity through its picture and through its name and found a lower bound on the rate of disagreement of 21.36, 28.10 and 20.53 percent for LLaVA at 7B, 13B and 34B, roughly constant with scale. Shao et al. (2024) found that GPT-4o's answers about a document agree with its own reading of the document in 75.26 percent of cases. In both, the model's two routes to the same content disagree before any external conflict is introduced, which means that a measured preference in an external-conflict experiment partly reflects which route the model would have used anyway.

The relevance to the anchor is that the rendered-passage result does not slot into a settled picture of image-versus-memory behaviour. Natural-image studies report both memory-dominant and image-dominant behaviour, and the mechanistic study suggests why: with a natural image, perception produces an answer early and memory must actively override it. With a rendered passage, the image contains a sentence rather than a scene, and the model must read it before it can answer from it; whether that reading engages the visual default or the textual pathway is exactly what the anchor's channel comparison probes, and the answer appears to depend on the family. The construction argument of the author's earlier record applies here too (Mahendrakar, 2026a): adoption rates against memory depend on how the conflicting evidence is built, and a comparison of image-memory rates across these studies would compare constructions.

10. A reliability-matched protocol

The preceding sections imply that a claim of modality bias needs more than a following ratio. Algorithm 1 states a protocol that would support such a claim, assembled from designs that individual studies have already used: the anchor's memory filter and counterbalanced orders, MC2's unimodal verification and option swapping, van Sprang et al.'s transcription check, and Zhang et al.'s graded difficulty and balance point. What is new is putting them in one design and adding the factorial over channel and content.

Steps 1 and 2 follow the anchor and the render-equivalence literature and remove the two confounds that cross-study comparisons most often mix in: a conflict with memory that the model does not hold, and a rendered input it cannot read. Step 3 is the step most studies omit; without it, steps 7 and 8 cannot be done. Step 4 is the factorial that Section 5 found missing: no published study crosses channel with content on the same items. Step 5 keeps the outcome Section 3.3 found hidden. Step 6 asks for the intervals the anchor does not report, at sample sizes it already has. Step 8 states the decision rule. Its effect is to make "bias" mean a departure from a reliability-matched baseline, as in perception research, rather than a majority share. Step 9 turns the conjectures in Sections 5.3 and 6.3 into tests: if rendering weakens instruction-treatment or shortens deliberation, the role manipulation and the reasoning-first prompt should move the gap in predictable directions.

The protocol is costly in items, since the factorial multiplies conditions and the balance-point fit needs items spread across relative difficulty. A minimal version, adequate to decide whether the anchor's reversal and the text-dominance results conflict, needs only the channel-by-content cells on a shared item set, one memory filter, both orders, and absolute rates with intervals.

11. Consequences

11.1 For multimodal retrieval

The practical case for caring about channel effects is retrieval over documents stored as page images (Faysse et al., 2025; Yu et al., 2025; Cho et al., 2024), and long-context compression through rendering (Cheng et al., 2025; Wei, Sun and Li, 2025; Li, Lan and Zhou, 2025). The anchor reports an encouraging side effect: when a gold passage is rendered and placed among nine typed distractors, answer quality improves, with the clearest benefit reported for the middle position, where one model gained 6.59 ROUGE-L points (Lee, Yoon and Kim, 2026, Section 5.1). The same property that makes a rendered gold passage easier to find among distractors would make a rendered counterfeit passage harder to ignore, and the anchor's own single-evidence result shows that rendered counter-memory evidence is adopted somewhat more often. A retrieval pipeline that mixes typed and rendered sources has therefore introduced a source-weighting decision that nobody chose, and its direction depends on the model family and on the order in which retrieved items are concatenated. Order should be randomised or fixed by design, and the pipeline should be tested with conflicts in both channels.

11.2 For safety

The safety consequence is the clearest. Rendering an instruction removes part of the protection that safety training provides, whether by the uneven coverage

Algorithm 1. Reliability-matched estimate of modality preference under conflict

```
Input: model M; items with a fact, a counter-fact and a query; channels {typed, rendered};
       content types {proposition, scene}; roles {evidence, instruction}
Step 1: Memory: ask M each query closed-book k times; keep items with a unanimous answer a_mem.
Step 2: Recoverability: for every rendered input, require M's transcription to match the source;
        for every scene, require M's unimodal answer to be stable across k samples.
Step 3: Unimodal reliability: for each source s on each item, record M's answer and entropy H(s)
        with s shown alone, and its accuracy against the counter-fact the source asserts.
Step 4: Factorial: build cells channel x content (Table 2); within each cell present
        (a) one source vs memory, (b) two conflicting sources, in both orders and both
        assignments of content to channel; add same-channel and same-content controls.
Step 5: Score per item: follow_A, follow_B, follow_memory, other; never drop "neither".
Step 6: Report absolute rates with intervals (bootstrap over items) for every cell and order;
        report normalized gaps only alongside the rates they are computed from.
Step 7: Fit P(follow A) against relative uncertainty (H(A) - H(B)) / mean(H(A), H(B));
        report the balance point and its interval per cell.
Step 8: Call the model biased toward channel c only if the balance point differs from zero
        beyond the same-channel control, with content held fixed; likewise for content type.
Step 9: Repeat with the role changed (evidence vs instruction) and with a reasoning-first prompt;
        report whether the gap shrinks when the model must reason before answering.
Output: per-cell rates, order effects, balance points, and which claims they license.
```

FigStep proposes (Gong et al., 2025), by the shrinkage of the refusal vector in middle layers reported for omni-modal models (Wang et al., 2026c), or by a mechanism not yet identified. The anchor's 6.99-point mean rise over the four MM-SafetyBench baselines (Text-only, SD, Typo and SD + Typo; +11.60 against Text-only, +4.97 against SD) adds that moving the whole instruction into the image is worse than embedding only the key phrase (Lee, Yoon and Kim, 2026). The author's earlier analysis of degraded renderings found that how far an image is degraded changes the sign of the effect across studies (Mahendrakar, 2026c), which means that a single rendering choice cannot certify a model as safe against this channel. The safety evaluation implied by Algorithm 1 is the role manipulation in step 9, run with harmful instructions, across renderings.

11.3 For mitigation

The anchor tested four mitigations and found that a system prompt asking for balanced use of both modalities was unreliable, that preference optimization and activation steering moved the gap in one order while worsening it in the other, and that conflict-aware supervised fine-tuning moved every gap toward zero, for example from +73.82 to +52.04 in the text-first order on ConflictQA, while slightly increasing adoption of reliable rendered evidence (Lee, Yoon and Kim, 2026, Tables 3, 4 and 12). Its distinction between controlling a preference and making it robust is the right one. Steering methods for modality preference are explicitly designed for control: MC2 steers representations toward a desired modality and reports gains on downstream tasks (Zhang et al., 2025a), and an instance-aware variant scales the steering per item to limit damage (Huang et al., 2026). Deng et al. (2025) report that fine-tuning with text augmentation raised accuracy under corrupted text from 28.69 to 71.25 percent.

The reliability framing changes what a successful mitigation should look like. The target is not a gap of zero; it is following that tracks reliability. In the anchor's single-evidence setting, where reliability is equal by construction, the target is indeed zero. In a setting like Deng et al.'s, where the image is the primary source, the target is a strong image preference on corrupted descriptions and a strong text preference where the description is right and the visual reading wrong. A mitigation evaluated only by how close it brings a pooled gap to zero can therefore be optimizing the wrong objective. The anchor's own caution, that a gap near zero could arise from following both sources less, is a special case of this.

11.4 For evaluation practice

Three reporting changes would make modality-preference results comparable. State which of the four quantities in Section 4 is measured. State the cell of Table 2: channel, content type, role, and reference point. Report absolute following rates for each source and for neither, by order, with intervals. The anchor reports most of this, which is what made the re-reading in Section 3 possible; most earlier studies report a single ratio in a single order.

12. What is not known

Whether text dominance is a preference for the token channel or for propositional content has not been tested on a single item set; the reading in Section 5 is an inference across studies. Whether the anchor's small single-evidence gaps for open models are distinguishable from zero is not known, because no intervals are reported. Why rendering raises adoption against memory, when rendered text is read almost perfectly and usually reasoned over worse, is not known; Section 6.3 offers a conjecture with two testable predictions. Whether rendering weakens the treatment of text as an instruction, which would link the safety, question-control and adoption results, is not known. Whether the family ordering reflects training data mix, image preprocessing, positional encoding or something else has not been tested with matched training, which is the

experiment the anchor names as its own next step. Whether the order effect's direction differs between model families because of chat-template conventions is not known. And for audio-visual models, where a 2026 study found answers committed late in the network to an internally preferred pattern under conflict (Sudheer et al., 2026), whether the channel-content distinction transfers is not known; the anchor's limitations section says the same of audio.

13. Conclusion

The claim that multimodal models trust text over images was measured mostly with typed propositions against perceived scenes. The 2026 result that rendered passages are adopted more readily than typed ones was measured with one proposition delivered through two channels. The two results answer different questions, and the typographic-attack literature, in which written words override pictures from inside the image, shows how to join them: the dominance documented so far attaches to propositional content more than to the token channel. Within the anchor's own data, the multi-evidence preference is predictable from order and family rather than arbitrary, the image-first order reproduces text dominance for most open models, most single-evidence gaps are no larger in magnitude than the study's controls, although the controls lean negative and on a signed comparison several gaps exceed them, and its strongest models are not counted as following either counterfactual passage on most items. Across studies, the direction of each family's preference replicates and its size does not. A preference measured without the reliability of each source is not yet a bias, and the protocol in Algorithm 1 states what would make it one.

References

Agrawal, A., Batra, D., Parikh, D., & Kembhavi, A. (2018). Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. CVPR 2018. arXiv:1712.00377. doi:10.48550/arXiv.1712.00377.

Alais, D., & Burr, D. (2004). The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. Current Biology, 14(3), 257-262. doi:10.1016/j.cub.2004.01.029.

Carragher, P., Rao, N., Jha, A., Raghav, R., & Carley, K. M. (2025). SegSub: Evaluating Robustness to Knowledge Conflicts and Hallucinations in Vision-Language Models. arXiv:2502.14908. doi:10.48550/arXiv.2502.14908. [PREPRINT]

Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., & Chang, B. (2024). An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. ECCV 2024. arXiv:2403.06764. doi:10.48550/arXiv.2403.06764.

Chen, L., Huang, M., Chen, Y., Lin, J., & Shi, Y. (2026). Majorization-Guided Test-Time Adaptation for Vision-Language Models under Modality-Specific Shift. ACM MM 2026. arXiv:2604.24602. doi:10.48550/arXiv.2604.24602.

Cheng, H., Xiao, E., Gu, J., Yang, L., Duan, J., Zhang, J., Cao, J., Xu, K., & Xu, R. (2024). Unveiling Typographic Deceptions: Insights of the Typographic Vulnerability in Large Vision-Language Model. ECCV 2024. arXiv:2402.19150. doi:10.48550/arXiv.2402.19150.

Cheng, J., Liu, Y., Zhang, X., Fei, Y., Hong, W., Lyu, R., Wang, W., Su, Z., Gu, X., Liu, X., Bai, Y., Tang, J., Wang, H., & Huang, M. (2025). Glyph: Scaling Context Windows via Visual-Text Compression. arXiv:2510.17800. doi:10.48550/arXiv.2510.17800. [PREPRINT]

Cho, J., Mahata, D., Irsoy, O., He, Y., & Bansal, M. (2024). M3DocRAG: Multi-modal Retrieval is What You Need for Multi-page Multi-document Understanding. arXiv:2411.04952. doi:10.48550/arXiv.2411.04952. [PREPRINT]

Deng, A., Cao, T., Chen, Z., & Hooi, B. (2025). Words or Vision: Do Vision-Language Models Have Blind Faith in Text? CVPR 2025. arXiv:2503.02199. doi:10.48550/arXiv.2503.02199.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415, 429-433. doi:10.1038/415429a.

Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., & Colombo, P. (2025). ColPali: Efficient Document Retrieval with Vision Language Models. ICLR 2025. arXiv:2407.01449. doi:10.48550/arXiv.2407.01449.

Gong, Y., Ran, D., Liu, J., Wang, C., Cong, T., Wang, A., Duan, S., & Wang, X. (2025). FigStep: Jailbreaking Large Vision-Language Models via Typographic Visual Prompts. AAAI 2025. arXiv:2311.05608. doi:10.48550/arXiv.2311.05608.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR 2017. arXiv:1612.00837. doi:10.48550/arXiv.1612.00837.

Guo, X., & Vosoughi, S. (2024). Serial Position Effects of Large Language Models. arXiv:2406.15981. doi:10.48550/arXiv.2406.15981. [PREPRINT]

Hua, T., Yun, T., & Pavlick, E. (2025). How Do Vision-Language Models Process Conflicting Information Across Modalities? arXiv:2507.01790. doi:10.48550/arXiv.2507.01790. [PREPRINT]

Huang, Z., Li, X., Surana, R., Yu, T., Wang, R., McAuley, J., Shang, J., & Wu, J. (2026). AMPS: Adaptive Modality Preference Steering via Functional Entropy. arXiv:2602.12533. doi:10.48550/arXiv.2602.12533. [PREPRINT]

Jia, Y., Jiang, K., Liang, Y., Ren, Q., Xin, Y., Yang, R., Feng, F., Chen, M., Lu, H., Wang, H., Qu, X., Liu, D., Cui, L., & Du, Y. (2025). Benchmarking Multimodal Knowledge Conflict for Large Multimodal Models. arXiv:2505.19509. doi:10.48550/arXiv.2505.19509.

Jin, Z., Cao, P., Chen, Y., Liu, K., Jiang, X., Xu, J., Li, Q., & Zhao, J. (2024). Tug-of-War Between Knowledge: Exploring and Resolving Knowledge Conflicts in Retrieval-Augmented Language Models. LREC-COLING 2024. arXiv:2402.14409. doi:10.48550/arXiv.2402.14409.

Kording, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. PLoS ONE, 2(9), e943. doi:10.1371/journal.pone.0000943.

Lee, J., Yoon, Y., & Kim, T. (2026). Same Semantics, Different Outcome: On the Modality Robustness of Multimodal LLMs under Knowledge Conflict. Findings of EMNLP 2026 (accepted). arXiv:2609.00550. doi:10.48550/arXiv.2609.00550.

Lee, K., Kim, M., Yoon, S., Kim, M., Lee, D., Koh, H., & Jung, K. (2025). VLind-Bench: Measuring Language Priors in Large Vision-Language Models. Findings of NAACL 2025. arXiv:2406.08702. doi:10.48550/arXiv.2406.08702.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., & Bing, L. (2024). Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. arXiv:2311.16922. doi:10.48550/arXiv.2311.16922.

Li, X., Lu, Y., Gan, Z., Gao, J., Wang, W. Y., & Choi, Y. (2023). Text as Images: Can Multimodal Large Language Models Follow Printed Instructions in Pixels? arXiv:2311.17647. doi:10.48550/arXiv.2311.17647. [PREPRINT]

Li, Y., Lan, Z., & Zhou, J. (2025). Text or Pixels? It Takes Half: On the Token Efficiency of Visual Text Inputs in Multimodal LLMs. Findings of EMNLP 2025. arXiv:2510.18279. doi:10.48550/arXiv.2510.18279.

Lietzow, N., Bitterman, D., Eickhoff, C., Rudman, W., & Golovanevsky, M. (2026). Vision-Default, Prior-Override: Causal Mechanisms of Perception-Knowledge Conflict in Vision-Language Models. arXiv:2606.28273. doi:10.48550/arXiv.2606.28273. [PREPRINT]

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024a). Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics. arXiv:2307.03172. doi:10.48550/arXiv.2307.03172.

Liu, Q., Feng, J., Wang, Y., Han, X., Cheng, Y., Zhu, Y., Diao, H., Zhuge, Y., & Lu, H. (2026). VISTA-Bench: Do Vision-Language Models Really Understand Visualized Text as Well as Pure Text? arXiv:2602.04802. doi:10.48550/arXiv.2602.04802. [PREPRINT]

Liu, X., Wang, W., Yuan, Y., Huang, J., Liu, Q., He, P., & Tu, Z. (2025). Insight Over Sight: Exploring the Vision-Knowledge Conflicts in Multimodal LLMs. ACL 2025. arXiv:2410.08145. doi:10.48550/arXiv.2410.08145.

Liu, X., Zhu, Y., Gu, J., Lan, Y., Yang, C., & Qiao, Y. (2024b). MM-SafetyBench: A Benchmark for Safety Evaluation of Multimodal Large Language Models. ECCV 2024. arXiv:2311.17600. doi:10.48550/arXiv.2311.17600.

Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., & Singh, S. (2021). Entity-Based Knowledge Conflicts in Question Answering. EMNLP 2021. arXiv:2109.05052. doi:10.48550/arXiv.2109.05052.

Mahendrakar, P. (2026a). A Conflict Is Constructed Before It Is Measured: Seven Design Choices That Set Which Way a Language Model Bends, and Why the Context-Memory Results Do Not Compare. Life of Research. doi:10.5281/zenodo.22239743.

Mahendrakar, P. (2026b). A Blind Score Is Not a Reliance Measure: What Answering Without the Audio Shows About Audio-Language Benchmarks, Why the Partial-Input Debate Already Settled Half of It, and the Counterfactual-Audio Test the Other Half Needs. Life of Research. doi:10.5281/zenodo.23381527.

Mahendrakar, P. (2026c). Hard to Read Is Not Hard to Audit: Why Degrading a Text Image Raises Jailbreak Success in One 2026 Study and Lowers It in Another, Four Accounts That Fit the Same Curve, and the Ablations That Would Separate Them. Life of Research. doi:10.5281/zenodo.23427759.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. ACL 2023. arXiv:2212.10511. doi:10.48550/arXiv.2212.10511.

Ngo, M. K., Cadieux, M. L., Sinnett, S., Soto-Faraco, S., & Spence, C. (2011). Reversing the Colavita visual dominance effect. Experimental Brain Research, 214, 607-618. doi:10.1007/s00221-011-2859-9.

Ortu, F., Jin, Z., Doimo, D., & Cazzaniga, A. (2025). When Seeing Overrides Knowing: Disentangling Knowledge Conflicts in Vision-Language Models. arXiv:2507.13868. doi:10.48550/arXiv.2507.13868.

Pezeshkpour, P., & Hruschka, E. (2023). Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. arXiv:2308.11483. doi:10.48550/arXiv.2308.11483. [PREPRINT]

Qraitem, M., Tasnim, N., Teterwak, P., Saenko, K., & Plummer, B. A. (2024). Vision-LLMs Can Fool Themselves with Self-Generated Typographic Attacks. arXiv:2402.00626. doi:10.48550/arXiv.2402.00626. [PREPRINT]

Shao, Z., Gao, F., Zhu, Z., Luo, C., Xing, H., Yu, Z., Zheng, Q., Yan, M., & Bu, J. (2024). Is Cognition Consistent with Perception? Assessing and Mitigating Multimodal Knowledge Conflicts in Document Understanding. arXiv:2411.07722. doi:10.48550/arXiv.2411.07722.

Sudheer, A., Li, D., Elbanna, O., Kodarapu, I., Bahuguna, A., & Sharma, V. (2026). Compositional Failure in Audio-Visual LLMs: Late-Layer Prior Dominance Under Cross-modal Conflict. arXiv:2608.27785. doi:10.48550/arXiv.2608.27785. [PREPRINT]

Sun, K., Yuan, X., Liu, H., Zhao, C., Zhang, C., Dredze, M., & Bai, F. (2026). Reading, Not Thinking: Understanding and Bridging the Modality Gap When Text Becomes Pixels in Multimodal LLMs. arXiv:2603.09095. doi:10.48550/arXiv.2603.09095. [PREPRINT]

Tian, X., Zou, S., Yang, Z., & Zhang, J. (2025). Identifying and Mitigating Position Bias of Multi-image Vision-Language Models. CVPR 2025. arXiv:2503.13792. doi:10.48550/arXiv.2503.13792.

van Sprang, A., Samson, L., Lucic, A., Acar, E., Ghebreab, S., & Asano, Y. M. (2026). Same Content, Different Answers: Cross-Modal Inconsistency in MLLMs. CVPR 2026. arXiv:2512.08923. doi:10.48550/arXiv.2512.08923.

Wang, C., Guo, J., Li, H., Tian, Y., Nie, Y., Xu, C., & Han, K. (2026a). Circle-RoPE: Cone-like Decoupled Rotary Positional Embedding for Large Vision-Language Models. ICML 2026. arXiv:2505.16416. doi:10.48550/arXiv.2505.16416.

Wang, Y., Zhou, R., Tang, Y., Zhu, Y., Zhao, X., Chang, X., & Liang, X. (2026b). When Prompts Become Pixels: Prompt-Region Grounding for Multimodal Reasoning. arXiv:2608.04726. doi:10.48550/arXiv.2608.04726. [PREPRINT]

Wang, K., Li, Z., Zhou, Z., Zhang, Y., Mi, Y., Yang, K., Zhang, Y., Dong, J., Sun, Z., Li, Q., & Liu, Y. (2026c). Omni-Safety under Cross-Modality Conflict: Vulnerabilities, Dynamics Mechanisms and Efficient Alignment. arXiv:2602.10161. doi:10.48550/arXiv.2602.10161. [PREPRINT]

Wang, Y., Feng, S., Wang, H., Shi, W., Balachandran, V., He, T., & Tsvetkov, Y. (2024). Resolving Knowledge Conflicts in Large Language Models. COLM 2024. arXiv:2310.00935. doi:10.48550/arXiv.2310.00935.

Wei, H., Sun, Y., & Li, Y. (2025). DeepSeek-OCR: Contexts Optical Compression. arXiv:2510.18234. doi:10.48550/arXiv.2510.18234. [PREPRINT]

Westerhoff, J., Purelku, E., Hackstein, J., Loos, J., Pinetzki, L., Rodner, E., & Hufe, L. (2025). SCAM: A Real-World Typographic Robustness Evaluation for Multimodal Foundation Models. CVPR 2025 Workshops. arXiv:2504.04893. doi:10.48550/arXiv.2504.04893.

Wu, H., Tang, M., Zheng, X., & Jiang, H. (2025). When Language Overrules: Revealing Text Dominance in Multimodal Large Language Models. arXiv:2508.10552. doi:10.48550/arXiv.2508.10552. [PREPRINT]

Xie, J., Zhang, K., Chen, J., Lou, R., & Su, Y. (2024). Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. ICLR 2024. arXiv:2305.13300. doi:10.48550/arXiv.2305.13300.

Yu, S., Tang, C., Xu, B., Cui, J., Ran, J., Yan, Y., Liu, Z., Wang, S., Han, X., Liu, Z., & Sun, M. (2025). VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents. ICLR 2025. arXiv:2410.10594. doi:10.48550/arXiv.2410.10594.

Zhang, Y., Ma, J., Hou, Y., Bai, X., Chen, K., Xiang, Y., Yu, J., & Zhang, M. (2025a). Evaluating and Steering Modality Preferences in Multimodal Large Language Model. arXiv:2505.20977. doi:10.48550/arXiv.2505.20977.

Zhang, Z., Wang, T., Gong, X., Shi, Y., Wang, H., Wang, D., & Hu, L. (2025b). When Modalities Conflict: How Unimodal Reasoning Uncertainty Governs Preference Dynamics in MLLMs. arXiv:2511.02243. doi:10.48550/arXiv.2511.02243. [PREPRINT]

Zhu, T., Liu, Q., Wang, F., Tu, Z., & Chen, M. (2024). Unraveling Cross-Modality Knowledge Conflicts in Large Vision-Language Models. arXiv:2410.03659. doi:10.48550/arXiv.2410.03659.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.23919442