VIGOR: VISUAL INTERVENTION-GROUNDED EVALUATION FOR OBSERVATION AND REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual reasoning benchmarks are read as evidence that a vision-language model (VLM) answers from the image, yet final-answer accuracy cannot show whether the image was used. Intervening on the image and reading the change in accuracy is the usual remedy, but an accuracy change can reflect a shift in the model's response prior as well as the removal of task-relevant content. We ask when such a delta supports a claim that the intended evidence was used. We introduce VIGOR, an audit that evaluates each item under five views and summarizes them as a validity profile, and VIGOR-Stress300, an evidence-annotated subset of MMStar, BLINK-Twice, and MMPerspective. Across open VLMs from three families, masking the annotated evidence lowers accuracy on MMStar and raises it on BLINK-Twice; read on its own, the gain would be taken as improved use of the image. Two requirements make the deltas interpretable. Under *score legibility*, a panel of option structure, label marginal, and answers without evidence shows that the BLINK-Twice gain is a response shift toward what the models answer for an occluded image, that a text-only polarity rule outscores every masked view, and that original-view predictions show no positive chance-adjusted agreement with the labels. Under *matched intervention*, MMStar's degradation survives label adjustment, six removal styles, location- and magnification-matched controls, and boxes drawn by independent auditors. On MMPerspective the tested interventions leave a small, model-dependent effect unresolved. We recommend that intervention-based audits report the legibility panel and the matched controls with every delta.

1 INTRODUCTION

A visual reasoning benchmark reports how often a vision-language model (VLM) selects the correct option. The claim attached to that number is stronger: that the model used the image, located the evidence the question is about, and answered from it. Accuracy does not license this inference. Models answer from question text and answer priors (Goyal et al., 2017; Agrawal et al., 2018), from benchmark regularities (Chen et al., 2024; Li et al., 2024), and from language priors that override what is shown (Guan et al., 2024; Lee et al., 2025; Luo et al., 2025). Benchmarks that target visual indispensability (MMStar (Chen et al., 2024), BLINK-Twice (Ye et al., 2025), and MMPerspective (Tang et al., 2025)) curate items so that the image should matter, but curation is a claim about items, not a measurement of how a given model answered them; Figure 1 outlines the audit that measures it.

Intervening on the image is the obvious next step. If a model answers from the evidence, removing that evidence should cost accuracy, whereas removing an equal amount of unrelated content should not; MMStar's multimodal gain takes the difference between the image and text-only conditions at benchmark level (Chen et al., 2024), and recent audits degrade the whole image or occlude regions and read the resulting drop (Lan et al., 2026). The delta these audits report is, however, a score of the same kind as the accuracy it was meant to explain. A difference of two accuracies depends on the label marginal of the items; a masked image differs from the original in more than the content it removes; a crop changes scale and context; and a control that is supposed to be matched may not be realizable on the items at hand. Whether a delta supports a claim that the intended evidence was used is therefore a question of its own, and the audits that report deltas rarely test them against a constant-predictor reference or a location-matched control.

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.

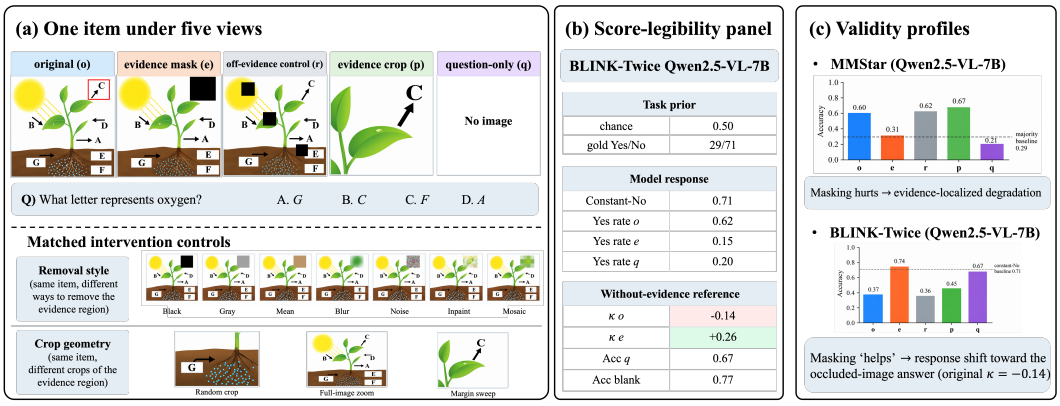


Figure 1: **The VIGOR audit.** (a) One item under the five views, with removal styles and crop geometries as controls. (b) The score-legibility panel, reported before any delta is interpreted. (c) The same intervention reads as evidence-localized degradation on MMStar and as a response shift toward the occluded-image answer on BLINK-Twice once the panel is consulted (Qwen2.5-VL-7B).

We ask that question here. VIGOR (Visual Intervention-Grounded evaluation for Observation and Reasoning) evaluates every benchmark item under five views (the original image, an evidence mask, an off-evidence control, an evidence crop, and question-only) and summarizes the item's behavior across views as a validity profile. We instantiate it on VIGOR-Stress300, 300 evidence-annotated items from MMStar, BLINK-Twice, and MMPerspective, and run four open VLMs through the full protocol under one prompt and one decoding setting. Masking the annotated evidence lowers MMStar accuracy for every model, raises BLINK-Twice accuracy for every model, and barely moves MMPerspective. Read on their own, these deltas certify MMStar as evidence-sensitive and the BLINK-Twice gain as improved use of the image.

The BLINK-Twice gain does not establish more faithful use of the intended evidence, and the response shift that produces it is the subject of this paper. A delta can be taken to support four claims, task accuracy, dependence on the image, dependence on the annotated region, and faithful use of the intended evidence, and interpreting a profile requires two things. Under *score legibility*, option structure, gold marginal, predicted-label distribution, and the model's answers without evidence are reported before any delta is interpreted; under *matched intervention*, the control is matched to the evidence intervention on location, style, and geometry, and its feasibility is checked on every item. Applied to the three sources, the two requirements leave one of the three readings unchanged, replace one, and leave the third unresolved.

Our contributions are:

- Evidence that intervention deltas, the usual evidence of visual-grounding audits, support a conclusion about evidence use on a public benchmark that the response distribution does not, across model families and scales, together with two requirements that separate the readings.
- VIGOR, an audit protocol that adds a score-legibility panel and three-axis matched controls to the intervention profile.
- VIGOR-Stress300, an evidence-annotated subset with a blind two-auditor re-annotation and a full-benchmark legibility panel for BLINK-Twice, and a cross-model audit that yields a corrected diagnosis of three visual reasoning benchmarks.

2 Related Work

Shortcuts, priors, and intervention-based audits. That accuracy need not reflect image use is established for VQA and for VLMs. VQA-CP shifts the answer distribution between train and test so that answer priors stop paying off (Agrawal et al., 2018); balanced VQA and NaturalBench pair each question with images that carry opposite answers, so that a blind solution scores at chance

(Goyal et al., 2017; Li et al., 2024); VLind-Bench and visual-prior probes measure how far a model follows a language prior against the image (Lee et al., 2025; Luo et al., 2025). HallusionBench pairs original images with edited variants (Guan et al., 2024), a design that BLINK-Twice's counterparts share and that we use as a control (§4.3). Interventions on the image itself take two forms. MMStar's multimodal gain subtracts a text-only condition from the image condition at benchmark level (Chen et al., 2024); Lan et al. (2026) degrade the whole image, occlude regions, rephrase questions, and inspect decision margins and no-image responses, and find predictions less sensitive to these changes than accuracy implies. Both read the resulting delta as a measure of visual dependence. To our knowledge, neither pairs the delta with a chance-referenced agreement statistic or with a control matched to the intervention's location, which is the gap the two requirements address. A separate line isolates capabilities through controlled inputs (Hou et al., 2025; Wang et al., 2026; Xu et al., 2026; 2025); VIGOR shares the move from a score to a controlled contrast and applies it to the evidence that existing benchmarks intend.

Measurement validity, confounds, and perturbation. Construct-validity theory asks whether a score justifies the interpretation placed on it, and measurement work in machine learning audits the gap between a construct and its operationalization (Cronbach & Meehl, 1955; Messick, 1995; Jacobs & Wallach, 2021). Dominguez-Olmedo et al. (2025) show that adjusting a benchmark for a confound changes conclusions about emergence; skill-level slicing shows that aggregate accuracy hides heterogeneous behavior (Moayeri et al., 2025); circular evaluation in MMBench removes option-position effects (Liu et al., 2024). The legibility panel plays the role of such an adjustment for intervention deltas. Attribution methods locate the regions a prediction depends on (Selvaraju et al., 2017; Petsiuk et al., 2018), and removal-based evaluations of attribution show that the removal itself carries information unless the masks are matched (Hooker et al., 2019; Rong et al., 2022); corruption and paired compositional tests show that aggregate accuracy hides perturbation sensitivity and hackability (Hendrycks & Dietterich, 2019; Ribeiro et al., 2020; Thrush et al., 2022; Hsieh et al., 2023). The matched controls are the removal baseline that an evidence intervention needs.

3 METHODS

3.1 EVALUATION VIEWS AND CONTROLS

Let item i consist of an image x_i, a question with answer options, and a gold option y_i. A model f returns a prediction $\hat{y}_i^c$ under view c, with correctness $z_i^c = \mathbb{1}[\hat{y}_i^c = y_i]$ scored against the original gold label under every view; accuracy under view c is $\mathrm{Acc}^c = \frac{1}{n}\sum_i z_i^c$, and Acc^o on the original view o is the benchmark score. The claim under audit is that f's correct answers are grounded in the visual evidence the benchmark intends. We operationalize the intended evidence with an evidence box b_i drawn on x_i (§3.3) and evaluate each item under four further views (Figure 1a; Table A1 in Appendix A): the *evidence mask* e replaces the content inside b_i; the *off-evidence control* r replaces pixels outside b_i, with zero overlap, over rectangles in the free bands around the box, with the same pixel budget as e whenever the box covers at most half the frame and the largest zero-overlap budget otherwise; the *evidence crop* p retains b_i with a 15% margin and passes it through the model's input pipeline; and *question-only* q removes the image. A *blank* view, the image replaced by a black canvas of the same size, serves with q as a reference for the model's answers without evidence. Every view is a deterministic function of the image and b_i; no saliency map or model internal is used, so the same views apply to any model.

The evidence views are realized under controls of three kinds, with a fourth that varies the annotation. The location control is r itself: an area-matched, non-overlapping removal placed outside the box. The style control realizes e with seven fills, six that remove the content (black, gray, mean color, Gaussian blur, random noise, inpainting) and one that retains its coarse structure (mosaic). The geometry control compares p with a random-location crop of the same size, a full-image zoom of matched magnification, and a sweep over the crop margin. The annotation control regenerates e, r, and p from boxes drawn by independent auditors (§3.3). If f answers from b_i, then z^e should fall below z^o on more items than z^r does, and z^p should recover what z^e loses; the audit asks when these predictions, once observed, can be read as confirmation (§3.2).

3.2 Profile Metrics and Legibility Panel

The validity profile of a model on a source collects the accuracies under the views and the contrasts between them. Evidence sensitivity $\mathrm{ES}_e = \mathrm{Acc}^o - \mathrm{Acc}^e$ is positive when masking hurts; the off-evidence sensitivity $\mathrm{ES}_r = \mathrm{Acc}^o - \mathrm{Acc}^r$ is its location-matched counterpart, and their difference, the specificity gap $\Delta_{\text{gap}} = \mathrm{ES}_e - \mathrm{ES}_r = \mathrm{Acc}^r - \mathrm{Acc}^e$, measures degradation beyond occlusion of the same amount elsewhere. Crop recovery $\mathrm{Acc}^p - \mathrm{Acc}^e$ measures how much the annotated region restores under the crop operation. Harmful and beneficial flips, $\mathrm{HF}_e = \frac{1}{n}\sum_i \mathbb{1}[z_i^o = 1, z_i^e = 0]$ and $\mathrm{BF}_e = \frac{1}{n}\sum_i \mathbb{1}[z_i^o = 0, z_i^e = 1]$, decompose $\mathrm{ES}_e = \mathrm{HF}_e - \mathrm{BF}_e$ into item-level directions. We call the profile computed from o, e, and p alone, as an audit would first compute it, the *unadjusted profile*.

Each contrast above is a difference of accuracies and depends on the label marginal of the items: a shift of predictions toward a frequent label raises accuracy without any use of evidence. We therefore report, alongside each accuracy, Cohen's κ between predictions and gold under each view (Cohen, 1960), the label-adjusted effect $\Delta\kappa = \kappa^o - \kappa^e$, and balanced accuracy. Any gold-independent strategy (a constant answer, an option-position preference, sampling from a fixed marginal) has expected $\kappa = 0$. The *score-legibility panel* collects what must be known before any delta is interpreted: option cardinality and chance level, gold marginal and majority baseline, the predicted-label distribution under each view, balanced accuracy, κ, the image-minus-question-only difference $\mathrm{Acc}^o - \mathrm{Acc}^q$ (MMStar's multimodal gain), and the model's answers without evidence in the question-only and blank views (definitions in Appendix A).

Two requirements. The audit interprets a delta only after two requirements are met; they are diagnostic requirements, and meeting them separates readings without certifying one. Under *score legibility*, the panel above is reported before any delta is interpreted, so that a shift toward a frequent label, or toward the model's answer without evidence, can be told from a change in evidence use. Under *matched intervention*, the evidence intervention is compared with an off-evidence control of zero overlap and matched budget, with content-removing edits of different style and, for crops, with a control of matched magnification, and the items on which each control is realizable are reported. Section 4 reports the profile first as an audit would compute it (§4.2) and then under each requirement (§4.3, §4.4).

3.3 Evidence Box Annotation

Each item carries one evidence box b_i in pixel coordinates, drawn by a human annotator. A single box is simple and reproducible; it is also an assumption, and the audit below tests it. Primary boxes for MMStar and MMPerspective were fixed before any prediction was generated; the BLINK-Twice boxes were drawn on the base images after the original and question-only views of those images had been run and before any masked, control, or crop view.

Two auditors who had not seen the primary boxes re-annotated a category-stratified sample of 50 items per source for MMStar and MMPerspective, and all 100 BLINK-Twice items, from the image, question, and gold answer alone (Appendix K). Mean IoU with the primary box is 0.59 / 0.69 / 0.58 for MMStar / BLINK-Twice / MMPerspective, and 79% of auditor boxes reach IoU ≥ 0.5 (71% / 84% / 78%). On MMStar the auditors agree with each other (0.73) more closely than with the primary box, so part of the gap is attributable to the primary boxes; on BLINK-Twice the two agreements coincide (0.68); on MMPerspective both auditors marked about half of the items as not localizable with one box. The auditors' boxes also serve as the annotation control (§4.4). A separate pass over the MMStar and MMPerspective items marks whether the evidence needed for the answer is one region, several disjoint regions, or the global layout: 14 / 8 items need several regions and 10 / 86 need the global layout, so a single box is, by the annotators' own judgment, a weak operationalization for MMPerspective.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks. VIGOR-Stress300 contains 300 items, 100 unique-image items from each of MMStar (Chen et al., 2024), BLINK-Twice (Ye et al., 2025), and MMPerspective (Tang et al., 2025); gold labels and question text match the upstream files for all 300 items (record fields in Appendix K). BLINK-Twice releases base images, to which its questions and labels refer, and for 103 of them an edited counterpart that upstream describes as visually similar with opposite factual content. We evaluate the base images and use the counterparts as a control (§4.3; Appendix L), so the 100 BLINK-Twice items are drawn, one question per image, from the 207 questions whose image has a counterpart. Three properties of the subset matter for its scores: BLINK-Twice items are binary Yes/No, with a gold marginal of 29 Yes / 71 No against 334 / 359 in the full source, and 19 of the 100 questions are negated, 18 of them with gold Yes; MMPerspective has four options with a 0.45 majority option and one option set in 71 items; MMStar has four options with a 0.29 majority label. The subset is a stress profile and does not estimate full-benchmark accuracy.

Models. Four open VLMs receive the full protocol: Qwen2.5-VL-3B and -7B (Bai et al., 2025), LLaVA-OneVision-7B (Li et al., 2025), and Idefics3-8B-Llama3 (Laurençon et al., 2024), three model families; Qwen2.5-VL-32B is run on the core views as a scale check (Appendix J).

Prompt and decoding. Every model and view uses one prompt that asks for a single option letter, greedy decoding, and a fixed extractor; an output with no valid letter (at most 1% of items in any cell) counts as incorrect. The primary four-model evaluation uses one run set of 21 views $\times$ 300 items; the full-benchmark, counterpart, auditor-box, polarity-flip, and 32B runs are listed separately (Appendices M and N).

Statistical analysis. Intervals are percentile bootstrap 95% intervals over items within each model$\times$source cell; paired views are compared with McNemar's exact test (McNemar, 1947); sign stability is measured over 1,000 random 50/50 item splits (§4.5).

4.2 INTERVENTION PROFILES

Table 1 reports the unadjusted profile of the four models on the three sources.

MMStar. Masking the annotated evidence lowers accuracy for every model, by 0.24 to 0.29, and every interval excludes zero. The change is driven by harmful flips: 0.29–0.34 of items go from correct to incorrect, against 0.01–0.06 in the other direction. Cropping to the annotated region restores the original level (0.60–0.67 versus 0.59–0.64).

Observation 1. *On MMStar, removing the annotated evidence lowers accuracy through harmful flips, and cropping to the evidence restores it. In the unadjusted profile this is the signature of evidence use.*

BLINK-Twice. The same intervention raises accuracy for every model: from 0.39 to 0.72, 0.37 to 0.74, 0.27 to 0.74, and 0.36 to 0.68. Beneficial flips dominate (0.43–0.55 against 0.08–0.12), and cropping to the annotated region leaves accuracy at the original level (0.27–0.45).

Observation 2. *On BLINK-Twice, original accuracy is below the 0.50 chance level of a binary task, and masking the annotated evidence raises it for every model to about 0.7, while cropping to the evidence does not. Read on its own, the gain would be taken as improved use of the image.*

MMPerspective. Evidence-mask effects are small and positive, 0.02–0.10; the interval excludes zero for one of the four models (uncorrected McNemar $p = 0.82$ / 0.022 / 0.23 / 0.076; 0.09 for the smallest after a Holm correction over the four), and crop accuracy stays at the original level.

Observation 3. *On MMPerspective, a single evidence box produces a small response, nominally significant for one model in four.*

Observations 1–3 differ across sources, and the unadjusted profile cannot say which of them to believe. BLINK-Twice items are binary, so chance is 0.50, and a constant-No predictor scores 0.71 on this subset, the level of every masked accuracy in Table 1; MMPerspective's 0.45 majority

Table 1: **Unadjusted intervention profiles.** Acc^o, Acc^e, Acc^p are accuracies under the original, evidence-mask (black fill), and evidence-crop views; $\mathrm{ES}_e = \mathrm{Acc}^o - \mathrm{Acc}^e$ (bold) with a percentile bootstrap 95% interval, positive when masking lowers accuracy and negative when it raises it; HF_e / BF_e are harmful / beneficial flip rates. $n = 100$ per cell.

Model	Acc^o	Acc^e	Acc^p	ES_e [95% CI]	HF_e	BF_e	$\mathrm{Acc}^p - \mathrm{Acc}^e$
MMStar							
Qwen2.5-VL-3B	0.60	0.36	0.67	$+\mathbf{0.24}$ [+0.14, +0.35]	0.30	0.06	+0.31
Qwen2.5-VL-7B	0.60	0.31	0.67	$+\mathbf{0.29}$ [+0.18, +0.39]	0.34	0.05	+0.36
LLaVA-OV-7B	0.64	0.36	0.61	$+\mathbf{0.28}$ [+0.19, +0.37]	0.29	0.01	+0.25
Idefics3-8B	0.59	0.32	0.60	$+\mathbf{0.27}$ [+0.16, +0.38]	0.33	0.06	+0.28
BLINK-Twice							
Qwen2.5-VL-3B	0.39	0.72	0.37	$-\mathbf{0.33}$ [−0.46, −0.19]	0.12	0.45	−0.35
Qwen2.5-VL-7B	0.37	0.74	0.45	$-\mathbf{0.37}$ [−0.49, −0.24]	0.10	0.47	−0.29
LLaVA-OV-7B	0.27	0.74	0.27	$-\mathbf{0.47}$ [−0.60, −0.34]	0.08	0.55	−0.47
Idefics3-8B	0.36	0.68	0.35	$-\mathbf{0.32}$ [−0.45, −0.19]	0.11	0.43	−0.33
MMPerspective							
Qwen2.5-VL-3B	0.25	0.23	0.24	$+\mathbf{0.02}$ [−0.07, +0.10]	0.11	0.09	+0.01
Qwen2.5-VL-7B	0.32	0.23	0.31	$+\mathbf{0.09}$ [+0.03, +0.16]	0.11	0.02	+0.08
LLaVA-OV-7B	0.28	0.23	0.28	$+\mathbf{0.05}$ [−0.01, +0.11]	0.08	0.03	+0.05
Idefics3-8B	0.34	0.24	0.38	$+\mathbf{0.10}$ [+0.00, +0.20]	0.18	0.08	+0.14

option exceeds every original accuracy. The interventions are also unmatched: the black fill is one realization of removal, the crop changes scale and context, and for the same-area random rectangle with IoU < 0.1 that we first used as a location control no placement was found within 4,000 draws on 68 / 46 / 82 of 100 items, because the boxes are large (Appendix E). The next two subsections apply the two requirements of §3.2.

4.3 Legibility Panel Analysis

BLINK-Twice, masked view. In the original view the models answer Yes on 62–84% of items against a 29% gold rate; after masking the Yes rate falls to 15–29% (Table 2, Figure 2a). Stratified by gold label, the evidence-mask effect is -0.56 to -0.73 on the 71 gold-No items and $+0.17$ to $+0.34$ on the 29 gold-Yes items, so masking helps wherever the label is No and hurts wherever it is Yes. Balanced accuracy rises from 0.39–0.49 to 0.61–0.67, so the gain is not only a shift toward the frequent label. Masked predictions agree with blank-image predictions on 78–91% of items; we call the gain an *occlusion-induced response shift*. The panel locates the label correlation in question polarity. Nineteen questions are negated and eighteen of them have gold Yes; with the evidence hidden the models answer Yes to negated questions and No to the rest (Appendix I), and the gain is concentrated on the 81 non-negated items (ES_e -0.43 to -0.56, against -0.11 to $+0.21$ on the 19 negated ones; Appendix I). A rule that answers Yes to a negated question and No otherwise, with no image and no model, scores 0.88 on the subset (0.80 on the full benchmark), above every masked and blank-image accuracy, and masked predictions agree with this rule on 74–80% of items (Appendix D). After masking, κ is $+0.22$ to $+0.36$, and polarity contributes to that agreement causally: flipping the polarity of the wording on the same masked image moves the answer with it, removing a negation lowers the masked Yes rate from 0.32–0.68 to 0.05–0.21 and adding one raises it from 0.11–0.22 to 0.27–0.35, whereas an added negation leaves the original-view answer unchanged (Appendix I). Wording accounts for part of the masked-view agreement, and none of it is evidence that the masked image is used.

BLINK-Twice, original view. In the original view κ is -0.02 to -0.15: chance-adjusted agreement with the labels is nonpositive, and an accuracy of 0.27–0.39 is what a Yes-heavy predictor scores on a No-heavy subset. On all 693 items of the benchmark, every model scores 0.42–0.45 with κ from -0.15 to -0.09, below the 0.50 chance level and 0.12–0.15 below its own question-only accuracy, with every interval excluding zero under an image-cluster bootstrap (Table 3); in the terms of MMStar's multimodal gain, the gain is negative. The edited counterparts show that the models respond to the region's appearance. Each counterpart depicts the misleading appearance of a base

Table 2: **Score-legibility panel for BLINK-Twice (base images, $n = 100$).** All items are binary (chance 0.50); subset gold marginal 29 Yes / 71 No, constant-No 0.71. Yes is the fraction of predictions equal to Yes; the two answers-without-evidence columns give accuracy in the question-only view (q) and with the image fully blacked out (blank); ES_e is stratified by gold label ($n = 29$ / 71). A rule that answers Yes to negated questions and No otherwise scores 0.88 with no image.

	Original o			Evidence mask e			Without evidence		ES_e by gold	
Model	Acc	Yes	κ	Acc	Yes	κ	q	blank	Yes	No
Qwen2.5-VL-3B	0.39	0.72	$\mathbf{-0.03}$	0.72	0.25	$+0.29$	0.62	0.76	$+0.24$	$\mathbf{-0.56}$
Qwen2.5-VL-7B	0.37	0.62	$\mathbf{-0.14}$	0.74	0.15	$+0.26$	0.67	0.77	$+0.17$	$\mathbf{-0.59}$
LLaVA-OV-7B	0.27	0.84	$\mathbf{-0.14}$	0.74	0.27	$+0.36$	0.42	0.83	$+0.17$	$\mathbf{-0.73}$
Idefics3-8B	0.36	0.81	$\mathbf{-0.02}$	0.68	0.29	$+0.22$	0.61	0.78	$+0.34$	$\mathbf{-0.59}$

image as fact, and a first-glance answer recorded by a human on the counterpart equals the flipped label on 92 of 100 items. The models give the same original answer on the base image and its counterpart on 71–82% of items and the same masked answer on 71–87%; scored against the flipped label, this behavior yields the MMStar pattern (Table 3; Appendix L): masking lowers accuracy by 0.36–0.49 (0.37–0.51 on the 92 confirmed items, every interval excluding zero), the off-evidence control is within 0.02 of zero, and the crop recovers 0.39–0.49. The behavior is the same on both images and the delta's sign is set by whether the region's appearance matches the label; the delta alone cannot tell the two cases apart, and the panel can.

MMStar. Balanced accuracy tracks accuracy within 0.01, original κ is $+0.45$ to $+0.51$, and $\Delta\kappa$ is $+0.33$ / $+0.38$ / $+0.39$ / $+0.37$ with every interval excluding zero (Appendix C). Question-only accuracy is 0.21–0.31 against a 0.25 chance level and a 0.29 majority label, and accuracy under permuted option order stays within 0.04 of the original, so Observation 1 is not a marginal effect.

MMPerspective. Original accuracies (0.25–0.34) fall below the 0.45 majority-option baseline and original κ is $+0.01$ to $+0.17$. One model places 81% of its original predictions on option A; question-only accuracy is 0.18–0.25, and with a blank image Idefics3 scores 0.50 by answering the majority option D on 77% of items, 0.16 above its own original accuracy. These scores do not distinguish limited task competence from a mismatch between a single box and the cues the questions require.

Read with the panel, the BLINK-Twice gain does not establish more faithful use of the intended evidence: the masked view is the model's occluded-image answer, its agreement with the labels is matched by a text-only polarity rule, the original view shows no positive chance-adjusted agreement, and on the full benchmark the image lowers accuracy below the question-only score. Score legibility changes the conclusion the delta suggests and leaves the MMStar reading intact.

4.4 Matched-Control Analysis

Location. We replace the infeasible same-area mask with the zero-overlap off-evidence control r, in which the pixel budget of b_i is spread over rectangles disjoint from b_i. The control's own effect ES_r lies between -0.07 and $+0.03$ in all twelve cells, whereas Δ_{gap} is $+0.24$ to $+0.31$ on MMStar and -0.35 to -0.40 on BLINK-Twice. The budget is matched exactly only when the box covers at most half the frame (37 / 74 / 24 items); on those items Δ_{gap} is $+0.27$ / $+0.30$ / $+0.32$ / $+0.27$ on MMStar and -0.31 / -0.41 / -0.42 / -0.32 on BLINK-Twice, with every interval excluding zero, and 0.00 to $+0.17$ on MMPerspective with every interval including it (Table A7 in Appendix E).

Style. Six content-removing fills (black, gray, mean color, blur, noise, inpainting (Telea, 2004)) and one structure-preserving edit, mosaic, are applied at the annotated region (Appendix F, Figure A1). The six content-removing styles preserve the sign in every MMStar and BLINK-Twice cell: ES_e is $+0.21$ to $+0.31$ on MMStar and -0.29 to -0.52 on BLINK-Twice. Mosaic, which retains coarse structure and color, attenuates MMStar to $+0.10$ to $+0.20$ and BLINK-Twice to -0.05 to -0.12. The effect follows the amount of removed content and does not depend on the fill.

Table 3: **BLINK-Twice beyond the subset.** Top: the full benchmark (693 items on 364 images, base images; intervals from an image-cluster bootstrap), with question-only accuracy and the share of items on which a model gives the same answer to a base image and to its edited counterpart. Bottom: the 100 counterparts scored against the flipped label, with mask, control, and crop built by the same procedure from the box annotated on the counterpart.

Full benchmark, base images ($n = 693$; gold 334 Yes / 359 No, constant 0.52)

Model	Acc^o [95% CI]	Yes^o	κ^o	Acc^q	$\text{Acc}^o - \text{Acc}^q$	same answer on base and counterpart (n=207)
Qwen2.5-VL-3B	0.42 [0.38, 0.46]	0.57	**−0.15**	0.55	**−0.13**	0.70
Qwen2.5-VL-7B	0.43 [0.39, 0.47]	0.49	**−0.14**	0.56	**−0.13**	0.66
LLaVA-OV-7B	0.43 [0.40, 0.47]	0.73	**−0.12**	0.56	**−0.12**	0.77
Idefics3-8B	0.45 [0.41, 0.48]	0.70	**−0.09**	0.60	**−0.15**	0.70

Edited counterparts, flipped label ($n = 100$)

Model	Acc^o	Acc^e	ES_e	ES_r	$\text{Acc}^p - \text{Acc}^e$
Qwen2.5-VL-3B	0.75	0.39	**+0.36**	−0.02	+0.39
Qwen2.5-VL-7B	0.86	0.37	**+0.49**	0.00	+0.49
LLaVA-OV-7B	0.81	0.41	**+0.40**	+0.01	+0.40
Idefics3-8B	0.84	0.41	**+0.43**	0.00	+0.41

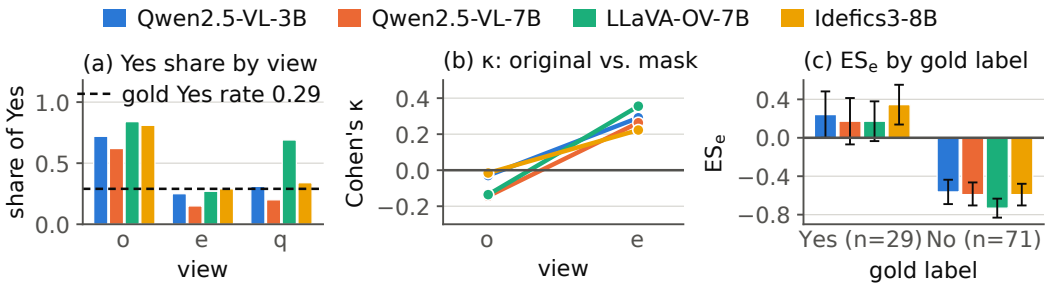


Figure 2: **Anatomy of the BLINK-Twice delta.** (a) Original predictions are dominated by Yes against a 29% gold rate; masking and question-only move the distribution toward No. (b) κ is nonpositive on the original view and positive after masking, consistent with the association between question polarity and the label. Error bars in (c) are bootstrap 95% intervals. (c) The evidence-mask effect is negative on gold-No items and positive on gold-Yes items.

Geometry. Crop recovery on MMStar is compared with a random-location crop of the same size, a full-image zoom of matched magnification, and a sweep over the crop margin (Appendix G). At matched magnification, the evidence crop exceeds the random-location crop by $+0.11$ to $+0.16$; and the tested full-image rescaling changes accuracy by -0.02 to $+0.01$ (Appendix G). Recovery is also item-level: of the MMStar items the mask breaks, 85–94% are correct again under the crop, against 3–26% of the items wrong under both views (Appendix G). For Qwen2.5-VL the processor resizes inputs to at most 384×384, so the rescaling control is uninformative there and the random-crop contrast carries the comparison. On BLINK-Twice the evidence crop stays at the original level (0.27–0.45), whereas a random-location crop of the same size, which leaves the annotated region out, raises it to 0.51–0.58: removing the region by cropping produces the same shift as masking it.

Annotation. With the evidence views rebuilt from either auditor's boxes (Table A16), the MMStar effect is $+0.24$ to $+0.38$ and the BLINK-Twice effect -0.34 to -0.54, against $+0.16$ to $+0.30$ and -0.32 to -0.47 with the primary boxes on the same items, and the specificity gap and crop recovery keep their sign. On MMPerspective three models are unchanged and Idefics3 changes sign under both auditors' boxes ($+0.10$ to -0.31 and -0.22), so the single-box profile on that source is not stable across annotators.

With the controls matched, MMStar's degradation is specific to the annotated region, holds under six removal styles and under independent boxes, and is not reproduced by magnification; the

BLINK-Twice response shift follows the annotated region, is attenuated when the region's structure is retained, and is reproduced under independent boxes. The unadjusted profile could not have distinguished either result from an artifact of the edit.

4.5 Stability Analysis

Over 1,000 random 50/50 splits, the signs of ES_e and Δ_{gap} agree across halves in every split for every model on MMStar and BLINK-Twice; on MMPerspective sign agreement of ES_e is 0.21 / 0.97 / 0.78 / 0.94 (Appendix H). Within every evidence-box area quartile the point estimate of the MMStar effect is positive and that of the BLINK-Twice effect negative for every model (Appendix I). On the 37 MMPerspective items with additional annotated regions, masking all of them gives $+0.03$ to $+0.05$ against 0.00 to $+0.05$ for the single box, and the evidence-extent pass classifies 86 of the 100 items as depending on the global layout, so the local interventions leave the MMPerspective effect unresolved. The 32B model, run with a longer generation budget, reproduces the MMStar and BLINK-Twice diagnoses (Appendix J).

5 Discussion

The two requirements change the reading each source receives: the MMStar diagnosis survives on controls and on independent boxes; the BLINK-Twice gain is a response shift, and its original view shows no positive agreement with the labels; the MMPerspective effect is left unresolved. Without the panel, a delta of -0.3 to -0.5 replicated across three model families would have been reported as improved use of the image.

A positive ES_e that survives $\Delta\kappa$, a near-zero off-evidence control, and several removal styles indicates that the score depends on the annotated region; a negative ES_e whose masked view agrees with the model's answer without evidence indicates a response shift; a small ES_e on a score below a majority baseline indicates a profile for which the tested interventions do not identify the source of the effect, and which of the three applies decides whether the benchmark supports a claim that the intended evidence was used.

Limitations. The audit covers 100 items per source and four open models from 3B to 8B, with a 32B model on the core views under a longer generation budget; frontier API models are not covered, and all runs use a direct-answer protocol with inputs resized to at most 384×384 for Qwen2.5-VL. The BLINK-Twice subset is more No-heavy (71%) than the full source (52%); the size of the masked-side gain is a property of the subset, whereas the nonpositive κ and the negative multimodal gain hold on the full benchmark. The primary boxes agree moderately with two blind auditors (mean IoU 0.59 / 0.69 / 0.58), and one BLINK-Twice auditor had seen the edited counterparts in an earlier task. The counterparts are scored against flipped labels that a human first-glance check supports on 92 of 100 items, the effects hold on those 92, and they are used as a control and are not offered as a benchmark. Validating the two requirements against predictors with known mechanisms is the next step.

In conclusion, reading an intervention delta as evidence that the intended evidence was used requires the same care as reading the accuracy it is meant to explain. Masking the annotated evidence produces opposite deltas on MMStar and BLINK-Twice; under score legibility the BLINK-Twice gain is a response shift matched by a text-only polarity rule, and under matched intervention MMStar's degradation is region-specific and style-robust. We recommend that both requirements accompany every intervention delta offered as evidence of grounded visual reasoning.

AI use statement

Generative AI tools were used solely to assist with language editing and to improve the clarity and readability of the manuscript. All scientific content, analyses, results, and conclusions were developed and verified by the authors.

Reproducibility statement

Views, metrics, and the panel are defined in §3 and Appendix A; models, prompt, decoding, and extraction in §4.1 and Appendix M; the annotation audit in Appendix K and the counterpart control in Appendix L. The annotation records and audit sheets, the intervention generator, the inference script, per-item predictions with raw outputs and run manifests (checkpoint revision, library versions) for every model$\times$view cell, and the scripts that generate every table in this paper from those predictions will be released publicly.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, 2024.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955.

Ricardo Dominguez-Olmedo, Florian E. Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence. In *International Conference on Learning Representations*, 2025.

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

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

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Yifan Hou, Buse Giledereli, Yilei Tu, and Mrinmaya Sachan. Do vision-language models really understand visual language? In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23910–23959, 2025.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, 2023.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Zixuan Lan, Luzhe Sun, Matthew R. Walter, and Jiawei Zhou. Seeing without looking: Do vision-language benchmarks really test vision? *arXiv preprint arXiv:2605.22903*, 2026.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: Insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. VLind-Bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, 2025.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. arXiv:2408.03326.

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

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in VLMs. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41120–41156, 2025.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Samuel Messick. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50 (9):741–749, 1995.

Mazda Moayeri, Vidhisha Balachandran, Varun Chandrasekaran, Safoora Yousefi, Thomas Fel, Soheil Feizi, Besmira Nushi, Neel Joshi, and Vibhav Vineet. Unearthing skill-level insights for understanding trade-offs of foundation models. In *International Conference on Learning Representations*, 2025.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*, 2018.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *International Conference on Machine Learning*, 2022.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, 2017.

Yunlong Tang, Pinxin Liu, Mingqian Feng, Zhangyun Tan, Rui Mao, Chao Huang, Jing Bi, Yunzhong Xiao, Susan Liang, Hang Hua, Ali Vosoughi, Luchuan Song, Zeliang Zhang, and Chenliang Xu. MMPerspective: Do MLLMs understand perspective? A comprehensive benchmark for perspective perception, reasoning, and robustness. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, 9(1):23–34, 2004.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Yucheng Wang, Yifan Hou, Aydin Javadov, Mubashara Akhtar, and Mrinmaya Sachan. Compose and fuse: Revisiting the foundational bottlenecks in multimodal reasoning. In *International Conference on Learning Representations*, 2026.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. VisuLogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In *International Conference on Learning Representations*, 2026.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? Assessing amplified hallucination in multimodal reasoning models. In *Advances in Neural Information Processing Systems*, 2025.

Junyan Ye, Dongzhi Jiang, Jun He, Baichuan Zhou, Zilong Huang, Zhiyuan Yan, Hongsheng Li, Conghui He, and Weijia Li. BLINK-Twice: You see, but do you observe? A reasoning benchmark on visual perception. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

A METRIC DEFINITIONS

For item i and view c the extractor maps the model response to an option label $\hat{y}_i^c$ and correctness $z_i^c = \mathbb{1}[\hat{y}_i^c = y_i]$:

$$\mathrm{Acc}^c = \tfrac{1}{n}\textstyle\sum_i z_i^c, \qquad \mathrm{ES}_e = \mathrm{Acc}^o - \mathrm{Acc}^e, \qquad \mathrm{ES}_r = \mathrm{Acc}^o - \mathrm{Acc}^r, \qquad \Delta_{\mathrm{gap}} = \mathrm{ES}_e - \mathrm{ES}_r,$$
$$\mathrm{HF}_e = \tfrac{1}{n}\textstyle\sum_i \mathbb{1}[z_i^o = 1, z_i^e = 0], \qquad \mathrm{BF}_e = \tfrac{1}{n}\textstyle\sum_i \mathbb{1}[z_i^o = 0, z_i^e = 1], \qquad \mathrm{ES}_e = \mathrm{HF}_e - \mathrm{BF}_e,$$
$$\kappa^c = \frac{p_{\mathrm{obs}}^c - p_{\mathrm{exp}}^c}{1 - p_{\mathrm{exp}}^c}, \qquad \Delta\kappa = \kappa^o - \kappa^e,$$

where $p_{\mathrm{obs}}^c = \mathrm{Acc}^c$ and $p_{\mathrm{exp}}^c = \sum_y \Pr(\hat{y}^c = y)\Pr(y_i = y)$. Balanced accuracy is the mean per-class recall. Negative ES_e is neither clipped nor discarded. The answers without evidence are taken in the question-only view q (no image) and in the blank view (the image replaced by a black canvas of the same size). Table A1 lists the views and controls with the contrast each one supports.

Table A1: **Views and controls.** Each view is a deterministic function of the image and the evidence box b_i. The last column gives the section of the main text that reports the contrast.

View or control	Change relative to o	Contrast (reads as)	§
Evidence mask e	content inside b_i replaced (black fill)	$\mathrm{ES}_e = \mathrm{Acc}^o - \mathrm{Acc}^e$, HF_e, BF_e (sensitivity to the annotated evidence)	4.2
Evidence crop p	content outside b_i removed; region resized	$\mathrm{Acc}^p - \mathrm{Acc}^e$ (crop recovery)	4.2
Question-only q; blank	image removed; image replaced by a black canvas	$\mathrm{Acc}^o - \mathrm{Acc}^q$ (multimodal gain); panel (answers without evidence)	4.3
Off-evidence control r	same pixel budget removed outside b_i, zero overlap	$\Delta_{\mathrm{gap}} = \mathrm{ES}_e - \mathrm{ES}_r$ (specificity to the location)	4.4
Style	fill of e varied over seven fills	ES per fill (independence from the fill)	4.4
Geometry	crop location, magnification, and margin varied	$\mathrm{Acc}^p - \mathrm{Acc}^{\text{random crop}}$; zoom $- \mathrm{Acc}^o$ (independence from magnification)	4.4
Annotation	e, r, p regenerated from auditors' boxes	profile regenerated (independence from the annotator)	4.4

B FULL PROFILE

Table A2 gives the full point estimates of the profile for every model and source, including the zero-overlap control r, the legacy same-area rectangle R, the flip rates, and McNemar's exact test.

Table A2: Full point estimates. r is the zero-overlap off-evidence control and R the infeasible same-area rectangle (Appendix E); Chg. is the fraction of items whose predicted option differs between o and e; p_{McN} is McNemar's exact test of o versus e.

Model	Source	Acc^o	Acc^e	Acc^r	Acc^R	Acc^p	ES_e	ES_r	Δ_{gap}	ES_R	HF_e	BF_e	Chg.	$\mathrm{Acc}^p - \mathrm{Acc}^e$	p_{McN}
Qwen2.5-VL-3B	MMStar	0.60	0.36	0.61	0.52	0.67	+0.24	−0.01	+0.25	+0.08	0.30	0.06	0.52	+0.31	7e-05
Qwen2.5-VL-3B	BLINK-Twice	0.39	0.72	0.37	0.54	0.37	−0.33	+0.02	−0.35	−0.15	0.12	0.45	0.57	−0.35	1.3e-05
Qwen2.5-VL-3B	MMPerspective	0.25	0.23	0.22	0.23	0.24	+0.02	+0.03	−0.01	+0.02	0.11	0.09	0.41	+0.01	0.82
Qwen2.5-VL-7B	MMStar	0.60	0.31	0.62	0.56	0.67	+0.29	−0.02	+0.31	+0.04	0.34	0.05	0.55	+0.36	2.4e-06
Qwen2.5-VL-7B	BLINK-Twice	0.37	0.74	0.36	0.49	0.45	−0.37	+0.01	−0.38	−0.12	0.10	0.47	0.57	−0.29	7.5e-07
Qwen2.5-VL-7B	MMPerspective	0.32	0.23	0.29	0.26	0.31	+0.09	+0.03	+0.06	+0.06	0.11	0.02	0.27	+0.08	0.022
LLaVA-OV-7B	MMStar	0.64	0.36	0.62	0.57	0.61	+0.28	+0.02	+0.26	+0.07	0.29	0.01	0.44	+0.25	5.8e-08
LLaVA-OV-7B	BLINK-Twice	0.27	0.74	0.34	0.46	0.27	−0.47	−0.07	−0.40	−0.19	0.08	0.55	0.63	−0.47	9.8e-10
LLaVA-OV-7B	MMPerspective	0.28	0.23	0.28	0.27	0.28	+0.05	+0.00	+0.05	+0.01	0.08	0.03	0.20	+0.05	0.23
Idefics3-8B	MMStar	0.59	0.32	0.56	0.43	0.60	+0.27	+0.03	+0.24	+0.16	0.33	0.06	0.47	+0.28	1.4e-05
Idefics3-8B	BLINK-Twice	0.36	0.68	0.33	0.48	0.35	−0.32	+0.03	−0.35	−0.12	0.11	0.43	0.54	−0.33	1.4e-05
Idefics3-8B	MMPerspective	0.34	0.24	0.34	0.25	0.38	+0.10	+0.00	+0.10	+0.09	0.18	0.08	0.49	+0.14	0.076

C SCORE-LEGIBILITY PANELS FOR MMSTAR AND MMPERSPECTIVE

MMStar: four options, chance 0.25, gold marginal 29 / 26 / 21 / 24, majority label 0.29. MMPerspective: four options, chance 0.25, gold marginal 26 / 21 / 8 / 45, majority option 0.45, one option

Table A3: Per-model panel. Pred. is the original-view predicted-label distribution over options A/B/C/D; Acc^{π} is accuracy with option order permuted and gold remapped.

Source	Model	Acc^{o}	Pred. A/B/C/D	κ^{o}	κ^{e}	Acc^{q}	Acc^{π}
MMStar	Qwen2.5-VL-3B	0.60	0.24/0.28/0.28/0.20	$+0.467$	$+0.141$	0.30	0.60
MMStar	Qwen2.5-VL-7B	0.60	0.27/0.28/0.22/0.23	$+0.465$	$+0.081$	0.21	0.64
MMStar	LLaVA-OV-7B	0.64	0.45/0.26/0.18/0.11	$+0.512$	$+0.127$	0.31	0.67
MMStar	Idefics3-8B	0.59	0.35/0.30/0.20/0.15	$+0.448$	$+0.079$	0.25	0.61
MMPerspective	Qwen2.5-VL-3B	0.25	0.41/0.34/0.12/0.13	$+0.005$	-0.001	0.18	0.24
MMPerspective	Qwen2.5-VL-7B	0.32	0.07/0.57/0.20/0.16	$+0.122$	-0.008	0.21	0.35
MMPerspective	LLaVA-OV-7B	0.28	0.81/0.07/0.09/0.03	$+0.045$	-0.044	0.23	0.39
MMPerspective	Idefics3-8B	0.34	0.18/0.09/0.51/0.22	$+0.169$	$+0.044$	0.25	0.27

Table A4: MMStar evidence-mask effect in accuracy, in balanced accuracy, and as $\Delta\kappa$ with a bootstrap 95% interval.

Model	ES_e (accuracy)	ES_e (balanced accuracy)	$\Delta\kappa$ [95% CI]
Qwen2.5-VL-3B	0.24	0.244	0.326 [0.183, 0.485]
Qwen2.5-VL-7B	0.29	0.281	0.384 [0.246, 0.518]
LLaVA-OV-7B	0.28	0.278	0.385 [0.260, 0.500]
Idefics3-8B	0.27	0.272	0.369 [0.240, 0.503]

set in 71 of 100 items. Table A3 reports per-model quantities. The MMStar evidence-mask effect after label adjustment is in Table A4.

D BALANCED ACCURACY AND THE POLARITY RULE ON BLINK-TWICE

A rule that answers Yes to a negated question and No otherwise uses no image and no model. With the 19 negated questions (18 gold Yes) and 81 non-negated questions (70 gold No) of the subset it scores 0.88 (balanced accuracy 0.80); on the 693 questions of the full benchmark (210 negated) it scores 0.80. Table A5 gives balanced accuracy under the original, evidence-mask, and blank views and the agreement of each model's predictions with the rule.

Table A5: BLINK-Twice ($n = 100$): balanced accuracy (mean per-class recall) under the original, evidence-mask, and blank views, and the fraction of predictions that equal the polarity rule's answer.

	Original o		Evidence mask e		Blank		Agreement with the rule		
Model	Acc	bal. Acc	Acc	bal. Acc	Acc	bal. Acc	e	blank	q
Qwen2.5-VL-3B	0.39	0.48	0.72	0.64	0.76	0.66	0.74	0.78	0.62
Qwen2.5-VL-7B	0.37	0.40	0.74	0.61	0.77	0.63	0.78	0.81	0.67
LLaVA-OV-7B	0.27	0.39	0.74	0.67	0.83	0.73	0.80	0.87	0.40
Idefics3-8B	0.36	0.49	0.68	0.61	0.78	0.64	0.74	0.86	0.65

E INFEASIBILITY OF THE SAME-AREA RANDOM MASK

The initial control sampled one rectangle per item with the area of the evidence box, inside the image, minimizing IoU with the box over 4,000 draws. Table A6 reports feasibility per source. The zero-overlap replacement distributes the same pixel budget over rectangles placed in the four bands around the box; exact budget matching is possible only when the box covers at most half the frame, and the achieved-to-target area ratio averages 0.77 / 0.89 / 0.75.

Table A6: Feasibility of the same-area random mask and its zero-overlap replacement, per source (100 items each); ranges are over the four models.

	MMStar	BLINK-Twice	MMPerspective
Median evidence-box area (fraction of frame)	0.54	0.32	0.57
Items with no IoU < 0.1 rectangle in 4,000 draws	68	46	82
Mean minimum IoU found	0.21	0.11	0.25
Items with exact zero-overlap budget match	37	74	24
ES_r, zero-overlap control	-0.02 to $+0.03$	-0.07 to $+0.03$	0.00 to $+0.03$
ES_R, same-area rectangle	$+0.04$ to $+0.16$	-0.19 to -0.12	$+0.01$ to $+0.09$
Δ_{gap}, all items	$+0.24$ to $+0.31$	-0.40 to -0.35	-0.01 to $+0.10$
Δ_{gap}, exactly matched items	$+0.27$ to $+0.32$	-0.42 to -0.31	0.00 to $+0.17$

Table A7: Specificity gap on the exactly budget-matched items (box area at most half the frame) and on the remaining items, per model; intervals are percentile bootstrap 95% intervals on the matched items.

Source	Model	n exact	ES_e exact	ES_r exact	Δ_{gap} exact [95% CI]	Δ_{gap} rest (n)	Δ_{gap} all
MMStar	Qwen2.5-VL-3B	37	$+0.24$	-0.03	$+0.27$ [$+0.08$, $+0.43$]	$+0.24$ (63)	$+0.25$
MMStar	Qwen2.5-VL-7B	37	$+0.22$	-0.08	$+0.30$ [$+0.11$, $+0.49$]	$+0.32$ (63)	$+0.31$
MMStar	LLaVA-OV-7B	37	$+0.38$	$+0.05$	$+0.32$ [$+0.16$, $+0.49$]	$+0.22$ (63)	$+0.26$
MMStar	Idefics3-8B	37	$+0.30$	$+0.03$	$+0.27$ [$+0.11$, $+0.43$]	$+0.22$ (63)	$+0.24$
BLINK-Twice	Qwen2.5-VL-3B	74	-0.28	$+0.03$	-0.31 [-0.46, -0.16]	-0.46 (26)	-0.35
BLINK-Twice	Qwen2.5-VL-7B	74	-0.39	$+0.01$	-0.41 [-0.54, -0.26]	-0.31 (26)	-0.38
BLINK-Twice	LLaVA-OV-7B	74	-0.47	-0.05	-0.42 [-0.55, -0.27]	-0.35 (26)	-0.40
BLINK-Twice	Idefics3-8B	74	-0.27	$+0.05$	-0.32 [-0.46, -0.18]	-0.42 (26)	-0.35
MMPerspective	Qwen2.5-VL-3B	24	$+0.12$	$+0.12$	$+0.00$ [-0.21, $+0.21$]	-0.01 (76)	-0.01
MMPerspective	Qwen2.5-VL-7B	24	$+0.12$	$+0.00$	$+0.12$ [-0.04, $+0.29$]	$+0.04$ (76)	$+0.06$
MMPerspective	LLaVA-OV-7B	24	$+0.12$	-0.04	$+0.17$ [$+0.00$, $+0.38$]	$+0.01$ (76)	$+0.05$
MMPerspective	Idefics3-8B	24	$+0.00$	$+0.00$	$+0.00$ [-0.21, $+0.21$]	$+0.13$ (76)	$+0.10$

F REMOVAL-STYLE SWEEP

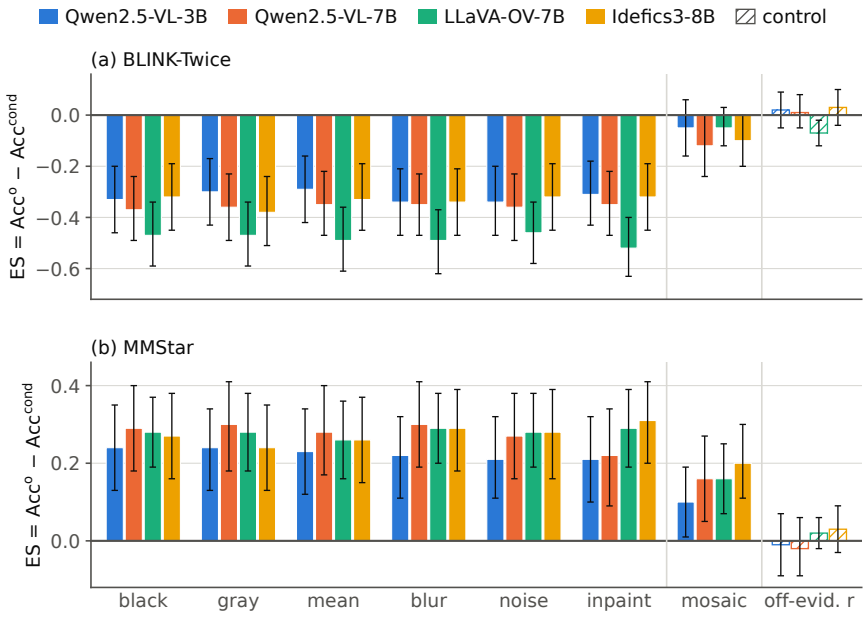


Figure A1: **Removal-style sweep.** Six content-removing edits at the annotated region reproduce the BLINK-Twice shift and the MMStar degradation; a structure-preserving mosaic attenuates both. The zero-overlap off-evidence control (hatched) stays near zero. Bars are ES with bootstrap 95% intervals.

Figure A1 plots the effect of each removal style, and Table A8 lists the values, including the zero-overlap control, the legacy same-area rectangle, and the blank image.

Table A8: ES $=$ Acc^o $-$ Acc$^{\text{cond}}$ per condition. Columns within each source: Qwen2.5-VL-3B, Qwen2.5-VL-7B, LLaVA-OV-7B, Idefics3-8B.

	BLINK-Twice				MMStar				MMPerspective			
Condition	Q3B	Q7B	LLaVA	Idef.	Q3B	Q7B	LLaVA	Idef.	Q3B	Q7B	LLaVA	Idef.
Black	−0.33	−0.37	−0.47	−0.32	+0.24	+0.29	+0.28	+0.27	+0.02	+0.09	+0.05	+0.10
Gray	−0.30	−0.36	−0.47	−0.38	+0.24	+0.30	+0.28	+0.24	+0.06	+0.09	+0.02	+0.10
Mean color	−0.29	−0.35	−0.49	−0.33	+0.23	+0.28	+0.26	+0.26	+0.04	+0.09	+0.01	+0.10
Gaussian blur	−0.34	−0.35	−0.49	−0.34	+0.22	+0.30	+0.29	+0.29	+0.08	+0.08	+0.00	+0.07
Random noise	−0.34	−0.36	−0.46	−0.32	+0.21	+0.27	+0.28	+0.28	+0.05	+0.10	+0.03	+0.06
Inpainting	−0.31	−0.35	−0.52	−0.32	+0.21	+0.22	+0.29	+0.31	+0.07	+0.00	−0.01	+0.21
Mosaic	−0.05	−0.12	−0.05	−0.10	+0.10	+0.16	+0.16	+0.20	+0.00	+0.03	−0.02	+0.07
Off-evidence r (zero overlap)	+0.02	+0.01	−0.07	+0.03	−0.01	−0.02	+0.02	+0.03	+0.03	+0.03	+0.00	+0.00
Same-area R	−0.15	−0.12	−0.19	−0.12	+0.08	+0.04	+0.07	+0.16	+0.02	+0.06	+0.01	+0.09
Blank image	−0.37	−0.40	−0.56	−0.42	+0.32	+0.36	+0.33	+0.38	+0.06	+0.12	+0.05	−0.16

G CROP-GEOMETRY CONTROLS

Table A9 reports the crop-geometry controls on MMStar: the evidence crop against a random-location crop of the same size, a full-image zoom of matched magnification, and the crop-margin sweep. On BLINK-Twice the evidence crop scores 0.27–0.45 and the random-location crop 0.51–0.58; on MMPerspective the two differ by at most 0.05.

Table A9: Crop-geometry controls on MMStar. The random-location crop has the size of the evidence crop and IoU < 0.1 with the box where achievable; the zoom upscales the whole image by the crop's linear magnification.

Condition	Qwen2.5-VL-3B	Qwen2.5-VL-7B	LLaVA-OV-7B	Idefics3-8B
Original Acc^o	0.60	0.60	0.64	0.59
Evidence crop Acc^p (margin 0.15)	0.67	0.67	0.61	0.60
Random-location crop	0.51	0.55	0.50	0.48
Full-image zoom	0.61	0.60	0.64	0.57
Margin sweep 0 / 0.15 / 0.5 / 1.0 (min–max)	0.61–0.67	0.65–0.68	0.61–0.63	0.54–0.60
Evidence − random crop	+0.16	+0.12	+0.11	+0.12
Zoom − original	+0.01	+0.00	+0.00	−0.02

Table A10 reports item-level recovery: among the items a model answers correctly in the original view and incorrectly under the evidence mask, the share it answers correctly under the evidence crop, against the same share among items wrong under both views.

Table A10: Item-level crop recovery (margin 0.15). $P(z^p{=}1 \mid z^o{=}1, z^e{=}0)$ with the number of such items, and $P(z^p{=}1 \mid z^o{=}0, z^e{=}0)$ as the comparison.

	$P(z^p{=}1 \mid z^o{=}1, z^e{=}0)$ (n)				$P(z^p{=}1 \mid z^o{=}0, z^e{=}0)$			
Source	Q3B	Q7B	LLaVA	Idef.	Q3B	Q7B	LLaVA	Idef.
MMStar	0.90 (30)	0.94 (34)	0.90 (29)	0.85 (33)	0.26	0.20	0.03	0.26
BLINK-Twice	0.67 (12)	0.80 (10)	1.00 (8)	0.73 (11)	0.25	0.19	0.11	0.24
MMPerspective	0.45 (11)	0.82 (11)	1.00 (8)	0.67 (18)	0.09	0.05	0.03	0.19

H STABILITY

Table A11 reports the sign agreement of ES_e across 1,000 random 50/50 item splits and a TOST test of equivalence to zero at a 0.10 margin, per model and source.

Table A11: Sign agreement of ES_e across 1,000 random 50/50 item splits; TOST p for equivalence to zero at a 0.10 margin.

	MMStar		BLINK-Twice		MMPerspective	
Model	sign	p_{TOST}	sign	p_{TOST}	sign	p_{TOST}
Qwen2.5-VL-3B	1.00	0.99	1.00	1.00	0.21	0.039
Qwen2.5-VL-7B	1.00	1.00	1.00	1.00	0.97	0.39
LLaVA-OV-7B	1.00	1.00	1.00	1.00	0.78	0.066
Idefics3-8B	1.00	1.00	1.00	1.00	0.94	0.50

I STRATIFIED EFFECTS

Table A12 reports the evidence-mask effect within gold-label, question-polarity, box-area, and multi-region strata.

Polarity-flipped wording. Each of the 100 BLINK-Twice questions was rewritten with the opposite polarity (the 19 negated questions made affirmative, the 81 affirmative questions negated; 93 by rule and 7 sentence-form questions by hand) and run on the same images, boxes, and views. Table A13 reports the Yes rate under the original and the flipped wording. Under the evidence mask the answer follows the wording in both directions; the change is asymmetric, since an added negation lifts the

Table A12: Evidence-mask effect by stratum; ranges are over the four models. Area quartiles are computed within source.

Source	Stratum	n	ES_e (range over models)
BLINK-Twice	gold = Yes	29	$+0.17$ to $+0.34$
BLINK-Twice	gold = No	71	-0.73 to -0.56
BLINK-Twice	negated question	19	-0.11 to $+0.21$
BLINK-Twice	non-negated question	81	-0.56 to -0.43
BLINK-Twice	box-area quartiles Q1–Q4	25 each	-0.60 to -0.12
MMStar	box-area quartiles Q1–Q4	25 each	$+0.04$ to $+0.44$
MMPerspective	several regions annotated, single box masked	37	0.00 to $+0.05$
MMPerspective	several regions annotated, all regions masked	37	$+0.03$ to $+0.05$
MMPerspective	remaining items	63	0.00 to $+0.13$

masked Yes rate to about a third, below the 0.32–0.68 of the originally negated items, so wording accounts for part of the masked-view agreement and the originally negated items differ in content as well. Under the original view an added negation leaves the Yes rate unchanged.

Table A13: Yes rate under the original and the polarity-flipped wording on the same images and views (negation removed: 19 items; negation added: 81 items). Intervals are bootstrap 95% intervals of the difference.

		Negation removed (19)		Negation added (81)	
Model	View	Yes rate orig. $\rightarrow$ flipped	Δ [95% CI]	Yes rate orig. $\rightarrow$ flipped	Δ [95% CI]
Qwen2.5-VL-3B	e	0.47 $\rightarrow$ 0.05	-0.42 [-0.68, -0.16]	0.20 $\rightarrow$ 0.28	$+0.09$ [-0.01, $+0.19$]
Qwen2.5-VL-7B	e	0.32 $\rightarrow$ 0.16	-0.16 [-0.42, $+0.11$]	0.11 $\rightarrow$ 0.27	$+0.16$ [$+0.09$, $+0.25$]
LLaVA-OV-7B	e	0.68 $\rightarrow$ 0.21	-0.47 [-0.74, -0.16]	0.17 $\rightarrow$ 0.33	$+0.16$ [$+0.05$, $+0.26$]
Idefics3-8B	e	0.58 $\rightarrow$ 0.05	-0.53 [-0.79, -0.26]	0.22 $\rightarrow$ 0.35	$+0.12$ [0.00, $+0.23$]
Qwen2.5-VL-3B	blank	0.42 $\rightarrow$ 0.21	-0.21 [-0.47, $+0.05$]	0.14 $\rightarrow$ 0.26	$+0.12$ [$+0.06$, $+0.20$]
Qwen2.5-VL-7B	blank	0.32 $\rightarrow$ 0.11	-0.21 [-0.47, $+0.05$]	0.07 $\rightarrow$ 0.17	$+0.10$ [$+0.04$, $+0.17$]
LLaVA-OV-7B	blank	0.58 $\rightarrow$ 0.00	-0.58 [-0.79, -0.37]	0.06 $\rightarrow$ 0.02	-0.04 [-0.10, $+0.01$]
Idefics3-8B	blank	0.42 $\rightarrow$ 0.00	-0.42 [-0.63, -0.21]	0.04 $\rightarrow$ 0.11	$+0.07$ [$+0.02$, $+0.14$]
Qwen2.5-VL-3B	o	0.58 $\rightarrow$ 0.79	$+0.21$ [-0.05, $+0.47$]	0.75 $\rightarrow$ 0.74	-0.01 [-0.14, $+0.11$]
Qwen2.5-VL-7B	o	0.21 $\rightarrow$ 0.68	$+0.47$ [$+0.16$, $+0.74$]	0.72 $\rightarrow$ 0.79	$+0.07$ [-0.05, $+0.19$]
LLaVA-OV-7B	o	0.68 $\rightarrow$ 0.89	$+0.21$ [-0.05, $+0.47$]	0.88 $\rightarrow$ 0.85	-0.02 [-0.11, $+0.06$]
Idefics3-8B	o	0.79 $\rightarrow$ 0.95	$+0.16$ [-0.05, $+0.37$]	0.81 $\rightarrow$ 0.90	$+0.09$ [-0.01, $+0.19$]

J MODEL SCALE

Qwen2.5-VL-32B does not follow the single-letter instruction within four tokens on MMStar (valid-output rate 0.71), so we report it with 256 generated tokens and extraction of the final option letter; BLINK-Twice and MMPerspective values are identical under both settings. Table A14 reports the core views.

Table A14: Qwen2.5-VL-32B on the core views (same prompt; 256 tokens; final-letter extraction).

Source	Acc^o	Acc^e	Acc^r	Acc^p	Acc^q	ES_e	Δ_{gap}	$\mathrm{ES}_{\mathrm{mosaic}}$
MMStar	0.62	0.33	0.55	0.65	0.19	$+0.29$	$+0.22$	$+0.17$
BLINK-Twice	0.41	0.78	0.40	0.43	0.78	-0.37	-0.38	-0.10
MMPerspective	0.43	0.33	0.47	0.42	0.21	$+0.10$	$+0.14$	$+0.04$

K ANNOTATION SCHEMA AND BLIND RE-ANNOTATION AUDIT

Each record stores the source benchmark, source item identifier, question, options, gold option, evidence box (x_1, y_1, x_2, y_2) in pixel coordinates, and notes. For the audit, 50 items per source were sampled with proportional allocation over source categories for MMStar and MMPerspective, and all 100 BLINK-Twice items were audited. Two auditors, working independently and without access to the primary boxes, drew one box per item from the image, question, and gold answer, recorded a confidence from 1 to 5, and could mark an item as not localizable with a single box; IoU is computed on the items for which the auditor drew a box. One auditor reported seeing the primary coordinates of one item before drawing; that record is excluded. Auditor A had seen the edited counterparts of the BLINK-Twice items in an earlier task and is therefore not fully blind on that source; auditor B is. Table A15 reports agreement. On six of the MMStar and MMPerspective items both auditors agree with each other (IoU ≥ 0.5) and disagree with the primary box (IoU < 0.3); in four of them the primary box is markedly larger or smaller than the auditors', and in two it marks a different region.

Table A15: Blind re-annotation audit. Primary–auditor values average the two auditors; n.l. is the number of items marked not localizable by auditor A / B.

		Primary vs. auditor			Auditor vs. auditor	
Source	n	Mean IoU	Median IoU	IoU ≥ 0.5	Mean IoU (IoU ≥ 0.5)	n.l.
MMStar	50	0.59	0.62	71%	0.73 (80%)	14 / 4
BLINK-Twice	100	0.69	0.73	84%	0.68 (81%)	1 / 4
MMPerspective	50	0.58	0.57	78%	0.82 (84%)	27 / 25
Overall	200	0.63	0.64	79%	0.73 (82%)	42 / 33

Interventions from the auditors' boxes. Table A16 rebuilds the evidence mask, the off-evidence control, and the evidence crop from each auditor's boxes on the audited items (MMStar and MMPerspective: 50 items each, one MMPerspective item excluded for the blindness breach; BLINK-Twice: 100 items on the base images) and compares the effects with those obtained from the primary boxes on the same items.

Table A16: Effects under the primary (P) and the two auditors' (A, B) boxes on the audited items. Box IoU: MMStar P–A 0.56, P–B 0.61, A–B 0.73; BLINK-Twice 0.68 / 0.70 / 0.68; MMPerspective 0.55 / 0.61 / 0.82.

			ES_e			Δ_{gap}			$\mathrm{Acc}^p - \mathrm{Acc}^e$		
Source	Model	n	P	A	B	P	A	B	P	A	B
MMStar	Qwen2.5-VL-3B	50	+0.16	+0.28	+0.32	+0.22	+0.32	+0.38	+0.20	+0.30	+0.40
MMStar	Qwen2.5-VL-7B	50	+0.24	+0.32	+0.34	+0.24	+0.42	+0.44	+0.32	+0.38	+0.44
MMStar	LLaVA-OV-7B	50	+0.18	+0.28	+0.24	+0.16	+0.28	+0.24	+0.16	+0.32	+0.24
MMStar	Idefics3-8B	50	+0.30	+0.34	+0.38	+0.28	+0.30	+0.34	+0.28	+0.28	+0.32
BLINK-Twice	Qwen2.5-VL-3B	100	−0.33	−0.40	−0.34	−0.35	−0.44	−0.35	−0.35	−0.42	−0.35
BLINK-Twice	Qwen2.5-VL-7B	100	−0.37	−0.41	−0.36	−0.38	−0.39	−0.35	−0.29	−0.39	−0.41
BLINK-Twice	LLaVA-OV-7B	100	−0.47	−0.54	−0.53	−0.40	−0.49	−0.48	−0.47	−0.50	−0.47
BLINK-Twice	Idefics3-8B	100	−0.32	−0.34	−0.36	−0.35	−0.34	−0.33	−0.33	−0.37	−0.39
MMPerspective	Qwen2.5-VL-3B	49	+0.08	+0.10	+0.06	+0.02	+0.12	+0.06	+0.02	+0.10	+0.02
MMPerspective	Qwen2.5-VL-7B	49	+0.10	+0.12	+0.16	+0.06	+0.12	+0.16	+0.10	+0.12	+0.14
MMPerspective	LLaVA-OV-7B	49	+0.04	+0.04	+0.04	+0.06	+0.06	+0.06	+0.08	+0.04	+0.10
MMPerspective	Idefics3-8B	49	+0.10	−0.31	−0.22	+0.08	−0.41	−0.27	+0.14	−0.37	−0.27

Evidence extent. One graduate-student annotator marked, for the MMStar and MMPerspective items and with the primary box shown, whether the evidence needed for the answer is a single region, several disjoint regions, or the global layout, and whether the primary box captures the most important region (Table A17).

Table A17: Evidence extent and adequacy of the primary box, per source (100 items each).

	Evidence extent			Primary box adequate		
Source	single	several	global	yes	partly	no
MMStar	76	14	10	81	17	2
MMPerspective	6	8	86	87	13	0

L BASE IMAGES AND EDITED COUNTERPARTS OF BLINK-TWICE

BLINK-Twice provides base images, to which its 693 questions and labels refer, and edited counterparts for 103 of them that depict the misleading appearance of the base image as fact; upstream describes each pair as visually similar with opposite factual content. For the 100 audited items the counterpart is scored against the flipped label. Two checks support that label. A first-glance answer recorded by a human on the counterpart, before judging the label, equals the flipped label on 92 of 100 items. The four models, which score 0.27–0.39 on the base images under the official label, score 0.75–0.86 on the counterparts under the flipped label.

Table A18 compares the profile on the two images. The interventions on each image are built from the box annotated on that image: the counterpart boxes were drawn on the counterparts and the base boxes on the base images, by different annotation passes. In relative coordinates the two boxes of an item overlap with mean IoU 0.51 (median 0.55), and the counterpart boxes are larger (median area 0.45 versus 0.32 of the frame). On the counterpart the profile is MMStar's: masking lowers accuracy, the off-evidence control is near zero, and the crop recovers. On the base image the same mask raises accuracy toward the occluded-image answer and the crop stays at the original level. On the 100 audited items the models give the same original answer on the base image and the counterpart on 71–82% of items and the same masked answer on 71–87%; across the 207 questions with a counterpart, a model gives the same original answer on 66–77% of items and answers both correctly, each under its own label, on 17–29%. On the 92 items whose flipped label the first-glance check confirms, ES_e is $+0.37$ [$+0.23$, $+0.50$] / $+0.51$ [$+0.39$, $+0.63$] / $+0.45$ [$+0.33$, $+0.57$] / $+0.46$ [$+0.33$, $+0.58$], ES_r within 0.03 of zero, and crop recovery $+0.40$ to $+0.52$; the eight disputed items do not carry the effect.

An earlier build of this audit attached the counterparts to the base labels; the panel showed $\kappa < 0$ in the original view for every model, and an item review located the mismatch. We report the case because accuracy alone would not have exposed it.

Table A18: Profile of the same 100 questions on the base image (official label) and on the edited counterpart (flipped label). Mask, control, and crop are generated by the same procedure from the box annotated on each image.

	Base image, official label					Edited counterpart, flipped label				
Model	Acc^o	Acc^e	ES_e	ES_r	$\mathrm{Acc}^p - \mathrm{Acc}^e$	Acc^o	Acc^e	ES_e	ES_r	$\mathrm{Acc}^p - \mathrm{Acc}^e$
Qwen2.5-VL-3B	0.39	0.72	-0.33	$+0.02$	-0.35	0.75	0.39	$+0.36$	-0.02	$+0.39$
Qwen2.5-VL-7B	0.37	0.74	-0.37	$+0.01$	-0.29	0.86	0.37	$+0.49$	0.00	$+0.49$
LLaVA-OV-7B	0.27	0.74	-0.47	-0.07	-0.47	0.81	0.41	$+0.40$	$+0.01$	$+0.40$
Idefics3-8B	0.36	0.68	-0.32	$+0.03$	-0.33	0.84	0.41	$+0.43$	0.00	$+0.41$

M PROMPT AND ANSWER EXTRACTION

The prompt is identical across models and image views:

```
Answer the multiple-choice question using the image.
Question:
<question>
Choices:
```

```
A. <option>
B. <option>...
Return only one option letter:  A, B, C, or D.
```

The question-only view drops the words "using the image". A control variant lists only the letters the item has ("A or B" for binary items), and a second control permutes the option order with the gold remapped; each changes BLINK-Twice original accuracy by at most 0.02 (Table A3 reports the permuted-order accuracy for the other sources). Decoding is greedy with four new tokens in half precision; Qwen2.5-VL inputs are resized to at most 384×384 pixels. The extractor returns the first option letter in the response; a letter outside the item's options, or no letter, is invalid and incorrect. Valid-output rates are 1.00 in 80 of 84 model$\times$view cells and at least 0.99 in the rest.

N RUN SETS AND ITEM-WISE AGREEMENT

The primary four-model evaluation comes from one run set produced with a single script, prompt, and decoding configuration; the full-benchmark run (693 base-image questions, original and question-only views), the counterpart run, the auditor-box runs, the polarity-flip run, and the 32B run are separate run sets with the same prompt and are recorded in the run manifest. An earlier, independent run of the o, e, and p views with the same prompt but different library versions provides a replication on MMStar and MMPerspective (Table A19); its BLINK-Twice rows are omitted because that run used the edited counterparts (Appendix L) and is not a replication of the base-image run. Predictions agree on 0.98–1.00 of original-view items for three models; for LLaVA-OneVision agreement is 0.72–0.73, with accuracy differences of up to 0.09, which we attribute to changes in the image preprocessing of the library. Within the run set, a view whose image is byte-identical to the evidence mask on 263 items reproduces the evidence-mask predictions on all of them.

Table A19: Item-wise agreement between the earlier run and the run set used in this paper: fraction of identical predictions, and accuracy in the earlier / current run.

		Original view			Evidence mask		
Model	Source	agree	Acc earlier	Acc current	agree	Acc earlier	Acc current
Qwen2.5-VL-3B	MMStar	1.00	0.60	0.60	0.91	0.39	0.36
Qwen2.5-VL-3B	MMPerspective	1.00	0.25	0.25	0.90	0.23	0.23
Qwen2.5-VL-7B	MMStar	0.99	0.60	0.60	0.92	0.30	0.31
Qwen2.5-VL-7B	MMPerspective	1.00	0.32	0.32	0.97	0.24	0.23
LLaVA-OV-7B	MMStar	0.72	0.56	0.64	0.69	0.37	0.36
LLaVA-OV-7B	MMPerspective	0.73	0.37	0.28	0.79	0.27	0.23
Idefics3-8B	MMStar	1.00	0.59	0.59	0.96	0.33	0.32
Idefics3-8B	MMPerspective	0.98	0.34	0.34	0.89	0.24	0.24