COUNTERFACTUAL AUDITING OF OBJECT PRESENCE AND QUANTITY JUDGMENTS IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) increasingly support systems that interpret visual content, making reliable evaluation essential. A basic capability is judging object presence, yet a correct answer on a static benchmark does not establish whether it relies on the relevant visual evidence. We introduce Counterfactual Object Visual Evidence (COVE), a paired-image benchmark that tests whether models update their answers when supporting object evidence is removed. Existence edits aim to remove the queried category, whereas quantity edits leave one instance, preserving category evidence while invalidating the at-least-two condition. The core datasets contain 160 existence pairs and 1,000 quantity pairs, complemented by 906 irrelevant-region existence controls. COVE measures evidence persistence: affirmative answers on both the original and target-edited image, stratified by intervention quality. Across six open VLMs, all-pair persistence ranges from 2.9–35.3% on 102 visually screened existence pairs and from 50.2–95.1% on a separate cohort of 329 metadata-screened two-to-one quantity pairs. More than 97% of eligible original-affirmative responses remain affirmative under irrelevant existence edits. Substantial count persistence recurs across prompts and editors; separate visual-audit aggregates provide a descriptive cross-check. These profiles show that low existence persistence can coexist with substantial difficulty updating quantity judgments. By combining instance-level removal records, quality strata, and paired response metrics, COVE provides a reusable resource for evaluation beyond static correctness.

1 INTRODUCTION

Vision-language models (VLMs) support image captioning, visual question answering, and multimodal assistants (Li et al., 2023a; Liu et al., 2023). Reliable evaluation must distinguish their apparent capabilities from unsupported outputs. Judging whether an object is present is central to hallucination evaluation, which treats affirmations of absent objects as false positives (Rohrbach et al., 2018; Li et al., 2023b; Fu et al., 2023). Yet a correct positive answer need not rely on the relevant evidence: it may reflect scene context or learned priors.

Resolving this ambiguity requires changing the evidence. Object removal asks the same question before and after target removal while retaining the surrounding scene (Moon et al., 2024; He et al., 2025). Category removal tests responses when the category disappears, but not quantity updating while it remains visible. Recent counterfactual counting benchmarks also study edited quantities (Vo et al., 2026; Alzahrani et al., 2026). COVE complements these approaches by organizing whole-instance removal around the nested requirements of presence and plurality, with explicit retained/removed instance records and edit-validity checks.

We introduce COVE (Counterfactual Object Visual Evidence), a paired-image benchmark for auditing object presence and quantity judgments (Fig. 1). Presence, termed *existence* in our evaluation, and the minimal quantity judgment of *at least two* form a natural hierarchy: one instance establishes presence; a second establishes plurality. Reducing multiple instances to one invalidates quantity while preserving category evidence and the binary answer format. *Existence-COVE* aims to remove every instance of the queried category, whereas *Count-COVE* retains one. Core resources contain

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.

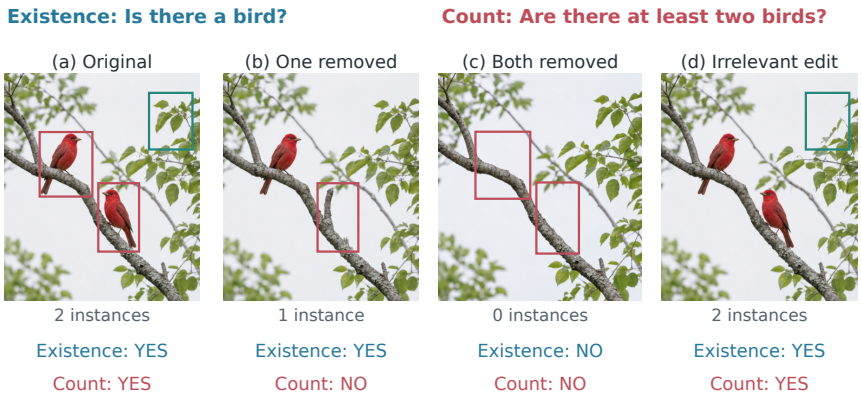

Figure 1: **Category evidence is not quantity evidence.** Expected answers, not predictions: removing one instance invalidates quantity only; removing both invalidates existence too. Red boxes mark target locations; green boxes mark the irrelevant region. Generated illustrations; empirical irrelevant controls evaluate existence only.

160 existence pairs and 1,000 count pairs, with 906 additional existence controls that edit irrelevant regions. Edit records support reuse with alternative editors and quality screens.

Evidence persistence measures YES on both views despite an intended change to NO. Failed removal can leave the original answer valid, so COVE stratifies persistence by quality and denominator: visual screening checks existence removal; metadata screening checks count bookkeeping and geometry. Irrelevant-region controls test stability when queried evidence is preserved.

Across six open VLMs, all-pair persistence ranges from 2.9–35.3% on visually screened existence pairs and from 50.2–95.1% on metadata-screened exact-two-to-one count pairs. Larger models reduce but do not eliminate count persistence within the tested families. Prompt and editor replications retain substantial persistence; separate visual-audit aggregates provide a descriptive cross-check. More than 97% of eligible original-affirmative existence responses remain affirmative under irrelevant edits. Low existence persistence thus coexists with difficulty updating quantity. Different cohorts and quality screens make this a comparison of diagnostic profiles, not a causal effect of changing only the question.

Our main contributions are:

- **A presence–quantity audit.** Nested evidence requirements motivate complementary category-removal and category-preserving quantity tests.
- **A reusable counterfactual resource and protocol.** COVE provides instance records, quality strata, explicit persistence denominators, and matched existence controls.
- **An empirical quantity-updating diagnostic.** Across six open VLMs, substantial screened two-to-one persistence coexists with low existence persistence, supported by prompt/editor replications and audit summaries.

2 Related Work

Hallucination and visual-evidence evaluation. CHAIR measures unsupported caption mentions (Rohrbach et al., 2018); POPE tests existence (Li et al., 2023b); MME includes presence and counting (Fu et al., 2023). MMHal-Bench and AMBER cover open-ended responses and hallucination types (Sun et al., 2024; Wang et al., 2024a); ODE covers open-set concepts (Tu et al., 2025). Static correctness does not establish whether an answer survives removal of its support. HallusionBench uses edited images and question pairs to diagnose hallucination and illusion (Guan et al., 2024). COVE instead separates presence and plurality evidence through whole-instance removal.

Interventional evaluation. BEAF tests transitions for removed and unremoved objects (Moon et al., 2024); ROHE curates removed-object VQA (He et al., 2025); VLind-Bench probes language priors with counterfactuals (Lee et al., 2025); Causal-HalBench targets co-occurrence bias (Xu et al., 2026). ReactBench covers edited relations, attributes, alteration tracing, and dense counting, with human-verified images and answers (Zhou et al., 2026). COVE builds on these interventions. Its instance records distinguish category removal from removal that leaves one category witness, while quality strata expose whether persistence survives a stricter validity screen (Table 1). Common-row BEAF/ROHE-inspired summaries in the appendix explain metric relationships, not performance on the original benchmarks.

Counterfactual counting. VLMBias alters familiar patterns and counts to test whether visible quantities override learned priors (Vo et al., 2026). CounterCount pairs factual and counterfactual attribute counts with localized evidence and studies attention-based corrections (Alzahrani et al., 2026). COVE crosses the minimal plurality boundary by removing whole instances until one remains: existence stays true, but at-least-two becomes false. Category-removal tests, reusable instance records, explicit denominators, and quality strata expose both evidence requirements in one binary-response protocol.

Table 1: **Intervention protocols and COVE: evaluation scope, not ranking.**

Benchmark	Established component	COVE comparison
BEAF	Removed/unremoved transitions	Original-yes conditioning; quality strata
ROHE	Removed-object VQA; curation	Raw vs visually screened rates
ReactBench	Edited tasks, including counting	At-least-two test under instance removal
CounterCount	Attribute/part-count edits	Whole-instance removal; category retained

3 THE COVE BENCHMARK

COVE separates three questions: what evidence makes a judgment true, whether an edit changes that truth, and whether the model updates its answer. We formalize two nested predicates, construct three complementary pair families, and define quality-stratified persistence and invariance metrics.

3.1 WHY QUANTITY, AND WHY AT LEAST TWO?

A minimal test beyond category recognition. Existence requires one visible witness; an at-least-two judgment requires a distinct second instance. This is the smallest quantity threshold that can change while category presence is preserved. After a valid two-to-one edit, treating category presence as sufficient for plurality would yield YES: a possible shortcut, not an established mechanism. The binary threshold shares existence's answer space without adding exact-numeral generation or large-count difficulty. It tests a necessary component of counting, not general counting ability.

Predicates and intended truth transitions. Let I_i be source image i, c_i its queried category, and $a_i \in \{\text{exist}, \text{count}\}$ the predicate type. Write $n(I, c)$ for the number of visually distinct instances of category c in image I. The binary truth functions are

$$\tau_{\text{exist}}(I, c) = \mathbb{1}\{n(I, c) \geq 1\}, \qquad \tau_{\text{count}}(I, c) = \mathbb{1}\{n(I, c) \geq 2\}, \tag{1}$$

where $\mathbb{1}\{\cdot\}$ is the indicator of a true condition. The fixed question $q_i = q_{a_i}(c_i)$ asks either *Is there a category in the image?* or *Are there at least two category objects in the image?* An original is selected to have intended truth $\tau_{a_i}(I_i, c_i) = 1$. The target transformation $T_{\text{tar},i}$ produces $I_i^{\text{tar}} = T_{\text{tar},i}(I_i)$: it aims to leave zero queried instances for existence and exactly one for count, giving intended truth zero. A control transformation $T_{\text{irr},i}$ instead preserves the queried evidence, producing I_i^{irr} with intended truth one. In this study, controls are evaluated for existence only.

Target tests expect YES/NO; irrelevant tests expect YES/YES. Continued YES on a target edit is persistence. Annotation counts specify intended transitions, but missed instances or residual parts can invalidate them, requiring quality checks. Count construction starts with two or more instances; the strict subset isolates two-to-one transitions.

Table 2: **Resource inventory before quality screening.** Each set has one pair per source image. Images recur across sets, so rows must not be summed as unique images. Strict count is nested within proxy-clean.

Resource	Pairs	Categories	Evaluation role
Existence core	160	5	Visual audit; 102 clean pairs
Count core	1,000	77	572 proxy-clean; 329 strict pairs
Irrelevant existence controls	906	78	Matched to the 942-pair existence set
Existence scale set A	942	78	Raw scale replication
Existence scale set B	1,711	79	Raw scale replication

3.2 Three Pair Families and Their Quality Records

To separate target sensitivity from response stability, we construct **category-removal pairs**, **category-preserving quantity pairs**, and **irrelevant-region control pairs** from COCO instance annotations (Lin et al., 2014). Annotation-guided batch editing avoids manually specifying every removal; no new generator is trained. Default mean-box edits fill target rectangles with an image-derived uniform color. Alternative editors change the fill, not the question.

Category-removal pairs. Editing all recorded boxes of the queried category aims to remove its evidence. The 160-pair core supports visual audit; the 942- and 1,711-pair sets test scale and broader coverage (Table 2). Intended labels do not depend on model answers.

Category-preserving quantity pairs. Among eligible non-crowd instances of a category, we retain the largest bounding box and remove the others. This deterministic rule preserves substantial category evidence without guaranteeing model recognition. Records contain retained/removed instance identifiers, boxes, eligible original count, and intended remaining count of one. Thus existence remains true while at-least-two becomes false; other editors can reuse the same test.

Irrelevant-region control pairs. We edit a non-target object with negligible target overlap and area close to the target region, falling back to a background block. The 906 controls (684 object and 222 background edits) link to the *same* image, category, and question in the 942-pair target set. Matching distinguishes evidence-specific retraction from generic editing sensitivity; region thresholds appear in Appendix A.2.

Why the source cohorts differ. Count requires multiple same-category instances and prioritizes category coverage; the existence core prioritizes detailed visual audit. They share 52 images but only 16 image–category pairs, so are not a matched task comparison. Shared rows support *within-task* model comparisons and *within-image* target/control tests. Cross-task profiles are descriptive: source composition and screening can contribute to their gap. Count's key result is persistence after its own truth-changing edit, not merely a higher rate than existence.

Quality strata: validity before behavior. Raw rates include completed original/edited pairs. For existence, the **visual-clean** event $A_i = 1$ requires a visible original target, successful removal, no residual target evidence, and good or minor-artifact edit quality. The archived audit mask retains 102 of 160 pairs. For count, **proxy-clean** requires consistent instance bookkeeping, a materialized edit, at most four eligible original instances, low retained/removed overlap, limited edited area, and a sufficiently large retained instance; it retains 572 of 1,000 pairs. **Strict** tightens the geometry limits and requires exactly two original instances, retaining 329. Exact thresholds are in Appendix A.2. These screens protect the retained instance and constrain edit extent, but cannot exclude unannotated instances, residual appearance, or grouping ambiguity. They are metadata screens, not human visual certification.

Separate visual-audit aggregates cover 100 count panels each for Qwen2-VL-7B, InternVL2-8B, and GLM-4V-Plus-0111, checking visible counts, residuals, quality, and grouping ambiguity. These cross-checks do not define the shared main mask; Appendix A.3 gives their provenance. Alternative editors—blur, mask-mean, mask inpainting, and LaMa—test editor sensitivity, not presumed quality rankings.

Reusable records and comparison consistency. Stable identifiers join images, instance regions, questions, editor settings, quality fields, answers, and request status. Models share edits; records expose coverage without claiming balanced or representative sampling. Separating construction, validation, and predictions supports new editors and quality masks. Reuse requires upstream images, which are not redistributed; image-level splits must account for cross-manifest reuse.

3.3 Metrics: Persistence, Quality, and Specificity

Response representation. Let f denote a VLM with fixed decoding and π its deterministic parser. For view $v \in \{\text{orig}, \text{tar}, \text{irr}\}$, define

$$Z^v_{f,i} = \pi(f(I^v_i, q_i)) \in \{\text{YES}, \text{NO}, \text{other}\}, \qquad Y^v_{f,i} = \mathbb{1}\{Z^v_{f,i} = \text{YES}\}, \tag{2}$$

where $I^{\text{orig}}_i = I_i$. Y records affirmation, not confidence or correctness. Views are queried independently with the same question and wrapper, not together in a conversation. Unrecognized answers remain *other*.

Two denominators answer different questions. For model f, predicate a, and quality stratum g, let $\mathcal{D}_{f,a,g}$ contain completed pairs and let $N_{f,a,g} = |\mathcal{D}_{f,a,g}|$. Persistence requires two affirmative answers, $P_{f,i} = Y^{\text{orig}}_{f,i} Y^{\text{tar}}_{f,i}$. We report

$$\widehat{\text{EP}}^{\text{all}}_{f,a,g} = \frac{\sum_{i \in \mathcal{D}_{f,a,g}} P_{f,i}}{N_{f,a,g}}, \tag{3}$$

$$\widehat{\text{EP}}^{+}_{f,a,g} = \frac{\sum_{i \in \mathcal{D}_{f,a,g}} P_{f,i}}{\sum_{i \in \mathcal{D}_{f,a,g}} Y^{\text{orig}}_{f,i}}, \qquad r^{+}_{f,a,g} = \frac{\sum_{i \in \mathcal{D}_{f,a,g}} Y^{\text{orig}}_{f,i}}{N_{f,a,g}}. \tag{4}$$

EP^{all} measures the prevalence of yes–yes outcomes among all evaluated pairs; EP^{+} measures failure to retract *among original affirmative answers*. Because $\widehat{\text{EP}^{\text{all}}} = r^{+}\widehat{\text{EP}^{+}}$, rarely affirming originals can lower EP^{all} without reliable updating. We report r^{+} and denominators; zero original-yes denominators are undefined. Indices are omitted only when fixed by the text or heading.

Quality conditioning. For visually clean existence edits the conditional estimand is

$$\text{EP}^{+}_{f,\text{exist,clean}} = \Pr(Y^{\text{tar}}_f = 1 \mid Y^{\text{orig}}_f = 1, A = 1), \tag{5}$$

where $A = 1$ is the clean-intervention event. Quality strata select different populations; screening is not model improvement. Unlike static false-positive rate (FPR) $\phi_f = \Pr(Y^{\text{orig}}_f = 1 \mid \tau_a(I, c) = 0)$, persistence starts from an affirmed positive image and removes support. Correlated rankings need not imply equal magnitudes.

Two-sided fidelity. Sensitivity to target removal alone rewards a model that retracts after any edit. On a nonempty set of shared original-yes existence triplets $\mathcal{C}_f$, define the *irrelevant-invariance score* (IIS) and two fidelity components:

$$\widehat{\text{IIS}}_f = \frac{1}{|\mathcal{C}_f|} \sum_{i \in \mathcal{C}_f} Y^{\text{irr}}_{f,i}, \qquad (S_{1,f}, S_{2,f}) = (1 - \widehat{\text{EP}^{+}}_f, \widehat{\text{IIS}}_f). \tag{6}$$

Here S_1 is target sensitivity and S_2 irrelevant invariance. Both use the same rows; if target/control runs repeat original inference, eligibility requires original YES in both. A constant-YES model gives $(0, 1)$, and one that retracts after every edit gives $(1, 0)$; the desired behavior is $(1, 1)$. The *evidence-dependence score*, $\text{EDS} = 2S_1S_2/(S_1+S_2)$, summarizes the two components and is zero when both are zero. It does not replace them or identify an internal causal mechanism. When a target answer is *other*, S_1 records non-affirmation, not a verified correct NO; such cases are reported separately.

4 Experiments

We compare quality-screened presence and quantity profiles, validate the interventions, test robustness, and analyze correlates of count persistence.

4.1 Experimental Setup

Models and data. The primary suite comprises BLIP-VQA (Li et al., 2022), Qwen2-VL-2B/7B (Wang et al., 2024b), InternVL2-2B/8B (Chen et al., 2024), and MiniCPM-V-2.6 (Yao et al., 2024). Qwen2.5-VL-3B and InternVL3-2B extend POPE calibration (Qwen Team, 2025; Zhu et al., 2025). GLM-4.6V is the API-served replication model; GLM-4V-Plus-0111 contributes only the separate count-audit aggregates. Endpoint versions are not pooled. The primary strata are the 102 visual-clean existence, 572 proxy-clean count, and 329 strict count pairs (Table 2); external tests use VOC2012 and nuImages (Lin et al., 2014; Everingham et al., 2010; Caesar et al., 2020). Models share rows within each split, but existence and count use separate image cohorts and quality screens.

Inference and statistics. Open models use deterministic decoding and fixed questions/wrappers across views; BLIP's VQA adapter receives bare questions. Refusals and parser failures remain *other* in raw/prompt denominators, never NO. Missing images and failed requests are excluded with status records. We report 95% intervals: 3,000-resample bootstrap percentiles for open-model split rates, and Wilson intervals for aggregate audits, API runs, and prompt rates. Shared-row model comparisons use Cochran's Q test followed by pairwise exact McNemar tests with Holm correction. Target/control and editor contrasts use exact McNemar tests (Holm correction over 24 editor contrasts). Regression covariance is image-clustered; probe intervals use paired bootstrap with Holm-corrected exact sign tests over 108 layers. Separate existence/count cohorts are never tested as paired.

4.2 Main Results: Presence and Quantity Updating

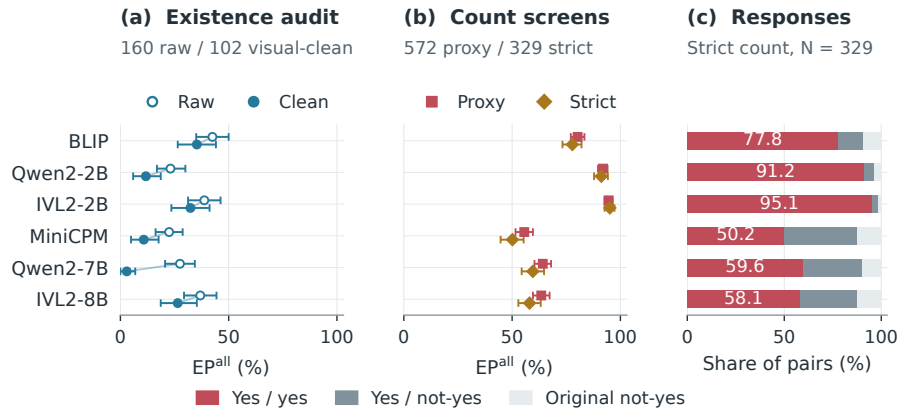

Figure 2: **Persistence under quality screening.** $\mathrm{EP}^{\mathrm{all}}$ for (a) existence and (b) metadata-screened count, with 95% bootstrap intervals. (c) Strict-count responses across 329 pairs. "Not-yes" need not be parsed NO. Separate cohorts preclude a paired predicate contrast.

Clean existence $\mathrm{EP}^{\mathrm{all}}$ ranges from 2.9–35.3%, versus 55.6–94.6% for proxy-clean count and 50.2–95.1% for strict count (Table 3). High original affirmative rates in both cohorts show that count persistence reflects frequent yes–yes responses, not predominantly failure on the original positive image.

Qwen2-VL-7B affirms all 102 clean existence originals and persists on only three removals, yet has 59.6% strict count $\mathrm{EP}^{\mathrm{all}}$. InternVL2-2B remains high on both (32.4% and 95.1%; Fig. 2). These profiles support predicate-specific reporting, not a causal predicate effect, because image content and quality screens differ.

Larger models improve within the tested families without eliminating count persistence. Qwen2-VL proxy $\mathrm{EP}^{\mathrm{all}}$ drops from 92.0% at 2B to 64.2% at 7B; InternVL2 drops from 94.6% at 2B to 63.5% at 8B. Shared-row differences across six models are significant for clean existence (Cochran's $Q(5) = 76.0$, $p < 10^{-14}$) and proxy count ($Q(5) = 512.7$, $p < 10^{-100}$). Respectively 9 and 13 of the 15 pairwise exact McNemar tests survive Holm correction. The Qwen2-VL-7B/InternVL2-8B count difference is not significant (adjusted $p = 0.811$); failure to detect a difference is not evidence of equivalence.

Table 3: **Quality-screened existence and count results (%).** r^+ is the model's original affirmative rate. All persistence columns report $\mathrm{EP}^{\mathrm{all}}$ with 95% bootstrap intervals. Existence is visually screened; count proxy and strict are metadata-screened, not human-clean. Separate cohorts preclude interpreting their contrast as a paired predicate effect.

	Existence clean ($N = 102$)		Count proxy ($N = 572$)		Count strict ($N = 329$)
Model	r^+	$\mathrm{EP}^{\mathrm{all}}$ [95% CI]	r^+	$\mathrm{EP}^{\mathrm{all}}$ [95% CI]	$\mathrm{EP}^{\mathrm{all}}$ [95% CI]
BLIP-VQA	90.2	35.3 [26.5, 44.1]	91.8	80.2 [77.1, 83.4]	77.8 [73.3, 82.1]
Qwen2-VL-2B	94.1	11.8 [5.9, 18.6]	97.0	92.0 [89.7, 94.1]	91.2 [87.8, 94.2]
InternVL2-2B	95.1	32.4 [23.5, 41.2]	98.1	94.6 [92.7, 96.3]	95.1 [92.7, 97.3]
MiniCPM-V-2.6	94.1	10.8 [4.9, 17.6]	90.9	55.6 [51.6, 59.6]	50.2 [44.7, 55.3]
Qwen2-VL-7B	100.0	**2.9** [0.0, 6.9]	92.3	64.2 [60.3, 68.0]	59.6 [54.4, 64.7]
InternVL2-8B	94.1	26.5 [18.6, 35.3]	90.7	63.5 [59.6, 67.3]	58.1 [52.9, 63.2]

4.3 Validating the Main Results

Intervention audit changes what persistence means. On all 160 existence pairs, Qwen2-VL-7B has 27.5% raw $\mathrm{EP}^{\mathrm{all}}$; on the shared 102 clean edits it has 2.9% [0.0, 6.9]. BLIP-VQA changes from 42.5% to 35.3%, while InternVL2-2B and 8B retain 32.4% and 26.5%. Thus raw persistence can overstate failure after valid removal to different extents across models. Screening changes the subset, not the model. Table 5 reports all raw and clean values.

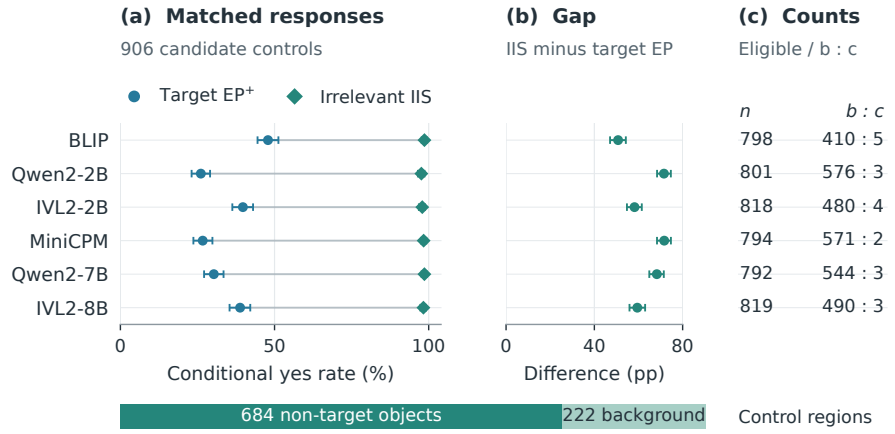

Figure 3: **Existence-specific response to removal.** (a) Matched target EP^+ and IIS; (b) their gap with 95% bootstrap intervals; (c) eligible denominators and discordant counts: b is target not-yes/control yes; c is the reverse. Models use the same eligible rows for each contrast. All exact McNemar tests give $p < 1.2 \times 10^{-113}$.

The author-reported 300-panel count audit supplies 88, 91, and 91 clean panels for Qwen2-VL-7B, InternVL2-8B, and GLM-4V-Plus-0111. Its eligible-subset persistence ratios are 50.0%, 43.0%, and 46.0% (Table 7). These model-specific aggregates cross-check, but do not replace, the shared metadata-screened cohorts. Their supplied eligibility differs from the original-yes denominator (Appendix A.3).

Irrelevant edits preserve existence answers. On matched controls, IIS is 97.6–98.6%, whereas target-removal EP^+ is 26.1–47.9% (Fig. 3). The resulting specificity gaps are 50.8–71.7 percentage points; all exact McNemar tests give $p < 1.2 \times 10^{-113}$. Target updates are therefore not indiscriminate retraction after localized edits. These existence controls use a larger cohort than the 102 clean edits; they neither establish count-specific invariance nor certify target-removal quality. IIS's complement is non-affirmation, not necessarily NO. Table 6 retains the two fidelity components and EDS.

4.4 Robustness and Generalization

Editors. The matched five-editor matrix evaluates mean-box, blur, mask-mean, mask-inpaint, and LaMa on 160 pairs per predicate for three models: 4,800 model–example evaluations, not 4,800 unique images (Table 11). Existence EP^+ ranges from 26.1% to 56.2%, and count from 59.6% to 98.1%. Count exceeds existence in every model–editor cell. Relative to mean-box, LaMa increases count persistence by 4.4–11.6 points; Holm-adjusted p-values are 0.0349 for Qwen2-VL-7B, 0.0347 for MiniCPM, and 0.1563 for InternVL2-2B. The ordering persists across editors, but no editor-specific visual audit establishes artifact elimination.

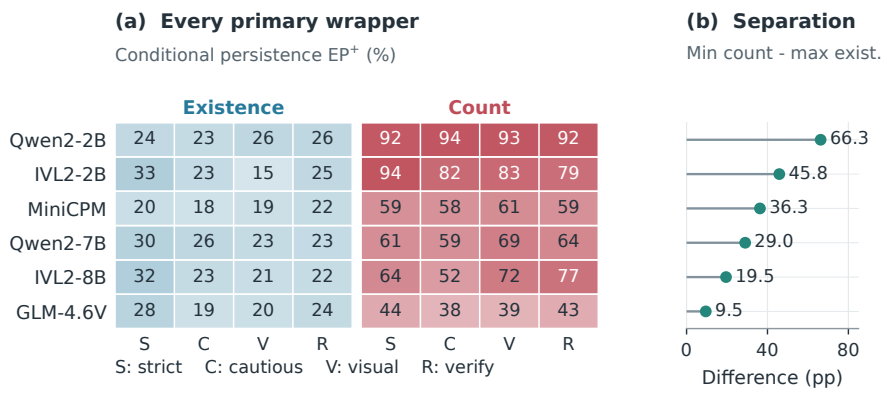

Figure 4: **Prompt robustness.** (a) All 48 primary cells on 0–100% scales. (b) Minimum count minus maximum existence EP^+, using unrounded rates. Gaps are descriptive: predicates use separate 200-image cohorts. Primary wrappers have at most 5% unrecognized answers; bare prompts are excluded. BLIP is not a wrapper sweep.

Prompts and parsing. Six instruction-following models receive five wrappers. The four primary wrappers (strict, cautious, visual-inspection, verification) have at most 5% unrecognized answers. Each model's minimum count EP^+ exceeds its maximum existence EP^+ (Fig. 4; Table 10). Bare prompts are a separate format stress test. BLIP receives identical questions across wrapper labels, so its duplicates do not test robustness. Parser coverage measures binary recognition, not semantic accuracy on ambiguous free-form answers.

Scale, domains, and API replication. Existence raw EP^{all} remains similar from 942 to 1,711 pairs (Qwen2-VL-7B: 26.1% to 25.9%; BLIP: 41.9% to 39.3%). External domains replicate existence persistence, not count (Fig. 7). API-served GLM-4.6V yields 14.6% clean existence EP^+, 47.8% count EP^+ on 999 valid pairs, and 99.1% IIS on 905 controls (Fig. 13; Table 15).

Calibration rather than metric independence. Clean existence EP^{all} and POPE FPR correlate across eight models ($\rho = 0.905$, permutation $p = 0.005$; Fig. 12). COVE adds intervention response and quality conditioning, not independent rankings. Appendices F–G provide BEAF/ROHE-inspired metric comparisons and verification transfer diagnostics, not official benchmark reruns or a correction-method contribution.

4.5 Understanding Count Persistence

Geometry and instance count. On 572 proxy-clean pairs, pooling the six models gives $EP^{all} = 81.0\%$ in the smallest retained-object area tertile and 66.0% in the largest. Controlling for model and edit geometry, larger retained area predicts lower persistence (odds ratio 0.81 per standardized log-area unit, pointwise 95% CI [0.68, 0.97], covariance clustered by image). Exact-two originals yield 72.0%, versus 81.2% for three or four (Fig. 10a,b). These associations identify correlates, not causes; the appendix reports the full regression.

Exploratory representation probe. A probe pairs 20 Qwen2-VL-7B-flip/InternVL2-2B-persist examples with 20 geometry-matched both-persist examples. The *same at-least-two question* along 2-

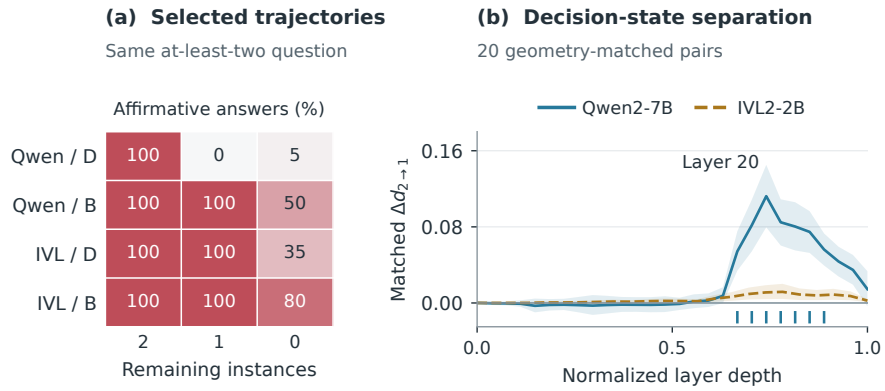

Figure 5: **Matched count probe.** (a) The at-least-two trajectory requires YES/NO/NO. D: disagreement; B: both-persist; 20 examples per row. (b) Decision-state separation; bands are pointwise 95% bootstrap intervals; ticks mark exact sign tests with Holm-adjusted $p < 0.05$ over 108 layers. The 20 behavior-selected pairs do not establish causality.

to-1-to-0 requires YES/NO/NO. Hidden-state cosine distances measure original-to-edited change; Δd is the mean paired difference between disagreement and both-persist strata. Qwen's largest decision-state effect occurs at layer 20 ($\Delta d = 0.1122$, 95% CI [0.0795, 0.1452], exact sign $p = 0.000402$, Holm-adjusted $p = 0.0427$ over 108 layers); seven decision layers survive correction. Qwen's vision peak and both InternVL peaks are nonsignificant. The late decision-state association (Fig. 5) is exploratory: behavioral selection and 20 matched pairs preclude a causal mechanism claim.

5 Discussion

Interpreting count persistence. Successful category-removal responses do not ensure reliable plurality judgments. Count persistence exposes this limitation without establishing that a model uses only one instance. Geometry and representation analyses identify correlates; residual evidence, instance grouping, and model-specific processing remain competing explanations.

Scope and limitations. COVE tests existence and a binary at-least-two predicate, not unrestricted counting. Separately constructed cohorts and different quality screens make their contrast descriptive, not a causal predicate effect. Controls establish specificity for existence; metadata-count screens cannot exclude all residual evidence or grouping ambiguity. The existence core covers five categories; larger scale sets broaden coverage but do not inherit its visual audit. Count-audit aggregates do not support reconstructing row-level agreement or exclusion decisions (Appendix A.3). One API provider and a behavior-selected probe limit generalization.

From diagnosis to supervision. COVE's reusable intervention records support new editors, audits, and model comparisons. Validated target/control pairs could jointly supervise retraction and invariance (Appendix H.4). Image-disjoint evaluation with held-out categories and editors would distinguish learned evidence dependence from editing cues. This is a prospective use, not a training result.

6 Conclusion

COVE audits nested presence and plurality requirements through category removal and category-preserving quantity edits. Across the evaluated cohorts, low visually screened existence persistence coexists with substantial metadata-screened two-to-one persistence. Instance records, quality strata, explicit denominators, and matched existence controls provide a reusable resource for evaluation beyond static correctness.

AI USE STATEMENT

AI tools assisted with literature discovery, research and experiment planning, implementation and analysis scripts, figure prototyping, and manuscript drafting, revision, and translation. Reported numerical results were checked against experimental records and executable analyses; citations were checked against primary sources. Generated illustrations are explanatory rather than benchmark observations. Audit provenance is reported in Appendix A.3. The authors are responsible for the final text, claims, results, and artifacts.

ETHICS STATEMENT

The benchmark uses established image datasets and evaluates model behavior; it does not collect new personal or sensitive participant data. Upstream images are not redistributed, and users must obtain them under their original terms. Counterfactual edits and model outputs may inherit biases from datasets and models. COVE is intended as a diagnostic rather than a basis for consequential decisions about individuals.

REPRODUCIBILITY STATEMENT

Section 3.2 inventories the data resources. The anonymous supplement provides construction rules, prompts, parser and failure policies, denominators, audit rubrics, confidence intervals, paired tests, and ablations. The accompanying anonymous archive contains portable JSONL manifests, evaluation scripts, source CSV/JSON tables, and a dataset card (Appendix H). The original existence audit mask supports the shared clean cohort, while the separate count visual audit is represented by author-reported aggregates. Appendix A.3 specifies which audit statistics are reconstructible from those records. Upstream images and model weights are referenced rather than redistributed.

REFERENCES

Reem Alzahrani, Hassan Alshanqiti, Bushra Bin Hemid, Zaid Alyafeai, Abdelrahman Eldesokey, and Bernard Ghanem. CounterCount: A diagnostic framework for counting bias in vision language models. *arXiv preprint arXiv:2605.17826*, 2026. doi: 10.48550/arXiv.2605.17826.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Hanqing Wang, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Yixiao He, Haifeng Sun, Pengfei Ren, Jingyu Wang, Huazheng Wang, Qi Qi, Zirui Zhuang, and Jing Wang. Evaluating and mitigating object hallucination in large vision-language models: Can they still see removed objects? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6841–6858, 2025. doi: 10.18653/v1/2025.naacl-long.349.

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. doi: 10.18653/v1/2025.findings-naacl.231.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023b. doi: 10.18653/v1/2023.emnlp-main.20.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Ye-Bin Moon, Hyeon-Woo Nam, Wonseok Choi, and Tae-Hyun Oh. BEAF: Observing BEfore-AFter changes to evaluate hallucination in vision-language models. In *European Conference on Computer Vision*, 2024.

Qwen Team. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018. doi: 10.18653/v1/D18-1437.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024. doi: 10.18653/v1/2024.findings-acl.775.

Yahan Tu, Rui Hu, and Jitao Sang. ODE: Open-set evaluation of hallucinations in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19836–19845, 2025.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.23941.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Zhe Xu, Zhicai Wang, Junkang Wu, Jinda Lu, and Xiang Wang. Causal-HalBench: Uncovering LVLMs object hallucinations through causal intervention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34169–34177, 2026. doi: 10.1609/aaai.v40i40.40712.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Yidong Chen, Hao Zhou, Zhensheng Zou, Haokun Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Shizhe Zhou, Bohan Jia, Kai Wu, Yan Shen, Tongyun Li, Yuyang Wu, and Shaohui Lin. ReactBench: A cause-driven benchmark for multimodal hallucination via systematic evaluation. *arXiv preprint arXiv:2605.29579*, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yiming Duan, Weijie He, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A BENCHMARK DETAILS

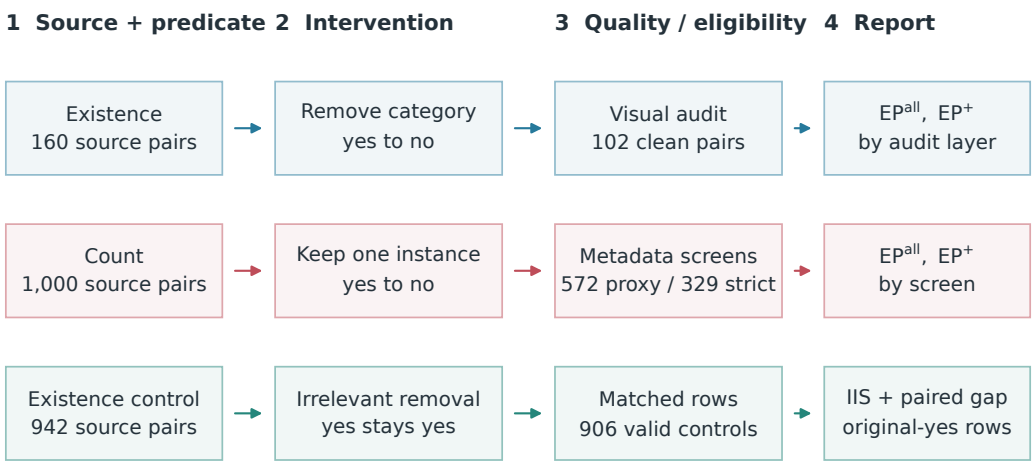

Figure 6: **Three traceable construction paths.** Existence removal, count reduction, and irrelevant-region controls have different validity checks and denominators. Each original/edited pair receives the same question. The middle column gives intended label transitions, not model predictions; quality screens do not guarantee those transitions visually. Counts identify evaluated resources, not a shared dual-predicate cohort. Records retain raw answers, parser status, edit regions, and quality labels.

A.1 PROMPTS AND PARSING

Existence-COVE uses the base question "Is there a *category* in the image?" Count-COVE uses "Are there at least two *category* objects in the image?" Within a model, the question wrapper is reused across views. The strict wrapper requests exactly one word; cautious adds an evidence-sufficiency instruction; visual asks the model to use only visibly present content; verify asks it to check the quoted question against the image; and bare supplies no response-format instruction. The first four form the primary prompt-robustness set for six instruction-following models. BLIP's VQA adapter passes only the base question regardless of the wrapper label: its ten archived cells are repeated fixed-input evaluations, not ten distinct prompts. Bare is a formatting stress condition for the instruction-following models.

Deterministic parsing maps explicit affirmative forms to YES, explicit denials to NO, and refusal, uncertainty, or unrecognized output to *other*. Evidence persistence requires an explicit YES on both the original and target-edited image. Parser failures, API failures, and content filters are stored separately. No *other* or failed request is recoded as NO.

Table 4: **Evaluation procedure.** Stage inputs and outputs share stable example identifiers.

1	Select an annotated image and predicate-valid target.
2	Query the original image and parse $Z_{f,i}^{\text{orig}}$ and $Y_{f,i}^{\text{orig}}$.
3	Materialize target edit $T_{\text{tar},i}(I_i)$; optionally materialize matched control $T_{\text{irr},i}(I_i)$.
4	Query each edited image with the unchanged prompt.
5	Store images, regions, answers, parser status, editor, and audit fields.
6	Report raw, clean/proxy-clean, conditional, and control rates with intervals.

A.2 TARGET AND CONTROL CONSTRUCTION

Existence targets are selected from instance-annotated COCO images. Their manifests specify the target-category boxes to be edited, with the intention of removing category evidence rather than only one arbitrarily selected instance. Any remaining visible target instance violates the clean-removal

criterion. Count construction groups eligible non-crowd annotations of the same category, retains the largest bounding box, and records the other instances as removal targets. Area filtering can omit small instances, so an annotation-derived remaining count of one is an intended label, not a visual guarantee. This is one reason to distinguish metadata screens from visual audits.

Composition and overlap. Each manifest in Table 2 contains one row per distinct source image within that set. The 160-pair existence core covers bowl, car, dining table, laptop, and oven; the 1,000-pair count core covers 77 categories. They share 52 source image identifiers, but were constructed separately and are not evaluated as a matched dual-predicate cohort. The 906 control rows link to corresponding target-removal rows in the 942-pair existence set: 684 edit non-target objects and 222 use background blocks. The 942- and 1,711-pair sets are not nested. Repeated source identifiers must therefore be resolved before any pooled analysis or train/test split.

Count quality screens. Proxy-clean first requires a materialized edit, at least two eligible original instances, exactly one intended remaining instance, and matching removed-instance identifiers and counts. It then requires at most four original instances, maximum retained/removed box intersection-over-union (IoU) below 0.05, summed removed-box area at most 35% of image area, and retained-box area at least 0.3%. Strict screening requires exactly two original instances, IoU below 0.02, removed area at most 25%, and retained area at least 0.5%. These rules yield 572 proxy-clean and 329 strict pairs. They test count bookkeeping and edit geometry, not residual appearance or visual grouping.

The irrelevant control first searches for a non-target COCO object that is not crowd-labeled, has bounding-box area at least 1,200 pixels, occupies 0.5–35% of the image, has target overlap IoU below 0.02, and has a center outside the target box. Among valid candidates, it chooses the area match closest to the target region. If no object qualifies, a background box of similar area is sampled with target IoU below 0.01. The random seed is 20260615. Failed control construction is logged rather than replaced silently.

A.3 Audit Rubrics

The 160-panel existence audit records target visibility in the original, target removal in the counterfactual, residual target evidence, edit quality, and a failure taxonomy. A pair is clean when the original target is visible, the target is removed, no residual evidence remains, and edit quality is good or minor. The released sheet is an author-verified audit mask, yielding the shared 102-pair clean subset. The authors separately report two-annotator Cohen's $\kappa = 0.862$.

The count rubric records: (i) at least two visible original instances; (ii) exactly one visible counterfactual instance; (iii) residual evidence of a second instance; (iv) edit quality; (v) grouping ambiguity; and (vi) the final clean-count decision. The authors report two annotators covering all 300 panels and Cohen's $\kappa = 0.831$. The released count records are model-specific aggregates for 100 panels each, rather than individual labels and adjudications.

Provenance and aggregate eligibility. The agreement values are author-reported statistics, not recomputed estimates from the released audit summaries. Those summaries support reconstruction of count persistence ratios and Wilson intervals, but not annotator agreement, sampling balance, or row-level exclusion decisions. For Qwen2-VL-7B, InternVL2-8B, and GLM-4V-Plus-0111, original-yes counts on the reported clean subsets are 87, 88, and 90, whereas the supplied eligible counts are 86, 86, and 87. The aggregates do not identify the excluded rows. Accordingly, Table 7 reports k/n_{elig} using the supplied persistent count k and eligible count n_{elig}; it does not substitute these denominators for the original-yes population in Equation 4. These data serve as descriptive supplementary evidence, separate from the shared cohorts and paired tests.

B FULL EXISTENCE RESULTS

Table 5: **Audit-conditioned Existence-COVE results.** All persistence columns are unconditional EP^{all}; 95% intervals are bootstrap percentiles. The same 102 clean intervention pairs are projected onto each model's outputs.

Model	Raw EP^{all}	Clean EP^{all}	Clean 95% CI	Clean original-yes	Clean persistent
BLIP-VQA	42.5	35.3	[26.5, 44.1]	90.2	36/102
Qwen2-VL-2B	23.1	11.8	[5.9, 18.6]	94.1	12/102
InternVL2-2B	38.8	32.4	[23.5, 41.2]	95.1	33/102
MiniCPM-V-2.6	22.5	10.8	[4.9, 17.6]	94.1	11/102
Qwen2-VL-7B	27.5	2.9	[0.0, 6.9]	100.0	3/102
InternVL2-8B	36.9	26.5	[18.6, 35.3]	94.1	27/102
Qwen2.5-VL-3B	23.8	16.7	[9.8, 23.6]	93.1	17/102
InternVL3-2B	34.4	21.6	[13.7, 29.4]	96.1	22/102

Qwen2-VL-7B answers YES on all 102 original images in the clean subset. This is not inconsistent with its 89.4% original-yes rate on all 160 images: the clean subset retains examples on which the target is visually unambiguous and, in this run, originally recognized by the model.

B.1 SCALE, DOMAINS, AND EDITORS

The Qwen2-VL scale-up remains close to the 160-pair estimate at 617 examples, and the expanded construction reaches 942 and 1,711 valid pairs. Held-out COCO, VOC2012, and nuImages evaluations preserve nonzero existence persistence for Qwen2-VL-7B and InternVL2-8B. These results test existence replication, not domain equivalence or a cross-domain count/existence ordering, because category distributions and edit difficulty differ.

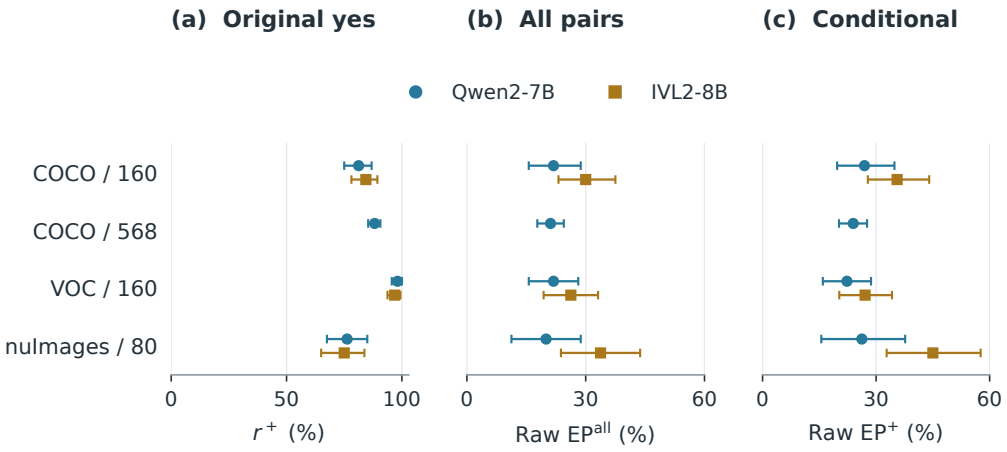

Figure 7: **External existence replication with explicit denominators.** (a) Original affirmative rate r^+. (b) Raw EP^{all}. (c) Conditional EP^+. Row labels give pair counts; bars show 95% bootstrap intervals. The 568-pair COCO holdout has Qwen2-VL-7B only. Category composition and edit difficulty differ across domains; this is not a domain-equivalence test.

Figure 8 shows why visual audit is required. Blur and mask filling can leave texture, boundaries, or parts that preserve evidence. The experiment rules out dependence on mean-box alone, but it does not establish a perfect editor.

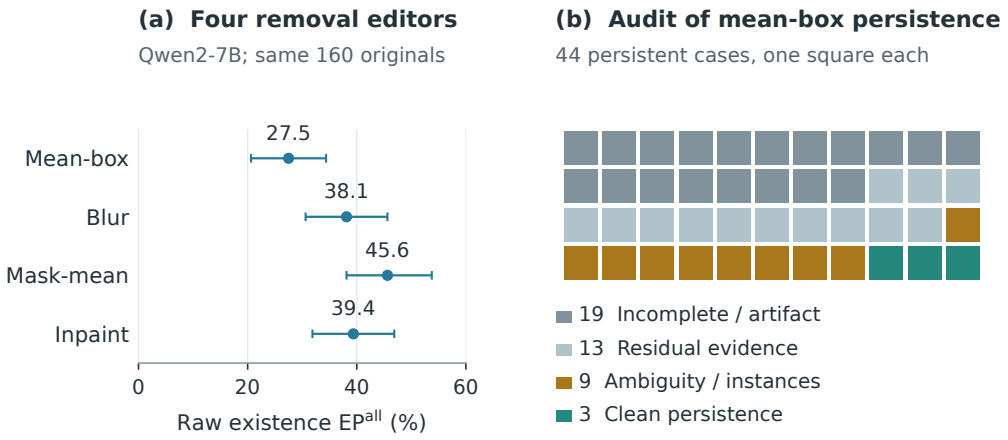

Figure 8: **Editor sensitivity and the need for audit.** (a) Qwen2-VL-7B raw existence $\mathrm{EP}^{\mathrm{all}}$ on the same 160 originals, with 95% bootstrap intervals. (b) Audit taxonomy for the 44 mean-box persistent cases; each square is one case. Only three remain clean persistence. The taxonomy applies to mean-box, not to the other three editors.

B.2 Interpreting Two-Sided Fidelity

Target sensitivity and irrelevant invariance rule out different degenerate behaviors. A constant-YES system appears perfectly invariant under irrelevant removal but never retracts after target removal. Conversely, a system that changes to NO after any visible edit can appear perfectly sensitive to target removal while failing the irrelevant control. This is why the main paper treats (S_1, S_2) as the identifiable report and EDS as a compact summary rather than a replacement metric.

On the matched existence controls, IIS is uniformly high, so most variation in EDS comes from target sensitivity. The EDS range for six open models is 68.2–84.1%. Qwen2-VL-2B and MiniCPM-V-2.6 have similar summaries (84.1 and 84.0); BLIP-VQA has the lowest EDS (68.2) because nearly half of its originally affirmative answers persist after target removal. These values should be read with the paired tests in Table 6, not as a leaderboard detached from the components. InternVL2-2B has two unparsed target answers among 818 eligible rows, so its complement includes 0.24 percentage points of non-affirmation without a recognized binary label.

Quality conditioning and two-sided fidelity answer separate questions. Audit asks whether $T_{\mathrm{tar}}(I)$ actually implements the intended semantic intervention. The irrelevant control asks how the same model responds when a similarly localized edit preserves the queried predicate. Neither can substitute for the other: a visually invalid target edit can inflate persistence even when IIS is high, while an unaudited control cannot certify that the target evidence disappeared.

C Full Specificity Control

Table 6 gives the common-row denominators and discordant pair counts behind the specificity comparison. Rates use the same eligible rows within each model.

Table 6: **Two-sided fidelity on the existence controls.** Target EP^{+} and IIS condition on original YES in both runs; $S_1 = 1 - \mathrm{EP}^{+}$, $S_2 = \mathrm{IIS}$, and EDS is their harmonic mean. b counts target-not-yes/control-yes pairs; c counts target-yes/control-not-yes pairs. "Matched" is the model-specific eligible denominator, not all 906 candidate controls.

Model	Matched	Target EP^{+}	S_1	IIS/S_2	EDS	Gap	(b, c)	Exact p
BLIP-VQA	798	47.9	52.1	98.6	68.2	50.8	(410, 5)	2.4×10^{-114}
Qwen2-VL-2B	801	26.1	73.9	97.6	84.1	71.5	(576, 3)	3.3×10^{-167}
InternVL2-2B	818	39.7	60.3	97.9	74.6	58.2	(480, 4)	9.1×10^{-137}
MiniCPM-V-2.6	794	26.7	73.3	98.4	84.0	71.7	(571, 2)	1.1×10^{-167}
Qwen2-VL-7B	792	30.3	69.7	98.6	81.7	68.3	(544, 3)	1.2×10^{-157}
InternVL2-8B	819	38.8	61.2	98.3	75.4	59.5	(490, 3)	1.6×10^{-141}
GLM-4.6V	842	33.4	66.6	99.2	79.7	65.8	(557, 3)	1.6×10^{-161}

D Full Count Results

Table 7: **Author-reported Count-COVE visual-audit aggregates.** Each model has 100 panels. Clean, eligible, and persistent counts are supplied; the ratio k/n_{elig} (%) and 95% Wilson intervals are calculated from those counts. Eligibility differs from the reported original-yes counts; see Appendix A.3.

Model	Clean	Eligible	Persist.	Ratio [95% CI]
Qwen2-VL-7B	88	86	43	50.0 [39.7, 60.3]
InternVL2-8B	91	86	37	43.0 [33.1, 53.6]
GLM-4V-Plus-0111	91	87	40	46.0 [35.9, 56.4]

Table 8: **Count-COVE results.** All persistence columns are unconditional $\mathrm{EP}^{\mathrm{all}}$. Proxy-clean contains 572 metadata-valid pairs; strict contains 329 exact-two-to-one pairs.

Model	Raw $\mathrm{EP}^{\mathrm{all}}$	Raw 95% CI	Proxy $\mathrm{EP}^{\mathrm{all}}$	Proxy 95% CI	Strict $\mathrm{EP}^{\mathrm{all}}$	Strict 95% CI
BLIP-VQA	80.8	[78.4, 83.2]	80.2	[77.1, 83.4]	77.8	[73.3, 82.1]
Qwen2-VL-2B	91.4	[89.6, 93.1]	92.0	[89.7, 94.1]	91.2	[87.8, 94.2]
InternVL2-2B	93.5	[91.9, 95.1]	94.6	[92.7, 96.3]	95.1	[92.7, 97.3]
MiniCPM-V-2.6	53.3	[50.1, 56.4]	55.6	[51.6, 59.6]	50.2	[44.7, 55.3]
Qwen2-VL-7B	60.4	[57.3, 63.5]	64.2	[60.3, 68.0]	59.6	[54.4, 64.7]
InternVL2-8B	59.4	[56.4, 62.3]	63.5	[59.6, 67.3]	58.1	[52.9, 63.2]

Table 9: **Paired cross-model tests.** The omnibus statistic is from Cochran's Q test. Pairwise differences use exact McNemar tests with Holm correction at $\alpha = 0.05$.

Split	N	Q	p	Sig.
Exist. clean	102	76.0	5.7×10^{-15}	9/15
Count proxy	572	512.7	1.5×10^{-108}	13/15

Table 10: **Prompt robustness ranges.** Min–max EP^+ on the same 200 examples within each predicate. BLIP is a fixed-input reference, not a template sweep.

Model	Existence	Count
BLIP-VQA (fixed)	50.3	85.1
Qwen2-VL-2B	22.7–25.6	91.9–93.9
Qwen2-VL-7B	22.6–29.8	58.8–69.5
InternVL2-2B	14.9–33.2	79.0–94.4
InternVL2-8B	21.5–32.5	52.0–76.8
MiniCPM-V-2.6	17.8–21.9	58.2–61.1
GLM-4.6V	19.1–28.2	37.7–43.5

Why paired tests are required. Every model is evaluated on the same images, so model outputs are repeated binary measurements rather than independent samples. Cochran's Q test provides an omnibus test of equal persistence on the shared rows; exact McNemar tests then localize pairwise disagreements, with Holm correction for family-wise error. This procedure is preferable to ANOVA, unpaired t-tests, or tests that treat model–image rows as independent. Statistical significance is reported alongside absolute rates because a small systematic difference can be significant on a large split without being practically decisive.

The matrix contains 70 named cells: 60 wrapper conditions for six instruction-following models, plus ten repeated BLIP cells. Strict, cautious, visual, and verify form the primary set because unrecognized-answer frequency is at most 5%. Bare is excluded from their ranges: it produces 49.8–53.5% parser failure for Qwen2-VL-2B and 92.8–100% for GLM-4.6V. Across the four wrappers, each instruction-following model's minimum count persistence exceeds its maximum existence persistence. The count and existence images are not matched across predicates. All 70 archived rates and Wilson intervals are in `prompt_matrix_metrics.csv`; the companion response inventory contains 1,568 unrecognized-answer candidates. These records measure recognition coverage under the deterministic parser. They do not supply human-verified semantic accuracy for ambiguous responses.

Table 11: **Full matched five-editor EP^+ matrix (%).** Every cell uses the same 160 examples for a model–predicate pair. MB: mean-box; BL: blur; MM: mask-mean; MI: mask-inpaint; LM: LaMa.

Model	Predicate	MB	BL	MM	MI	LM
Qwen2-VL-7B	Exist.	30.8	42.7	51.0	44.1	35.0
Qwen2-VL-7B	Count	60.7	68.0	76.7	70.7	71.3
InternVL2-2B	Exist.	42.5	48.6	56.2	47.3	47.3
InternVL2-2B	Count	93.7	94.3	96.2	98.1	98.1
MiniCPM-V-2.6	Exist.	26.1	38.0	42.3	35.9	36.6
MiniCPM-V-2.6	Count	59.6	63.7	80.1	70.5	71.2

All editor comparisons against mean-box use the exact McNemar test on shared original-YES rows. For count, LaMa versus mean-box gives nominal $p = 0.00249$ (Qwen2-VL-7B), 0.0156 (InternVL2-2B), and 0.00232 (MiniCPM-V-2.6). After Holm correction over all 24 comparisons, these become 0.0349, 0.1563, and 0.0347, respectively. Full intervals and nominal tests are in the original editor CSVs; adjusted tests are in `source_data/iclr_review_20260905/editor_matrix_paired_holm.csv`. The matrix is not visually audited editor by editor, so it establishes sensitivity and qualitative replication rather than an editor-quality ranking.

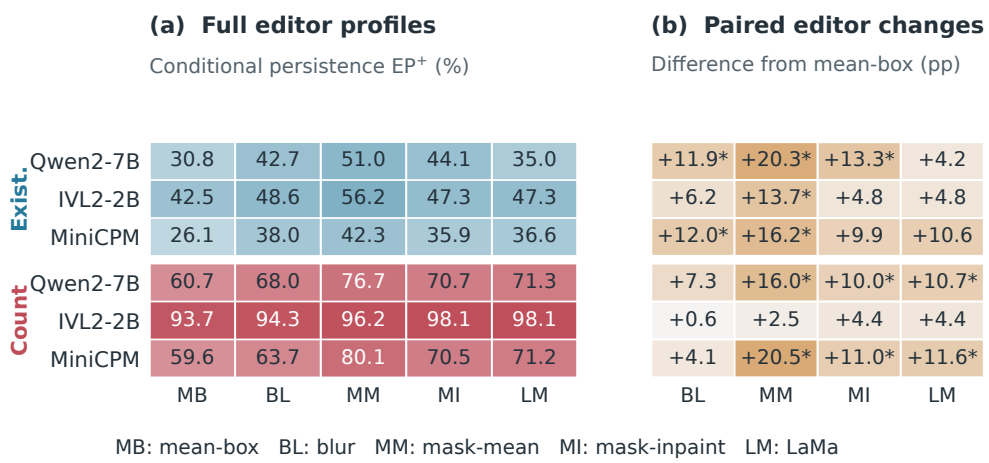

Figure 9: **Five-editor robustness and paired changes.** (a) All 30 model–predicate–editor rates, using a 0–100% scale. (b) Differences from mean-box on shared original-yes rows; asterisks denote exact McNemar tests with Holm-adjusted $p < 0.05$ across 24 contrasts. The difference color scale is symmetric at ± 25 percentage points. Each model/predicate uses the same 160 rows across editors. No editor-specific visual audit is implied.

Figure 10 separates geometry profiles by model rather than showing only pooled rates. The fitted logistic model uses the yes–yes persistence indicator and includes model indicators, standardized log retained area, standardized log total removed area, log original and removed instance counts, and maximum removed/retained IoU. Covariance is clustered by 572 source images, preserving dependence among six responses per image. Retained-area odds are 0.81 [0.68, 0.97] and removed-area odds are 0.72 [0.60, 0.85]. These are pointwise intervals for an exploratory regression, not multiplicity-adjusted confirmatory tests. The strongly correlated instance-count coefficients remain uncertain. The corrected data and covariance specification are in `source_data/iclr_review_20260905/`.

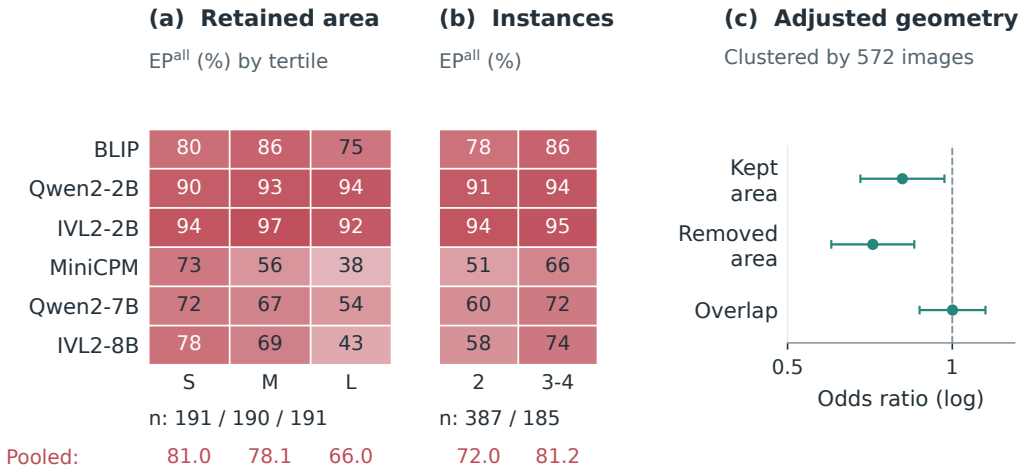

Figure 10: **Count persistence by model and scene geometry.** (a) Retained-area tertiles: small, medium, large. (b) Exact-two versus three/four-instance originals. Heatmaps use a 0–100% scale; pooled rates below them average six responses per image, and n counts images. (c) Selected geometry coefficients from the adjusted logistic model, with pointwise 95% intervals and covariance clustered by 572 images. Coefficients are per standardized predictor unit. The analysis is associational, not a causal explanation.

Figure 11 aligns the archived responses on the same strict pairs and shows two selected image pairs. This sample-level view complements average rates without treating metadata screening as visual certification.

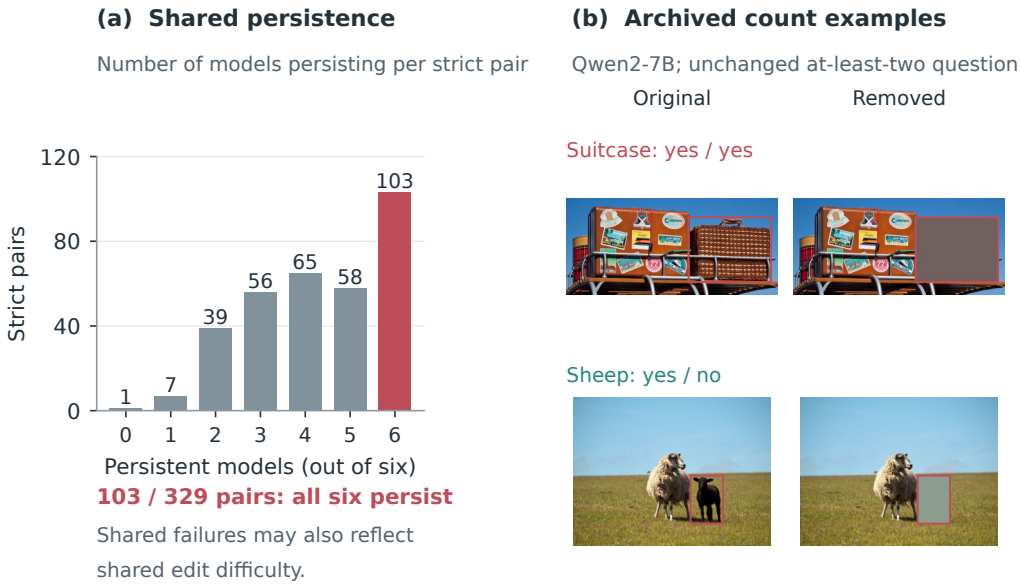

Figure 11: **Shared count persistence and inspectable examples.** (a) Number of the six models persisting on each of the same 329 strict pairs, recomputed from archived responses; all six persist on 103 pairs. (b) Two selected pairs with archived Qwen2-VL-7B answers. Red boxes mark the removed regions. Original photographs are from COCO; edited displays are reconstructions using the archived mean-box settings, not newly queried outputs or claims of byte-identical server images. Strict screening is metadata-based, and shared failures may reflect shared edit difficulty.

E Matched Count-Mechanism Probe

The mechanism probe contains 20 geometry-matched pairs of exact-two examples. Each pair contains one Qwen-flip/InternVL-persist disagreement and one example where both models persist after the 2-to-1 intervention. A second intervention removes the remaining instance, producing a 2-to-1-to-0 trajectory. Table 12 reports the behavioral checks; Table 13 reports the strongest layerwise matched effects.

The at-least-two question is unchanged across all three views, so the correct answer sequence is YES/NO/NO, not YES/YES/NO. For layer ℓ, the extraction script records its first forward-pass output: the last input-token state for the language stack and the token-averaged state for the vision stack. Let I_j denote the view with j intended remaining instances. Writing the pooled vectors as $h_\ell(I)$, we compute $d_{2\to1,\ell} = 1 - \cos(h_\ell(I_2), h_\ell(I_1))$ and average the within-matched-pair difference between the disagreement and both-persist strata. Indices are zero-based. Bootstrap intervals estimate the mean difference; the exact sign test instead tests the balance of positive and negative pair differences.

Table 12: **Behavioral 2-to-1-to-0 probe (%).** Each row contains 20 examples. Columns 2+, 1, and 0 give affirmative rates. The last column is the fraction that retracts only at zero; this is a diagnostic of late retraction, not correct at-least-two behavior.

Model	Stratum	2+	1	0	1→0
InternVL2-2B	Both persist	100	100	80	20
InternVL2-2B	Disagreement	100	100	35	65
Qwen2-VL-7B	Both persist	100	100	50	50
Qwen2-VL-7B	Disagreement	100	0	5	0

Table 13: **Peak layerwise matched effects.** Positive $\Delta d_{2\to1}$ means the disagreement member changes more than its geometry-matched both-persist partner. Intervals are pointwise paired bootstrap estimates; p is an exact sign test and p_H applies Holm correction across all 108 scanned layers.

Model	Stack/layer	Δd	95% CI	p	p_H
InternVL2-2B	Decision 18	.0117	[.0042,.0201]	.263	1.000
InternVL2-2B	Vision 13	.0059	[.0010,.0123]	.263	1.000
Qwen2-VL-7B	Decision 20	.1122	[.0795,.1452]	.000402	.0427
Qwen2-VL-7B	Vision 03	.0033	[-.0040,.0118]	.503	1.000

Holm correction over all 108 scanned layers retains Qwen decision layers 18–24: adjusted $p = 0.00433$ at layers 18–19 and 0.0427 at layers 20–24. No other layer passes 0.05. This guards against reporting only a favorable peak, but does not remove behavioral selection bias or turn a 20-pair probe into a causal intervention on the network. The release preserves all original layer rows and provides the adjusted values in `source_data/iclr_review_20260905/e12_layer_effects_holm.csv`.

F Calibration and API Replication

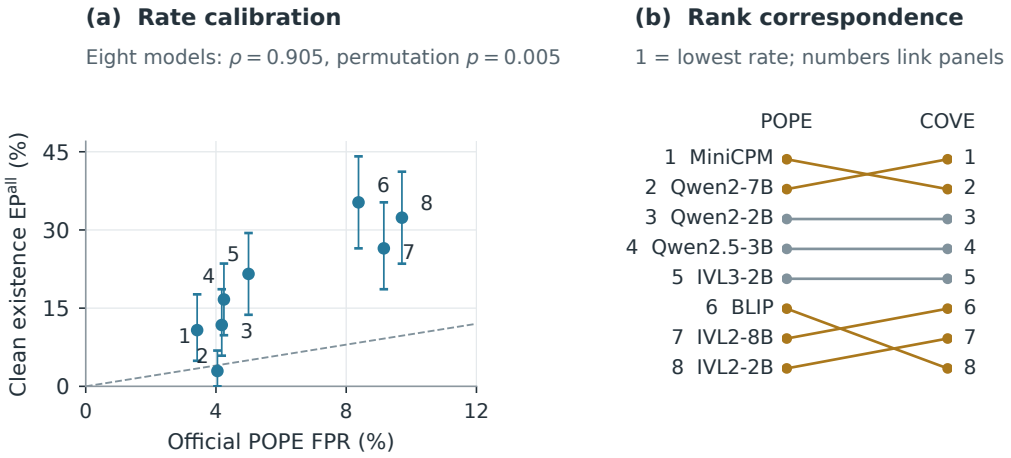

Figure 12: **POPE calibration in magnitude and rank.** (a) Eight-model rates with clean-existence 95% bootstrap intervals; dashed line: equal rates. Point numbers identify the models in (b), which links their rank positions. Gold marks changed ranks. Rank association remains strong ($\rho = 0.905$, permutation $p = 0.005$); the figure does not support metric independence.

Table 14: **Common-subset transition summaries (%).** Update is original-yes/target-not-yes over all matched rows; stay-yes is original-yes/target-yes; change is original/control affirmation disagreement. H is the harmonic mean of update and $1 - \text{change}$. These are BEAF/ROHE-inspired score translations on COVE, not official benchmark reruns.

Model	Matched	Update	Stay-yes	Change	H	COVE EP$^+$	IIS	EDS
BLIP-VQA	906	45.9	42.2	3.0	62.3	47.9	98.6	68.2
Qwen2-VL-2B	906	65.3	23.1	3.4	77.9	26.1	97.6	84.1
Qwen2-VL-7B	906	60.9	26.5	3.6	74.7	30.3	98.6	81.7
InternVL2-2B	906	54.4	35.9	2.5	69.8	39.7	97.9	74.6
InternVL2-8B	906	55.3	35.1	3.1	70.4	38.8	98.3	75.4
MiniCPM-V-2.6	906	64.2	23.4	2.8	77.4	26.7	98.4	84.0
GLM-4.6V	905	62.3	31.0	2.2	76.1	33.4	99.2	79.7

Table 15: **GLM-4.6V API check.** EP^+ and IIS are conditional on original YES. The human-clean row applies the existing 102-edit audit mask.

Split	Valid	Orig. YES	Metric
Existence main	159	89.3	32.4 EP^+
Existence clean	102	94.1	14.6 EP^+
Existence 1,000	941	93.5	32.2 EP^+
Count 1,000	999	93.1	47.8 EP^+
Irrelevant 906	905	93.4	99.1 IIS

The same safety-filtered COCO image produces one unavailable request in each applicable GLM-4.6V split; no returned answer fails parsing. The row is excluded only from the relevant denominator and remains in the failure logs. GLM's standalone control IIS is $837/845 = 99.1\%$. Requiring original-yes in both target and control runs yields $835/842 = 99.2\%$ in Table 6; the difference is an eligibility change, not a conflicting measurement. GLM-4V-Plus-0111 is not pooled with GLM-4.6V and appears only in the reported count-audit aggregates here.

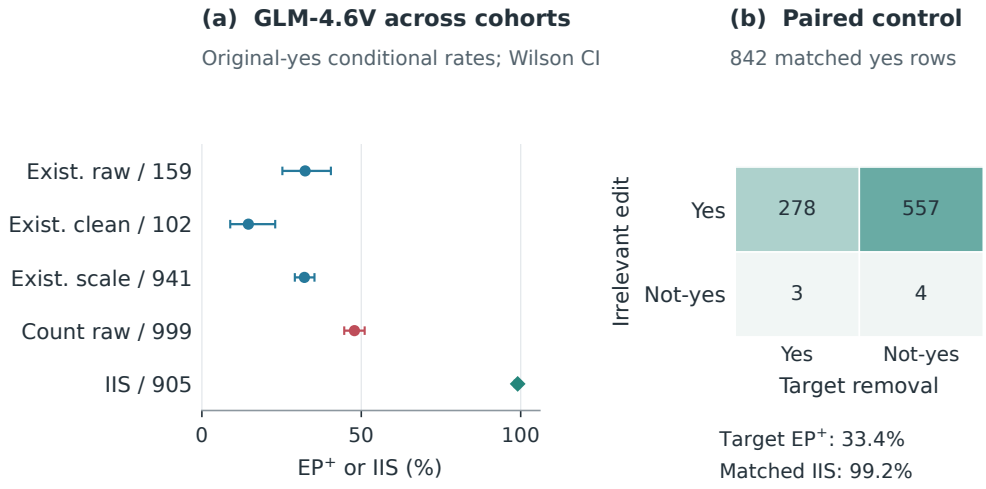

Figure 13: **API replication and matched-row accounting.** (a) GLM-4.6V rates with 95% Wilson intervals; row labels give valid pair counts, not original-yes denominators. Count is raw and only the designated existence row is visually screened. (b) Target/control affirmation counts on 842 common original-yes rows. The paired 99.2% IIS differs from standalone 99.1% because eligibility differs. Endpoint versions are not pooled.

G Diagnostic Verification Transfer

We evaluated Faster R-CNN, OWL-ViT, segmentation-mask quality, and an evidence-overlap proxy as post-hoc gates. Lower thresholds preserve more original affirmative answers but leave more residual EP^{all}; stricter thresholds reduce EP^{all} while discarding valid positives. On official POPE and CHAIR transfer, this conservatism appears as lower recall, more abstention, or removal of true object mentions. These results explain why verification is not framed as the paper's method contribution.

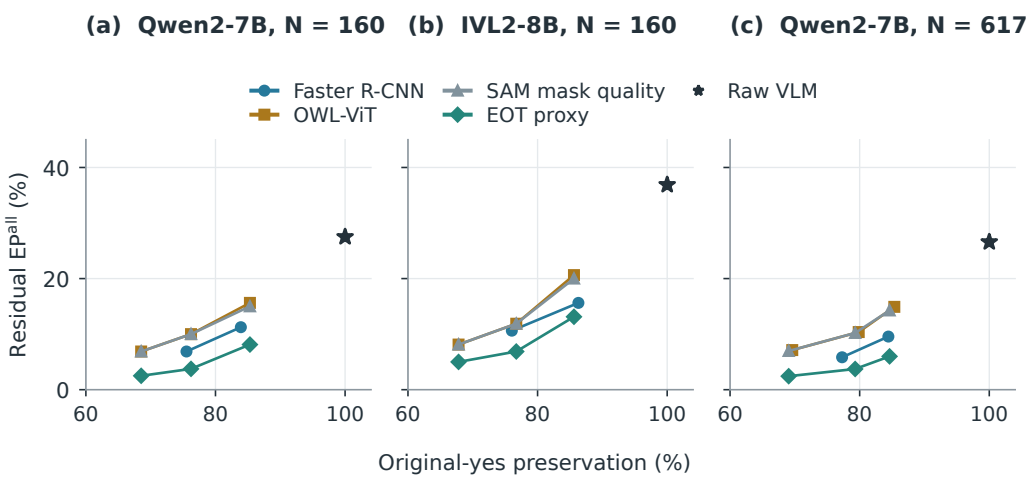

Figure 14: **Verification trade-offs across gates and scale.** Each panel shows archived threshold sweeps for four non-oracle gates on the specified model and cohort; stars mark the raw model. Lower residual EP^{all} comes at a cost to preservation of original affirmative answers. Connected points are measured thresholds, not fitted response curves. Oracle-region gates are excluded from this deployable-gate comparison.

H Reproducibility Artifacts

The accompanying archive contains JSONL manifests for the 160/942/1,711 existence splits, 1,000 count pairs, 906 irrelevant controls, and official POPE; model-output summaries; construction/evaluation scripts; an audit rubric; and a dataset card. It also contains source tables for the GLM-4.6V endpoint, common-row protocol translation, 70-cell prompt matrix, 30-cell editor matrix, and full layerwise representation probe. Upstream dataset images are referenced by identifiers and require separate download. Manifests resolve image paths from a configurable data root. The existence audit mask and model-specific count-audit aggregates have the evidential scope described in Appendix A.3.

H.1 Manifest Schema

Each JSONL row has a stable example identifier, upstream dataset/image identifier, target category, image dimensions, prompt, intended original and counterfactual labels, edit mode, and one or more intervention regions. Existence rows include the recorded target-category boxes. Count rows additionally record the retained instance, removed instance identifiers, eligible original instance count, and intended counterfactual remaining count. Control rows retain the queried category while recording the non-target object or sampled background region. Counterfactual manifests add the materialized image basename and final pixel-space edit boxes.

Image files are not embedded in the release. Manifests store portable basenames and the evaluation scripts resolve them under user-supplied original and counterfactual image roots. This keeps the benchmark independent of the construction server and avoids redistributing datasets under incompatible licenses. The dataset card lists the required upstream sources and split names.

H.2 Reuse of the Removal Resource

The released manifests fix the exact evaluated cohorts. To repeat an edit, users resolve its source image, apply the stored intervention regions, and retain the example identifier when joining predictions and quality masks. Changing an editor or sampling rule creates a new intervention condition; its validity should be checked independently rather than inherited from a different edited image. Model outputs are not used as evidence that removal succeeded. For training extensions, partitioning by upstream image identifier across all manifests prevents source-image reuse from leaking across splits.

These practices make the resource useful for new models and editors while preserving the meaning of each reported denominator.

H.3 Failure and Denominator Policy

Inference outputs preserve the raw answer, parsed label, request status, and resolved paths. Explicit affirmative and negative answers map to YES and NO; refusals, ambiguous outputs, parser failures, missing images, and API failures remain separate statuses. No failed row is silently converted to NO. Summary scripts report total, valid, original-affirmative, and persistent denominators so that $\mathrm{EP}^{\mathrm{all}}$ and EP^{+} can be recomputed without guessing which rows were excluded.

Figure source CSVs contain the plotted values rather than pixels extracted from rendered figures. The paired-model analysis additionally releases its JSON/CSV test outputs, and the two-sided fidelity table has a dedicated CSV with S_1, S_2, EDS, specificity gaps, and exact p-values. These artifacts separate scientific values from presentation choices and permit independent replotting.

H.4 Prospective Training Use

A model with trainable parameters θ could use validated triplets to learn both relevant updating and irrelevant invariance. Let $p_\theta(y \mid I, q)$ be its binary answer distribution, with original label $y = 1$ and target-edited label $y_t = 0$. One possible objective is

$$\begin{aligned}\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{task}}(\theta) - \lambda_t \log p_\theta(y_t \mid T_{\mathrm{tar}}(I), q) \\ + \lambda_r D_{\mathrm{KL}}(p_\theta(\cdot \mid I, q) \,\|\, p_\theta(\cdot \mid T_{\mathrm{irr}}(I), q))\,.\end{aligned} \tag{7}$$

Here $\mathcal{L}_{\mathrm{task}}$ is a supervised loss on original images and general visual question answering examples; $\lambda_t, \lambda_r \geq 0$ weight the target and control terms. D_{KL} is Kullback–Leibler divergence between the two answer distributions. The target term penalizes failure to retract, while the control term penalizes changes under a valid irrelevant edit. Only validated edits should supply these labels. Training and evaluation would require source-image-disjoint splits across all manifests, along with held-out categories and editors to test for shortcuts. This objective describes a possible reuse of COVE, not an evaluated training method.