SAME SCORE, DIFFERENT BLIND SPOTS: WHAT OBJECT-HALLUCINATION BENCHMARKS CAN AND CANNOT CERTIFY

Anonymous authors
Paper under double-blind review

ABSTRACT

Object-hallucination benchmarks aggregate heterogeneous queries into pooled scores that do not reveal how false-positive risk varies across objects. We present an object-conditioned audit that assigns violation, unresolved, or certified verdicts and assesses whether benchmark allocation supports simultaneous object-wise certification. We evaluate eight open VLM checkpoints on POPE and DASH-B at a nominal 10% pooled false-positive budget. Kimi-VL and Molmo 2-O differ by 0.004 in pooled AUROC yet make 20 versus 49 false assertions on the same 147 POPE-popular cup negatives. The pair-level difference survives a selection-aware permutation test and persists under the models' native decisions. Using split-half held-out calibration and simultaneous exact bounds, we identify violations in 20 of 29 model benchmark cells in each direction, with 15 of 29 violating in both. A conservative calibration sensitivity analysis retains violations in a majority of cells in both directions. In the descriptive full-data census, no cell earns a simultaneous certificate covering all eligible objects. Under this convention, sparse allocation can preclude certification even with zero observed errors. An Open Images positive control on a checkpoint subset yields class-level certificates under frozen labels. A pre-specified readout gate voids Gemma 4's POPE cells, showing that statistical resolution alone does not establish valid model attribution. We recommend object-wise counts and uncertainty bounds, explicit unresolved outcomes, and validity checks alongside pooled scores.

1 INTRODUCTION

Object-hallucination benchmarks reduce thousands of heterogeneous object queries to a small set of pooled metrics, such as accuracy, F_1, or AUROC (Li et al., 2023; Augustin et al., 2025). These metrics summarize aggregate performance or discrimination, but do not establish how false-positive risk is distributed across objects. In particular, they do not show whether any object violates a fixed error budget, whether the apparent risk is diffuse or concentrated, or whether the benchmark contains enough evidence to establish the converse claim that all eligible objects satisfy that budget. A benchmark may therefore contain enough samples to estimate a pooled score precisely while containing too few examples per semantic group to resolve the object-conditioned claims that pooled reporting leaves open. The distinction between average and subgroup risk is well established (Cherian & Candès, 2024; Ferrara et al., 2025); our question is empirical and benchmark-level: which object-conditioned risk statements do current hallucination benchmarks actually contain enough evidence to support?

This limitation is visible in a direct model comparison. On POPE-popular, Kimi-VL (Kimi Team et al., 2025) and Molmo 2-O (Deitke et al., 2025) differ by only 0.004 in pooled AUROC but make 20 and 49 false assertions on the same 147 "cup" negatives at a matched pooled false-positive budget (Figure 1). The pair survives selection-aware testing, and the contrast persists under native hard decisions at 14 versus 31. The complementary InternVL3.5–Molmo comparison shows that an object-wise gap can also depend on the operating point (Section 4.1).

We study how much such evidence current benchmarks provide across eight open VLM checkpoints from six released VLM-system families on POPE (Li et al., 2023) and DASH-B (Augustin

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.

et al., 2025). At a nominal $\alpha = 0.10$ pooled false-positive budget, simultaneous lower and upper bounds classify each model benchmark cell as violation, unresolved, or certified. Split-half inference identifies violations in $20/29$ cells in each calibration/counting direction and in $15/29$ in both. A bootstrap-tightened calibration sensitivity retains $16/29$ and $15/29$, respectively. The full-data descriptive census finds $27/29$ violations, yet no POPE/DASH-B cell certifies that all eligible objects satisfy the budget.

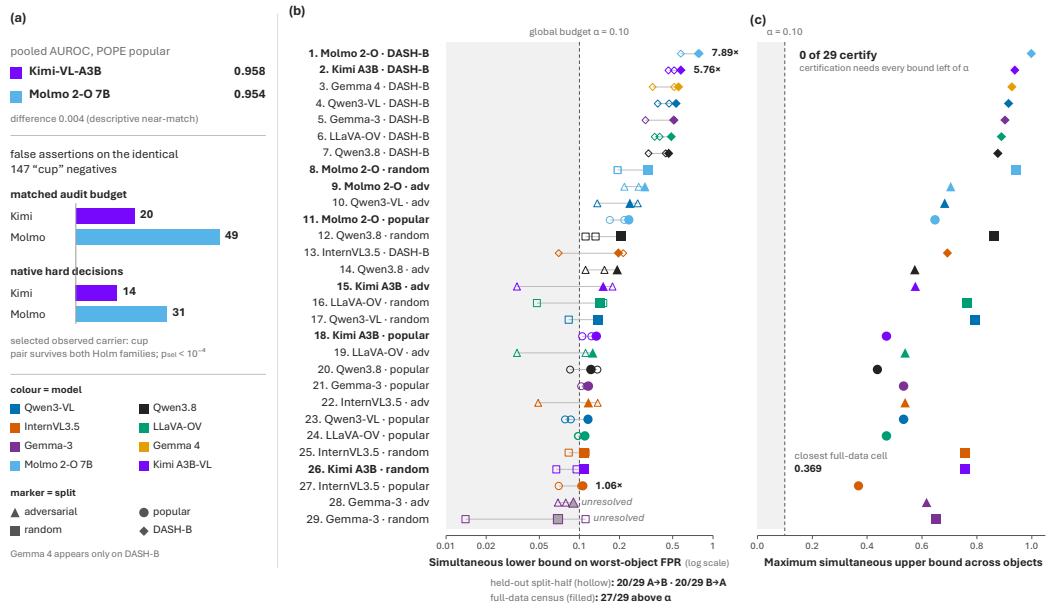

Figure 1: **Similar pooled discrimination can conceal different object-conditioned risks, while violation and certification require different evidence.** (a) Kimi-VL and Molmo 2-O have near-matched pooled AUROC but make 20 and 49 false assertions on the same 147 POPE-popular "cup" negatives at the matched audit budget. The contrast persists under their native hard decisions, at 14 versus 31. The pair-level claim survives a selection-aware permutation test that re-runs object selection inside the null; "cup" is the selected carrier rather than a separately selection-adjusted named-object effect. (b) In the descriptive full-data census, the largest within-cell simultaneous lower bound exceeds α in $27/29$ cells. (c) The corresponding maximum simultaneous upper bound exceeds α in every cell, yielding no whole-cell certificates.

The absence of whole-cell certificates is partly a property of benchmark allocation, not simply evidence of poor models. We summarize this design-side limitation with best-case certifiable coverage (BCC), a count-only measure of how much of the stated object family could be certified at zero observed errors. BCC can be computed before running a model and separates certification failures already blocked by allocation from those caused by observed errors. A separate six-model Open Images positive control (Kuznetsova et al., 2020) yields class-level certificates when per-class allocation clears the required floor.

However, statistical resolution is insufficient for model attribution. A pre-specified readout gate voids Gemma 4's POPE cells but passes on DASH-B; RePOPE corrections preserve the POPE conclusion for all seven checkpoints that pass the readout gate (Neuhaus & Hein, 2025). A separately disclosed, single-inspector post-hoc Open Images review suggests a semantic-reference concern but is not treated as adjudicated ground truth or corrected labels. These checks separate statistical detection from behavioural and semantic validity (Sections 7.1–7.2).

Our contributions are threefold. We show that near-tied pooled scores can conceal different object-wise risks, use simultaneous bounds and BCC to distinguish observed violations from allocation limits on certification, and examine whether reference labels and behavioural readouts support the resulting interpretations.

2 RELATED WORK

Object-hallucination benchmarks and fine-grained evaluation. CHAIR (Rohrbach et al., 2018) introduced a metric for object hallucination in image captions, while POPE (Li et al., 2023) formulated object presence as binary question answering, typically evaluated using pooled accuracy and F1. Subsequent benchmarks examine a broader range of errors and provide finer diagnostic detail. BEAF (Moon et al., 2024) reports object-wise errors before and after object removal, together with uncertainty intervals. DASH and HOPE mine object-specific hard negatives (Augustin et al., 2025; Xie et al., 2025), while RePOPE (Neuhaus & Hein, 2025) corrects annotation errors in POPE. FINER (Xiao et al., 2026) evaluates multi-object, attribute, and relation hallucinations using fine-grained negative queries. MIOH (Min et al., 2026) extends evaluation to existence, counting, attributes, and position in multi-image settings. THRONE (Kaul et al., 2024) highlights the distinction between hallucination in free-form captions and errors in targeted object-presence queries; our study focuses on the latter. We examine whether benchmark evidence supports simultaneous object-risk claims at a shared operating point and whether its allocation permits certification.

Subgroup inference and finite-sample certifiability. Cherian & Candès (2024) formulate subgroup auditing as a multiple-testing problem, with simultaneous guarantees for flagging groups with excessive error and certifying adequate performance. Size-adaptive hypothesis testing (Ferrara et al., 2025) and empirical Bayes shrinkage (Fogliato et al., 2024) address uncertainty in small-group evaluation. TRON (Wang et al., 2025a) controls marginal miscoverage in sampled response sets. SAFER (Wang et al., 2026) calibrates sampling and filtering under a fixed generation cap, abstaining when calibration cannot establish the desired miscoverage bound. We apply subgroup auditing to fixed object categories and examine whether benchmark allocations support simultaneous false-positive constraints and whole-cell certification.

Benchmark and readout validity. RePOPE (Neuhaus & Hein, 2025) shows that reference-label errors can change the interpretation of POPE results. This illustrates a broader measurement-validity concern: whether an evaluation captures the property it is intended to measure (Jacobs & Wallach, 2021). The choice of readout also matters, since first-token probabilities can disagree with a model's generated answer (Wang et al., 2024). These concerns motivate checking both reference labels and behavioural readouts when interpreting object-wise risk.

Decoding-time mitigation methods such as VCD and OPERA are commonly evaluated with pooled POPE metrics (Leng et al., 2024; Huang et al., 2024); Appendix A.13 uses VCD only as an auxiliary case study of operating-point dependence, consistent with recent concerns that aggregate VCD gains need not imply improved visual grounding (Yin et al., 2025).

The underlying statistical primitives are established; our contribution lies in how they are assembled to diagnose object-conditioned risk, certifiability, and validity in hallucination benchmarks. Appendix A.1 details the statistical tools used in our audit and their sources.

3 SETUP AND AUDIT PROTOCOL

3.1 SETTING AND ESTIMAND

Each query pairs an image B with an object-presence claim c, posed as a yes/no question. A *cell* denotes one model checkpoint evaluated on one benchmark split. The VLM assigns a yes/no confidence score $s(B, c)$ using the readout defined in Section 3.4. Our audit *asserts* the claim when $s(B, c) > \tau$. Unless explicitly identified as a native hard readout, "assertion" refers to this thresholded audit decision rather than a generated answer.

Queries are grouped by the object named in the claim. For object group g, the false-positive risk at threshold τ is

$$R_g = \Pr(s(B, c) > \tau \mid c \text{ false}, g) .$$

We compare cells at a nominal pooled false-positive budget α, using $\alpha = 0.10$ as an illustrative audit setting. For each cell, we set τ to the sample $(1 - \alpha)$ quantile of the pooled negative scores in the calibration data. Finite-sample effects and score ties can make the calibration FPR differ from α;

scores equal to τ are not asserted. This calibration does not guarantee that pooled FPR on held-out data is at most α. Appendix A.3 specifies the quantile and tie conventions.

3.2 Inference and verdicts

For *primary inference*, we deterministically divide images into two halves, A and B, keeping all queries from an image in the same half. We calibrate τ on pooled negatives from A and count false assertions in B, then reverse the roles. Results are reported separately for the two directions. The *full-data census* calibrates the threshold and counts errors on the same data and is used for descriptive reporting. Object eligibility depends only on negative counts in the full benchmark inventory. Let N_g denote this count for object g. We include objects with $N_g \geq 20$ on POPE and $N_g \geq 15$ on DASH-B. The resulting eligible family $\mathcal{G}$ contains $G = |\mathcal{G}| = 18, 6, 36$ objects for POPE adversarial, popular, and random, respectively, and $G = 34$ for DASH-B. The cutoffs were fixed before inspecting the corresponding results, and the family remains fixed in both split-half directions. Appendix A.9 reports sensitivity analyses for eligibility cutoffs of 15, 20, 25, and 30 negatives per object. All verdicts apply only to eligible objects.

For each analysis, let n_g be the number of evaluated negative queries in group g and x_g the number asserted. These counts use the evaluation half for split-half inference and the full inventory for the census, where $n_g = N_g$. We compute two-sided Clopper–Pearson bounds $[\underline{R}_g, \overline{R}_g]$ with Bonferroni correction at family-wise level $\delta = 0.05$, assigning $\delta/(2G)$ to each tail. The cell-level verdict is

$$\text{violation} \iff \max_{g \in \mathcal{G}} \underline{R}_g > \alpha,$$
$$\text{certified} \iff \max_{g \in \mathcal{G}} \overline{R}_g \leq \alpha,$$

and unresolved otherwise. An individual object is certified when its upper bound satisfies $\overline{R}_g \leq \alpha$. Whole-cell certification requires this condition for every eligible object.

For held-out inference, we assume that assertion indicators are i.i.d. within each group conditional on the calibration data, under the stated benchmark-construction model. Negatives within a group come from distinct images; this alone does not establish independence. Under this model, the bounds provide simultaneous coverage of at least $1-\delta$ within a cell for each fixed split direction and budget. Bonferroni allows dependence across object groups. Coverage is not controlled jointly over the 29-cell matrix. Appendices A.3, A.5, and A.6 report leave-one-group-out and matrix-wide diagnostics.

Observed pooled FPR on the evaluation halves ranges from 0.067 to 0.135, so α remains a nominal budget. Appendix A.11 reports a sensitivity analysis using bootstrap-tightened calibration. Independence diagnostics are reported in Appendix A.7. We also evaluate budgets $\alpha \in \{0.02, 0.04, \ldots, 0.30\}$.

3.3 Design-only certifiability

Best-case certifiable coverage (BCC) is the fraction of eligible objects whose full-inventory counts permit certification with zero errors under the same two-sided Bonferroni convention:

$$\mathrm{BCC}(\alpha, \delta) = \frac{1}{G} \sum_{g \in \mathcal{G}} \mathbf{1}\left[\overline{R}_g(0, N_g) \leq \alpha\right].$$

Here $\overline{R}_g(0, N_g)$ is the upper bound for zero errors among N_g negatives. BCC can be computed from the counts, budget, and inference convention before running a model; $\mathrm{BCC} < 1$ precludes whole-cell certification even with zero errors. Accordingly, BCC is a convention-conditional design diagnostic, not a model-performance metric or an inference-procedure-independent limit on certifiability.

3.4 Readout applicability and audit matrix

We append the instruction `Answer Yes or No.` to each POPE question (Li et al., 2023) and use the model's own chat template. Let p_Y denote the summed first-token probability of `Yes`,

leading-space `Yes`, and `yes` under the full-vocabulary distribution, and define p_N analogously for the corresponding No variants. The audit score and answer mass are

$$s(B, c) = \frac{p_Y}{p_Y + p_N}, \qquad a(B, c) = p_Y + p_N.$$

We fixed the readout gate's directional criterion and evaluation timing before inspecting per-query answer mass. A cell is voided only when all four conditions hold:

1. At least 1% of its negative queries have answer mass $a(B, c) < 0.5$.
2. Per-object mean answer mass is negatively correlated with the empirical false-positive rate, with two-sided Spearman $p < 0.05$.
3. The forced binary readout identifies a violation.
4. Restricting evaluation to answered queries either removes the violation or reduces the originally selected object's lower bound to at most α.

A voided cell is excluded from the audit matrix without a claim of compliance.

We evaluate eight checkpoints from six released VLM families: Qwen3-VL (Bai et al., 2025), Qwen3.8-27B (Qwen Team, 2026), Kimi-VL-A3B-Instruct (Kimi Team et al., 2025), InternVL3.5 (Wang et al., 2025b), LLaVA-OV (Li et al., 2024), Gemma-3 (Gemma Team et al., 2025), Gemma 4 (Gemma Team, 2026), and Molmo 2-O (Deitke et al., 2025). Qwen3.8 and Kimi-VL were added after the initial analysis using inclusion criteria based on procedural validity rather than risk results. The matrix-level pattern was already present before these additions: the original 21-cell matrix yielded $14/21$ held-out violations in each direction and $19/21$ violations in the descriptive census (Appendix A.15). Token checks, readout sensitivities, model configurations, and the addition chronology are documented in Appendices A.10, A.15, and A.2.

All eight checkpoints are evaluated on DASH-B. Gemma 4 fails the POPE readout gate, leaving seven checkpoints across the three POPE splits and 29 primary cells overall. DASH-B is treated as a mined hard-negative stress test rather than an independent replication because of its overlapping model lineages (Appendix A.7).

4 Pooled scores conceal object-level violations

4.1 Near-tied pooled scores, different object-wise errors

On POPE-popular, Kimi-VL and Molmo 2-O differ by only 0.004 in pooled AUROC but make 20 and 49 false assertions on the same 147 "cup" negatives at the full-data operating point with nominal pooled FPR budget $\alpha = 0.10$ (Figure 1). Their paired discordances are 8 Kimi-only and 37 Molmo-only false assertions, and the gap persists under native hard decisions at 14 versus 31. The InternVL3.5–Molmo comparison illustrates a different relationship between pooled discrimination and object-wise errors: although their pooled-AUROC difference is unresolved ($\Delta = 0.0022$, image-clustered 95% CI $[-0.0032, 0.0074]$), they make 19 and 49 false assertions on the same "cup" negatives at the matched budget, with 3 InternVL-only and 33 Molmo-only errors. Under native hard decisions, these counts are much closer at 34 versus 31, showing that a similar pooled AUROC does not determine whether an object-wise error difference persists across operating points.

To account for paired observations and object selection, we screen model pairs by pooled AUROC and use image-level permutation tests that independently swap the models' complete prediction vectors within each image and repeat object selection in every draw. Both matched-budget comparisons remain significant after Holm correction within the screened family and when the screen is removed (Appendix A.6). The selection-adjusted inference concerns the model pairs, with "cup" reported as the selected object rather than as a separately selection-adjusted named-object effect; native-decision sensitivities are reported in Appendix A.9.

4.2 Violations across the audit matrix

Across the 29 cells, pooled AUROC ranges from 0.87 to 0.98. Our primary inferential result is the held-out split-half analysis, which identifies at least one object-level violation in $20/29$ cells in

each direction, with $15/29$ violating in both. Separately, the descriptive full-data census flags $27/29$ cells despite pooled FPR remaining between 0.092 and 0.100; among the flagged cells, the largest simultaneous lower bounds range from 1.06 to 7.89 times the nominal budget. The maximum occurs for Molmo 2-O on DASH-B, where $48/50$ false assertions for "orange" yield an empirical FPR of 0.96 and a simultaneous lower bound of 0.789 (Figure 1; Appendix A.5). Molmo 2-O was not among the models used to mine or transfer-filter DASH-B (Appendix A.7).

The highest-risk objects often overlap across models: in the full-data census, "chair" has the largest simultaneous lower bound in $5/7$ POPE-popular cells, while "cell phone" does so in $5/7$ POPE-random cells. Such overlap is compatible with the model-dependent error magnitudes and operating-point sensitivity observed above. Permutation tests against diffuse-error nulls provide additional evidence of object-wise concentration in most POPE cells and, using the maximum-risk statistic, in all eight DASH-B cells (Appendix A.6).

4.3 Robustness of the matrix-level conclusion

Bootstrap-tightened calibration reduces the number of evaluations whose held-out pooled FPR exceeds α from 29 to 8 out of 58, while retaining violations in $16/29$ cells for $A \to B$ and $15/29$ for $B \to A$ (Appendix A.11). Applying a single Bonferroni correction across all 692 eligible cell–object combinations retains $18/29$ violations in the full-data census (Appendix A.6), and the 90th percentile of empirical FPR across eligible objects, q_{90}, exceeds pooled FPR at every tested budget $\alpha \in \{0.02, 0.04, \ldots, 0.30\}$ in $26/29$ cells. On RePOPE-adversarial, the worst-object rates remain above budget for all seven evaluated checkpoints, with the implications for reference validity examined in Section 7.1. Additional analyses of native decisions, prompt and readout variants, text-only controls, and oracle performance are reported in Appendices A.9, A.14, and A.8.

5 Benchmark allocation limits what can be certified

5.1 Object certificates do not imply whole-cell certification

The descriptive full-data census yields 27 violations, 2 unresolved cells, and 0 **whole-cell certificates**. Whole-cell certification requires every eligible object's simultaneous upper bound to be at most α, but the smallest value of $\max_{g \in \mathcal{G}} \overline{R}_g$ across the 29 cells is 0.369. Individual objects can nevertheless receive certificates: under the same within-cell convention, the 692 eligible cell–object groups comprise 58 violations, 21 certificates, and 613 unresolved groups. These object certificates occur in $13/29$ cells. The leave-one-group-out, one-sided-bound, and expanded-family diagnostics also yield no whole-cell certificates (Appendices A.5, A.3, and A.8).

5.2 Zero-error floors diagnose allocation limits

The benchmark allocation already prevents whole-cell certification, even before model errors are observed. Under the two-sided Bonferroni convention used for the audit, an object with zero observed errors requires at least

$$N_{\min}(G, \alpha, \delta) = \left\lceil \frac{\log(\delta/(2G))}{\log(1-\alpha)} \right\rceil$$

negative observations to receive a certificate. At $\alpha = 0.10$ and $\delta = 0.05$, this requirement is 63, 53, and 70 observations for POPE-adversarial, popular, and random, respectively, and 69 for DASH-B. POPE-random's largest group contains only 29 negatives, so none of its objects can meet the required floor of 70, even with zero observed errors.

Applying BCC to the full benchmark inventories gives $5/18$, $5/6$, $0/36$, and $0/34$ for the eligible families on POPE-adversarial, popular, random, and DASH-B, respectively. Expanding each family to include every object with at least one negative and recomputing the Bonferroni correction gives $5/54$, $5/10$, $0/79$, and $0/70$. Because every split has at least one eligible object below its floor, **none of the 29 cells could receive a whole-cell certificate even with zero observed errors**. At the object level, however, the implications differ across splits: POPE-popular has enough observations to certify five of its six eligible objects if no errors occur, whereas POPE-random and DASH-B cannot certify any eligible object under this convention.

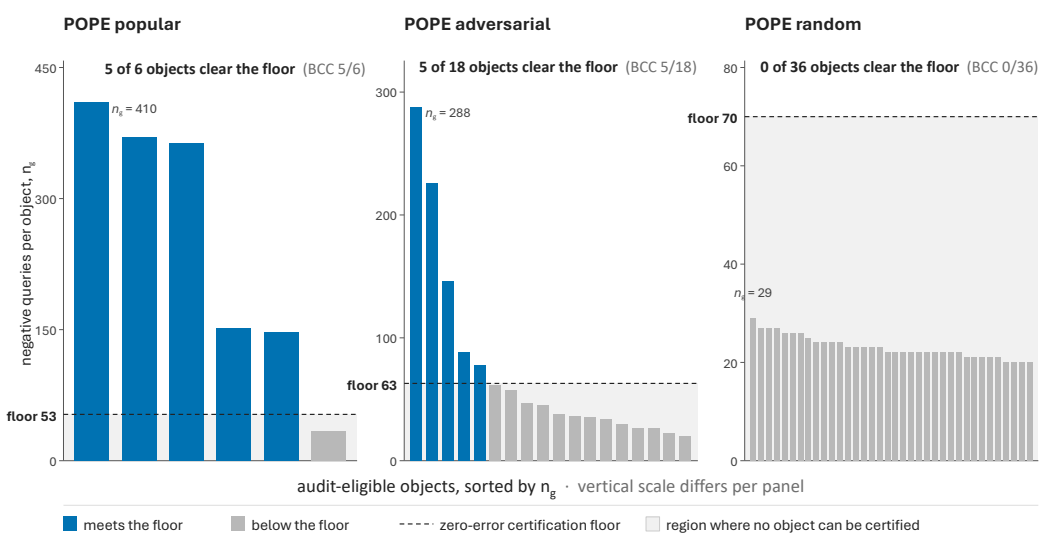

Figure 2: Negative-query allocation and zero-error certification on POPE. Each bar represents an audit-eligible object, sorted by its negative count; blue bars meet the two-sided Bonferroni zero-error requirement, and grey bars fall below it. At $\alpha = 0.10$ and $\delta = 0.05$, the floors are 53, 63, and 70 for popular, adversarial, and random, respectively, giving BCC values of $5/6$, $5/18$, and $0/36$. Every panel contains objects below its floor, preventing whole-cell certification even with zero observed errors.

The family-wise one-sided sensitivity increases adversarial BCC to $7/18$ but still precludes whole-cell certification in all 29 cells (Appendix A.8).

Redistributing the same total number of negatives can change these limits, with different effects across splits. When each POPE split's $1{,}500$ negatives are distributed evenly across all objects with negative queries, BCC increases from $5/10$ to $10/10$ on popular, decreases from $5/54$ to $0/54$ on adversarial, and remains $0/79$ on random. Equal allocation therefore enables whole-cell certification on popular but leaves every object below the required floor on adversarial and random. Figure 2 shows the observed allocations; Appendix A.8 reports the equalisation analysis and its effects on estimation precision.

5.3 The allocation limit recurs across benchmark designs

We examined whether the same allocation limit occurs beyond POPE and DASH-B by scanning six atomic object-presence variants from four additional public benchmark designs: AMBER (Wang et al., 2023), BEAF (Moon et al., 2024), HOPE (Xie et al., 2025), and POPEv2 (Li et al., 2026). The registry and adapter rules were fixed before count extraction, and the included variants cover natural, counterfactual, and mined negatives. Under the same $\alpha = 0.10$, $\delta = 0.05$ two-sided family-wise convention, every variant has $\mathrm{BCC} < 1$, so each contains at least one object whose allocation precludes zero-error certification (Table 6, Appendix A.8). These results show that the allocation limit recurs across benchmark designs, although differences in native ontologies preclude a direct ranking and the nested BEAF and HOPE variants limit the independence of the comparisons.

All allocation limits reported here apply to the stated object-wise Clopper–Pearson bounds and Bonferroni correction under the within-object sampling assumptions in Appendix A.7.

6 When allocation supports certification: an Open Images positive control

To test whether the audit yields class certificates when allocation is sufficient, we evaluated a pre-specified Open Images V7 slice (Kuznetsova et al., 2020; Google, 2022) containing 173 physical-

object classes. After excluding images that overlap with COCO, each class has 100–150 human-verified negatives, exceeding the zero-error floor of 84 under the same two-sided family-wise convention and giving $\mathrm{BCC} = 1.00$. The labels were proposed automatically and verified by human annotators, and the negatives were not mined using any audited model. The class family, sampling caps, prompt template, and analysis were fixed before any model output; we then calibrated τ on a separate validation split and applied the audit to the full test split (Appendix A.12). Open Images scores were collected for six checkpoints before Qwen3.8 and Kimi-VL were added to the main audit.

Under the frozen benchmark labels, each of the six models receives **class-level certificates for** 32–52 **of the** 173 **classes**. Each model also has 12–19 classes with violations and 107–124 unresolved classes, so none receives a whole-cell certificate. All six models select "shirt" as the class with the largest simultaneous lower bound, ranging from 0.54 to 0.69, even though pooled AUROC is 0.86–0.90 and observed pooled test FPR is 0.070–0.091.

A sensitivity analysis that adjusts calibration for assumed label contamination retains some class certificates, with the original test labels unchanged (Appendix A.12). The shared concentration on "shirt" also motivates closer inspection of the reference labels: Section 7.1 examines potential mismatches between those labels and the intended object-presence claims.

7 Statistical resolution is not model attribution

7.1 Reference validity

Interpreting a detected violation as object hallucination requires reference labels that match the intended object-presence question. Applying RePOPE's corrections at the frozen operating points leaves the cross-fitted worst-object false-positive rate above α for all seven POPE-audited checkpoints. The rate increases for the five models in the original valid-readout subset, remains essentially unchanged for Qwen3.8 ($0.368 \rightarrow 0.367$), and decreases slightly for Kimi-VL ($0.318 \rightarrow 0.310$). On POPE-popular, removing four invalid "cup" questions changes the false-assertion counts for InternVL3.5 and Molmo from 19 and 49 out of 147 to 16 and 46 out of 143, respectively. Discordant errors remain unchanged at 3 for InternVL3.5 alone and 33 for Molmo alone, preserving the paired contrast under these reference corrections (Appendix A.6).

In Open Images, inspection of the flagged classes raises concerns about reference semantics. A single inspector conducted a post-hoc review of grids of selected negative examples from each of the 32 classes flagged by at least one of the six models. Under the review rubric, 29 grids were dominated by patterns consistent with sibling-class or ordinary-language mismatches, 3 were mixed, and none was classified as cleanly attributable to model hallucination. Without a second annotator or adjudication, these classifications remain exploratory and do not provide corrected ground truth. The frozen-label audit results are unchanged, and the targeted inspection does not estimate a global contamination rate; the rubric, classification summary, and sensitivity to assumed calibration contamination are reported in Appendix A.12.

7.2 Readout validity

Seven checkpoints pass the POPE readout gate, while Gemma 4 exhibits non-answer behaviour concentrated in particular object groups (Appendices A.10 and A.15). On the full POPE-random gate split, 7.0% of its greedy first tokens fall outside the pre-specified Yes/No token set. The eight worst objects under the forced readout have a mean non-answer share of 38%, compared with 0% for the other 28, and per-object mean answer mass is negatively correlated with empirical false-positive rate (Spearman $\rho = -0.73$, $p < 10^{-4}$).

For the selected object, the forced readout gives $12/24$ false assertions and a simultaneous lower bound of 0.193, whereas restricting evaluation to answered queries gives $0/12$ and a lower bound of 0.000. This restriction changes the evaluated population to queries on which the model answers, so the filtered result cannot replace the estimate for the original query population. The results trigger the pre-specified readout gate on POPE-random (Section 3.4); all three Gemma 4 POPE cells are excluded from the primary matrix. On DASH-B, Gemma 4 passes the same check: 2.7% of nega-

tives have low answer mass, the correlation with object-wise risk is weak ($\rho = +0.09$, $p = 0.63$), and restricting to answered queries leaves the selected object and verdict unchanged.

Consistent with prior observations that first-token probabilities can disagree with generated answers (Wang et al., 2024), parsing Gemma 4's generated responses reduces the POPE-random non-answer share from 7.0% to 0.3%, with rates of 0.3–0.7% across the three POPE splits, and no longer shows the negative mass–risk association. These parsed hard decisions are evaluated at their native operating points and are reported separately from the threshold-calibrated audit matrix. Figure 3 and Appendix A.10 provide the parser comparisons and results for the comparison models.

8 Discussion

Pooled AUROC and FPR summarize overall discrimination and false-positive behaviour, but they do not establish whether every object satisfies a risk budget. Our results show that the same benchmark can support object-level violation claims while lacking the evidence needed for whole-cell certification, because these claims place different demands on the available observations. An explicit unresolved outcome prevents non-detection from being reported as compliance, or failure to certify from being reported as a violation (Lakens, 2017).

When certification is an evaluation goal, benchmark designers should specify the object family, risk budget, confidence level, and simultaneous inference procedure before allocating negatives. The resulting zero-error floor gives a minimum sample requirement for each object, while BCC identifies the fraction of objects whose allocation meets that requirement; observed errors can still prevent certification. The equalisation analysis shows why total sample size and balance alone are insufficient: redistributing the same observations can improve or reduce certifiability depending on whether the resulting object-wise counts meet the family-adjusted floor (Appendix A.8).

Benchmark reports should therefore pair pooled metrics with per-object sample and error counts, simultaneous uncertainty bounds (Snell et al., 2023; Zollo et al., 2024), and an explicit unresolved outcome, adding BCC when certification is a target. Calibration and risk estimation should use separate data (Angelopoulos et al., 2025), with per-query outputs or code that regenerates them available for independent checks.

More observations can narrow uncertainty bounds without resolving a mismatch between reference labels, the behavioural readout, and the intended object-presence question. The exploratory Open Images inspection raises concerns about label semantics, while the Gemma 4 analysis shows that readout applicability can differ across benchmarks for the same model. These findings make reference and readout checks necessary for interpreting statistically resolved errors as model hallucinations.

Our headline results use a nominal budget of $\alpha = 0.10$ and within-cell simultaneous bounds at $\delta = 0.05$ for fixed eligible object families. The exact-binomial guarantees depend on the within-object sampling assumptions in Appendix A.7, and the allocation limits are specific to the chosen inference procedure. Alternative shrinkage or risk-control procedures may improve efficiency under different assumptions or targets (Fogliato et al., 2024; Angelopoulos et al., 2025). The audited checkpoints and benchmarks are not an independent population sample, and the certificates apply to the specified object families, benchmark distributions, and decision rules; they do not extend to arbitrary objects, distribution shifts, free-form generation, or system safety. The audit makes these boundaries explicit by reporting which object-risk claims the available evidence supports and which remain unresolved.

Ethics Statement

This work analyzes outputs from publicly released vision–language models on established public benchmarks. It involves no recruitment of human participants, collection of new personal data, or release of new identifiable data. We use models and datasets under their stated access conditions and report limitations arising from benchmark labels, readout validity, and finite sample allocation. The results characterize object-conditioned benchmark behaviour and should not be interpreted as certifying any audited model for deployment, safety-critical use, or unrestricted object populations.

We did not identify additional ethical risks specific to this study beyond the documented limitations of the models and benchmark data.

Reproducibility Statement

The main text and appendices specify the model checkpoints, benchmark versions, prompts and readouts, eligibility rules, operating-point calibration, data splits, simultaneous bounds, and sensitivity analyses used in this study. The anonymous supplementary code provides environment specifications and scripts for model scoring and the analyses reported in the paper, including the BCC calculations and the Open Images and VCD case studies. Model weights and benchmark data are not redistributed and can be obtained from their official releases at the revisions identified in the code. We cross-checked the analysis code against the saved model outputs and recomputed the key reported counts and results for consistency.

AI Use Statement

Generative AI tools were used to assist with manuscript editing for structure, clarity, and language, and with preparing the supplementary code for release, including code organization, cleanup, comments, and consistency checks. They were not used to formulate the research questions or hypotheses, develop the conceptual framework or statistical methodology, design the experiments, or interpret the results. The authors reviewed all AI-assisted manuscript edits and code changes, verified numerical claims and citations against the underlying analysis artifacts and original sources, and take responsibility for the final text, code, analyses, and conclusions.

References

Douglas G. Altman and J. Martin Bland. Absence of evidence is not evidence of absence. *BMJ*, 311:485, 1995. Statistics Notes series; PMID 7647644.

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners. *The Quarterly Journal of Economics*, 139(1):305–358, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Maximilian Augustin, Yannic Neuhaus, and Matthias Hein. DASH: Detection and assessment of systematic hallucinations of VLMs. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

John J. Cherian and Emmanuel J. Candès. Statistical inference for fairness auditing. *Journal of Machine Learning Research*, 25(149):1–49, 2024. arXiv:2305.03712; code at github.com/jjcherian/fairaudit.

Matt Deitke, Christopher Clark, Sangho Lee, et al. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Tiffany Ding, Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Ryan J. Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.09335.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Antonio Ferrara, Francesco Cozzi, Alan Perotti, André Panisson, and Francesco Bonchi. Size-adaptive hypothesis testing for fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.10586.

Riccardo Fogliato, Pratik Patil, Nil-Jana Akpinar, and Mathew Monfort. Precise model benchmarking with only a few observations. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Gemma Team, Aishwarya Kamath, Johan Ferret, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Google. Open Images Dataset V7: Download. Official dataset release and download documentation, 2022. URL `https://storage.googleapis.com/openimages/web/download_v7.html`. Released October 2022; accessed September 1, 2026.

James A. Hanley and Abby Lippman-Hand. If nothing goes wrong, is everything all right? interpreting zero numerators. *JAMA*, 249(13):1743–1745, 1983.

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

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

Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xAUC metric. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, C. J. Taylor, and Stefano Soatto. THRONE: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27228–27238, 2024.

Kimi Team, Angang Du, Bohong Yin, et al. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128:1956–1981, 2020.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, et al. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Yifan Li, Kun Zhou, Xin Zhao, Lei Fang, and Ji-Rong Wen. Analyzing and mitigating object hallucination: A training bias perspective. In *AAAI Conference on Artificial Intelligence*, 2026.

Joonki Min, Chaeyun Kim, Hyungwook Choi, Yejin Kim, Kihyun Kim, Yohan Jo, and Joonseok Lee. Fine-grained multi image object hallucination benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18295–18305, 2026.

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 (ECCV)*, 2024.

Yannic Neuhaus and Matthias Hein. RePOPE: Impact of annotation errors on the POPE benchmark. *arXiv preprint arXiv:2504.15707*, 2025.

Qwen Team. Qwen3.8-27B. Hugging Face model card, 2026. URL `https://huggingface.co/Qwen/Qwen3.8-27B`. Revision 1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0; accessed September 2, 2026.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Jake C. Snell, Thomas P. Zollo, Zhun Deng, Toniann Pitassi, and Richard Zemel. Quantile risk control: A flexible framework for bounding the probability of high-loss predictions. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2212.13629.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

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*, 2023.

Qingni Wang, Tiantian Geng, Zhiyuan Wang, Teng Wang, Bo Fu, and Feng Zheng. Sample then identify: A general framework for risk control and assessment in multimodal large language models. In *International Conference on Learning Representations (ICLR)*, 2025a.

Qingni Wang, Yue Fan, and Xin Eric Wang. SAFER: Risk-constrained sample-then-filter in large language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL `https://openreview.net/forum?id=kJmLmOvwLC`.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "My answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Rui Xiao, Sanghwan Kim, Yongqin Xian, Zeynep Akata, and Stephan Alaniz. FINER: MLLMs hallucinate under fine-grained negative queries. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36235–36244, 2026.

Ming-Kun Xie, Jia-Hao Xiao, Gang Niu, Lei Feng, Zhiqiang Kou, Min-Ling Zhang, and Masashi Sugiyama. What makes "good" distractors for object hallucination evaluation in large vision-language models? *arXiv preprint arXiv:2508.06530*, 2025.

Hao Yin, Guangzong Si, and Zilei Wang. The mirage of performance gains: Why contrastive decoding fails to mitigate object hallucinations in MLLMs? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. doi: 10.52202/085713-1110.

Thomas P. Zollo, Todd Morrill, Zhun Deng, Jake C. Snell, Toniann Pitassi, and Richard Zemel. Prompt risk control: A rigorous framework for responsible deployment of large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

A APPENDIX

Appendix roadmap. Appendices A.1–A.4 document inherited tools, model specifications, formal statements, and pooled-versus-within-object diagnostics. Appendices A.5–A.11 report the complete census, inferential assumptions, certifiability analyses, and robustness checks. Appendices A.12–A.15 cover the Open Images positive control, VCD and oracle case studies, and reproducibility.

A.1 CREDIT AND DELTA

Table 1: Statistical ingredients inherited from prior work and their role in our benchmark audit.

PRIOR WORK	WHAT WE TAKE	WHAT THIS PAPER ADDS
Cherian & Candès (2024)	flag-and-certify subgroup audit on per-group FPR	exact CP where resampling is uninformative (n_g down to 1); symmetric convention; the census read as a design property
Ferrara et al. (2025)	small-subgroup brittleness, minimum size	floors recomputed at our own convention ($63/53/70$; 69); split-specific allocation prescriptions
Hanley & Lippman-Hand (1983); Altman & Bland (1995)	zero-numerator rule; unresolved $\neq$ safe	matched-convention floors; typed verdicts
Vovk et al. (2005); Ding et al. (2023); Barber et al. (2021)	reference-class dilemma; small-class resolution loss; attainability limits	the dilemma measured (Gini; BCC); scoped correctly (finite partition, counting limit)
Kallus & Zhou (2019)	within/cross AUC decomposition	diagnostic use for claim-evaluation only
Andrews et al. (2024)	winner's-curse framing	conservative unconditional route, no dominance claim
Dror et al. (2018)	paired tests on shared items	selection rule re-run inside the permutation null
Jacobs & Wallach (2021)	measurement validity	a readout gate whose failure correlates with audited groups

A.2 AUDITED CHECKPOINTS AND MODEL STACKS

Table 2: Primary checkpoint specifications from the released configs and model reports (Bai et al., 2025; Qwen Team, 2026; Kimi Team et al., 2025; Wang et al., 2025b; Li et al., 2024; Gemma Team et al., 2025; Gemma Team, 2026; Deitke et al., 2025). For MoE systems, total and active scales are both shown where reported. "Six families" in the text counts released VLM systems; it is not a claim of six independent language-backbone or training-data lineages.

Checkpoint	Scale	Vision stack	Language backbone	Primary coverage
Qwen3-VL-30B-A3B-Instruct	30B / 3B active MoE	Qwen3-VL ViT	Qwen3 MoE	POPE + DASH-B
Qwen3.8-27B	27B dense	Qwen3.8 visual stack	Qwen3.5 hybrid dense	POPE + DASH-B
Kimi-VL-A3B-Instruct	16B / 2.8B active decoder	MoonViT	Kimi MoE	POPE + DASH-B
InternVL3.5-8B-HF	8B dense	InternViT	Qwen3-8B	POPE + DASH-B
LLaVA-OneVision-Qwen2-7B-OV-HF	7B dense	SigLIP	Qwen2-7B	POPE + DASH-B
Gemma-3-12B-IT	12B dense	SigLIP	Gemma-3	POPE + DASH-B
Gemma-4-12B-IT	12B dense	unified patch projection	Gemma-4	DASH-B only
Molmo2-O-7B	7B dense	Molmo/PixMo ViT	OLMo-7B	POPE + DASH-B

A.3 FORMAL STATEMENTS

Score transformations. Strictly increasing score transformations preserve rankings and leave threshold decisions unchanged when the threshold is transformed from τ to $h(\tau)$. The implementation uses NumPy's linear sample quantile with the strict assertion rule $s > \tau$; recomputing this interpolated quantile after a nonlinear transformation can change held-out decisions.

Leave-one-group-out exactness. Fix group g and let τ_{-g} be computed from negatives on images disjoint from g's images. Conditional on τ_{-g}, g's indicators are i.i.d. $\mathrm{Bern}(R_g(\tau_{-g}))$ across its images under the stated benchmark-construction sampling model (one negative per image, asserted in code). This i.i.d. model is an inferential assumption, not a consequence of image distinctness; conditional on it, the CP interval is exact, and marginal coverage follows by integration. The Bonferroni union over g needs no cross-group independence. Empirically the LOIO bound is $\geq$ the in-sample bound in $28/29$ cells; Kimi-VL/popular is the sole decrease ($0.134 \to 0.125$) and remains a violation. The LOIO census establishes $28/29$ violations and no certificates, with pooled FPR at τ_{-g} spanning 0.095–0.142 — a diagnostic, not a budget-holding operating point.

Floors. A zero-error one-sided bound certifies $R_g \leq \alpha$ at level δ' iff $(1-\alpha)^{n_g} \leq \delta'$, i.e. $n_g \geq \ln \delta' / \ln(1-\alpha)$: 29 at δ'=0.05 (loose), and at the paper's two-sided family convention $\delta' = \delta/2G$: $63/53/70$ on POPE, 69 on DASH-B. Under the familywise one-sided analogue $\delta' = \delta/G$, the corresponding floors are $56/46/63/62$.

A.4 Auxiliary pooled-versus-within-object AUROC diagnostic

Across the primary 29 cells, macro within-object AUROC lies within -0.038 to $+0.049$ of pooled AUROC, although within-object pairs are only 1.1%–5.4% of all positive–negative comparisons; we therefore use the decomposition only as a diagnostic and do not attribute the primary result to systematic pooled-AUROC inflation. Across model pairs, agreement on which objects are risky has Spearman range 0.35–1.00 on POPE and 0.18–0.69 on DASH-B.

A.5 The full census

Table 3: All 29 cells at the in-sample operating point: pooled AUROC, bound-selected worst object with its exact count, and the simultaneous two-sided 95% CP lower and max-upper bounds ($\delta/2G$ per tail). Cells with lower bound $>$ α=0.10 are the 27 violations; Gemma-3 adv/ran are the two unresolved cells.

MODEL	SPLIT	AUROC	WORST OBJ	x/n_g	LOWER	MAX-UPPER
Qwen3-VL	adv	0.949	backpack	20/44	0.239	0.683
Qwen3.8	adv	0.946	handbag	21/57	0.192	0.574
Kimi-VL	adv	0.944	backpack	15/44	0.151	0.576
InternVL3.5	adv	0.941	truck	16/61	0.117	0.539
LLaVA-OV	adv	0.948	handbag	16/57	0.126	0.539
Gemma-3	adv	0.911	backpack	11/44	0.090	0.617
Molmo 2-O	adv	0.939	handbag	29/57	0.310	0.705
Qwen3-VL	pop	0.965	bottle	10/34	0.116	0.533
Qwen3.8	pop	0.961	chair	62/363	0.122	0.438
Kimi-VL	pop	0.958	chair	67/363	0.134	0.471
InternVL3.5	pop	0.956	chair	55/363	0.106	0.369
LLaVA-OV	pop	0.961	chair	57/363	0.110	0.471
Gemma-3	pop	0.924	chair	60/363	0.117	0.533
Molmo 2-O	pop	0.954	cup	49/147	0.235	0.648
Qwen3-VL	ran	0.984	cell phone	9/20	0.137	0.795
Qwen3.8	ran	0.981	cell phone	11/20	0.205	0.863
Kimi-VL	ran	0.980	cell phone	8/20	0.108	0.758
InternVL3.5	ran	0.983	cell phone	8/20	0.108	0.758
LLaVA-OV	ran	0.984	orange	10/23	0.143	0.764
Gemma-3	ran	0.941	book	7/23	0.069	0.653
Molmo 2-O	ran	0.968	cell phone	14/20	0.327	0.942
Qwen3-VL	dashb	0.964	shipping box	38/50	0.531	0.916
Qwen3.8	dashb	0.915	watch	35/50	0.467	0.877
Kimi-VL	dashb	0.908	orange	40/50	0.576	0.939
InternVL3.5	dashb	0.870	hanger	20/50	0.196	0.693
LLaVA-OV	dashb	0.900	watch	36/50	0.488	0.890
Gemma-3	dashb	0.875	shipping box	37/50	0.509	0.903
Gemma 4	dashb	0.915	hanger	39/50	0.553	0.928
Molmo 2-O	dashb	0.900	orange	48/50	0.789	0.999

Table 4: Held-out layers, breadth, and utility per cell: split-half bounds in each direction (half the sample), the LOIO diagnostic bound with the pooled FPR at τ_{-g} in parentheses (its disclosure), the number of objects whose simultaneous lower bound clears α (breadth — not a point-estimate count), and pooled TPR at the operating point.

MODEL	SPLIT	HALF A→B	HALF B→A	LOIO (FPR)	#OBJ$>\alpha$	TPR
Qwen3-VL	adv	0.273	0.136	0.277 (0.108)	2	0.861
Qwen3.8	adv	0.154	0.111	0.206 (0.107)	2	0.859
Kimi-VL	adv	0.177	0.034	0.165 (0.106)	3	0.846
InternVL3.5	adv	0.137	0.049	0.129 (0.108)	2	0.845
LLaVA-OV	adv	0.111	0.034	0.126 (0.106)	1	0.869
Gemma-3	adv	0.079	0.069	0.111 (0.105)	0	0.774
Molmo 2-O	adv	0.278	0.217	0.342 (0.113)	3	0.845
Qwen3-VL	pop	0.078	0.086	0.116 (0.105)	1	0.916
Qwen3.8	pop	0.085	0.136	0.136 (0.109)	2	0.896
Kimi-VL	pop	0.105	0.123	0.125 (0.095)	1	0.877
InternVL3.5	pop	0.070	0.103	0.106 (0.100)	1	0.877
LLaVA-OV	pop	0.097	0.098	0.127 (0.115)	1	0.903
Gemma-3	pop	0.113	0.103	0.136 (0.109)	2	0.784
Molmo 2-O	pop	0.169	0.217	0.393 (0.142)	2	0.882
Qwen3-VL	ran	0.136	0.083	0.137 (0.105)	3	0.959
Qwen3.8	ran	0.132	0.111	0.205 (0.106)	3	0.946
Kimi-VL	ran	0.095	0.067	0.123 (0.105)	2	0.945
InternVL3.5	ran	0.083	0.111	0.108 (0.101)	1	0.958
LLaVA-OV	ran	0.048	0.151	0.143 (0.105)	1	0.957
Gemma-3	ran	0.014	0.111	0.092 (0.105)	0	0.844
Molmo 2-O	ran	0.193	0.193	0.327 (0.109)	3	0.909
Qwen3-VL	dashb	0.469	0.386	0.623 (0.128)	3	0.902
Qwen3.8	dashb	0.329	0.443	0.599 (0.127)	2	0.777
Kimi-VL	dashb	0.513	0.464	0.700 (0.130)	3	0.790
InternVL3.5	dashb	0.213	0.070	0.212 (0.107)	2	0.657
LLaVA-OV	dashb	0.399	0.366	0.700 (0.130)	4	0.702
Gemma-3	dashb	0.514	0.312	0.509 (0.122)	3	0.657
Gemma 4	dashb	0.513	0.353	0.576 (0.126)	2	0.778
Molmo 2-O	dashb	0.773	0.575	0.866 (0.133)	3	0.673

Each direction yields $20/29$ violations; 15 cells violate in both directions and 25 in at least one.

A.6 Nulls, multiplicity, and the claims ledger

The *cluster null* holds each image's false-positive count fixed and reassigns which objects they land on (preserving image-level error clustering; degenerate on DASH-B, which poses one object per image); the *label-permutation null* permutes object labels of negatives (preserving group sizes); the per-split *omnibus* shares one permutation across a split's models. Monte-Carlo seeds are deterministic per (model, split) via CRC32; concentration P values use $B = 100{,}000$. Under a single matrix-wide family of all 692 eligible tests, $\delta/692$ two-sided, 18 of 29 cells still violate. Across the 21 POPE cells, 17 concentration tests remain significant under Benjamini–Hochberg q=0.05; the exceptions are Gemma-3/random, InternVL/popular, LLaVA/popular, and Qwen3.8/popular with raw $P_{q90} = 0.05485$. The seven-model joint omnibus gives $P = 1/20{,}001$ on each split. The DASH-B label-permutation q_{90} test rejects in $7/8$ cells, with Kimi-VL the exception ($P = 0.0871$), while its maximum statistic rejects in $8/8$.

Model-wide label correction. On RePOPE's relabelled adversarial split, each of the five original models with a valid POPE readout has a higher cross-fitted worst-object rate than under the original labels (0.263–0.509 $\rightarrow$ 0.276–0.571); the pre-specified Qwen3-VL/backpack cell moves $0.432 \rightarrow 0.500$. Qwen3.8's selected handbag rate is essentially unchanged ($0.368 \rightarrow 0.367$), while Kimi-VL's worst rate decreases $0.318 \rightarrow 0.310$ and its carrier changes from backpack to bowl; all seven current rows remain above the budget. Thus five rates rise and two do not. These are rates at the frozen operating point, not a new benchmark-wide familywise claim.

Table 5: Claims ledger. "Named carrier" tests describe the selected objects; the selection-adjusted inferential claim remains at the model-pair level.

CLAIM	UNIT	ROLE	SELECTION HANDLING	MULTIPLICITY
Cell violation	eligible objects in one cell	primary	simultaneous lower bounds	within-cell G
Whole-cell certificate	eligible objects in one cell	primary	simultaneous upper bounds	within-cell G
Focal near-tied pair	split–model pair	supporting	object selection rerun in every draw	Holm over 37 screened pairs
All-pairs check	split–model pair	sensitivity	same selection-aware permutation	Holm over all 63 pairs
Named carrier contrast	selected object	descriptive	no independent selection adjustment	Bonferroni over 58 tests
Concentration	model–split cell	supporting	dependence-preserving nulls	BH over 21 POPE cells; omnibus
Bootstrap-tightened calibration	split direction	sensitivity	threshold fixed before counting	within-cell G
BCC	object allocation	design diagnostic	zero-error counterfactual; no model selection	family-specific floor

Selection-aware paired null. For each of the 37 screened split–model-pair combinations, every randomisation independently swaps the two models' full false-assertion vectors within each image, preserving dependence among that image's queries. It then reruns the complete object-selection rule on both pseudo-models, takes the union of selected objects, and records the minimum McNemar p-value. The randomisation-tail probability is Holm-corrected over pairs ($B = 20{,}000$; floor 5×10^{-5}); a second analysis reruns the same procedure for all 63 split–pair combinations. Five of 37 screened pairs and eight of all 63 survive their respective Holm corrections; four screened survivors also contain a contrast passing Bonferroni over the 58 selected-object tests. The complementary InternVL–Molmo pair remains at the Monte Carlo floor. The Kimi-VL/Molmo popular pair, from a separately developed VLM family, survives both Holm families ($p = 2/20{,}001$), and its "cup" contrast (20 versus 49; 8:37 discordance; named-object $p = 1.54 \times 10^{-5}$) clears the $0.05/58$ Bonferroni threshold. Conversely, the Qwen3-VL/LLaVA-OV "backpack" pair is descriptive only: its selection-aware $p = 0.00235$ misses both expanded Holm families, and its named contrast misses the $0.05/58$ threshold.

InternVL exhibit under corrected labels. The model-wide sensitivity above uses RePOPE-adversarial. RePOPE's popular labels are used only for this separately fixed-carrier "cup" recount. RePOPE removes 4 of the 147 "cup" negatives as invalid questions and relabels *none* of them as present; on the 143 surviving negatives the exhibit is unchanged — 16 vs 46 (11.2% vs 32.2%), the same 33:3 discordant split, descriptive named-object McNemar $p = 2.3 \times 10^{-7}$. The selection-adjusted claim remains the pair-level test of Section 4.1. The 19-vs-49 contrast is therefore not an artifact of POPE's known annotation errors.

A.7 Independence and clustering

Within an eligible object group, duplicate (image, object) pairs among negatives: 0 (random), 0 (popular), 1 (adversarial, in a non-eligible group). POPE poses three negatives per image (500 images per split); DASH-B poses one object per image (1341 distinct images). The exact-binomial assumption operates within a group across distinct images under the stated benchmark-construction sampling model; distinctness eliminates repeated pairs but does not prove independence. The cluster null operates across objects within an image, so the two procedures govern different axes.

DASH-B's 1,341 negatives were mined against PaliGemma and two LLaVA-NeXT models, retained only when the failure transferred to Qwen2-VL and Llama 3.2-VL, and human-verified by two-labeler consensus. Thus LLaVA-OV, both Qwen models, and the Gemma models overlap the mining or transfer lineages; Molmo 2-O and Kimi-VL are non-target lineages. DASH-B is therefore a stress test, not an independent benchmark replication.

A.8 Certification and resolution derivations

Floors as in Appendix A.3; below even the loose unadjusted one-sided exact diagnostic floor of 29 sit 74%/40%/99% of groups (adv/pop/ran). At the six-checkpoint freeze, pooled AUROC 95% half-widths are $\pm 0.009/0.008/0.006$ while per-object rate half-widths span $\pm 0.15/0.05/0.20$ — uncertainty ratios $16.2\times$ /$6.5\times$ /$34.4\times$, worst on the split with the highest pooled AUROC. Negative-allocation Gini: $0.71/0.57/0.13$. Equalisation counterfactual at fixed 1,500 negatives, with each all-group family recomputed under the paper's two-sided Bonferroni convention: popular $1.29\times$ median-group estimation precision with BCC $5/10 \to 10/10$; adversarial $1.96\times$ median-group es-

timation precision with $5/54 \to 0/54$; random essentially unchanged ($0.97\times$; $0/79 \to 0/79$). BCC as in Section 5; every split carries at least one eligible object below its floor, so no cell could certify under a zero-error run. The model-output census is less stark at the object level: under each cell's own simultaneous family, $58/692$ eligible cell–object groups are violations, $21/692$ are certificates, and $613/692$ are unresolved. The certificates occur in $13/29$ cells (adversarial: 10; popular: 11; random and DASH-B: 0), but no cell certifies its entire eligible family. DASH-B's 1,341 negatives span 70 objects with count min/median/max $3/11.5/50$; the 34 eligible objects span $15/29.5/50$, and only 11 groups have $n_g = 50$. Hence the displayed worst groups at 50 do not represent the allocation as a whole; no eligible group reaches the matched two-sided zero-error floor of 69.

Table 6: Frozen cross-benchmark allocation scan. A negative trial is one unique provenance-cluster–native-object pair; $N_- = \sum_g n_g$. Floors and BCC use $\alpha = 0.10$, $\delta = 0.05$, two-sided Clopper–Pearson bounds, and within-variant Bonferroni correction over every native negative-bearing group. Native inventories are not harmonised; BEAF's two arms and HOPE's two mining modes are nested variants and are not pooled or counted as independent families.

Variant	Construction	G	N_-	n_g min/q1/med/q3/max	Gini	Floor	BCC (K/G)
AMBER existence	natural	188	4,855	1/1/3.5/18.25/538	.787	85	17/188
BEAF original distractor	natural	79	3,675	12/19/24/40.5/413	.503	77	8/79
BEAF removed target	counterfactual	79	1,738	1/7.5/15/23/309	.508	77	1/79
HOPE content-CLIP	mined	357	12,000	1/7/17/43/416	.585	91	31/357
HOPE union	mined	359	36,000	1/20/48/115.5/1,234	.624	91	106/359
POPEv2 counterfactual	counterfactual	80	500	3/5/6/7/12	.137	77	0/80

This bounded scan follows fixed registry and adapter rules but is not an exhaustive survey. The below-one pattern persists for every variant across the one- and two-sided $\alpha \in \{0.05, 0.10, 0.20\}$ grids and under the raw-row optimistic diagnostic. It also persists in every nonempty family recomputed after filtering to $n_g \geq 20$ (POPEv2 has no qualifying group), and among the three archival variants (BEAF's two arms and POPEv2). AMBER and HOPE are preprints; BEAF and POPEv2 are archival. As a descriptive check rather than a simultaneous certificate, the median n_g exceeds the loose unadjusted exact floor of 29 only for HOPE-union among the six scan variants. AMBER, BEAF, and POPEv2 have partial COCO lineage, whereas HOPE uses Objects365; these are four benchmark designs, not statistically independent datasets. The AMBER adapter excludes 63 scene–object keys with conflicting native annotations. PhD was excluded because its native targets are non-atomic strings, MME because the available metadata slice could not be shown equivalent to the author release, and Causal-HalBench because its annotations could not be pinned separately. These values diagnose convention-conditional zero-error allocation attainability, not model performance, label validity, benchmark quality, or information-theoretic impossibility.

A.9 Robustness details

Uncalibrated hard-decision census (pre-specified). Assert iff the model's own first-token yes/no mass exceeds 0.5 — no calibration; where a generated-answer parser exists this decision agrees with it on 97.8–100% of queries (Kimi-VL: exactly). Same eligible families, two-sided $\delta/2G$ convention. The worst object's simultaneous lower bound exceeds the cell's own pooled uncalibrated FPR in $22/29$ cells and the illustrative 0.10 budget in $17/29$; the closest strict case is LLaVA-OV/popular (0.034415 versus pooled FPR 0.034000), which rounds to a tie in the table. The q_{90} exceeds the pooled rate in $28/29$. On the identical 147 popular "cup" negatives the uncalibrated counts are InternVL 34, Kimi-VL 14, Molmo 31 (discordance 16:13 and 13:30 against Molmo): the InternVL–Molmo contrast is specific to the matched budget, while the Kimi–Molmo split persists uncalibrated, at near-tied accuracy.

Qwen successor comparison. From Qwen3-VL to Qwen3.8, the four-setting held-out tally rises from $3/4$ to $4/4$ in at least one split direction and from $2/4$ to $3/4$ in both. Each checkpoint nevertheless has nine full-data violating object groups across the same 94 cell–object groups, and the bound-selected carrier changes in three of four settings. Qwen3.8's pooled AUROC is lower in all four settings even as held-out violation resolution increases. This is a within-family successor comparison, not a scale law: it shows that pooled ordering, resolution, and the location of concentrated risk need not move together.

Table 7: All 29 cells at the uncalibrated hard decision ($p > 0.5$): pooled uncalibrated FPR, bound-selected worst object with exact count, simultaneous lower bound, and q_{90}.

MODEL	SPLIT	FPR	WORST OBJ	x/n_g	LOWER	q_{90}
Qwen3-VL	adv	0.069	handbag	16/57	0.126	0.185
Qwen3-VL	pop	0.034	chair	18/363	0.024	0.048
Qwen3-VL	ran	0.012	book	2/23	0.002	0.040
Qwen3-VL	DASH-B	0.329	shipping box	48/50	0.789	0.556
InternVL3.5	adv	0.226	handbag	31/57	0.342	0.460
InternVL3.5	pop	0.161	chair	90/363	0.191	0.271
InternVL3.5	ran	0.045	book	5/23	0.032	0.086
InternVL3.5	DASH-B	0.592	parallel bars	45/50	0.700	0.894
LLaVA-OV	adv	0.059	chair	10/78	0.041	0.124
LLaVA-OV	pop	0.034	chair	23/363	0.034	0.061
LLaVA-OV	ran	0.011	book	2/23	0.002	0.019
LLaVA-OV	DASH-B	0.688	orange	50/50	0.866	0.968
Gemma-3	adv	0.231	book	15/26	0.279	0.467
Gemma-3	pop	0.199	bottle	21/34	0.380	0.479
Gemma-3	ran	0.134	book	8/23	0.092	0.280
Gemma-3	DASH-B	0.742	hanger	48/50	0.789	0.941
Molmo 2-O	adv	0.114	handbag	31/57	0.342	0.363
Molmo 2-O	pop	0.067	cup	31/147	0.130	0.223
Molmo 2-O	ran	0.034	cell phone	7/20	0.081	0.086
Molmo 2-O	DASH-B	0.410	orange	50/50	0.866	0.758
Qwen3.8	adv	0.066	handbag	17/57	0.138	0.176
Qwen3.8	pop	0.035	chair	25/363	0.039	0.064
Qwen3.8	ran	0.011	book	4/23	0.018	0.000
Qwen3.8	DASH-B	0.511	watch	50/50	0.866	0.911
Kimi-VL	adv	0.094	backpack	14/44	0.135	0.240
Kimi-VL	pop	0.061	chair	41/363	0.073	0.115
Kimi-VL	ran	0.017	bowl	2/21	0.002	0.047
Kimi-VL	DASH-B	0.585	orange	50/50	0.866	0.914
Gemma 4	DASH-B	0.342	watch	46/50	0.728	0.650

The Qwen3.8/random $q_{90} = 0$ uses the same 36-object family as every random cell. Only three eligible objects have any native false positive (six errors total among eligible groups); the other ten native false positives fall in groups below the $n_g \geq 20$ cutoff, so at least 90% of the eligible rates are zero.

Operating-point sweep: q_{90} remains above pooled FPR at all 15 grid points in $26/29$ cells; the LLaVA-OV/DASH-B and Molmo/DASH-B rows miss only at $\alpha = .02$, while Kimi-VL/DASH-B misses at $\alpha = .02, .04, .06$ and crosses above thereafter. Prompt/readout battery (Qwen3-VL/POPE-adversarial): three prompt templates crossed with four variants — normalised Yes/No probability, its rank-equivalent raw log-odds, full-vocabulary Yes probability, and the native hard decision — produce 12 q_{90} tests; 11 survive BH correction, with only one hard-decision prompt non-significant ($p = 0.072$). Eligibility-cutoff sweep over $\{15, 20, 25, 30\}$: the multiplicity-corrected worst object is cutoff-robust where significant (e.g. Qwen backpack $P_{\max} = 2 \times 10^{-4}$ at every cutoff), while the 90th-percentile per-object rate q_{90} is cutoff-sensitive by construction (a fixed percentile over a variable family). Text-only control on the six-checkpoint subset: pooled AUROC is 0.35–0.48 without the image and 0.87–0.98 with it. This control excludes Qwen3.8 and Kimi-VL and is not a 29-cell denominator.

A.10 Readout validity in full

In the 150-query adversarial preflight, Gemma 4 places its greedy first token on yes/no in 91.3% of queries; the five comparison models do so in 100%. Parser sensitivity at Gemma 4's natural operating point: pooled parsed-yes on negatives $7.4\%/6.1\%/1.6\%$ (adv/pop/ran), worst-object bounds $0.083/0.097/0.002$ — unresolved, and not matrix-comparable (a parser has no threshold to cali-

brate). All six tokenizers in that battery encode the frozen Yes/No surfaces as single tokens, so sequence likelihood coincides with the first-token readout.

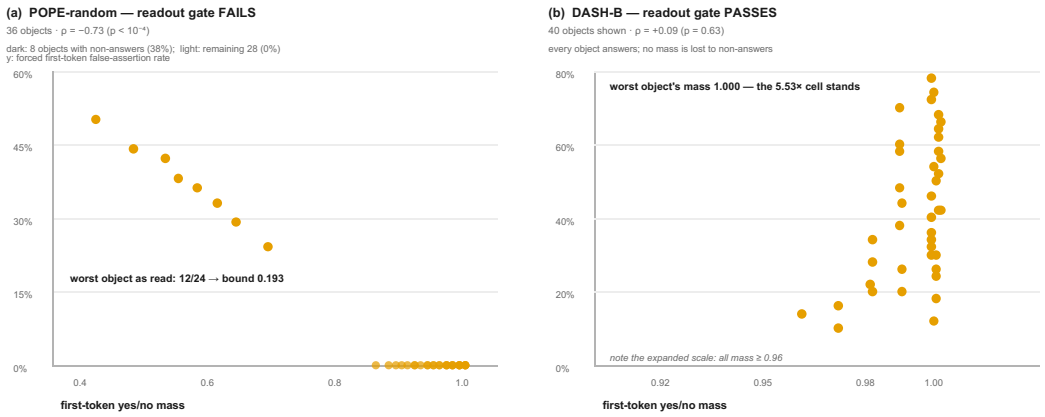

Figure 3: Readout validity is model-specific and detectable. (a) On POPE-random (the gate split), Gemma 4's first-token yes/no mass anti-correlates with per-object rate ($\rho = -0.73$): forced and answered-only reads give opposite empirical readings, neither a valid matrix verdict, so no number is reported. (b) The identical pre-specified check passes on DASH-B ($\rho = +0.09$), so that cell stands. Bottom: on POPE-random the non-answer share is 7.0% under the first-token read and 0.3% under generated-answer parsing (parser range 0.3–0.7% across the three POPE splits), locating the discrepancy in the forced readout.

A.11 Tightening the budget under calibration uncertainty

The held-out layer sets τ at the calibration half's sample $(1 - \alpha)$ quantile. This targets, rather than guarantees, pooled FPR α on the independent counting half; realised test-half FPR exceeds α in 29 of 58 direction-cells. The pre-specified bootstrap-tightened calibration sensitivity instead takes the largest k whose *image-clustered bootstrap* upper bound on the pooled rate still falls under α and puts τ so exactly k calibration negatives exceed it; everything downstream (family, $\delta/2G$, both directions) is unchanged. Clustering matters in principle here — each POPE image carries three negative queries — but has little aggregate effect: the measured design effect has median 1.05 (range 0.91–1.41), because an image's three queries concern three *different* objects and it is object identity, not image identity, that drives false assertion.

The median selected calibration fraction moves $0.101 \to 0.082$, overshoot falls to 8 of 58, and the cluster-aware tally becomes $16/29$ cells for A→B and $15/29$ for B→A (11 in both), with surviving bounds 0.105–0.706; a query-level binomial alternative gives $16/29$ and $14/29$ (11 in both). Nine direction-cells drop out, including marginal cases such as InternVL/adversarial $0.137 \to 0.095$. Across all 58 directions, seventeen retain the same selected carrier, count, and bound; thirteen of these remain violations. This tally is a different quantity from the matrix-wide single-family tally of Section 4.2, which now reads 18 of 29.

A.12 The Open Images external audit

Pre-specified design. The audit includes 173 physical-object classes with ≥ 100 verified test negatives and ≥ 30 verified validation negatives each (COCO-overlap images removed; person/body-part classes excluded, with the inclusive G=190 read as a sensitivity); per class, up to 150 test negatives and up to 150 positives (the realised negative range is 100–150) in deterministic CRC32 order; τ = the 0.9 quantile of a 10,000-query pooled validation-negative set; questions from a fixed head-noun template; readout, budget, and convention identical to Section 3.2. The quantile uses NumPy's linear rule and assertion is strict ($s > \tau$), so scores tied exactly at τ are not asserted. Zero images failed to download; zero queries were dropped. This auxiliary covers the six checkpoints for which OI scores were collected; Qwen3.8 and Kimi-VL are not included.

Table 8: Alternative-readout battery on the six-checkpoint subset: 21 primary cells plus three Gemma 4 POPE diagnostic cells. The table reports the share of generated answers that are non-answers, query-level agreement between parsed and first-token decisions (answered queries), and Spearman between per-object rates under the two readouts. It is not part of the 29-cell primary tally. Low ρ values occur at small G (popular has 6 objects) and differing operating points; they are descriptive.

MODEL	SPLIT	NON-ANS	AGREE	ρ
Qwen3-VL	adv	0.0000	0.998	0.97
InternVL3.5	adv	0.0000	0.997	0.83
LLaVA-OV	adv	0.0000	0.994	0.83
Gemma-3	adv	0.0000	1.000	0.84
Molmo 2-O	adv	0.0000	1.000	0.97
Qwen3-VL	pop	0.0000	0.997	0.14
InternVL3.5	pop	0.0000	0.997	0.94
LLaVA-OV	pop	0.0000	0.995	1.00
Gemma-3	pop	0.0000	1.000	0.94
Molmo 2-O	pop	0.0000	1.000	0.83
Qwen3-VL	ran	0.0000	0.998	0.46
InternVL3.5	ran	0.0000	0.998	0.58
LLaVA-OV	ran	0.0000	0.996	0.44
Gemma-3	ran	0.0000	1.000	0.90
Molmo 2-O	ran	0.0000	1.000	0.68
Qwen3-VL	dashb	0.0000	0.996	0.63
InternVL3.5	dashb	0.0000	0.986	0.73
LLaVA-OV	dashb	0.0000	0.978	0.44
Gemma-3	dashb	0.0000	0.999	0.39
Gemma 4	dashb	0.0000	0.994	0.71
Molmo 2-O	dashb	0.0000	1.000	0.68
Gemma 4*	adv	0.0053	0.976	0.83
Gemma 4*	pop	0.0070	0.975	0.66
Gemma 4*	ran	0.0030	0.988	0.42

Table 9: Open Images census for the six-model OI positive-control subset, at the held-out validation threshold. "Worst" is the bound-selected worst class with its simultaneous lower bound; the worst class is *shirt* for all six models. Verdicts at $\alpha = 0.10$, $\delta/2G$ two-sided, $G = 173$.

MODEL	AUROC	FPR	TPR	VIOL	CERT	UNRES	WORST (shirt)
InternVL3.5	0.888	0.084	0.636	15	36	122	124/150, 0.693
Molmo 2-O	0.894	0.075	0.645	14	52	107	118/150, 0.647
Qwen3-VL	0.901	0.081	0.664	13	36	124	119/150, 0.654
LLaVA-OV	0.897	0.091	0.676	19	32	122	103/150, 0.538
Gemma-3	0.857	0.077	0.570	17	46	110	107/150, 0.566
Gemma 4	0.876	0.070	0.586	12	51	110	106/150, 0.559

Readout gates. Four models have no non-answer tail at all and positive or null mass–rate correlations. Qwen3-VL's correlation is negative ($\rho = -0.224$, $p = 0.003$) but its minimum yes/no mass over all 25,950 test negatives is 0.998 — no non-answer tail exists, the answered-only restriction is vacuous, and certificates survive its exclusion — so the cell remains valid under the gate. Gemma 4's non-answer tail *recurs* (7.3% of negatives, minimum mass 10^{-8}) but is uncorrelated with the audited classes ($\rho = +0.17$; no flip), so it passes under the frozen gate conjunction: with POPE ($\rho = -0.73$, fail) and DASH-B (pass), the criterion demonstrably discriminates the correlated pathology, not the model. Gemma 4's τ sits in a saturated score region (top-of-range mass), disclosed here; the census uses counts, not margins.

The exploratory label-validity review was performed by a single inspector after seeing the results; it is post-hoc and does not modify the pre-specified census. Selection rule: for each class whose simultaneous lower bound exceeded α for any model (32 classes over the six-model union), a grid of

negatives false-asserted by consensus of models (CRC32 order, 6–9 images) was visually reviewed, with never-asserted controls for the worst class. Under the recorded rubric, the inspector coded 29 classes as dominated by patterns consistent with label/ontology artifacts, 3 as mixed, and 0 as clean hallucination. No second annotator or adjudication was used; these categories are qualitative diagnostic evidence, not corrected ground truth. The same inspector's review of 27 random above-τ validation negatives found $\sim$65% patterns consistent with label artifacts and $\sim$35% apparent genuine false assertions (a kitchen prompting "sink", balloons as "ball") — genuine over-assertion exists but is *diffuse*, while patterns consistent with label artifacts are *class-concentrated*, explaining why the per-class census preferentially flags the latter.

Certificates under calibration contamination. Under the sensitivity model, mislabeled positives have assertion probability no lower than true negatives, so conditional on τ they inflate a class's measured rate in expectation. The threshold need not be conservative because the validation pool shares the labels. Recomputing τ at deflated quantiles $(1-m)(1-\alpha)$: certificates per model go from 32–52 (m=0) to 22–39 (m=0.02), 9–26 (m=0.05), and 2–10 (m=0.10); the targeted tail inspection does not estimate a global m, because non-tail negatives have zero inclusion probability. The sweep is therefore indexed by assumed m and is not a corrected-label certificate.

A.13 Aggregate gain and false-positive risk in a published mitigation

This case study is auxiliary and is not used to support the primary prevalence, certification, or benchmark-allocation claims.

POPE splits are balanced, so any published (precision, recall) pair determines the false-positive rate, $\mathrm{FPR} = R(1-P)/P$, cross-checkable through accuracy as $1-(2\cdot\mathrm{Acc}-R)$ (the two agree within rounding, <0.7pp, on every cell below). Applying this to the VCD paper's own POPE table (Leng et al., 2024):

Table 10: Derived from the published VCD table (arXiv:2311.16922v1, Table 1, MS-COCO): F1 rises on all nine cells while the false-positive rate rises on all six LLaVA-1.5 and Qwen-VL cells and falls on all three InstructBLIP cells. A pooled F1 gain is sign-blind to the false-positive-risk movement. This is the pooled-level observation only; the object-level acceptance test at a matched operating point is the protocol of Section 3.2.

MODEL	SPLIT	F1 REG→VCD	FPR REG	FPR VCD	ΔFPR
LLaVA-1.5	random	81.3→87.2	6.2%	7.8%	+1.6
LLaVA-1.5	popular	80.1→85.1	9.1%	12.5%	+3.4
LLaVA-1.5	adversarial	77.6→81.3	14.8%	21.5%	+6.7
Qwen-VL	random	82.7→87.8	3.3%	4.6%	+1.3
Qwen-VL	popular	82.1→86.4	4.4%	7.6%	+3.2
Qwen-VL	adversarial	80.4→83.9	8.1%	13.5%	+5.4
InstructBLIP	random	80.4→83.7	17.8%	10.3%	−7.5
InstructBLIP	popular	78.4→81.1	22.4%	16.4%	−6.0
InstructBLIP	adversarial	76.6→79.5	27.3%	20.3%	−7.0

Reproduction at the paper's POPE setting. We reran the released code and weights at the configuration the paper specifies for POPE (α=1, β=0.1, diffusion step T=999), five seeds. Both rows land inside seed noise: regular accuracy 83.39 ± 0.37 and F1 81.48 ± 0.45 against the published 83.29/81.33; VCD accuracy 87.79 ± 0.33 and F1 87.24 ± 0.33 against 87.73/87.16. The precision movement reproduces in direction and size ($92.03 \to 91.35$ against $92.13 \to 91.42$), as does the ten-point recall gain, and so therefore does the risk: the derived pooled FPR rises $6.33\% \to 7.91\%$ against $6.22\% \to 7.82\%$ published. The reproduction therefore confirms that, on this model, a 5.8-point F1 gain arrives with a 1.6-point rise in exactly the rate a deployment budget constrains.

The released code defaults to diffusion step T=500, which the paper specifies for its *other* benchmarks. At that default the published POPE row does not reproduce (accuracy 85.1–85.8) and the risk direction reverses (derived FPR $6.33\% \to 4.8$–5.8%), so we use the paper-specified T=999. A neutral run (α=β=0) through the same code path reproduces the regular row answer-for-answer ($0/3000$ differ), and the contrast scales monotonically in α.

Nothing here evaluates VCD as a method: it improves the aggregate metrics it reports, both in its table and in our reproduction. The narrow claim we rely on is that an aggregate gain does not carry the sign of the false-positive-risk movement — positive on the two models above, negative on InstructBLIP — and that no per-object counts are published that would say which objects absorb the additional false assertions.

At a matched budget. To separate operating-point movement from ranking quality, both arms are calibrated separately to $\alpha = 0.10$ and read deterministically, the contrasted arm from the contrasted logits at three query-keyed noise seeds. This is a *ranking-level* audit; the sampled decoding above remains the record of deployed behaviour. At pooled FPR 0.099–0.100 the contrast buys no discrimination on any split: pooled TPR falls $0.892 \to 0.868$, $0.845 \to 0.827$, $0.770 \to 0.746$ and pooled AUROC falls $0.956 \to 0.941$, $0.943 \to 0.928$, $0.914 \to 0.898$ (random / popular / adversarial). The published sampled-decoding gain therefore arrives without improved discrimination under this readout — consistent with a more assertive operating point rather than a better ranking of claims. We do not claim to identify the sampler's mechanism: the two readouts differ, though they are linked — at temperature 1 the released sampler draws its first token from exactly the contrasted distribution whose yes/no mass this readout scores.

Per object at that matched budget, the *descriptive* census moves: on adversarial the bound-selected worst object shifts from "truck" (0.141) to "bottle" (0.171–0.188) in all three seeds, most objects tightening (breadth $2 \to 1.3$, certificates $2 \to 2.7$) while one worsens. We type that strictly: it does *not* reproduce at the held-out layer, where the two calibration directions disagree with each other and with the full-data census, and no paired McNemar comparison on the identical negatives reaches significance ($p \geq 0.062$; the direction is consistent across seeds). This is the paper's own resolution argument applied to itself — at 47–61 negatives per object a half-sample comparison cannot resolve a difference this size — which is why the matched-budget conclusion we rely on is the pooled one. On popular the worst object is unchanged; on random both arms are *unresolved* (breadth 0), the verdict the typology exists to express. The plausibility mask fires on 61–71% of queries but leaves the budget reachable (top atom $\leq 0.9\%$ of negatives), so the officially masked arm is auditable too and agrees within seed spread.

The opposite-direction model. We also reproduce InstructBLIP because its published table moves the other way (derived FPR $17.8\% \to 10.3\%$). At the paper-stated setting, the reproduction matches the published row: regular accuracy is 81.28 ± 0.68 against 80.71, VCD accuracy is 84.74 ± 0.65 against 84.53, and the pooled sampled rate falls $16.8\% \to 10.2\%$ against the derived $17.8\% \to 10.3\%$. The released script's actual behaviour (its `--cd_alpha` is declared but never forwarded, so the contrast runs at the sampler's internal fallback α=0.5, at the T=500 default) misses the published row by 1.3–1.7 points: a second instance of the release-versus-paper gap, resolved the same way. At the matched budget the LLaVA-1.5 conclusion repeats: pooled TPR falls $0.865 \to 0.832$, $0.803 \to 0.776$, $0.754 \to 0.745$ and pooled AUROC falls $0.950 \to 0.923$, $0.927 \to 0.901$, $0.907 \to 0.888$, with the worst object unchanged on two splits and no bound rising. So in all six model×split cells — on the model whose published false-assertion rate rises with the gain *and* on the one whose rate falls — the contrasted ranking buys no discrimination at a matched budget; what the aggregate gain reflects in both cases is consistent with where the operating point sits, in opposite directions, on a slightly weaker ranking.

A.14 The oracle threshold's decision cost

An ex-post per-group threshold (each group held at $\approx 10\%$ on its own test negatives) removes almost all worst-group excess at a recall change between a 2.3-point loss and a 13.7-point gain, with a gain in 20 of 29 cells (seven losses and two unchanged). It is fit on test labels, hence not deployable, and per-group thresholds are not certifiable at these counts; it demonstrates only that the disparity is not an inherent accuracy–utility trade-off.

A.15 Reproducibility

All runs use batch size 1 because BF16 numerical differences at larger batch sizes affected some small-count analyses. Monte Carlo seeds are derived deterministically from the CRC32 hash of each cell identifier. The anonymous supplement provides environment specifications, model and dataset identifiers, and scripts for scoring and analysis, including the verdict census and per-split

BCC. Model weights, benchmark data, cached outputs, and generated results are not included; the supplied scripts regenerate the outputs and results from the official releases.

The added checkpoints are `Qwen/Qwen3.8-27B` at revision `1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0` and `moonshotai/Kimi-VL-A3B-Instruct` at revision `398eede0903cd983a2bfa0cc634e9ac1d843f375`. Both use BF16, one greedy step, and the common six-surface readout. Qwen3.8 uses its released non-thinking chat template, while Kimi-VL uses pinned remote code and eager attention. Under the pinned eager MoonViT implementation, Kimi-VL execution at larger batch sizes is not equivalent to batch-1 execution and was not used.

Both checkpoints were added under admission criteria independent of their risk results. The seven checks covered finite logits for single-token answer surfaces, non-degenerate scores, an image-over-text visual gap, answer-token and answer-mass validity, exact repeat determinism, agreement between one-step scoring and generation logits at batch size 1, and an active image path. Kimi-VL passed these checks before the primary analysis; its score-uniqueness fraction was $135/150 = 0.900$, equal to the admission threshold. Before these additions, the matrix yielded $14/21$ held-out violations in each direction and $19/21$ violations in the descriptive census.